PRISM: When to Trust the Teacher? Prioritized Instructive Signal Modulation for On-Policy Distillation

Anonymous ACL submission

Abstract

On-policy distillation (OPD) has emerged as an effective paradigm for transferring complex reasoning capabilities from strong teacher models to smaller student. Current methods improve OPD by calibrating where teacher signals are applied. However, they fail to distinguish holistic quality differences across sequences and provide fine-grained modulation within them. To address these limitations, we present PRISM, a three-level teacher-trust modulation framework for single-trajectory OPD. At the outcome level, PRISM applies a binary gate that preserves full teacher correction on incorrect trajectories while mitigating intervention on correct ones. At the quality level, it leverages the teacher's token-averaged negative log-likelihood (NLL) to prioritize incorrect trajectories. Specifically, it up-weights low-NLL errors that remain distributionally close to the teacher. At the temporal level, curriculum temperature annealing is employed to sharpen this NLL-based prioritization. This process transitions the training from broad correction early on to targeted refinement as student errors become more differentiated. Empirical results validate that such fine-grained teacher signals lead to significant performance gains. Notably, in 30B-4B distillation, PRISM achieves mean@32 scores of 60.63 and 54.69 on AIME 2024/2025, respectively.

1 Introduction

Knowledge distillation (Buciluǎ et al., 2006; Hinton et al., 2015), originally developed for model compression and later extended to sequence generation (Kim and Rush, 2016; Sanh et al., 2019; Hsieh et al., 2023; Ko et al., 2024), has become a central route for transferring the reasoning capabilities of frontier-scale language models (Achiam et al., 2023; Touvron et al., 2023; Guo et al., 2025; Yang et al., 2024, 2025) into deployable students while reducing inference cost. Conventional offline distillation trains on fixed teacher trajectories or static logits. This creates a distribution mismatch when the student generates its own long reasoning chains (Wen et al., 2023; Xu et al., 2025b). In contrast, on-policy distillation (OPD) mitigates this mismatch by querying the teacher on student-generated rollouts (Gu et al., 2024; Agarwal et al., 2024; Yang et al., 2026), providing supervision on the states the student actually visits.

Recent OPD variants have sought to refine teacher supervision from multiple angles, either by integrating teacher guidance with reinforcement learning objectives (Yang et al., 2026; Xu et al., 2025a; Zhang et al., 2026; Yu et al., 2026) or dynamically filtering teacher signals at the token level (Ko et al., 2026; Jin et al., 2026; Cheng et al., 2026; Jia et al., 2026). Other efforts focus on optimizing distillation targets, context utilization, or training infrastructure (Jang et al., 2026; Ye et al., 2026; Zhao et al., 2026; Sheng et al., 2025). Despite this progress, existing methods mainly decide how or where teacher signals are applied, while often treating complete reasoning trajectories uniformly. In single-trajectory OPD, the central challenge is not only how to construct teacher advantages but also when and on which complete rollouts the teacher should be trusted.

This challenge is especially pronounced in computationally efficient single-trajectory ($n = 1$) OPD. Without multiple candidate responses per prompt, teacher trust must be allocated from the single observed rollout itself. The student's outcome correctness provides a reliable but coarse first signal: correct trajectories should be protected from excessive teacher intervention, whereas incorrect ones require correction. Furthermore, among the student's failed rollouts, coherent near-misses and severe logical failures warrant distinct levels of attention. This lack of granularity is highly problematic for mathematical reasoning. In this domain, benchmark progress depends heavily on long-horizon search, verification, and structured reason-

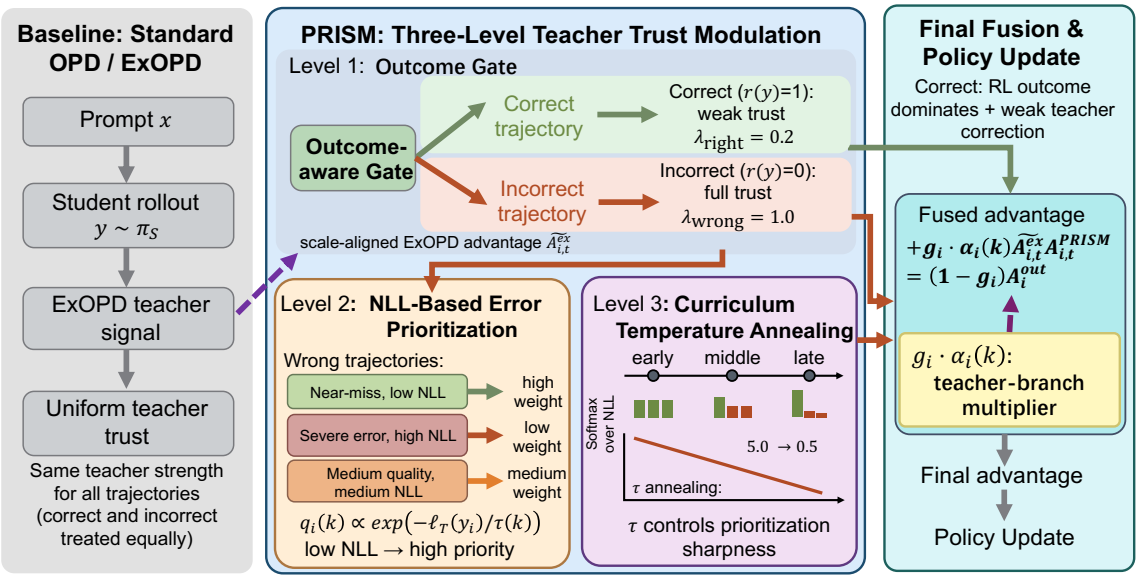

Figure 1: Overview of PRISM. Unlike standard OPD/ExOPD, which applies a uniform teacher signal to all rollouts, PRISM modulates teacher trust through three levels of decision making: outcome, quality, and time. An outcome-aware gate weakens teacher intervention on correct trajectories, NLL-based prioritization focuses correction on coherent near-miss errors, and curriculum temperature annealing gradually sharpens this prioritization during training. The resulting trajectory-level trust coefficient mixes outcome reward with the scale-aligned ExOPD signal for policy optimization.

ing trajectories (Cobbe et al., 2021; Hendrycks et al., 2021).

To address these limitations, we propose Prioritized Instructive Signal Modulation (PRISM), a three-level framework designed for single-trajectory OPD. PRISM dynamically calibrates teacher supervision across three dimensions: outcome, quality, and time. **First**, PRISM applies an outcome-level gate to mitigate teacher intervention on correct student rollouts. This mechanism safeguards the student's existing capabilities. **Second**, for incorrect rollouts, PRISM leverages the teacher's token-averaged negative log-likelihood (NLL) as a holistic trajectory-quality signal. Low-NLL errors indicate that the student's flawed reasoning still closely aligns with the teacher's distribution. Consequently, these errors are prioritized for their higher correction value. **Finally**, this quality signal can be noisy in early training phases due to poor student performance. To address this, PRISM implements a curriculum temperature annealing mechanism. This enables a smooth transition from broad, exploratory correction early on to sharp, targeted refinement of near-miss trajectories as training progresses. In summary, our main contributions are as follows:

- We identify sequence-level teacher trust allocation as a critical missing dimension in single-trajectory OPD, shifting the focus from traditional token-level filtering to holistic trajectory evaluation.
- We propose PRISM, a novel three-level framework that dynamically calibrates teacher trust across outcome, quality, and temporal dimensions through adaptive gating, NLL-based prioritization, and curriculum annealing.
- Through extensive evaluations across two distillation regimes and six benchmarks, we demonstrate that PRISM achieves state-of-the-art performance, particularly on challenging long-horizon reasoning tasks like AIME.

2 Related Work

On-policy distillation. Knowledge distillation transfers large-model behavior to smaller models (Hinton et al., 2015; Gou et al., 2021), with extensions to sequence generation and LLM reasoning (Sanh et al., 2019; Jiao et al., 2020; Hsieh et al., 2023; Ko et al., 2024; Wen et al., 2023). OPD reduces the distribution mismatch of offline methods by querying the teacher on student-generated contexts (Gu et al., 2024; Agarwal et al., 2024), and ExOPD further strengthens correction through

reward extrapolation (Yang et al., 2026). Recent variants refine teacher supervision from multiple angles: KDRL and RLAD integrate teacher guidance with RL objectives or trust-region ratios (Xu et al., 2025a; Zhang et al., 2026; Yu et al., 2026; Sheng et al., 2025); REOPOLD, EOPD, and AOPD reformulate targets or select token positions using entropy, reward clipping, or advantage signs (Jang et al., 2026; Ko et al., 2026; Jin et al., 2026; Jia et al., 2026); context distillation and self-distillation provide complementary on-policy directions (Ye et al., 2026; Zhao et al., 2026). These methods primarily decide *how* or *where* teacher signals are applied at the token level; PRISM instead addresses *which* complete rollouts should receive teacher trust and *when* that trust should sharpen.

Trajectory-quality signals and curriculum learning. For mathematical reasoning, solution quality depends on intermediate reasoning paths and process-level supervision, not only final answers (Cobbe et al., 2021; Hendrycks et al., 2021; Lightman et al., 2024; Yao et al., 2023; He et al., 2024; Rein et al., 2023). Teacher NLL offers a compact trajectory-quality proxy: low-NLL incorrect rollouts are more likely to be coherent near-misses than distributionally divergent failures. However, such quality signals are noisy early in training when student errors are undifferentiated, motivating a curriculum-style schedule (Bengio et al., 2009, 2015). Unlike data curricula that order training examples by difficulty, PRISM schedules the *sharpness* of a teacher-derived quality signal, adding trajectory-level granularity and temporal calibration to OPD.

3 Preliminaries

We briefly define the training signals used by PRISM. For prompt x_i, let $y_i = (y_{i,1}, \ldots, y_{i,L_i})$ be a rollout sampled from the old student policy $\pi_{\theta_{\text{old}}}$, with token context $c_{i,t} = (x_i, y_{i,<t})$. In OPD, the teacher π_T is queried on these student-induced contexts, yielding

$$\begin{aligned} A^{\text{OPD}}_{i,t} &= \log \pi_T(y_{i,t} \mid c_{i,t}) \\ &\quad - \log \pi_{\theta_{\text{old}}}(y_{i,t} \mid c_{i,t}), \end{aligned} \tag{1}$$

PRISM builds on ExOPD (Yang et al., 2026), which extrapolates the teacher direction relative to a fixed reference policy π_{ref}; in our setting, π_{ref} is the fixed student initialization before on-policy training. Its teacher advantage is

$$\begin{aligned} A^{\text{ex}}_{i,t} &= \lambda_{\text{ex}} \log \pi_T(y_{i,t} \mid c_{i,t}) \\ &\quad + (1 - \lambda_{\text{ex}}) \log \pi_{\text{ref}}(y_{i,t} \mid c_{i,t}) \\ &\quad - \log \pi_{\theta_{\text{old}}}(y_{i,t} \mid c_{i,t}), \end{aligned} \tag{2}$$

where $\lambda_{\text{ex}} \geq 0$ is the extrapolation coefficient; standard ExOPD uses $\lambda_{\text{ex}} > 1$. This coefficient is distinct from the trajectory-level gate strengths λ_{right} and λ_{wrong} used by PRISM.

We focus on the single-trajectory setting. Each prompt produces one rollout with outcome $r_i \in \{0, 1\}$ and outcome-derived advantage $A^{\text{out}}_i = r_i$. Unlike multi-sample methods such as GRPO (Shao et al., 2024), this setting provides no within-prompt comparison set for estimating group-relative advantages, so failed rollouts cannot be ranked by reward alone. To distinguish incorrect trajectories, PRISM uses the token-averaged teacher negative log-likelihood (NLL):

$$\ell_T(y_i) = -\frac{1}{L_i} \sum_{t=1}^{L_i} \log \pi_T(y_{i,t} \mid c_{i,t}). \tag{3}$$

Lower ℓ_T indicates higher teacher consistency and suggests a coherent near-miss rather than a chaotic failure.

4 PRISM

This section introduces PRISM, a trajectory-level teacher-trust modulation framework for single-trajectory OPD. PRISM keeps the underlying ExOPD advantage fixed and deterministically modulates how strongly this advantage should be trusted for each rollout. Given the outcome-derived advantage A^{out}_i, the ExOPD advantage $A^{\text{ex}}_{i,t}$, and the teacher NLL score $\ell_T(y_i)$ defined in Section 3, PRISM constructs the training advantage as

$$\begin{aligned} A^{\text{PRISM}}_{i,t} &= (1 - g_i) A^{\text{out}}_i \\ &\quad + g_i \alpha_i(k) \widetilde{A}^{\text{ex}}_{i,t}, \end{aligned} \tag{4}$$

where $g_i \in [0, 1]$ is an outcome-level teacher gate, $\alpha_i(k)$ is a non-negative priority multiplier, and $\widetilde{A}^{\text{ex}}_{i,t}$ is the scale-aligned ExOPD advantage whose formal construction is given in Section 4.4. The gate g_i controls the interpolation between the outcome-derived and teacher-derived branches, while $\alpha_i(k)$ only redistributes teacher correction among incorrect trajectories. This separation prevents a priority multiplier larger than one from being misinterpreted as a convex-combination coefficient.

4.1 Level 1: Outcome-Level Teacher Gate

The first level separates correct and incorrect trajectories. A correct trajectory already reaches the target answer, so strong imitation may unnecessarily pull the student away from its own valid reasoning path. An incorrect trajectory, in contrast, receives no fine-grained corrective information from the outcome-derived advantage alone and therefore requires stronger teacher correction. We define the outcome gate as

$$g_i = \lambda_{\text{wrong}}(1 - r_i) + \lambda_{\text{right}} r_i, \tag{5}$$

where $0 \leq \lambda_{\text{right}} \leq \lambda_{\text{wrong}} \leq 1$. By default, we set $\lambda_{\text{wrong}} = 1.0$ and $\lambda_{\text{right}} = 0.2$. Thus, incorrect trajectories keep full teacher intervention, whereas correct trajectories retain only weak teacher guidance.

This gate is intentionally trajectory-level. In the single-trajectory setting, there is no within-prompt comparison set for estimating group-relative advantages. PRISM therefore uses the most reliable supervision available at this stage: whether the complete reasoning trajectory reaches the correct answer.

4.2 Level 2: NLL-Based Error Prioritization

Outcome gating distinguishes correct from incorrect trajectories but still treats all incorrect rollouts equally. However, incorrect reasoning chains are not equally useful for learning. A low-ℓ_T incorrect trajectory remains relatively likely under the teacher and is more likely to be a coherent near-miss, whereas a high-ℓ_T trajectory is more strongly rejected by the teacher and may correspond to a chaotic failure.

Let $\mathcal{W} = \{i : r_i = 0\}$ denote the set of incorrect trajectories in a batch. When $\mathcal{W}$ is non-empty, PRISM assigns each $i \in \mathcal{W}$ an NLL-based priority

$$q_i(k) = |\mathcal{W}| \frac{\exp\left(-\ell_T(y_i)/\tau(k)\right)}{\sum_{j \in \mathcal{W}} \exp\left(-\ell_T(y_j)/\tau(k)\right)}. \tag{6}$$

If $\mathcal{W}$ is empty, this prioritization step is skipped. Otherwise, the factor $|\mathcal{W}|$ makes $q_i(k)$ a batch-normalized relative redistributor with unit mean over incorrect trajectories. Thus, it reallocates teacher correction within the current error set rather than changing the total correction strength; when $|\mathcal{W}| = 1$, it naturally reduces to $q_i(k) = 1$. Since $q_i(k)$ is not a probability and may exceed one, it is used only as a teacher-branch multiplier:

$$\alpha_i(k) = \begin{cases} 1, & r_i = 1, \\ q_i(k), & r_i = 0. \end{cases} \tag{7}$$

This design confines NLL-based modulation to the error path. Correct trajectories are controlled only by the bounded outcome gate g_i, preventing gradient destabilization from the non-linear coupling of trajectory rewards and token likelihoods.

4.3 Level 3: Curriculum Temperature Annealing

The priority $q_i(k)$ depends on the temperature $\tau(k)$. A large temperature makes priorities close to uniform, while a small temperature sharply emphasizes low-ℓ_T near-miss trajectories. This sharpness should not be fixed throughout training: early NLL differences are often noisy, whereas later differences better reflect meaningful error quality.

PRISM therefore uses a linear curriculum schedule

$$\tau(k) = \tau_{\text{start}} + (\tau_{\text{end}} - \tau_{\text{start}}) \frac{k}{K}, \tag{8}$$

where k is the current training step, K is the total number of training steps, $\tau_{\text{start}} = 5.0$, and $\tau_{\text{end}} = 0.5$. Early in training, the large temperature keeps teacher correction broad and conservative. As training progresses, the decayed temperature gradually sharpens prioritization, shifting correction toward coherent low-ℓ_T errors.

4.4 Scale Alignment Between Advantages

The two branches in Equation 4 have different granularities and magnitudes: A_i^{out} is a trajectory-level scalar, whereas $A_{i,t}^{\text{ex}}$ is a dense token-level advantage. Directly mixing them can cause one branch to dominate the update. We therefore align the scale of the ExOPD advantage before applying the gate and priority multiplier.

For each trajectory, let

$$\bar{A}_i^{\text{ex}} = \frac{1}{L_i} \sum_{t=1}^{L_i} A_{i,t}^{\text{ex}}. \tag{9}$$

Given a batch of size B, we estimate the response-level magnitude of the ExOPD branch as

$$m_{\text{ex}} = \frac{1}{B} \sum_{i=1}^{B} |\bar{A}_i^{\text{ex}}|. \tag{10}$$

For the outcome branch, we use a floored target magnitude

$$m_{\text{out}} = \max\left(\frac{1}{B}\sum_{i=1}^{B}|A_i^{\text{out}}|, \delta_{\text{out}}\right), \tag{11}$$

where $\delta_{\text{out}} > 0$ is a positive floor; for binary outcome-derived advantages, we use the natural unit floor $\delta_{\text{out}} = 1.0$. This prevents all-incorrect batches from setting the teacher advantage to zero. The scale-aligned ExOPD advantage is then

$$\widetilde{A}_{i,t}^{\text{ex}} = sA_{i,t}^{\text{ex}}, \quad s = \frac{m_{\text{out}}}{m_{\text{ex}} + \epsilon}, \tag{12}$$

where $\epsilon > 0$ is a small constant for numerical stability. This preserves the relative token-level structure of ExOPD while preventing the dense teacher advantage from overwhelming the outcome-derived branch.

4.5 Policy Optimization

PRISM uses $A_{i,t}^{\text{PRISM}}$ as the advantage in the same PPO-Clip objective used by the reproduced baselines. Let

$$\rho_{i,t}(\theta) = \frac{\pi_\theta(y_{i,t} \mid c_{i,t})}{\pi_{\theta_{\text{old}}}(y_{i,t} \mid c_{i,t})}, \tag{13}$$

Let $N = \sum_i L_i$ denote the total token count, $\hat{A}_{i,t} = A_{i,t}^{\text{PRISM}}$ the PRISM advantage, and $\bar{\rho}_{i,t} = \text{clip}(\rho_{i,t}(\theta), 1-\epsilon_c, 1+\epsilon_c)$ the clipped policy ratio with $\epsilon_c = 0.2$. The PRISM loss is

$$\begin{aligned} \mathcal{L}_{\text{PRISM}}(\theta) &= -\frac{1}{N}\sum_{i,t}\ell_{i,t}^{\text{clip}}, \\ \ell_{i,t}^{\text{clip}} &= \min\left(\rho_{i,t}(\theta)\hat{A}_{i,t}, \bar{\rho}_{i,t}\hat{A}_{i,t}\right). \end{aligned} \tag{14}$$

PRISM therefore changes neither the policy-gradient optimizer nor the ExOPD teacher advantage itself; it only modulates trajectory-level teacher trust before optimization.

4.6 Interpretation

PRISM yields different update regimes depending on trajectory outcome. For a correct trajectory ($r_i = 1$), we have $g_i = \lambda_{\text{right}}$ and $\alpha_i(k) = 1$, so the mixed advantage becomes

$$\begin{aligned} A_{i,t}^{\text{PRISM}} &= (1-\lambda_{\text{right}})A_i^{\text{out}} \\ &\quad + \lambda_{\text{right}}\widetilde{A}_{i,t}^{\text{ex}}, \end{aligned} \tag{15}$$

so learning is dominated by the outcome-derived advantage with only weak teacher correction. For an incorrect trajectory ($r_i = 0$), we have $g_i = \lambda_{\text{wrong}}$ and $\alpha_i(k) = q_i(k)$, while the outcome-derived advantage vanishes. The update becomes

$$A_{i,t}^{\text{PRISM}} = \lambda_{\text{wrong}} q_i(k)\widetilde{A}_{i,t}^{\text{ex}}. \tag{16}$$

Thus, incorrect trajectories are learned through teacher correction, but the amount of correction is redistributed toward coherent near-miss errors and sharpened progressively during training. Importantly, $q_i(k) > 1$ does not imply a negative coefficient on the outcome branch, because the interpolation coefficient is the bounded gate g_i, whereas $q_i(k)$ only scales the teacher branch on incorrect trajectories.

In summary, PRISM converts OPD from uniform teacher imitation into outcome-aware, quality-aware, and time-aware teacher-trust modulation. It preserves student autonomy when the student succeeds, prioritizes informative failures when the student fails, and delays sharp quality-based prioritization until the teacher NLL score becomes reliable.

5 Experiments

Following recent OPD studies, we evaluate PRISM under a unified protocol and organize the section as follows: experimental setup, main benchmark results, component ablations, curriculum-mechanism analysis, and stability/diversity diagnostics (Yang et al., 2026; Zhang et al., 2026; Ko et al., 2026; Jin et al., 2026). Capacity-boundary and additional diagnostic results are provided in the appendix.

5.1 Experimental Setup

Models and distillation regimes. We evaluate two distillation regimes. In the same-size setting, the teacher is a Qwen3-4B RL-trained math model and the student is a Qwen3-4B base student without math-specific RL post-training. In the strong-to-weak setting, the teacher is Qwen3-30B-A3B-Instruct and the student remains the same Qwen3-4B base student. These two settings test whether trajectory-level teacher-trust modulation is useful both when teacher and student have comparable capacity and when the teacher is substantially stronger.

Training configuration. All methods share the same student initialization, training data, and verl/vLLM training pipeline. We re-use the hyperparameters from verl/vLLM; detailed configurations are provided in Appendix A.

Method	AIME24	AIME25	MATH-500	AMC23	HMMT-Feb	HMMT-Nov	Avg.
Same-size distillation: 4B → 4B							
OPD	59.17	55.73	94.00	94.92	31.25	38.65	62.29
ExOPD	61.46	55.63	95.31	94.53	**34.79**	37.92	63.27
RLAD	58.02	53.65	94.53	94.69	29.06	36.35	61.05
REOPOLD	**62.71**	55.73	93.77	94.38	34.58	39.27	63.41
EOPD	62.50	57.08	94.09	95.00	34.69	38.33	63.62
PRISM	**62.71**	**58.67**	**95.45**	**95.23**	32.19	**39.30**	**63.93**
Strong-to-weak distillation: 30B → 4B							
OPD	57.08	49.79	93.01	91.41	30.31	38.85	60.08
ExOPD	57.29	50.52	93.90	92.89	28.44	38.85	60.32
RLAD	54.06	45.21	93.30	92.11	30.63	37.60	58.82
REOPOLD	56.15	50.42	**94.21**	**93.59**	30.21	38.75	60.56
EOPD	59.38	50.83	93.46	92.11	30.00	41.25	61.17
PRISM	**60.63**	**54.69**	93.42	91.09	**33.35**	**41.56**	**62.46**

Table 1: Main results on mathematical reasoning benchmarks. All benchmark numbers are mean@32 under the same evaluation protocol. Avg. denotes the arithmetic mean over the six benchmarks.

Evaluation. We evaluate on six mathematical reasoning benchmarks: AIME 2024, AIME 2025, MATH-500, AMC 2023, HMMT-Feb 2025, and HMMT-Nov 2025. We report mean@32 for all benchmarks using the same checkpoint-selection and evaluation protocol.

Baselines. We compare PRISM against five representative baselines: OPD (Gu et al., 2024; Agarwal et al., 2024), ExOPD (Yang et al., 2026), RLAD (Zhang et al., 2026), REOPOLD (Ko et al., 2026), and EOPD (Jin et al., 2026). OPD and ExOPD represent standard reverse-KL and reward-extrapolated on-policy distillation. RLAD represents reinforcement-aware teacher integration. REOPOLD and EOPD represent recent token-selective and entropy-aware OPD variants. Because prior papers use different teachers, students, data, and decoding temperatures, we reproduce the baselines in the same training framework whenever possible rather than directly borrowing incomparable numbers. As a sanity check, our reproduced ExOPD results are close to the originally reported numbers on overlapping benchmarks, with small deviations attributable to evaluation protocol and implementation differences. For baselines whose original papers use different model families, training data, or evaluation metrics, our numbers should be interpreted as controlled adaptations under our unified protocol rather than exact reproductions of the original settings.

5.2 Main Results

Table 1 reports the main results. PRISM obtains the strongest overall performance in both distillation regimes, with particularly clear gains on AIME 2024/2025 and HMMT-Nov. The gains are most pronounced in the strong-to-weak 30B-to-4B setting, where a substantially stronger teacher can over-constrain the student if applied uniformly.

In the same-size 4B → 4B setting, PRISM achieves the best average performance and leads on AIME25, MATH-500, AMC23, and HMMT-Nov, while tying REOPOLD on AIME24. In the strong-to-weak 30B → 4B setting, PRISM improves the average score by 1.29 points over the strongest non-PRISM baseline and leads on AIME24, AIME25, HMMT-Feb, and HMMT-Nov. PRISM is not uniformly best on saturated or easier benchmarks such as MATH-500 and AMC23, but this pattern is expected: algorithmic differences are most visible on challenging benchmarks where the student still has substantial room for improvement.

5.3 Ablation Study

We next isolate the contribution of each design choice in the 30B → 4B setting. Table 2 groups the ablations by the three levels of PRISM: outcome awareness, quality awareness, and temporal calibration.

Our ablation analysis in Table 2 reveals four critical insights. First, **outcome selectivity is indispensable**: setting $\lambda_{\text{right}} = 1.0$ keeps full teacher pressure on correct trajectories and reduces AIME25 from 54.69 to 51.88, while removing teacher correction entirely on correct traces ($\lambda_{\text{right}} = 0.0$) also degrades performance. This shows that correct trajectories should receive weakened, but not zero, teacher guidance. More importantly, uniform trust weakening ($\lambda = 0.2$) collapses learning to 31.25/26.98, confirming that selective

Ablation Group	Variant	AIME24	AIME25
Full Model	**PRISM (Ours)**	**60.63**	**54.69**
Outcome Gate	$\lambda_{\text{right}} = 1.0$	59.38	51.88
	$\lambda_{\text{right}} = 0.0$	57.81	51.25
	Symmetric $\lambda = 0.5$	58.30	51.50
	Uniform $\lambda = 0.2$	31.25	26.98
Wrong-path Strength	$\lambda_{\text{wrong}} = 0.5$	52.70	46.50
Token Gate	Log-probability gate	55.56	51.78
	Teacher-entropy gate	56.49	49.31
Temperature τ	Static NLL weighting	56.56	53.54
	Adaptive cap schedule	57.71	53.23

Table 2: Ablation results of PRISM variants in the 30B $\rightarrow$ 4B setting.

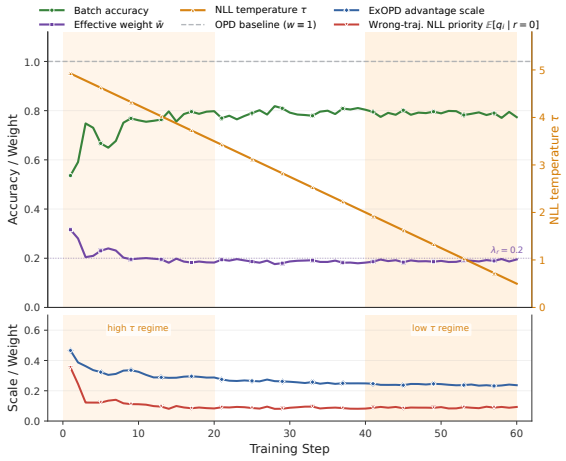

Figure 2: Adaptive pressure induced by PRISM. The outcome gate separates correct and incorrect trajectories, while curriculum temperature annealing gradually concentrates teacher correction on low-NLL incorrect trajectories as training progresses.

weakening is fundamentally different from globally reducing teacher pressure. Second, **incorrect trajectories require uncompromised correction**: reducing wrong-path strength to $\lambda_{\text{wrong}} = 0.5$ substantially hurts both AIME24 and AIME25, indicating that failed rollouts still need strong teacher intervention. Third, **token-level gates create a granularity mismatch**: both log-probability gating and teacher-entropy gating underperform PRISM, suggesting that local token uncertainty is too noisy to replace trajectory-level outcome information under single-trajectory settings ($n = 1$). Finally, **the timing of quality weighting matters**: static NLL weighting (56.56/53.54) and adaptive cap scheduling (57.71/53.23) remain below the full PRISM curriculum (60.63/54.69), supporting our claim that teacher-derived quality signals should be sharpened gradually rather than trusted uniformly throughout training.

Second, **incorrect trajectories require uncompromised correction**: reducing the wrong-path strength to $\lambda_{\text{wrong}} = 0.5$ substantially hurts both AIME24 and AIME25, indicating that failed rollouts still necessitate robust teacher intervention. Third, **token-level gates introduce a granularity mismatch**: both the log-probability gating and teacher-entropy gating variants underperform PRISM. This gap suggests that local token uncertainty is too noisy to substitute for trajectory-level outcome information in single-trajectory ($n = 1$) settings. Finally, **the timing of quality weighting is crucial**: static NLL weighting (56.56/53.54) and adaptive cap scheduling (57.71/53.23) both fall short of the full PRISM curriculum (60.63/54.69), strongly supporting our hypothesis that teacher-derived quality signals should be sharpened gradually rather than trusted uniformly throughout the entire training lifecycle.

5.4 When Does Teacher NLL Become Reliable?

PRISM uses teacher NLL only through a curriculum schedule. This design is motivated by the hypothesis that teacher-derived trajectory-quality signals are unreliable early in training but become useful once student errors become more differentiated. Table 2 supports this hypothesis: static NLL weighting is worse than PRISM, and more complex adaptive schedules do not outperform linear annealing.

Figure 2 visualizes the mechanism. Early in training, large $\tau(k)$ keeps the priority distribution close to uniform, preventing noisy NLL differences from dominating the update. Later in training, smaller $\tau(k)$ sharpens the priority distribution and focuses correction on coherent near-miss errors. This supports PRISM's central principle: the usefulness of a quality signal depends not only on what the signal measures, but also on when it is trusted.

5.5 Training Stability and Solution Diversity

Final accuracy alone does not fully capture the behavior of on-policy distillation. We therefore provide training-stability and solution-diversity diagnostics in Appendix B. These analyses show that PRISM mitigates the variance of policy exploration and maintains competitive reasoning diversity without collapsing to a narrow reasoning template.

Additional diagnostics. We further report response-length distributions in the appendix.

Across five methods, response lengths remain within a stable range, indicating that PRISM's gains are not caused by uncontrolled length expansion.

6 Discussion

Trajectory-level trust as a missing dimension in OPD. Our results suggest that teacher guidance in on-policy distillation should not be treated as uniformly beneficial. A correct rollout, a coherent near-miss, and a severe failure expose fundamentally different learning needs. Uniform teacher pressure can over-constrain correct trajectories, while uniform weakening removes correction from failures that need it most. PRISM addresses this mismatch by making trajectory-level trust allocation an explicit decision before applying token-level teacher advantages. This reframes OPD from merely designing stronger teacher signals to deciding when, and for which complete rollouts, those signals should be trusted.

How PRISM complements token-level selectivity. Prior selective OPD methods mainly operate at the token level, for example by filtering high-entropy or low-confidence positions. Such mechanisms can be useful, but they do not decide whether the complete rollout should be corrected, preserved, or prioritized. PRISM makes this coarser decision first: outcome correctness determines the base level of teacher intervention, and teacher NLL further prioritizes incorrect trajectories. Token-level advantages are then applied under this trajectory-level trust decision. This separation helps explain why token-level gates alone are weaker in our ablations: local uncertainty is informative only after the learning role of the whole trajectory has been identified.

Why single-trajectory OPD needs coarse but reliable signals. In the single-trajectory setting, the learner cannot rely on group-relative comparisons across multiple sampled responses. This makes fine-grained token-level gates highly unstable: token probabilities or entropy may reflect local semantic uncertainty rather than the learning value of the complete reasoning path. Our ablations show that log-probability and teacher-entropy gates underperform PRISM, suggesting that outcome correctness serves as a more reliable first-order signal for deciding whether to preserve student autonomy or apply teacher correction.

Where the gains appear. PRISM does not improve every benchmark uniformly. Its advantage is most visible on harder benchmarks such as AIME and HMMT, where the student has room to benefit from selective correction. On more saturated benchmarks such as MATH-500 and AMC23, the strongest baselines are already close to ceiling performance, leaving less space for trajectory-level trust modulation to help. This pattern is consistent with the role of PRISM: it is most useful when failures are frequent enough to require teacher correction, but structured enough for teacher NLL to distinguish near-misses from less useful errors.

Timing matters for teacher-derived quality signals. Teacher NLL is useful only when student errors become sufficiently differentiated. Applying NLL weighting too early makes the update sensitive to noisy likelihood differences, whereas curriculum temperature annealing delays sharp prioritization until the signal becomes more meaningful. This explains why static NLL weighting and adaptive schedules underperform the full PRISM schedule.

7 Conclusion

We presented PRISM, a trajectory-level teacher-trust modulation framework for single-trajectory on-policy distillation. Rather than treating teacher supervision as uniformly beneficial, PRISM separates correct and incorrect rollouts, prioritizes coherent near-miss errors using teacher NLL, and schedules the sharpness of this quality signal over training. Across mathematical reasoning benchmarks and controlled ablations, PRISM shows that selectively reducing teacher intervention on correct trajectories while preserving strong correction on errors improves average performance over strong OPD baselines. The broader implication is that strong-to-weak distillation is not only a question of constructing a stronger teacher signal. It also requires deciding when that signal should be trusted and when the student should retain autonomy over its own valid reasoning paths. This perspective helps explain why uniform teacher pressure, uniform weakening, and purely token-level gates are insufficient under single-trajectory OPD. We hope PRISM encourages future work to study teacher trust as an explicit part of on-policy distillation, especially in settings where dense teacher guidance must be balanced against student exploration and capacity constraints.

Limitations

This work has several limitations. First, our experiments focus on mathematical reasoning benchmarks, where outcome verification is relatively reliable. Whether the same trajectory-level teacher-trust principle transfers to open-ended instruction following, coding, or multilingual generation requires further study.

Second, our experiments are primarily based on Qwen-series teachers and students. Although this controlled setting isolates the effect of teacher-trust modulation under a consistent open-weight model family, broader validation across model families would strengthen the generality of the conclusions.

Third, PRISM is designed for single-trajectory on-policy distillation. Extending the same principle to multi-sample RL settings, where group-relative advantages are available, remains an important direction.

Finally, PRISM may face a capacity boundary in more extreme compression settings. The method is most effective when the student has enough capacity to benefit from selective teacher-trust modulation; when the student is too small to absorb dense teacher correction, adaptive trust allocation alone may be insufficient. We provide supporting evidence in Appendix C.

References

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, and 1 others. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Rishabh Agarwal, Nino Vieillard, Yongchao Zhou, Piotr Stanczyk, Sabela Ramos Garea, Matthieu Geist, and Olivier Bachem. 2024. On-policy distillation of language models: Learning from self-generated mistakes. In *International Conference on Learning Representations*, volume 2024, pages 21246–21263.

Samy Bengio, Oriol Vinyals, Navdeep Jaitly, and Noam Shazeer. 2015. Scheduled sampling for sequence prediction with recurrent neural networks. *Advances in neural information processing systems*, 28.

Yoshua Bengio, Jérôme Louradour, Ronan Collobert, and Jason Weston. 2009. Curriculum learning. In *Proceedings of the 26th annual international conference on machine learning*, pages 41–48.

Cristian Buciluǎ, Rich Caruana, and Alexandru Niculescu-Mizil. 2006. Model compression. In *Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 535–541.

Daixuan Cheng, Shaohan Huang, Xuekai Zhu, Bo Dai, Xin Zhao, Zhenliang Zhang, and Furu Wei. 2026. Reasoning with exploration: An entropy perspective. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pages 30377–30385.

Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, and 1 others. 2021. Training verifiers to solve math word problems. *arXiv preprint arXiv:2110.14168*.

Jianping Gou, Baosheng Yu, Stephen J Maybank, and Dacheng Tao. 2021. Knowledge distillation: A survey. *International journal of computer vision*, 129(6):1789–1819.

Yuxian Gu, Li Dong, Furu Wei, and Minlie Huang. 2024. Minillm: Knowledge distillation of large language models. In *International Conference on Learning Representations*, volume 2024, pages 32694–32717.

Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Peiyi Wang, Qihao Zhu, Runxin Xu, Ruoyu Zhang, Shirong Ma, Xiao Bi, and 1 others. 2025. Deepseek-r1 incentivizes reasoning in llms through reinforcement learning. *Nature*, 645(8081):633–638.

Chaoqun He, Renjie Luo, Yuzhuo Bai, Shengding Hu, Zhen Thai, Junhao Shen, Jinyi Hu, Xu Han, Yujie Huang, Yuxiang Zhang, and 1 others. 2024. Olympiadbench: A challenging benchmark for promoting agi with olympiad-level bilingual multimodal scientific problems. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3828–3850.

Dan Hendrycks, Collin Burns, Saurav Kadavath, Akul Arora, Steven Basart, Eric Tang, Dawn Song, and Jacob Steinhardt. 2021. Measuring mathematical problem solving with the math dataset. *arXiv preprint arXiv:2103.03874*.

Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. 2015. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*.

Cheng-Yu Hsieh, Chun-Liang Li, Chih-Kuan Yeh, Hootan Nakhost, Yasuhisa Fujii, Alex Ratner, Ranjay Krishna, Chen-Yu Lee, and Tomas Pfister. 2023. Distilling step-by-step! outperforming larger language models with less training data and smaller model sizes. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 8003–8017.

Ijun Jang, Jewon Yeom, Juan Yeo, Hyunggu Lim, and Taesup Kim. 2026. Stable on-policy distillation through adaptive target reformulation. *arXiv preprint arXiv:2601.07155*.

Nan Jia, Haojin Yang, Xing Ma, Jiesong Lian, Shuailiang Zhang, Weipeng Zhang, Ke Zeng, Xunliang Cai, and Zequn Sun. 2026. Asymmetric on-policy distillation: Bridging exploitation and imitation at the token level. *arXiv preprint arXiv:2605.06387*.

Xiaoqi Jiao, Yichun Yin, Lifeng Shang, Xin Jiang, Xiao Chen, Linlin Li, Fang Wang, and Qun Liu. 2020. Tinybert: Distilling bert for natural language understanding. In *Findings of the association for computational linguistics: EMNLP 2020*, pages 4163–4174.

Woogyeol Jin, Taywon Min, Yongjin Yang, Swanand Ravindra Kadhe, Yi Zhou, Dennis Wei, Nathalie Baracaldo, and Kimin Lee. 2026. Entropy-aware on-policy distillation of language models. *arXiv preprint arXiv:2603.07079*.

Yoon Kim and Alexander M Rush. 2016. Sequence-level knowledge distillation. In *Proceedings of the 2016 conference on empirical methods in natural language processing*, pages 1317–1327.

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

Jongwoo Ko, Sungnyun Kim, Tianyi Chen, and Se-Young Yun. 2024. Distillm: Towards streamlined distillation for large language models. *arXiv preprint arXiv:2402.03898*.

Hunter Lightman, Vineet Kosaraju, Yuri Burda, Harrison Edwards, Bowen Baker, Teddy Lee, Jan Leike, John Schulman, Ilya Sutskever, and Karl Cobbe. 2024. Let's verify step by step. In *International Conference on Learning Representations*, volume 2024, pages 39578–39601.

David Rein, Betty Li Hou, Asa Cooper Stickland, Jackson Petty, Richard Yuanzhe Pang, Julien Dirani, Julian Michael, and Samuel R Bowman. 2023. Gpqa: A graduate-level google-proof q&a benchmark. *arXiv preprint arXiv:2311.12022*.

Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. 2019. Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter. *arXiv preprint arXiv:1910.01108*.

Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, YK Li, Yang Wu, and 1 others. 2024. Deepseekmath: Pushing the limits of mathematical reasoning in open language models. *arXiv preprint arXiv:2402.03300*.

Guangming Sheng, Chi Zhang, Zilingfeng Ye, Xibin Wu, Wang Zhang, Ru Zhang, Yanghua Peng, Haibin Lin, and Chuan Wu. 2025. Hybridflow: A flexible and efficient rlhf framework. In *Proceedings of the Twentieth European Conference on Computer Systems*, pages 1279–1297.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, and 1 others. 2023. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.

Yuqiao Wen, Zichao Li, Wenyu Du, and Lili Mou. 2023. F-divergence minimization for sequence-level knowledge distillation. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 10817–10834.

Hongling Xu, Qi Zhu, Heyuan Deng, Jinpeng Li, Lu Hou, Yasheng Wang, Lifeng Shang, Ruifeng Xu, and Fei Mi. 2025a. Kdrl: Post-training reasoning llms via unified knowledge distillation and reinforcement learning. *arXiv preprint arXiv:2506.02208*.

Wenda Xu, Rujun Han, Zifeng Wang, Long Le, Dhruv Madeka, Lei Li, William Wang, Rishabh Agarwal, Chen-Yu Lee, and Tomas Pfister. 2025b. Speculative knowledge distillation: Bridging the teacher-student gap through interleaved sampling. In *International Conference on Learning Representations*, volume 2025, pages 64616–64646.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, and 1 others. 2025. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*.

An Yang, Baosong Yang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Zhou, Chengpeng Li, Chengyuan Li, Dayiheng Liu, Fei Huang, and 1 others. 2024. Qwen2 technical report. *Preprint*, arXiv:2407.10671.

Wenkai Yang, Weijie Liu, Ruobing Xie, Kai Yang, Saiyong Yang, and Yankai Lin. 2026. Learning beyond teacher: Generalized on-policy distillation with reward extrapolation. *arXiv preprint arXiv:2602.12125*.

Shunyu Yao, Dian Yu, Jeffrey Zhao, Izhak Shafran, Tom Griffiths, Yuan Cao, and Karthik Narasimhan. 2023. Tree of thoughts: Deliberate problem solving with large language models. *Advances in neural information processing systems*, 36:11809–11822.

Tianzhu Ye, Li Dong, Xun Wu, Shaohan Huang, and Furu Wei. 2026. On-policy context distillation for language models. *arXiv preprint arXiv:2602.12275*.

Qiying Yu, Zheng Zhang, Ruofei Zhu, Yufeng Yuan, Xiaochen Zuo, Yu Yue, Weinan Dai, Tiantian Fan, Gaohong Liu, Lingjun Liu, and 1 others. 2026. Dapo: An open-source llm reinforcement learning system at scale. *Advances in Neural Information Processing Systems*, 38:113222–113244.

Zhaoyang Zhang, Shuli Jiang, Yantao Shen, Yuting Zhang, Dhananjay Ram, Shuo Yang, Zhuowen Tu,

Wei Xia, and Stefano Soatto. 2026. Reinforcement-aware knowledge distillation for llm reasoning. *arXiv preprint arXiv:2602.22495*.

Siyan Zhao, Zhihui Xie, Mengchen Liu, Jing Huang, Guan Pang, Feiyu Chen, and Aditya Grover. 2026. Self-distilled reasoner: On-policy self-distillation for large language models. *arXiv preprint arXiv:2601.18734*.

A Detailed Training Configurations

All experiments are implemented with the same verl/vLLM-based training pipeline. We train on DeepMath-103K with batch size 1024, rollout number $n = 1$, learning rate 1×10^{-5}, maximum response length 16K, sampling temperature 1.0, and PPO clipping threshold 0.2. For PRISM, we use $\lambda_{\text{wrong}} = 1.0$, $\lambda_{\text{right}} = 0.2$, $\tau_{\text{start}} = 5.0$, and $\tau_{\text{end}} = 0.5$. All training jobs are run on a single 8×NVIDIA H20 server with 96GB memory per GPU. A complete training run takes approximately 25 wall-clock hours, corresponding to about 200 H20 GPU-hours. Hardware allocation, Ray cluster settings, optimizer details, and vLLM generation settings follow the same configuration across all compared methods. Because our controlled reproductions use DeepMath-103K and a unified six-benchmark evaluation protocol, they are not exact reruns of the original ExOPD training-data setting; small differences from the numbers reported in the ExOPD paper are therefore expected.

For RLAD, we follow the TRRD formulation with mixing coefficient $\alpha = 0.5$, which is supported by the original RLAD ablation and balances the policy-ratio term and the teacher-ratio term. Under our unified training setup, we adopt conservative asymmetric clipping bounds $[\epsilon_{\text{low}}, \epsilon_{\text{high}}] = [0.20, 0.28]$, corresponding to the clipping interval $[0.80, 1.28]$. The base teacher advantage uses $\lambda = 1.0$, i.e., the standard reverse-KL teacher signal. All other training hyperparameters, including learning rate, batch size, rollout number, and response length, follow the unified setup above.

REOPOLD. For REOPOLD, we use the two-phase selective distillation procedure. In the first phase, we apply reward-based masking with $\lambda = 0.2$, a conservative threshold that filters tokens whose advantage falls below $\log(\lambda/(1-\lambda)) = -1.39$. In the second phase, we apply entropy-based token selection with percentile coefficient $\beta = 0.2$, retaining the top-20% high-entropy tokens, consistent with the main REOPOLD hyperparameter setting. The phase switch is set to $T^* = 15$ to perform an early transition from exploration to refinement under our shorter training budget. Other optimization hyperparameters are kept identical to the unified setup above.

B Additional Ablations and Diagnostics

We provide additional analyses to visualize the effects of outcome-aware trust modulation and curriculum temperature scheduling. These diagnostics complement the main results by isolating component-level effects, identifying benchmark-wise gain patterns, and examining whether PRISM improves consistency while preserving diverse reasoning behaviors.

C Capacity Boundary Results

We examine a stronger compression setting, 30B $\rightarrow$ 1.7B, to clarify a boundary condition of PRISM.

Method	AIME24	AIME25
ExOPD	**37.60**	**32.10**
PRISM	36.56	29.17

Table 3: Appendix result for 30B $\rightarrow$ 1.7B distillation. PRISM does not improve in this lower-capacity setting, suggesting that the student must have sufficient capacity to benefit from adaptive teacher-trust modulation.

PRISM underperforms ExOPD in the 30B $\rightarrow$ 1.7B setting. We interpret this as a capacity boundary: when the student is too small to absorb the teacher's dense corrective signal, adaptive trust allocation alone is insufficient. This result delineates where PRISM is expected to help: difficult but still learnable strong-to-weak distillation, such as 30B $\rightarrow$ 4B.

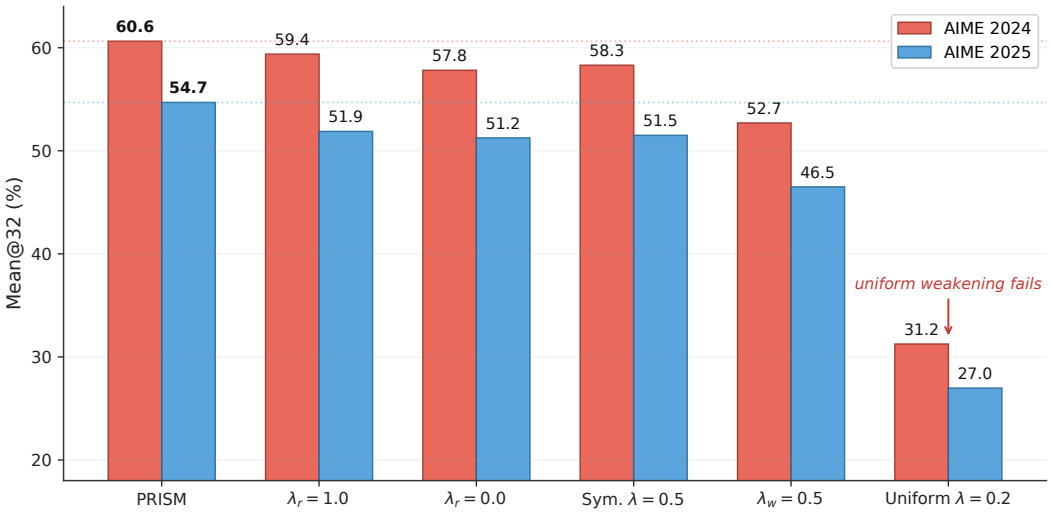

Figure 3: Outcome-aware trust ablation. We visualize key variants from Table 2. Uniformly weakening teacher trust causes severe degradation, while reducing teacher correction on incorrect trajectories also hurts performance. PRISM performs best by weakening teacher intervention only on correct rollouts while preserving strong correction on failed rollouts. Dotted lines indicate PRISM performance.

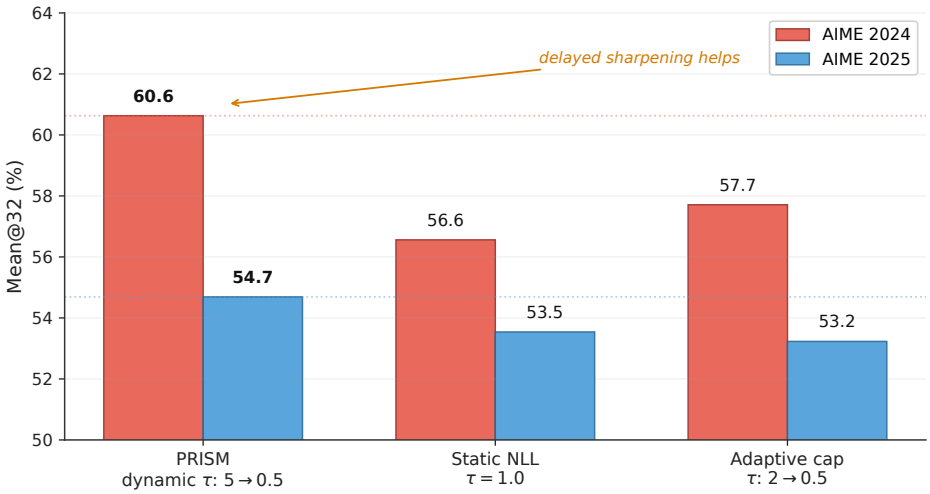

Figure 4: Temperature-schedule ablation. We compare PRISM's curriculum annealing against static NLL weighting and an alternative adaptive-cap schedule. Dynamic annealing from high to low temperature yields the strongest performance, supporting the need to delay sharp NLL-based prioritization until teacher-derived quality signals become reliable. Dotted lines indicate PRISM performance.

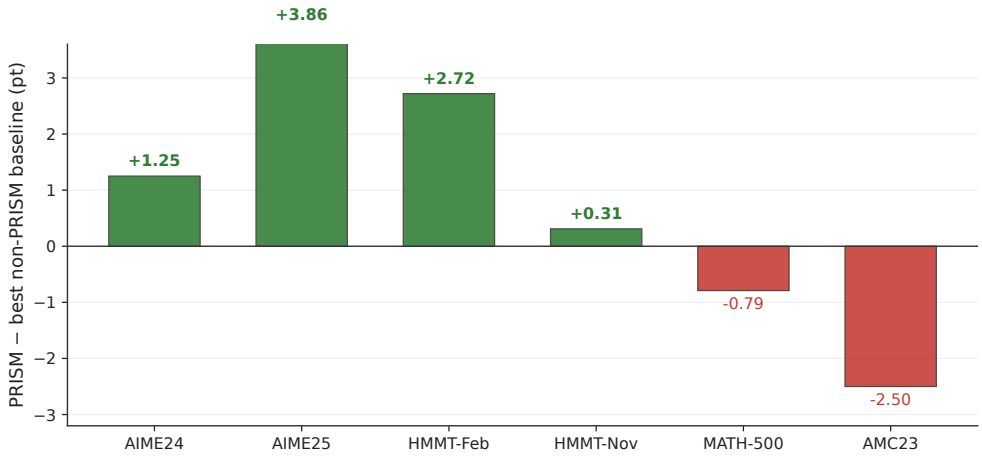

Figure 5: Benchmark-wise gains over the strongest non-PRISM baseline in the 30B $\rightarrow$ 4B setting. We report PRISM minus the best competing baseline under the same evaluation protocol. PRISM's gains concentrate on challenging benchmarks such as AIME and HMMT, while saturated benchmarks such as MATH-500 and AMC23 leave less room for improvement.

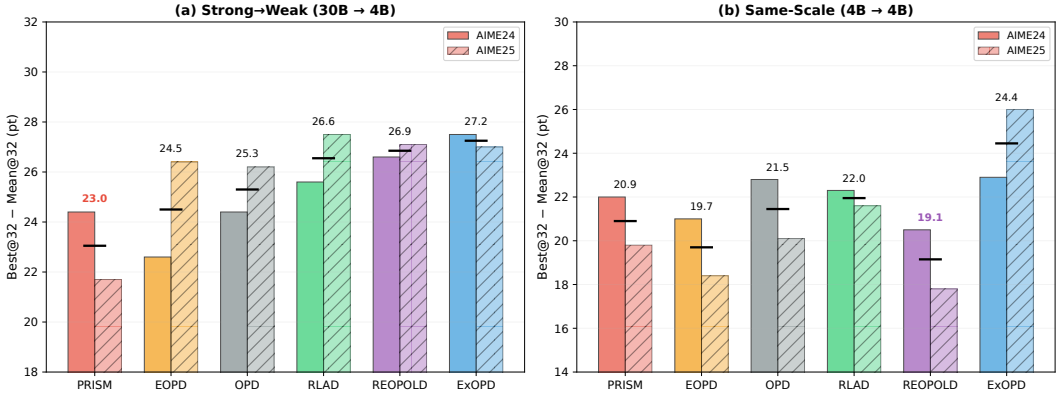

Figure 6: Best–mean consistency gap across distillation settings. We report the gap between Best@32 and Mean@32, where a smaller gap indicates more consistent sampled performance rather than reliance on rare high-scoring generations. In the strong-to-weak 30B $\rightarrow$ 4B setting, PRISM achieves the smallest gap, suggesting more stable reasoning behavior across samples.

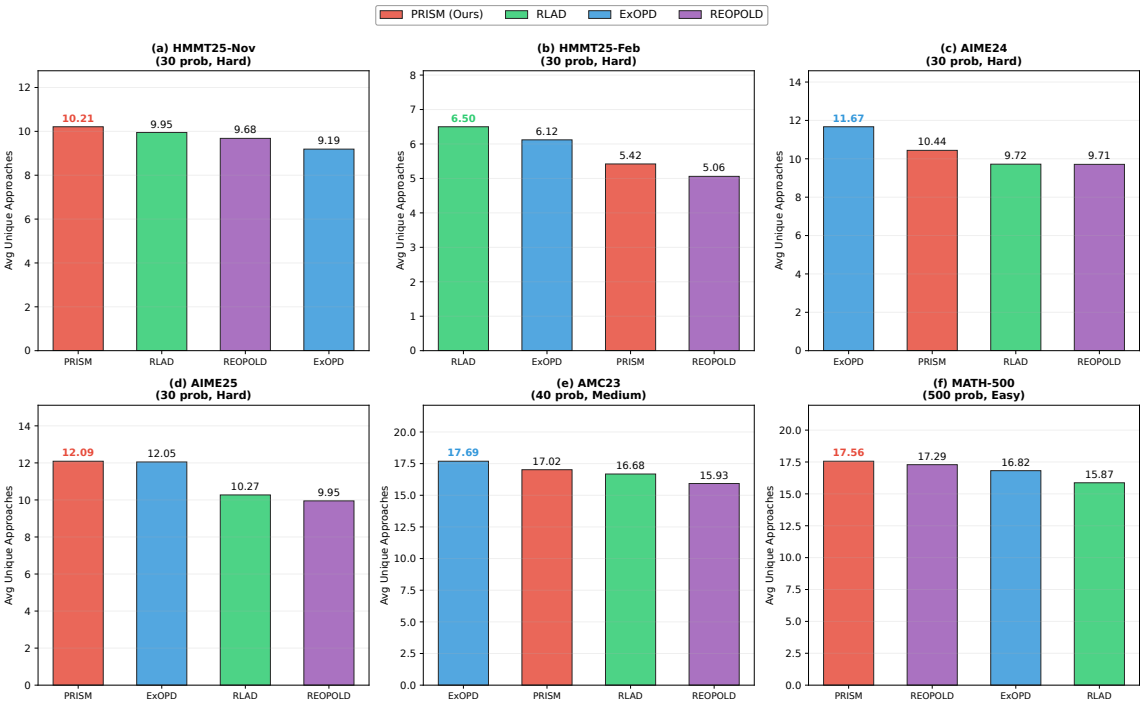

Figure 7: Solution diversity across six mathematical reasoning benchmarks. We report the average number of unique reasoning approaches per problem under $n = 32$ sampling. PRISM maintains competitive diversity across benchmarks and achieves the highest diversity on several challenging settings.