# THE GRAVITY OF PRIVILEGE: TILTING LEARNING MEASURES IN ON-POLICY SELF-DISTILLATION

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

## ABSTRACT

On-Policy Self-Distillation (OPSD) is an emerging paradigm for LLM post-training, in which a model learns from a privileged view of itself through dense supervision on student-generated trajectories. However, existing OPSD methods incorporate privileged information through the conditional teacher distribution, while leaving the empirical measure over student-generated trajectories unchanged. This induces a structural asymmetry: privilege determines what the student learns at each visited trajectory, but not the measure under which different trajectories contribute to learning, even though privileged conditioning can vary substantially across sampled trajectories. To address this limitation, we propose Privileged Measure Self-Distillation (PMSD), which leverages privilege not only to construct distillation targets, but also to adaptively reweight the learning measure across training problems. Specifically, PMSD contrasts the same frozen teacher's support for each trajectory with and without privileged information, aggregating log-likelihood ratios into a policy-dependent signal for problem-level allocation. We formulate this adaptation as a relative-entropy trust-region optimization that assigns greater learning mass to problems receiving stronger support from privilege, while bounding the KL divergence from the original empirical measure. This mechanism intrinsically favors problems with larger estimated privileged likelihood contrasts while preventing extreme polarization in the learning measure. The resulting Gibbs-form solution is governed by a scalar dual variable, with the KL constraint directly bounding the induced shift in population gradient aggregation for the exact solution under bounded per-problem gradients. To obtain a more reliable estimate of the privilege-induced effect under the current policy, PMSD additionally uses multiple on-policy rollouts per problem. Extensive experiments on mathematical reasoning across Qwen3-1.7B/4B/8B show that PMSD consistently surpasses selected baselines by 0.84–2.50 percentage points in Avg@12 while maintaining superior token efficiency. Our code is available on `https://anonymous.4open.science/r/pmsd-paper-code-B260`.

## 1 INTRODUCTION

Large language models (LLMs) have made substantial progress in complex reasoning, while effective post-training remains central to further improving their capabilities. Reinforcement-learning-based methods improve reasoning through on-policy exploration and outcome feedback, but typically rely on sparse sequence-level signals and substantial rollout budgets (Schulman et al., 2017; Shao et al., 2024; Guo et al., 2025). Knowledge distillation offers a complementary approach by transferring dense predictive distributions from a teacher to a student (Hinton et al., 2015; Kim & Rush, 2016; Furlanello et al., 2018; Gu et al., 2024). For autoregressive reasoning, however, conventional distillation is commonly performed on fixed or teacher-generated sequences, creating a mismatch between the training distribution and the states actually visited by the student. On-policy distillation addresses this mismatch by evaluating teacher supervision directly along student-generated trajectories, coupling token-level feedback with the evolving student policy (Agarwal et al., 2024).

As illustrated in Figure 1, On-Policy Self-Distillation (OPSD) further removes the need for a separate stronger teacher. A single model serves as both student and teacher under different information conditions: the student generates trajectories from its current policy, while a frozen teacher view is additionally conditioned on training-only privileged information, such as a refer-



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

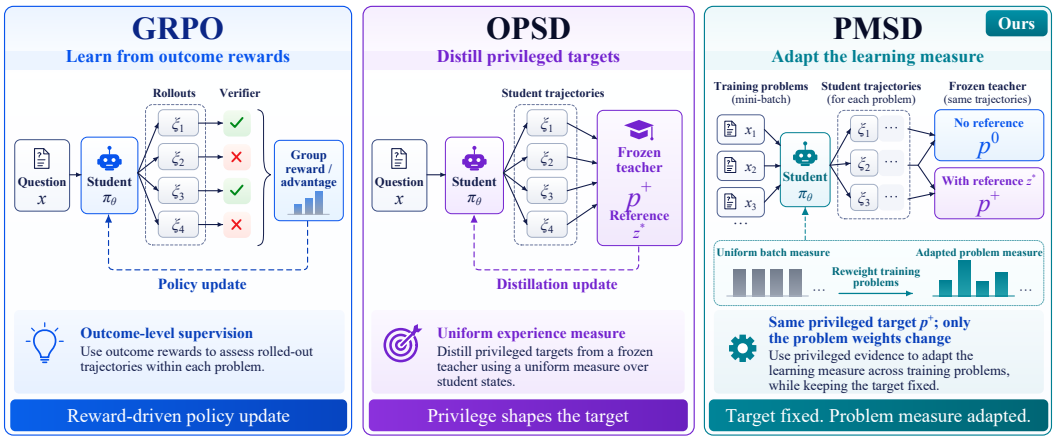



Figure 1: **From privileged targets to privileged learning measures.** GRPO learns from outcome feedback; OPSD introduces privileged targets on student-generated experience but leaves its empirical learning measure unchanged. PMSD adapts the measure across training problems.

ence solution, and provides dense supervision along those trajectories (Zhao et al., 2026b). This construction combines on-policy state visitation with privileged token-level supervision and has been shown to achieve competitive reasoning performance at substantially lower generation cost than reinforcement-learning-based post-training. Subsequent work has improved OPD and OPSD through selective supervision, uncertainty modeling, teacher calibration, trajectory filtering, prefix-level allocation, and adaptive budgeting (Yang et al., 2026a; Zheng et al., 2026; Hou et al., 2026; Zhu et al., 2026; Ke et al., 2026; Li et al., 2026; Zhao et al., 2026a; Zhou et al., 2026). These demonstrate the value of adaptive supervision. Despite these advances, privileged information is still used primarily to determine the supervision applied to sampled experience, rather than the empirical measure under which that experience enters optimization.

This distinction reveals a second degree of freedom in OPSD. For a set of student-generated experience, OPSD specifies a privileged conditional target at each visited state, while the empirical contribution of different training problems remains determined by the original sampling measure. Privileged information therefore changes what is distilled, but not how learning mass is distributed over the sampled experience. This separation becomes consequential when privileged conditioning affects different problems unequally: revealing the reference may substantially increase the teacher's support for the student-generated trajectory on some problems, while inducing little or even opposite change on others. Uniformly preserving their original contributions ignores this variation exposed by the privileged teacher. For instance, if a challenging problem receives poor supervision under standard conditions but gains exceptionally high-quality guidance when the reference is revealed, it logically warrants a larger learning weight. *This raises a natural question: can privileged information be used not only to construct the distillation target, but also to adapt the empirical measure under which on-policy experience contributes to learning?*

To answer this question, we introduce Privileged Measure Self-Distillation (PMSD) that extends privileged self-distillation from target construction to learning-measure adaptation, as shown in Figure 1. For each student-generated trajectory, PMSD evaluates the same frozen teacher under paired privileged and non-privileged conditions and computes the log-likelihood ratio of the tokens actually generated by the student. This contrast isolates how revealing privileged information changes the teacher's support for the observed trajectory. We aggregate these trajectory-level contrasts into a problem-level score and use it to reweight the empirical measure across on-policy training problems, while retaining the original privileged teacher distribution as the distillation target. We formulate this measure adaptation as a relative-entropy trust-region optimization, maximizing privilege-induced preference subject to controlled deviation from the original empirical measure. The resulting variational problem admits a Gibbs-form solution with a scalar dual variable; moreover, the KL constraint directly bounds the induced displacement of gradient aggregation when per-example gradients are bounded. To make the problem-level estimate less dependent on a single sampled trajectory, PMSD additionally uses multiple on-policy rollouts per problem. In summary, our main contributions are:

- **Methodology:** We propose PMSD, an algorithm that extends privileged self-distillation from target construction to adaptive, problem-level learning measure reweighting.
- **Theoretical Grounding:** We formulate this measure adaptation as a relative-entropy trust-region optimization, mathematically demonstrating that the KL constraint strictly bounds the induced displacement of gradient aggregation.
- **Empirical Performance:** Extensive evaluations on mathematical reasoning across the Qwen3 family (1.7B to 8B) demonstrate that PMSD consistently surpasses strong baselines like OPSD by 0.84-2.50% while maintaining superior token efficiency.

## 2 Related Works

### 2.1 Knowledge Transfer

Knowledge distillation transfers predictive information through teacher distributions (Hinton et al., 2015), sequence-level targets (Kim & Rush, 2016), or successive models of the same capacity (Furlanello et al., 2018). For language generation, the conditioning history and divergence determine what is transferred. MiniLLM studies reverse-KL distillation (Gu et al., 2024), while GKD trains on student-generated sequences (Agarwal et al., 2024). The importance of learner-induced states has earlier roots in imitation learning and sequence prediction, including DAgger and scheduled sampling (Ross et al., 2011; Bengio et al., 2015). PMSD uses this on-policy support and studies the measure used to aggregate its supervision. Training-only references fit the privileged-information setting of Vapnik & Vashist (2009) and its connection to distillation developed by Lopez-Paz et al. (2015).

### 2.2 Privileged Feedback

OPSD conditions a self-teacher on reference solutions and distills its predictions at student prefixes (Zhao et al., 2026b). Subsequent methods refine the quality, granularity, and exposure of this feedback. TIP combines uncertainty and teacher–student discrepancy to estimate token importance (Xu et al., 2026). SCOPE uses correctness-dependent routing and calibrated weighting (Zheng et al., 2026). EGRSD combines outcome-grounded direction with likelihood-based and uncertainty-sensitive signals (Ke et al., 2026). FiRe-OPD filters trajectories before reweighting tokens (Li et al., 2026). USD jointly adapts supervision weights and privileged exposure according to student learning capacity (Yang et al., 2026b). These methods establish that adaptive allocation is already an active part of on-policy distillation. Other work changes the teacher signal itself. AR-OPD constructs anchored residual guidance (Zhang, 2026), and RLCSD contrasts correct and incorrect hints to isolate aspects of privileged feedback (Pan et al., 2026). Purified OPSD uses a reference-only teacher to construct a residual-based PMI target anchored to a question-only base distribution (Shen et al., 2026). Analyses also expose prefix dependence (Zhu et al., 2026), performance degradation in studied thinking-model settings (Kaur et al., 2026), and context-induced changes that need not reflect reference-specific information (Ichihara et al., 2026). The distinction between privileged likelihood and outcome value is especially relevant to our evidence definition (Nguyen et al., 2026).

### 2.3 Measure Adaptation

Relative-entropy regularization yields exponential changes of measure and tractable dual optimization, as in Relative Entropy Policy Search (Peters et al., 2010). Policy trust-region methods instead constrain or approximate changes in action distributions (Schulman et al., 2015; 2017). PMSD applies a KL constraint to the finite empirical measure over attempts from prompts. This scope matters: its constraint directly bounds redistribution of distillation mass, rather than policy divergence after an optimizer step. The derivation uses standard convex duality (Vandenberghe & Boyd, 2004). Our evidence also differs from a symmetric divergence between teacher views. Jensen–Shannon divergence (Lin, 1991) quantifies distributional disagreement at a prefix without specifying whether the realized student token gains or loses support. The signed log-likelihood ratio retains this directional information. In contrast to verifier or process-supervision approaches (Cobbe et al., 2021; Lightman et al., 2024), it requires no correctness label for a generated attempt. This removes a source of supervision, but also limits interpretation: a bounded measure change does not establish that the favored trajectories contain better reasoning.

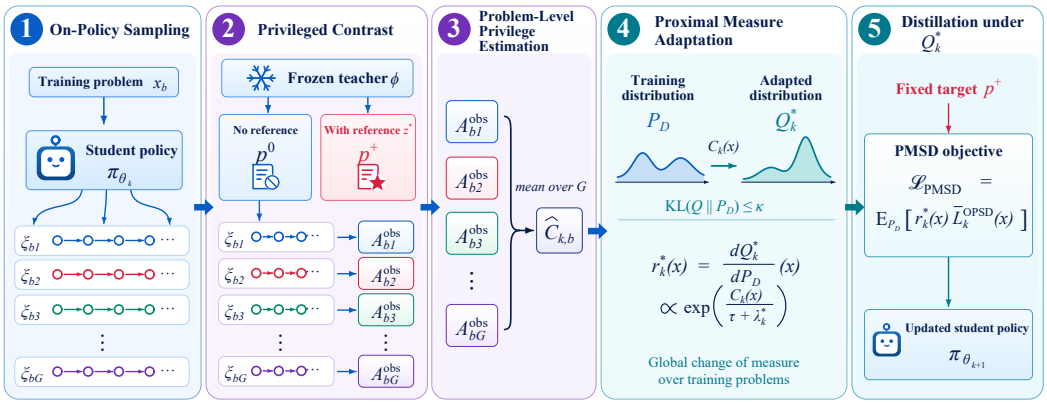



Figure 2: **The overview of PMSD.** Trajectory-level privilege evidence is aggregated into problem-level scores to adapt the on-policy learning measure over training problems.

## 3 METHOD

### 3.1 PROBLEM SETUP AND OVERVIEW

Let $P_D$ denote the training distribution over problems, where each problem $x$ is paired with training-only privileged information $z^\star$. At optimization step $k$, the current student policy $\pi_{\theta_k}$ induces an on-policy trajectory distribution $\mu_{\theta_k}(\xi \mid x)$, where $\xi = (y_1, \ldots, y_T)$ is generated without access to $z^\star$. Given a realized prefix $H_t = (x, y_{<t})$, OPSD evaluates a frozen teacher under the privileged condition $p_t^+ = p_\phi(\cdot \mid H_t, z^\star)$, and distills this distribution into the student prediction $p_{\theta,t}^S = p_\theta^S(\cdot \mid H_t)$. For a sampled trajectory $\xi$, the trajectory-normalized OPSD loss is

$$L^{\mathrm{OPSD}}(x, \xi; \theta) = \frac{1}{T} \sum_{t=1}^{T} \ell_{\mathrm{OPSD}} \left(p_t^+, p_{\theta,t}^S\right). \tag{1}$$

The corresponding population objective is

$$\mathcal{L}_{\mathrm{OPSD}}(\theta) = \mathbb{E}_{x \sim P_D} \mathbb{E}_{\xi \sim \mu_{\theta_k}(\cdot|x)} \left[L^{\mathrm{OPSD}}(x, \xi; \theta)\right]. \tag{2}$$

Thus, privilege determines the conditional target $p_t^+$ at visited student states, while the outer problem distribution remains $P_D$.

As shown in Figure 2, PMSD extends privileged self-distillation from target specification to learning-measure adaptation. While OPSD retains the training distribution $P_D$ and uses privilege only through the conditional target $p^+$, PMSD additionally constructs a policy-dependent change of measure over training problems. For each problem, multiple trajectories are sampled from the current policy. The same frozen teacher evaluates their realized prefixes with and without the reference solution, yielding a signed trajectory-level contrast. Averaging these contrasts estimates the problem response $C_k(x)$ under the current policy. PMSD then uses $C_k(x)$ to construct a relative-entropy-constrained distribution $Q_k^\star$ over training problems.

### 3.2 PRIVILEGED CONTRAST ESTIMATION

Let $\mathcal{B}_k = \{(x_{k,b}, z_{k,b}^\star)\}_{b=1}^{B_k}$ denote the training problems sampled at optimization step $k$. For each $x_{k,b}$, the student policy $\pi_{\theta_k}$ generates $G$ independent trajectories

$$\xi_{k,bg} = (y_{k,bg1}, \ldots, y_{k,bgT_{k,bg}}) \sim \mu_{\theta_k}(\cdot \mid x_{k,b}), \quad g = 1, \ldots, G, \tag{3}$$

where $T_{bg}$ is the number of valid generated tokens and $\mu_{\theta_k}$ denotes the actual decoding law.

For each realized prefix $H_{bgt} = (x_b, y_{bg,<t})$, we evaluate the same frozen teacher under two paired information conditions, $p_{bgt}^0 = p_\phi(\cdot \mid H_{bgt})$ and $p_{bgt}^+ = p_\phi(\cdot \mid H_{bgt}, z_b^\star)$. The teacher parameters and realized student prefix are held fixed; only access to $z_b^\star$ differs. This paired evaluation isolates

the change in teacher support induced by privileged conditioning on the states actually visited by the current policy.

Rather than measuring an unconditional discrepancy between $p^0_{bgt}$ and $p^+_{bgt}$, PMSD associates the privileged shift with the realized student action. For each generated token $y_{bgt}$, we define the observed log-likelihood ratio

$$\imath_{bgt} = \log \frac{p^+_{bgt}(y_{bgt})}{p^0_{bgt}(y_{bgt})}, \tag{4}$$

Its sign preserves the direction of the privileged shift: positive values indicate increased support for the realized token under privileged conditioning, while negative values indicate decreased support.

We aggregate these local contrasts along the observed trajectory as

$$A^{\mathrm{obs}}_{bg} = \frac{1}{T_{bg}} \sum_{t=1}^{T_{bg}} \imath_{bgt}. \tag{5}$$

The length normalization yields a trajectory-level rate of privileged support rather than an accumulated quantity that scales mechanically with generation length. For truncated generations, $A^{\mathrm{obs}}_{bg}$ is defined on the observed prefix only. This contrast also admits a direct information-theoretic interpretation. For a fixed prefix and student token law $\mu$,

$$\mathbb{E}_{y\sim\mu}\left[\log \frac{p^+(y)}{p^0(y)}\right] = D_{\mathrm{KL}}(\mu\|p^0) - D_{\mathrm{KL}}(\mu\|p^+). \tag{6}$$

Hence, $A^{\mathrm{obs}}_{bg}$ characterizes how privileged conditioning changes the teacher's relative support for learner-generated behavior.

### 3.3 Problem-Level Privilege Estimation

For a training problem $x$, we define its response under the current policy as

$$C_k(x) = \mathbb{E}_{\xi\sim\mu_{\theta_k}(\cdot|x)}\left[A^{\mathrm{obs}}(x,\xi)\right]. \tag{7}$$

For a sampled problem $x_b$, this quantity is estimated from $G$ independent on-policy trajectories, $\widehat{C}_{k,b} = \frac{1}{G}\sum_{g=1}^{G} A^{\mathrm{obs}}_{bg}$.

Since $C_k(x)$ is defined with respect to $\mu_{\theta_k}(\cdot \mid x)$, it is specific to the current policy rather than an intrinsic property of the problem. Multiple rollouts provide a Monte Carlo estimate of this quantity and reduce the variance associated with individual sampled trajectories. They are used for estimating the problem-level response; the subsequent measure adaptation acts across training problems. For $G = 1$, the estimator remains well defined with $\widehat{C}_{k,b} = A^{\mathrm{obs}}_{b1}$.

### 3.4 Proximal Measure Adaptation

PMSD tilts this distribution according to the policy-dependent response $C_k(x)$:

$$Q^\star_k = \arg\max_{Q\ll P_D} \{\mathbb{E}_{x\sim Q}[C_k(x)] - \tau D_{\mathrm{KL}}(Q\|P_D)\} \quad \text{s.t.} \quad D_{\mathrm{KL}}(Q\|P_D) \le \kappa. \tag{8}$$

The temperature $\tau > 0$ determines the nominal strength of the exponential tilt, while $\kappa$ specifies the admissible relative-entropy displacement from $P_D$.

The optimizer is absolutely continuous with respect to $P_D$ and has density ratio

$$r^\star_k(x) \equiv \frac{dQ^\star_k}{dP_D}(x) = \frac{\exp(C_k(x)/(\tau+\lambda^\star_k))}{Z_k(\lambda^\star_k)}, \tag{9}$$

with

$$Z_k(\lambda) = \mathbb{E}_{x\sim P_D}\left[\exp(C_k(x)/(\tau+\lambda))\right]. \tag{10}$$

Here $\lambda^\star_k \ge 0$ is the multiplier associated with the KL constraint. If the nominal tilt at $\lambda^\star_k = 0$ is feasible, the constraint is inactive; otherwise $\lambda^\star_k$ satisfies $D_{\mathrm{KL}}(Q^\star_k\|P_D) = \kappa$. The derivation of this

optimal measure follows from the Lagrangian formulation and convex duality, and is provided in Appendix A.1.

The rollout conditional is left unchanged. For each problem, the $G$ on-policy trajectories are averaged uniformly. Hence trajectory sampling serves to estimate $C_k(x)$, whereas $r_k^\star(x)$ determines the cross-problem change of measure. Uniform multi-rollout OPSD is recovered when $Q_k^\star = P_D$, including the cases of constant $C_k(x)$, $\kappa = 0$, and the limit $\tau \to \infty$. Additional properties and limiting cases of the adapted measure are discussed in Appendix A.2.

**Stochastic approximation.** The population solution requires the partition function $Z_k(\lambda_k^\star)$ and the dual variable $\lambda_k^\star$. We track them online with two global iterates, $\widehat{\psi}_k \approx \log Z_k(\lambda_k^\star)$ and $\widehat{\lambda}_k \approx \lambda_k^\star$. For a sampled problem $x_b$, we first form the plug-in density ratio

$$\widehat{r}_{k,b} = \exp\left(\frac{\widehat{C}_{k,b}}{\tau + \widehat{\lambda}_k} - \widehat{\psi}_k\right). \tag{11}$$

The current on-policy samples provide stochastic estimates of the two population moments

$$\widehat{R}_k = \frac{1}{B_k}\sum_{b=1}^{B_k} \widehat{r}_{k,b}, \qquad \widehat{D}_k = \frac{1}{B_k}\sum_{b=1}^{B_k} \widehat{r}_{k,b}\log \widehat{r}_{k,b}. \tag{12}$$

We update the global normalizer and dual variable by

$$\widehat{\psi}_{k+1} = \widehat{\psi}_k + \eta_\psi\left(\widehat{R}_k - 1\right), \tag{13}$$

$$\widehat{\lambda}_{k+1} = \left[\widehat{\lambda}_k + \eta_\lambda\left(\widehat{D}_k - \kappa\right)\right]_+. \tag{14}$$

The first update tracks the normalization condition $\mathbb{E}_{P_D}[r_k(x)] = 1$, while the second tracks the relative-entropy constraint $\mathbb{E}_{P_D}[r_k(x)\log r_k(x)] \le \kappa$. The ratios used for the current student update are computed from $(\widehat{\psi}_k, \widehat{\lambda}_k)$ before these statistics are advanced, and are not renormalized within the minibatch.

### 3.5 Distillation under the Adapted Measure

Let

$$L^{\mathrm{OPSD}}(x,\xi;\theta) = \frac{1}{T_\xi}\sum_{t=1}^{T_\xi} \ell_{\mathrm{OPSD}}\left(p_\phi^+(\cdot \mid x,\xi_{<t}), p_\theta^S(\cdot \mid x,\xi_{<t})\right) \tag{15}$$

denote the trajectory-normalized OPSD loss, and define

$$\bar{L}_k^{\mathrm{OPSD}}(x;\theta) = \mathbb{E}_{\xi\sim\mu_{\theta_k}(\cdot|x)}\left[L^{\mathrm{OPSD}}(x,\xi;\theta)\right]. \tag{16}$$

PMSD optimizes

$$\mathcal{L}_{\mathrm{PMSD}}(\theta) = \mathbb{E}_{x\sim P_D}\left[\mathrm{sg}[r_k^\star(x)]\bar{L}_k^{\mathrm{OPSD}}(x;\theta)\right]. \tag{17}$$

The target distribution remains $p^+$; the density ratio changes only the relative contribution of training problems. In implementation, $r_k^\star(x)$ and the expectations above are replaced by their stochastic estimates, without minibatch renormalization of the density ratios. The complete training procedure is summarized in Appendix B.

### 3.6 Properties of Proximal Measure Adaptation

Define the per-problem surrogate gradient at the current iterate as

$$\mathcal{G}_k(x) = \nabla_\theta \bar{L}_k^{\mathrm{OPSD}}(x;\theta)\big|_{\theta=\theta_k}. \tag{18}$$

The change in population gradient relative to uniform OPSD is

$$\Delta\mathcal{G}_k = \mathbb{E}_{x\sim P_D}\left[(r_k^\star(x) - 1)\,\mathcal{G}_k(x)\right]. \tag{19}$$

Since $\mathbb{E}_{P_D}[r_k^\star(x)] = 1$,

$$\Delta\mathcal{G}_k = \mathrm{Cov}_{P_D}\left(r_k^\star(x), \mathcal{G}_k(x)\right). \tag{20}$$

If $\|\mathcal{G}_k(x)\| \le M$ for all $x$, Pinsker's inequality (Cover et al., 1991) gives

$$\|\Delta\mathcal{G}_k\| \le M\sqrt{2D_{\mathrm{KL}}(Q_k^\star\|P_D)} \le M\sqrt{2\kappa}. \tag{21}$$

Thus, the KL radius directly bounds the change in gradient aggregation induced by the adapted problem measure.

# 4 EVALUATION

**Models and Datasets.** We evaluate PMSD across three model scales in the Qwen3 family: Qwen3-1.7B, Qwen3-4B, and Qwen3-8B (Yang et al., 2025). Training is conducted on a mathematical reasoning subset derived from OpenThoughts (Guha et al., 2026), where reference solutions are available only to the privileged teacher view and are never exposed to the student during trajectory generation. We evaluate mathematical reasoning on AIME 2024, AIME 2025, and HMMT 2025, each comprising 30 problems. For each problem, we sample 12 independent responses and report the mean correctness across samples, denoted Avg@12. The overall score is the macro-average of Avg@12 over the three benchmarks.

**Baselines.** We compare PMSD with the pretrained BASE model, Supervised Fine-Tuning (SFT), Group Relative Policy Optimization (GRPO) (Shao et al., 2024), OPSD (Zhao et al., 2026b), and Purified OPSD (P-OPSD) (Shen et al., 2026). Among these methods, OPSD constitutes the primary control, as it uses the same privileged teacher target and on-policy student experience while retaining the original learning measure. We further construct a matched multi-rollout OPSD (MR-OPSD) baseline with the same number of sampled rollouts per problem, teacher configuration, token-level distillation loss, optimizer, and training budget as PMSD. Implementation details of baselines and evaluation protocols are provided in Appendix C.2. Detailed training configurations, including optimization settings, rollout parameters, and implementation choices, are provided in Appendix C.1.

## 4.1 OVERALL PERFORMANCE

Table 1: Overall mathematical reasoning performance (Avg@12, %). Avg. denotes the macro-average over AIME24, AIME25, and HMMT25. P-OPSD denotes Purified OPSD.

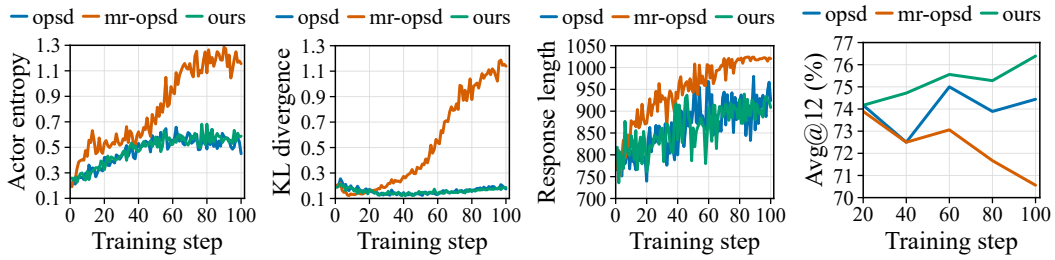
| Method | Qwen3-8B | | | Qwen3-4B | | | Qwen3-1.7B | | |
|---|---|---|---|---|---|---|---|---|---|
| | AIME24 | AIME25 | HMMT25 | AIME24 | AIME25 | HMMT25 | AIME24 | AIME25 | HMMT25 |
| Base | 76.39 | 66.94 | 45.56 | 73.33 | 65.83 | 42.78 | 47.50 | 38.61 | 23.33 |
| SFT | 73.89 | 66.11 | 43.89 | 69.44 | 60.83 | 41.11 | 50.56 | 36.11 | 25.00 |
| GRPO | 76.94 | 68.89 | 45.83 | 75.28 | 65.56 | 40.28 | 49.72 | 33.61 | 22.22 |
| OPSD | 77.22 | 69.44 | 47.50 | 74.44 | 65.28 | 44.44 | 55.00 | 40.83 | 25.56 |
| P-OPSD | 76.94 | 68.61 | 46.39 | 73.89 | 63.61 | 44.44 | 52.22 | 38.33 | 25.28 |
| **PMSD** | **78.33** | **70.28** | **48.06** | **76.39** | **67.22** | **45.00** | **58.61** | **41.67** | **28.61** |

Table 1 reports reasoning performance across three Qwen3 model scales. PMSD achieves the highest macro-average at all scales, improving over the strongest competing baseline by 0.84, 1.48, and 2.50 % on Qwen3-8B, Qwen3-4B, and Qwen3-1.7B, respectively. In all three cases, the strongest baseline is OPSD, which provides the most direct comparison because it uses privileged teacher supervision on student-generated trajectories while retaining the original learning measure. The improvement is consistent across model scales but is not uniform across benchmarks. On Qwen3-8B, PMSD improves over OPSD from 64.72 to 65.56 in macro-average, with gains of 1.11 % on AIME24, 0.84 % on AIME25, and 0.56 % on HMMT25. On Qwen3-4B, the corresponding improvement is larger, from 61.39 to 62.87, driven by gains of 1.95 and 1.94 % on AIME24 and AIME25, respectively, while HMMT25 improves more moderately by 0.56 %. The largest relative benefit appears at the 1.7B scale: PMSD raises the average from 40.46 to 42.96, including improvements of 3.61 % on AIME24 and 3.05 % on HMMT25.

Interestingly, the gain over OPSD becomes larger as model scale decreases. We treat this as an empirical trend rather than evidence of a general scaling law. One plausible interpretation is that weaker models induce more heterogeneous privileged responses across training problems, leaving

more room for measure adaptation to redistribute learning mass. Establishing this explanation, however, would require direct measurements of the score distribution $C_k(x)$ across model scales. The comparison with the remaining post-training baselines further suggests that the gain does not arise merely from additional optimization. SFT and GRPO are inconsistent across scales: both can improve individual benchmarks while reducing the macro-average relative to the pretrained model. In contrast, OPSD provides a stronger and substantially more stable baseline, and PMSD improves upon it at every model scale. This pattern is consistent with the central design of PMSD: the privileged target itself is not replaced; instead, PMSD changes how strongly different training problems contribute to the same OPSD objective. Overall, these results isolate the main empirical effect of PMSD: privileged information is useful not only for defining what should be distilled, but also for deciding where the corresponding distillation updates should place learning mass.

### 4.2 Training Dynamics




(a) Actor entropy. (b) KL divergence. (c) Response length. (d) AIME24 accuracy.

Figure 3: Training dynamics of different methods on Qwen3-4B.

Figure 3 compares PMSD with standard OPSD and MR-OPSD, where MR-OPSD uses the same $G = 4$ rollout budget as PMSD but retains uniform learning mass. This controls for the effect of additional on-policy samples. Here, KL divergence denotes the predictive divergence between the student and privileged teacher, rather than the measure KL $D_{\mathrm{KL}}(Q_k^\star \| P_D)$.

MR-OPSD progressively departs from the other two methods: actor entropy rises, student–teacher KL increases, and response length approaches the generation limit, while AIME24 accuracy deteriorates in the latter half of training. Standard OPSD is more stable in these statistics, but its accuracy fluctuates across checkpoints. PMSD stays close to OPSD in entropy, predictive KL, and response length, while showing a steadier improvement in accuracy and reaching 76.39% at the final checkpoint. Since MR-OPSD uses the same rollout budget, the comparison suggests that the gain comes from how multi-rollout experience is weighted rather than from additional sampling alone.

### 4.3 Ablations

Table 2: Ablation study of PMSD on Qwen3-4B across mathematical reasoning benchmarks (Avg@12, %). Avg. denotes the macro-average over AIME24, AIME25, and HMMT25. Arrows indicate changes relative to full PMSD in percentage points.

| Variant | AIME24 | AIME25 | HMMT25 | Avg. |
|---|---|---|---|---|
| MR-OPSD | 70.56 (↓ 5.83) | 65.28 (↓ 1.94) | 43.89 (↓ 1.11) | 59.91 (↓ 2.96) |
| w/o KL Constraint | 74.72 (↓ 1.67) | 66.39 (↓ 0.83) | 43.61 (↓ 1.39) | 61.57 (↓ 1.30) |
| Trajectory-Level PMSD | 75.56 (↓ 0.83) | 66.39 (↓ 0.83) | 44.44 (↓ 0.56) | 62.13 (↓ 0.74) |
| Divergence-Based PMSD | 75.00 (↓ 1.39) | 66.39 (↓ 0.83) | 44.17 (↓ 0.83) | 61.85 (↓ 1.02) |

Table 2 examines the main components of PMSD. Removing measure adaptation, as in MR-OPSD, produces the largest drop, from 62.87 % to 59.91 % Avg@12 ($-2.96$ % points), despite using the same $G = 4$ rollout budget. Removing the KL constraint reduces performance by 1.30 points, indicating that the Gibbs tilt benefits from explicitly limiting deviation from the original training measure. Applying the measure update at the trajectory level yields 62.13 ($-0.74$ %), supporting

Table 3: Hyperparameter sensitivity on Qwen3-4B (Avg@12, %). Arrows indicate changes relative to the default PMSD setting ($\tau = 0.10$, $\kappa = 0.05$) in percentage points.

| Parameter | Value | AIME24 | AIME25 | HMMT25 | Avg. |
|---|---|---|---|---|---|
| Temperature $\tau$ | 0.20 | 75.83 ($\downarrow$ 0.56) | 66.67 ($\downarrow$ 0.55) | 44.72 ($\downarrow$ 0.28) | 62.41 ($\downarrow$ 0.46) |
| | 0.50 | 75.28 ($\downarrow$ 1.11) | 66.11 ($\downarrow$ 1.11) | 44.17 ($\downarrow$ 0.83) | 61.85 ($\downarrow$ 1.02) |
| KL budget $\kappa$ | 0.01 | 75.56 ($\downarrow$ 0.83) | 66.39 ($\downarrow$ 0.83) | 44.44 ($\downarrow$ 0.56) | 62.13 ($\downarrow$ 0.74) |
| | 0.10 | 75.83 ($\downarrow$ 0.56) | 66.67 ($\downarrow$ 0.55) | 44.17 ($\downarrow$ 0.83) | 62.22 ($\downarrow$ 0.65) |

the use of repeated rollouts to estimate the problem-level response $C_k(x)$ rather than treating each trajectory as an independent allocation unit. Replacing the signed log-likelihood ratio with an unsigned teacher-view divergence further reduces the average to 61.85 % ($-1.02$ %), showing that the direction of the privilege-induced change matters in addition to its magnitude. Overall, the ablations support the three main choices in PMSD: bounded measure adaptation, problem-level aggregation, and signed privilege evidence.

### 4.4 Hyperparameters

Table 3 studies the two hyperparameters governing the measure update. The default setting uses $\tau = 0.10$ and $\kappa = 0.05$, yielding 62.87 % Avg@12. Increasing $\tau$ weakens performance: $\tau = 0.20$ and $0.50$ reduce the macro-average by 0.46 and 1.02 %, respectively. This is consistent with the Gibbs form $r_k^\star(x) \propto \exp(C_k(x)/(\tau+\lambda_k^\star))$, where a larger temperature attenuates the effect of $C_k(x)$ and moves the adapted measure closer to the original distribution. Since only larger temperatures are evaluated, these results support the need for nontrivial measure tilting rather than establishing $\tau = 0.10$ as globally optimal.

The KL radius shows a complementary pattern. Tightening the budget from $\kappa = 0.05$ to $0.01$ decreases the average by 0.74 %, while relaxing it to $0.10$ decreases it by 0.65 %. This is consistent with the role of the trust region: an overly small radius limits redistribution, whereas an overly large one weakens proximal control. Performance remains within roughly one point of the default across all tested settings, indicating that PMSD is not highly sensitive to either hyperparameter.

## 5 Limitation and Discussion

PMSD shows that privileged information can shape both the distillation target and the learning measure over on-policy problems. Across model scales, the results and ablations support bounded problem-level measure adaptation, while matched multi-rollout OPSD shows that additional rollouts alone do not explain the gains. However, several limitations also remain. The privilege contrast measures changes in teacher support rather than reasoning correctness and may capture non-semantic effects. Its problem-level estimate is based on few rollouts, and the KL constraint controls measure redistribution rather than global policy evolution. Our evaluation is also limited to Qwen3 models, mathematical reasoning, and reference-solution privilege.

## AI Use Statement

We acknowledge the use of AI tools as supplementary aids for code generation, software debugging, and manuscript preparation, including language polishing, grammar correction, and improving clarity and presentation. All core methodological ideas, algorithm design, experimental setup, execution, data analysis, and reported results were developed, conducted, and independently verified by the authors.

## Ethics Statement

This work studies post-training methods for mathematical reasoning and does not involve human subjects or newly collected personal or sensitive data. The experiments use existing language models, training data, and mathematical reasoning benchmarks. Relevant considerations include the

provenance and licensing of training data and reference solutions, potential benchmark contamination, and the limitations of automatic answer-based evaluation.

## Reproducibility Statement

We provide the complete PMSD objective and optimization procedure in Section 3, together with the model, dataset, baseline, and evaluation protocols in Section 4. The appendix contains additional derivations, the training algorithm, and implementation details, including the main hyperparameters and optimization settings. Code, configurations, and evaluation scripts necessary to reproduce the experiments will be provided as anonymized supplementary material.

## 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 *International Conference on Learning Representations*, volume 2024, pp. 21246–21263, 2024.

Samy Bengio, Oriol Vinyals, Navdeep Jaitly, and Noam Shazeer. Scheduled Sampling for Sequence Prediction with Recurrent Neural Networks. volume 28, 2015.

Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, et al. Training Verifiers to Solve Math Word Problems. *arXiv preprint arXiv:2110.14168*, 2021.

Thomas M Cover, Joy A Thomas, and John Kieffer. *Elements of Information Theory*, volume 2. wiley New York, 1991.

Tommaso Furlanello, Zachary Lipton, Michael Tschannen, Laurent Itti, and Anima Anandkumar. Born Again Neural Networks. In *International Conference on Learning Representations*, pp. 1607–1616. PMLR, 2018.

Yuxian Gu, Li Dong, Furu Wei, and Minlie Huang. MiniLLM: Knowledge Distillation of Large Language Models. In *International Conference on Learning Representations*, volume 2024, pp. 32694–32717, 2024.

Etash Guha, Ryan Marten, Sedrick Keh, Negin Raoof, Georgios Smyrnis, Hritik Bansal, Marianna Nezhurina, Jean Mercat, Trung Vu, Zayne Sprague, et al. OpenThoughts: Data Recipes for Reasoning Models. In *International Conference on Learning Representations*, volume 2026, pp. 108059–108130, 2026.

Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Peiyi Wang, Qihao Zhu, Runxin Xu, Ruoyu Zhang, Shirong Ma, Xiao Bi, et al. DeepSeek-R1 Incentivizes Reasoning in LLMs through Reinforcement Learning. *Nature*, 645(8081):633–638, 2025.

Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. Distilling the Knowledge in a Neural Network. *arXiv preprint arXiv:1503.02531*, 2015.

Wenjin Hou, Shangpin Peng, Weinong Wang, Zheng Ruan, Yue Zhang, Zhenglin Zhou, Mingqi Gao, Yifei Chen, Kaiqi Wang, Hongming Yang, et al. Uni-OPD: Unifying On-Policy Distillation with a Dual-Perspective Recipe. *arXiv preprint arXiv:2605.03677*, 2026.

Yuki Ichihara, Naoto Iwase, Mohammad Atif Quamar, and Junpei Komiyama. Privileged Solutions or Context-Induced Teacher Behavior? Dissecting On-Policy Self-Distillation. *arXiv preprint arXiv:2608.09228*, 2026.

Simran Kaur, Narutatsu Ri, Yinghui He, Liam Fowl, and Sanjeev Arora. Rethinking On-Policy Self-Distillation for Thinking Models. *arXiv preprint arXiv:2607.05184*, 2026.

Junlong Ke, Zichen Wen, Weijia Li, Conghui He, and Linfeng Zhang. Respecting Self-Uncertainty in On-Policy Self-Distillation for Efficient LLM Reasoning. *arXiv preprint arXiv:2605.13255*, 2026.

Yoon Kim and Alexander M Rush. Sequence-Level Knowledge Distillation. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pp. 1317–1327, 2016.

Yuying Li, Leqi Zheng, Yongzi Yu, Wenrui Zhou, Xuchang Zhong, Xing Hu, Jing Jin, Hangjie Yuan, and Tao Feng. Filter, Then Reweight: Rethinking Optimization Granularity in On-Policy Distillation. *arXiv preprint arXiv:2606.02684*, 2026.

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

Jianhua Lin. Divergence Measures Based on the Shannon Entropy. *IEEE Transactions on Information theory*, 37(1):145–151, 1991.

David Lopez-Paz, Léon Bottou, Bernhard Schölkopf, and Vladimir Vapnik. Unifying Distillation and Privileged Information. 2015.

Xuan-Phi Nguyen, Shrey Pandit, Yiran Zhao, Anurag Koul, Zeyu Liu, and Shafiq Joty. Privileged Likelihood Is Not Automatically Value: Three Checks for Token Credit in On-Policy Self-Distillation. *arXiv preprint arXiv:2608.09263*, 2026.

Leyi Pan, Shuchang Tao, Yunpeng Zhai, Lingzhe Zhang, Zhaoyang Liu, Bolin Ding, Aiwei Liu, and Lijie Wen. RLCSD: Reinforcement Learning with Contrastive On-Policy Self-Distillation. *arXiv preprint arXiv:2606.11709*, 2026.

Jan Peters, Katharina Mulling, and Yasemin Altun. Relative Entropy Policy Search. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 24, pp. 1607–1612, 2010.

Stéphane Ross, Geoffrey Gordon, and Drew Bagnell. A Reduction of Imitation Learning and Structured Prediction to No-Regret Online Learning. In *Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics*, pp. 627–635. JMLR Workshop and Conference Proceedings, 2011.

John Schulman, Sergey Levine, Pieter Abbeel, Michael Jordan, and Philipp Moritz. Trust Region Policy Optimization. In *International Conference on Machine Learning*, pp. 1889–1897. Pmlr, 2015.

John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. Proximal Policy Optimization Algorithms. *arXiv preprint arXiv:1707.06347*, 2017.

Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, YK Li, Yang Wu, et al. DeepSeekMath: Pushing the Limits of Mathematical Reasoning in Open Language Models. *arXiv preprint arXiv:2402.03300*, 2024.

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.

Lieven Vandenberghe and Stephen Boyd. *Convex Optimization*, volume 1. Cambridge university press Cambridge, 2004.

Vladimir Vapnik and Akshay Vashist. A New Learning Paradigm: Learning Using Privileged Information. *Neural Networks*, 22(5-6):544–557, 2009.

Yuanda Xu, Hejian Sang, Zhengze Zhou, Ran He, Zhipeng Wang, and Alborz Geramifard. TIP: Token Importance in On-Policy Distillation. *arXiv preprint arXiv:2604.14084*, 2026.

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.

Wenkai Yang, Weijie Liu, Ruobing Xie, Kai Yang, Saiyong Yang, and Yankai Lin. Learning Beyond Teacher: Generalized On-Policy Distillation with Reward Extrapolation. *arXiv preprint arXiv:2602.12125*, 2026a.

Yongkang Yang, Zhezheng Hao, Hong Zhang, Yi Liu, Xiankun Lin, Wence Ji, Fanjunduo Wei, Jiarui Yu, Qiang Lin, Xiaoyun Liang, et al. Matching Supervision to the Student's Learning Capacity: A Unified Framework for On-Policy Self-Distillation. *arXiv preprint arXiv:2608.08176*, 2026b.

Wenhao Zhang. Beyond Absolute Imitation: Anchored Residual Guidance for Privileged On-Policy Distillation. *arXiv preprint arXiv:2606.10385*, 2026.

Qingfei Zhao, Huan Song, Shuyu Tian, Jiawei Shao, and Xuelong Li. Prefix-Guided On-Policy Distillation: Mining Golden Trajectories from Rollouts. *arXiv preprint arXiv:2606.21994*, 2026a.

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. *arXiv preprint arXiv:2601.18734*, 2026b.

Binbin Zheng, Xing Ma, Yiheng Liang, Jingqing Ruan, Xiaoliang Fu, Kepeng Lin, Benchang Zhu, Ke Zeng, and Xunliang Cai. SCOPE: Signal-Calibrated On-Policy Distillation Enhancement with Dual-Path Adaptive Weighting. *arXiv preprint arXiv:2604.10688*, 2026.

Yuhang Zhou, Kai Zheng, Haoling Li, Dengyun Peng, Can Xu, and Jingjing Chen. TurnOPD: Making On-Policy Distillation Turn-Aware for Efficient Long-Horizon Agent Training. *arXiv preprint arXiv:2607.05804*, 2026.

Siqi Zhu, Xuyan Ye, Hongyu Lu, Weiye Shi, and Ge Liu. The Many Faces of On-Policy Distillation: Pitfalls, Mechanisms, and Fixes. *arXiv preprint arXiv:2605.11182*, 2026.

# A Additional Method Details

## A.1 Derivation of the Optimal Measure

We derive the closed-form solution of the proximal measure adaptation problem in Eq. (8). Let $r(x) = \frac{dQ(x)}{dP_D(x)}$ denote the density ratio between the adapted measure and the original training measure. The optimization problem can be rewritten as

$$\max_r \mathbb{E}_{x\sim P_D}\left[r(x)C_k(x) - \tau r(x)\log r(x)\right], s.t. \mathbb{E}_{P_D}[r(x)] = 1 \text{ and } \mathbb{E}_{P_D}[r(x)\log r(x)] \le \kappa.$$

Introducing a Lagrange multiplier $\lambda \ge 0$ for the KL constraint and $\eta$ for the normalization constraint gives the Lagrangian

$$\mathcal{J}(r) = \mathbb{E}_{P_D}\left[r(x)C_k(x) - (\tau+\lambda)r(x)\log r(x) - \eta r(x)\right] + \lambda\kappa + \eta.$$

Taking the functional derivative with respect to $r(x)$ yields

$$C_k(x) - (\tau+\lambda)(1+\log r(x)) - \eta = 0.$$

Therefore,

$$r(x) = \exp\left(\frac{C_k(x) - \eta'}{\tau+\lambda}\right),$$

where $\eta'$ is the normalization constant. Applying $\mathbb{E}_{P_D}[r(x)] = 1$ gives

$$r_k^\star(x) = \frac{\exp(C_k(x)/(\tau+\lambda_k^\star))}{Z_k(\lambda_k^\star)}.$$

The multiplier $\lambda_k^\star$ is determined by the complementary slackness condition. When the unconstrained exponential tilt satisfies the KL constraint, $\lambda_k^\star = 0$; otherwise it increases until the KL boundary is reached.

## A.2 Properties of the Adapted Measure

**Recovery of uniform OPSD.** PMSD reduces to standard OPSD when the adapted measure remains identical to the original training distribution. This occurs when the privilege response is constant across training problems: $C_k(x) = c,\ \forall x$. In this case,

$$r_k^\star(x) = \frac{\exp(c/(\tau+\lambda))}{\mathbb{E}_{P_D}[\exp(c/(\tau+\lambda))]} = 1,$$

and therefore

$$Q_k^\star = P_D.$$

Hence, PMSD does not introduce additional reweighting when privileged information provides no relative preference among training problems.

# B Training Algorithm

Algorithm 1 summarizes the complete training procedure of PMSD. At each optimization step, PMSD first samples a batch of problems from the original training distribution and generates $G$ on-policy responses from the current student policy. For each sampled trajectory, the frozen teacher is evaluated under two conditions: without access to the reference solution and with access to the reference solution. The difference between these two teacher views produces the trajectory-level privilege evidence. The trajectory-level signals are then averaged across the $G$ rollouts of the same problem to estimate the problem-level privilege response $C_k(x)$. PMSD uses this response to compute the adapted density ratio and reweights the original training measure. The weighted OPSD objective is subsequently used to update the student policy. Following the optimization formulation in Section 3.4, the normalizer and the dual variable are updated after the student optimization step using the current estimate of the adapted measure.

**Algorithm 1** Training procedure of PMSD

**Require:** Student policy $\pi_\theta$, frozen teacher $\phi$, training distribution $P_D$, rollout number $G$, temperature $\tau$, KL budget $\kappa$

**for** each training iteration $k$ **do**
  Sample a batch of problems $\{x_i\} \sim P_D$
  **for** each problem $x_i$ **do**
    Generate $G$ trajectories $\{\xi_{i,g}\}_{g=1}^{G} \sim \pi_{\theta_k}$
    **for** each trajectory $\xi_{i,g}$ **do**
      Compute teacher probabilities $p_\phi^0$ and $p_\phi^+$
      Compute observed privilege contrast $A^{obs}(x_i, \xi_{i,g})$
    **end for**
    Estimate problem response $C_k(x_i) = \frac{1}{G}\sum_g A^{obs}(x_i, \xi_{i,g})$
  **end for**
  Compute adapted density ratio $r_k(x)$
  Optimize weighted OPSD objective using $r_k(x)$
  Update normalizer and dual variables
**end for**

# C Implementation Details

## C.1 Configuration

Table 4 summarizes the implementation configurations used in all experiments. Unless otherwise specified, all methods are trained with the same student initialization, optimizer, batch configuration, and training budget. PMSD-specific hyperparameters, including the rollout number, measure temperature, and KL budget, are also reported.

Table 4: Implementation configurations.

| Configuration | Value |
|---|---|
| Base models | Qwen3-1.7B, Qwen3-4B, and Qwen3-8B |
| Teacher model | Frozen base checkpoint at the corresponding scale |
| Training data | Official OPSD OpenThoughts-derived math training split |
| Student prompting | Privilege-free; thinking disabled |
| Teacher prompting | Privileged/native views; thinking enabled |
| Rollout number $G$ | 4 per problem |
| Measure temperature $\tau$ | 0.10 |
| KL budget $\kappa$ | 0.05 |
| Sampling temperature | 1.1 |
| Sampling top-$p$ / top-$k$ | 0.95 / 20 |
| Maximum generation length | 1024 tokens |
| Optimizer | AdamW |
| Learning rate | $5 \times 10^{-6}$ |
| LoRA rank / scaling | $r = 64$, $\alpha = 128$ |
| Precision | bf16 |
| Batch size | 32 problems |
| Training budget | 100 steps |
| Hardware | 8× NVIDIA A100 GPUs, 40 GB per GPU |
| Random seed | 42 |

## C.2 Baseline and Evaluation Details

All baselines use the same student initialization, training data, optimization budget, and evaluation protocol as PMSD. The only difference lies in how the training objective incorporates privileged information or sampled trajectories. For baselines with publicly available implementations, we follow the released code and recommended configurations. When an official implementation is unavailable,

we reproduce the method following the algorithmic description and experimental protocol reported in the original paper as closely as possible.

**OPSD.** OPSD follows the original privileged self-distillation formulation. The student generates on-policy trajectories, and the frozen teacher provides privileged and non-privileged supervision. Unlike PMSD, OPSD uses the original training measure and assigns uniform importance to all training problems.

**MR-OPSD.** MR-OPSD is designed as a matched multi-rollout baseline. It uses the same rollout number $G$, teacher evaluations, decoding configuration, and distillation objective as PMSD. The only difference is that all sampled problems receive identical weights, removing the measure adaptation component.

**P-OPSD.** Purified OPSD modifies the token-level distillation target. On student-generated trajectories, it evaluates the same frozen base model under three views: question plus reference, reference only with the question omitted, and question only without the reference. At each shared trajectory prefix, the log-probability difference between the privileged and reference-only views provides a residual that is combined with the question-only base distribution to construct a PMI target. The student is distilled toward this purified target. PMSD instead retains the original privileged teacher target and uses the privileged-versus-question-only contrast to compute problem weights.

**Evaluation protocol.** We evaluate all methods on AIME24, AIME25, and HMMT25. For each benchmark, the model generates 12 independent responses, and the reported metric is Avg@12. The final score is computed as the macro average over the three benchmarks. Answers are evaluated using exact matching after standard answer extraction.