SR-OPSD: SELF-REFERENCED ON-POLICY SELF-DISTILLATION

Anonymous authors
Paper under double-blind review

ABSTRACT

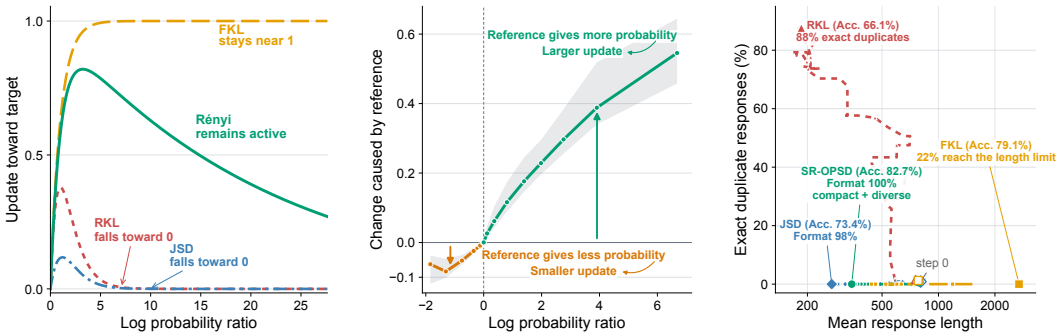
On-policy self-distillation (OPSD) converts feedback into dense token-level supervision on student-generated trajectories, complementing reinforcement learning with sparse outcome rewards. Its self-teacher, derived from the student's current or exponentially averaged parameters and conditioned on additional context, evolves alongside the student and its rollout context distribution. The benefit of modifying this moving target depends on how target–student probability mismatches translate into updates. We propose *Self-Referenced On-Policy Self-Distillation (SR-OPSD)*, which constructs a normalized geometric target from the self-teacher and a frozen initial policy, then minimizes the forward Rényi divergence from this target to the student. The interpolation coefficient controls the self-teacher's contribution, while the Rényi order controls the power weighting of target-to-student probability ratios in the gradient. For fixed contexts and target components, we establish a conditional variational characterization and derive the exact token-logit gradient, revealing how anchoring and projection jointly shape the effective update target. Experiments across scientific reasoning, tool use, mathematical reasoning, and code generation demonstrate strong performance across multiple model families and scales. Ablations further show that reference anchoring can improve or degrade performance depending on the projection objective, supporting the joint design of target construction and projection geometry.

1 INTRODUCTION

Post-training of large language models (LLMs) encompasses supervised instruction tuning, preference optimization, and reinforcement learning (Wei et al., 2022; Sanh et al., 2022; Ouyang et al., 2022; Rafailov et al., 2023; He et al., 2026; Wu et al., 2026). Among reinforcement-learning methods, RLHF methods commonly optimize a learned reward model using proximal policy optimization (PPO) (Schulman et al., 2017; Ouyang et al., 2022), whereas recent reasoning systems increasingly employ reinforcement learning with verifiable rewards (RLVR), using algorithms such as Group Relative Policy Optimization (GRPO) (Shao et al., 2024) and its variants (Yu et al., 2025). Methods such as Direct Preference Optimization (DPO) (Rafailov et al., 2023) optimize preference data without explicit reward modeling. In terms of reasoning, RLVR is particularly attractive because the correctness of the outcome can often be assessed automatically through external checks, unit tests, or rule-based verifiers. However, these methods assign a scalar-valued reward to an entire response. Although this reward induces policy-gradient updates over the generated tokens, it does not directly identify which intermediate tokens or reasoning steps caused success or failure, resulting in coarse and sparse credit assignment over long sequences.

On-policy distillation (OPD) offers an alternative approach by utilizing the dense token-level supervision on trajectories sampled from the policy to be optimized (also known as the student policy) (Gu et al., 2023; Agarwal et al., 2024; Zhu et al., 2026; Jin et al., 2026). A teacher policy evaluates student-generated prefixes, reducing the state-distribution mismatch associated with offline distillation. Existing on-policy distillation methods can be viewed through three design choices including: the divergence objective, the source of supervision, and the mechanism used to stabilize policy optimization. These choices are often designed largely in isolation, which is relatively benign in conventional offline distillation where both the teacher and the training policy are fixed. This benefit, however, generally requires continued access to an extra teacher policy. On-policy self-distillation (OPSD) instead uses two views of the same evolving model: the policy we aim to optimize is

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.



(a) Divergence-dependent token correction.

(b) Reference-induced change in correction.

(c) Observed training trajectories.

Figure 1: Motivation and empirical behavior of SR-OPSD. π_θ, $\pi_{\bar{\theta}}$, and π_{ref} denote the student, self-teacher, and reference policies. **(a)** Binary token-logit correction versus $\log(\pi_{\bar{\theta}}(a)/\pi_\theta(a))$, with $\pi_{\bar{\theta}}(a) = 0.1$. **(b)** Change in token-logit correction induced by reference anchoring, $\pi^\star_\alpha \propto \pi^\alpha_{\bar{\theta}} \pi^{1-\alpha}_{\text{ref}}$, versus $\log(\pi_{\text{ref}}(a)/\pi_{\bar{\theta}}(a))$. The "update" here is the local correction applied to a token logit returned by π_θ. Points and shading indicate bin medians and interquartile ranges. **(c)** Training trajectories on SciKnowEval Physics, showing mean response length and exact-response duplication. SR-OPSD combines concise responses and low duplication with the highest final accuracy and full format validity.

conditioned on the original prompt, whereas the so-called self-teacher policy is the same policy conditioned on feedback, privileged information, or richer contextual signals (Zhao et al., 2026a; Hübotter et al., 2026). OPSD can therefore convert feedback, reflections, verified traces, or successful responses into dense token-level supervision.

Motivation We consider on-policy self-distillation with an evolving feedback-conditioned self-teacher. The policy to be optimized π_θ generates its own training contexts, while the supervisory target (also known as the self-teacher policy) $\pi_{\bar{\theta}}$ is updated from it. Both the supervisory target and the rollout context distribution therefore change throughout training. At each visited prefix, target construction determines the probability discrepancies presented to the student, and the divergence determines how these discrepancies translate into updates. This raises a central design question:

How should the supervisory target and the student's response to it be controlled jointly?

The distinction between forward KL (FKL) (Zhao et al., 2026a) and reverse KL (RKL) (Hübotter et al., 2026) illustrates the role of the projection objective. In the binary illustration in Figure 1a, the target assigns a fixed positive probability to a token. As the student's probability of that token approaches zero, the FKL logit correction approaches a nonzero plateau, whereas the RKL and JSD corrections vanish. Forward Rényi divergence provides an order-dependent response: for $0 < \rho < 1$, its correction decays as a power of the student probability, more slowly than under RKL or JSD in this regime. The order ρ thus controls the response to tokens that the student strongly underweights relative to its target. Reference-policy anchoring provides a complementary control through target construction (Yang et al., 2026). Incorporating a reference policy changes the target probabilities and can alter both the direction and magnitude of the resulting correction, as illustrated in Figure 1b. Because reference anchoring changes the probability ratios presented to the student policy, its effect depends on how the divergence weights those ratios. The runs in Figure 1c exhibit distinct patterns: RKL produces more duplicate responses, FKL generates longer responses, and SR-OPSD combines higher accuracy with concise outputs and low duplication.

We therefore propose *Self-Referenced On-Policy Self-Distillation (SR-OPSD)*, which anchors the supervisory target to a frozen copy of the student's own initial policy and matches this target through forward Rényi projection. Specifically, using the frozen initial policy as a reference policy π_{ref}, SR-OPSD constructs the normalized geometric target $\pi^\star_\alpha \propto \pi^\alpha_{\bar{\theta}} \pi^{1-\alpha}_{\text{ref}}$ and minimizes $D_\rho(\pi^\star_\alpha \| \pi_\theta)$ on student-generated prefixes. The coefficient $\alpha \in [0, 1]$ controls *where* the target lies along the geometric interpolation between the self-teacher and reference policies, while $\rho \in (0, 1)$ controls

how target-to-student probability ratios are weighted in the logit update. Together, they control how reference information and evolving self-teacher supervision shape student learning.

Main Contributions Our contributions are threefold. (i) We propose SR-OPSD, which jointly controls target construction and policy updates by coupling a normalized geometric target, formed from the self-teacher and the frozen initial student policy, with forward Rényi projection. (ii) At fixed student-generated contexts, we establish a conditional variational characterization and derive the exact token-logit gradient. The gradient identifies the effective update target as a normalized geometric mixture of the self-teacher, reference, and current student, making the roles of α and ρ in the update explicit. (iii) We demonstrate competitive performance across scientific reasoning, tool use, mathematical reasoning, and code generation, covering different feedback sources, model families, and scales.

2 Background

Notation We denote the policy to be optimized by π_θ, the self-teacher policy by $\pi_{\bar{\theta}}$, and the reference policy by π_{ref}. Given a prompt $x \sim \mathcal{D}$ and an on-policy response $y = (y_1, \ldots, y_L)$, the student prefix at position t is $h_t = (x, y_{<t})$. We write h_t^{sf} for the corresponding self-feedback or privileged context supplied to the self-teacher and h_t^{ref} for the context supplied to the reference policy. When a statement is token-level and the contexts are fixed, we suppress the position index and write h, h_{sf}, and h_{ref}. Thus, h, h_{sf}, and h_{ref} always denote three possibly different fixed contexts. In this work, $h_t^{\text{sf}} = h_t^{\text{ref}} = (x^+, y_{<t})$, where x^+ augments x with successful examples or feedback. All token-level distributions are defined over the same next-token vocabulary $\mathcal{V}$. For the theoretical derivations, all compared distributions are normalized on a common support. The operator $\text{sg}[\cdot]$ denotes stop-gradient.

On-Policy Self-Distillation Given a self-teacher policy $\pi_{\bar{\theta}}$, on-policy self-distillation trains the student on its own sampled prefixes, for example through the reverse-KL loss

$$\mathbb{E}_{\substack{x\sim\mathcal{D} \\ y\sim\pi_{\theta_{\text{old}}}(\cdot|x)}} \left[\sum_{t=1}^{L} D_{\text{KL}}(\pi_\theta(\cdot \mid h_t) \,\|\, \text{sg}[\pi_{\bar{\theta}}]) \right].$$

Here $\pi_{\theta_{\text{old}}}$ denotes the rollout policy used to collect the current on-policy batch; it coincides with the current student before the optimization step. Such a self-teacher policy is often an exponential-moving-average of π_θ with additional context information. The additional context h_t^{sf} can contain textual feedback, verified solutions, reflections, or successful sibling responses that are unavailable in the ordinary student context h_t. This construction also supports forward KL and JSD, evaluated between conditional token distributions at the same visited prefixes.

Self-Teacher Policy and Its Updates The self-teacher policy $\pi_{\bar{\theta}}$ may be a stopped copy, a lagged snapshot, or an exponential-moving-average version of the policy to be optimized π_θ. In this work, we follow (Hübotter et al., 2026) and use $\beta \in [0, 1)$ as the exponential moving average coefficient and adopt the convention $\bar{\theta}_{k+1} = \beta\bar{\theta}_k + (1-\beta)\theta_{k+1}$. Thus, larger β produces a slower-moving self-teacher policy. During each update, $\bar{\theta}_k$ and the resulting self-teacher policy are held fixed.

Rényi divergence For positive probability distributions p and q, the Rényi divergence of order $0 < \rho < 1$ is $D_\rho(p\|q) = \frac{1}{\rho-1} \log \sum_{a\in\mathcal{V}} p(a)^\rho q(a)^{1-\rho}$. We focus on this order range and define $D_1(p\|q) = D_{\text{KL}}(p\|q)$ by continuity (van Erven & Harremoës, 2014). When p is the target and q the student, we refer to $D_\rho(p\|q)$ as the forward projection. The two argument orders satisfy $D_\rho(p\|q) = \frac{\rho}{1-\rho} D_{1-\rho}(q\|p)$. This identity relates the projection direction to the divergence order.

3 Method

3.1 Revisiting On-Policy Self-Distillation: Reference-Regularized On-Policy Self-Distillation

We now consider the following token-level formulation of the reverse-KL-based on-policy self-distillation objective at fixed contexts:

$$\max_\theta \; \mathbb{E}_{\substack{x\sim\mathcal{D} \\ y\sim\pi_{\theta_{\text{old}}}(\cdot|x)}} \left[\sum_{t=1}^{L} -D_{\text{KL}}\big(\pi_\theta(\cdot \mid h_t) \,\big\|\, \text{sg}\big[\pi_{\bar{\theta}}(\cdot \mid h_t^{\text{sf}})\big]\big)\right].$$

At each token position, let (h, h_{sf}) denote the student, self-teacher contexts. During the inner update, these contexts and the self-teacher policy are held fixed. Under regular conditions, the following objective

$$\mathcal{J}_\alpha(\theta; h, h_{\text{sf}}, h_{\text{ref}}) := \mathbb{E}_{a\sim\pi_\theta(\cdot|h)}\left[\alpha \log \frac{\text{sg}[\pi_{\bar{\theta}}(a|h_{\text{sf}})]}{\pi_{\text{ref}}(a|h_{\text{ref}})}\right] - D_{\text{KL}}(\pi_\theta(\cdot \mid h) \,\|\, \pi_{\text{ref}}(\cdot \mid h_{\text{ref}})) \tag{1}$$

satisfies $\mathcal{J}_\alpha(\theta; h, h_{\text{sf}}, h_{\text{ref}}) = -D_{\text{KL}}(\pi_\theta(\cdot \mid h) \,\|\, \pi^\star_\alpha(\cdot \mid h_{\text{sf}}, h_{\text{ref}})) + \log Z_\alpha(h_{\text{sf}}, h_{\text{ref}})$ where $\pi^\star_\alpha(a \mid h_{\text{sf}}, h_{\text{ref}}) \propto \text{sg}[\pi_{\bar{\theta}}(a \mid h_{\text{sf}})]^\alpha \, \pi_{\text{ref}}(a \mid h_{\text{ref}})^{1-\alpha}$. See Appendix B.1 for details of our derivation. Consequently, for fixed contexts and frozen target components, maximizing $\mathcal{J}_\alpha$ with respect to θ is equivalent to minimizing the reverse KL from the student to $\pi^\star_\alpha$. When $\alpha = 1$, $Z_1 = 1$ and $\pi^\star_1(\cdot \mid h_{\text{sf}}, h_{\text{ref}}) = \text{sg}[\pi_{\bar{\theta}}(\cdot \mid h_{\text{sf}})]$, recovering the token-level reverse-KL-based on-policy self-distillation objective. For $0 < \alpha < 1$, this formulation incorporates reference regularization into on-policy self-distillation through the supervisory target (Hübotter et al., 2026; Yang et al., 2026). The geometric target $\pi^\star_\alpha$ interpolates between the self-teacher and reference policies.

3.2 Self-Referenced On-policy Self-distillation with Rényi Projection on Student Rollouts

Figures 1a and 1b illustrate how projection choice and reference anchoring shape the token-logit correction. We use the Rényi family to parameterize the projection, first expressing the objectives against the self-teacher and then present how the forward objective being connected to the geometric target $\pi^\star_\alpha$.

During the inner update, the sampled contexts and the teacher and reference policies are held fixed. For $0 < \rho < 1$ and strictly positive token probabilities, the reverse Rényi projection maximizes the following objective with respect to θ:

$$-\textstyle\sum_t D_\rho\big(\pi_\theta(\cdot \mid h_t) \,\big\|\, \text{sg}\big[\pi_{\bar{\theta}}(\cdot \mid h_t^{\text{sf}})\big]\big) = \sum_t \frac{1}{1-\rho} \log \mathbb{E}_{a\sim\pi_\theta(\cdot|h_t)}\big[\big(\tfrac{\text{sg}[\pi_{\bar{\theta}}(a|h_t^{\text{sf}})]}{\pi_\theta(a|h_t)}\big)^{1-\rho}\big].$$

The forward Rényi projection maximizes the corresponding objective

$$-\textstyle\sum_t D_\rho\big(\text{sg}\big[\pi_{\bar{\theta}}(\cdot \mid h_t^{\text{sf}})\big] \,\big\|\, \pi_\theta(\cdot \mid h_t)\big) = \sum_t \frac{1}{1-\rho} \log \mathbb{E}_{a\sim\pi_\theta(\cdot|h_t)}\big[\big(\tfrac{\text{sg}[\pi_{\bar{\theta}}(a|h_t^{\text{sf}})]}{\pi_\theta(a|h_t)}\big)^{\rho}\big]. \tag{2}$$

Both objectives admit a student-expectation representation, with respective powers $1-\rho$ and ρ applied to the teacher-to-student probability ratio. We adopt the forward parameterization, in which ρ directly controls this power.

SR-OPSD applies forward Rényi projection to the normalized geometric target $\pi^\star_\alpha$, combining reference anchoring through α with projection control through ρ. To connect this projection to the reference-regularized formulation above, we construct a Rényi-aggregated token functional from the same teacher-to-reference log-ratio. The following proposition shows that combining this functional with a Rényi penalty relative to π_{ref} yields an objective equivalent to minimizing $D_\rho(\pi^\star_\alpha \| \pi_\theta)$ at fixed contexts.

Proposition 3.1 (Token-Level Variational Characterization of SR-OPSD). *Fix h, h_{sf}, and h_{ref}. Let $\alpha \in [0, 1]$ and $\rho > 0$, $\rho \neq 1$, and assume that all token probabilities are strictly positive. Define the token-level pseudo-reward $r_\alpha(a; h_{\text{sf}}, h_{\text{ref}}) := \alpha \log \frac{\text{sg}[\pi_{\bar{\theta}}(a|h_{\text{sf}})]}{\pi_{\text{ref}}(a|h_{\text{ref}})}$. Define the Rényi-aggregated token functional $\mathcal{R}_{\alpha,\rho}(\theta; h, h_{\text{sf}}, h_{\text{ref}}) := -\frac{1}{\rho-1} \log \frac{\sum_{a\in\mathcal{V}} \pi_{\text{ref}}(a|h_{\text{ref}})^\rho \pi_\theta(a|h)^{1-\rho} \exp(\rho r_\alpha(a; h_{\text{sf}}, h_{\text{ref}}))}{\sum_{a\in\mathcal{V}} \pi_\theta(a|h)^\rho \pi_{\text{ref}}(a|h_{\text{ref}})^{1-\rho}}$. Let $\mathcal{J}_{\alpha,\rho}(\theta; h, h_{\text{sf}}, h_{\text{ref}}) := \mathcal{R}_{\alpha,\rho}(\theta; h, h_{\text{sf}}, h_{\text{ref}}) - D_\rho(\pi_\theta(\cdot \mid h) \,\|\, \pi_{\text{ref}}(\cdot \mid h_{\text{ref}}))$. Then*

$$\mathcal{J}_{\alpha,\rho}(\theta; h, h_{\text{sf}}, h_{\text{ref}}) = -D_\rho(\pi^\star_\alpha(\cdot \mid h_{\text{sf}}, h_{\text{ref}}) \,\|\, \pi_\theta(\cdot \mid h)) - \frac{\rho}{\rho-1} \log Z_\alpha(h_{\text{sf}}, h_{\text{ref}}). \tag{3}$$

Therefore, for fixed contexts and frozen target components, maximizing $\mathcal{J}_{\alpha,\rho}$ with respect to θ is equivalent to minimizing the forward Rényi projection from $\pi^\star_\alpha(\cdot \mid h_{\text{sf}}, h_{\text{ref}})$ to $\pi_\theta(\cdot \mid h)$.

See Appendix B.2 for the proof. In this variational formulation, $\mathcal{R}_{\alpha,\rho}$ depends on the current student policy. When $\alpha = 1$, we have $Z_1 = 1$, and the objective reduces to on-policy self-distillation with forward Rényi divergence, with the self-teacher $\pi_{\bar{\theta}}$ as the sole target.

Algorithm 1 SR-OPSD

Require: Initial policy π_{θ_0}
Initialize the self-teacher: $\bar{\theta}_0 \leftarrow \theta_0$
Set the frozen reference: $\pi_{\text{ref}} \leftarrow \text{sg}[\pi_{\theta_0}]$
for $k = 0, 1, \ldots, K-1$ **do**
  Sample prompts $x \sim \mathcal{D}$ and responses $y \sim \pi_{\theta_k}(\cdot \mid x)$
  for $t = 1, \ldots, |y|$ **do**
    Construct $h_t = (x, y_{<t})$ and $h_t^{\text{sf}} = h_t^{\text{ref}} = (x^+, y_{<t})$.
    Form the frozen target $\pi^\star_{\alpha,k}(\cdot \mid h_t^{\text{sf}}, h_t^{\text{ref}})$
  end for
  Update the policy
  $\theta_{k+1} \leftarrow \text{Update}\left(\theta_k,\ -\eta \nabla_\theta \sum_t D_\rho\left(\pi^\star_{\alpha,k}(\cdot \mid h_t^{\text{sf}}, h_t^{\text{ref}}) \,\middle\|\, \pi_\theta(\cdot \mid h_t)\right)\Big|_{\theta=\theta_k}\right)$
  Update the self-teacher: $\bar{\theta}_{k+1} \leftarrow \beta\bar{\theta}_k + (1-\beta)\theta_{k+1}$
end for
return π_{θ_K}

The proposed method, SR-OPSD, utilizes this formulation at each student-generated position t, using the geometric target $\pi^\star_\alpha(a \mid h_t^{\text{sf}}, h_t^{\text{ref}}) \propto \text{sg}[\pi_{\bar{\theta}}(a \mid h_t^{\text{sf}})]^\alpha \pi_{\text{ref}}(a \mid h_t^{\text{ref}})^{1-\alpha}$. We use a frozen copy of the initial policy as the reference policy, $\pi_{\text{ref}} := \text{sg}[\pi_{\theta_0}]$, where θ_0 denotes the initial parameters. Since the reference comes from the student itself, we refer to this anchoring scheme as *self-referenced*. The corresponding token-level objective is:

$$\max_\theta \mathbb{E}_{\substack{x\sim\mathcal{D} \\ y\sim\pi_{\theta_{\text{old}}}(\cdot|x)}} \sum_{t=1}^{L} \frac{1}{1-\rho}\Big[\log \sum_{a\in\mathcal{V}} \exp\big((1-\rho)\log \pi_\theta(a \mid h_t) + \rho\alpha \log \text{sg}\big[\pi_{\bar{\theta}}(a \mid h_t^{\text{sf}})\big] + \rho(1-\alpha)\log \pi_{\text{ref}}(a \mid h_t^{\text{ref}})\big)\Big]. \tag{4}$$

See Algorithm 1 for the pseudo-code, where $\text{Update}(\cdot)$ denotes an optimizer such as stochastic gradient descent or Adam. The algorithm states the normalized, unclipped objective. See Appendix D.2 for more details. The coefficient α determines the self-referenced target: $\alpha = 1$ recovers the self-teacher, whereas $\alpha = 0$ recovers the reference policy. The order ρ controls the projection toward this target.

3.3 How Anchoring and Projection Determine the Update

To characterize how anchoring and projection jointly shape the student update, we derive the exact token-logit gradient of the SR-OPSD objective. The following proposition expresses this gradient as the difference between the student distribution and an effective target that combines $\pi^\star_\alpha$ with the current student.

Proposition 3.2 (Logit Gradient of SR-OPSD). *Fix h, h_{sf}, and h_{ref}, and treat $\pi^\star_\alpha(\cdot \mid h_{\text{sf}}, h_{\text{ref}})$ as frozen during the student update. Let $z_\theta(a \mid h)$ denote the student logit such that $\pi_\theta(a \mid h) = \frac{\exp(z_\theta(a|h))}{\sum_{b\in\mathcal{V}}\exp(z_\theta(b|h))}$. For $\rho > 0$, $\rho \neq 1$, define $\widetilde{\pi}_{\alpha,\rho,\theta}(a \mid h, h_{\text{sf}}, h_{\text{ref}}) \propto \pi_\theta(a \mid h)^{1-\rho}\, \pi^\star_\alpha(a \mid h_{\text{sf}}, h_{\text{ref}})^\rho$. Then $\frac{\partial}{\partial z_\theta(a|h)} D_\rho(\pi^\star_\alpha(\cdot \mid h_{\text{sf}}, h_{\text{ref}}) \,\|\, \pi_\theta(\cdot \mid h)) = \pi_\theta(a \mid h) - \widetilde{\pi}_{\alpha,\rho,\theta}(a \mid h, h_{\text{sf}}, h_{\text{ref}})$.*

For $0 < \rho < 1$, the effective target $\widetilde{\pi}_{\alpha,\rho,\theta}$ is a normalized geometric mixture of the self-teacher, reference, and current student policies, with exponents $\alpha\rho$, $(1-\alpha)\rho$, and $1-\rho$, respectively. Thus, α sets the balance between self-teacher supervision and reference anchoring, while ρ controls their combined weight relative to the current student. Together, these parameters determine the effective target toward which the token-logit update moves the student. This expression links the two controls in Figures 1a and 1b: reference anchoring changes the target-to-student probability ratios, and the Rényi order determines how these ratios enter the token-logit correction.

4 Experiments

We evaluate the proposed method across various representative scenarios that capture different feedback regimes and levels of supervision, including: (i) Science Q&A and tool use benchmarks without rich environment feedback; (ii) mathematical reasoning on challenging problems with access to ground-truth solutions; and (iii) scaling behavior on coding tasks with rich execution-based feedback; and (iv) ablation study on the effectiveness of each component of the proposed method.

4.1 Performance with Bootstrapping on Generated Responses as Feedback

Setting We compare our method against SDPO (Hübotter et al., 2026), on-policy GRPO, which samples trajectories from the current policy, and GRPO, which reweights trajectories from previous policies using importance sampling. For all baselines, we follow the settings in (Hübotter et al., 2026). We evaluate on two sets of tasks: (1) scientific reasoning on undergraduate-level problems in chemistry, physics, biology, and materials science, using the L3 reasoning subset of SciKnowEval (Feng et al., 2024); and (2) tool use, i.e., mapping a tool-API specification and user request to the correct tool call, using ToolAlpaca (Tang et al., 2023). We split the data into training and test sets to evaluate in-domain generalization. We initialize the policies from Qwen3-8B (Yang et al., 2025) and OLMo-3-7B-Instruct (Olmo et al., 2025), and report Avg@16 in Table 1 in terms of wall-clock training time. We also report Maj@16 accuracy and Best@16 accuracy in Figure 5 and Figure 6. All experiments are conducted using the `verl` framework on NVIDIA H200 GPUs to ensure fair comparisons with SDPO (Hübotter et al., 2026). See Table 6 for detailed experimental configurations.

Table 1: Performance on SciKnowEval and Tool Use benchmarks. We report Avg@16 accuracy. The best and second-best results are shown in bold and underlined, respectively.

	Chemistry			Physics			Biology			Materials			Tool use		
	5h	10h	15h	5h	10h	15h	5h	10h	15h	5h	10h	15h	5h	10h	15h
Qwen3-8B		41.2			59.2			30.8			58.9			57.5	
+ GRPO	63.1	68.3	72.6	62.7	72.0	74.1	35.2	**52.5**	53.5	74.3	75.9	79.3	62.6	**66.3**	66.3
+ GRPO (on-policy)	51.8	62.9	68.1	60.8	62.2	62.6	33.8	35.5	36.4	70.4	73.1	74.8	60.1	62.5	65.5
+ SDPO	**80.5**	80.5	80.5	71.1	73.4	73.4	**50.1**	50.1	50.1	76.7	78.3	78.3	**63.8**	63.8	63.8
+ SR-OPSD (ours)	80.2	**82.8**	**82.8**	**82.7**	**82.7**	**82.7**	50.0	50.0	**54.0**	**77.7**	**79.6**	**80.9**	63.6	64.4	**67.5**
Olmo3-7B-Instruct		22.8			37.7			16.2			36.7			39.3	
+ GRPO	39.1	45.4	53.4	58.9	61.4	62.8	33.1	38.5	42.6	71.4	73.3	74.2	56.9	61.0	61.0
+ GRPO (on-policy)	37.5	40.3	44.4	54.3	59.1	60.9	29.8	31.6	34.9	66.4	70.7	73.3	55.3	57.0	59.6
+ SDPO	**79.6**	79.6	79.6	67.6	70.8	70.8	46.7	46.7	48.4	**77.4**	**77.4**	**77.4**	**58.8**	60.4	60.4
+ SR-OPSD (ours)	78.7	**79.9**	**79.9**	**68.1**	**72.3**	**74.8**	**50.1**	**51.1**	**51.1**	76.2	76.2	76.2	58.1	**61.4**	**62.7**

Results Analysis As shown in Table 1, SR-OPSD achieves the highest 15-hour Avg@16 accuracy in nine of the ten model–task settings and outperforms both GRPO variants in all ten. For Qwen3-8B, it leads across all five domains, achieving an average of 73.58, compared with 69.22 for SDPO and 69.16 for GRPO. For OLMo3-7B-Instruct, it leads in four domains, with an average of 68.94, exceeding SDPO (67.32) and GRPO (58.80). The only exception is OLMo3-7B-Instruct on Materials, where SR-OPSD trails SDPO by 1.2 percentage points. Across the reported training budgets, the number of settings in which SR-OPSD ranks first increases from four at 5 hours to seven at 10 hours and nine at 15 hours.

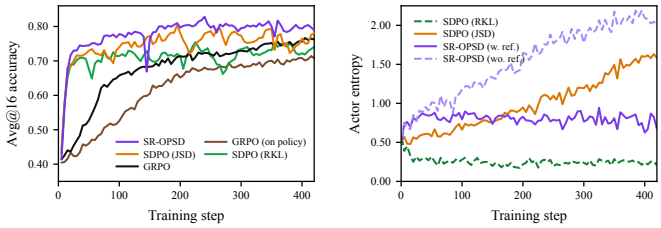


Figure 2: Investigation of Entropy of the Policy on Science Q&A (Chemistry). Left: Avg@16 accuracy. Right: Training entropy of the policy.

In Figure 3, we report Avg@16 test accuracy over training steps on five reasoning-related benchmarks for Qwen3-8B and Olmo3-7B-Instruct. Compared with the SDPO (with JSD) baseline, SR-OPSD generally maintains more stable long-budget performance, whereas SDPO often reaches an early peak and then degrades. These results suggest that the proposed method provides a more favorable performance and training stability.

4.2 Performance on Mathematical Reasoning Benchmarks

Setting We then evaluate the performance in a setting where we have access to ground-truth solutions as feedback on very hard mathematical reasoning problems. Following the setting of Agrawal et al. (2026), we use a training collection of 758 particularly challenging mathematics problems. We benchmark our approach against OPSD (forward KL) (Zhao et al., 2026a), SDPO (Hübotter et al., 2026) (reverse KL), and GRPO on AIME 2024, AIME 2025, HMMT 2025, AMC

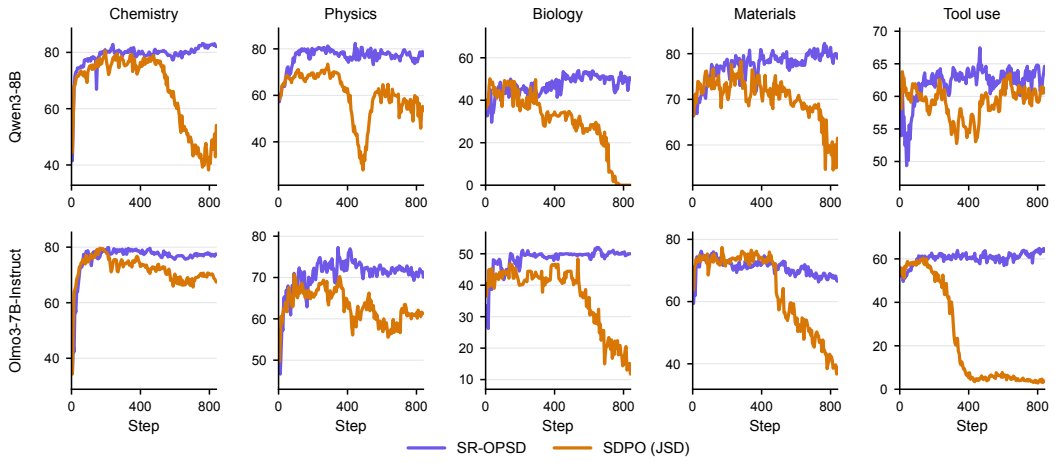


Figure 3: Training dynamics of SDPO and SR-OPSD: the x-axis shows the training step, and the y-axis shows the test Avg@16 accuracy.

2023, and Minerva, with all methods initialized from Qwen3-4B-Instruct-2507 or Qwen3-8B. See Table 7 for detailed experimental configurations.

Table 2: Performance on mathematical reasoning benchmarks. The best and second-best results within each model size are shown in bold and underlined.

Model	Method	AIME24		AIME25		HMMT25		AMC23		Minerva	
		Avg@16	Pass@16	Avg@16	Pass@16	Avg@16	Pass@16	Avg@16	Pass@16	Avg@16	Pass@16
Qwen3-4B	Base	61.4	82.0	50.3	69.5	30.3	48.8	**93.8**	98.8	43.2	48.6
	+ GRPO	62.5	**86.7**	51.1	**80.0**	34.3	60.0	89.7	**100.0**	46.7	**59.2**
	+ SDPO	65.0	**86.7**	54.0	**80.0**	36.3	63.3	92.2	**100.0**	47.0	58.5
	+ OPSD	61.9	83.3	52.1	73.3	33.5	56.7	89.8	**100.0**	46.9	58.8
	+ SR-OPSD (Ours)	**66.3**	**86.7**	**54.6**	**80.0**	**36.5**	**66.7**	93.1	**100.0**	**47.4**	**59.2**
Qwen3-8B	Base	75.9	86.8	66.9	79.6	44.7	65.6	95.6	**100.0**	48.9	56.3
	+ GRPO	76.6	**93.3**	70.1	**93.3**	46.0	80.0	96.2	**100.0**	49.2	61.0
	+ SDPO	77.1	**93.3**	70.0	**93.3**	45.8	76.7	96.7	**100.0**	49.7	60.7
	+ OPSD	77.2	**93.3**	70.3	**93.3**	46.2	**83.3**	**97.3**	**100.0**	49.4	61.4
	+ SR-OPSD (Ours)	**77.7**	**93.3**	**71.9**	**93.3**	**46.5**	80.0	**97.3**	**100.0**	**50.0**	**63.2**

Results analysis Table 2 shows that SR-OPSD achieves the highest mean Avg@16 across the five benchmarks at both model sizes: 59.58 for Qwen3-4B and 68.68 for Qwen3-8B. These scores exceed SDPO by 0.68 and 0.82 percentage points and OPSD by 2.74 and 0.60 percentage points, respectively. For Qwen3-4B, SR-OPSD outperforms all three trained baselines in Avg@16 on every benchmark and achieves the best or joint-best Pass@16 on all five. Its mean Pass@16 reaches 78.52, compared with 77.70 for SDPO and 74.42 for OPSD, showing gains in both average accuracy and multi-sample problem coverage. For Qwen3-8B, SR-OPSD achieves the highest Avg@16 on four benchmarks and ties with OPSD on AMC23. On AIME25, it reaches 71.9, exceeding the strongest baseline, OPSD, by 1.6 percentage points. It also achieves the best or joint-best Pass@16 on four benchmarks, including 63.2 on Minerva, 1.8 percentage points above OPSD. On HMMT25, SR-OPSD and OPSD obtain Pass@16 scores of 80.0 and 83.3, respectively. Overall, SR-OPSD improves Avg@16 over GRPO and SDPO across all five tasks at both model sizes, while matching or exceeding OPSD. Its best or joint-best Pass@16 in nine of the ten model–benchmark combinations further shows that these broad gains in average accuracy are accompanied by strong multi-sample coverage.

4.3 Performance on Code Generation under Model Scaling

Setting We next assess our method on code-generation tasks. Programming provides a representative setting with informative feedback signals, including execution errors and unsuccessful unit tests. Success in this domain depends on effective credit assignment, since the model must locate the precise source of an error and use that information to avoid making the same mistake in subsequent attempts. Following the protocol of (Hübotter et al., 2026), we focus our evaluation on the LCBv6

subset. We use the Qwen3 model family for this investigation and train all methods for 100 iterations. Results are averaged over two random seeds, and are reported with standard deviations. See Table 8 for detailed experimental configurations.

Results Analysis As shown in Table 3, SR-OPSD improves over the base models at all four scales, with gains ranging from 7.7 to 21.8 percentage points. Its mean score increases monotonically from 16.9 at 0.6B to 48.0 at 8B. At 1.7B, 4B, and 8B, SR-OPSD achieves the highest mean scores, exceeding the strongest baseline at each scale by 4.0, 10.1, and 3.8 percentage points, respectively.

Table 3: Performance on LiveCodeBench v6 across Qwen3 model scales. The best result in each column is shown in bold.

Method	Qwen3-0.6B	Qwen3-1.7B	Qwen3-4B	Qwen3-8B
Base	9.2	14.7	23.1	27.9
GRPO	15.1 ± 0.6	23.7 ± 0.9	34.8 ± 2.2	37.1 ± 2.2
SDPO	$\mathbf{18.4} \pm 1.1$	27.8 ± 1.4	33.5 ± 0.1	44.2 ± 0.2
SR-OPSD	16.9 ± 1.0	$\mathbf{31.8} \pm 0.3$	$\mathbf{44.9} \pm 0.7$	$\mathbf{48.0} \pm 2.9$

4.4 Ablation Studies

We examine how the effect of reference anchoring depends on the projection objective, and how target interpolation interacts with the Rényi order. Using Qwen3-8B on SciKnowEval Physics, we first compare JSD (Hübotter et al., 2026), forward KL (Zhao et al., 2026a), and forward Rényi with and without reference anchoring. We then jointly vary α and ρ within SR-OPSD to examine their effects on performance, as shown in Table 4. All variants share the same initialization, training data, and evaluation protocol. For each configuration, we report the Avg@16 accuracy achieved within a 10-hour training budget.

Table 4: Ablation on SciKnowEval Physics (Avg@16). Top: effect of self-reference, with $\alpha = 0.90$ for self-referenced variants; Δ denotes the change in percentage points. Bottom: joint effect of α and ρ. The best result in each block is bold.

Effect of reference anchoring

Objective	No reference	With reference	Δ
JSD	73.4	77.2	+3.8
Forward KL	79.1	76.3	−2.8
Forward Rényi ($\rho = 0.50$)	76.5	79.9	+3.4
Forward Rényi ($\rho = 0.70$)	75.9	**82.7**	+6.8
Forward Rényi ($\rho = 0.90$)	79.7	77.2	−2.5
Forward Rényi ($\rho = 0.95$)	77.7	<u>81.2</u>	+3.5

Joint effect of α and ρ

	Rényi order ρ			
α	0.50	0.70	0.90	0.95
0.50	78.8	78.8	78.2	78.4
0.70	76.3	82.0	77.0	80.7
0.90	79.9	**82.7**	77.2	81.2

Does self-reference consistently improve performance? The effect of reference anchoring depends on the projection objective. Self-reference improves JSD from 73.4 to 77.2, a gain of 3.8 percentage points, but reduces forward KL from 79.1 to 76.3, a drop of 2.8 points. These opposing effects show that adding a reference target does not consistently improve performance, motivating consideration of how the student is projected toward that target.

Does Rényi projection alone improve over forward KL? Without self-reference, forward Rényi achieves 75.9 at $\rho = 0.70$ and 77.7 at $\rho = 0.95$, both below forward KL at 79.1. Adding self-reference raises these scores to 82.7 and 81.2, respectively. In particular, at $\rho = 0.70$, self-reference alone and Rényi projection alone each underperform the unanchored forward-KL baseline, whereas their combination exceeds it by 3.6 percentage points. This reversal supports jointly designing the supervisory target and projection geometry.

How do α and ρ interact? The effect of changing α depends on ρ. Increasing α from 0.50 to 0.90 improves Avg@16 from 78.8 to 82.7 at $\rho = 0.70$, but reduces it from 78.2 to 77.2 at $\rho = 0.90$. Sensitivity to ρ also varies with α: scores span only 78.2–78.8 at $\alpha = 0.50$, compared with 77.2–82.7 at $\alpha = 0.90$. The best tested configuration is $(\alpha, \rho) = (0.90, 0.70)$, achieving 82.7, followed by $(0.70, 0.70)$ at 82.0. In summary, neither self-reference nor Rényi projection alone consistently improves performance. See Figure 7 in Appendix D.3 for discussions on responses lengths.

Memory and Computation Cost We separately benchmark resource costs on SciKnowEval Physics with Qwen3-8B, giving each run exclusive access to a single node with eight NVIDIA H200 GPUs. For each run, we report the maximum logged single-device PyTorch-allocated memory across all GPU workers and steps in decimal GB, and compute the average step time.

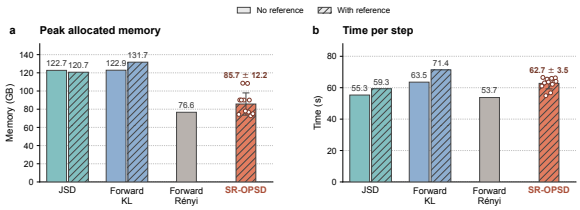


Figure 4: Memory Consumption and Empirical Computational Cost.

Across the 12 evaluated (α, ρ) configurations, SR-OPSD uses 85.7 ± 12.2 GB and 62.7 ± 3.5 seconds per step, where $\pm$ denotes the population standard deviation across configurations. Compared with self-reference-anchored JSD and forward KL, its mean allocated memory is lower by 29.0% and 34.9%, respectively, while its mean step time is 5.7% higher and 12.2% lower. Relative to unanchored forward Rényi, SR-OPSD uses 11.9% more memory and takes 16.8% longer per step.

5 Related Work

On-Policy Distillation On-policy distillation uses student-generated trajectories for training, allowing the teacher policy to provide dense supervision on prefixes visited by the student (Agarwal et al., 2024; Wen et al., 2023; Gu et al., 2023; Lu & Lab, 2025; Jin et al., 2026; Luo et al., 2026; Yang et al., 2026; Stein et al., 2026; Oh et al., 2026; Kim et al., 2026). Gu et al. (2023) optimizes sequence-level reverse KL through policy-gradient estimators, while Agarwal et al. (2024) evaluates forward KL, reverse KL, and JSD on student rollouts. Later work studies entropy-dependent divergence selection, reference-policy constraints, and rollout mixing to improve stability (Jin et al., 2026; Luo et al., 2026). Yang et al. (2026) further formulates external-teacher on-policy distillation as KL-regularized policy optimization, where a teacher-to-reference log-probability ratio defines the dense reward and controls interpolation or extrapolation between the two policies. See (Song & Zheng, 2026) for a detailed review.

On-Policy Self-Distillation These methods replace the external teacher with a self-teacher policy derived from the policy being optimized and conditioned on privileged or auxiliary context available only during training. Existing work uses ground-truth solutions, environment feedback, successful sibling responses, or document context to construct the self-teacher (Zhao et al., 2026a; Hübotter et al., 2026; Penaloza et al., 2026; Ye et al., 2026; Sang et al., 2026; Zhao et al., 2026b; Wang et al., 2026). Subsequent work controls the amount of additional information exposed to the teacher, restricts distillation to selected spans, or gates token-level supervision according to teacher reliability and uncertainty (Han et al., 2026; Wang et al., 2026; Liu et al., 2026; Ke et al., 2026). Other studies analyze the effects of additional context information, Top-K approximation, and update schemes of self-teacher policies (Zhu et al., 2026; Guo et al., 2026). Existing objectives mainly use forward KL, reverse KL, or JSD for direct distribution matching (Zhao et al., 2026a; Hübotter et al., 2026), while preference-based self-distillation replaces direct matching with reward-regularized preference optimization (Yu et al., 2026).

6 Conclusion

Our study shows that reference anchoring can improve or degrade self-distillation accuracy depending on the projection objective. This dependence highlights the need to consider target construction and the student's response to the target jointly. SR-OPSD brings these choices into a single objective by combining a geometric target anchored to the frozen initial policy with forward Rényi projection. At fixed student-generated contexts, our variational characterization and exact token-logit gradient reveal an effective update target that geometrically combines the self-teacher, reference, and current student policies. Experiments across scientific reasoning, tool use, mathematical reasoning, and code generation demonstrate strong performance.

Understanding how this interaction evolves throughout training remains an open question. Our analysis conditions on fixed rollout contexts and frozen target components, leaving the joint evolution of the student, self-teacher, and rollout distribution beyond its scope. Future work could extend the analysis to these coupled dynamics and adapt α and ρ to teacher reliability or policy drift.

AI USE STATEMENT

Generative AI tools were used to polish the language, prepare figures, and plan experiments. We have reviewed the AI-assisted material and take responsibility for the final content of this work, including any text, claims, or artifacts produced with the aid of generative AI.

REPRODUCIBILITY STATEMENT

SR-OPSD objective and training procedure are provided in Section 3 and Algorithm 1. Appendix B provides the derivations for our theoretical results. Section 4 describes the benchmarks, models, baselines, and evaluation metrics. Appendix D.2 (Tables 6–8) reports the training hyperparameters, rollout and evaluation settings, and method-specific implementation choices. The measurement and aggregation procedures for memory and computation cost are described in Section 4.4. Code is also provided as supplemental materials.

ETHICS STATEMENT

This work studies on-policy self-distillation using existing language-model checkpoints and benchmark datasets, whose sources are cited in the paper. We describe our experimental settings, evaluation procedures, and computational costs to support transparent assessment and reproducibility. The resulting models may inherit biases, privacy risks, and harmful behaviors from their underlying models and training data. Improved reasoning and code-generation capabilities may also enable misuse. Benchmark improvements alone do not establish fairness, factual reliability, or deployment safety. Use of the proposed method should therefore respect the licenses and usage restrictions of the underlying resources and include application-specific safety evaluation, appropriate safeguards, and human oversight.

REFERENCES

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

Rishabh Agrawal, Jacob Fein-Ashley, and Paria Rashidinejad. Reinforcement learning from rich feedback with distributional DAgger. *arXiv preprint arXiv:2606.05152*, 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.

Haowei Guo, Baolong Bi, Ruicheng Zhang, Bingqian Sun, and Wentao Zhang. When should the teacher move? temporal coupling and stability in self on-policy distillation. *arXiv preprint arXiv:2606.03532*, 2026.

Zihao Han, Tiangang Zhang, Huaibin Wang, and Yilun Sun. Adaptive teacher exposure for self-distillation in llm reasoning. *arXiv preprint arXiv:2605.11458*, 2026.

Li He, Qiang Qu, He Zhao, Stephen Wan, Dadong Wang, Lina Yao, and Tongliang Liu. Unifying stable optimization and reference regularization in RLHF. In *The Fourteenth International Conference on Learning Representations*, 2026.

Jonas Hübotter, Frederike Lübeck, Lejs Deen Behric, Anton Baumann, Marco Bagatella, Daniel Marta, Ido Hakimi, Idan Shenfeld, Thomas Kleine Buening, Carlos Guestrin, and Andreas Krause. Reinforcement learning via self-distillation. In *Forty-third International Conference on Machine Learning*, 2026.

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.

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.

Yeongmin Kim, Donghyeok Shin, Mina Kang, Byeonghu Na, and Il-Chul Moon. Distillation of large language models via concrete score matching. In *International Conference on Learning Representations*, volume 2026, pp. 16649–16682, 2026.

Xiaogeng Liu, Xinyan Wang, Yingzi Ma, Yechao Zhang, and Chaowei Xiao. When are teacher tokens reliable? position-weighted on-policy self-distillation for reasoning. *arXiv preprint arXiv:2605.21606*, 2026.

Kevin Lu and Thinking Machines Lab. On-policy distillation. *Thinking Machines Lab: Connectionism*, 2025. doi: 10.64434/tml.20251026.

Feng Luo, Yu-Neng Chuang, Guanchu Wang, Zicheng Xu, Xiaotian Han, Tianyi Zhang, and Vladimir Braverman. Demystifying opd: Length inflation and stabilization strategies for large language models. *arXiv preprint arXiv:2604.08527*, 2026.

Minjae Oh, Sangjun Song, Gyubin Choi, Yunho Choi, and Yohan Jo. Kl for a kl: On-policy distillation with control variate baseline. *arXiv preprint arXiv:2605.07865*, 2026.

Team Olmo, Allyson Ettinger, Amanda Bertsch, Bailey Kuehl, David Graham, David Heineman, Dirk Groeneveld, Faeze Brahman, Finbarr Timbers, Hamish Ivison, et al. Olmo 3. *arXiv preprint arXiv:2512.13961*, 2025.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul F Christiano, Jan Leike, and Ryan Lowe. Training language models to follow instructions with human feedback. In S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh (eds.), *Advances in Neural Information Processing Systems*, volume 35, pp. 27730–27744. Curran Associates, Inc., 2022. doi: 10.52202/068431-2011.

Emiliano Penaloza, Dheeraj Vattikonda, Nicolas Gontier, Alexandre Lacoste, Laurent Charlin, and Massimo Caccia. Privileged information distillation for language models. *arXiv preprint arXiv:2602.04942*, 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 *Advances in Neural Information Processing Systems*, volume 36, pp. 53728–53741, 2023. doi: 10.52202/075280-2338.

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.

Victor Sanh, Albert Webson, Colin Raffel, Stephen Bach, Lintang Sutawika, Zaid Alyafeai, Antoine Chaffin, Arnaud Stiegler, Arun Raja, Manan Dey, M Saiful Bari, Canwen Xu, Urmish Thakker, Shanya Sharma Sharma, Eliza Szczechla, Taewoon Kim, Gunjan Chhablani, Nihal Nayak, Debajyoti Datta, Jonathan Chang, Mike Tian-Jian Jiang, Han Wang, Matteo Manica, Sheng Shen, Zheng Xin Yong, Harshit Pandey, Rachel Bawden, Thomas Wang, Trishala Neeraj, Jos Rozen, Abheesht Sharma, Andrea Santilli, Thibault Fevry, Jason Alan Fries, Ryan Teehan, Teven Le Scao, Stella Biderman, Leo Gao, Thomas Wolf, and Alexander M Rush. Multitask prompted training enables zero-shot task generalization. In *International Conference on Learning Representations*, 2022.

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, Y. K. Li, Y. Wu, et al. DeepSeekMath: Pushing the limits of mathematical reasoning in open language models. *arXiv preprint arXiv:2402.03300*, 2024.

Mingyang Song and Mao Zheng. A survey of on-policy distillation for large language models. *arXiv preprint arXiv:2604.00626*, 2026.

Alex Stein, Furong Huang, and Tom Goldstein. GATES: Self-distillation under privileged context with consensus gating. *arXiv preprint arXiv:2602.20574*, 2026.

Qiaoyu Tang, Ziliang Deng, Hongyu Lin, Xianpei Han, Qiao Liang, Boxi Cao, and Le Sun. Toolalpaca: Generalized tool learning for language models with 3000 simulated cases. *arXiv preprint arXiv:2306.05301*, 2023.

Tim van Erven and Peter Harremoës. Rényi divergence and kullback-leibler divergence. *IEEE transactions on information theory*, 60(7):3797–3820, 2014. doi: 10.1109/TIT.2014.2320500.

Jiaxuan Wang, Xuan Ouyang, Zhiyu Chen, Yulan Hu, Zheng Pan, Xin Li, and Lan-Zhe Guo. TRACE: Distilling where it matters via token-routed self on-policy alignment. *arXiv preprint arXiv:2605.10194*, 2026.

Jason Wei, Maarten Bosma, Vincent Zhao, Kelvin Guu, Adams Wei Yu, Brian Lester, Nan Du, Andrew M. Dai, and Quoc V Le. Finetuned language models are zero-shot learners. In *International Conference on Learning Representations*, 2022.

Yuqiao Wen, Zichao Li, Wenyu Du, and Lili Mou. f-divergence minimization for sequence-level knowledge distillation. In Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki (eds.), *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 10817–10834, Toronto, Canada, July 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.acl-long.605.

Di Wu, Chengshuai Shi, Jing Yang, and Cong Shen. f-divergence regularized RLHF: Two tales of sampling and unified analyses. In *Forty-third International Conference on Machine Learning*, 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*, 2026.

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

Qiying Yu, Zheng Zhang, Ruofei Zhu, Yufeng Yuan, Xiaochen Zuo, Yu Yue, Weinan Dai, Tiantian Fan, Gaohong Liu, Juncai Liu, LingJun Liu, Xin Liu, Haibin Lin, Zhiqi Lin, Bole Ma, Guangming Sheng, Yuxuan Tong, Chi Zhang, Mofan Zhang, Ru Zhang, Wang Zhang, Hang Zhu, Jinhua Zhu, Jiaze Chen, Jiangjie Chen, Chengyi Wang, Hongli Yu, Yuxuan Song, Xiangpeng Wei, Hao Zhou, Jingjing Liu, Wei-Ying Ma, Ya-Qin Zhang, Lin Yan, Yonghui Wu, and Mingxuan Wang. DAPO: An open-source LLM reinforcement learning system at scale. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025. doi: 10.52202/085713-3775.

Xin Yu, Liuchen Liao, Yiwen Zhang, Yingchen Yu, Lingzhou Xue, and Qinzhen Guo. Preference-based self-distillation: Beyond kl matching via reward regularization. *arXiv preprint arXiv:2605.05040*, 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. In *Forty-third International Conference on Machine Learning*, 2026a.

Zhengyang Zhao, Lu Ma, and Wentao Zhang. Training with harnesses: On-policy harness self-distillation for complex reasoning. *arXiv preprint arXiv:2605.08741*, 2026b.

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.

APPENDIX

A NOTATION TABLE

Table 5: Summary of the notation.

Symbol	Meaning
$x, \mathcal{D}$	Prompt and prompt distribution, with $x \sim \mathcal{D}$.
$y, y_t, y_{<t}$	On-policy response, token at position t, and prefix before t.
$a, \mathcal{V}, \mathcal{C}_t$	Generic next-token action, full vocabulary, and candidate set at position t.
h_t	Student prefix context, $h_t = (x, y_{<t})$.
$h_t^{\mathrm{sf}}, h_t^{\mathrm{ref}}$	Context information for the self-teacher policy and reference policy at position t, respectively.
$h, h_{\mathrm{sf}}, h_{\mathrm{ref}}$	Fixed-context shorthand obtained by suppressing the position index t.
π_θ, θ	Policy to be optimized (student policy) and its parameters.
$\pi_{\bar{\theta}}, \bar{\theta}$	Frozen or EMA self-teacher policy and its parameters.
$\pi_{\theta_{\mathrm{old}}}$	Rollout policy used to sample the current on-policy batch.
π_{ref}	Reference policy.
θ_0	Parameters of the student policy before optimization.
π_T	External teacher policy, when on-policy distillation is discussed.
$\mathrm{sg}[\cdot]$	Stop-gradient operator.
$\pi_\alpha^\star, Z_\alpha$	Reference-anchored self-teacher target and its normalizing constant.
α, ρ	Target-interpolation coefficient and Rényi order.
D_{KL}, D_ρ	KL divergence and Rényi divergence.
β	EMA coefficient.

B Theoretical Results and Proofs

B.1 Proof of Token-Level Variational Characterization of On-policy Self-distillation

Proposition B.1 (Token-Level Variational Characterization of OPSD). *Fix h, h_{sf}, and h_{ref}, and let $\alpha \in [0,1]$. Assume that, for every token $a \in \mathcal{V}$, $\pi_\theta(a \mid h) > 0$, $\mathrm{sg}[\pi_{\bar\theta}(a \mid h_{\mathrm{sf}})] > 0$, $\pi_{\mathrm{ref}}(a \mid h_{\mathrm{ref}}) > 0$. Define $Z_\alpha(h_{\mathrm{sf}}, h_{\mathrm{ref}}) := \sum_{a\in\mathcal{V}} \mathrm{sg}[\pi_{\bar\theta}(a \mid h_{\mathrm{sf}})]^\alpha \, \pi_{\mathrm{ref}}(a \mid h_{\mathrm{ref}})^{1-\alpha}$ and $\pi^\star_\alpha(a \mid h_{\mathrm{sf}}, h_{\mathrm{ref}}) := \frac{\mathrm{sg}[\pi_{\bar\theta}(a|h_{\mathrm{sf}})]^\alpha \pi_{\mathrm{ref}}(a|h_{\mathrm{ref}})^{1-\alpha}}{Z_\alpha(h_{\mathrm{sf}},h_{\mathrm{ref}})}$. Then the conditional token-level objective*

$$\mathcal{J}_\alpha(\theta; h, h_{\mathrm{sf}}, h_{\mathrm{ref}}) := \mathbb{E}_{a\sim\pi_\theta(\cdot|h)}\left[\alpha \log \frac{\mathrm{sg}[\pi_{\bar\theta}(a \mid h_{\mathrm{sf}})]}{\pi_{\mathrm{ref}}(a \mid h_{\mathrm{ref}})}\right] - D_{\mathrm{KL}}(\pi_\theta(\cdot \mid h) \,\|\, \pi_{\mathrm{ref}}(\cdot \mid h_{\mathrm{ref}}))$$

satisfies $\mathcal{J}_\alpha(\theta; h, h_{\mathrm{sf}}, h_{\mathrm{ref}}) = -D_{\mathrm{KL}}(\pi_\theta(\cdot \mid h) \,\|\, \pi^\star_\alpha(\cdot \mid h_{\mathrm{sf}}, h_{\mathrm{ref}})) + \log Z_\alpha(h_{\mathrm{sf}}, h_{\mathrm{ref}})$. Consequently, for fixed contexts and frozen target components, maximizing $\mathcal{J}_\alpha$ with respect to θ is equivalent to minimizing the reverse KL from the student to $\pi^\star_\alpha$. When $\alpha = 1$, $Z_1 = 1$ and $\pi^\star_1(\cdot \mid h_{\mathrm{sf}}, h_{\mathrm{ref}}) = \mathrm{sg}[\pi_{\bar\theta}(\cdot \mid h_{\mathrm{sf}})]$, recovering the token-level objective.

Proof. Fix h, h_{sf}, and h_{ref}, and abbreviate

$$q(a) := \pi_\theta(a \mid h), \qquad p(a) := \mathrm{sg}[\pi_{\bar\theta}(a \mid h_{\mathrm{sf}})], \qquad r(a) := \pi_{\mathrm{ref}}(a \mid h_{\mathrm{ref}}).$$

The conditional objective expands as

$$\begin{aligned}\mathcal{J}_\alpha &= \sum_a q(a)\left[\alpha \log \frac{p(a)}{r(a)} - \log \frac{q(a)}{r(a)}\right] \\ &= \sum_a q(a)\left[-\log q(a) + \alpha \log p(a) + (1-\alpha)\log r(a)\right].\end{aligned}$$

By definition,

$$\log \pi^\star_\alpha(a \mid h_{\mathrm{sf}}, h_{\mathrm{ref}}) = \alpha \log p(a) + (1-\alpha)\log r(a) - \log Z_\alpha(h_{\mathrm{sf}}, h_{\mathrm{ref}}).$$

Substituting this identity yields

$$\begin{aligned}\mathcal{J}_\alpha &= \sum_a q(a)\left[-\log q(a) + \log \pi^\star_\alpha(a \mid h_{\mathrm{sf}}, h_{\mathrm{ref}}) + \log Z_\alpha(h_{\mathrm{sf}}, h_{\mathrm{ref}})\right] \\ &= -D_{\mathrm{KL}}(\pi_\theta(\cdot \mid h) \,\|\, \pi^\star_\alpha(\cdot \mid h_{\mathrm{sf}}, h_{\mathrm{ref}})) + \log Z_\alpha(h_{\mathrm{sf}}, h_{\mathrm{ref}}),\end{aligned}$$

where the last step uses $\sum_a q(a) = 1$. Since the contexts and target components are fixed, Z_α is constant with respect to the inner optimization variable θ.

When $\alpha = 1$,

$$Z_1(h_{\mathrm{sf}}, h_{\mathrm{ref}}) = \sum_a p(a) = 1$$

and

$$\pi^\star_1(a \mid h_{\mathrm{sf}}, h_{\mathrm{ref}}) = p(a) = \mathrm{sg}[\pi_{\bar\theta}(a \mid h_{\mathrm{sf}})].$$

The objective therefore reduces to the reverse-KL OPSD objective. The constancy statement is conditional on the fixed contexts: after averaging over a context distribution that itself changes with the policy, the expected normalizer need not be constant. □

B.2 Proof of Proposition 3.1

Proof. Fix h, h_{sf}, and h_{ref}. To simplify the notation, define

$$\begin{aligned}q(a) &:= \pi_\theta(a \mid h), \\ p(a) &:= \mathrm{sg}[\pi_{\bar\theta}(a \mid h_{\mathrm{sf}})], \\ r(a) &:= \pi_{\mathrm{ref}}(a \mid h_{\mathrm{ref}}).\end{aligned}$$

Because all token probabilities are strictly positive, all logarithms, probability ratios, and real-valued powers appearing below are well-defined.

Under these abbreviations, the normalizing constant is

$$Z_\alpha := Z_\alpha(h_{\text{sf}}, h_{\text{ref}}) = \sum_{a \in \mathcal{V}} p(a)^\alpha r(a)^{1-\alpha}.$$

The normalized self-referenced target satisfies

$$\pi^\star_\alpha(a \mid h_{\text{sf}}, h_{\text{ref}}) = \frac{p(a)^\alpha r(a)^{1-\alpha}}{Z_\alpha}.$$

Multiplying both sides of above equation by Z_α gives

$$p(a)^\alpha r(a)^{1-\alpha} = Z_\alpha \pi^\star_\alpha(a \mid h_{\text{sf}}, h_{\text{ref}}). \tag{5}$$

We first simplify the numerator of $\mathcal{R}_{\alpha,\rho}$. By the definition of the pseudo-reward r_α,

$$r_\alpha(a; h_{\text{sf}}, h_{\text{ref}}) = \alpha \log \frac{p(a)}{r(a)}.$$

Multiplying both sides by ρ yields

$$\rho r_\alpha(a; h_{\text{sf}}, h_{\text{ref}}) = \rho\alpha \log \frac{p(a)}{r(a)}. \tag{6}$$

Exponentiating both sides of Equation (6) gives

$$\begin{aligned}
\exp(\rho r_\alpha(a; h_{\text{sf}}, h_{\text{ref}})) &= \exp\left(\rho\alpha \log \frac{p(a)}{r(a)}\right) \\
&= \left(\frac{p(a)}{r(a)}\right)^{\rho\alpha} \\
&= p(a)^{\rho\alpha} r(a)^{-\rho\alpha}.
\end{aligned} \tag{7}$$

Consider one summand in the numerator of $\mathcal{R}_{\alpha,\rho}$. Using Equation (7), we obtain

$$\begin{aligned}
&r(a)^\rho q(a)^{1-\rho} \exp(\rho r_\alpha(a; h_{\text{sf}}, h_{\text{ref}})) \\
&= r(a)^\rho q(a)^{1-\rho} p(a)^{\rho\alpha} r(a)^{-\rho\alpha}.
\end{aligned} \tag{8}$$

Combining the two powers of $r(a)$ gives

$$\begin{aligned}
r(a)^\rho r(a)^{-\rho\alpha} &= r(a)^{\rho-\rho\alpha} \\
&= r(a)^{\rho(1-\alpha)}.
\end{aligned} \tag{9}$$

Substituting Equation (9) into Equation (8) yields

$$\begin{aligned}
&r(a)^\rho q(a)^{1-\rho} \exp(\rho r_\alpha(a; h_{\text{sf}}, h_{\text{ref}})) \\
&= q(a)^{1-\rho} p(a)^{\rho\alpha} r(a)^{\rho(1-\alpha)}.
\end{aligned} \tag{10}$$

Therefore, Equation (10) becomes

$$\begin{aligned}
&r(a)^\rho q(a)^{1-\rho} \exp(\rho r_\alpha(a; h_{\text{sf}}, h_{\text{ref}})) \\
&= q(a)^{1-\rho} \left[p(a)^\alpha r(a)^{1-\alpha}\right]^\rho.
\end{aligned} \tag{11}$$

Applying Equation (5) to Equation (11) gives

$$\begin{aligned}
&r(a)^\rho q(a)^{1-\rho} \exp(\rho r_\alpha(a; h_{\text{sf}}, h_{\text{ref}})) \\
&= q(a)^{1-\rho} \left[Z_\alpha \pi^\star_\alpha(a \mid h_{\text{sf}}, h_{\text{ref}})\right]^\rho \\
&= Z_\alpha^\rho \pi^\star_\alpha(a \mid h_{\text{sf}}, h_{\text{ref}})^\rho q(a)^{1-\rho}.
\end{aligned} \tag{12}$$

Summing both sides of Equation (12) over $a \in \mathcal{V}$ gives

$$\begin{aligned}
&\sum_{a \in \mathcal{V}} r(a)^\rho q(a)^{1-\rho} \exp(\rho r_\alpha(a; h_{\text{sf}}, h_{\text{ref}})) \\
&= \sum_{a \in \mathcal{V}} Z_\alpha^\rho \pi^\star_\alpha(a \mid h_{\text{sf}}, h_{\text{ref}})^\rho q(a)^{1-\rho}.
\end{aligned} \tag{13}$$

Because Z_α does not depend on the summation variable a, it can be taken outside the sum:

$$\sum_{a\in\mathcal{V}} r(a)^\rho q(a)^{1-\rho} \exp(\rho r_\alpha(a; h_{\text{sf}}, h_{\text{ref}})) = Z_\alpha^\rho \sum_{a\in\mathcal{V}} \pi_\alpha^\star(a \mid h_{\text{sf}}, h_{\text{ref}})^\rho q(a)^{1-\rho}. \tag{14}$$

By the definition of Rényi divergence,

$$D_\rho(\pi_\alpha^\star(\cdot \mid h_{\text{sf}}, h_{\text{ref}}) \,\|\, q) = \frac{1}{\rho-1} \log \sum_{a\in\mathcal{V}} \pi_\alpha^\star(a \mid h_{\text{sf}}, h_{\text{ref}})^\rho q(a)^{1-\rho}. \tag{15}$$

Multiplying both sides of Equation (15) by $\rho - 1$ gives

$$(\rho-1) D_\rho(\pi_\alpha^\star(\cdot \mid h_{\text{sf}}, h_{\text{ref}}) \,\|\, q) = \log \sum_{a\in\mathcal{V}} \pi_\alpha^\star(a \mid h_{\text{sf}}, h_{\text{ref}})^\rho q(a)^{1-\rho}. \tag{16}$$

Exponentiating both sides of Equation (16) gives

$$\exp((\rho-1) D_\rho(\pi_\alpha^\star(\cdot \mid h_{\text{sf}}, h_{\text{ref}}) \,\|\, q)) = \sum_{a\in\mathcal{V}} \pi_\alpha^\star(a \mid h_{\text{sf}}, h_{\text{ref}})^\rho q(a)^{1-\rho}. \tag{17}$$

Substituting Equation (17) into Equation (14) yields

$$\sum_{a\in\mathcal{V}} r(a)^\rho q(a)^{1-\rho} \exp(\rho r_\alpha(a; h_{\text{sf}}, h_{\text{ref}})) = Z_\alpha^\rho \exp((\rho-1) D_\rho(\pi_\alpha^\star(\cdot \mid h_{\text{sf}}, h_{\text{ref}}) \,\|\, q)). \tag{18}$$

We next simplify the denominator of $\mathcal{R}_{\alpha,\rho}$. By the definition of Rényi divergence,

$$D_\rho(q\|r) = \frac{1}{\rho-1} \log \sum_{a\in\mathcal{V}} q(a)^\rho r(a)^{1-\rho}. \tag{19}$$

Multiplying both sides of Equation (19) by $\rho - 1$ gives

$$(\rho-1) D_\rho(q\|r) = \log \sum_{a\in\mathcal{V}} q(a)^\rho r(a)^{1-\rho}. \tag{20}$$

Exponentiating both sides of Equation (20) gives

$$\exp((\rho-1) D_\rho(q\|r)) = \sum_{a\in\mathcal{V}} q(a)^\rho r(a)^{1-\rho}. \tag{21}$$

Using Equations (18) and (21) in the definition of $\mathcal{R}_{\alpha,\rho}$ gives

$$\mathcal{R}_{\alpha,\rho}(\theta; h, h_{\text{sf}}, h_{\text{ref}}) = -\frac{1}{\rho-1} \log \frac{Z_\alpha^\rho \exp((\rho-1) D_\rho(\pi_\alpha^\star(\cdot \mid h_{\text{sf}}, h_{\text{ref}}) \,\|\, q))}{\exp((\rho-1) D_\rho(q\|r))}. \tag{22}$$

We then obtain

$$\mathcal{R}_{\alpha,\rho}(\theta; h, h_{\text{sf}}, h_{\text{ref}}) = -\frac{1}{\rho-1} \Big[\log(Z_\alpha^\rho \exp((\rho-1) D_\rho(\pi_\alpha^\star(\cdot \mid h_{\text{sf}}, h_{\text{ref}}) \,\|\, q))) - \log(\exp((\rho-1) D_\rho(q\|r))) \Big]. \tag{23}$$

Expanding the logarithms and suppressing the fixed context arguments gives

$$\mathcal{R}_{\alpha,\rho}(\theta) = -\frac{\rho}{\rho-1}\log Z_\alpha - D_\rho(\pi_\alpha^\star\|q) + D_\rho(q\|r).$$

Subtracting the reference penalty therefore yields

$$\begin{aligned}\mathcal{J}_{\alpha,\rho}(\theta) &= \mathcal{R}_{\alpha,\rho}(\theta) - D_\rho(q\|r)\\ &= -D_\rho(\pi_\alpha^\star\|q) - \frac{\rho}{\rho-1}\log Z_\alpha.\end{aligned}$$

For fixed h, h_{sf}, and h_{ref}, the distributions p and r are frozen during the inner student update. Thus,

$$Z_\alpha(h_{\mathrm{sf}}, h_{\mathrm{ref}}) = \sum_{a\in\mathcal{V}} p(a)^\alpha r(a)^{1-\alpha}$$

is independent of θ, so the normalizer term is constant. Restoring $q = \pi_\theta(\cdot \mid h)$, we obtain

$$\arg\max_\theta \mathcal{J}_{\alpha,\rho}(\theta) = \arg\min_\theta D_\rho(\pi_\alpha^\star(\cdot \mid h_{\mathrm{sf}}, h_{\mathrm{ref}}) \,\|\, \pi_\theta(\cdot \mid h)).$$

This proves the stated conditional variational equivalence.

When $\alpha = 1$, $Z_1 = \sum_a p(a) = 1$ and $\pi_1^\star = p$. Hence,

$$\mathcal{J}_{1,\rho}(\theta; h, h_{\mathrm{sf}}, h_{\mathrm{ref}}) = -D_\rho(\mathrm{sg}[\pi_{\bar\theta}(\cdot \mid h_{\mathrm{sf}})] \,\|\, \pi_\theta(\cdot \mid h)).$$

□

B.3 Proof of Proposition 3.2

Proof. By the definition of Rényi divergence,

$$\begin{aligned}&D_\rho(\pi_\alpha^\star(\cdot \mid h_{\mathrm{sf}}, h_{\mathrm{ref}}) \,\|\, \pi_\theta(\cdot \mid h))\\ &= \frac{1}{\rho-1}\log\sum_{b\in\mathcal{V}} \pi_\alpha^\star(b \mid h_{\mathrm{sf}}, h_{\mathrm{ref}})^\rho \pi_\theta(b \mid h)^{1-\rho}.\end{aligned} \tag{24}$$

Define

$$S_{\alpha,\rho,\theta} := \sum_{b\in\mathcal{V}} \pi_\alpha^\star(b \mid h_{\mathrm{sf}}, h_{\mathrm{ref}})^\rho \pi_\theta(b \mid h)^{1-\rho}. \tag{25}$$

Then

$$D_\rho = \frac{1}{\rho-1}\log S_{\alpha,\rho,\theta}.$$

We first compute the derivative of the student probability with respect to the student logit. By the softmax definition,

$$\pi_\theta(b \mid h) = \frac{\exp(z_\theta(b \mid h))}{\sum_{c\in\mathcal{V}}\exp(z_\theta(c \mid h))}.$$

Therefore,

$$\begin{aligned}\frac{\partial \pi_\theta(b \mid h)}{\partial z_\theta(a \mid h)} &= \frac{\mathbf{1}\{a=b\}\exp(z_\theta(b \mid h))}{\sum_c \exp(z_\theta(c \mid h))}\\ &\quad - \frac{\exp(z_\theta(b \mid h))\exp(z_\theta(a \mid h))}{\left(\sum_c \exp(z_\theta(c \mid h))\right)^2}\\ &= \pi_\theta(b \mid h)\left[\mathbf{1}\{a=b\} - \pi_\theta(a \mid h)\right].\end{aligned} \tag{26}$$

Since $\pi_\alpha^\star$ is frozen during the student update, only π_θ depends on z_θ. Hence,

$$\begin{aligned}\frac{\partial S_{\alpha,\rho,\theta}}{\partial z_\theta(a \mid h)} &= \sum_{b\in\mathcal{V}} \pi_\alpha^\star(b \mid h_{\mathrm{sf}}, h_{\mathrm{ref}})^\rho \frac{\partial \pi_\theta(b \mid h)^{1-\rho}}{\partial z_\theta(a \mid h)}\\ &= (1-\rho)\sum_{b\in\mathcal{V}} \pi_\alpha^\star(b \mid h_{\mathrm{sf}}, h_{\mathrm{ref}})^\rho \pi_\theta(b \mid h)^{-\rho}\frac{\partial \pi_\theta(b \mid h)}{\partial z_\theta(a \mid h)}.\end{aligned}$$

Substituting Equation (26) gives

$$\frac{\partial S_{\alpha,\rho,\theta}}{\partial z_\theta(a \mid h)} = (1-\rho) \sum_{b \in \mathcal{V}} \pi^\star_\alpha(b \mid h_{\mathrm{sf}}, h_{\mathrm{ref}})^\rho \pi_\theta(b \mid h)^{1-\rho} \times \left[\mathbf{1}\{a = b\} - \pi_\theta(a \mid h)\right].$$

Separating the two terms,

$$\begin{aligned}\frac{\partial S_{\alpha,\rho,\theta}}{\partial z_\theta(a \mid h)} &= (1-\rho)\pi^\star_\alpha(a \mid h_{\mathrm{sf}}, h_{\mathrm{ref}})^\rho \pi_\theta(a \mid h)^{1-\rho} \\ &\quad - (1-\rho)\pi_\theta(a \mid h) \sum_{b \in \mathcal{V}} \pi^\star_\alpha(b \mid h_{\mathrm{sf}}, h_{\mathrm{ref}})^\rho \pi_\theta(b \mid h)^{1-\rho} \\ &= (1-\rho)\left[\pi^\star_\alpha(a \mid h_{\mathrm{sf}}, h_{\mathrm{ref}})^\rho \pi_\theta(a \mid h)^{1-\rho} - \pi_\theta(a \mid h) S_{\alpha,\rho,\theta}\right]. \qquad (27)\end{aligned}$$

Next, differentiating the Rényi divergence gives

$$\frac{\partial D_\rho}{\partial z_\theta(a \mid h)} = \frac{1}{\rho - 1} \frac{1}{S_{\alpha,\rho,\theta}} \frac{\partial S_{\alpha,\rho,\theta}}{\partial z_\theta(a \mid h)}.$$

Substituting Equation (27),

$$\frac{\partial D_\rho}{\partial z_\theta(a \mid h)} = \frac{1-\rho}{\rho - 1}\left[\frac{\pi^\star_\alpha(a \mid h_{\mathrm{sf}}, h_{\mathrm{ref}})^\rho \pi_\theta(a \mid h)^{1-\rho}}{S_{\alpha,\rho,\theta}} - \pi_\theta(a \mid h)\right].$$

Since

$$\frac{1-\rho}{\rho-1} = -1,$$

we obtain

$$\begin{aligned}\frac{\partial D_\rho}{\partial z_\theta(a \mid h)} &= \pi_\theta(a \mid h) - \frac{\pi^\star_\alpha(a \mid h_{\mathrm{sf}}, h_{\mathrm{ref}})^\rho \pi_\theta(a \mid h)^{1-\rho}}{S_{\alpha,\rho,\theta}} \\ &= \pi_\theta(a \mid h) - \widetilde{\pi}_{\alpha,\rho,\theta}(a \mid h, h_{\mathrm{sf}}, h_{\mathrm{ref}}).\end{aligned}$$

Finally, by definition,

$$\pi^\star_\alpha(a \mid h_{\mathrm{sf}}, h_{\mathrm{ref}}) = \frac{\mathrm{sg}[\pi_{\bar{\theta}}(a \mid h_{\mathrm{sf}})]^\alpha \, \pi_{\mathrm{ref}}(a \mid h_{\mathrm{ref}})^{1-\alpha}}{Z_\alpha(h_{\mathrm{sf}}, h_{\mathrm{ref}})}.$$

Raising both sides to the power ρ gives

$$\pi^\star_\alpha(a \mid h_{\mathrm{sf}}, h_{\mathrm{ref}})^\rho = \frac{\mathrm{sg}[\pi_{\bar{\theta}}(a \mid h_{\mathrm{sf}})]^{\alpha\rho} \, \pi_{\mathrm{ref}}(a \mid h_{\mathrm{ref}})^{(1-\alpha)\rho}}{Z_\alpha(h_{\mathrm{sf}}, h_{\mathrm{ref}})^\rho}.$$

Then, the common factor $Z_\alpha(h_{\mathrm{sf}}, h_{\mathrm{ref}})^{-\rho}$ appears in both the numerator and denominator and therefore cancels. Hence,

$$\begin{aligned}\widetilde{\pi}_{\alpha,\rho,\theta}(a \mid h, h_{\mathrm{sf}}, h_{\mathrm{ref}}) &\propto \mathrm{sg}[\pi_{\bar{\theta}}(a \mid h_{\mathrm{sf}})]^{\alpha\rho} \\ &\quad \times \pi_{\mathrm{ref}}(a \mid h_{\mathrm{ref}})^{(1-\alpha)\rho} \pi_\theta(a \mid h)^{1-\rho}.\end{aligned}$$

□

C Qualitative Comparison of Generations

The examples below compare selected Qwen3-8B responses from a synthetic stress test, MMLU, and SciKnowEval. For each method, we present excerpts from Sample 0, preserving the original wording while removing formatting tags and bold emphasis. Omissions are marked by [...], and final answers are listed separately. Red text highlights errors or contradictions. Alongside the responses, we provide brief quality assessments and discuss what each example reveals.

CASE 1 **Correct calculation, wrong option** *Math*

Question. An item costs $89. It is discounted by 30%, then a 12% sales tax is applied to the discounted price. What is the final price?
A. $99.68 B. $69.78 C. $62.3 D. $71

Gold. B **Source.** Synthetic stress test

Method	Model response (excerpt)	Quality
SR-OPSD (w. ref.)	[. . .] A 30% discount reduces the price to $62.30. A 12% sales tax is then applied to this discounted price, which results in a final price of $62.30 + $7.48 = $69.78. [. . .] Therefore, the correct solution is B, which corresponds to the final price of $69.78. **Answer:** B	✓ **Consistent**
SR-OPSD (w/o ref.)	[. . .] - Final price after tax = $62.30 + $7.476 = $69.776, or approximately $69.78 [. . .] - Option B is not sufficiently aligned with the tax calculation. [. . .] - Only Option D accounts for both the discount and the tax, making it the most scientific and logical solution. [. . .] **Answer:** D	× **Wrong option**
SDPO (RKL)	The original price of the item is $89. When a 30% discount is applied, the price of the item is reduced by 30%. Therefore, the discounted price is $89 * (1 - 0.30) = $62.3. Then, a 12% sales tax is applied to the discounted price. Therefore, the final price is $62.3 * (1 + 0.12) = $69.78. **Answer:** D	× **Wrong option**

Quality analysis. All three responses obtain approximately $69.78, but only SR-OPSD with reference selects the corresponding option B. The no-reference response explicitly rejects B after computing that value; RKL also selects D. In both incorrect responses, the final answer does not follow from the calculation.

CASE 2 **Conflicting conclusions** *Physics*

Question. A certain element emits 1 alpha particle, and its products then emit 2 beta particles in succession. The atomic number of the resulting element is changed by
A. zero B. minus 1 C. minus 2 D. plus 1

Gold. A **Source.** MMLU / Conceptual Physics

Method	Model response (excerpt)	Quality
SR-OPSD (w. ref.)	[. . .] When we combine these two effects, the net change in the atomic number is: -2 (from the alpha particle) + 2 (from the beta particles) = 0. Thus, the atomic number of the resulting element is unchanged, and the correct answer is A. **Answer:** A	✓ **Consistent**
SR-OPSD (w/o ref.)	[. . .] - Option D (plus 1) reflects a gain in atomic number, which is not consistent with the actual sequence of events. [. . .] Only Option D correctly captures the net effect of the beta decay as a gain in atomic number, which is the only scientifically valid outcome in this context. [. . .] **Answer:** D	× **Contradiction**
SDPO (RKL)	The original element emits 1 alpha particle. When an alpha particle is emitted, the atomic number of the element is decreased by 1. The products then emit 2 beta particles in succession. When a beta particle is emitted, the atomic number of the element is increased by 1. Therefore, the atomic number of the resulting element is changed by a decrease of 1 and an increase of 1. The net change is a decrease of 1. **Answer:** D	× **Incorrect rule**

Quality analysis. The no-reference response first rejects D, then endorses it. RKL states an incorrect alpha-decay rule and concludes a decrease of 1, yet selects D (plus 1). SR-OPSD obtains zero and selects A.

CASE 3 **Contradictory answer selection** *Anatomy*

Question. In the spinal cord, motor neuron cell bodies are located in
A. Lamina I. B. Lamina II. C. Lamina V. D. Lamina IX
Gold. D **Source.** MMLU / Anatomy

Method	Model response (excerpt)	Quality
SR-OPSD (w. ref.)	The spinal cord is organized into distinct layers, with motor neuron cell bodies located in lamina IX. This region is known as the ventral horn and is responsible for controlling voluntary movements. [. . .] **Answer:** D	✓ Correct location
SR-OPSD (w/o ref.)	[. . .] - Lamina IX, however, is not a part of the spinal cord's functional motor system, as it lies outside the typical range of motor neuron localization. Given this, the correct solution must be Option A, Lamina I, which does not accurately reflect the known anatomical distribution of motor neurons. [. . .] **Answer:** A	× Contradiction
SDPO (RKL)	The spinal cord is divided into several laminae. Motor neuron cell bodies are located in the laminae. The laminae are divided into different layers. The motor neuron cell bodies are located in the lamina II. **Answer:** B	× Factual error

Quality analysis. The no-reference response calls A correct while stating that it does not match the anatomical distribution. RKL repeats general statements about laminae and identifies the wrong location. SR-OPSD identifies lamina IX and selects its matching option.

CASE 4 **Repeated reasoning, wrong answer** *Nutrition*

Question. How many kcal are there in one gram of ethanol?
A. 23.5 kJ or 5.8 kcal per g B. 19.7 kJ or 4.1 kcal per g C. 29.7 kJ or 7.1 kcal per g D. 36.5 kJ or 8.1 kcal per g
Gold. C **Source.** MMLU / Nutrition

Method	Model response (excerpt)	Quality
SR-OPSD (w. ref.)	[. . .] Among the given choices, Option C (29.7 kJ or 7.1 kcal per g) is the only one that accurately reflects the energy content of ethanol. It aligns with the known caloric value and is the most scientifically accurate answer. **Answer:** C	✓ Consistent
SR-OPSD (w/o ref.)	[. . .] This makes Option A, which suggests 23.5 kJ or 5.8 kcal per gram, the most scientifically and realistically accurate choice. [. . .] - Option C, in contrast, overestimates the energy output, making it less accurate in this context. [. . .] **Answer:** A	× Wrong value
SDPO (RKL)	[. . .] The heat of combustion of ethanol is approximately 29.7 kJ per gram. Therefore, the number of kcal in one gram of ethanol is determined by the heat of combustion. The heat of combustion of ethanol is approximately 29.7 kJ per gram. Therefore, the number of kcal in one gram of ethanol is determined by the heat of combustion. **Answer:** B	× Repetition

Quality analysis. RKL repeats the same two sentences about heat of combustion but selects B instead of the matching option C. The no-reference response endorses the wrong value in A. SR-OPSD links the stated energy value to C, whereas the repeated RKL reasoning adds no support for its final choice.

CASE 5 **Answer selection without a numerical derivation** *Chemistry*

Question. What is the molar weight (g/mol) of the molecule with the the IUPAC name '(2S)-2-[[(2S)-2-[[(2S)-2-[[(2S)-5-amino-2-[[(2S)-2-[[(2S)-2-[[(2S)-1-[(2S)-2-[[(2S)-2-[[(2S)-2-[[(2S)-1-[(2S)-2-amino-3-methylbutanoyl]pyrrolidine-2-carbonyl]amino]-3-(1H-indol-3-yl)propanoyl]amino]-4-methylsulfanylbutanoyl]amino]-4-carboxybutanoyl]pyrrolidine-2-carbonyl]amino]propanoyl]amino]-3-(4-hydroxyphenyl)propanoyl]amino]-5-oxopentanoyl]amino]-5-(diaminomethylideneamino)pentanoyl]amino]-3-phenylpropanoyl]amino]-4-methylpentanoic acid'?
A: 1517.700 B: 1589.900 C: 1536.800 D: 1490.700
Please reason step by step.

Gold. C **Source.** SciKnowEval / Chemistry

Method	Model response (excerpt)	Quality
SR-OPSD (w. ref.)	[. . .] The solution correctly concludes that Option C (1536.800) is the most accurate representation of the molecule's molar weight, based on its composition and the absence of any indication of a significantly different molecular structure. [. . .] **Answer:** C	✓ **Correct choice**
SR-OPSD (w/o ref.)	[. . .] - Option A (1517.700), with its lower and more moderate weight, is the most accurate solution in line with the molecule's actual composition. - Option C (1536.800), while not incorrect, is too high in its estimation of the molecule's mass. [. . .] **Answer:** A	× **Unsupported choice**
SDPO (RKL)	The molar weight of a molecule is determined by summing the atomic weights of all the atoms in the molecule. The given IUPAC name indicates that the molecule contains several hydrophilic and hydrophobic groups. These groups are consistent with a hydrophilic molecule. Therefore, the molar weight of the molecule is determined by summing the atomic weights of all the atoms in the molecule. The hydrophilic groups are more consistent with a hydrophilic molecule. Therefore, the molar weight of the molecule is determined by summing the atomic weights of all the atoms in the molecule. **Answer:** D	× **Off-topic repetition**

Quality analysis. SR-OPSD selects C, matching the answer key, but does not show the atom counts or mass calculation. The no-reference response prefers a lower mass without calculating it. RKL repeats a general calculation rule and discusses hydrophilicity without deriving the requested mass.

D Extra Experimental Details

D.1 Extra Results for Science Q&A

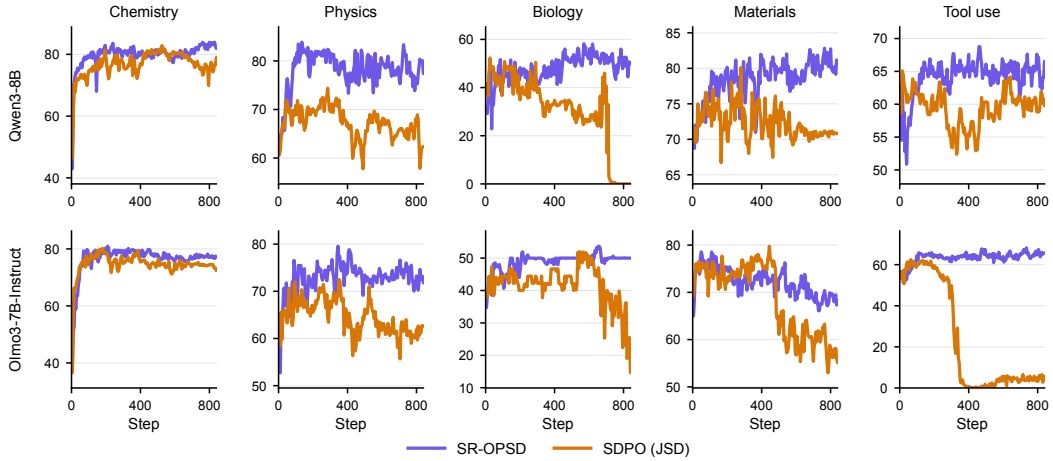


Figure 5: **Maj@16 test accuracy over training.** We report Maj@16 test accuracy across training steps on five reasoning-related benchmarks for Qwen3-8B and Olmo3-7B-Instruct.

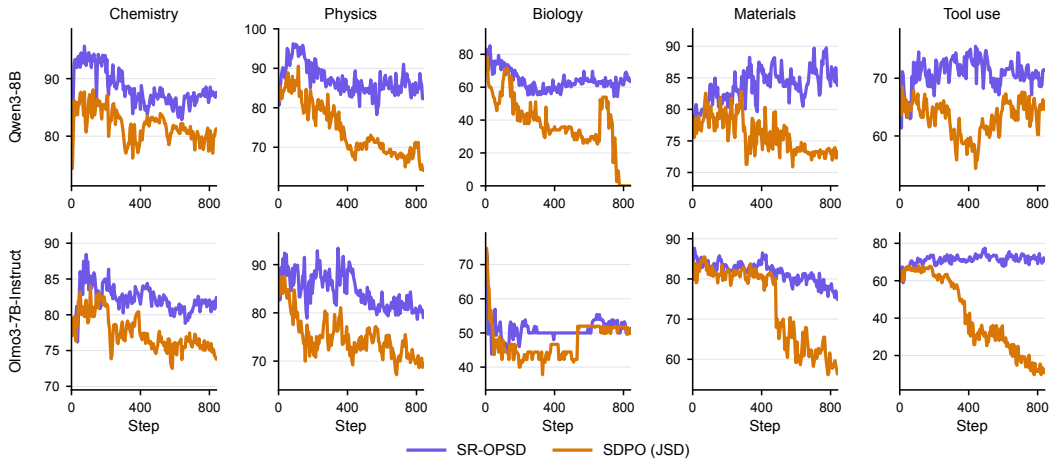


Figure 6: **Best@16 test accuracy over training.** We report Best@16 test accuracy across training steps on five reasoning-related benchmarks for Qwen3-8B and Olmo3-7B-Instruct.

D.2 Experimental Configurations

Table 6: Hyperparameters for the SciKnowEval Science Q&A and Tool-use experiments.

Parameter	GRPO	SDPO (JSD)	SR-OPSD (Ours)
General			
Benchmark	SciKnowEval	SciKnowEval	SciKnowEval
Models	Qwen3-8B and OLMo-3-7B-Instruct		
Primary evaluation metrics	Avg@16	Avg@16	Avg@16
Evaluation thinking mode	Disabled	Disabled	Disabled
Data, rollout, and evaluation			
Maximum prompt length	2048	2048	2048
Maximum training response length	8192	8192	8192
Question batch / PPO mini-batch size	32 / 8	32 / 32	32 / 32
Per-device batch / gradient accumulation	1 / 16	1 / 64	1 / 64
Training rollouts per question	8	8	8
Training temperature / top-p / top-k	1.0 / 1.0 / -1	1.0 / 1.0 / -1	1.0 / 1.0 / -1
Evaluation samples per question	16	16	16
Evaluation temperature / top-p / top-k	0.6 / 0.95 / -1	0.6 / 0.95 / -1	0.6 / 0.95 / -1
Method-specific objective			
Policy/distillation objective	GRPO	Jensen–Shannon	Forward Rényi
PPO clipping ϵ_{low} / ϵ_{high}	0.20 / 0.28	–	–
Rényi order ρ	–	–	0.9/0.95
Self-reference coefficient α	–	–	0.7/0.90
Frozen-reference anchoring	No	No	Yes
Reference policy	–	–	Frozen initial policy
Top-K distillation / tail bucket	–	100 / Yes	100 / Yes
Teacher EMA update rate	–	0.05	0.05
Distillation IS clip	–	2.0	2.0
Token-loss clip	–	–	–
Optimization			
Optimizer	AdamW	AdamW	AdamW
Learning rate	5×10^{-6}	1×10^{-5}	1×10^{-5}
Learning-rate schedule / warmup steps	constant / 10	constant / 10	constant / 10
Weight decay / gradient clipping norm	0.01 / 1.0	0.01 / 1.0	0.01 / 1.0

Table 7: Hyperparameters for the Qwen3-4B and Qwen3-8B mathematical-reasoning experiments.

Parameter	GRPO	SDPO (RKL)	OPSD (FKL)	SR-OPSD (Ours)
General				
Model	Qwen3-4B and Qwen3-8B			
Benchmark	AIME24, AIME25, HMMT25, AMC23, Minerva			
Primary evaluation metrics	Avg@16 and Pass@16			
Evaluation thinking mode	Enabled	Enabled	Enabled	Enabled
Data, rollout, and evaluation				
Maximum training response length	16384	16384	16384	16384
Question batch size	8	8	64	8
Training rollouts per question	8	8	1	8
Training rollout thinking mode	Disabled	Disabled	Disabled	Disabled
Training temperature / top-p / top-k	0.7 / 0.95 / 20			
Evaluation temperature / top-p / top-k (4B)	0.7 / 0.95 / 20			
Evaluation temperature / top-p / top-k (8B)	1.0 / 1.0 / -1			
Method-specific objective				
Policy/distillation objective	GRPO	Reverse KL	Forward KL	Forward Rényi
Rényi order ρ (4B / 8B)	–	–	–	0.7 / 0.9
Self-reference coefficient α (4B / 8B)	–	–	–	0.9 / 0.9
Frozen-reference anchoring	No	No	No	Yes
Reference policy	–	–	–	Frozen initial policy
Top-K distillation / tail bucket	–	100 / No	100 / No	100 / No
Teacher EMA update rate	–	0.05	–	0.05
Token-loss clip (4B / 8B)	–	0.05 / 0.05	0.05 / 0.06	0.05 / 0.05
GRPO policy-ratio clip	0.2	–	–	–
Optimization				
Optimizer	AdamW	AdamW	AdamW	AdamW
Learning rate	5×10^{-6}			
Learning-rate schedule / warmup steps	linear / 0			
Weight decay / gradient clipping norm	0 / 0.1			

Table 8: Hyperparameters for the Qwen3 model-scaling experiments on LiveCodeBench v6. Training uses eight rollouts per prompt, whereas test uses 16 sampled solutions per problem.

Parameter	GRPO	SDPO	SR-OPSD (Ours)
General			
Models	Qwen3-0.6B, Qwen3-1.7B, Qwen3-4B, Qwen3-8B		
Benchmark	LiveCodeBench v6		
Training and evaluation thinking mode	Disabled	Disabled	Disabled
Data, rollout, and test			
Maximum prompt / response length	2048 / 8192	2048 / 8192	2048 / 8192
Question batch / PPO mini-batch size	32 / 32	32 / 32	32 / 32
Training rollouts per question	8	8	8
Training temperature / top-p / top-k	1.0 / 1.0 / -1	1.0 / 1.0 / -1	1.0 / 1.0 / -1
Test frequency / samples per question	5 / 16	5 / 16	5 / 16
Test temperature / top-p / top-k	0.6 / 0.95 / -1	0.6 / 0.95 / -1	0.6 / 0.95 / -1
Method-specific objective			
Policy/distillation objective	GRPO	Reverse KL	Forward Rényi
Environment feedback in reprompting	–	Yes	Yes
Rényi order	–	–	0.95
Self-reference coefficient	–	–	0.90
Frozen-reference anchoring	–	No	Yes
Reference policy	–	–	Frozen initial policy
Top-K distillation / tail bucket	–	100 / Yes	100 / Yes
Teacher EMA update rate	–	0.05	0.05
Rollout IS correction / distillation IS clip	token / –	token / 2.0	token / 2.0
Optimization			
Optimizer	AdamW	AdamW	AdamW
Learning rate / schedule	1×10^{-5} / constant	1×10^{-5} / constant	1×10^{-5} / constant
Warmup steps	10	10	10
Weight decay / gradient clipping norm	0.01 / 1.0	0.01 / 1.0	0.01 / 1.0
PPO epochs / loss aggregation	1 / token mean	1 / token mean	1 / token mean

D.3 Extra Details of Ablation Study

Table 9: Ablation study on Science Q&A Physics. For SR-OPSD, α denotes the self-reference coefficient and ρ denotes the Rényi order. Each row reports the optimal Avg@16 within 10 compute hours.

Model	Method	Reference	Rényi		Physics			
			α	ρ	Avg@16	Pass@16	Maj@16	Format
Qwen3-8B	JSD	No	–	–	73.4	88.3	81.7	99.8
	JSD	Yes	0.9	–	77.2	84.6	79.4	99.8
	FKL	No	–	–	79.1	91.9	81.8	100.0
	FKL	Yes	0.9	–	76.3	82.3	78.3	100.0
	Forward Rényi	No	–	0.95	77.7	92.6	81.8	100.0
	SR-OPSD (Ours)	Yes	0.50	0.50	78.8	91.6	82.7	99.6
				0.70	78.8	92.6	81.2	99.5
				0.90	78.2	94.5	81.9	99.7
				0.95	78.4	92.0	80.9	99.1
			0.70	0.50	76.3	85.5	77.6	99.9
				0.70	82.0	94.3	85.0	99.9
				0.90	77.0	93.9	78.0	99.3
				0.95	80.7	94.2	81.6	99.5
			0.90	0.50	79.9	88.9	80.0	100.0
				0.70	82.7	91.0	84.2	100.0
				0.90	77.2	90.4	79.8	99.9
				0.95	81.2	86.8	82.4	100.0

On SciKnowEval Physics, SR-OPSD maintains relatively stable response lengths, whereas JSD and forward KL exhibit pronounced length growth in later training stages. Self-Reference anchoring does not prevent this growth under JSD, while the self-reference anchored forward-KL variant produces even longer responses than its unanchored counterpart. These trajectories complement the accuracy ablation, further indicating that the effect of self-referenced anchoring depends on the projection objective.

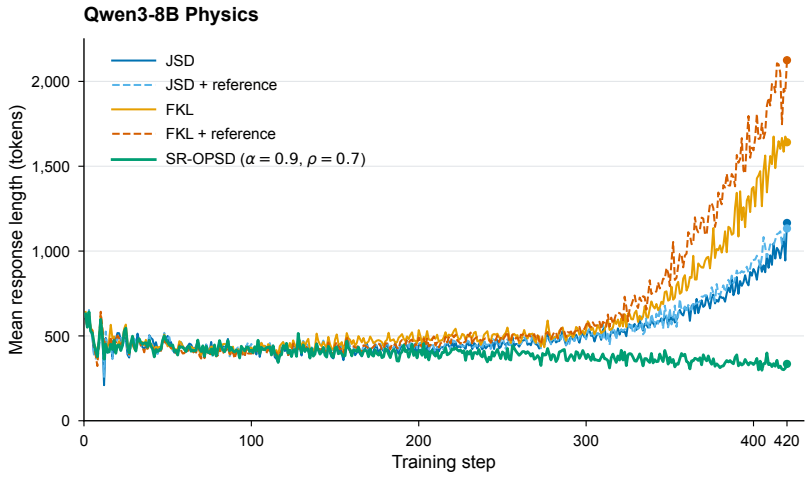


Figure 7: Investigation of response lengths.