REFERENCE-FREE ON-POLICY SELF-DISTILLATION VIA SELF-REFLECTION

Anonymous authors
Paper under double-blind review

ABSTRACT

On-policy self-distillation (OPSD) provides dense, token-level supervision through a teacher policy conditioned on privileged information (PI), typically in the form of external reference solutions that require additional annotation or generation. Besides this acquisition cost, mismatches in reasoning paths and styles between these references and the student's own rollouts can impair student learning, motivating the use of PI that better aligns with the student's own reasoning. Reflection offers a natural source of such PI: it revisits a student's responses to provide feedback tailored to the reasoning they contain. We therefore investigate whether conditioning the teacher on self-reflections generated by the shared policy can improve on-policy self-distillation by providing more effective token-level supervision. We propose **R**eference-**F**ree **S**elf-**D**istillation via Self-**Ref**lection (RF-SD), where a single policy serves as the student, reflector, and teacher. Given a question and a student rollout, the reflector is routed by the verifier outcome to summarize useful reasoning in correct trajectories or diagnose errors in incorrect ones. The resulting reflection replaces external references as the teacher's PI for token-level supervision on the student's own rollout. We further introduce RF-SD-e, which evolves the privileged information by optimizing the reflector using the correctness of reflection-conditioned teacher responses as rewards. Experiments on challenging mathematical reasoning benchmarks show that RF-SD outperforms strong baselines across different model scales, with further gains from reflector evolution.

1 INTRODUCTION

On-policy self-distillation (OPSD) trains a student model on its own sampled trajectories using token-level feedback from a teacher formed by conditioning the same model on additional privileged information (PI), e.g., an external reference solution, rather than relying on a separate teacher model (Zhao et al., 2026a; Hübotter et al., 2026). By matching the teacher's token distributions along student-generated rollouts, the student gains fine-grained guidance for reasoning that can complement learning from outcome-based rewards (Liu et al., 2026). Constructing a reliable teacher that provides informative yet transferable guidance is therefore a key design challenge in OPSD.

Existing methods commonly construct PI from externally annotated or generated reference solutions (Zhao et al., 2026a; Liu et al., 2026), ground-truth answers (Yang et al., 2026a), or correct peer rollouts (Tan & Hong, 2026; Zhao et al., 2026b). Reference solutions incur additional acquisition costs, while peer-based context is unavailable for the uniformly incorrect group (He et al., 2026). Moreover, a reference can follow a reasoning path or style different from the student's own rollout, encouraging imitation over error-specific correction (Zhao et al., 2026b). Recent evidence highlights this distinction: in long-budget evaluations of thinking models, reference-conditioned teachers can retain high accuracy while their distilled students degrade (Kaur et al., 2026). The performance gain obtained by conditioning a teacher on a correct reference does not necessarily transfer to the student. Thus, reference-conditioned OPSD may provide supervision shaped by the reference rather than tailored to the student's own rollout.

These limitations motivate privileged information grounded in the reasoning the student actually produces. Self-reflection offers a natural source: it revisits a completed response to identify useful steps, missing justifications, and potential errors. Prior work has used self-generated feedback to refine responses at test time (Madaan et al., 2023; Shinn et al., 2023). We instead ask whether

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.

self-reflection can serve as reference-free PI for OPSD, enabling the teacher to provide supervision tailored to the student's current rollout.

We propose Reference-Free Self-Distillation via Self-Reflection (RF-SD), in which a single policy serves three roles: student, reflector, and teacher. The *student* samples a rollout, the *reflector* analyzes the question and rollout, and the *teacher* utilizes the reflection as PI to provide token-level supervision along the same rollout. To construct the rollout-grounded PI, the reflector must distinguish what to preserve from what to correct: an error-seeking prompt can produce misleading criticism of correct work, while an overly broad prompt can yield generic checks. RF-SD therefore uses the verifier result to route the reflector toward summarizing useful reasoning in correct rollouts or diagnosing errors in failed ones. Apart from correctness-guided routing, neither the reflector nor the teacher is exposed to an expert reference solution, or a correct peer rollout. The training objective of RF-SD combines reflection-conditioned self-distillation with outcome-based GRPO (Shao et al., 2024), pairing token-level guidance along the student's rollout with verifier feedback on the final answer.

Building on the student-reflector-teacher loop, we extend RF-SD to RF-SD-e by optimizing the reflector so that the teacher's PI can evolve alongside training. For a fixed student rollout, the reflector samples candidate reflections and rewards them using the correctness of the corresponding reflection-conditioned teacher answers as a proxy for reflection quality. The reflector's GRPO update is applied only to reflector-specific parameters and kept separate from student training to preserve role separation. The reflections generated by the improved reflector serve as more effective PI for the teacher, strengthening its supervision of the student.

To verify the effectiveness of our approach, we evaluate RF-SD and RF-SD-e on Qwen3 backbones spanning 1.7B to 8B across challenging mathematical reasoning benchmarks. Under the same training and evaluation settings, RF-SD outperforms GRPO, OPSD, RLSD, and SDPG on average at every scale, without relying on external reference solutions. Ablations support the verifier-routed reflection design, while online student-teacher diagnostics show higher incorrect-to-correct and lower correct-to-incorrect transition rates with RF-SD than with a generic reflector, which reflects in a general manner regardless of rollout correctness. RF-SD-e further improves average student performance, indicating that the teacher's reflection-based PI can itself be improved through learning.

2 PRELIMINARIES

Reinforcement learning with verifiable rewards. For each question, a behavior policy $\pi_{\theta_{\text{old}}}$ samples N rollouts, each scored by an outcome verifier. For a sampled question-rollout pair (x_i, y_i), GRPO (Shao et al., 2024) computes an advantage $\hat{A}_i$ by normalizing its reward within the question group and minimizes

$$\mathcal{L}_{\text{GRPO}}(\theta) = -\mathbb{E}_{i,t}\Big[\min\Big(w_{i,t}(\theta)\hat{A}_i,\ \text{clip}(w_{i,t}(\theta), 1-\varepsilon_{\text{low}}, 1+\varepsilon_{\text{high}})\hat{A}_i\Big)\Big], \tag{1}$$

where $w_{i,t}(\theta) = \pi_\theta(y_{i,t} \mid x_i, y_{i,<t}) / \pi_{\theta_{\text{old}}}(y_{i,t} \mid x_i, y_{i,<t})$ is the policy probability ratio and $\mathbb{E}_{i,t}$ averages over response tokens in the sampled batch. Each token in a rollout receives the same outcome advantage, without separate feedback on individual reasoning steps.

On-policy self-distillation. OPSD obtains dense token-level feedback from a teacher that shares the student's parameters but receives additional privileged information c_i (Zhao et al., 2026a). At each student prefix $y_{i,<t}$, let $p_{i,t} = \pi_\theta(\cdot \mid x_i, y_{i,<t})$ and $q_{i,t} = \pi_\theta(\cdot \mid x_i, c_i, y_{i,<t})$ denote the student and teacher token distributions. Using reverse KL and taking gradients only through the student distribution, the distillation objective is

$$\mathcal{L}_{\text{OPSD}}(\theta) = \mathbb{E}_{i,t}[D_{\text{KL}}(p_{i,t} \,\|\, q_{i,t})]. \tag{2}$$

This distillation signal can serve as an implicit process reward alongside outcome-based rewards in RLVR, as in SDPG (Liu et al., 2026).

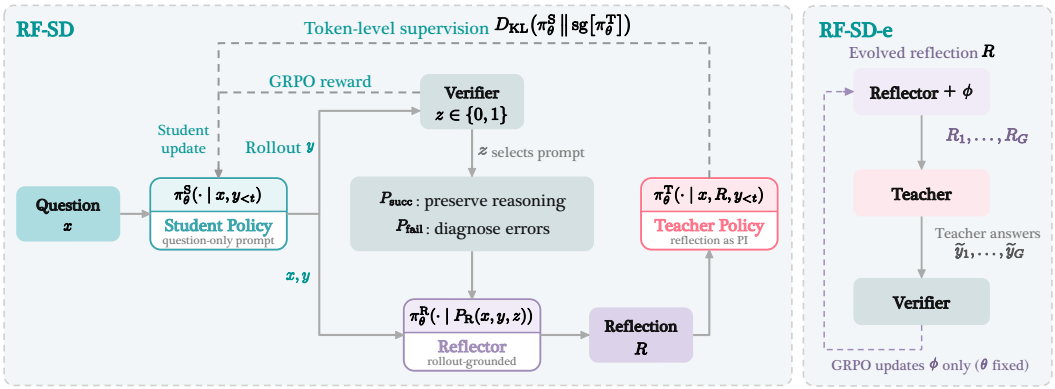

Figure 1: **Overview of RF-SD and RF-SD-e.** RF-SD uses self-reflection as reference-free PI for on-policy self-distillation. Ground-truth verification determines whether the reflection focuses on preserving useful reasoning or diagnosing errors, without exposing the answer to the reflector or teacher. RF-SD-e further optimizes the reflector using teacher-answer correctness to improve the reflections used as teacher privileged information.

3 RF-SD: Reference-Free Self-Distillation via Self-Reflection

3.1 Bridging Student and Teacher through Self-Reflection

RF-SD builds on on-policy self-distillation, using a reflection of the student's own rollout as the teacher's privileged information. A single policy serves as the *student* π_θ^{S}, *reflector* π_θ^{R}, and *teacher* π_θ^{T}: the student samples a rollout for a question; the reflector analyzes the question and rollout to produce a reflection; and the teacher conditions on this reflection to grade the tokens of the same student trajectory. These roles share model parameters θ but receive different role-specific prompts.

For question x_i, the student samples a complete response y_i, whose individual tokens are denoted by $y_{i,t}$. The response and its verifier label are

$$y_i \sim \pi_{\theta_{\mathrm{old}}}^{\mathrm{S}}(\cdot \mid P_{\mathrm{S}}(x_i)), \qquad z_i = \mathrm{Ver}(x_i, y_i) \in \{0, 1\}, \tag{3}$$

where $P_{\mathrm{S}}(x_i)$ denotes the student prompt containing the question x_i without additional privileged information. Here $z_i = 1$ denotes verifier acceptance and an incorrect or invalid answer receives $z_i = 0$. The label supplies the outcome reward for RLVR and selects the reflection route. The reflector and teacher receive no reference solution, or correct peer rollout. RF-SD purely relies on its own on-policy reflection as the privileged information to mitigate the teacher-student discrepancy. Fig. 1 summarizes the framework.

3.2 Verifier-routed Dual Reflection

RF-SD employs its self-played reflector to give per-rollout feedback. However, self-critique can be superficial or mistakenly revise an already correct answer (Tang et al., 2025; Huang et al., 2024). RF-SD therefore *detaches the correctness justification from the generic reflection process* and uses the verifier result z_i from Eq. (3) to select a success- or failure-specific reflection prompt:

$$P_{\mathrm{R}}(x_i, y_i, z_i) = \begin{cases} P_{\mathrm{succ}}(x_i, y_i), & z_i = 1, \\ P_{\mathrm{fail}}(x_i, y_i), & z_i = 0. \end{cases} \tag{4}$$

Here P_{succ} and P_{fail} are fixed prompts for reflecting on successful and failed responses, respectively, both including the question x_i and student rollout y_i. Conditioned on the selected prompt, the reflector generates a reflection:

$$R_i \sim \pi_{\theta_{\mathrm{old}}}^{\mathrm{R}}(\cdot \mid P_{\mathrm{R}}(x_i, y_i, z_i)). \tag{5}$$

Thus z_i determines how the reflector analyzes (x_i, y_i), and the resulting R_i supplies the teacher's privileged information. We show the key excerpts of the actual reflector's prompts below, with full details in Appendix B:

Reflection from Success P_{succ} $(z_i = 1)$	**Reflection from Failure** P_{fail} $(z_i = 0)$
"Extract the most valuable reasoning steps, identities, constraints, or checks that are explicitly present in the trajectory."	"Act as the critic for a failed or incomplete trajectory. Locate the first or highest-impact mathematical mistake, [...] visible in the trajectory."

Reflection from Success. For accepted student responses, the reflector summarizes useful reasoning and notes where further justification is needed, rather than reassessing correctness. The resulting teacher policy $\pi_\theta^{\mathrm{T}}(\cdot \mid P_{\mathrm{T}}(x_i, R_i), y_{i,<t})$ can thus favor tokens consistent with the useful reasoning identified in the student's response.

Reflection from Failure. For incorrect or invalid rollouts, the reflector is prompted to identify a high-impact error or specify a step to verify when no clear error can be located. Conditioned on this error-focused reflection, the teacher $\pi_\theta^{\mathrm{T}}(\cdot \mid P_{\mathrm{T}}(x_i, R_i), y_{i,<t})$ is shaped to provide error-focused token-level supervision associated with the identified mistake via the distillation objective.

Examples of generated reflections, including an unsuccessful case, are provided in Appendix G.

3.3 Reflection-conditioned Self-distillation

The teacher prompt $P_{\mathrm{T}}(x_i, R_i)$ combines the question x_i with the routed reflection R_i as privileged information, while the student retains its question-only prompt $P_{\mathrm{S}}(x_i)$. The teacher grades the student's existing rollout token by token, using the student's preceding tokens $y_{i,<t}$ rather than generating a replacement solution.

RF-SD internalizes this guidance by combining distillation with outcome-based GRPO (Liu et al., 2026). Instantiating Eq. (2) with $c_i = R_i$, the student objective is

$$\mathcal{L}_{\mathrm{S}}(\theta) = \mathcal{L}_{\mathrm{GRPO}}(\theta) + \beta\, \mathbb{E}_{i,t}\Big[D_{\mathrm{KL}}\Big(\pi_\theta^{\mathrm{S}}(\cdot \mid P_{\mathrm{S}}(x_i), y_{i,<t}) \,\Big\|\, \mathrm{sg}\big[\pi_\theta^{\mathrm{T}}(\cdot \mid P_{\mathrm{T}}(x_i, R_i), y_{i,<t})\big]\Big)\Big], \quad (6)$$

where $y_i \sim \pi_{\theta_{\mathrm{old}}}^{\mathrm{S}}(\cdot \mid P_{\mathrm{S}}(x_i))$ is sampled from the behavior student policy. The reverse KL is computed over the full vocabulary at each response position and supplies an implicit process reward alongside GRPO's outcome reward. Here sg denotes stop-gradient, so distillation gradients flow only through the student branch. Both success and failure routes contribute to this term, regardless of the sign of the rollout's GRPO advantage.

The coefficient β controls the strength of distillation. Its value and other training settings are given in Appendix C. At test time, the student answers directly without the reflector or teacher.

Reducing reference-induced mismatch. Self-reflection analyzes the student's own rollout rather than a separate solution. To examine how this may reduce teacher-student mismatch, we consider an idealized model in which the two teacher families share the same mean logits, but reference-specific deviations are scaled down in the reflection-conditioned teacher. The following proposition supports our design: under the assumed logit model, reflection-based PI can reduce reference-induced teacher-student mismatch, as measured by expected reverse KL.

Proposition 1 (Reverse-KL reduction under logit attenuation)**.** *Fix the model, student instance, and response position, and let p be the student distribution. Let $v_\star$ be a fixed mean logit vector and ξ a zero-mean random deviation associated with the reference context. Define*

$$q_s(\xi) = \mathrm{softmax}(v_\star + s\xi), \qquad \mathbb{E}\xi = 0, \quad s \in [0, 1]. \quad (7)$$

Assume finite logits on a finite vocabulary and $\mathbb{E}\|\xi\| < \infty$. For a fixed $\lambda \in [0, 1]$, suppose the reference-conditioned teacher q_C and the reflection-conditioned teacher q_R over valid reflections have the same distributions as $q_1(\xi)$ and $q_\lambda(\xi)$, respectively. Writing $q_\star = \mathrm{softmax}(v_\star)$, we have

$$\mathbb{E}_C D_{\mathrm{KL}}(p\|q_C) - \mathbb{E}_R D_{\mathrm{KL}}(p\|q_R) \geq (1-\lambda)\mathbb{E}_\xi D_{\mathrm{KL}}(q_\star\|q_1(\xi)) \geq 0. \quad (8)$$

The reduction is strict if $\lambda < 1$ and ξ changes a token-to-token logit difference with positive probability.

Here λ scales reference-specific variation, and $q_\star$ may differ from the student distribution. Appendix H provides the proof and gradient analysis.

3.4 Co-evolving Student and Teacher

On-policy self-distillation based on generated reflection grounds the teacher's privileged information in the student's rollout. This student-reflector-teacher loop also makes the reflector itself an optimization target, allowing the privileged information, the reflection, to evolve alongside the student. We extend RF-SD to RF-SD-e as an option, using teacher's correctness as the proxy to reward the reflector. The evolved reflections can strengthen the teacher's distillation guidance and, in turn, the student's learning. To preserve role separation, the reflector's auxiliary GRPO update optimizes only its LoRA parameters ϕ, with the shared backbone θ held fixed during that update. Student training continues to update θ with the adapter disabled.

Learning from teacher correctness. For each student rollout group, the reflector samples G candidate reflections for one selected rollout (x, y, z) using the routed prompt from Eq. (4). Each reflection conditions a fresh teacher answer, whose correctness rewards that reflection. For brevity, we omit the rollout index; θ, ϕ denote pre-update parameters in these sampling equations:

$$\begin{aligned} R_g &\sim \pi^{\mathrm{R}}_{\theta,\phi}(\cdot \mid P_{\mathrm{R}}(x, y, z)), \\ \widetilde{y}_g &\sim \pi^{\mathrm{T}}_{\theta}(\cdot \mid P_{\mathrm{T}}(x, R_g)), \\ r^{\mathrm{R}}_g &= \mathrm{Ver}(x, \widetilde{y}_g) \in \{0, 1\}, \qquad g = 1, \ldots, G. \end{aligned} \tag{9}$$

Unlike the student's update in Section 3.3, this teacher pass generates a complete answer solely to provide rewards for the reflector's update. The resulting rewards r^{R} are normalized within each fixed-rollout group to compute $\hat{A}^{\mathrm{R}}$, the advantage for the generated reflection tokens. The reflector uses the same clipped GRPO surrogate as in Eq. (1), applied to reflection tokens with reflector-specific clipping and aggregation:

$$\mathcal{L}_{\mathrm{R}}(\phi) = \mathcal{L}_{\mathrm{GRPO}}(\phi). \tag{10}$$

Alternating updates. Each iteration first updates the reflector by minimizing $\mathcal{L}_{\mathrm{R}}(\phi)$ with θ fixed, using teacher-correctness rewards from samples collected before either update. The student then updates the shared backbone by minimizing $\mathcal{L}_{\mathrm{S}}(\theta)$ with the adapter disabled for both student and teacher, reusing the collected reflections as teacher context. The complete training procedure, including optional reflector evolution, is summarized in Algorithm 1 in Appendix A.

4 Experiments

4.1 Experimental Setup

Models and data. We train Qwen3-1.7B, Qwen3-4B, and Qwen3-8B (Yang et al., 2025) on the English subset from the DAPO math dataset (Yu et al., 2026), with around 14k samples. OPSD and SDPG additionally use verified expert reference solutions, sourced primarily from a public dataset generated by Gemini 3 Flash (Appendix C).

Training details. For RF-SD and baselines, we train for 300 steps with 128 queries per step and $N = 8$ rollouts per query group. The maximum response length is set to 4096 tokens. Unless otherwise stated, the student, teacher, and reflector operate in non-thinking mode, which is also retained for student evaluation. We follow DAPO-style asymmetric ratio clipping with $\varepsilon_{\mathrm{low}} = 0.20$ and $\varepsilon_{\mathrm{high}} = 0.28$. The distillation coefficient is configured as $\beta = 10^{-3}$ unless otherwise specified. Ablations on the distillation objective and coefficient are reported in Appendix F. Further optimization and sampling settings are provided in Appendix C.

Evaluation and baselines. We evaluate on AIME 2024, AIME 2025, HMMT 2025 and AMC 2023 using the same decoding settings for all methods. We report Avg@12 as *Final / Best*: final denotes step 300, and best denotes the checkpoint with the highest average within 300 steps. We compare against four baselines. **GRPO** (Shao et al., 2024) optimizes verifier rewards with a frozen-reference KL regularizer. **RLSD** (Yang et al., 2026a) uses an answer-conditioned teacher to reweight the policy advantage. **OPSD** (Zhao et al., 2026a) performs on-policy distillation from a reference-solution-conditioned teacher, without a verifier-reward objective. **SDPG** (Liu et al., 2026) combines GRPO with self-distillation from a teacher conditioned on the verified reference solution.

Table 1: **Main results.** Avg@12 (%), averaged over the four benchmarks. *Final* reports step 300, while *Best* selects each run's checkpoint with the highest four-benchmark average through step 300. The best result is shown in **bold** and the second-best is underlined.

	AIME24		AIME25		HMMT25		AMC23		Average	
Method	Final	Best	Final	Best	Final	Best	Final	Best	Final	Best
Qwen3-1.7B										
Base	12.78	12.78	7.78	7.78	5.28	5.28	45.21	45.21	17.76	17.76
GRPO	27.78	27.78	19.17	21.67	8.06	13.06	66.46	67.29	30.36	32.45
OPSD	18.06	19.72	13.33	18.61	6.94	9.17	54.17	57.29	23.12	26.20
RLSD	23.61	**32.22**	**26.11**	23.61	13.06	**16.94**	64.79	71.25	31.89	36.01
SDPG	26.39	24.17	21.39	**26.39**	12.50	13.06	70.42	67.92	32.67	32.88
RF-SD	**31.94**	31.94	25.00	25.00	**16.11**	16.11	**75.63**	**75.63**	**37.17**	**37.17**
Qwen3-4B										
Base	21.39	21.39	18.89	18.89	13.06	13.06	66.88	66.88	30.05	30.05
GRPO	36.39	40.00	33.06	37.50	20.28	20.28	85.63	85.83	43.84	45.90
OPSD	34.44	41.94	27.50	36.94	17.78	23.61	72.50	77.71	38.05	45.05
RLSD	40.28	45.28	35.00	37.22	16.11	16.11	89.17	88.54	45.14	46.79
SDPG	40.00	41.67	34.72	40.28	16.67	17.78	85.83	88.96	44.31	47.17
RF-SD	**42.50**	**47.50**	**39.44**	**40.56**	**23.06**	**24.44**	**90.83**	**89.79**	**48.96**	**50.57**
Qwen3-8B										
Base	25.00	25.00	17.22	17.22	12.22	12.22	67.92	67.92	30.59	30.59
GRPO	38.89	46.67	31.39	31.67	16.39	15.28	85.21	90.21	42.97	45.95
OPSD	36.39	44.44	32.78	36.11	21.39	20.00	70.42	78.96	40.25	44.88
RLSD	48.89	**51.39**	**37.78**	41.39	16.39	18.89	89.58	**91.88**	48.16	50.89
SDPG	45.83	48.61	35.00	37.78	18.06	**26.67**	90.00	86.88	47.22	49.98
RF-SD	**52.78**	50.28	37.22	**43.61**	**25.00**	23.06	**90.63**	90.83	**51.41**	**51.94**

4.2 Main Results

Among the reported results, RF-SD achieves the highest final and best four-benchmark averages at all three model scales (Table 1). Its final/best gains over the strongest baseline are 4.50/1.16, 3.82/3.40, and 3.25/1.06 points on Qwen3-1.7B, 4B, and 8B, respectively, without requiring the verified reference solutions available to SDPG. On the most challenging HMMT 2025 (Balunovic et al., 2026), RF-SD achieves the highest final score at every scale, exceeding the strongest baseline by 2.78 to 3.61 points. On both AIME 2024 and AIME 2025, its final and best scores remain competitive with answer- and solution-conditioned baselines across all three model scales. These cross-scale results support reflection as an effective source of guidance beyond terminal verifier rewards on challenging contest problems with different levels of difficulty from HMMT 2025 to AMC 2023's broader high-school problems. RLSD also performs well using the answer-conditioned teacher to reweight verifier-driven advantages rather than directly matching its distribution, a design intended to limit privileged-information leakage (Yang et al., 2026a). SDPG's gains over GRPO across all three scales further support distillation as a useful complement to outcome-based rewards.

Training and evaluation dynamics. Fig. 2 shows the training and evaluation dynamics of RF-SD and the baselines. On Qwen3-4B, RF-SD's verifier reward exhibits rapid early growth followed by gradual consistent gains. RF-SD shares a similar reward trajectory with SDPG but exhibits higher token entropy, particularly in the second half of training. Meanwhile, RF-SD maintains strong benchmark performance while producing the shortest responses on average over training, with its per-step mean falling to 1,686 tokens at step 259. At step 300, its responses average 2,091 tokens, compared with 2,123 for GRPO, 2,169 for RLSD, and 2,219 for SDPG. Full training and evaluation dynamics are provided in Appendix E, showing that RF-SD leads the four-benchmark average from step 240 to 300 across all three model scales.

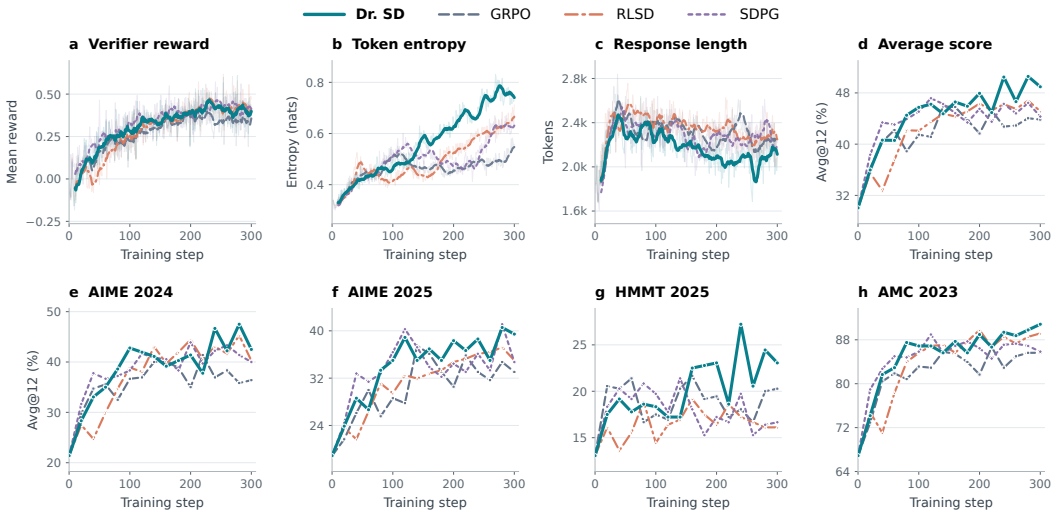

Figure 2: **Training and evaluation dynamics on Qwen3-4B.** Top row: (a) verifier reward, (b) token entropy, (c) mean response length, and (d) Avg@12 over the four benchmarks. Bottom row: evaluation dynamics on (e) AIME 2024, (f) AIME 2025, (g) HMMT 2025, and (h) AMC 2023.

Table 2: **Reflection-routing ablation** on Qwen3-4B. We report the average result of four benchmarks (Avg@12) to investigate the effect of reflection routing. The success-only and failure-only variants of RF-SD retain one reflection route and disable the other, whereas the generic reflector uses a shared reflection prompt regardless of rollout correctness.

Variant	Correct-rollout reflection	Failed-rollout reflection	Average (Final / Best)
RF-SD	success reflection	failure reflection	**48.96 / 50.57**
RF-SD (Failure only)	-	failure reflection	46.74 / 48.80
RF-SD (Success only)	success reflection	-	45.75 / 50.31
Single generic reflector	shared prompt	shared prompt	46.04 / 47.38
Single generic reflector (hinted)	shared prompt (hinted)	shared prompt (hinted)	47.01 / 49.13

4.3 Understanding Dual Reflection

To examine how reflection design affects self-distillation, we ablate verifier-guided routing and reference access on Qwen3-4B. Specifically, we compare the full design with single-route and shared-prompt variants (Table 2).

Why two reflection routes? A single generic reflector, using a shared prompt regardless of rollout correctness, lowers the RF-SD final/best average performance from 48.96/50.57 to 46.04/47.38 (Table 2). The generic reflector receives the question and student rollout without the verifier status, whereas the full method uses the verifier outcome to select a specialized reflection prompt. The generic reflection can offer broad checks rather than step-specific feedback. The hinted variant additionally receives an answer and verified reference solution, improving its performance but without closing the gap to dual reflection (Table 2). The dual-reflection design thus combines correctness-specific guidance with reference-free privileged information.

Complementarity of the two routes. RF-SD (Failure only) and RF-SD (Success only) restrict reflection-based distillation to failed and correct rollouts, respectively. In both variants, rollouts on the disabled route contribute no distillation loss but remain in the token-count denominator and the GRPO objective. The two variants achieve final/best averages of 46.74/48.80 and 45.75/50.31, respectively. Although RF-SD (Success only) approaches the full method's best score, using both routes gives the strongest final and best results. These results suggest that preserving useful reasoning in successful rollouts and diagnosing errors in failed ones provide complementary guidance for self-distillation. The success route also broadens distillation coverage, as failed rollouts account for

Table 3: **Online student-teacher diagnostics on Qwen3-4B.** Results are aggregated over 300 training steps. *Student* and *Teacher* report answer accuracy, while *Gap* reports the teacher's accuracy improvement over the student. *Rescue* and *Regression* are the percentages of all rollouts that transition from incorrect to correct and from correct to incorrect, respectively.

Reflection	Student	Teacher	Gap	Rescue	Regression
Single generic reflector	64.58	65.45	+0.87	5.91	5.04
RF-SD	65.26	**69.35**	+4.09	**6.56**	**2.47**

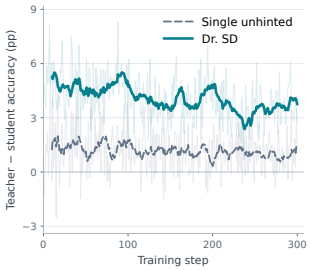

Figure 3: Teacher-student accuracy gap on Qwen3-4B.

Table 4: **The results of RF-SD-e with reflector optimization** across model scales. Average results of the four benchmarks are reported (Avg@12).

Model	RF-SD (Final / Best)	RF-SD-e (Final / Best)	Δ (Final / Best)
Qwen3-1.7B	37.17 / 37.17	**37.95 / 37.95**	+0.78 / +0.78
Qwen3-4B	48.96 / 50.57	**49.74 / 51.75**	+0.78 / +1.18
Qwen3-8B	51.41 / 51.94	**52.67 / 53.04**	+1.27 / +1.09

only about one third of student rollouts in the entire training. Both single-route variants achieve higher best-checkpoint averages than the generic reflector, which applies reflection to all rollouts, suggesting that supervision from rollout-relevant, informative teacher context can be more valuable than generic feedback with full coverage.

Teacher diagnostics: error correction and answer preservation. To examine how reflection affects teacher behavior, we additionally sample a complete reflection-conditioned teacher answer for each student rollout. As shown in Table 3, with a single generic reflector, the teacher corrects failed student answers on 5.91% of all rollouts but changes correct answers to incorrect ones on another 5.04%, almost canceling the benefit of correction. RF-SD increases the rescue rate to 6.56% while reducing regression to 2.47%, improving both error correction and the preservation of successful answers. These results support the complementary roles of the two reflection routes: helping the teacher address mistakes in failed rollouts and retain useful reasoning in successful ones. The teacher-student accuracy gap over training is shown in Fig. 3.

4.4 Reflector Evolution

Optimizing the reflector in RF-SD-e provides additional gains beyond the base design as shown in Table 4 across different model scales. Fig. 4 includes the reward dynamics of RF-SD and RF-SD-e of student policy on Qwen3-4B, showing that evolved reflection of RF-SD-e provides consistent improvement over RF-SD, and the largest gaps occur at around step 100 and 300. Reward dynamics comparison between RF-SD and RF-SD-e for all three scales are provided in Fig. 5 of Appendix D. These gains suggest that teacher-answer correctness can guide reflector optimization toward more effective PI for the teacher.

Role isolation during RF-SD-e's training. During the auxiliary reflector update, RF-SD-e optimizes only the reflector-specific LoRA parameters while holding the shared backbone fixed. The student continues to update the backbone through its own training objective. In Table 5, we also report the results of updating the reflector and student with shared full model. On Qwen3-4B, this variant achieves the lowest final/best score 44.24/44.44, compared with 48.96/50.57 for the base RF-SD and 49.74/51.75 for reflector-specific updates of RF-SD-e. This comparison supports separating reflector optimization from student updates.

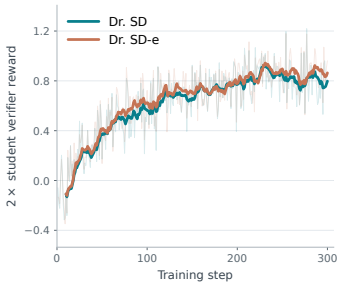

Figure 4: Training reward dynamics comparison between RF-SD and RF-SD-e on Qwen3-4B.

Table 5: **Role isolation ablation** on Qwen3-4B for reflector updates. Both reflector-update variants of RF-SD-e use the same teacher-correctness reward to update either the shared model or a reflector-specific adapter.

Method	Reflector update	Average (Final / Best)
RF-SD	-	48.96 / 50.57
RF-SD-e	Shared parameters	44.24 / 44.44
RF-SD-e	Reflector-specific LoRA	**49.74 / 51.75**

5 Related Work

On-policy self-distillation. On-policy distillation supervises student-generated rollouts (Agarwal et al., 2024; Li et al., 2026); OPSD constructs the teacher through privileged conditioning of a shared policy (Zhao et al., 2026a). SDPG combines full-vocabulary reverse-KL distillation with GRPO, whereas RLSD reweights policy advantages (Liu et al., 2026; Yang et al., 2026a). Other work refines teacher signals and credit assignment (Sun et al., 2026; Yang et al., 2026b; Tian et al., 2026; Wu et al., 2026; Nguyen et al., 2026), or examines harmful effects of privileged supervision (Kaur et al., 2026; Kim et al., 2026; Shen et al., 2026). We focus on constructing and improving the teacher's PI.

Privileged information for reasoning. PI sources include verified solutions, ground-truth answers, environment feedback, and correct peers (Zhao et al., 2026a; Liu et al., 2026; Yang et al., 2026a; Hübotter et al., 2026; Tan & Hong, 2026; Huang et al., 2026). SD-Zero first trains a reviser, then distills its outcome-conditioned supervision (He et al., 2026); ROSD uses correct peers for failed rollouts and ground-truth answers in both reflection prompts (Zhao et al., 2026b); AMR-SD converts diagnostic hints into token-level advantage modulation (Wei et al., 2026). RF-SD generates reflection from the question and current rollout without exposing reference solutions, ground-truth answers, or correct peers to the reflector or teacher. The verifier selects the reflection pattern, and reflection remains available even when all sampled rollouts fail.

Reflection and learned critique. Self-Refine and Reflexion use feedback for test-time refinement (Madaan et al., 2023; Shinn et al., 2023). RefCritic trains critics using revision feedback (Tang et al., 2025), while ICRL internalizes critique-guided gains into a critique-free solver (Lin et al., 2026). SCoRe optimizes multi-turn self-correction (Kumar et al., 2025). RF-SD instead distills reflection-conditioned teacher distributions on the original rollout. In RF-SD-e, teacher-answer correctness rewards reflector-specific updates.

6 Conclusion

We introduced RF-SD, which uses verifier-routed self-reflection as reference-free privileged information for on-policy self-distillation. Success- and failure-focused reflections ground the teacher's token-level supervision in the student's own reasoning, without exposing ground-truth answers or reference solutions to either the reflector or the teacher. Across four mathematical reasoning benchmarks and Qwen3 backbones spanning 1.7B to 8B, RF-SD consistently improves average performance over strong reinforcement learning and self-distillation baselines. Routing ablations support the complementary roles of preserving useful reasoning and diagnosing errors. Our RF-SD-e extension further improves student performance through reflector-specific optimization, allowing the teacher's reflection-based PI to evolve during training. Together, these findings show that useful privileged information need not come from an external solution: it can be generated from the student's own rollout and improved through learning to better support self-distillation.

AI USE STATEMENT

In this work, we used generative AI tools for writing refinement and assistance in coding. We have not used generative AI tools for conceptual frameworks, formulating or refining research hypotheses, designing experiments, interpreting results. And cleaning and reformating dataset, supporting qualitative and thematic data analysis are not applicable to this work. Additionally, we used generative AI tools for grammar, wording, sentence flow, readability, and brainstorming. We have reviewed all AI-assisted work. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

REPRODUCIBILITY STATEMENT

The main text describes the RF-SD framework, training objectives, and experimental setup. Appendix A provides the training pseudocode and detailed objectives. Appendix B documents the reflector and teacher prompts, and Appendix C specifies the data sources, training settings, benchmarks, decoding parameters, and checkpoint-selection procedure.

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.

Mislav Balunovic, Jasper Dekoninck, Ivo Petrov, Nikola Jovanović, and Martin Vechev. Matharena: Evaluating llms on uncontaminated math competitions. *Advances in Neural Information Processing Systems*, 38, 2026.

Yinghui He, Simran Kaur, Adithya Bhaskar, Yongjin Yang, Jiarui Liu, Narutatsu Ri, Liam Fowl, Abhishek Panigrahi, Danqi Chen, and Sanjeev Arora. Self-distillation zero: Self-revision turns binary rewards into dense supervision. *arXiv preprint arXiv:2604.12002*, 2026.

Jie Huang, Xinyun Chen, Swaroop Mishra, Huaixiu Steven Zheng, Adams Yu, Xinying Song, and Denny Zhou. Large language models cannot self-correct reasoning yet. In *International conference on learning representations*, volume 2024, pp. 32808–32824, 2024.

Zhilin Huang, Hang Gao, Ziqiang Dong, Yuan Chen, Yifeng Luo, Chujun Qin, Jingyi Wang, Yang Yang, and Guanjun Jiang. Learning from your own mistakes: Constructing learnable micro-reflective trajectories for self-distillation. *arXiv preprint arXiv:2606.18844*, 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.

Simran Kaur, Narutatsu Ri, Yinghui He, Liam Fowl, and Sanjeev Arora. Rethinking on-policy self-distillation for thinking models. *arXiv preprint arXiv:2607.05184*, 2026.

Jeonghye Kim, Jiwon Jeon, Dongsheng Li, and Yuqing Yang. Rebellious student: Reversing teacher signals for reasoning exploration with self-distilled rlvr. *arXiv preprint arXiv:2605.10781*, 2026.

Aviral Kumar, Vincent Zhuang, Rishabh Agarwal, Yi Su, John D Co-Reyes, Avi Singh, Kate Baumli, Shariq Iqbal, Colton Bishop, Rebecca Roelofs, et al. Training language models to self-correct via reinforcement learning. In *International Conference on Learning Representations*, volume 2025, pp. 54523–54549, 2025.

Yaxuan Li, Yuxin Zuo, Bingxiang He, Jinqian Zhang, Chaojun Xiao, Cheng Qian, Tianyu Yu, Huanang Gao, Wenkai Yang, Zhiyuan Liu, et al. Rethinking on-policy distillation of large language models: Phenomenology, mechanism, and recipe. *arXiv preprint arXiv:2604.13016*, 2026.

Jianbo Lin, Xiaomin Yu, Yi Xin, Yifu Guo, Zhuosong Jiang, Zhongqi Yue, Weishi Wang, Heqing Zou, Chengwei Qin, and Hui Xiong. Icrl: Learning to internalize self-critique with reinforcement learning. *arXiv preprint arXiv:2605.15224*, 2026.

Yifeng Liu, Shiyuan Zhang, Yifan Zhang, and Quanquan Gu. Self-distilled policy gradient. *arXiv preprint arXiv:2606.04036*, 2026.

Aman Madaan, Niket Tandon, Prakhar Gupta, Skyler Hallinan, Luyu Gao, Sarah Wiegreffe, Uri Alon, Nouha Dziri, Shrimai Prabhumoye, Yiming Yang, et al. Self-refine: Iterative refinement with self-feedback. *Advances in neural information processing systems*, 36:46534–46594, 2023.

Duy Nguyen, Hanqi Xiao, Archiki Prasad, Zaid Khan, Anirban Das, Austin Zhang, Sambit Sahu, Hyunji Lee, Elias Stengel-Eskin, and Mohit Bansal. Avsd: Adaptive-view self-distillation by balancing consensus and teacher-specific privileged signals. *arXiv preprint arXiv:2605.20643*, 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*, 2024.

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*, 2026.

Noah Shinn, Federico Cassano, Ashwin Gopinath, Karthik Narasimhan, and Shunyu Yao. Reflexion: Language agents with verbal reinforcement learning. *Advances in neural information processing systems*, 36:8634–8652, 2023.

Zhuo Sun, Entong Li, Yanlong Zhao, Xiaoyuan Cheng, Wenxuan Yuan, Kaiyu Li, Che Liu, Huihang Liu, Harrison Bo Hua Zhu, and Li Zeng. Sr-opsd: Self-referenced on-policy self-distillation. *arXiv preprint arXiv:2608.09745*, 2026.

Zhiquan Tan and Yinrong Hong. Self-supervised on-policy distillation for reasoning language models. *arXiv preprint arXiv:2605.17497*, 2026.

Qiaoyu Tang, Hao Xiang, Le Yu, Bowen Yu, Hongyu Lin, Yaojie Lu, Xianpei Han, Le Sun, and Junyang Lin. Refcritic: Training long chain-of-thought critic models with refinement feedback. *arXiv preprint arXiv:2507.15024*, 2025.

Yang Tian, Rui Wang, Xumeng Wen, Junjie Li, Shizhao Sun, Lei Song, Jiang Bian, and Bo Zhao. Pbsd: Privileged bayesian self-distillation for long-horizon credit assignment. *arXiv preprint arXiv:2606.09348*, 2026.

Zhenlin Wei, Pu Jian, Yingzhuo Deng, Xiaohan Wang, Jiajun Chai, Zhexin Hu, Wei Lin, Shanbin Zhang, and Guojun Yin. Amr-sd: Asymmetric meta-reflective self-distillation for token-level credit assignment. *arXiv preprint arXiv:2605.18529*, 2026.

Xingjian Wu, Junlin Liu, Xingchen Liu, Xuhang Zhu, Jianing Wang, Linsen Guo, Xiaoyu Li, Xuezhi Cao, and Xunliang Cai. Contrastive reinforced policy optimization via privileged self-distillation. *arXiv preprint arXiv:2607.28026*, 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.

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.

Yuxiao Yang, Xiaoyun Wang, and Weitong Zhang. Ogls-sd: On-policy self-distillation with outcome-guided logit steering for llm reasoning. *arXiv preprint arXiv:2605.12400*, 2026b.

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

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

A TRAINING ALGORITHM

Algorithm 1: RF-SD

Input: dataset $\mathcal{D}$; verifier Ver; group sizes N, G; steps K; evolution flag e

// Student and teacher always use the adapter-off backbone

Initialize θ and, if $e = 1$, ϕ; otherwise $\phi = \varnothing$

for $k = 1, \ldots, K$ **do**

  $\theta_{\text{old}} \leftarrow \theta$, $\phi_{\text{old}} \leftarrow \phi$

  Sample questions from $\mathcal{D}$

  Sample N rollouts per question: $y_i \sim \pi^{\text{S}}_{\theta_{\text{old}}}(\cdot \mid P_{\text{S}}(x_i))$

  Compute $z_i = \text{Ver}(x_i, y_i)$ and group-normalized $\hat{A}_i$ (Eq. (1))

  for each (x_i, y_i, z_i) **do**

    Sample $R_i \sim \pi^{\text{R}}_{\theta_{\text{old}},\phi_{\text{old}}}(\cdot \mid P_{\text{R}}(x_i, y_i, z_i))$ (Eq. (4))

    Store $P_{\text{T}}(x_i, R_i)$

  end

  if $e = 1$ **then** optional RF-SD-e

    Randomly select one rollout (x_j, y_j, z_j) per question

    for each *selected rollout* j *and candidate* $g = 1, \ldots, G$ **do**

      Sample $R_{j,g} \sim \pi^{\text{R}}_{\theta_{\text{old}},\phi_{\text{old}}}(\cdot \mid P_{\text{R}}(x_j, y_j, z_j))$

      Sample $\widetilde{y}_{j,g} \sim \pi^{\text{T}}_{\theta_{\text{old}}}(\cdot \mid P_{\text{T}}(x_j, R_{j,g}))$

      $r^{\text{R}}_{j,g} \leftarrow \text{Ver}(x_j, \widetilde{y}_{j,g})$

    end

    Compute $\hat{A}^{\text{R}}_{j,g}$ from group-normalized $r^{\text{R}}_{j,g}$

    if *some* $\hat{A}^{\text{R}}_{j,g} \neq 0$ **then**

      Update ϕ with $\mathcal{L}_{\text{R}}(\phi)$ (Eq. (10)); hold θ fixed

    end

  end

  Compute teacher targets $\text{sg}[\pi^{\text{T}}_{\theta}(\cdot \mid P_{\text{T}}(x_i, R_i), y_{i,<t})]$

  Update θ with $\mathcal{L}_{\text{S}}(\theta)$ (Eq. (6))

end

Output: student π^{S}_{θ} with the reflector adapter disabled

B REFLECTION PROMPTS AND PARSING

Both routes receive user content of the form `Task:\n{question}\n\n<label>:\n{student_response}` and nothing else. The success route labels the trajectory neutrally as "Solver trajectory"; the failure route labels it "Failed or incomplete solver trajectory". The failure route also states this failure status in its system prompt; neither route exposes the ground-truth answer, and the success route does not reveal the positive verdict.

Reflection from Success (system). Act as a grounded solution-note extractor. Extract the most valuable reasoning steps, identities, constraints, or checks that are explicitly present in the trajectory. Give one to four concise, actionable tips, grounding every tip in a specific step or expression visible in the trajectory. If a visible step needs more support, say what justification or verification it needs *without declaring that step wrong*. Do not decide or state whether the trajectory or the final answer is correct. Do not introduce any formula, number, intermediate result, candidate answer, alternative derivation, or final answer that is absent from the trajectory. Output exactly one `<critic>...</critic>` block.

Reflection from Failure (system). Act as the critic for a failed or incomplete trajectory. Locate the first or highest-impact mathematical mistake, missing constraint, unsupported inference, or final-answer-format problem visible in the trajectory, and ground every criticism in a specific step or expression. If no exact mistake can be located, identify the step that must be re-derived and what property needs checking; do not guess an error. Do not introduce unsupported formulas, numbers, intermediate results, candidate answers, or final answers. Output exactly one `<critic>...</critic>` block.

Teacher context. The teacher prompt is `{question}\n\nA separate reflection process supplied the following grounded guidance:\n{reflection}\n\nUse the guidance to verify your reasoning, but solve the problem yourself and show the full solution process.`

Reflection parsing. Reflection generation stops at `</critic>` or a maximum of 512 tokens. A strict parser extracts the content enclosed by `<critic>...</critic>`, and only this extracted text enters the teacher context. During reflector evolution, the policy-gradient update is applied to the full generated completion, including formatting tokens.

Invalid reflections. A reflection is valid if it is nonempty, well-formed, and successfully incorporated into the teacher context. Invalid reflections contribute zero distillation loss without changing the token-count denominator, while the corresponding student rollouts still contribute to RLVR. During reflector evolution, invalid candidates remain in their groups and receive zero reward.

C Reproducibility details

Data sources. We use the English subset of DAPO-Math-17k[1] and source expert reference solutions primarily from the public Gemini 3 Flash distillation dataset.[2]

Training configuration. We train Qwen3-1.7B, Qwen3-4B, and Qwen3-8B on the English subset of DAPO-Math-17k for 300 optimizer steps. Each global batch contains 128 questions, with $N = 8$ student rollouts per question and a maximum response length of 4,096 tokens. GRPO clipping and loss aggregation follow Section 4. We use a distillation coefficient of $\beta = 10^{-3}$ unless otherwise stated, without soft-overlong reward shaping or dynamic group filtering. The learning rate is set as 10^{-6}.

Evaluation. We evaluate on AIME 2024, AIME 2025, HMMT 2025, and AMC 2023, containing 30, 30, 30, and 40 questions, respectively. We generate 12 responses per question, yielding 1,560 evaluated responses per checkpoint. All evaluations use the official Qwen chat template, temperature 0.7, top-p 0.8, top-k 20, and a maximum sequence length of 32,768 tokens. Responses are scored using the same fixed grader. Avg@12 denotes the fraction of correct responses on each benchmark, and the overall score is the unweighted mean across the four benchmarks. *Best* reports the checkpoint with the highest overall score at or before training step 300, with all benchmark results taken from that same checkpoint.

Reflector-evolution configuration. For each selected student rollout, the reflector generates $G = 8$ candidate reflections. Each candidate conditions one teacher answer sampled at temperature 1.0, whose correctness provides the reflector reward. The auxiliary GRPO update uses a clipping parameter of $\varepsilon = 0.2$ and optimizes only the reflector's LoRA adapter, with the shared backbone held fixed during this update. The student and teacher use the backbone without the adapter, and evaluation uses the adapter-free student. The learning rate is set as 10^{-6}.

D Additional reflector-evolution results

Table 6 expands the average scores in Table 4. The LoRA rank and scaling are set as 64 and 128.

Fig. 5 compares the student reward trajectories of RF-SD and RF-SD-e at all three scales.

[1] https://huggingface.co/datasets/BytedTsinghua-SIA/DAPO-Math-17k

[2] https://huggingface.co/datasets/RyanYr/DAPO_Math_17k-gemini3-flash

Table 6: **Per-benchmark reflector-evolution results.** Avg@12 (%), reported as Final / Best for the configurations in Table 4. Final is step 300; Best uses one checkpoint selected by the four-benchmark average.

Method	AIME24	AIME25	HMMT25	AMC23	Average	Step
Qwen3-1.7B						
RF-SD	31.94 / 31.94	25.00 / 25.00	16.11 / 16.11	75.63 / 75.63	37.17 / 37.17	300 / 300
RF-SD-e	33.61 / 33.61	26.94 / 26.94	15.00 / 15.00	76.25 / 76.25	37.95 / 37.95	300 / 300
Qwen3-4B						
RF-SD	42.50 / 47.50	39.44 / 40.56	23.06 / 24.44	90.83 / 89.79	48.96 / 50.57	300 / 280
RF-SD-e	42.22 / 47.50	43.33 / 46.11	22.78 / 22.78	90.63 / 90.63	49.74 / 51.75	300 / 180
Qwen3-8B						
RF-SD	52.78 / 50.28	37.22 / 43.61	25.00 / 23.06	90.63 / 90.83	51.41 / 51.94	300 / 240
RF-SD-e	56.11 / 56.39	41.11 / 41.67	22.22 / 23.06	91.25 / 91.04	52.67 / 53.04	300 / 280

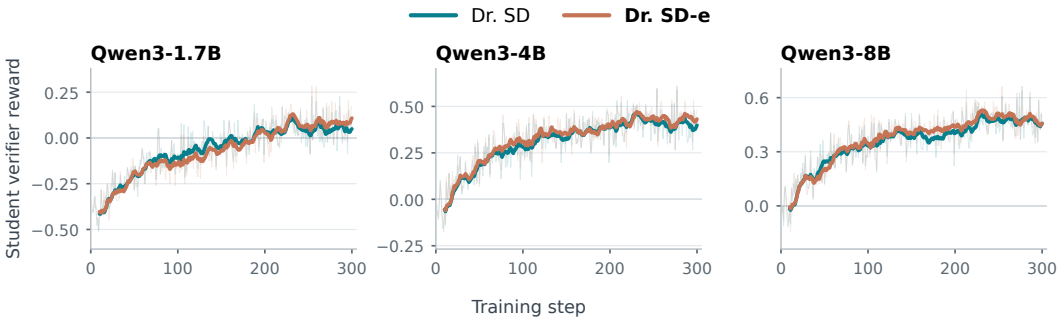

Figure 5: **Student reward with and without reflector evolution.**

E Training and Evaluation Dynamics

Figs. 7 and 8 extend the Qwen3-4B results in Fig. 2 to Qwen3-1.7B and Qwen3-8B. Figs. 6, 9 and 10 provide cross-scale comparisons of training metrics and benchmark performance. Evaluations are conducted every 20 steps, starting from the shared untrained model at step 0.

F Distillation Objective and Strength

At $\beta = 10^{-3}$, reverse KL achieves the highest final and best averages among the tested divergences (Table 7). Increasing its coefficient to 10^{-2} reduces the final average from 48.96 to 37.43. With $\beta = 10^{-1}$, the run reaches its best average of 36.01 at step 20 but collapses to 0.00 by step 300. These results support the selection of $\beta = 10^{-3}$ in our experiments.



Figure 6: **Average evaluation score across model scales of RF-SD.** Each point is the average of the four benchmark Avg@12 scores at one checkpoint. RF-SD leads at every evaluated checkpoint from step 240 through 300 at all three scales.

Table 7: **Distillation coefficient and divergence** on Qwen3-4B. Average is the mean of the four benchmark scores (Avg@12). Final / Best reports step 300 / each run's checkpoint with the highest average through step 300. Only the coefficient β and divergence objective vary.

β	Divergence	Average (Final / Best)
10^{-3}	reverse KL (ours)	**48.96 / 50.57**
10^{-3}	forward KL	47.17 / 48.40
10^{-3}	Jensen-Shannon	43.92 / 48.28
10^{-2}	reverse KL	37.43 / 40.14
10^{-1}	reverse KL	0.00 / 36.01

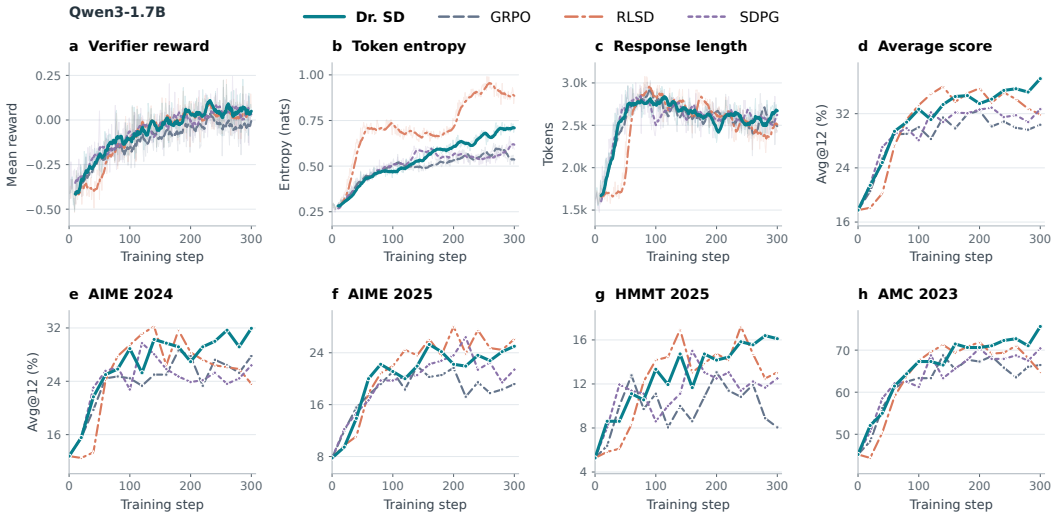

Figure 7: **Training and evaluation dynamics on Qwen3-1.7B.** Panels show verifier reward, token entropy, response length, and average and individual benchmark scores.

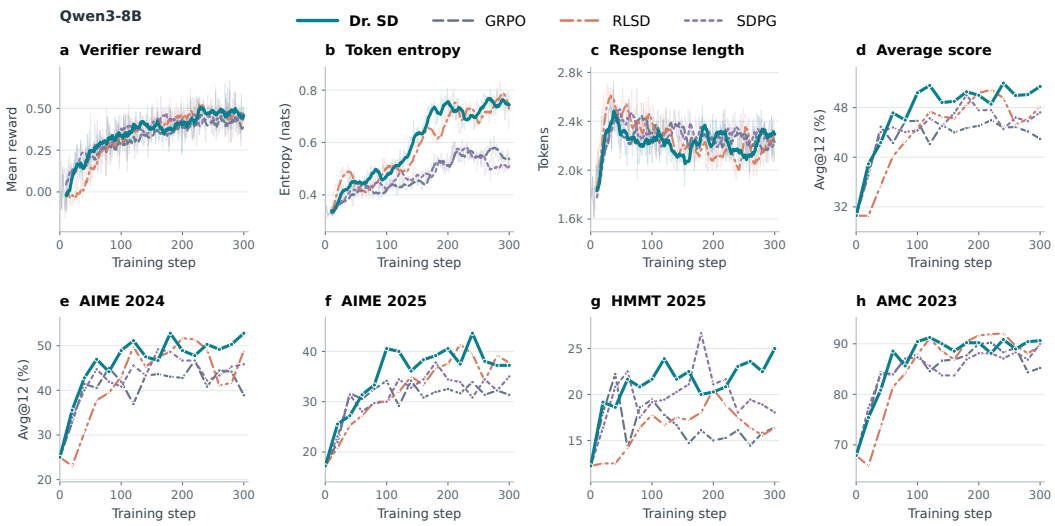

Figure 8: **Training and evaluation dynamics on Qwen3-8B.** Panel layout and curve processing match Fig. 7.

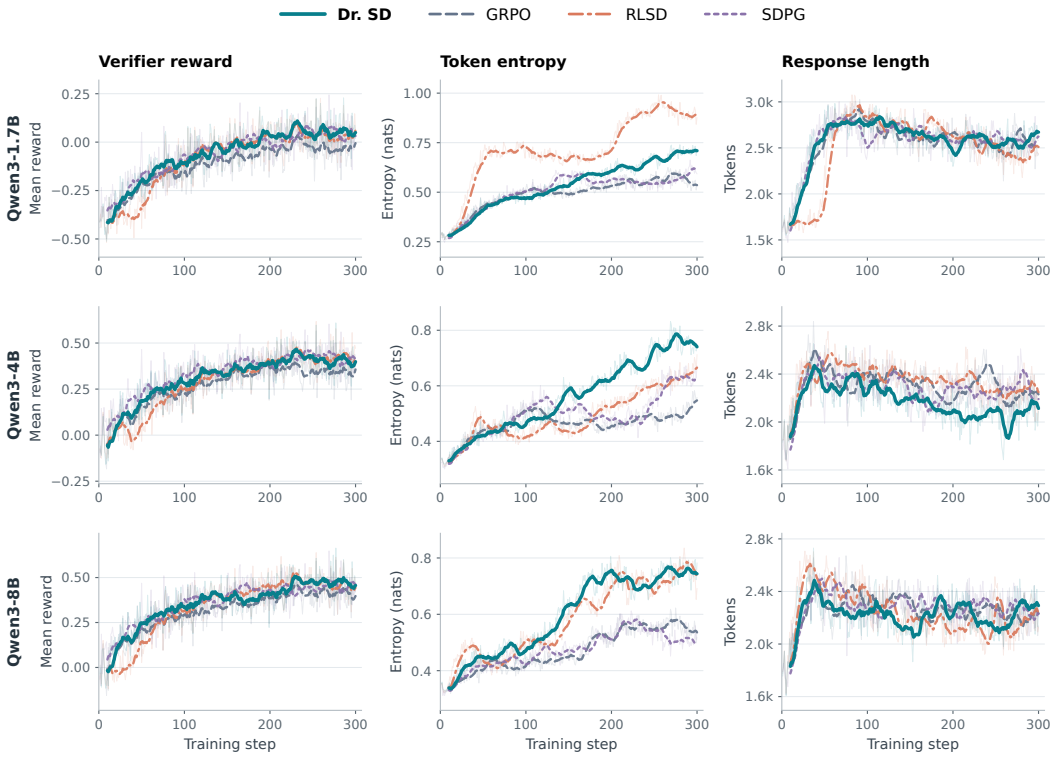

Figure 9: **Training dynamics across model scales.** Rows correspond to Qwen3-1.7B, 4B, and 8B.

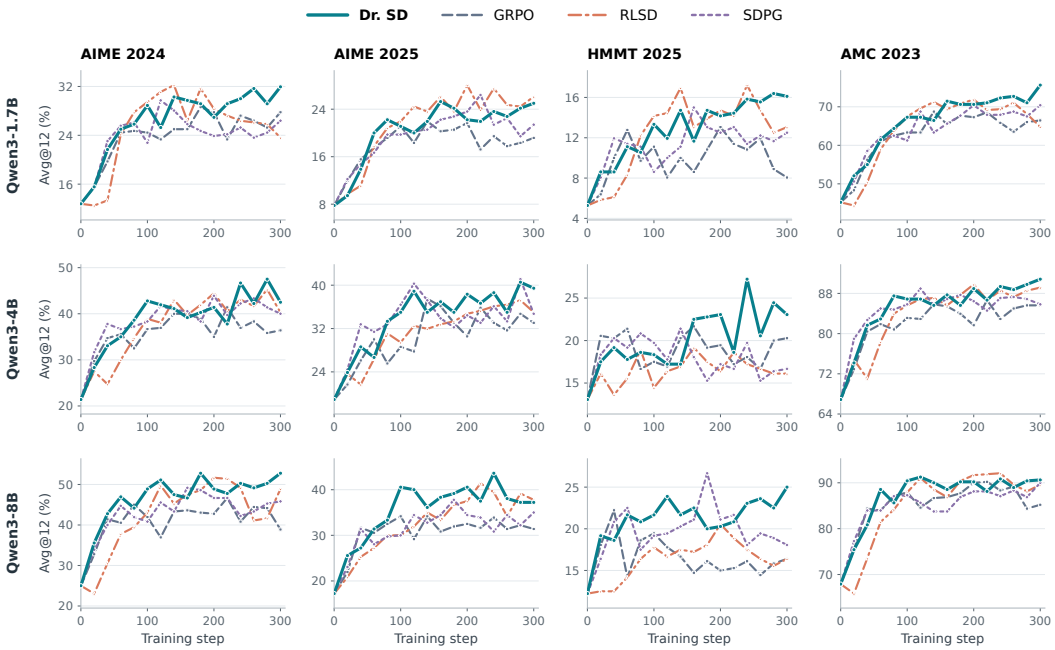

Figure 10: **Benchmark trajectories across model scales.** Every point is a measured Avg@12 score. Within each row, step 0 is the same untrained model for all methods.

G GENERATED REFLECTION EXAMPLES

We present examples of self-reflection generated during training with Qwen3-4B. Student and teacher responses are excerpted, with omissions marked by [. . .], while reflections are reproduced in full. Ground-truth answers were not provided to the reflectors.

G.1 REFLECTION FROM SUCCESSFUL ROLLOUTS

The following examples compare generic and success-route reflections at training step 1. The two runs generated different student responses to the same question, both of which were correct. The generic reflection suggests verifying the answer, whereas the success-route reflection summarizes the reasoning steps in its corresponding student response.

Question

Find x, where x is a square number which is divisible by four, and between 39 and 80.

Generic reflection

Student response (excerpts).

[. . .] - $6^2 = 36$
- $7^2 = 49$
- $8^2 = 64$
- $9^2 = 81$

[. . .] - **49 ÷ 4 = 12.25 → Not divisible by 4**
- **64 ÷ 4 = 16 → Divisible by 4**

[. . .] $\boxed{64}$

Generated reflection.

Check that 64 is indeed a square number and lies between 39 and 80.

RF-SD: success-route reflection

Student response (excerpts).

[. . .] - $6^2 = 36$
- $7^2 = 49$
- $8^2 = 64$
- $9^2 = 81$

[. . .] - $49 \div 4 = 12.25 \rightarrow$ **not divisible by 4**
- $64 \div 4 = 16 \rightarrow$ **divisible by 4**

[. . .] $\boxed{64}$

Generated reflection.

Tip 1: Identify square numbers between 39 and 80 by squaring integers from 7 to 9, as $6^2 = 36$ and $9^2 = 81$ are outside the range.

Tip 2: Check divisibility by 4 by dividing each square number by 4 and verifying if the result is an integer.

Tip 3: The only square number between 39 and 80 that is divisible by 4 is 64, as $64 \div 4 = 16$.

G.2 Reflection from failed rollouts

At training step 300, the generic-reflector and RF-SD runs generated different student responses to the same geometry problem, but both gave the incorrect answer 576. Both reflection-conditioned teachers subsequently produced the correct answer, 288.

Question (text as logged)

The diagram below shows a parallelogram $ABCD$ with $AB = 36$ and $AD = 60$. Diagonal BD is perpendicular to side AB. Points E and F bisect sides AD and BC, respectively. Points G and H are the intersections of BD with AF and CE, respectively. Find the area of quadrilateral $EGFH$.

Generic reflection

Student response (excerpts).

[. . .] E(18,24), G(36,16), H(36,32), F(54,24)

List in order and repeat the first at the end:

(18,24), (36,16), (36,32), (54,24), (18,24)

[. . .] **Area = $\frac{1}{2}$ |3600 - 3600| = $\frac{1}{2}$ * 0 = 0**

[. . .] $\boxed{576}$ [. . .]

Generated reflection.

Check the order of points in the shoelace formula; the area calculation may be incorrect due to an incorrect polygon ordering. Use vectors or coordinate decomposition instead.

RF-SD: failure-route reflection

Student response (excerpts).

[. . .] Let me list the coordinates as:

E(18,24), G(36,16), H(36,32), F(54,24)

[. . .] Total sum is 0? That can't be, that would be 0 area. That indicates my point order is incorrect.

[. . .] So the figure EGFH is a trapezoid with bases at 24 (EF of length 54-18=36), and the vertical segment GH of length 16 (from y=16 to y=32)

[. . .] Final Answer: $\boxed{576}$ (Assuming it's roughly 36*16=576) from coordinates. [. . .]

Generated reflection.

Tip 1: The shoelace formula was applied with an incorrect point ordering, leading to a zero area result. The points must be ordered sequentially around the perimeter of the quadrilateral without crossing.

Tip 2: The geometric interpretation of the quadrilateral as a trapezoid with bases EF (horizontal) and GH (vertical) is flawed. The area should be calculated using proper decomposition or correct ordering for the shoelace formula.

RF-SD: teacher response (excerpts)

[. . .] Try E, G, F, H

[. . .] Area = 1/2 * |576| = 288

So the area is **288**

Final Answer $\boxed{288}$

Both reflections identify the incorrect vertex ordering in the shoelace calculation. The failure-route reflection additionally identifies the mistaken treatment of EF and GH as trapezoid bases. These segments are perpendicular diagonals of lengths 36 and 16, yielding an area of $\frac{1}{2} \cdot 36 \cdot 16 = 288$.

G.3 Reflections at different training stages

The following examples were sampled at steps 1 and 300 of the same RF-SD-e training run. They concern the same question but different student responses, both of which were correct.

Question

Evaluate $2000^3 - 1999 \cdot 2000^2 - 1999^2 \cdot 2000 + 1999^3$

Early: step 1

Student response (excerpts).

[...] $2000^2(2000 - 1999) = 2000^2(1) = 2000^2$

[...] $1999^2(1999 - 2000) = 1999^2(-1) = -1999^2$

[...] $2000^2 - 1999^2 = (2000 - 1999)(2000 + 1999)$

$= (1)(3999) = 3999$ [...]

Generated reflection.

Tip 1: Group terms with common factors to simplify the expression. This helps identify patterns and reduces complexity.

Tip 2: Factor each group using the distributive property, which is crucial for simplifying large expressions.

Tip 3: Recognize the difference of squares identity $a^2 - b^2 = (a - b)(a + b)$ to further simplify the expression.

Late: step 300

Student response (excerpts).

[...] $(2000^3 - 1999 \cdot 2000^2) - (1999^2 \cdot 2000 - 1999^3)$

[...] Now, note that the second group is subtracted:

So the expression becomes:

$2000^2 - 1999^2$

[...] $(2000 - 1999)(2000 + 1999) = (1)(3999) = 3999$ [...]

Generated reflection.

Tip 1: Group the terms as (2000^3 - 1999·2000^2) - (1999^2·2000 - 1999^3) to factor out common factors from each group. This step reveals that the first group factors to 2000^2 and the second group factors to 1999^2.

Tip 2: Recognize that the expression simplifies to 2000^2 - 1999^2 after factoring, which can be further simplified using the difference of squares formula.

Tip 3: Apply the difference of squares formula to 2000^2 - 1999^2, resulting in (2000 - 1999)(2000 + 1999) = 1 $\times$ 3999 = 3999. This step relies on correctly identifying the values of a and b in the formula.

The early reflection summarizes the main solution techniques, including factoring and the difference-of-squares identity. The late reflection refers more explicitly to the intermediate expressions and final result in its corresponding student response.

G.4 AN UNSUCCESSFUL REFLECTION

This example is drawn from the RF-SD run at training step 300. Both the student and the reflection-conditioned teacher produced the incorrect answer 2415, whereas the correct answer is 142.

Question

The Purple Comet! Math Meet runs from April 27 through May 3, so the sum of the calendar dates for these seven days is $27 + 28 + 29 + 30 + 1 + 2 + 3 = 120$. What is the largest sum of the calendar dates for seven consecutive Fridays occurring at any time in any year?

Student response (excerpts)

[. . .] Let me have the day of the year for the first date d be x, then x+42 $\leq$ 365

Thus x $\leq$ 365-42 = 323

[. . .] Then sum is 7*324 +147

Compute:

7*324 = 2268

2268+147 = 2415

[. . .] **Final Answer:** $\boxed{2415}$

Failure-route reflection

Tip 1: The calculation of the maximum day of the year (323 or 324) assumes that the date on that day is a Friday, which is not verified in the trajectory. The solver assumes it without confirming that the day of the year 323 or 324 corresponds to a Friday, which is a critical missing constraint.

Tip 2: The derivation of the sum formula (7d + 147) is mathematically correct, but the assumption that d can be as high as 323 or 324 in a leap year is not grounded in the actual calendar structure. The solver does not check whether the 7th Friday (d+42) falls within the year, but this is implicitly assumed without explicit validation.

Teacher response (excerpt)

[. . .] Thus, final boxed answer:

$\boxed{2415}$

The student incorrectly interprets calendar dates as day-of-year values rather than day-of-month values. The reflection overlooks this error and questions the year-boundary condition even though the student has already imposed $x + 42 \leq 365$. The teacher's response retains the incorrect final answer.

H SELF-REFLECTION AND TEACHER-STUDENT MISMATCH

We analyze the expected teacher-student reverse KL and its gradients for the base RF-SD, with model parameters, the student instance, and teacher-context sampling distributions held fixed.

H.1 A MODEL OF REFERENCE-SPECIFIC LOGIT VARIATION

Fix a student instance (x, y, z) and response position t, with $p(a) = \pi_\theta^{\mathrm{S}}(a \mid P_{\mathrm{S}}(x), y_{<t})$. Let q_C and q_R denote the teacher distributions induced by reference solutions and valid reflections, respectively, for the same question and student prefix. We assume that a valid reflection is generated with positive probability.

Following Eq. (7), we compare the two teacher families through

$$q_C \overset{\mathrm{d}}{=} q_1(\xi), \qquad q_R \overset{\mathrm{d}}{=} q_\lambda(\xi), \qquad q_s(\xi) = \mathrm{softmax}(v_\star + s\xi), \tag{11}$$

where $s \in [0, 1]$ and $\lambda \in [0, 1]$ is fixed independently of ξ. We use centered logit vectors and assume finite logits on a finite vocabulary, $\mathbb{E}\xi = 0$, and $\mathbb{E}\|\xi\| < \infty$.

Here $v_\star$ is the shared mean logit vector, and ξ represents zero-mean reference-specific variation. The model assumes that the reflection-conditioned teacher retains the same mean logits while scaling these deviations by λ. These distributional identities define the analytical comparison between the two teacher families.

H.2 Proof of the mismatch comparison

Proof of Proposition 1. Write $A(v) = \log \sum_a e^{v_a}$ and $F(s) = \mathbb{E}_\xi D_{\mathrm{KL}}(p \| q_s(\xi))$. Finite vocabulary and $\mathbb{E}\|\xi\| < \infty$ ensure that these quantities are finite. Expanding the divergence and using $\mathbb{E}\xi = 0$ gives

$$\begin{aligned} F(s) &= \sum_a p(a) \log p(a) - p^\top v_\star - s\, p^\top \mathbb{E}\xi + \mathbb{E}A(v_\star + s\xi) \\ &= \sum_a p(a) \log p(a) - p^\top v_\star + \mathbb{E}A(v_\star + s\xi). \end{aligned} \tag{12}$$

By convexity of A,

$$F(\lambda) \le (1-\lambda)F(0) + \lambda F(1), \qquad F(0) \le F(1), \tag{13}$$

where the second inequality follows from Jensen's inequality. Writing $q_\star = \mathrm{softmax}(v_\star)$, we also have

$$\begin{aligned} F(1) - F(0) &= \mathbb{E}[A(v_\star + \xi) - A(v_\star)] \\ &= \mathbb{E}\big[A(v_\star + \xi) - A(v_\star) - q_\star^\top \xi\big] \\ &= \mathbb{E} D_{\mathrm{KL}}(q_\star \| q_1(\xi)), \end{aligned} \tag{14}$$

where the second equality uses $\mathbb{E}\xi = 0$.

Combining Eqs. (13) and (14) yields

$$F(1) - F(\lambda) \ge (1-\lambda)\mathbb{E} D_{\mathrm{KL}}(q_\star \| q_1(\xi)) \ge 0. \tag{15}$$

The distributional identities in Eq. (11) identify $F(1)$ and $F(\lambda)$ with the reference- and reflection-conditioned teacher expectations in Eq. (8).

If ξ changes a token-to-token logit difference with positive probability, then $q_1(\xi) \neq q_\star$ with positive probability. The expected KL on the right-hand side is therefore strictly positive, establishing strict reduction when $\lambda < 1$. □

Gradient interpretation. By Eq. (12),

$$F(s) = D_{\mathrm{KL}}(p \| q_\star) + c_s,$$

where $c_s = \mathbb{E}A(v_\star + s\xi) - A(v_\star)$ is independent of p. Thus, the change in expected KL does not change its gradient with respect to the student logits.

More explicitly, write $p = \mathrm{softmax}(w)$ and $J_p = \mathrm{diag}(p) - pp^\top$. Treating each sampled teacher distribution as stop-gradient and using $J_p \mathbf{1} = 0$ gives

$$G_s := \nabla_w D_{\mathrm{KL}}(p \| q_s(\xi)) = J_p(\log p - v_\star - s\xi).$$

If $\mathbb{E}\|\xi\|^2 < \infty$, then

$$\begin{aligned} \mathbb{E}G_s &= J_p(\log p - v_\star), \\ \mathrm{Cov}(G_s) &= s^2 J_p\, \mathrm{Cov}(\xi) J_p. \end{aligned} \tag{16}$$

These expectations and covariances are taken over teacher-context draws at the fixed student prefix. Scaling the deviations from ξ to $\lambda\xi$ therefore preserves the mean gradient with respect to student logits while reducing its covariance by a factor of λ^2.