FAIR CREDIT ACROSS CORRECT MODES: MITIGATING CORRECT-PATH COVERAGE COLLAPSE IN ON-POLICY SELF-DISTILLATION FOR LANGUAGE MODELS

Anonymous authors
Paper under double-blind review

ABSTRACT

On-policy self-distillation (OPSD) is a promising approach for improving reasoning in large language models through dense supervision from a context-augmented version of the model. However, OPSD with sampled demonstrations (OPSD-SD) can achieve strong Pass@1 accuracy while showing limited gains from additional rollouts, a phenomenon we call *correct-path coverage collapse*. Beyond general mode-seeking tendencies, demonstration conditioning in OPSD-SD can exacerbate this collapse through pointwise conditional mutual information (PCMI)-induced sharpening, further favoring already-probable correct solutions. To this end, we propose $\texttt{C}^2\texttt{OPSD}$, an OPSD framework that combines PCMI Debiasing (PD) and Diversity-Aware Exploration (DAE). PD corrects demonstration-dependent preferences among equally correct responses. Under idealized conditions, we show that this correction exactly cancels the additional PCMI-induced sharpening between correct responses. DAE then assigns credit according to each correct path's marginal contribution to diversity, encouraging underrepresented paths. Experiments across multiple datasets demonstrate that $\texttt{C}^2\texttt{OPSD}$ mitigates correct-path coverage collapse while outperforming strong baselines on both Pass@1 and Pass@k. Code is available at `https://anonymous.4open.science/r/C2OPSD-0612`.

1 INTRODUCTION

Reinforcement learning with verifiable rewards (RLVR) has substantially improved large language model reasoning by optimizing policies using scalar rewards derived from automatically verified outcomes (Shao et al., 2024a; Guo et al., 2025). However, these response-level scalar rewards do not provide supervision for intermediate reasoning steps, resulting in a sparse credit-assignment bottleneck (Chan et al., 2024). On-policy distillation (OPD) addresses this limitation by providing dense teacher feedback on student-generated trajectories, thereby reducing the mismatch between training and inference (Agarwal et al., 2024; Gu et al., 2023; Hou et al., 2026). On-policy self-distillation (OPSD) (Zhao et al., 2026; Hübotter et al., 2026) further removes the need for an external teacher by assigning the same model distinct teacher and student roles.

Despite this significant progress, recent work (Nicolicioiu et al., 2026) shows that on-policy self-distillation with sampled demonstrations (OPSD-SD) can improve Pass@1 yet reduce rollout diversity and flatten Pass@k curves. Like on-policy reinforcement learning (RL), OPSD-SD exhibits mode-seeking, whereby optimization concentrates rollouts on a smaller subset, but its optimal policy can additionally sharpen imbalances among equally rewarded responses (Nicolicioiu et al., 2026). To distinguish this additional contraction from generic mode-seeking, we refer to this phenomenon as *correct-path coverage collapse*, which captures the diminished coverage of distinct correct reasoning paths. According to the analysis in Nicolicioiu et al. (2026), correct-path coverage collapse arises from two compounding mechanisms: reverse KL-induced mode-seeking and pointwise conditional mutual information (PCMI)-induced sharpening. First, reverse KL-induced mode-seeking concentrates updates on high-probability rollouts sampled from the current policy, leaving lower-probability alternatives with fewer opportunities for reinforcement. Second, OPSD-SD samples

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.

both a student rollout and a correct demonstration, making common solutions more likely to align with the teacher context. This double-sampling bias increases their expected PCMI, causing the optimal policy to tilt further toward these common solutions. As a result, this likely-get-likelier effect compounds generic mode seeking, progressively widening probability gaps, suppressing rarer valid modes, and shrinking correct-path coverage under the same sampling budget.

Existing methods (Cheng et al., 2026; Jin et al., 2026; Li et al., 2026c) have focused primarily on mitigating reverse KL-induced mode-seeking, whereas PCMI-induced sharpening remains underexplored. Specifically, their core strategy is to preserve entropy during optimization, either by modifying advantage estimates with entropy-based signals (Cheng et al., 2026) or by adopting entropy-aware adaptive objectives (Jin et al., 2026). Such interventions reshape optimization dynamics, but they do not correct biases encoded in the target optimum. As shown in Nicolicioiu et al. (2026), PCMI-induced sharpening operates at this latter level by reweighting responses that receive identical correctness rewards according to their expected PCMI with the sampled demonstration. Consequently, existing methods may alleviate generic concentration, but they cannot mitigate PCMI-induced sharpening or restore base-policy probability ratios among correct responses.

To mitigate correct-path coverage collapse, a principled solution should first neutralize PCMI-induced sharpening and then address any residual mode-seeking tendency. Accordingly, two primary challenges must be addressed. (1) **Debiasing PCMI-induced sharpening**. Because this sharpening arises in the optimal policy induced by the sampled-demonstration objective, it cannot generally be eliminated by tuning KL regularization or other optimization heuristics alone. (2) **Expanding correct-path coverage**. Quantifying the diversity of correct paths is inherently difficult in LLM reasoning tasks. Yet such quantification is necessary to construct supervisory signals that incentivize the model to explore a broader range of correct paths.

To address these challenges, we propose C^{2}OPSD, a self-distillation framework designed to mitigate correct-path coverage collapse. Building on OPSD, C^{2}OPSD first mitigates PCMI-induced sharpening and then broadens correct-path coverage. For the first challenge, C^{2}OPSD constructs a PCMI-based debiasing advantage that downweights high-PCMI correct rollouts while favoring low-PCMI ones. Our theoretical analysis further shows that, under idealized conditions, setting the debiasing coefficient γ to 1 eliminates PCMI-induced sharpening. For the second challenge, C^{2}OPSD introduces a diversity metric that quantifies differences among correct reasoning paths and assigns greater credit to underrepresented paths. Our main contributions are as follows:

- We propose C^{2}OPSD, an OPSD framework that mitigates correct-path coverage collapse through two components: PCMI debiasing and diversity-aware exploration.
- We theoretically analyze PCMI debiasing by deriving its optimal policy under the OPSD-SD-KL objective. Under idealized conditions, we prove that PCMI debiasing with $\gamma = 1$ exactly cancels PCMI-induced sharpening among correct responses. We characterize how standardization, clipping, and length normalization affect the effective debiasing strength.
- We conduct experiments across several reasoning datasets to verify the effectiveness of C^{2}OPSD. The experimental results show that C^{2}OPSD outperforms strong baselines and improves exploration of underrepresented paths while maintaining Pass@1.

2 Preliminaries

2.1 On-Policy Self-Distillation

Let x denote an input question, π_0 a base policy, and π_θ the current student policy. For each x, the student samples a group of N on-policy rollouts,

$$\mathcal{Y}(x) = \{y_i\}_{i=1}^{N}, \qquad y_i = (y_{i,1}, \ldots, y_{i,T_i}) \sim \pi_\theta(\cdot \mid x), \tag{1}$$

where $y_{i,<t}$ is the prefix preceding token $y_{i,t}$. A verifier assigns a reward $r(x, y_i)$, and $\mathcal{C}(x) \subseteq \mathcal{Y}(x)$ denotes the subset of verified-correct rollouts. We write $y^{\mathrm{corr}} \sim p_{\mathrm{corr}}(\cdot \mid x)$ for a correct demonstration sampled from a distribution supported on $\mathcal{C}(x)$. The regularized teacher is denoted by $\pi_{\bar{\theta}}$, where $\bar{\theta}$ is typically an exponential moving average (EMA) of θ. We use $D_{\mathrm{KL}}(p\|q)$ for the Kullback–Leibler divergence and $\mathrm{sg}[\cdot]$ for the stop-gradient operator.

On-policy self-distillation uses the same model in two roles: the student generates an initial response, whereas a context-augmented version of the model acts as a teacher that retrospectively evaluates the student's token choices. Specifically, the student first samples $\mathcal{Y}(x)$ and receives verifier or environment feedback. The privileged teacher context may contain tokenized environment feedback or a successful response from the same rollout group. According to Hübotter et al. (2026), OPSD-SD provides the EMA teacher with a verified-correct rollout y^{corr} generated by the student. With $q_{i,t}^{\text{corr}}$ defined as $\pi_{\bar{\theta}}(\cdot \mid x, y^{\text{corr}}, y_{i,<t})$, the student is optimized toward these next-token distributions through the reverse-KL objective (Hübotter et al., 2026; Nicolicioiu et al., 2026),

$$\mathcal{L}_{\text{SD}}(\theta; x) = \frac{1}{N} \sum_{y_i \in \mathcal{Y}(x)} \mathbb{E}_{y^{\text{corr}} \sim p_{\text{corr}}(\cdot|x)} \left[\sum_{t=1}^{T_i} D_{\text{KL}}\Big(\pi_\theta(\cdot \mid x, y_{i,<t}) \parallel \text{sg}[q_{i,t}^{\text{corr}}] \Big) \right]. \tag{2}$$

The stop-gradient prevents the teacher from being optimized toward the unconditioned student. Equivalently, the log-probability ratio between the teacher and student provides a dense, token-dependent advantage: a token favored after conditioning on the privileged context receives positive credit, whereas a disfavored token receives negative credit.

2.2 Collapse in On-Policy Self-Distillation

To explain correct-path coverage collapse, we recall the tilted distribution of the optimal policy for KL-regularized RL and OPSD-SD with the reverse KL objective (Nicolicioiu et al., 2026). Further details are provided in Appendix A.2.1. For a sequence-level reward $r_x(y)$ and coefficient $\beta_{\text{RL}} > 0$, according to prior works (Rafailov et al., 2023; Korbak et al., 2022), the optimal policy satisfies

$$\pi^\star_{\text{RL}}(y \mid x) \propto \pi_0(y \mid x) \exp\left(\frac{r_x(y)}{\beta_{\text{RL}}} \right). \tag{3}$$

Consequently, if two correct responses receive the same reward, the reward tilt preserves their probability ratio under π_0. We then recall the tilted distribution of the optimal policy under OPSD-SD with the reverse KL objective (OPSD-SD-KL) according to Nicolicioiu et al. (2026). Specifically, the PCMI and its expectation over correct demonstrations are defined as

$$i(y; y^{\text{corr}} \mid x) := \log \frac{\pi_0(y \mid x, y^{\text{corr}})}{\pi_0(y \mid x)}, I_x(y) := \mathbb{E}_{y^{\text{corr}} \sim p_{\text{corr}}(\cdot|x)}[i(y; y^{\text{corr}} \mid x)]. \tag{4}$$

These quantities measure how much conditioning on a correct demonstration changes the base policy's log-probability of a candidate response, for one demonstration and on average, respectively. Thus, the optimal OPSD-SD-KL policy is

$$\pi^\star_{\text{OPSD-SD-KL}}(y \mid x) \propto \pi_0(y \mid x) \exp\left(\frac{I_x(y)}{1+\beta} \right). \tag{5}$$

Thus, unlike the reward tilt in Eq. (3), OPSD-SD-KL additionally reweights responses according to their expected PCMI with sampled demonstrations. For two equally correct responses y_a and y_b,

$$\frac{\pi^\star_{\text{OPSD-SD-KL}}(y_a \mid x)}{\pi^\star_{\text{OPSD-SD-KL}}(y_b \mid x)} = \frac{\pi_0(y_a \mid x)}{\pi_0(y_b \mid x)} \exp\left(\frac{I_x(y_a) - I_x(y_b)}{1+\beta} \right). \tag{6}$$

Therefore, even when y_a and y_b are equally correct, unequal expected PCMI changes their relative probability beyond the base-policy imbalance. In particular, when the already more probable response also has higher expected PCMI, the exponential factor further shifts probability toward this common response, producing a rich-get-richer concentration among correct paths. We refer to this mechanism as PCMI-induced sharpening, formalized as follows.

Definition 1 (PCMI-induced sharpening)**.** *For two responses y_a and y_b, we define their sharpening under π relative to the base policy π_0 as*

$$\mathcal{S}_{ab}(\pi) := \log \frac{\pi(y_a \mid x)}{\pi(y_b \mid x)} - \log \frac{\pi_0(y_a \mid x)}{\pi_0(y_b \mid x)}.$$

$\mathcal{S}_{ab}(\pi)$ measures the change in log-probability imbalance relative to π_0.

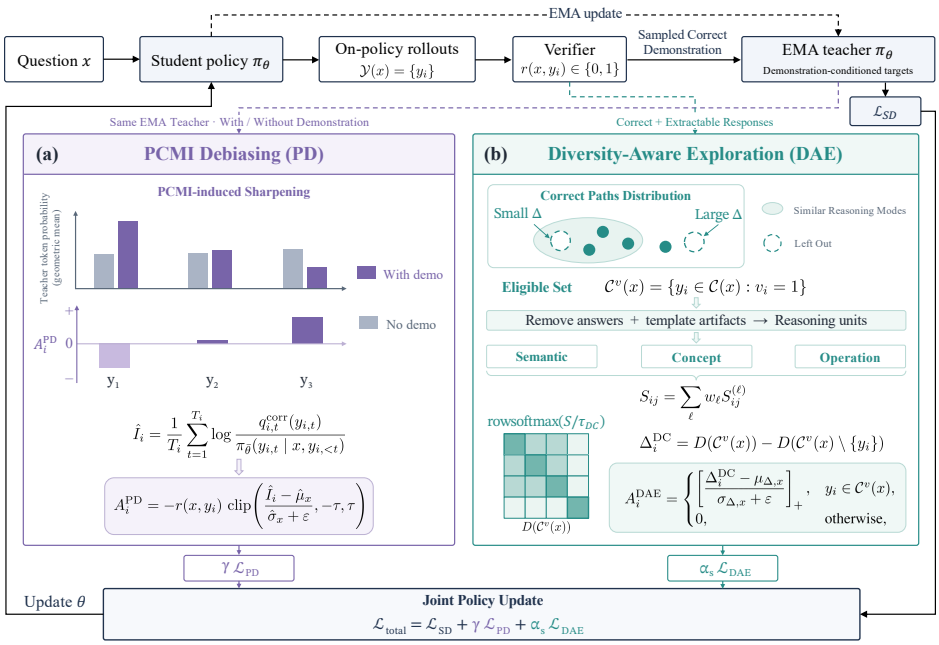


Figure 1: Overview of C^{2}OPSD. (a) PCMI Debiasing corrects demonstration-induced preference. (b) Diversity-Aware Exploration assigns credit to complementary correct rationales.

3 Methodology

3.1 Overview

To address the three challenges identified in Section 1, we propose C^{2}OPSD, illustrated in Figure 1. Building on the OPSD paradigm described in Section 2.1, C^{2}OPSD introduces two components: *PCMI Debiasing (PD)* and *Diversity-Aware Exploration (DAE)* First, debiasing PCMI-induced sharpening requires correcting demonstration-dependent preferences among equally correct responses. PD therefore downweights high-PCMI correct responses and favors low-PCMI ones. Meanwhile, our theoretical analysis characterizes the conditions for ideal cancellation of this bias. Second, expanding correct-path coverage requires a measurable signal that distinguishes responses with complementary reasoning from those with redundant reasoning. DAE provides this signal by quantifying each correct response's marginal contribution to representation-based coverage.

3.2 PCMI Debiasing

In the OPSD setting, a model conditioned on a sampled correct demonstration can favor responses aligned with the demonstration's reasoning, even when alternative responses are equally correct. This preference is quantified by the PCMI-induced sharpening defined in Definition 1, which measures changes in the log-probability ratio between correct responses relative to the base policy. To mitigate this sharpening, PD downweights high-PCMI correct responses and upweights low-PCMI ones, thereby separating demonstration-induced preference from verifier-assessed correctness.

PD first estimates the demonstration-induced preference of each response using a sampled, length-normalized online analogue of the PCMI in Eq. (4). Specifically, we compare each response's token probabilities under the same EMA teacher with and without a sampled correct demonstration. Using $q_{i,t}^{\mathrm{corr}}$ defined in Section 2.1, we compute

$$\hat{I}_i = \frac{1}{T_i}\sum_{t=1}^{T_i} \log \frac{q_{i,t}^{\mathrm{corr}}(y_{i,t})}{\pi_{\bar{\theta}}(y_{i,t} \mid x, y_{i,<t})}. \tag{7}$$

By isolating the effect of demonstration conditioning while controlling for response length, $\hat{I}_i$ measures the current teacher's relative demonstration-induced preference for response y_i.

Based on $\hat{I}_i$, we construct a correction advantage to mitigate the model's PCMI-induced sharpening. Let $\hat{\mu}_x$ and $\hat{\sigma}_x$ denote the mean and standard deviation of $\hat{I}_i$ over $\mathcal{C}(x)$. With binary verifier rewards $r(x, y_i) \in \{0, 1\}$, we define

$$A_i^{\mathrm{PD}} = -r(x, y_i)\ \mathrm{clip}\left(\frac{\hat{I}_i - \hat{\mu}_x}{\hat{\sigma}_x + \varepsilon}, -\tau, \tau\right), \tag{8}$$

where $\varepsilon > 0$ stabilizes normalization and $\tau > 0$ bounds the correction magnitude. This advantage assigns negative corrections to verified-correct responses with above-average $\hat{I}_i$ and positive corrections to those with below-average $\hat{I}_i$. Computing the group statistics over $\mathcal{C}(x)$ and masking the advantage with $r(x, y_i)$ prevent incorrect responses from receiving positive credit merely because their PCMI is low. In practice, the advantage is treated as a fixed training signal and set to zero when fewer than two correct responses are available.

Furthermore, we theoretically show that PD can exactly eliminate PCMI-induced sharpening under idealized conditions, as stated in Theorem 2. Due to space constraints, we defer the correction-strength analyses under standardization alone and under combined standardization and clipping to Appendix A.2, specifically Propositions 2 and 3, respectively.

Theorem 1 (Exact cancellation of PCMI-induced sharpening)**.** *Under idealized PCMI Debiasing without standardization or clipping, setting the correction coefficient to $\gamma = 1$ yields, for any two correct responses $y_a, y_b \in \mathcal{C}_x$,*

$$\mathcal{S}_{ab}(\pi_1^\star) = 0.$$

Therefore, ideal PCMI Debiasing exactly cancels the additional sharpening introduced by demonstration-conditioned self-distillation between correct modes.

3.3 Diversity-Aware Exploration

While PD mitigates PCMI-induced sharpening, reverse KL-induced mode-seeking can still concentrate updates on common correct reasoning paths, leaving underrepresented correct paths with less reinforcement. PD therefore needs to be complemented by an explicit exploration incentive to broaden correct-path coverage. To encourage exploration of underrepresented correct paths, DAE assigns credit to each correct response according to its marginal contribution to reasoning-path diversity within the sampled group. Responses that contribute complementary reasoning receive greater credit than those that repeat reasoning already represented in the group.

To quantify each correct response's marginal contribution to reasoning-path diversity, DAE first aims to capture differences in the underlying reasoning logic rather than superficial differences in wording or expression style. We therefore combine three complementary views that characterize what a response means, which knowledge it uses, and how its reasoning proceeds. Let $v_i \in \{0, 1\}$ indicate whether reasoning content can be successfully extracted from y_i after removing final answers and template artifacts, and define $\mathcal{C}^v(x) = \{y_i \in \mathcal{C}(x) : v_i = 1\}$. We divide the reasoning content extracted from each correct response in $\mathcal{C}^v(x)$ into reasoning units using line breaks and sentence boundaries. These units are then used to construct the following representations:

- **Semantic view.** We average and normalize reasoning-unit embeddings from a frozen encoder to capture reasoning meaning beyond surface wording.
- **Concept view.** We use term frequency–inverse document frequency (TF-IDF) features (Salton & Buckley, 1988) of content unigrams and bigrams, including formulas, numerical values, and units, to approximate the explicit knowledge used in reasoning.
- **Reasoning-operation view.** We assign rule-based operation labels, such as causal reasoning, comparison, and elimination, to reasoning units, then use label frequencies and adjacent-label transitions to characterize reasoning structure.

Based on these views, we obtain feature vectors $z_i^{(\ell)}$, where $\ell \in \{\mathrm{semantic}, \mathrm{concept}, \mathrm{operation}\}$, for each correct response $y_i \in \mathcal{C}^v(x)$. We compute the cosine similarity $S_{ij}^{(\ell)}$ between the corresponding nonzero feature vectors and combine the view-specific similarities as

$$S_{ij} = \sum_\ell w_\ell S_{ij}^{(\ell)},\ w_\ell \geq 0,\ \sum_\ell w_\ell = 1. \tag{9}$$

We use default weights of 0.60, 0.25, and 0.15, respectively. Overall, the combined similarity aims to capture shared reasoning content with less sensitivity to wording and expression style. Based on S, we use DCScore (Zhu et al., 2025) to quantify reasoning-path diversity within $\mathcal{C}^v(x)$ and derive response-level marginal contributions:

$$D(\mathcal{C}^v(x)) = \mathrm{tr}[\mathrm{softmax}(S/\tau_{\mathrm{DC}})]\,, \tag{10}$$

$$\Delta_i^{\mathrm{DC}} = D(\mathcal{C}^v(x)) - D(\mathcal{C}^v(x) \setminus \{y_i\}),\ y_i \in \mathcal{C}^v(x), \tag{11}$$

where $\mathrm{tr}(\cdot)$ denotes the matrix trace operation, $\mathrm{softmax}(\cdot)$ is applied row-wise, and $\tau_{\mathrm{DC}} > 0$ controls the sensitivity of the score to similarity.

According to DCScore (Zhu et al., 2025), the row-wise softmax compares each response's self-similarity with its similarities to other responses in $\mathcal{C}^v(x)$. Responses with fewer similar reasoning paths receive larger diagonal probabilities, whereas those with more similar paths receive smaller ones. The marginal contribution Δ_i^{DC} is the decrease in group diversity after removing y_i, with repeated samples treated as distinct indexed occurrences. For this leave-one-out calculation, we remove only the row and column corresponding to y_i from S. We then recompute the row-wise softmax on the resulting principal submatrix to obtain $D(\mathcal{C}^v(x) \setminus \{y_i\})$.

We convert the marginal contributions into a nonnegative exploration advantage:

$$A_i^{\mathrm{DAE}} = \begin{cases} \left[\dfrac{\Delta_i^{\mathrm{DC}} - \mu_{\Delta,x}}{\sigma_{\Delta,x} + \varepsilon}\right]_+, & y_i \in \mathcal{C}^v(x), \\ 0, & \text{otherwise}, \end{cases} \tag{12}$$

where $\mu_{\Delta,x}$ and $\sigma_{\Delta,x}$ denote the mean and population standard deviation of Δ_i^{DC} over $\mathcal{C}^v(x)$, $\varepsilon > 0$ stabilizes normalization, and $[u]_+ = \max(u, 0)$. Additionally, this advantage is set to zero when fewer than two eligible responses are available or all marginals are identical. This advantage assigns positive credit to responses with above-average marginal diversity contributions and zero credit to those at or below the group average. The nonnegative formulation avoids penalizing common correct solutions, whose marginal contributions are often below average because similar reasoning is already represented in the sampled group. Without $[\cdot]_+$, these solutions would receive negative credit despite being correct. Such penalties could shift the policy excessively toward less common solutions, potentially degrading pass@1 and destabilizing training. By retaining only positive credit, DAE encourages exploration of underrepresented correct paths without explicitly penalizing common correct solutions for their lower marginal diversity contributions.

3.4 Unified Objective and Training Algorithm

C^{2}OPSD combines the distillation objective with PD and DAE:

$$\mathcal{L}_{\mathrm{total}} = \mathcal{L}_{\mathrm{SD}} + \gamma \mathcal{L}_{\mathrm{PD}} + \alpha_s \mathcal{L}_{\mathrm{DAE}}, \tag{13}$$

where γ controls debiasing strength, and α_s follows an exploration warm-up and ramp schedule For a batch $\mathcal{B}$ of question–response pairs (x_i, y_i), let $\ell_i(\theta) = \sum_{t=1}^{T_i} \log \pi_\theta(y_{i,t} \mid x_i, y_{i,<t})$ and $\bar{T}_{\mathcal{B}} = |\mathcal{B}|^{-1} \sum_{i \in \mathcal{B}} T_i$. The auxiliary losses are

$$\mathcal{L}_z = -\frac{\sum_{i \in \mathcal{B}} \mathrm{sg}[A_i^z] \ell_i(\theta)}{|\mathcal{B}| \bar{T}_{\mathcal{B}}},\ z \in \{\mathrm{PD}, \mathrm{DAE}\}, \tag{14}$$

with advantages computed within each question group. The quality loss $\mathcal{L}_Q$ retains separate PPO clipping. Limited by space, we provide the training algorithm of C^{2}OPSD in the Appendix A.3.

4 Experiments

4.1 Experimental Setup

Datasets. We conduct experiments on four science question-answering datasets from SciKnowEval (Feng et al., 2024), covering Biology, Chemistry, Materials, and Physics. These datasets consist of undergraduate-level multiple-choice science questions. We split each dataset into training and test sets following the protocol used in SDPO (Hübotter et al., 2026).

Table 1: Pass@1 and Pass@16 at the step 200 for Qwen3-1.7B and Qwen3-4B. Mean ± std across 3 seeds. The best results are highlighted in **bold**.

Method	Biology		Chemistry		Materials		Physics		Average	
	Avg@16	Pass@16	Avg@16	Pass@16	Avg@16	Pass@16	Avg@16	Pass@16	Avg@16	Pass@16
Qwen3-1.7B										
GRPO	$41.63_{\pm 1.32}$	$81.33_{\pm 5.03}$	$58.09_{\pm 5.24}$	$84.13_{\pm 3.24}$	$\mathbf{69.35_{\pm 4.46}}$	$\mathbf{84.04_{\pm 11.06}}$	$56.02_{\pm 1.41}$	$\mathbf{89.17_{\pm 0.72}}$	$56.27_{\pm 3.11}$	$84.67_{\pm 5.01}$
SDPO	$50.83_{\pm 5.26}$	$74.67_{\pm 9.87}$	$72.48_{\pm 2.26}$	$86.03_{\pm 0.73}$	$62.59_{\pm 2.89}$	$71.99_{\pm 2.21}$	$56.64_{\pm 2.31}$	$80.42_{\pm 1.44}$	$60.64_{\pm 3.18}$	$78.28_{\pm 3.56}$
C²OPSD	$\mathbf{51.04_{\pm 1.23}}$	$\mathbf{90.67_{\pm 3.06}}$	$\mathbf{73.60_{\pm 0.35}}$	$\mathbf{88.41_{\pm 1.67}}$	$62.70_{\pm 2.26}$	$80.85_{\pm 1.84}$	$\mathbf{57.21_{\pm 1.76}}$	$80.83_{\pm 4.02}$	$\mathbf{61.14_{\pm 1.40}}$	$\mathbf{85.19_{\pm 2.65}}$
Qwen3-4B										
GRPO	$41.13_{\pm 1.19}$	$71.33_{\pm 4.16}$	$64.83_{\pm 0.36}$	$76.19_{\pm 0.95}$	$\mathbf{77.59_{\pm 1.09}}$	$\mathbf{84.04_{\pm 2.13}}$	$62.11_{\pm 1.35}$	$87.50_{\pm 4.33}$	$61.41_{\pm 1.00}$	$79.77_{\pm 2.89}$
SDPO	$47.75_{\pm 4.82}$	$62.00_{\pm 6.00}$	$75.10_{\pm 1.83}$	$\mathbf{94.13_{\pm 1.53}}$	$67.46_{\pm 4.42}$	$75.53_{\pm 4.64}$	$\mathbf{67.63_{\pm 1.44}}$	$\mathbf{87.92_{\pm 1.91}}$	$64.49_{\pm 3.13}$	$79.89_{\pm 3.52}$
C²OPSD	$\mathbf{50.17_{\pm 2.82}}$	$\mathbf{82.00_{\pm 4.00}}$	$\mathbf{76.43_{\pm 1.26}}$	$93.49_{\pm 3.17}$	$70.88_{\pm 2.20}$	$75.53_{\pm 3.19}$	$63.91_{\pm 1.84}$	$84.17_{\pm 5.05}$	$\mathbf{65.34_{\pm 2.03}}$	$\mathbf{83.79_{\pm 3.85}}$

Baselines and Models. We compare C²OPSD with two baselines, GRPO and SDPO. GRPO (Shao et al., 2024b) is an RLVR method that optimizes policies using group-relative rewards without requiring an explicit critic. SDPO (Hübotter et al., 2026) performs on-policy self-distillation, with the same model serving as both student and teacher. We use Qwen3-1.7B and Qwen3-4B (Yang et al., 2025) as the backbone models.

Implementation Details. We report results averaged over three seeds, using the model checkpoint at training step 200 for the final evaluation. Experiments are conducted on six NVIDIA RTX 6000D GPUs using AdamW with a learning rate of 5×10^{-6} and 10 warm-up steps. Each training batch contains 8 prompts, with 8 responses sampled per prompt at a temperature of 1.0. The maximum prompt and response lengths are 2,048 and 8,192 tokens, respectively. For evaluation, we sample 16 responses per prompt with a temperature of 0.6 and top-p of 0.95 to compute pass@k. We provide more experimental details in the Appendix A.5.

4.2 Comparison Study

As shown in Table 1, we compare C²OPSD with two on-policy training baselines, GRPO and SDPO. Averaged across datasets, C²OPSD outperforms both baselines in Pass@1 and Pass@16, which suggests greater benefits from additional rollouts. These findings are consistent with the intended roles of its two components in mitigating PCMI-induced sharpening and encouraging exploration of diverse reasoning paths. At the dataset level, C²OPSD shows substantial advantages on Biology and Chemistry. The gains are particularly pronounced on Biology, where C²OPSD outperforms the second-best method by just over 10 percentage points in Pass@16 with Qwen3-1.7B. Although C²OPSD does not achieve the best results on Materials and Physics, it improves both metrics over the OPSD baseline SDPO in most cases. These improvements further support the effectiveness of C²OPSD. We provide more detailed results and further analysis in the Appendix A.4.1.

4.3 Ablation Study

To assess the contribution of each component, we remove PD and DAE separately, yielding C²OPSD w/o PD and C²OPSD w/o DAE, respectively. We also remove both components of C²OPSD to yield C²OPSD w/o PD&DAE, which is equivalent to SDPO. Table 2 reports the ablation results on the Biology and Physics datasets. We observe that both PD and DAE improve Pass@16 while maintaining or improving Pass@1, indicating the mitigation of correct-path collapse. Furthermore, PD contributes more to the Pass@16 improvement than DAE.

Table 2: Ablation results for Qwen3-1.7B on Biology and Physics datasets. Mean ± std across 3 seeds. The best results are highlighted in **bold**.

Method	Biology		Physics	
	Avg@16	Pass@16	Avg@16	Pass@16
C²OPSD w/o PD	$50.88_{\pm 1.84}$	$76.00_{\pm 3.46}$	$56.20_{\pm 1.18}$	$80.42_{\pm 1.91}$
C²OPSD w/o DAE	$50.58_{\pm 1.45}$	$78.67_{\pm 4.62}$	$55.55_{\pm 1.88}$	$\mathbf{81.88_{\pm 2.65}}$
C²OPSD w/o PD&DAE	$50.83_{\pm 5.26}$	$74.67_{\pm 9.87}$	$56.64_{\pm 2.31}$	$80.42_{\pm 1.44}$
C²OPSD	$\mathbf{51.04_{\pm 1.23}}$	$\mathbf{90.67_{\pm 3.06}}$	$\mathbf{57.21_{\pm 1.76}}$	$80.83_{\pm 4.02}$

4.4 Hyperparameter Sensitivity

Figure 2 presents the sensitivity of C²OPSD to γ and α_s on Biology and Physics. To isolate each component, we vary γ with only PD enabled and α_s with only DAE enabled. On Biology, increasing

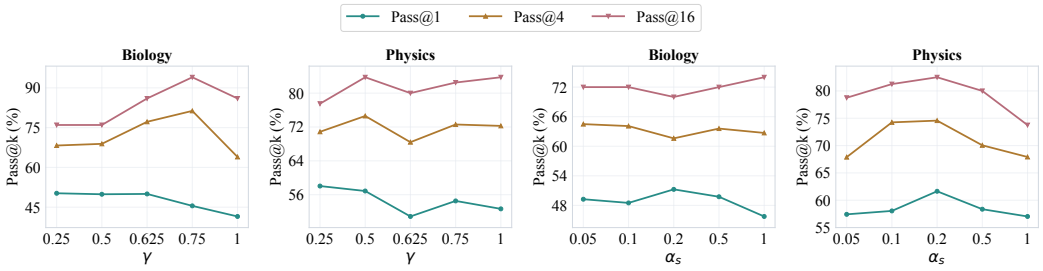


Figure 2: Sensitivity analysis of the hyperparameters γ and α_s on the Biology and Physics datasets.

γ generally improves Pass@16 while reducing Pass@1, indicating a reduced preference for common solutions. When γ exceeds 0.75, however, both metrics decline, suggesting that excessive emphasis on uncommon solutions may hinder learning. Physics also shows a trade-off between Pass@1 and Pass@16, although its Pass@16 trend is less consistent across γ values. In contrast, performance is less sensitive to α_s on Biology, whereas Physics performs best around $\alpha_s = 0.2$ and degrades at larger values. Notably, the effect of α_s on performance varies across datasets.

4.5 Visualization

PCMI Distribution. Figure 3 compares the PCMI distributions of the initial model before training, the model trained with SDPO, and the model trained with C^{2}OPSD using PD only. According to Figure 3, we observe that PD increases the representation of lower-PCMI correct responses while retaining a high-PCMI component. SDPO training shifts the correct-response distribution toward positive values, increasing its mean PCMI from -0.747 to 0.193. This shift suggests that the SDPO-trained model is more inclined to generate common solutions. Compared with SDPO, adding PD reduced this mean to 0.060, whereas the mean across all valid responses changed only slightly, from -0.034 to -0.053. These findings suggest that PD rebalances the distribution of correct responses rather than uniformly lowering PCMI, consistent with mitigating PCMI-induced sharpening while preserving correct responses strongly supported by demonstrations.

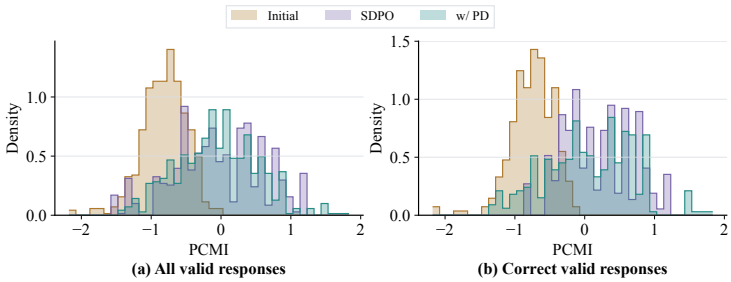


Figure 3: PCMI distributions for the initial model, SDPO, and C^{2}OPSD with PD only (w/ PD). These histograms show the PCMI scores of (a) all valid responses and (b) correct valid responses.

Diversity Distribution. Figure 4 presents the per-question rollout diversity distributions of the initial model, SDPO, and C^{2}OPSD with DAE only. Compared with the initial model, C^{2}OPSD consistently improves the overall diversity distribution across all three metrics, whereas SDPO alone tends to reduce rollout diversity despite increasing surface-level lexical variation. Specifically, mean $1 - \text{Self-BLEU}$ increased from 0.286 to 0.420 with SDPO and 0.417 with C^{2}OPSD. However, SDPO reduced mean cosine distance from 0.173 to 0.117 and DCScore from 1.366 to 1.115, with approximately 60% of questions concentrated near their lower bounds. This discrepancy indicates that greater lexical variation can coexist with semantic collapse. In contrast, DAE increased mean cosine distance to 0.460 and DCScore to 1.644 while preserving the higher lexical diversity observed after training. These results show that DAE promotes the exploration of alternative solution paths rather than merely varying surface wording, thereby improving rollout diversity.

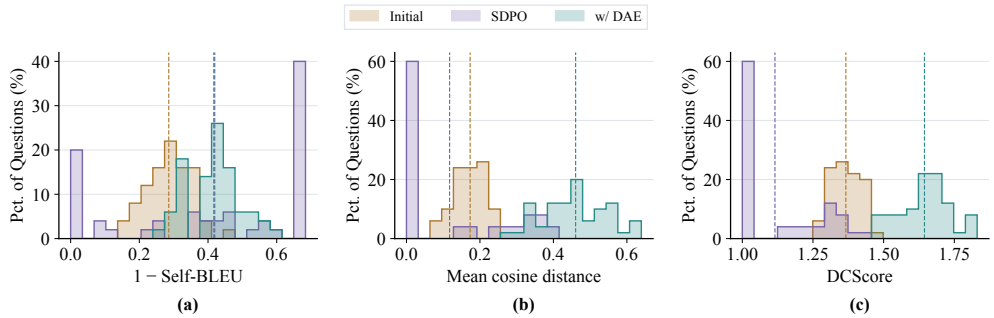


Figure 4: Diversity distributions for the initial model, SDPO, and C^{2}OPSD with DAE only (w/ DAE). These histograms show the per-question rollout diversity measured by (a) 1−Self-BLEU, (b) mean pairwise cosine distance, and (c) DCScore. Dashed lines indicate distribution means.

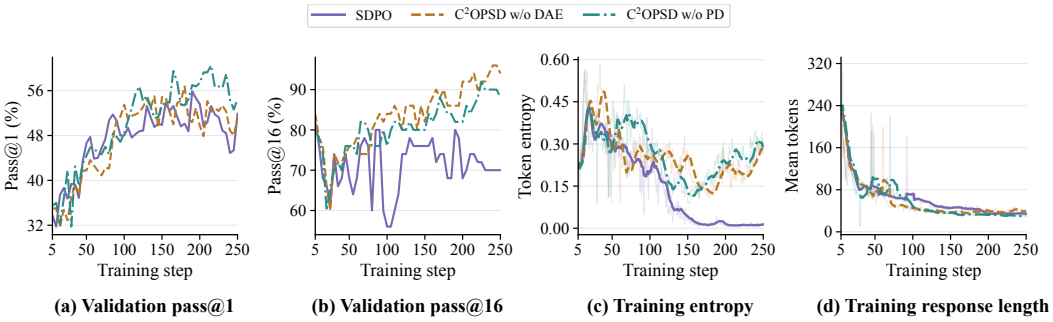


Figure 5: Training dynamics of C^{2}OPSD on the Biology dataset.

Case Study. We compare responses generated by SDPO and C^{2}OPSD to investigate their reasoning behavior. Due to space constraints, we present the examples and detailed analysis in Appendix A.4.2. These examples suggest that C^{2}OPSD explores a broader range of correct reasoning paths, whereas SDPO exhibits diversity collapse and appears to memorize answers.

4.6 Training Dynamics

We present the training dynamics of its individual components in Figure 5. Both PD and DAE improve Pass@16 while maintaining Pass@1, suggesting that they mitigate correct-path collapse. Specifically, standard SDPO exhibits a nearly monotonic decline in entropy and response length, while its validation performance plateaus and eventually deteriorates. In contrast, PD and DAE maintain higher entropy, consistent with reduced concentration of the response distribution. Despite these shared trends, PD and DAE contribute differently to performance. PD mitigates PCMI-induced sharpening and reduces the preference for common solutions, but requires carefully balancing Pass@1 and Pass@16. DAE uses nonnegative exploration bonuses to encourage underrepresented solutions while preserving coverage of common solutions. This yields a more favorable balance between Pass@1 and Pass@16. Moreover, unlike the entropy-based approach in (Cheng et al., 2026), both components of C^{2}OPSD improve Pass@16 without increasing response length.

5 Conclusion

In this work, we investigate correct-path collapse in on-policy self-distillation and propose C^{2}OPSD, which combines PCMI Debiasing and Diversity-Aware Exploration. PCMI Debiasing mitigates PCMI-induced sharpening, while Diversity-Aware Exploration encourages exploration of underrepresented correct reasoning paths. Meanwhile, our theoretical analysis establishes conditions under which PCMI Debiasing exactly cancels PCMI-induced sharpening among correct responses in an idealized setting. Experiments across multiple datasets show that C^{2}OPSD improves Pass@k while maintaining Pass@1, providing empirical evidence that it mitigates correct-path collapse.

AI USE STATEMENT

We used large language models to polish the manuscript by improving grammatical accuracy, clarity, and readability, and to assist in checking and correcting mathematical proofs by identifying potential gaps in logical reasoning and inconsistencies in notation. We also used generative AI tools to support data visualization, including assistance with plotting code and refinement of figure layouts, labels, and formatting.

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.

Alex J Chan, Hao Sun, Samuel Holt, and Mihaela Van Der Schaar. Dense reward for free in reinforcement learning from human feedback. *arXiv preprint arXiv:2402.00782*, 2024.

Zhipeng Chen, Xiaobo Qin, Youbin Wu, Yue Ling, Qinghao Ye, Wayne Xin Zhao, and Guang Shi. Pass@ k training for adaptively balancing exploration and exploitation of large reasoning models. *arXiv preprint arXiv:2508.10751*, 2025.

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

Kehua Feng, Xinyi Shen, Weijie Wang, Xiang Zhuang, Yuqi Tang, Qiang Zhang, and Keyan Ding. Sciknoweval: Evaluating multi-level scientific knowledge of large language models. *arXiv preprint arXiv:2406.09098*, 2024.

Yuxian Gu, Li Dong, Furu Wei, and Minlie Huang. Minillm: On-policy distillation of large language models. *arXiv preprint arXiv:2306.08543*, 2023.

Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Peiyi Wang, Qihao Zhu, Runxin Xu, Ruoyu Zhang, Shirong Ma, Xiao Bi, et al. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*, 2025.

Anthony GX-Chen, Jatin Prakash, Jeff Guo, Rob Fergus, and Rajesh Ranganath. Kl-regularized reinforcement learning for generative modelling is designed to mode collapse. In *International Conference on Learning Representations*, volume 2026, pp. 50665–50693, 2026.

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.

Jonas Hübotter, Frederike Lübeck, Lejs Behric, Anton Baumann, Marco Bagatella, Daniel Marta, Ido Hakimi, Idan Shenfeld, Thomas Kleine Buening, Carlos Guestrin, et al. Reinforcement learning via self-distillation. *arXiv preprint arXiv:2601.20802*, 2026.

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

Jeonghye Kim, Xufang Luo, Minbeom Kim, Sangmook Lee, Dohyung Kim, Jiwon Jeon, Dongsheng Li, and Yuqing Yang. Why does self-distillation (sometimes) degrade the reasoning capability of llms? *arXiv preprint arXiv:2603.24472*, 2026.

Tomasz Korbak, Ethan Perez, and Christopher L. Buckley. RL with KL penalties is better viewed as bayesian inference. In Yoav Goldberg, Zornitsa Kozareva, and Yue Zhang (eds.), *Findings of the Association for Computational Linguistics: EMNLP 2022, Abu Dhabi, United Arab Emirates, December 7-11, 2022*, volume EMNLP 2022 of *Findings of ACL*, pp. 1083–1091. Association for Computational Linguistics, 2022. doi: 10.18653/V1/2022.FINDINGS-EMNLP.77. URL https://doi.org/10.18653/v1/2022.findings-emnlp.77.

Chenyi Li, Yuan Zhang, Bo Wang, Guoqing Ma, Wei Tang, Haoyang Huang, and Nan Duan. Setpo: Set-level policy optimization for diversity-preserving llm reasoning. *arXiv preprint arXiv:2602.01062*, 2026a.

Long Li, Zhijian Zhou, Jiaran Hao, Jason Liu, Yanting Miao, Wei Pang, Xiaoyu Tan, Wei Chu, Zhe Wang, Shirui Pan, et al. The choice of divergence: A neglected key to mitigating diversity collapse in reinforcement learning with verifiable reward. In *International Conference on Learning Representations*, volume 2026, pp. 64742–64768, 2026b.

Tianjian Li, Yiming Zhang, Ping Yu, Swarnadeep Saha, Daniel Khashabi, Jason Weston, Jack Lanchantin, and Tianlu Wang. Jointly reinforcing diversity and quality in language model generations. *arXiv preprint arXiv:2509.02534*, 2025.

Xiaozhe Li, Yang Li, Xinyu Fang, Shengyuan Ding, Peiji Li, Yongkang Chen, Yichuan Ma, Tianyi Lyu, Linyang Li, Dahua Lin, et al. Beyond mode collapse: Distribution matching for diverse reasoning. *arXiv preprint arXiv:2605.19461*, 2026c.

Yijiang Li, Bingyang Wang, Yijun Liang, Yunjie Tian, Di Fu, and Nuno Vasconcelos. On-policy self-distillation without any supervision. *arXiv preprint arXiv:2608.06296*, 2026d.

Yunhe Li, Hao Shi, Wenhao Liu, Mengzhe Ruan, Hanxu Hou, Zhongxiang Dai, Shuang Qiu, and Linqi Song. Demopsd: Disagreement-modulated policy self-distillation. *arXiv preprint arXiv:2607.02502*, 2026e.

Andrei Liviu Nicolicioiu, Mohammad Pezeshki, and Aaron Courville. On-policy self-distillation with sampled demonstrations reduces output diversity. *arXiv preprint arXiv:2606.26091*, 2026.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D. Manning, Stefano Ermon, and Chelsea Finn. Direct preference optimization: Your language model is secretly a reward model. In Alice Oh, Tristan Naumann, Amir Globerson, Kate Saenko, Moritz Hardt, and Sergey Levine (eds.), *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*, 2023. URL http://papers.nips.cc/paper_files/paper/2023/hash/a85b405ed65c6477a4fe8302b5e06ce7-Abstract-Conference.html.

Gerard Salton and Chris Buckley. Term-weighting approaches in automatic text retrieval. *Inf. Process. Manag.*, 24(5):513–523, 1988. doi: 10.1016/0306-4573(88)90021-0. URL https://doi.org/10.1016/0306-4573(88)90021-0.

Hejian Sang, Yuanda Xu, Zhengze Zhou, Ran He, Zhipeng Wang, and Jiachen Sun. Crisp: Compressed reasoning via iterative self-policy distillation. *arXiv preprint arXiv:2603.05433*, 2026.

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*, 2024a.

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*, 2024b.

Guobin Shen, Xiang Cheng, Chenxiao Zhao, Lei Huang, Jindong Li, Dongcheng Zhao, and Xing Yu. Anti-self-distillation for reasoning rl via pointwise mutual information. *arXiv preprint arXiv:2605.11609*, 2026a.

Guobin Shen, Lei Huang, Xiang Cheng, Chenxiao Zhao, Jindong Li, Dongcheng Zhao, and Xing Yu. From generic correlation to input-specific credit in on-policy self distillation. *arXiv preprint arXiv:2605.11613*, 2026b.

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

Idan Shenfeld, Mehul Damani, Jonas Hübotter, and Pulkit Agrawal. Self-distillation enables continual learning. *arXiv preprint arXiv:2601.19897*, 2026.

Zhongwei Wan, Yun Shen, Zhihao Dou, Donghao Zhou, Yu Zhang, Xin Wang, Hui Shen, Jing Xiong, Chaofan Tao, Zixuan Zhong, Peizhou Huang, and Mi Zhang. Dsdr: Dual-scale diversity regularization for exploration in llm reasoning, 2026. URL https://arxiv.org/abs/2602.19895.

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.

Chenxu Yang, Chuanyu Qin, Qingyi Si, Minghui Chen, Naibin Gu, Dingyu Yao, Zheng Lin, Weiping Wang, Jiaqi Wang, and Nan Duan. Self-distilled rlvr. *arXiv preprint arXiv:2604.03128*, 2026a.

Run Yang, Runpeng Dai, Jie Sun, Jielei Zhang, Fan Zhou, Hongtu Zhu, Peiyi Li, and Longwen Gao. Influence-directed distillation: Solving the diversity bottleneck in sampled-token on-policy distillation. *arXiv preprint arXiv:2608.29846*, 2026b.

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

Dongxu Zhang, Zhichao Yang, Sepehr Janghorbani, Jun Han, Andrew Ressler II, Qian Qian, Gregory D Lyng, Sanjit Singh Batra, and Robert E Tillman. Fast and effective on-policy distillation from reasoning prefixes. In *Findings of the Association for Computational Linguistics: ACL 2026*, pp. 25553–25569, 2026.

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*, 2026.

Yaoming Zhu, Sidi Lu, Lei Zheng, Jiaxian Guo, Weinan Zhang, Jun Wang, and Yong Yu. Texygen: A benchmarking platform for text generation models. In *The 41st international ACM SIGIR conference on research & development in information retrieval*, pp. 1097–1100, 2018.

Yuchang Zhu, Huizhe Zhang, Bingzhe Wu, Jintang Li, Zibin Zheng, Peilin Zhao, Liang Chen, and Yatao Bian. Measuring diversity in synthetic datasets. In Aarti Singh, Maryam Fazel, Daniel Hsu, Simon Lacoste-Julien, Felix Berkenkamp, Tegan Maharaj, Kiri Wagstaff, and Jerry Zhu (eds.), *Forty-second International Conference on Machine Learning, ICML 2025, Vancouver, BC, Canada, July 13-19, 2025*, volume 267 of *Proceedings of Machine Learning Research*. PMLR / OpenReview.net, 2025. URL https://proceedings.mlr.press/v267/zhu25ac.html.

A Appendix

A.1 Related Work

A.1.1 On-Policy Distillation

OPD provides dense teacher feedback on student-generated trajectories, as exemplified by MiniLLM (Gu et al., 2023) and GKD (Agarwal et al., 2024). Building on this paradigm, subsequent work explores contextual knowledge transfer (Ye et al., 2026) and efficient distillation from reasoning prefixes (Zhang et al., 2026). Despite this progress, conventional OPD typically requires a separate, often larger teacher. To remove this dependency while exploiting available reference solutions, OPSD assigns the same model distinct teacher and student roles, using privileged context to provide dense supervision on student-generated rollouts (Zhao et al., 2026). Related approaches extend this principle to continual learning through demonstration-conditioned self-teachers (Shenfeld et al., 2026) and reasoning improvement through environmental feedback (Hübotter et al., 2026). Further developments explore reasoning compression (Sang et al., 2026) and self-generated supervision (Li et al., 2026d). However, recent analyses identify potential reasoning degradation (Kim et al., 2026) and show that OPSD-SD can improve pass@1 while reducing rollout diversity and flattening pass@k curves (Nicolicioiu et al., 2026). Motivated by these findings, C^{2}OPSD addresses this issue by mitigating PCMI-induced sharpening and encouraging exploration of underrepresented correct paths while maintaining pass@1.

A.1.2 Diversity Collapse in On-Policy Learning

In on-policy learning, diversity collapse occurs when probability mass concentrates on a small set of reasoning paths, limiting the gains from additional sampling even as pass@1 improves (Nicolicioiu et al., 2026; Yang et al., 2026b). Prior analysis shows that OPSD-SD exhibits reverse KL-induced mode-seeking and can additionally amplify probability imbalances among equally correct responses through PCMI-induced sharpening (Nicolicioiu et al., 2026). To mitigate mode-seeking, existing methods adjust either the optimization objective or the update signal. The former modifies divergence losses and reference distributions to retain broader coverage (Li et al., 2026b; Jin et al., 2026), or rescales rewards to favor target distributions that cover multiple high-quality modes (GX-Chen et al., 2026). The latter reduces the strength of updates that lower policy entropy (Yang et al., 2026b). Other methods explicitly expand solution coverage by rewarding semantic or set-level diversity (Li et al., 2025; 2026a), combining diversity across correct trajectories with within-trajectory entropy regularization (Wan et al., 2026), or using pass@k as the reward (Chen et al., 2025).

PCMI-induced sharpening introduces an additional challenge: equally correct responses receive different reinforcement according to their compatibility with sampled demonstrations (Nicolicioiu et al., 2026). Existing methods do not explicitly correct this tilt induced by unequal expected PCMI, although several address related biases in demonstration- or feedback-conditioned teacher supervision. These approaches filter reference-induced or input-generic components (Shen et al., 2026c;b), or reverse distillation incentives (Shen et al., 2026a). Others regulate teacher influence using teacher-student disagreement (Li et al., 2026e) or verifier feedback (Yang et al., 2026a). Whereas existing methods typically focus on either mode-seeking or biases in conditioned supervision, C^{2}OPSD jointly mitigates reverse KL-induced mode-seeking and PCMI-induced sharpening by coupling PD with DAE. This combination corrects demonstration-dependent preferences among equally correct responses while encouraging exploration of underrepresented correct paths. Because this exploration may broaden correct-path coverage at the expense of pass@1, C^{2}OPSD further uses PCPO to target pass@1 non-inferiority relative to the original performance without the proposed modules.

A.2 Theoretical Analysis

We analyze PCMI Debiasing under the same frozen-base-policy setting as Nicolicioiu et al. (2026). Let π_0 denote the base policy, which is used both as the unconditional reference policy and as the demonstration-conditioned teacher. For an input x, a correct demonstration is sampled as

$$y^{\mathrm{corr}} \sim p_{\mathrm{corr}}(\cdot \mid x).$$

The PCMI between a candidate response y and a correct demonstration y^{corr} is

$$i(y; y^{\mathrm{corr}} \mid x) := \log \frac{\pi_0(y \mid x, y^{\mathrm{corr}})}{\pi_0(y \mid x)}.$$

We use

$$I_x(y) := \mathbb{E}_{y^{\mathrm{corr}} \sim p_{\mathrm{corr}}(\cdot|x)} \left[i(y; y^{\mathrm{corr}} \mid x) \right]$$

to denote the expected PCMI of y.

Let $\mathcal{C}_x$ be the set of correct responses for x, and let

$$r_x(y) = \mathbf{1}\{y \in \mathcal{C}_x\}$$

be the correctness indicator. We define the mean expected PCMI over correct responses as

$$\mu_x = \frac{1}{|\mathcal{C}_x|} \sum_{y \in \mathcal{C}_x} I_x(y).$$

Throughout the analysis, I_x, μ_x, and all correction statistics are treated as frozen quantities, consistent with the stop-gradient implementation of the PCMI advantage.

To measure the additional probability imbalance introduced by PCMI, we define PCMI-induced sharpening as follows.

Definition 2 (PCMI-induced sharpening)**.** *For two responses y_a and y_b, we define their sharpening under π relative to the base policy π_0 as*

$$\mathcal{S}_{ab}(\pi) := \log \frac{\pi(y_a \mid x)}{\pi(y_b \mid x)} - \log \frac{\pi_0(y_a \mid x)}{\pi_0(y_b \mid x)}.$$

Thus, $\mathcal{S}_{ab}(\pi)$ measures the change in the log-probability ratio beyond the imbalance already present in π_0.

A.2.1 Standard KL-Regularized RL and OPSD-SD-KL

We recall the sequence-level baseline results from Propositions 1–2 and Remark 1 of Nicolicioiu et al. (2026). For distributions P and Q over the considered response support, we use

$$D_{\mathrm{KL}}(P \| Q) := \sum_y P(y) \log \frac{P(y)}{Q(y)}$$

to denote the Kullback–Leibler (KL) divergence of P relative to Q. Under the common-support assumptions of Nicolicioiu et al. (2026), standard KL-regularized RL with the binary correctness reward r_x solves

$$\max_{\pi} \ \mathbb{E}_{y \sim \pi(\cdot|x)} \left[r_x(y) \right] - \beta_{\mathrm{RL}} D_{\mathrm{KL}} \left(\pi(\cdot \mid x) \| \pi_0(\cdot \mid x) \right), \ \beta_{\mathrm{RL}} > 0.$$

Its optimizer is

$$\pi^{\star}_{\mathrm{RL}}(y \mid x) \propto \pi_0(y \mid x) \exp\left(\frac{r_x(y)}{\beta_{\mathrm{RL}}} \right).$$

Hence, for any two correct responses $y_a, y_b \in \mathcal{C}_x$, their reward tilts are identical, and

$$\mathcal{S}_{ab}\left(\pi^{\star}_{\mathrm{RL}}\right) = 0.$$

OPSD-SD-KL instead distills from the demonstration-conditioned teacher while regularizing the student toward the unconditional base policy. Thus, the OPSD-SD-KL objective is:

$$\min_{\pi} \ \mathbb{E}_{y^{\mathrm{corr}} \sim p_{\mathrm{corr}}(\cdot|x)} \left[D_{\mathrm{KL}} \left(\pi(\cdot \mid x) \| \pi_0(\cdot \mid x, y^{\mathrm{corr}}) \right) \right] + \beta D_{\mathrm{KL}} \left(\pi(\cdot \mid x) \| \pi_0(\cdot \mid x) \right), \ \beta \geq 0.$$

As shown in Nicolicioiu et al. (2026), the optimizer is

$$\pi^{\star}_{\text{OPSD-SD-KL}}(y \mid x) \propto \pi_0(y \mid x) \exp\left(\frac{I_x(y)}{1+\beta} \right).$$

Therefore, for two correct responses $y_a, y_b \in \mathcal{C}_x$,

$$\mathcal{S}_{ab}\left(\pi^{\star}_{\text{OPSD-SD-KL}}\right) = \frac{I_x(y_a) - I_x(y_b)}{1+\beta}.$$

Unlike standard KL-regularized RL with the binary correctness reward, OPSD-SD-KL can thus alter the relative probability of equally correct responses whenever their expected PCMI values differ.

A.2.2 OPSD-SD-KL UNDER PCMI DEBIASING

Theorem 2 (Exact cancellation of PCMI-induced sharpening)**.** *Under idealized PCMI Debiasing without standardization or clipping, setting the correction coefficient to $\gamma = 1$ yields, for any two correct responses $y_a, y_b \in \mathcal{C}_x$,*

$$\mathcal{S}_{ab}(\pi_1^\star) = 0.$$

Therefore, ideal PCMI Debiasing exactly cancels the additional sharpening introduced by demonstration-conditioned self-distillation between correct modes.

Theorem 2 restores the relative probability of correct responses to the ratio under the base policy. We next derive the optimal policy underlying this result.

Let $h_x(y)$ be any frozen PCMI correction cost, and the objective of the OPSD-SD-KL under PCMI Debiasing is

$$\min_\pi \ \mathbb{E}_{y^{\mathrm{corr}}\sim p_{\mathrm{corr}}(\cdot|x)}\left[D_{\mathrm{KL}}\left(\pi(\cdot \mid x)\|\pi_0(\cdot \mid x, y^{\mathrm{corr}})\right)\right] + \beta D_{\mathrm{KL}}\left(\pi(\cdot \mid x)\|\pi_0(\cdot \mid x)\right) + \gamma \mathbb{E}_{y\sim\pi(\cdot|x)}\left[h_x(y)\right],\ \beta \geq 0. \tag{15}$$

Proposition 1 (Optimal policy for OPSD-SD-KL under PCMI Debiasing)**.** *The optimal policy of Eq. (15) is the following tilted distribution:*

$$\pi_\gamma^\star(y \mid x) \propto \pi_0(y \mid x) \exp\left(\frac{I_x(y) - \gamma h_x(y)}{1+\beta}\right).$$

Thus, PCMI Debiasing changes the score governing PCMI-induced sharpening from $I_x(y)$ to $I_x(y) - \gamma h_x(y)$.

Proof. Fix x and suppress it from the notation. Expanding the two reverse-KL terms in Eq. (15) gives

$$\min_\pi \ \mathbb{E}_{y^{\mathrm{corr}}}\left[\sum_y \pi(y)\log\frac{\pi(y)}{\pi_0(y \mid y^{\mathrm{corr}})}\right] + \beta\sum_y \pi(y)\log\frac{\pi(y)}{\pi_0(y)} + \gamma\sum_y \pi(y)h(y), \tag{16}$$

subject to

$$\sum_y \pi(y) = 1.$$

After interchanging the expectation and summation and collecting terms, the Lagrangian is

$$\begin{aligned}\mathcal{L}(\pi,\lambda) = (1+\beta)\sum_y \pi(y)\log\pi(y) &- \sum_y \pi(y)\mathbb{E}_{y^{\mathrm{corr}}}\left[\log\pi_0(y \mid y^{\mathrm{corr}})\right] \\ &- \beta\sum_y \pi(y)\log\pi_0(y) + \gamma\sum_y \pi(y)h(y) \\ &+ \lambda\left(\sum_y \pi(y) - 1\right).\end{aligned} \tag{17}$$

Differentiating with respect to $\pi(y)$ yields

$$\begin{aligned}\frac{\partial\mathcal{L}}{\partial\pi(y)} = (1+\beta)(\log\pi(y)+1) &- \mathbb{E}_{y^{\mathrm{corr}}}\left[\log\pi_0(y \mid y^{\mathrm{corr}})\right] - \beta\log\pi_0(y) \\ &+ \gamma h(y) + \lambda.\end{aligned} \tag{18}$$

Setting the derivative to zero and using

$$\mathbb{E}_{y^{\mathrm{corr}}}\left[\log \pi_0(y \mid y^{\mathrm{corr}})\right] = \log \pi_0(y) + I(y)$$

gives

$$\log \pi(y) = \log \pi_0(y) + \frac{I(y) - \gamma h(y)}{1+\beta} + c,$$

where c is independent of y.

Restoring the conditioning on x and exponentiating the preceding equality yields

$$\pi^\star_\gamma(y \mid x) \propto \pi_0(y \mid x) \exp\left(\frac{I_x(y) - \gamma h_x(y)}{1+\beta}\right).$$

The proportionality constant is uniquely determined by the normalization constraint

$$\sum_y \pi^\star_\gamma(y \mid x) = 1.$$

Finally, the Hessian with respect to the policy probabilities is

$$(1+\beta)\,\mathrm{diag}\left(\frac{1}{\pi(y)}\right),$$

which is positive definite on the interior of the probability simplex. Therefore, the exponentially tilted policy above is the unique global minimizer.

□

Proof of Theorem 2. For the ideal correction,

$$h^{\mathrm{raw}}_x(y) = r_x(y)\big(I_x(y) - \mu_x\big).$$

By Proposition 1,

$$\pi^\star_\gamma(y \mid x) \propto \pi_0(y \mid x) \exp\left(\frac{I_x(y) - \gamma r_x(y)(I_x(y) - \mu_x)}{1+\beta}\right).$$

For two correct responses $y_a, y_b \in \mathcal{C}_x$, we have $r_x(y_a) = r_x(y_b) = 1$. Their probability ratio is therefore

$$\log \frac{\pi^\star_\gamma(y_a \mid x)}{\pi^\star_\gamma(y_b \mid x)} = \log \frac{\pi_0(y_a \mid x)}{\pi_0(y_b \mid x)} + \frac{(1-\gamma)\,(I_x(y_a) - I_x(y_b))}{1+\beta}. \tag{19}$$

Consequently,

$$\mathcal{S}_{ab}(\pi^\star_\gamma) = (1-\gamma)\frac{I_x(y_a) - I_x(y_b)}{1+\beta}.$$

Setting $\gamma = 1$ gives

$$\mathcal{S}_{ab}(\pi^\star_1) = 0,$$

which proves the result. □

Eq. (19) also characterizes the correction when $\gamma \neq 1$. In particular, $0 < \gamma < 1$ attenuates the SDSD-induced sharpening, whereas $\gamma > 1$ may reverse the demonstration-induced ordering between correct responses.

Let σ_x denote the standard deviation of the expected PCMI over correct responses:

$$\sigma_x = \left(\frac{1}{|\mathcal{C}_x|}\sum_{y \in \mathcal{C}_x} \left(I_x(y) - \mu_x\right)^2\right)^{1/2}.$$

Let $\varepsilon > 0$ be a small constant introduced for numerical stability, and define the standardization scale as

$$s_x := \sigma_x + \varepsilon.$$

The standardized but unclipped PCMI correction is then

$$h^{\mathrm{std}}_x(y) = r_x(y)\frac{I_x(y) - \mu_x}{s_x}.$$

Proposition 2 (Correction strength under PCMI Debiasing with standardization)**.** *Under PCMI standardization, the correction strength for any two correct responses changes from γ to*

$$\gamma_{\mathrm{eff},x}^{\mathrm{std}} = \frac{\gamma}{\sigma_x + \varepsilon}.$$

Exact cancellation occurs when

$$\gamma = \sigma_x + \varepsilon,$$

rather than universally at $\gamma = 1$.

Proof. Substituting h_x^{std} into Proposition 1 gives

$$\pi^{\star}_{\gamma,\mathrm{std}}(y \mid x) \propto \pi_0(y \mid x) \exp\left(\frac{I_x(y) - \frac{\gamma r_x(y)}{s_x}(I_x(y) - \mu_x)}{1+\beta}\right).$$

For two correct responses, the common mean term cancels in their log-probability ratio, yielding

$$\mathcal{S}_{ab}\left(\pi^{\star}_{\gamma,\mathrm{std}}\right) = \frac{\left(1 - \frac{\gamma}{s_x}\right)\left(I_x(y_a) - I_x(y_b)\right)}{1+\beta}.$$

□

Let

$$z_x(y) = \frac{I_x(y) - \mu_x}{\sigma_x + \varepsilon}$$

and define the standardized and clipped correction

$$h_x^{\mathrm{clip}}(y) = r_x(y)\psi_\tau(z_x(y)),$$

where

$$\psi_\tau(z) = \mathrm{clip}(z, -\tau, \tau).$$

For two correct responses $y_a, y_b \in \mathcal{C}_x$, define

$$\Delta I_{ab} = I_x(y_a) - I_x(y_b)$$

and

$$\Delta\psi_{ab} = \psi_\tau(z_x(y_a)) - \psi_\tau(z_x(y_b)).$$

Proposition 3 (Correction strength under PCMI Debiasing with standardization and clipping)**.** *Under PCMI Debiasing with standardization and clipping, the sharpening between two correct responses is*

$$\mathcal{S}_{ab}\left(\pi^{\star}_{\gamma,\mathrm{clip}}\right) = \frac{\Delta I_{ab} - \gamma\Delta\psi_{ab}}{1+\beta}.$$

For $\Delta I_{ab} \neq 0$, the pair-dependent correction strength is

$$\gamma_{\mathrm{eff},ab}^{\mathrm{clip}} = \gamma\frac{\Delta\psi_{ab}}{\Delta I_{ab}},$$

such that

$$\mathcal{S}_{ab}\left(\pi^{\star}_{\gamma,\mathrm{clip}}\right) = \left(1 - \gamma_{\mathrm{eff},ab}^{\mathrm{clip}}\right)\frac{\Delta I_{ab}}{1+\beta}.$$

Moreover, if $\Delta I_{ab} > 0$, then

$$0 \leq \Delta\psi_{ab} \leq \min\left\{\frac{\Delta I_{ab}}{\sigma_x + \varepsilon}, 2\tau\right\}.$$

Consequently,

$$0 \leq \gamma_{\mathrm{eff},ab}^{\mathrm{clip}} \leq \gamma\min\left\{\frac{1}{\sigma_x + \varepsilon}, \frac{2\tau}{\Delta I_{ab}}\right\}.$$

Thus, clipping bounds the magnitude of the correction but generally prevents a single global value of γ from exactly cancelling the sharpening for all pairs of correct responses.

Proof. Proposition 1 directly gives

$$\pi^{\star}_{\gamma,\mathrm{clip}}(y \mid x) \propto \pi_0(y \mid x) \exp\left(\frac{I_x(y) - \gamma r_x(y)\psi_\tau(z_x(y))}{1+\beta}\right).$$

Taking the log-probability ratio of two correct responses produces

$$\mathcal{S}_{ab}\left(\pi^{\star}_{\gamma,\mathrm{clip}}\right) = \frac{\Delta I_{ab} - \gamma \Delta\psi_{ab}}{1+\beta}.$$

The clipping function ψ_τ is monotone and 1-Lipschitz. Because

$$z_x(y_a) - z_x(y_b) = \frac{\Delta I_{ab}}{\sigma_x + \varepsilon},$$

we have

$$0 \le \Delta\psi_{ab} \le \frac{\Delta I_{ab}}{\sigma_x + \varepsilon}.$$

Furthermore, since the range of ψ_τ is $[-\tau, \tau]$,

$$\Delta\psi_{ab} \le 2\tau.$$

Combining the two bounds proves the result. □

Remark 1 (Piecewise behavior induced by clipping)**.** *The pair-dependent correction in Proposition 3 admits three representative regimes:*

1. *If both responses lie in the linear region of the clipping function, then*

$$\gamma^{\mathrm{clip}}_{\mathrm{eff},ab} = \frac{\gamma}{\sigma_x + \varepsilon},$$

   recovering Proposition 2.

2. *If both responses saturate at the same clipping boundary, then $\Delta\psi_{ab} = 0$, and PCMI Debiasing does not alter their relative probability.*

3. *If the two responses saturate at opposite boundaries, then $\Delta\psi_{ab} = 2\tau$, giving*

$$\gamma^{\mathrm{clip}}_{\mathrm{eff},ab} = \frac{2\gamma\tau}{\Delta I_{ab}}.$$

Exact pairwise cancellation occurs if and only if

$$\gamma\Delta\psi_{ab} = \Delta I_{ab}.$$

Because this condition depends on the response pair after clipping, a single γ generally provides approximate attenuation rather than universal cancellation.

Remark 2 (Length-normalized implementation)**.** *The preceding results assume that PCMI Debiasing is constructed from the same sequence-level quantity $I_x(y)$ that determines the OPSD-SD exponential tilt. In the implementation, the correction is computed from the length-normalized score*

$$\bar{I}_x(y) = \frac{I_x(y)}{|y|}.$$

Accordingly, the exact implementation-level sharpening is

$$\mathcal{S}_{ab} = \frac{\Delta I^{\mathrm{seq}}_{ab} - \gamma\left(g_x(\bar{I}_x(y_a)) - g_x(\bar{I}_x(y_b))\right)}{1+\beta}.$$

When the two responses have equal length L and clipping is inactive, Proposition 2 holds with effective strength

$$\gamma_{\mathrm{eff},x} = \frac{\gamma}{L(\sigma_{\bar{I},x} + \varepsilon)}.$$

For unequal-length responses, the general pairwise form should be retained.

Algorithm 1 Training Algorithm of $\texttt{C}^2\texttt{OPSD}$

Require: Training set $\mathcal{D}$, initial policy parameters θ_0, training steps S, rollouts per question N, debiasing weight γ, exploration weight schedule $\{\alpha_s\}$, EMA update rate κ_{EMA}
Ensure: Trained student policy π_{θ_S}

Initialize the EMA teacher: $\bar{\theta}_0 \leftarrow \theta_0$
for $s = 0, \ldots, S-1$ **do**
// On-policy sampling and verification
  $\mathcal{X}_s \sim \mathcal{D}$, $\mathcal{Y}_s(x) \sim \pi_{\theta_s}^{\otimes N}(\cdot \mid x)$ $(x \in \mathcal{X}_s)$
  $\mathcal{B}_s \leftarrow \{(x, y) : x \in \mathcal{X}_s,\ y \in \mathcal{Y}_s(x)\}$
  $c_i \leftarrow r(x_i, y_i)$, $\mathcal{C}(x) \leftarrow \{y_i \in \mathcal{Y}_s(x) : c_i = 1\}$
// Self-distillation targets
  Select y_i^{corr} from $\mathcal{C}(x_i)$ using the baseline demonstration-selection and masking rules.
  $q_{i,t}^{\text{corr}} \leftarrow \text{sg}[\pi_{\bar{\theta}_s}(\cdot \mid x_i, y_i^{\text{corr}}, y_{i,<t})]$ (eligible i)
// PCMI Debiasing (PD)
  $\hat{I}_i \leftarrow$ Eq. (7), $\quad A_i^{\text{PD}} \leftarrow$ Eq. (8)
// Diversity-Aware Exploration (DAE)
  $A_i^{\text{DAE}} \leftarrow 0$ $(\forall i \in \mathcal{B}_s)$
  if $\alpha_s > 0$ **then**
    for $x \in \mathcal{X}_s$ **do**
      Extract rationales to obtain v_i and construct the three feature views as in Section 3.3.
      $\mathcal{C}^v(x) \leftarrow \{y_i \in \mathcal{C}(x) : v_i = 1\}$ ▷ Correct, valid rationales
      if $|\mathcal{C}^v(x)| \geq 2$ **then**
        Compute cosine similarities $S_{ij}^{(\ell)}$ and $S_{ij} \leftarrow \sum_\ell w_\ell S_{ij}^{(\ell)}$.
        $\Delta_i^{\text{DC}} \leftarrow D(\mathcal{C}^v(x)) - D(\mathcal{C}^v(x) \setminus \{y_i\})$
        Compute $\mu_{\Delta,x}$ and $\sigma_{\Delta,x}$ over $\{\Delta_i^{\text{DC}} : y_i \in \mathcal{C}^v(x)\}$.
        if $\sigma_{\Delta,x} > 0$ **then** ▷ Skip degenerate groups
          $A_i^{\text{DAE}} \leftarrow \left[(\Delta_i^{\text{DC}} - \mu_{\Delta,x})/(\sigma_{\Delta,x} + \varepsilon)\right]_+$
        end if
      end if
    end for
  end if
// Joint optimization
  $\bar{T}_{\mathcal{B}_s} \leftarrow |\mathcal{B}_s|^{-1} \sum_{i \in \mathcal{B}_s} T_i$, $\quad \ell_i(\theta) \leftarrow \sum_{t=1}^{T_i} \log \pi_\theta(y_{i,t} \mid x_i, y_{i,<t})$
  Compute $\mathcal{L}_{\text{SD}}$ using the baseline distillation objective, teacher targets, and masks.
  $\mathcal{L}_z \leftarrow -\dfrac{\sum_{i \in \mathcal{B}_s} \text{sg}[A_i^z]\ell_i(\theta)}{|\mathcal{B}_s|\bar{T}_{\mathcal{B}_s}}$ $(z \in \{\text{PD}, \text{DAE}\})$
  $\mathcal{L}_{\text{total}} \leftarrow \mathcal{L}_{\text{SD}} + \gamma\mathcal{L}_{\text{PD}} + \alpha_s\mathcal{L}_{\text{DAE}}$
  Update θ_s to θ_{s+1} by minimizing $\mathcal{L}_{\text{total}}$, with teacher targets and auxiliary advantages fixed.
// EMA teacher update after successful optimization
  $\bar{\theta}_{s+1} \leftarrow (1 - \kappa_{\text{EMA}})\bar{\theta}_s + \kappa_{\text{EMA}}\theta_{s+1}$
end for
return π_{θ_S}

A.3 Training Algorithm

Algorithm 1 summarizes the complete training procedure, from on-policy sampling and self-distillation to the joint policy update. Here θ_s denotes the student parameters after s updates, so π_{θ_S} is the final instance of the student policy π_θ.

A.4 Additional Experiments

In this subsection, we provide additional experiments to further understand the reasoning behavior of $\texttt{C}^2\texttt{OPSD}$. Specifically, we present Pass@k performance with larger k and a case study in the following subsection.

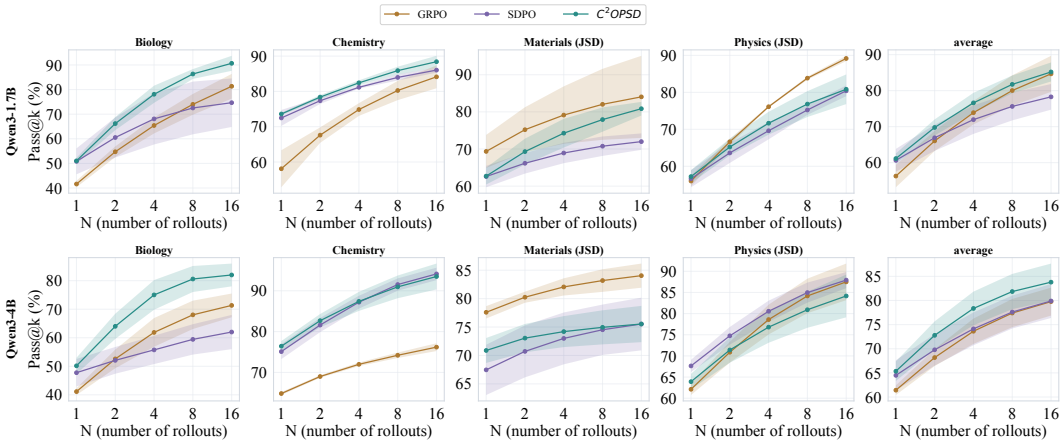


Figure 6: Pass@k performance comparison of C^{2}OPSD with baselines.

A.4.1 Pass@k Performance

Figure6 shows Pass@k performance across four datasets. Across Qwen3-1.7B and Qwen3-4B, C^{2}OPSD achieves the highest average Pass@1 and Pass@16, indicating improvements at both rollout budgets. Its advantage over SDPO is particularly pronounced in Biology and generally widens with additional rollouts, suggesting greater benefits from additional sampling. In Chemistry, C^{2}OPSD remains comparable to SDPO, with both outperforming GRPO. However, GRPO retains an advantage in Materials, and C^{2}OPSD does not consistently lead in Physics. One possible explanation is that tasks with more constrained solution spaces offer less scope for exploration to uncover additional correct solutions.

A.4.2 Case Study

To investigate the reasoning behavior of C^{2}OPSD, we sample 16 rollouts per question from SDPO and C^{2}OPSD on the Biology test set. As shown in Figure 7, C^{2}OPSD generates varied explanatory structures for the same question, whereas SDPO repeats an identical response across all 16 rollouts. Specifically, SDPO consistently states "Reaction is correct" without explaining how it obtains the answer. This behavior more closely resembles supervised fine-tuning-style answer memorization than explicit reasoning.

In contrast, the variation in C^{2}OPSD suggests broader exploration of explanatory strategies during reasoning. Specifically, Rollout 1 provides a direct justification, attributing the predicted "reaction" relation to a biological interaction between the two proteins. Rollout 13 builds on the same premise by attempting to exclude a competing answer, explicitly stating that option D is incorrect. Thus, the two rollouts differ in their explanatory structure: one offers a direct justification, while the other supplements it with option exclusion. Despite arriving at the same correct answer, these rollouts exhibit more varied explanatory structures than SDPO's identical responses.

A.5 Experiment Details

A.5.1 Hyperparameters

Table 3 lists the hyperparameters used for GRPO, SDPO, and C^{2}OPSD. For C^{2}OPSD, we select γ and α_s by searching over $\{0.1, 0.25, 0.5, 0.625, 0.75, 1.0\}$ and $\{0.05, 0.1, 0.2, 0.5, 1\}$, respectively.

A.5.2 Settings for Further Probe

For **PCMI Distribution**, we used Qwen3-1.7B as the initial model and compared it with SDPO from the comparison study and C^{2}OPSD with PD only ($\gamma = 0.625$), both trained from the initial model using the same seed. We randomly sampled 50 questions from the Biology test set. For each

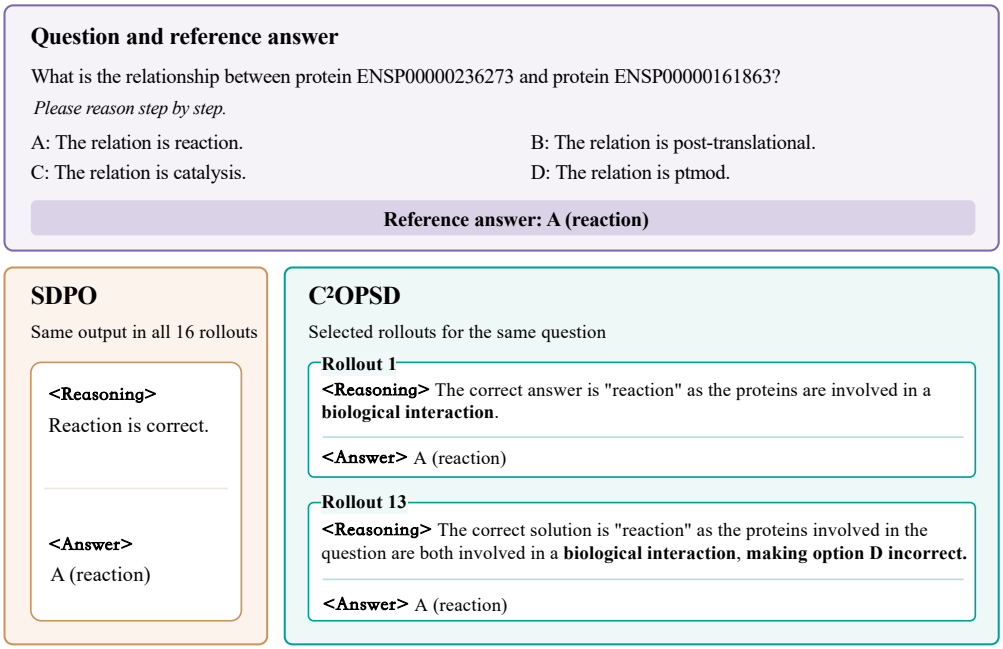


Figure 7: Output examples of SDPO and C^{2}OPSD on a Biology question with 16 rollouts per method. SDPO repeats the same response, whereas C^{2}OPSD exhibits varied explanations through direct justification and option exclusion.

question, the initial model generated 16 demonstration candidates, from which we selected the first verifier-accepted response. We retained the 44 questions with a correct demonstration and independently generated 16 evaluation responses per question for each model, yielding 704 responses per model. All responses were scored using the frozen initial model, fixed demonstrations, and a scoring temperature of 1.0. PCMI was computed as the token-averaged log-probability difference between demonstration-conditioned and original contexts, without group-wise standardization or clipping. We plotted density-normalized histograms for all valid responses and correct valid responses using shared bin edges. Statistics were aggregated at the response level and the correct-response subset was not reweighted to give each question equal weight.

For **Diversity Distribution**, we used Qwen3-1.7B as the initial model and compared it with SDPO from the comparison study and C^{2}OPSD with DAE only ($\alpha_s = 0.5$), both trained from the initial model using the same seed. We sampled 50 questions from the Biology test set and generated 16 rollouts per question for each model. For visualization, we use 20 shared bins for each diversity metric and normalize each model distribution to 100%.

Diversity was computed using only the reasoning portion of each rollout y_i. We denote the extracted reasoning text by

$$\tilde{y}_i = g(y_i),\ \widetilde{\mathcal{Y}}(x) = \{\tilde{y}_i\}_{i=1}^{N},$$

where $g(\cdot)$ extracts the text enclosed by the <reasoning> tags. When these tags are absent, $g(\cdot)$ uses the text preceding the <answer> block. The final answer and formatting tags are excluded. We retain all valid rollouts in $\mathcal{Y}(x)$ without correctness filtering or deduplication. All 50 questions are valid across the three model groups and are included in the comparison. We use the following three diversity metrics.

1-Self**-**BLEU**.** We measure lexical diversity using one minus the average leave-one-out Self-BLEU (Zhu et al., 2018):

$$D_{\mathrm{SB}}(x) = 1 - \frac{1}{N}\sum_{i=1}^{N} \mathrm{BLEU}_4\left(\tilde{y}_i, \widetilde{\mathcal{Y}}(x) \setminus \{\tilde{y}_i\}\right),$$

Table 3: Hyperparameters used for GRPO, SDPO, and C^{2}OPSD. A dash denotes a parameter that is not applicable or an operation that is not used.

Parameters	GRPO	SDPO	C^{2}OPSD
General			
Model	Qwen3-1.7B; Qwen3-4B	Qwen3-1.7B; Qwen3-4B	Qwen3-1.7B; Qwen3-4B
Thinking	False	False	False
Data			
Max. prompt length	2048	2048	2048
Max. response length	8192	8192	8192
Batching			
Question batch size	8	8	8
Mini batch size (questions)	8	8	8
Number of rollouts	8	8	8
Rollout			
Inference engine	vLLM	vLLM	vLLM
Temperature	1.0	1.0	1.0
Rollout importance sampling clip	2	2	2
Validation			
Number of rollouts	16	16	16
Temperature	0.6	0.6	0.6
Top-p	0.95	0.95	0.95
Self-distillation Loss			
Top-K distillation	-	100	100
Distillation divergence	-	Generalized JSD	Generalized JSD
Clip advantages	-	-	-
Teacher-EMA update rate	-	0.05 (Biology); 0.01 (Physics, Chemistry, Materials)	0.05 (Biology); 0.01 (Physics, Chemistry, Materials)
PCMI Debiasing (PD)			
PD advantage clip	-	-	2
Numerical stabilizer ε	-	-	1×10^{-6}
Diversity-Aware Exploration (DAE)			
DCScore temperature τ_{DC}	-	-	1.0
DAE warmup steps W	-	-	10
DAE ramp steps R	-	-	20
View weights (semantic, concept, operation)	-	-	$(0.60, 0.25, 0.15)$
Semantic encoder	-	-	all-MiniLM-L6-v2
Max. encoder input length	-	-	256
Encoder batch size	-	-	16
Max. reasoning units	-	-	32
Numerical stabilizer ε	-	-	1×10^{-6}
Training			
Optimizer	AdamW	AdamW	AdamW
Learning rate	1×10^{-5}	5×10^{-6}	5×10^{-6}
Learning-rate schedule	Constant	Constant	Constant
Warmup steps	10	10	10
Weight decay	0.01	0.01	0.01
Gradient clip norm	1.0	1.0	1.0

where each $\tilde{y}_i$ is treated as the hypothesis and the remaining reasoning texts as references. We use equally weighted one- to four-gram precision and NLTK smoothing method 1. A higher $D_{\mathrm{SB}}(x)$ indicates greater lexical variation.

Cosine Distance. Let $z_i = E(\tilde{y}_i)$ denote the representation produced by the frozen `all-MiniLM-L6-v2` encoder. We measure semantic diversity using the mean pairwise cosine distance:

$$D_{\cos}(x) = \frac{2}{N(N-1)} \sum_{i<j} \left(1 - \frac{z_i^\top z_j}{\|z_i\|_2 \|z_j\|_2}\right).$$

A higher $D_{\cos}(x)$ indicates that the rollouts are less similar in semantic space.

DCScore. DCScore jointly measures diversity from semantic, conceptual, and operational views. Let $z_i^{(\ell)}$ denote the feature representation of rollout i under view $\ell \in \{\text{semantic}, \text{concept}, \text{operation}\}$. We compute the view-specific cosine similarity as

$$S_{ij}^{(\ell)} = \frac{{z_i^{(\ell)}}^\top z_j^{(\ell)}}{\|z_i^{(\ell)}\|_2 \|z_j^{(\ell)}\|_2},$$

and construct the fused similarity matrix as

$$S_{ij} = \sum_{\ell \in \{\text{semantic},\text{concept},\text{operation}\}} w_\ell S_{ij}^{(\ell)},$$

where $(w_{\text{semantic}}, w_{\text{concept}}, w_{\text{operation}}) = (0.60, 0.25, 0.15)$.

The set-level DCScore is then computed as

$$D_{DC}(x) = \sum_{i=1}^{N} \frac{\exp\left(S_{ii}/\tau_{\text{DC}}\right)}{\sum_{j=1}^{N} \exp\left(S_{ij}/\tau_{\text{DC}}\right)}, \ \tau_{\text{DC}} = 1.$$

A higher $D_{\text{DC}}(x)$ indicates that individual rollouts are more distinguishable from the remaining members of the set across the three views.

For the **case study**, we use the SDPO and `C`^{2}`OPSD` models from the comparison study, both based on Qwen3-1.7B and trained on the Biology dataset, to generate rollouts for questions from the Biology test set.