# PRIVILEGED SOLUTIONS OR CONTEXT-INDUCED TEACHER BEHAVIOR? DISSECTING ON-POLICY SELF-DISTILLATION

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

## ABSTRACT

On-Policy Self-Distillation (OPSD) is commonly interpreted as the transfer of privileged information: a teacher observes the verified solution to the target problem and supervises the student's trajectory. However, this interpretation conflates two effects. The reference solution not only reveals the answer to the current instance but also changes the context under which the teacher provides token-level supervision. We investigate the role of target-specific privilege with OP$^2$SD (***O**n-**P**olicy **S**elf-**D**istillation from Other **P**roblems*), which replaces the paired reference with a problem and solution from a different example, while preserving the student rollout, teacher, and distillation objective. Across three models and three mathematics benchmarks, OP$^2$SD improves over the base model, remains competitive with OPSD. The success of OP$^2$SD implies that OPSD gains do not necessarily come from access to the reference solution, and that the teacher's context-induced behavior is an important factor.

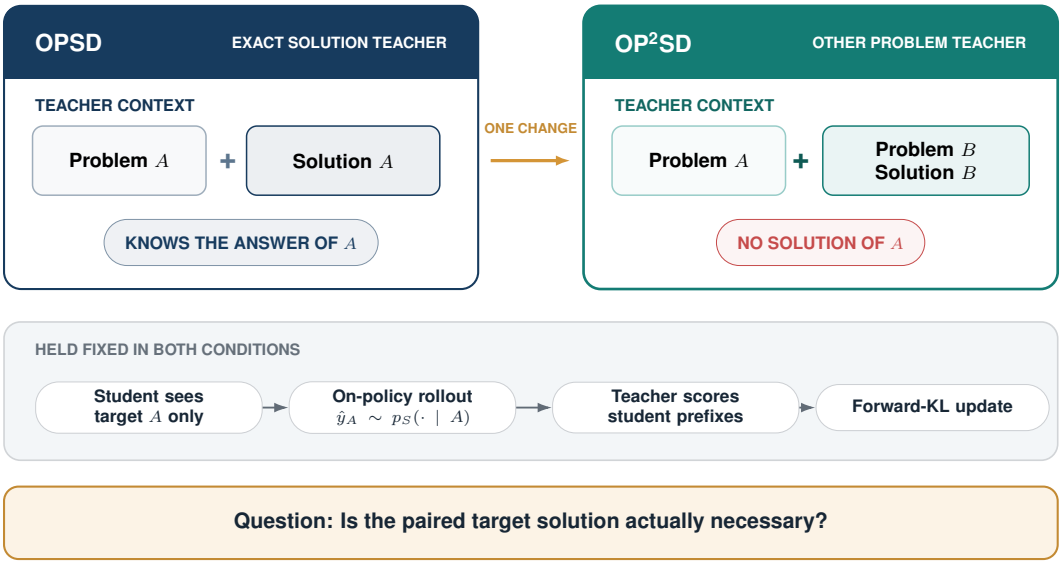


Figure 1: OPSD and OP$^2$SD share the student prompt, on-policy rollout, frozen self-teacher fixed to the base model, and loss. The only difference is the teacher-only context: OPSD supplies the verified solution to the target problem $A$, whereas OP$^2$SD supplies a worked solution to a different problem $B \neq A$. In this way, OP$^2$SD withholds the target solution while retaining the teacher-only worked-solution context.

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

# 1 INTRODUCTION

On-policy distillation (OPD) (Gu et al., 2024; Agarwal et al., 2024) has recently emerged as a practical approach to distilling autoregressive language models. Unlike distillation on a fixed dataset of teacher-generated outputs, OPD samples trajectories from the current student and obtains token-level supervision from the teacher along the resulting student-generated prefixes. Agarwal et al. (2024) demonstrated this approach on summarization, machine translation, arithmetic reasoning, and task-agnostic instruction tuning. The same principle has since been incorporated into the training pipelines of released LLM families: Gemma 2 performs teacher distillation on the student's distribution during supervised fine-tuning, while Qwen3 combines off-policy and on-policy strong-to-weak distillation in the post-training of its smaller models (Team et al., 2024; Yang et al., 2025).

On-Policy Self-Distillation (OPSD) is a recent instantiation of OPD that does not require a distinct, typically larger, teacher model. Instead, the trained student and the teacher are initialized from the same language model. The student conditions only on the target problem and generates an on-policy response, while the teacher additionally conditions on the target's reference solution and provides next-token distributions along the student-generated prefixes. Across three Qwen3 models and several mathematics benchmarks, OPSD improves over the base model and supervised fine-tuning, while matching or exceeding GRPO with substantially fewer sampled training tokens (Zhao et al., 2026a). Conditioning the teacher on the solution may therefore produce more informative token-level targets for the student who does not see the solution. **This suggests an explanation for OPSD's gains: they arise because the teacher has privileged access to the answer and reasoning trace for the exact problem being solved by the student.**

This explanation is plausible, but the usual comparison does not highlight it. When a verified solution is placed in the teacher prompt, two things change simultaneously. First, it reveals target-specific privileged information: the derivation and answer for the current problem. Second, it provides a complete worked solution within the teacher's context. This context may alter the teacher's next-token distribution due to its reasoning structure, notation, style, and effects on continuation and termination, even if it does not solve the current instance. Therefore, an improvement over the base model does not establish that OPSD transfers the privileged answer.

We ask whether OPSD depends on the identity of the solution, or whether part of its effect arises from behavior induced by the teacher's additional context. To probe this distinction, we introduce $\mathrm{OP}^2\mathrm{SD}$ (***O**n-**P**olicy **S**elf-**D**istillation from Other **P**roblems*). For a target problem $A$, standard OPSD gives the teacher the verified solution to $A$. $\mathrm{OP}^2\mathrm{SD}$ instead gives it a different problem $B$ and its solution, while explicitly stating that the example is neither a solution nor a hint for $A$. The student still sees only $A$. We otherwise preserve the on-policy rollout, token-level objective, and optimization procedure. This intervention breaks the target and reference pairing while retaining the teacher-only worked-solution context (Figure 1). Across the evaluated Qwen3-1.7B, Qwen3-4B, and Qwen3-8B settings, $\mathrm{OP}^2\mathrm{SD}$ improves accuracy over the corresponding base model on AIME 2024, AIME 2025, and HMMT 2025 (Section 4.1). The Qwen3-1.7B setting follows the asymmetric configuration of prior work, with a non-thinking student and a thinking teacher, whereas the Qwen3-4B and Qwen3-8B settings are mode-matched in non-thinking mode.

Furthermore, we found that not all additional context yields a useful teacher. When evaluating the model on algebraic questions, we find that a worked solution drawn from the broader mathematical domain, rather than algebra specifically, provides a similar advantage. In contrast, replacing mathematical examples with physics problems and solutions significantly reduces accuracy and causes many generations to fall into repetitive, non-terminating trajectories. Thus, the identity and form of the teacher context matter, even though exact alignment with the target solution does not appear necessary.

Our contribution is therefore primarily diagnostic. We identify a confounding factor in the standard interpretation of OPSD, introduce a controlled intervention that decouples the target-reference pairing, and demonstrate that a significant amount of improvement persists despite this intervention. These findings suggest that OPSD should not be interpreted solely as a privileged-answer transfer: context-induced changes in teacher behavior may also constitute an important part of the signal being distilled.

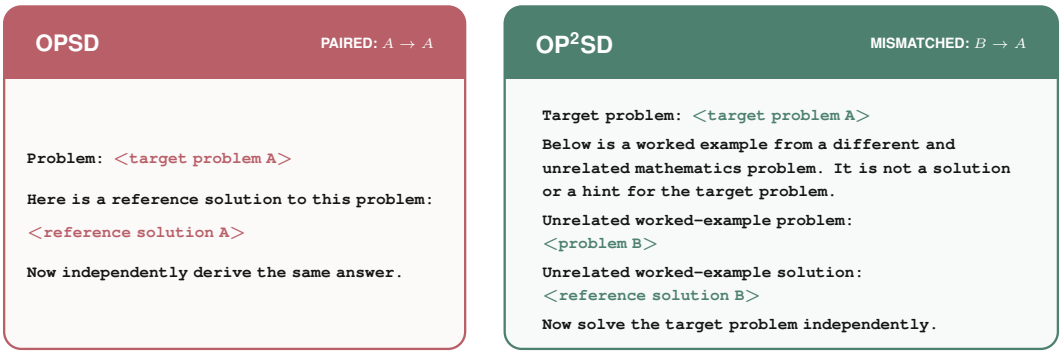


Figure 2: **Teacher prompt comparison.** OPSD conditions the teacher on the reference paired with target $A$; OP$^2$SD substitutes a worked solution to a different problem $B$. The target, student rollout, teacher model, and token-level objective are unchanged.

## 2 Problem Formulation

### 2.1 On-Policy Self-Distillation

On-Policy Self-Distillation (OPSD) trains a language model using token-level supervision from a frozen copy of its own initialization (Zhao et al., 2026a). The student and teacher, therefore, originate from the same base model, but they receive different conditioning information. The trainable student sees only the target problem and generates the trajectory used for training. The teacher sees the same target problem together with its reference solution and evaluates the student's trajectory under this additional context. The method is *on-policy* because supervision is provided on prefixes generated by the current student, and it is *self-distillation* because the student and teacher are derived from the same model rather than from models of different sizes.

Let $\mathcal{D} = \{(x_i, y_i^\star)\}_{i=1}^N$ denote a dataset of mathematical problems $x_i$ paired with reference solutions $y_i^\star$. Let $p_0$ be the base model and $p_\theta$ the trainable student initialized from $p_0$. For target problem $x_i$, the student policy conditions only on the problem, whereas the OPSD teacher additionally conditions on the paired reference:

$$p_S(\cdot \mid x_i) = p_\theta(\cdot \mid x_i), \qquad p_T^{\text{target}}(\cdot \mid x_i, y_i^\star) = p_0(\cdot \mid x_i, y_i^\star). \tag{1}$$

The teacher does not generate an alternative solution. Thus, the teacher's advantage does not derive from having a larger parameter count, but from its privileged access to $y_i^\star$. This reference is available only to the teacher during training; the student must solve the problem without it both when generating training trajectories and at inference time. Given the two context-conditioned policies above, an OPSD training step proceeds in two stages.

First, the current student samples a complete response to the target problem:

$$\hat{y}_i \sim p_\theta(\cdot \mid x_i). \tag{2}$$

Second, to compute the OPSD loss, the same output is replayed under both the student and the teacher. Note that the student prompt $x_i$ and the teacher prompt $(x_i, y_i^\star)$ are independently right-padded within their respective minibatches before $\hat{y}_i$ is appended; details are provided in Appendix E.

### 2.2 Target-Specific Privilege or Context-Induced Teacher Behavior?

Existing accounts of OPSD are consistent with the broader literature on learning using privileged information (Vapnik & Vashist, 2009), typically attribute its gains to the teacher's access to privileged information about the current problem (Zelikman et al., 2022; Mitra & Ulukus, 2025; Qi et al., 2025; Zhao et al., 2026a; Hübotter et al., 2026). Under this interpretation, the teacher receives the verified (partial) answer and reasoning trace for the same problem (also known as "reference solution" or "target solution") that the student is attempting to solve, and the student learns from the teacher's predictions under this target-specific information.

However, providing a reference solution changes not only what the teacher knows about the target, but also the context under which it produces token-level supervision. A worked solution may alter the teacher's next-token distribution through its reasoning structure, notation, style, and effects on continuation and termination, even when it conveys no information about the current instance. We therefore distinguish *target-specific privilege* from *context-induced teacher behavior*. Our central question is whether OPSD requires privileged information about the current target, or whether a teacher-only worked-solution context can induce a useful supervisory distribution without revealing the target's solution.

# 3 OP$^2$SD On-Policy Self-Distillation from Other Problems

## 3.1 Replacing the Paired Reference

OP$^2$SD retains the on-policy training procedure of OPSD but changes the source of the teacher-only reference context. In OPSD, the teacher receives the reference solution paired with the current target. In OP$^2$SD, that paired reference is replaced by a problem and solution pair from a different training instance. This intervention is designed to test whether instance-level alignment between the target problem and the privileged reference is necessary for the OPSD gain.

Formally, for each target problem $x_i$, we select another problem $x_j$ and its reference solution $y_j^\star$, subject to $j \neq i$. We refer to the pair $(x_j, y_j^\star)$ as the *worked-example*. The student policy is unchanged; on the other hand, the teacher instead conditions on both the target problem and the worked-example (Figure 2):

$$p_T^{\text{other}}(\cdot \mid x_i, x_j, y_j^\star) = p_0(\cdot \mid x_i, x_j, y_j^\star). \tag{3}$$

As in the target solution, OPSD, the teacher does not generate a separate response. The paired target solution $y_i^\star$ is not included in the conditioning context of either policy. The teacher's prompt explicitly identifies $(x_j, y_j^\star)$ as a worked example from a different problem and states that it should not be treated as a solution or hint for $x_i$.

## 3.2 Distillation Objective

OPSD and OP$^2$SD use the same rollout construction and token-level distillation objective. To express both conditions in a common form, let

$$m \in \{\text{target}, \text{other}\}, \qquad c_i^{\text{target}} = y_i^\star, \qquad c_i^{\text{other}} = (x_j, y_j^\star).$$

In either condition, the current student first samples a response, and the student and teacher then evaluate the same set of student-generated prefixes. The corresponding objective is

$$\mathcal{L}_m(\theta) = \mathbb{E}_{(x_i, y_i^\star)\sim\mathcal{D}} \mathbb{E}_{\hat{y}_i \sim p_S(\cdot|x_i)} \left[ \frac{1}{|\hat{y}_i|} \sum_{t=1}^{|\hat{y}_i|} D(p_T^m(\cdot \mid x_i, c_i^m, \hat{y}_{i,<t}) \,\|\, p_S(\cdot \mid x_i, \hat{y}_{i,<t})) \right]. \tag{4}$$

Here, $D$ denotes the divergence at the token-level between the next-token distributions of the teacher and student. In our experiments, both conditions use the same clipped Forward-KL surrogate. The teacher distribution is treated as a fixed target and the gradients are propagated only through the student. Thus, the two conditions differ in the teacher-only context $c_i^m$: OPSD uses the reference paired with the target, whereas OP$^2$SD uses the worked-example from another problem. The student input, rollout construction, teacher model, distillation objective, and optimization procedures are otherwise unchanged.

# 4 Experiments

We evaluate OP$^2$SD using Qwen3-1.7B ($p_S$: non-thinking; $p_T$: thinking), Qwen3-4B ($p_S$, $p_T$: non-thinking), and Qwen3-8B ($p_S$, $p_T$: non-thinking) (Yang et al., 2025). All main experiments are conducted on the OpenThoughts Math dataset (`siyanzhao/Openthoughts_math_30k_opsd`), following the experimental setup of Zhao et al. (2026a). For each model setting, we compare OP$^2$SD against OPSD and the base model (Base). In OP$^2$SD, each target problem is paired with a worked-example drawn from the same training dataset. The worked-example pool consists of entries labeled

Table 1: **Main benchmark results.** For Qwen3-1.7B, the student is non-thinking, and the teacher is thinking, following the original OPSD setup; Qwen3-4B and Qwen3-8B use mode-matched non-thinking student and teacher policies. Avg@12 is reported as the mean (±) corrected Monte Carlo standard error, while Pass@12 and Vote@12 are reported as the mean (±) empirical decoding-seed standard error. Within each model and benchmark, the highest Avg@12 point estimate is shown in **bold**, and the second-highest is underlined. $\text{OP}^2\text{SD}$ achieves the highest Avg@12 point estimate in eight of the nine model–benchmark groups. Comparisons with SFT and GRPO are not repeated here, as Zhao et al. (2026a) already reports that OPSD outperforms SFT and matches or exceeds GRPO under their experimental setup; our focus is on isolating the effect of the teacher-side context.

| Model | Benchmark | Method | Avg@12 | Pass@12 | Vote@12 |
|---|---|---|---|---|---|
| Qwen3-1.7B | AIME 2024 | Base | $49.86 \pm 0.94$ | $76.67 \pm 0.00$ | $70.00 \pm 2.72$ |
| | | OPSD | $\mathbf{55.42 \pm 0.82}$ | $78.33 \pm 0.96$ | $65.00 \pm 1.67$ |
| | | $\text{OP}^2\text{SD}$ | $55.35 \pm 0.83$ | $75.83 \pm 0.83$ | $70.00 \pm 2.36$ |
| | AIME 2025 | Base | $37.36 \pm 0.81$ | $70.00 \pm 3.04$ | $48.33 \pm 0.96$ |
| | | OPSD | $40.35 \pm 0.77$ | $65.00 \pm 2.15$ | $50.00 \pm 1.36$ |
| | | $\text{OP}^2\text{SD}$ | $\mathbf{40.69 \pm 0.74}$ | $63.33 \pm 2.36$ | $52.50 \pm 1.60$ |
| | HMMT 2025 | Base | $23.61 \pm 0.64$ | $52.50 \pm 2.50$ | $28.33 \pm 0.96$ |
| | | OPSD | $25.76 \pm 0.63$ | $50.83 \pm 2.10$ | $29.17 \pm 0.83$ |
| | | $\text{OP}^2\text{SD}$ | $\mathbf{27.57 \pm 0.66}$ | $51.67 \pm 1.67$ | $30.83 \pm 0.83$ |
| Qwen3-4B | AIME 2024 | Base | $23.19 \pm 0.72$ | $50.83 \pm 1.60$ | $33.33 \pm 2.36$ |
| | | OPSD | $30.76 \pm 0.77$ | $59.17 \pm 0.83$ | $45.83 \pm 1.60$ |
| | | $\text{OP}^2\text{SD}$ | $\mathbf{31.53 \pm 0.80}$ | $60.83 \pm 1.60$ | $43.33 \pm 1.36$ |
| | AIME 2025 | Base | $21.11 \pm 0.64$ | $46.67 \pm 2.72$ | $26.67 \pm 2.36$ |
| | | OPSD | $23.06 \pm 0.65$ | $50.00 \pm 1.36$ | $28.33 \pm 0.96$ |
| | | $\text{OP}^2\text{SD}$ | $\mathbf{30.62 \pm 0.71}$ | $55.00 \pm 2.15$ | $40.83 \pm 1.60$ |
| | HMMT 2025 | Base | $11.67 \pm 0.49$ | $22.50 \pm 2.10$ | $16.67 \pm 0.00$ |
| | | OPSD | $15.42 \pm 0.57$ | $35.83 \pm 2.85$ | $20.83 \pm 0.83$ |
| | | $\text{OP}^2\text{SD}$ | $\mathbf{16.11 \pm 0.60}$ | $38.33 \pm 2.15$ | $20.00 \pm 1.92$ |
| Qwen3-8B | AIME 2024 | Base | $28.47 \pm 0.75$ | $58.33 \pm 3.19$ | $41.67 \pm 3.97$ |
| | | OPSD | $45.00 \pm 0.86$ | $74.17 \pm 1.60$ | $61.67 \pm 2.15$ |
| | | $\text{OP}^2\text{SD}$ | $\mathbf{55.83 \pm 0.87}$ | $80.00 \pm 1.36$ | $71.67 \pm 2.15$ |
| | AIME 2025 | Base | $20.97 \pm 0.63$ | $44.17 \pm 2.10$ | $28.33 \pm 0.96$ |
| | | OPSD | $32.15 \pm 0.76$ | $58.33 \pm 2.89$ | $40.00 \pm 1.36$ |
| | | $\text{OP}^2\text{SD}$ | $\mathbf{42.85 \pm 0.71}$ | $62.50 \pm 2.10$ | $52.50 \pm 2.50$ |
| | HMMT 2025 | Base | $11.81 \pm 0.51$ | $23.33 \pm 0.00$ | $16.67 \pm 0.00$ |
| | | OPSD | $16.18 \pm 0.60$ | $39.17 \pm 1.60$ | $20.00 \pm 1.36$ |
| | | $\text{OP}^2\text{SD}$ | $\mathbf{25.56 \pm 0.72}$ | $55.00 \pm 1.67$ | $30.83 \pm 2.10$ |

`amc_aime` or `aops_forum` that do not include AIME2024, AIME2025, and HMMT Feb 2025 in the OpenThoughts Math dataset, and each selected example includes both a problem and its reference solution. The details of the experiments are provided in Appendix D. We evaluated these methods on AIME 2024, AIME 2025, and HMMT Feb 2025 (Balunovic et al., 2025). For each problem, we generate 12 outputs under each of four different seeds.

### 4.1 Main results

We report Avg@12 as the mean over four different seeds, together with a corrected Monte Carlo standard error. Pass@12 and Vote@12 are reported as means together with empirical seed standard errors. The definitions of Avg@12, Pass@12, and Vote@12 are given in Appendix C.

Table 1 shows the main benchmark results. For Qwen3-1.7B, OPSD and $\text{OP}^2\text{SD}$ improve Avg@12 over Base on all three benchmarks. These results suggest that access to paired target references is not necessary to achieve Avg@12 gains comparable to OPSD's. Avg@12 and Pass@12 do not always exhibit the same pattern. Although $\text{OP}^2\text{SD}$ improves Avg@12 over Base on all three benchmarks for Qwen3-1.7B, its Pass@12 point estimates are slightly lower. This pattern is consistent with

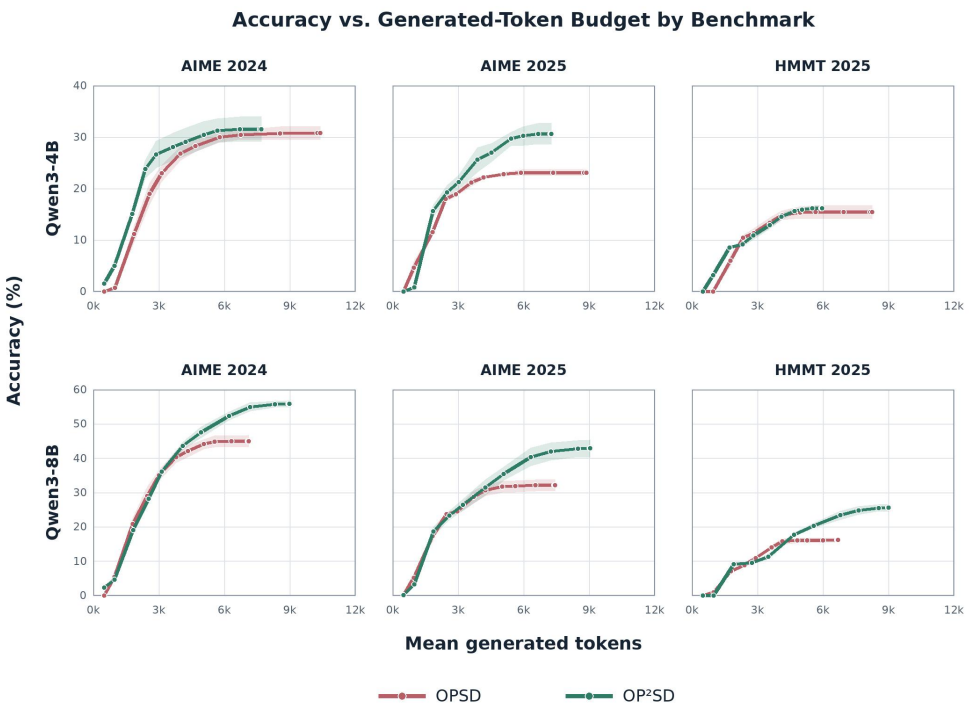


Figure 3: **Accuracy under generated-token budgets.** Each point is obtained by truncating the saved response at a fixed token cap and reapplying the original answer extractor and grader. If the model has not produced an answer by the truncation point, the response is counted as incorrect. The horizontal coordinate is the actual mean number of retained tokens after accounting for responses that terminate before the cap.

$\text{OP}^2\text{SD}$ increasing the probability of generating correct responses for problems that the base model can already solve occasionally, rather than uniformly expanding the set of problems solved within 12 samples (Appendix I). The Qwen3-4B and Qwen3-8B results further demonstrate the effectiveness of $\text{OP}^2\text{SD}$ at larger model scales, as it achieves the highest Avg@12 point estimate on all three benchmarks for both model sizes. For Qwen3-8B, it outperforms OPSD in Avg@12 by 10.83, 10.70, and 9.38 points on AIME 2024, AIME 2025, and HMMT 2025, respectively. We next examine whether these gains require additional generation.

#### 4.1.1 Accuracy under fixed token budgets

Using the saved Qwen3-4B and Qwen3-8B outputs, we truncate each response at budgets from 500 to 38,912 tokens and reapply the original boxed answer evaluator; prefixes without a complete boxed answer are counted as incorrect. Figure 3 shows accuracy against the actual mean number of retained tokens, accounting for responses that terminate before the specified budget is reached.

Figure 3 shows different scaling behavior. For Qwen3-4B, $\text{OP}^2\text{SD}$ remains above OPSD over nearly the entire budget range and reaches a higher final accuracy with fewer mean generated tokens. Thus, the 4B gain is not explained by longer responses. For Qwen3-8B, the two methods are comparable at small budgets, after which $\text{OP}^2\text{SD}$ continues to improve while OPSD saturates. Its final output is longer, but a substantially higher pooled accuracy accompanies the additional budget. Hence, $\text{OP}^2\text{SD}$ does not uniformly shorten reasoning across model scales; rather, it achieves higher accuracy for a given token budget once the budget is sufficiently large.

The results reported so far show that paired target solutions are not necessary to obtain improvements comparable to those of OPSD. We next investigate which properties of the teacher signal account for this result. We first disentangle the effect of solution conditioning from that of the student-teacher

Table 2: **Worked-solution controls with Qwen3-4B.** Target-only provides no additional information beyond the target problem, whereas Answer-only additionally provides the verified final answer without a derivation. Unlike in the Qwen3-1.7B setting, neither control improves over Base: Target-only substantially degrades Avg@12 on all three benchmarks, and Answer-only partially recovers performance but remains below Base. These results indicate that the gains observed in the asymmetric 1.7B controls are not observed when student and teacher modes are matched.

| Teacher information | AIME 2024 | AIME 2025 | HMMT 2025 |
|---|---|---|---|
| Target-only | $1.18 \pm 0.28$ | $0.49 \pm 0.18$ | $0.76 \pm 0.22$ |
| Answer-only | $18.68 \pm 0.55$ | $17.85 \pm 0.55$ | $9.24 \pm 0.49$ |
| Base | $23.19 \pm 0.72$ | $21.11 \pm 0.64$ | $11.67 \pm 0.49$ |

thinking-mode asymmetry. We then consider a matched non-thinking setting and progressively examine whether the observed gains depend on diverse mathematical examples, broad domain alignment, a correct auxiliary derivation, or simply the presence of arbitrary additional context.

**A mode-asymmetry inherited from the 1.7B OPSD setup.** For Qwen3-1.7B, we adopted the original OPSD settings of Zhao et al. (2026a), in which the student generates in non-thinking mode while the frozen teacher scores the same trajectory in thinking mode. The asymmetry is therefore inherited from the baseline recipe rather than introduced by $\mathrm{OP}^2\mathrm{SD}$, and retaining it enables a direct comparison with the original setting. It nevertheless confounds the role of worked-solution context: an improvement could arise from the solution, the stronger teacher mode, generic on-policy distillation, or their interaction.

We introduce two controls to separate this mode asymmetry from the effect of worked-solution context. In *Target-only*, the teacher receives the same problem-only content as the student, with no reference solution. In *Answer-only*, the teacher additionally receives the final answer, but not its derivation. We confirm that both Target-only and Answer-only improve the base model's accuracy across all three benchmarks (see Appendix G). Thus, in the 1.7B setting, improvements can arise even when the teacher receives neither a worked solution nor additional mathematical context. Therefore, these gains cannot be exclusively attributed to the content of the privileged solution and may instead reflect, at least in part, the distillation of the behavior induced by the teacher's thinking mode. Complete results and prompt details are reported in Appendix G. To remove this asymmetry, we evaluate the controls using Qwen3-4B, with both the student and the teacher in non-thinking mode.

As shown in Table 2, the Qwen3-4B non-thinking setting behaves differently from the 1.7B setting. Target-only performs substantially below Base on all three benchmarks. Providing the verified final answer recovers part of this loss, but Answer-only remains $4.51$, $3.26$, and $2.43$ percentage points below Base on AIME 2024, AIME 2025, and HMMT 2025, respectively. In contrast, both target-solution OPSD and worked-problem $\mathrm{OP}^2\mathrm{SD}$ improve over Base, as shown in Table 1.

Thus, in this mode-matched configuration, improvement is observed only when the teacher is provided with a complete mathematical worked solution. We therefore next investigate which properties of this worked-solution context are responsible for the improvement.

### 4.2 What Properties of the Other-Problem Context Matter?

**Is diversity among worked examples necessary?** The $\mathrm{OP}^2\mathrm{SD}$ condition exposes the teacher to examples drawn from different problems. One possibility is that its benefit arises from the diversity of these examples. To test this hypothesis, we remove this diversity by conditioning the teacher on a single fixed algebra problem from Omni-MATH (Gao et al., 2025) throughout training. We refer to this condition as *Fixed correct*. The exact problem and solution are provided in Appendix L. Table 3 shows that the fixed-example run attains Avg@12 point estimates of $35.83$, $31.94$, and $17.71$ on AIME 2024, AIME 2025, and HMMT 2025, compared with $31.53$, $30.62$, and $16.11$ for varying-example $\mathrm{OP}^2\mathrm{SD}$ with Qwen3-4B. Thus, an OPSD-like improvement can occur even when the auxiliary context is reduced to one repeated example.

Table 3: **Effect of diversity and form of the mathematical teacher context.** Replacing the varying worked examples with a single fixed algebra problem and a concise, correct solution preserves and slightly improves Avg@12 across all three benchmarks with Qwen3-4B, showing that diversity among worked examples is unnecessary, whereas a substantially more verbose correct solution yields lower Avg@12 and shorter student responses. In contrast, a fixed trivial $1 + 1$ example performs poorly, indicating that an arbitrary correct mathematical example is not sufficient. Solution tokens denote the length of the teacher-provided solution, and Mean tokens denote the average student response length across all generations. The exact fixed contexts are provided in Appendix L.

| Teacher context | Solution tokens | AIME 2024 | AIME 2025 | HMMT 2025 | Mean tokens |
|---|---|---|---|---|---|
| Varying mathematics examples | – | $31.53 \pm 0.80$ | $30.62 \pm 0.71$ | $16.11 \pm 0.60$ | 6,979 |
| Fixed correct | 204 | $35.83 \pm 0.87$ | $31.94 \pm 0.79$ | $17.71 \pm 0.62$ | 6,428 |
| Fixed locally corrupted | 204 | $36.04 \pm 0.83$ | $34.86 \pm 0.78$ | $17.99 \pm 0.63$ | 6,442 |
| Fixed verbose correct | 742 | $33.89 \pm 0.85$ | $28.40 \pm 0.74$ | $15.76 \pm 0.68$ | 5,773 |
| Fixed trivial correct $(1 + 1)$ | 18 | $5.21 \pm 0.52$ | $5.28 \pm 0.49$ | $2.29 \pm 0.35$ | 19,029 |

We next hold the fixed worked example constant while varying its solution. The Fixed locally corrupted condition uses a solution with the same length as the concise correct version (204 tokens) but introduces a local error that changes the final answer. The Fixed verbose correct condition instead preserves the correct answer while deliberately making the derivation verbose by increasing it to 742 tokens. The Fixed locally corrupted condition yields Avg@12 point estimates comparable to or higher than those of the Fixed correct condition across all three benchmarks. This result suggests that $\mathrm{OP}^2\mathrm{SD}$ does not require a fully correct solution to achieve strong performance. The verbose correct context exhibits a different pattern. Relative to the concise correct context, it lowers Avg@12 by $1.94$, $3.54$, and $1.95$ points while reducing the pooled mean student response length by approximately $10.2\%$. In this experiment, these results suggest that a concise, focused solution example may provide a more effective teacher signal than a substantially expanded derivation. More broadly, the form and degree of compression of the teacher-only context appear to be associated with both accuracy and student response length. However, these comparisons do not identify the underlying causal mechanism.

As an additional control, we condition the teacher on a fixed, trivially correct (1+1) example (see Appendix L). This condition performs poorly, achieving Avg@12 scores of only $5.21$, $5.28$, and $2.29$ on AIME 2024, AIME 2025, and HMMT 2025, respectively. Thus, although diversity is not necessary, merely providing an arbitrary correct mathematical example is insufficient to achieve the observed gains.

**Is broad domain matching necessary?** The fixed-example result implies that $\mathrm{OP}^2\mathrm{SD}$ does not necessarily require the teacher to observe a diverse stream of worked-problem solutions to attain strong performance. However, the successful fixed example is still mathematical and may benefit from belonging to the same broad domain as the target problems. We therefore next ask whether the auxiliary problem must match the target's mathematical domain.

Using Omni-MATH (Gao et al., 2025), we train on 1,280 Algebra problems and evaluate on 30 held-out Algebra problems. The teacher receives another Algebra problem and its solution, or a Geometry problem and its solution. The two other problem sets are matched by source, difficulty, and context length; construction and leakage checks are provided in Appendix F. From Table 4, the Geometry condition is $1.67$ points higher, but this apparent advantage is driven by one target near the boundary between Algebra and analytic Geometry. Excluding that problem reverses the ordering: Algebra obtains $59.41\%$ and Geometry obtains $58.98\%$ Avg@12. We therefore find no clear evidence that matching a coarse domain label is necessary.

**Is mathematical context necessary?** The preceding controls relax target alignment, example diversity, broad domain matching, and, in one pilot, exact correctness without uniformly eliminating the gain. This does not imply that any additional text produces a useful teacher. For this cross-subject control, we sample 2,768 problem and solution pairs from the external CAMEL Physics dataset (Li et al., 2023), matching the size of the mathematical example pool. We keep the OpenThoughts

Table 4: **Effect of coarse mathematical domain matching.** Using Qwen3-4B in the matched non-thinking setting, we train on Algebra targets and compare teacher-only contexts containing either another Algebra problem and solution or a matched Geometry problem and solution. Geometry context performs comparably to, and slightly better than, Algebra context on the 30 held-out Algebra problems, providing no clear evidence that matching the target's coarse mathematical domain is necessary. Dataset construction and leakage checks are provided in Appendix F.

| Teacher-only context | Avg@12 | Pass@12 |
|---|---|---|
| Another Algebra problem and solution | $57.43 \pm 0.71$ | $79.17 \pm 0.83$ |
| Another Geometry problem and solution | $59.10 \pm 0.73$ | $80.00 \pm 1.36$ |

Table 5: **Cross-subject control for mathematical solution context.** Using Qwen3-4B in the matched non-thinking setting, we replace only the teacher's mathematical worked examples with CAMEL Physics problem and solution pairs. Mathematical contexts improve Avg@12 over Base on all three benchmarks, whereas physics contexts fall below Base throughout. Thus, an arbitrary cross-subject context does not preserve the $\mathrm{OP}^2\mathrm{SD}$ gain, suggesting that the mathematical solution context, or an associated property, is important. Additional metrics and implementation details are provided in Appendix K.

| Condition | AIME 2024 | AIME 2025 | HMMT 2025 |
|---|---|---|---|
| Base | $23.19 \pm 0.72$ | $21.11 \pm 0.64$ | $11.67 \pm 0.49$ |
| $\mathrm{OP}^2\mathrm{SD}$ (mathematics) | $31.53 \pm 0.80$ | $30.62 \pm 0.71$ | $16.11 \pm 0.60$ |
| $\mathrm{OP}^2\mathrm{SD}$ (physics) | $18.68 \pm 0.69$ | $20.21 \pm 0.61$ | $8.68 \pm 0.50$ |

Math target stream unchanged and replace only the teacher-only worked-example context. Data construction and the complete teacher-prompt template are reported in Appendix K.

As shown in Table 5, replacing mathematical worked solutions with physics examples reduces Avg@12 from 31.53 to 18.68 on AIME 2024, from 30.62 to 20.21 on AIME 2025, and from 16.11 to 8.68 on HMMT 2025. The physics condition is also below Base on all three benchmarks. Appendix K reports the corresponding Pass@12, Vote@12, and answer-format validity, together with the actual teacher-prompt template; these diagnostics show that the accuracy drop is accompanied by substantially less reliable answer formatting. Together, these results suggest that the mathematical solution context, or some property associated with it, is important for obtaining the $\mathrm{OP}^2\mathrm{SD}$ gain. However, because the physics condition differs from the mathematics condition in several respects, this experiment does not establish that mathematical subject matter itself is the causal factor.

Overall, the gains persist after removing instance-level alignment, example diversity, and coarse domain matching, but are substantially reduced when mathematical worked solutions are replaced with solutions from another subject. This suggests that substantive mathematical worked-solution context, rather than privileged-answer transfer or arbitrary extra context, is important for $\mathrm{OP}^2\mathrm{SD}$.

## 5 Conclusion

We investigated whether OPSD requires the teacher to observe a verified solution to the same problem as the student. Across various mathematical benchmarks and models, $\mathrm{OP}^2\mathrm{SD}$ remains competitive with OPSD, showing that paired target solutions are not necessary for OPSD-like gains. Most of the OPSD gain is retained, or even reinforced, after the reference solution to the target problem is replaced by the solution to a different problem ($\mathrm{OP}^2\mathrm{SD}$). Moreover, the gain persists when the diversity of the problem-solution pair given to the teacher is reduced to a single fixed example. This shows that neither diverse auxiliary information nor exact contextual information is required. At the same time, trivial mathematical and cross-subject physics contexts fail to preserve the improvement, indicating that arbitrary additional text is insufficient. The verbose contexts also reduced the effectiveness of the teacher. Taken together, these results imply that the major source of the observed gain is not the privileged answer to the target problem, but the change in the teacher's token-level behavior induced by a mathematical context that works.

## AI USE STATEMENT

We used generative AI tools to polish the sentences and to code during the preparation of this work. AI-assisted text was reviewed and edited by the authors, and AI-assisted code was manually inspected and tested for correctness before being used in our experiments. The authors take full responsibility for the experimental design, analysis, claims, and final content of the paper.

## REPRODUCIBILITY STATEMENT

Experimental details, control experiments, checkpoint analyses, and problem-level analyses are provided in the appendices. We also report the prompts and configurations used for the different teacher-context conditions to facilitate reproduction of the comparisons in this work.

## ETHICS STATEMENT

This work studies self-distillation methods for language models using existing models and mathematical reasoning datasets. We do not conduct human-subject experiments or collect new personal data.

## REFERENCES

Rishabh Agarwal, Nino Vieillard, Yongchao Zhou, Piotr Stanczyk, Sabela Ramos Garea, Matthieu Geist, and Olivier Bachem. On-Policy Distillation of Language Models: Learning from Self-Generated Mistakes. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=3zKtaqxLhW.

Mislav Balunovic, Jasper Dekoninck, Ivo Petrov, Nikola Jovanović, and Martin Vechev. Matharena: Evaluating LLMs on uncontaminated math competitions. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, 2025. URL https://openreview.net/forum?id=y0zL9IZxZ7.

Bofei Gao, Feifan Song, Zhe Yang, Zefan Cai, Yibo Miao, Qingxiu Dong, Lei Li, Chenghao Ma, Liang Chen, Runxin Xu, Zhengyang Tang, Benyou Wang, Daoguang Zan, Shanghaoran Quan, Ge Zhang, Lei Sha, Yichang Zhang, Xuancheng Ren, Tianyu Liu, and Baobao Chang. Omni-MATH: A Universal Olympiad Level Mathematic Benchmark for Large Language Models. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=yaqPf0KAlN.

Yuxian Gu, Li Dong, Furu Wei, and Minlie Huang. MiniLLM: Knowledge Distillation of Large Language Models. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=5h0qf7IBZZ.

Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. Distilling the knowledge in a neural network. *Deep Learning and Representation Learning Workshop in Conjunction with NIPS*, 2014.

Cheng-Yu Hsieh, Chun-Liang Li, Chih-kuan Yeh, Hootan Nakhost, Yasuhisa Fujii, Alex Ratner, Ranjay Krishna, Chen-Yu Lee, and Tomas Pfister. Distilling step-by-step! outperforming larger language models with less training data and smaller model sizes. In Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki (eds.), *Findings of the Association for Computational Linguistics: ACL 2023*, pp. 8003–8017, Toronto, Canada, July 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.findings-acl.507. URL https://aclanthology.org/2023.findings-acl.507/.

Jonas Hübotter, Frederike Lübeck, Lejs Deen Behric, Anton Baumann, Marco Bagatella, Daniel Marta, Ido Hakimi, Idan Shenfeld, Thomas Kleine Buening, Carlos Guestrin, and Andreas Krause. Reinforcement learning via self-distillation. In *Forty-third International Conference on Machine Learning*, 2026. URL https://openreview.net/forum?id=QkfkxyRizZ.

Simran Kaur, Narutatsu Ri, Yinghui He, Liam H Fowl, and Sanjeev Arora. Rethinking On-Policy Self-Distillation for Thinking Models. In *ICML 2026 Workshop on Foundations of Deep Generative Models: Understanding Memorization, Generalization, and Reasoning*, 2026. URL https://openreview.net/forum?id=VhCJItwQHn.

Yoon Kim and Alexander M. Rush. Sequence-level knowledge distillation. In Jian Su, Kevin Duh, and Xavier Carreras (eds.), *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pp. 1317–1327, Austin, Texas, November 2016. Association for Computational Linguistics. doi: 10.18653/v1/D16-1139. URL https://aclanthology.org/D16-1139/.

Jongwoo Ko, Sungnyun Kim, Tianyi Chen, and Se-Young Yun. DistiLLM: Towards Streamlined Distillation for Large Language Models. In *Forty-first International Conference on Machine Learning*, 2024. URL https://openreview.net/forum?id=lsHZNNoC7r.

Guohao Li, Hasan Hammoud, Hani Itani, Dmitrii Khizbullin, and Bernard Ghanem. CAMEL: Communicative agents for" mind" exploration of large language model society. *Advances in neural information processing systems*, 36:51991–52008, 2023.

Xiaogeng Liu, Xinyan Wang, Yingzi Ma, Yechao Zhang, and Chaowei Xiao. When Are Teacher Tokens Reliable? Position-Weighted On-Policy Self-Distillation for Reasoning. *arXiv preprint arXiv:2605.21606*, 2026.

David Lopez-Paz, Léon Bottou, Bernhard Schölkopf, and Vladimir Vapnik. Unifying distillation and privileged information. *arXiv preprint arXiv:1511.03643*, 2015.

Purbesh Mitra and Sennur Ulukus. Semantic soft bootstrapping: Long context reasoning in llms without reinforcement learning. *arXiv preprint arXiv:2512.05105*, 2025.

Emiliano Penaloza, Dheeraj Vattikonda, Nicolas Gontier, Alexandre Lacoste, Laurent Charlin, and Massimo Caccia. Privileged information distillation for language models. In *Forty-third International Conference on Machine Learning*, 2026. URL https://openreview.net/forum?id=ebZcMQImhG.

Siyuan Qi, Bangcheng Yang, Kailin Jiang, Xiaobo Wang, Jiaqi Li, Yifan Zhong, Yaodong Yang, and Zilong Zheng. In-context editing: Learning knowledge from self-induced distributions. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=w6rHCuN3YG.

Zhanming Shen, Jintao Tong, Shaotian Yan, Chen Shen, Hao Chen, Wentao Ye, Xiaomeng Hu, Rui Miao, Haobo Wang, Junbo Zhao, et al. Purified OPSD: On-Policy Self-Distillation Without Losing How to Think. *arXiv preprint arXiv:2607.02234*, 2026.

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

Charlie Snell, Dan Klein, and Ruiqi Zhong. Learning by distilling context. *arXiv preprint arXiv:2209.15189*, 2022.

Zhiquan Tan and Yinrong Hong. Self-Supervised On-Policy Distillation for Reasoning Language Models. *arXiv preprint arXiv:2605.17497*, 2026.

Gemma Team, Morgane Riviere, Shreya Pathak, Pier Giuseppe Sessa, Cassidy Hardin, Surya Bhupatiraju, Léonard Hussenot, Thomas Mesnard, Bobak Shahriari, Alexandre Ramé, et al. Gemma 2: Improving open language models at a practical size. *arXiv preprint arXiv:2408.00118*, 2024.

Vladimir Vapnik and Akshay Vashist. A new learning paradigm: Learning using privileged information. *Neural Networks*, 22(5):544–557, 2009. ISSN 0893-6080. doi: https://doi.org/10.1016/j.neunet.2009.06.042. URL https://www.sciencedirect.com/science/article/pii/S0893608009001130. Advances in Neural Networks Research: IJCNN2009.

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

Tianzhu Ye, Li Dong, Xun Wu, Shaohan Huang, and Furu Wei. On-Policy Context Distillation for Language Models. *arXiv preprint arXiv:2602.12275*, 2026.

Eric Zelikman, Yuhuai Wu, Jesse Mu, and Noah Goodman. STar: Bootstrapping reasoning with reasoning. In Alice H. Oh, Alekh Agarwal, Danielle Belgrave, and Kyunghyun Cho (eds.), *Advances in Neural Information Processing Systems*, 2022. URL https://openreview.net/forum?id=_3ELRdg2sgI.

Siyan Zhao, Zhihui Xie, Mengchen Liu, Jing Huang, Guan Pang, Feiyu Chen, and Aditya Grover. Self-Distilled Reasoner: On-Policy Self-Distillation for Large Language Models. In *Forty-third International Conference on Machine Learning*, 2026a. URL https://openreview.net/forum?id=Jpxfof0EaS.

Ziqi Zhao, Xinyu Ma, Liu Yang, Yujie Feng, Daiting Shi, Jingzhou He, Xin Xin, Zhaochun Ren, and Xiao-Ming Wu. ROSD: Reflective on-policy self-distillation for language model reasoning across domains. *arXiv preprint arXiv:2605.28014*, 2026b.

# A LIMITATIONS

Our study is intentionally narrow. All target-side training and evaluation tasks are mathematical, and the experiments use only Qwen3-1.7B, Qwen3-4B, and Qwen3-8B. Each trained condition is represented by one training run. The four seeds measure sampling variation for a fixed trained model but do not capture training-run variation. This limitation is particularly important for the fixed, corrupted, and verbose contexts, which are obtained from separate training runs. Finally, we observe deterioration under longer $\mathrm{OP}^2\mathrm{SD}$ training, showing that a paired target solution is not required for this failure mode, but we do not determine why longer OPSD-family training degrades. Establishing that mechanism will require repeated training runs and controlled token-level interventions.

# B RELATED WORK

**On-policy distillation.** Knowledge distillation conventionally trains a student to match a teacher's predictive distribution (Hinton et al., 2014). Sequence-level distillation instead trains an autoregressive student on sequences decoded by the teacher (Kim & Rush, 2016), but fixed teacher trajectories do not cover all prefixes that the student encounters at inference time. MiniLLM addresses this mismatch by optimizing a reverse-KL objective on student samples (Gu et al., 2024). Generalized Knowledge Distillation (GKD) likewise queries the teacher on student-generated sequences and supports different divergences and mixtures of on-policy student outputs and fixed off-policy sequences (Agarwal et al., 2024). DistiLLM combines a skew-KL objective with an adaptive off-policy scheme for using student-generated outputs more efficiently (Ko et al., 2024).

**Reasoning supervision and privileged information.** One way to transfer reasoning is to use teacher rationales as explicit supervised targets, as in Distilling Step-by-Step (Hsieh et al., 2023). OPSD instead evaluates the student's own trajectory with a token-level teacher distribution. Its asymmetric context is closely related to learning using privileged information, where additional training-time features are unavailable at inference (Vapnik & Vashist, 2009; Lopez-Paz et al., 2015). In OPSD, a copy of the model receives the verified solution to the current problem while the student receives only the problem (Zhao et al., 2026a). Related work on privileged-information distillation studies action-only transfer in multi-turn agentic environments, proposing both a jointly trained privileged teacher–student objective and a reverse-KL-regularized on-policy RL alternative, which the authors also call OPSD (Penaloza et al., 2026). These approaches retain privileged information that is relevant to the current instance or trajectory.

**Context-conditioned self-teachers.** Context distillation trains a model to retain behavior induced by instructions, examples, or scratchpads after that context is removed (Snell et al., 2022). Self-Distillation Fine-Tuning (SDFT) brings this idea on-policy, using a demonstration-conditioned copy of the model to provide supervision for continual skill and knowledge acquisition (Shenfeld et al., 2026). On-Policy Context Distillation (OPCD) similarly distills a context-conditioned teacher, but constructs its context from extracted experience or optimized instructions (Ye et al., 2026). OPCD reports that directly substituting raw previous-problem traces for extracted experience can reduce math validation accuracy. Self-Supervised On-Policy Distillation (SSOPD), by contrast, needs no external solution trace: within a mixed group of rollouts, it conditions the teacher on a short correct completion and distills that distribution into prefixes of a persistent wrong completion (Tan & Hong, 2026).

**Recent analyses and variants of OPSD.** Several recent variants reconsider which parts of the privileged teacher signal should be distilled. Purified OPSD decomposes the teacher distribution into a reference-induced component and a question-conditioned component intended to remain useful at inference time, and suppresses the former using an additional reference-only teacher (Shen et al., 2026). ROSD argues that globally imitating a reference-conditioned teacher may encourage trajectory imitation rather than correction of the student's specific error, and instead applies distillation locally around an identified error (Zhao et al., 2026b). PW-OPSD similarly questions uniform token-level supervision and assigns position-dependent weights according to the estimated reliability of the privileged teacher (Liu et al., 2026). These methods modify or filter the signal obtained from the target-aligned reference solution. Kaur et al. (2026) report that privileged-context on-policy distillation can degrade the long-budget reasoning performance of thinking models, particularly when

the teacher receives a full reference solution. These methods modify or filter supervision from a target-aligned reference. In contrast, OP$^2$SD retains the same distillation rule and intervenes on whether the worked solution is paired with the target.

## C Evaluation Metrics and Uncertainty

Let $M$ be the number of evaluation problems, $S$ the number of decoding seeds, and $N$ the number of generations per problem and seed. In our main experiments, $M = 30$, $S = 4$, $N = 12$, and hence each problem has $R = SN = 48$ pooled generations.

We report Avg@12 together with its corrected Monte Carlo standard error, computed as follows:

$$\text{Monte Carlo standard error} = \frac{100}{M}\sqrt{\frac{1}{R-1}\sum_{i=1}^{M}\hat{p}_i(1-\hat{p}_i)}, \tag{5}$$

where $M$ is the number of problems, $R$ is the number of generations per problem, and $\hat{p}_i$ is the fraction of correct generations for problem $i$. In particular, $M = 30, R = 4 \times 12$ in our experiments.

Pass@12 is computed separately for each seed as the percentage of problems with at least one correct generation among that seed's $N = 12$ outputs. Vote@12 is likewise computed per seed as the percentage of problems for which the most frequent extracted, formatted answer among the 12 outputs is graded correct. Let $Z_s$ denote either the seed-level percentage and let $\bar{Z} = S^{-1}\sum_{s=1}^{S} Z_s$. We report $\bar{Z}$ together with the empirical decoding-seed standard error

$$\widehat{\mathrm{SE}}_{\mathrm{seed}}(\bar{Z}) = \sqrt{\frac{1}{S(S-1)}\sum_{s=1}^{S}(Z_s-\bar{Z})^2}, \tag{6}$$

which is the sample standard deviation across decoding seeds divided by $\sqrt{S}$. These uncertainty estimates describe stochastic decoding for a fixed trained model and fixed benchmark. They do not include variation across training runs or uncertainty from sampling a different problem set.

## D Detailed Experimental Settings

All experiments are conducted with H100 $\times 4$ GPUs.

Tables 6–8 summarize the model-specific training configurations.

Table 6: Qwen3-1.7B training configuration shared by OPSD and OP$^2$SD.

| Component | Setting |
|---|---|
| Base model | Qwen3-1.7B |
| Training dataset | `siyanzhao/Openthoughts_math_30k_opsd` |
| Training seed | 42 |
| Per-device batch / gradient accumulation | 4 / 2 |
| Student and teacher mode | Non-thinking / thinking |
| Maximum student completion | 1,024 tokens |
| Student sampling | temperature 1.1, top-$p$ 0.95, top-$k$ 20 |
| Learning rate | $5 \times 10^{-6}$, linear decay |
| Maximum gradient norm | 0.1 |
| Adapter | LoRA $r = 64$, $\alpha = 128$ |
| Teacher | Frozen base model; LoRA disabled |
| Objective | Forward KL, temperature 1.1, $\beta = 0$ |
| KL clip | $\tau = 0.05$ |

Table 7: Qwen3-4B non-thinking training configuration shared by OPSD and $\mathrm{OP}^2\mathrm{SD}$.

| Component | Setting |
|---|---|
| Base model | Qwen3-4B |
| Training dataset | `siyanzhao/Openthoughts_math_30k_opsd` |
| Training seed | 42 |
| Per-device batch / gradient accumulation | 4 / 2 |
| Student and teacher mode | Non-thinking / non-thinking |
| Maximum student completion | 1,024 tokens |
| Student sampling | temperature 1.1, top-$p$ 0.95, top-$k$ 20 |
| Learning rate | $5 \times 10^{-6}$, constant |
| Maximum gradient norm | 0.1 |
| Adapter | LoRA $r = 64$, $\alpha = 128$ |
| Teacher | Frozen base model; LoRA disabled |
| Objective | Forward KL, temperature 1.1, $\beta = 0$ |
| KL clip | $\tau = 10^{-6}$ |

Table 8: Qwen3-8B non-thinking training configuration shared by OPSD and $\mathrm{OP}^2\mathrm{SD}$.

| Component | Setting |
|---|---|
| Base model | Qwen3-8B |
| Training dataset | `siyanzhao/Openthoughts_math_30k_opsd` |
| Training seed | 42 |
| Per-device batch / gradient accumulation | 2 / 4 |
| Student and teacher mode | Non-thinking / non-thinking |
| Maximum student completion | 1,024 tokens |
| Student sampling | temperature 1.1, top-$p$ 0.95, top-$k$ 20 |
| Learning rate | $5 \times 10^{-6}$, constant |
| Maximum gradient norm | 0.1 |
| Adapter | LoRA $r = 64$, $\alpha = 128$ |
| Teacher | Frozen base model; LoRA disabled |
| Objective | Forward KL, temperature 1.1, $\beta = 0$ |
| KL clip | $\tau = 10^{-7}$ |

## E Details of the Padding

The main experiments follow the implementation of OPSD (Zhao et al., 2026a). Student prompts are right-padded to the longest student prompt in each minibatch, while teacher prompts are padded independently to the longest teacher prompt. The sampled response is then appended after the padded prompt. Consequently, shorter examples contain masked padding between the prompt and response, and the amount of padding can differ between the student and teacher inputs. Although these padding tokens are masked out of attention, they affect the position from which the response is scored. We therefore confirm the matched non-thinking Qwen3-4B comparison using left-padded student and teacher prompts with the same checkpoint as Table 1. This shifts the padding to the start of each sequence, restoring adjacency between the unpadded prompt and the sampled response for both policies. This intervention removes padding from the boundary between the prompt and response.

Table 9: **Qwen3-4B non-thinking results with left-padded loss replay.** $\mathrm{OP}^2\mathrm{SD}$ retains a substantial improvement after right padding is removed.

| Benchmark | Method | Avg@12 | Pass@12 | Vote@12 |
|---|---|---|---|---|
| AIME 2024 | Base | $23.19 \pm 0.72$ | $50.83 \pm 1.60$ | $33.33 \pm 2.36$ |
| | OPSD | $27.78 \pm 0.88$ | $64.17 \pm 1.60$ | $52.50 \pm 0.83$ |
| | $\mathrm{OP}^2\mathrm{SD}$ | $41.11 \pm 0.89$ | $73.33 \pm 1.36$ | $57.50 \pm 0.83$ |
| AIME 2025 | Base | $21.11 \pm 0.64$ | $46.67 \pm 2.72$ | $26.67 \pm 2.36$ |
| | OPSD | $19.44 \pm 0.53$ | $36.67 \pm 2.36$ | $23.33 \pm 2.36$ |
| | $\mathrm{OP}^2\mathrm{SD}$ | $33.89 \pm 0.71$ | $55.83 \pm 2.85$ | $41.67 \pm 0.96$ |
| HMMT 2025 | Base | $11.67 \pm 0.49$ | $22.50 \pm 2.10$ | $16.67 \pm 0.00$ |
| | OPSD | $12.85 \pm 0.49$ | $29.17 \pm 2.50$ | $16.67 \pm 0.00$ |
| | $\mathrm{OP}^2\mathrm{SD}$ | $17.36 \pm 0.64$ | $45.00 \pm 3.97$ | $22.50 \pm 1.60$ |

As shown in Table 9, $\mathrm{OP}^2\mathrm{SD}$ continues to yield substantial improvements under left padding, whereas the gains from OPSD are reduced, especially in AIME 2025. Therefore, this setting also supports our main claim that access to the exact reference solution paired with the target is unnecessary.

## F Detail of Result of Table 4

The worked-problem domain control uses a deterministic split of Omni-MATH (Gao et al., 2025), which does not provide an official train/test split for this experiment. We select 1,280 Algebra-tagged, non-Geometry training targets and 30 disjoint held-out targets, balanced across lower, middle, and upper annotated difficulty. Each $\mathrm{OP}^2\mathrm{SD}$ arm uses a 320-example worked-example pool: Algebra/non-Geometry for $\mathrm{OP}^2\mathrm{SD}$-ID and Geometry-only for $\mathrm{OP}^2\mathrm{SD}$-OOD. Worked-examples share source provenance and are aligned by difficulty and Qwen-tokenized context length; both use the same deterministic target-to-context-example index mapping. The 1.7B setting uses training seed 42, 100 updates, an effective batch of 32, and the remaining optimization settings in Table 6. This gives 3,200 target exposures, or 2.5 passes over the target set. We evaluate checkpoint 100 with 12 samples per problem and 4 seeds.

We repeat the control with Qwen3-4B while disabling thinking for the student, teacher, and evaluation. Table 10 reports the complete results for both model settings.

With Qwen3-1.7B, $\mathrm{OP}^2\mathrm{SD}$-ID exceeds $\mathrm{OP}^2\mathrm{SD}$-OOD by only $0.63$ points in Avg@12 and wins, ties, and loses on 6, 19, and 5 targets, respectively. The ordering reverses with Qwen3-4B, where OOD is $1.67$ points higher. This apparent OOD advantage is concentrated in one target asking about the intersections of a circle and a parabola: OOD solves it in 30 of 48 samples, whereas ID solves it in none. Although labeled Algebra, the problem lies near the boundary with analytic Geometry. After excluding it, ID obtains $59.41\%$ Avg@12 and OOD obtains $58.98\%$. We therefore find no consistent evidence that matching these coarse domain labels is necessary.

Table 10: **Complete coarse domain-matching control on a held-out Omni-MATH Algebra split.** We evaluate Qwen3-1.7B and Qwen3-4B on 30 held-out Algebra problems, comparing teacher-only contexts drawn from either Algebra ($OP^2SD$-ID) or Geometry ($OP^2SD$-OOD). The worked-example pools are matched by source, difficulty, and context length. For Qwen3-1.7B, $OP^2SD$-ID exceeds $OP^2SD$-OOD by only 0.63 Avg@12 points, whereas for Qwen3-4B, $OP^2SD$-OOD is 1.67 points higher. The latter difference is driven by a single problem near the boundary between Algebra and analytic Geometry; excluding it reverses the ordering. Overall, the results provide no consistent evidence that matching the teacher-only context to the target's coarse mathematical domain is necessary.

| Method | Teacher-only context | Avg@12 | Pass@12 | Vote@12 |
|---|---|---|---|---|
| **Qwen3-1.7B** | | | | |
| Base | None | $64.72 \pm 0.65$ | $77.50 \pm 1.60$ | $70.00 \pm 2.36$ |
| OPSD | Current target solution | $66.74 \pm 0.56$ | $76.67 \pm 0.00$ | $70.00 \pm 1.36$ |
| $OP^2SD$-ID | Another Algebra problem and solution | $68.06 \pm 0.59$ | $80.83 \pm 0.83$ | $70.00 \pm 1.36$ |
| $OP^2SD$-OOD | Another Geometry problem and solution | $67.43 \pm 0.56$ | $76.67 \pm 0.00$ | $70.83 \pm 0.83$ |
| **Qwen3-4B** | | | | |
| Base | None | $56.60 \pm 0.76$ | $79.17 \pm 2.10$ | $65.83 \pm 2.50$ |
| OPSD | Current target solution | $55.69 \pm 0.65$ | $70.83 \pm 1.60$ | $60.83 \pm 1.60$ |
| $OP^2SD$-ID | Another Algebra problem and solution | $57.43 \pm 0.71$ | $79.17 \pm 0.83$ | $64.17 \pm 1.60$ |
| $OP^2SD$-OOD | Another Geometry problem and solution | $59.10 \pm 0.73$ | $80.00 \pm 1.36$ | $62.50 \pm 0.83$ |

Table 11: **Target-only and Answer-only controls across student–teacher mode settings.** For Qwen3-1.7B, the student operates in non-thinking mode while the teacher evaluates the same trajectory in thinking mode. Under this asymmetric setting, both Target-only and Answer-only improve Avg@12 over Base on all three benchmarks, and revealing the verified final answer provides no consistent advantage over Target-only. For Qwen3-4B, both student and teacher use non-thinking mode; in this matched setting, neither control improves over Base, although Answer-only substantially outperforms Target-only. These results indicate that the gains of the 1.7B controls can arise without a worked-solution context and may depend on the student and teacher mode asymmetry.

| Model | Benchmark | Base | Target-only | Answer-only | OPSD | $OP^2SD$ |
|---|---|---|---|---|---|---|
| Qwen3-1.7B | AIME 2024 | $49.86 \pm 0.94$ | $54.51 \pm 0.86$ | $53.82 \pm 0.88$ | $\mathbf{55.42 \pm 0.82}$ | $55.35 \pm 0.83$ |
| | AIME 2025 | $37.36 \pm 0.81$ | $\mathbf{41.81 \pm 0.76}$ | $40.07 \pm 0.77$ | $40.35 \pm 0.77$ | $40.69 \pm 0.74$ |
| | HMMT 2025 | $23.61 \pm 0.64$ | $26.67 \pm 0.64$ | $24.38 \pm 0.63$ | $25.76 \pm 0.63$ | $\mathbf{27.57 \pm 0.66}$ |
| Qwen3-4B | AIME 2024 | $23.19 \pm 0.72$ | $1.18 \pm 0.28$ | $18.68 \pm 0.55$ | $30.76 \pm 0.77$ | $\mathbf{31.53 \pm 0.80}$ |
| | AIME 2025 | $21.11 \pm 0.64$ | $0.49 \pm 0.18$ | $17.85 \pm 0.55$ | $23.06 \pm 0.65$ | $\mathbf{30.62 \pm 0.71}$ |
| | HMMT 2025 | $11.67 \pm 0.49$ | $0.76 \pm 0.22$ | $9.24 \pm 0.49$ | $15.42 \pm 0.57$ | $\mathbf{16.11 \pm 0.60}$ |

## G Target-Only and Answer-Only Teacher Controls

Following Zhao et al. (2026a), in the Qwen3-1.7B setting, the student generates in non-thinking mode while the teacher scores the same trajectory in thinking mode. This asymmetry suggests an alternative explanation for the gain: the student may learn from the behavior of a stronger inference mode even when the teacher has no access to privileged solution information. We test this explanation with two controls. Namely, in the Target-only setting, the teacher receives only the target problem, whereas in the Answer-only setting, the teacher additionally receives the verified final answer $y_i^\star$ but withholds its derivation. The student prompt remains unchanged in each condition.

Target-only remains above Base on all three Qwen3-1.7B. Adding the verified answer does not improve over Target-only; its Avg@12 point estimate is lower by 0.69, 1.74, and 2.29 percentage points on AIME 2024, AIME 2025, and HMMT 2025, respectively. Thus, a substantial part of the 1.7B improvement does not require a worked solution and is consistent with generic on-policy distillation or transfer from the teacher's thinking mode. This control does not establish that thinking capability itself is transferred: the resulting models are evaluated with thinking enabled, and the comparison does not isolate the mode change from the generic self-distillation.

In contrast, in Qwen3-4B, in which both student and teacher are non-thinking, the results show a very different pattern: Target-only substantially underperforms Base, and Answer-only also falls below Base on all three benchmarks, while both OPSD and $\mathrm{OP}^2\mathrm{SD}$ outperform Base. In this configuration, successful distillation is therefore associated with a complete mathematical worked-solution context, but not with exact target alignment: $\mathrm{OP}^2\mathrm{SD}$ provides a solution to a different problem and still improves. We regard this contrast as descriptive evidence only, since it does not isolate the factor responsible for the difference across model settings.

## H Effect of Extended Training

We further train $\mathrm{OP}^2\mathrm{SD}$ for 400 updates and evaluate its final checkpoint on AIME 2024. Table 12 reports the four-seed averages for Base and the 100- and 400-update $\mathrm{OP}^2\mathrm{SD}$ conditions. The 400-update checkpoint obtains $51.39$ Avg@12, a decrease of $3.96$ percentage points from the 100-update result. It nevertheless remains $1.53$ points above Base. The change is not a uniform collapse: Pass@12 increases from $75.83$ to $79.17$, and answer-format validity remains above $99\%$, whereas Vote@12 falls from $70.00$ to $64.17$. The observed deterioration is therefore specific to sampled accuracy and voting consensus in this evaluation. Kaur et al. (2026) reported that standard on-policy distillation can degrade with prolonged training, particularly when the teacher receives a complete reference solution. Although $\mathrm{OP}^2\mathrm{SD}$ provides the teacher with a complete mathematical solution, that solution corresponds to a different problem rather than the target itself. Similar to OPSD, $\mathrm{OP}^2\mathrm{SD}$'s performance declines with extended training.

Table 12: **Long-horizon checkpoint sensitivity of $\mathrm{OP}^2\mathrm{SD}$ on AIME 2024 with Qwen3-1.7B.** Extending training from 100 to 400 updates reduces Avg@12 from $55.35$ to $51.39$ and Vote@12 from $70.00$ to $64.17$, although the 400-update checkpoint remains above Base in Avg@12. In contrast, Pass@12 and answer-format validity increase at the later checkpoint, indicating that the decline is not uniform across metrics.

| Condition | Avg@12 | Pass@12 | Vote@12 | Format |
|---|---|---|---|---|
| Base | $49.86 \pm 0.94$ | 76.67 | 70.00 | 98.68 |
| $\mathrm{OP}^2\mathrm{SD}$, checkpoint 100 | $55.35 \pm 0.83$ | 75.83 | 70.00 | 99.03 |
| $\mathrm{OP}^2\mathrm{SD}$, checkpoint 400 | $51.39 \pm 0.90$ | 79.17 | 64.17 | 99.51 |

## I Problem-Level Analysis of Accuracy Gains

In this section, we analyze the performance gains of OPSD and $\mathrm{OP}^2\mathrm{SD}$ over Base on a problem-by-problem basis. Aggregate Avg@12 can improve either (a) because a method solves problems that Base never solves or (b) because it increases the probability of producing a correct answer on problems that Base already solves occasionally. We examine these possibilities for Qwen3-1.7B by pooling results across the four seeds. For each problem $i$ and method $m$, let $c_i^{(m)} \in \{0, \ldots, 48\}$ denote the number of correct responses among 12 samples under each of four seeds. Table 13 reports the Pearson correlation between these problem-level correct counts for every pair of methods. It also reports how often the first method has a higher, equal, or lower count than the second method across all 90 benchmark problems. For methods $a$ and $b$ over $N$ problems, we define

$$r(a,b) = \frac{\sum_{i=1}^{N}(c_i^{(a)} - \bar{c}^{(a)})(c_i^{(b)} - \bar{c}^{(b)})}{\sqrt{\sum_{i=1}^{N}(c_i^{(a)} - \bar{c}^{(a)})^2}\sqrt{\sum_{i=1}^{N}(c_i^{(b)} - \bar{c}^{(b)})^2}},$$

where $\bar{c}^{(m)}$ is the mean correct count of method $m$ over those problems.

The methods largely agree on which problems are easy and difficult. Most notably, OPSD and $\mathrm{OP}^2\mathrm{SD}$ have a correlation of $0.990$ after pooling all 90 problems, and $\mathrm{OP}^2\mathrm{SD}$ wins, ties, and loses on 31, 39, and 20 problems, respectively. Target-only and $\mathrm{OP}^2\mathrm{SD}$ are similarly correlated at $0.989$. Replacing or removing the paired target solution, therefore, does not substantially reorder problem-level difficulty in these runs. This agreement does not establish a shared reasoning mechanism:

Table 13: **Problem-level agreement among Qwen3-1.7B methods.** Pearson correlations r are computed from correct counts out of 48 responses per problem. "All" pools the 90 problems from the three benchmarks. W/T/L compares the first-named method with the second across those 90 problems.

| Pair (first vs. second) | AIME24 $r$ | AIME25 $r$ | HMMT25 $r$ | All $r$ | W/T/L |
|---|---|---|---|---|---|
| OPSD vs. Base | 0.942 | 0.974 | 0.974 | 0.966 | 39/30/21 |
| OP$^2$SD vs. Base | 0.953 | 0.976 | 0.970 | 0.969 | 40/31/19 |
| Target-only vs. Base | 0.967 | 0.964 | 0.983 | 0.973 | 39/33/18 |
| OP$^2$SD vs. OPSD | 0.988 | 0.987 | 0.992 | 0.990 | 31/39/20 |
| Target-only vs. OPSD | 0.970 | 0.967 | 0.978 | 0.973 | 26/42/22 |
| Target-only vs. OP$^2$SD | 0.989 | 0.989 | 0.988 | 0.989 | 25/36/29 |

common benchmark difficulty can itself produce high correlations, and the pooled correlation also contains between-benchmark variation. We next bin problems by the observed Base count $c_i^{(\mathrm{Base})}$. Table 14 reports the mean change $c_i^{(m)} - c_i^{(\mathrm{Base})}$ within each bin. The unit is additional correct responses per problem out of 48, rather than percentage points.

Table 14: **Accuracy-mass shifts by Base success count.** Problems are binned by the number of correct Base responses out of 48. Each method column gives the mean number of additional correct responses per problem relative to Base.

| Base correct count | Problems | OPSD | OP$^2$SD | Target-only |
|---|---|---|---|---|
| 0 | 22 | $+0.14$ | $+0.14$ | $+0.14$ |
| 1–4 | 13 | $+0.54$ | $+0.85$ | $+0.46$ |
| 5–12 | 12 | $+1.25$ | $+0.92$ | $+0.92$ |
| 13–24 | 11 | $+5.00$ | $+5.09$ | $+4.64$ |
| 25–36 | 12 | $+4.50$ | $+7.08$ | $+7.58$ |
| 37–47 | 10 | $+2.50$ | $+2.50$ | $+1.70$ |
| 48 | 10 | $-0.50$ | $-0.70$ | $-0.40$ |
| All | 90 | $+1.71$ | $+2.04$ | $+1.94$ |

Table 15: **Accuracy-mass shifts by Base success count for Qwen3-4B non-thinking.** Problems are binned by the number of correct Base responses out of 48. Each method column gives the mean number of additional correct responses per problem relative to Base.

| Base correct count | Problems | OPSD | OP$^2$SD |
|---|---|---|---|
| 0 | 45 | $+1.13$ | $+3.22$ |
| 1–4 | 14 | $+3.64$ | $+2.07$ |
| 5–12 | 10 | $+4.10$ | $+8.00$ |
| 13–24 | 6 | $-2.33$ | $+5.50$ |
| 25–36 | 3 | $+9.33$ | $+5.67$ |
| 37–47 | 12 | $+2.83$ | $+1.42$ |
| 48 | 0 | — | — |
| All | 90 | $+2.12$ | $+3.57$ |

All three distilled methods obtain their largest gains on problems for which Base already succeeds at an intermediate rate. By contrast, each method adds only three correct responses in total across the 22 Base-zero problems, or $0.14$ responses per problem. The main empirical finding here is that the method increases the likelihood of producing a correct answer that was already accessible, rather than consistently uncovering solutions that Base sampling could not generate. This descriptive analysis does not rule out individual newly solved problems.

Table 16: **Problem-level agreement among Qwen3-4B non-thinking methods.** Pearson correlations are computed from correct counts out of 48 responses per problem. "All" pools the 90 problems from the three benchmarks. W/T/L compares the first-named method with the second across those 90 problems.

| Pair (first vs. second) | AIME24 $r$ | AIME25 $r$ | HMMT25 $r$ | All $r$ | W/T/L |
|---|---|---|---|---|---|
| OPSD vs. Base | 0.938 | 0.975 | 0.928 | 0.948 | 42/36/12 |
| OP$^2$SD vs. Base | 0.838 | 0.831 | 0.966 | 0.870 | 43/34/13 |
| OP$^2$SD vs. OPSD | 0.883 | 0.857 | 0.922 | 0.882 | 37/29/24 |

## J Matched Thinking-Mode Control

The main Qwen3-1.7B experiments use a non-thinking student rollout and a thinking teacher. We test whether the target-solution OPSD gain persists after removing this mode asymmetry. In this control, both the student rollout and the frozen teacher use thinking mode during training; the training dataset, optimization budget, and distillation settings are otherwise unchanged. All conditions in Table 17 are evaluated in thinking mode on AIME 2024.

Table 17: **Qwen3-1.7B matched thinking-mode control on AIME 2024.** The both-thinking condition does not reproduce the Avg@12 gain observed with the asymmetric training configuration. Accuracy metrics are four-seed means in percentages. Avg@12 is reported with corrected Monte Carlo standard error; Pass@12 and Vote@12 are reported with empirical standard error.

| Condition | Student | Teacher | Avg@12 | Pass@12 | Vote@12 | Format |
|---|---|---|---|---|---|---|
| Base | – | – | $49.86 \pm 0.94$ | $76.67 \pm 0.00$ | $70.00 \pm 2.72$ | 98.68 |
| OPSD | Non-thinking | Thinking | $\mathbf{55.42 \pm 0.82}$ | $78.33 \pm 0.96$ | $65.00 \pm 1.67$ | 99.17 |
| OPSD | Thinking | Thinking | $49.24 \pm 0.91$ | $77.50 \pm 0.83$ | $67.50 \pm 1.60$ | 85.90 |

The both-thinking condition obtains $49.24$ Avg@12, $0.62$ percentage points below Base and $6.18$ points below the asymmetric OPSD condition. It therefore does not reproduce the main 1.7B improvement. Pass@12 remains close to the other conditions, but per-sample accuracy and answer-format validity are lower; in particular, format validity falls to $85.90\%$. This result suggests that the observed 1.7B benefit is not invariant to the student–teacher mode pairing. Because each condition uses one training seed, this control does not by itself identify whether the difference is caused by mode asymmetry, rollout-state changes, or optimization variance.

## K OP$^2$SD with physics References

Using Qwen3-4B in non-thinking mode, we test whether the teacher context remains effective when drawn from a non-mathematical domain. We replace the teacher-only context pool with 2,768 problem–solution pairs sampled from the external CAMEL Physics dataset (Li et al., 2023). This matches the number of worked examples in the mathematics dataset. Both the student and teacher operate in non-thinking mode, and the two OP$^2$SD conditions are evaluated at the same checkpoint.

The following is a teacher prompt made during training.

**Actual teacher prompt with a physics reference**

```
Target mathematics problem:
Find $4^{-1} \pmod{35}$, as a residue modulo 35.
(Give an answer between 0 and 34, inclusive.)

Below is a worked example from a different and unrelated physics
problem. It is provided only to demonstrate careful scientific
reasoning and solution structure. It is not a solution or hint for the
target mathematics problem above. Do not reuse its final answer,
problem-specific quantities, units, laws, assumptions, or
```

```
domain-specific facts.

=== Unrelated physics Worked Example Problem Begin ===
A sample of radioactive material has an initial mass of 1 gram and a
half-life of 10 days. After 30 days, what will be the mass of the sample?
=== Unrelated physics Worked Example Problem End ===
=== Unrelated physics Worked Example Solution Begin ===
To find the mass of the radioactive material after 30 days, we can use
the formula:

Final mass = Initial mass * (1/2)^(time elapsed / half-life)

In this case, the initial mass is 1 gram, the half-life is 10 days, and
the time elapsed is 30 days. Plugging these values into the formula,
we get:

Final mass = 1 * (1/2)^(30 / 10)
Final mass = 1 * (1/2)^3
Final mass = 1 * (1/8)
Final mass = 0.125 grams

After 30 days, the mass of the radioactive sample will be 0.125 grams.
=== Unrelated physics Worked Example Solution End ===

Now solve the target mathematics problem independently. Use the worked
example only as a general demonstration of clear, step-by-step
problem-solving. Please reason step by step, verify your reasoning, and
put your final answer within \boxed{}.
```

Table 18: $OP^2SD$ **with physics problems.** Replacing worked-problem mathematics solutions with physics solutions does not retain the $OP^2SD$ gain on any of the three benchmarks and substantially reduces valid-answer formatting.

| Benchmark | Condition | Avg@12 | Pass@12 | Vote@12 | Format |
|---|---|---|---|---|---|
| AIME 2024 | Base | $23.19 \pm 0.72$ | $50.83 \pm 1.60$ | $33.33 \pm 2.36$ | 97.57 |
| | $OP^2SD$ (mathematics) | $\mathbf{31.53 \pm 0.80}$ | $60.83 \pm 1.60$ | $43.33 \pm 1.36$ | 89.58 |
| | $OP^2SD$ (physics) | $18.68 \pm 0.69$ | $40.83 \pm 1.60$ | $36.67 \pm 1.36$ | 46.04 |
| AIME 2025 | Base | $21.11 \pm 0.64$ | $46.67 \pm 2.72$ | $26.67 \pm 2.36$ | 99.17 |
| | $OP^2SD$ (mathematics) | $\mathbf{30.62 \pm 0.71}$ | $55.00 \pm 2.15$ | $40.83 \pm 1.60$ | 93.61 |
| | $OP^2SD$ (physics) | $20.21 \pm 0.61$ | $44.17 \pm 2.10$ | $29.17 \pm 2.50$ | 52.71 |
| HMMT 2025 | Base | $11.67 \pm 0.49$ | $22.50 \pm 2.10$ | $16.67 \pm 0.00$ | 99.58 |
| | $OP^2SD$ (mathematics) | $\mathbf{16.11 \pm 0.60}$ | $38.33 \pm 2.15$ | $20.00 \pm 1.92$ | 94.93 |
| | $OP^2SD$ (physics) | $8.68 \pm 0.50$ | $20.83 \pm 2.10$ | $15.00 \pm 0.96$ | 50.76 |

The physics-reference condition obtains $18.68$, $20.21$, and $8.68$ Avg@12 on AIME 2024, AIME 2025, and HMMT 2025, respectively. These values are below Base by $4.51$, $0.90$, and $2.99$ percentage points, and below $OP^2SD$ with mathematics references by $12.85$, $10.41$, and $7.43$ points. Its Pass@12 is also below both comparison conditions on all three benchmarks, while valid-answer formatting falls to approximately half of the generations. Thus, the failure to retain the mathematics reference gain is consistent across the three evaluations rather than being specific to AIME 2024. In this configuration, an arbitrary worked solution from a different scientific subject is not sufficient to reproduce the benefit of the other-problem mathematical context.

This result does not isolate which property of the physics context causes the decline. The mathematics and physics pools differ in source, subject matter, notation, units, solution style, context length, and subject-specific prompt wording. Moreover, the CAMEL solutions are synthetic and filtered but not independently verified, and the external physics assignment is deterministic but not pairwise aligned. Accordingly, we treat this as a descriptive cross-subject ablation rather than evidence that mathematical semantics alone explain $OP^2SD$.

# L EXACT FIXED MATHEMATICAL CONTEXTS

Table 3 compares three versions of one fixed mathematical context. All three conditions use the same auxiliary problem. Within a condition, the displayed problem–solution pair is inserted verbatim into the worked-problem fields of the OP$^2$SD teacher prompt and is repeated across all training rows. This appendix reproduces the exact textual content of the `problem` and `solution` fields used in those runs; line wrapping below is typographic.

## L.1 SHARED FIXED PROBLEM

**Fixed auxiliary problem used in all three conditions**

```
Let $a, b, c$ be positive integers such that
$\frac{a}{77}+\frac{b}{91}+\frac{c}{143}=1$.
What is the smallest possible value of $a+b+c$?
```

## L.2 CONCISE CORRECT SOLUTION

This is the 204-token correct solution used by the *Concise correct* condition.

**Concise correct solution (204 tokens)**

```
Multiplying the equation by $1001=7\cdot 11\cdot 13$ gives
$13a+11b+7c=1001$. Let $S=a+b+c$. Then
$13(S-77)=2b+6c=2(b+3c)$. Because $13$ is coprime to $2$, the
positive integer $b+3c$ must be divisible by $13$, so $b+3c\ge 13$.
Hence $S-77=2(b+3c)/13\ge 2$, and therefore $S\ge 79$. Equality is
attained by $b=10$, $c=1$, and $a=68$: indeed,
$13\cdot68+11\cdot10+7\cdot1=1001$. Thus the smallest possible
value is $\boxed{79}$.
```

## L.3 LOCALLY CORRUPTED SOLUTION

The corrupted condition keeps the same problem and nearly the same solution form, but replaces the valid lower bound $b + 3c \geq 13$ with the false claim that positivity requires the *second* positive multiple of 13. It consequently reports the feasible but nonminimal value 81.

**Locally corrupted solution (204 tokens)**

```
Multiplying the equation by $1001=7\cdot 11\cdot 13$ gives
$13a+11b+7c=1001$. Let $S=a+b+c$. Then
$13(S-77)=2b+6c=2(b+3c)$. Because $13$ is coprime to $2$,
positivity makes $b+3c$ at least the second positive multiple of
$13$, namely $26$. Hence $S-77=2(b+3c)/13\ge 4$, and therefore
$S\ge 81$. Equality is attained by $b=23$, $c=1$, and $a=57$.
This bound is sharp. Indeed,
$13\cdot57+11\cdot23+7\cdot1=1001$. Thus the smallest possible
value is $\boxed{81}$.
```

## L.4 VERBOSE CORRECT SOLUTION

The verbose condition retains the correct argument and answer, but expands the derivation to 742 solution tokens.

**Verbose correct solution (742 tokens)**

```
Let
\[
```

```
S=a+b+c.
\]
We remove the fractions and translate the resulting equation into a
restriction on $S$. The least common multiple of $77$, $91$, and $143$
is $1001$. Multiplying the given equation by $1001$ gives
\[
13a+11b+7c=1001. \tag{1}
\]
Because $1001=13\cdot77$, we compare equation (1) with thirteen times
$S$. We have
\[
13S=13a+13b+13c.
\]
Subtracting equation (1) from this identity yields
\[
13S-1001=2b+6c=2(b+3c).
\]
Therefore
\[
13(S-77)=2(b+3c). \tag{2}
\]
Equation (2) implies that $13$ divides $2(b+3c)$. Since
$\gcd(2,13)=1$, it follows that
\[
13\mid(b+3c).
\]
Thus, for some integer $k$,
\[
b+3c=13k. \tag{3}
\]
The variables $b$ and $c$ are positive, so $b+3c$ is positive. It
follows that $k$ is a positive integer and in particular $k\ge1$.

Substituting (3) into (2), we get
\[
13(S-77)=2\cdot13k.
\]
Dividing by $13$ gives
\[
S=77+2k. \tag{4}
\]
Since $k\ge1$, equation (4) proves the universal lower bound
\[
S\ge77+2=79.
\]
It remains to show that equality is attainable. Equality requires
$k=1$, so equation (3) becomes
\[
b+3c=13.
\]
Take $c=1$. Then $b=10$, and equation (4) gives $S=79$. The
remaining variable is
\[
a=S-b-c=79-10-1=68,
\]
which is positive. Hence $(a,b,c)=(68,10,1)$ is an admissible
candidate if it satisfies the original constraint.

We can verify it first in the cleared-denominator equation:
\[
13\cdot68+11\cdot10+7\cdot1=884+110+7=1001.
\]
Equivalently, verification in the original equation gives
\[
\frac{68}{77}+\frac{10}{91}+\frac{1}{143}
=\frac{884}{1001}+\frac{110}{1001}+\frac{7}{1001}
=1.
\]
Thus this triple is valid and its sum is
\[
68+10+1=79.
\]
Equation (4) shows that every possible sum is $77+2k$ for some
positive integer $k$, so no valid triple has sum less than $79$. The
explicit triple above attains $79$. Therefore the smallest possible
value is
\[
\boxed{79}.
```

```
\]
```

### L.5 Trivial Correct Context

The trivial control replaces the fixed Omni-MATH example with the following arithmetic example.

**Teacher prompt with the fixed trivial context**

```
Compute $1+1$.

Adding one and one gives two. Therefore, the final answer is \boxed{2}.
```

## M Fixed-Context Replication with Qwen3-1.7B

We repeat the fixed-context comparison with Qwen3-1.7B. The worked example here is the concise correct `omni02169` problem and solution pair (Section L), and the same pair is supplied to the teacher for every target problem. The student rollout is collected in non-thinking mode, the frozen teacher scores it in thinking mode, and evaluation uses the same thinking-mode protocol as the main Qwen3-1.7B comparison. Table 19 compares this condition with Base, target-solution OPSD, and the standard OP$^2$SD condition that varies the worked-problem example across training targets.

Table 19: **Fixed mathematical context with Qwen3-1.7B.** Repeating one concise, correct mathematical example matches the standard OP$^2$SD condition on AIME 2024 and HMMT 2025 and has a higher Avg@12 point estimate on AIME 2025. Within each benchmark, the highest Avg@12 point estimate is shown in **bold** and the second-highest is underlined.

| Benchmark | Condition | Avg@12 | Pass@12 | Vote@12 |
|---|---|---|---|---|
| AIME 2024 | Base | $49.86 \pm 0.94$ | $76.67 \pm 0.00$ | $70.00 \pm 2.72$ |
| | OPSD | $\underline{55.42 \pm 0.82}$ | $78.33 \pm 0.96$ | $65.00 \pm 1.67$ |
| | OP$^2$SD | $55.35 \pm 0.83$ | $75.83 \pm 0.83$ | $70.00 \pm 2.36$ |
| | OP$^2$SD (fixed) | $\mathbf{55.56 \pm 0.85}$ | $76.67 \pm 1.36$ | $70.00 \pm 1.36$ |
| AIME 2025 | Base | $37.36 \pm 0.81$ | $70.00 \pm 3.04$ | $48.33 \pm 0.96$ |
| | OPSD | $40.35 \pm 0.77$ | $65.00 \pm 2.15$ | $50.00 \pm 1.36$ |
| | OP$^2$SD | $\underline{40.69 \pm 0.74}$ | $63.33 \pm 2.36$ | $52.50 \pm 1.60$ |
| | OP$^2$SD (fixed) | $\mathbf{43.40 \pm 0.73}$ | $65.83 \pm 1.60$ | $53.33 \pm 2.36$ |
| HMMT 2025 | Base | $23.61 \pm 0.64$ | $52.50 \pm 2.50$ | $28.33 \pm 0.96$ |
| | OPSD | $25.76 \pm 0.63$ | $50.83 \pm 2.10$ | $29.17 \pm 0.83$ |
| | OP$^2$SD | $\underline{27.57 \pm 0.66}$ | $51.67 \pm 1.67$ | $30.83 \pm 0.83$ |
| | OP$^2$SD (fixed) | $\mathbf{27.78 \pm 0.66}$ | $55.00 \pm 2.15$ | $31.67 \pm 1.67$ |

The fixed condition obtains 55.56, 43.40, and 27.78 Avg@12 on AIME 2024, AIME 2025, and HMMT 2025. Relative to varying-example OP$^2$SD, the corresponding point-estimate differences are $+0.21$, $+2.71$, and $+0.21$ percentage points. It also matches or exceeds target-solution OPSD on all three benchmarks. The result therefore extends the fixed-context observation beyond the Qwen3-4B non-thinking setting: exposure to a diverse stream of worked examples is not necessary for obtaining an OPSD-like gain in this Qwen3-1.7B run. It does not establish that a fixed example is generally preferable.

## N THINKING-MODE EVALUATION AFTER QWEN3-4B NON-THINKING TRAINING

We test whether their behavior transfers across inference modes by evaluating Base and the same trained models with thinking enabled, without further training.

Table 20: **Thinking-mode evaluation after Qwen3-4B non-thinking training.** $\mathrm{OP}^2\mathrm{SD}$ matches or exceeds OPSD in Avg@12 on all three benchmarks, but neither distilled condition improves uniformly over the thinking-mode Base model. Within each benchmark, the highest Avg@12 point estimate is shown in **bold** and the second-highest is underlined; tied highest values are both bolded.

| Benchmark | Condition | Avg@12 | Pass@12 | Vote@12 |
|---|---|---|---|---|
| AIME 2024 | Base | $\mathbf{72.85 \pm 0.69}$ | $86.67 \pm 0.00$ | $80.83 \pm 0.83$ |
| | OPSD | $\underline{70.90 \pm 0.73}$ | $85.00 \pm 2.15$ | $80.83 \pm 1.60$ |
| | $\mathrm{OP}^2\mathrm{SD}$ | $\mathbf{72.85 \pm 0.70}$ | $88.33 \pm 1.67$ | $79.17 \pm 0.83$ |
| AIME 2025 | Base | $\mathbf{67.36 \pm 0.73}$ | $82.50 \pm 0.83$ | $77.50 \pm 0.83$ |
| | OPSD | $64.03 \pm 0.81$ | $82.50 \pm 1.60$ | $76.67 \pm 1.36$ |
| | $\mathrm{OP}^2\mathrm{SD}$ | $\underline{64.17 \pm 0.82}$ | $83.33 \pm 1.36$ | $78.33 \pm 0.96$ |
| HMMT 2025 | Base | $43.47 \pm 0.72$ | $60.83 \pm 1.60$ | $55.00 \pm 2.89$ |
| | OPSD | $\underline{44.44 \pm 0.74}$ | $68.33 \pm 2.15$ | $55.83 \pm 0.83$ |
| | $\mathrm{OP}^2\mathrm{SD}$ | $\mathbf{45.69 \pm 0.71}$ | $68.33 \pm 2.15$ | $55.83 \pm 1.60$ |

Table 20 shows that the gains under non-thinking evaluation do not consistently transfer when thinking is enabled at inference. This suggests that teacher supervision should be produced in the same mode used for evaluation. Because both the student rollout and the teacher were non-thinking during training, we interpret this result as evidence for mode alignment rather than teacher-mode mismatch alone.