TEACHING WITH HINDSIGHT: REFLECTION-INDUCED CORRECTION FOR ON-POLICY DISTILLATION

Anonymous authors
Paper under double-blind review

ABSTRACT

On-policy distillation (OPD) improves language-model reasoning by transferring teacher knowledge on student-generated trajectories, aligning supervision with the reasoning patterns and errors encountered under the student's own policy. However, standard OPD relies on prefix-conditioned teacher predictions without explicitly incorporating the teacher's retrospective diagnosis of why a completed reasoning attempt failed. To address this gap, we propose Reflection-Corrected On-Policy Distillation (RC-OPD), which translates the teacher's reasoning about student failures into corrective supervision on the student's original trajectories. For verifier-identified failures, a frozen teacher generates reflective feedback and evaluates the same response under standard and reflection-conditioned contexts. Token-level log-probability differences quantify the changes in teacher assessment induced by reflection, making its additional guidance explicit relative to the teacher's existing predictions. RC-OPD selectively incorporates these adjustments into base OPD over diagnosed reasoning suffixes and bounds the combined distillation signal before integrating it with outcome-based reinforcement learning. This allows retrospective diagnosis to refine predictive supervision while retaining standard guidance elsewhere. The student is optimized solely on the original problem and its sampled response, requiring neither diagnostic feedback nor additional reflection at inference. Experiments on Qwen3 models across mathematical reasoning and code generation show that RC-OPD consistently outperforms strong reinforcement-learning and distillation baselines across all tested scales, demonstrating the value of retrospective diagnosis as a complement to predictive teacher supervision. Code will be released upon acceptance.

1 INTRODUCTION

Large-scale pretraining and reinforcement learning have substantially advanced the language and reasoning capabilities of large language models (LLMs) (OpenAI, 2023; DeepSeek-AI, 2025; DeepSeek-AI et al., 2026; GLM-5-Team et al., 2026; Team et al., 2026). Knowledge distillation provides an efficient way to transfer these capabilities to smaller models (Hinton et al., 2015; Hsieh et al., 2023; Shridhar et al., 2023; Gu et al., 2024). However, conventional distillation often learns from teacher-generated responses (Kim & Rush, 2016), creating a distribution mismatch between teacher trajectories seen during training and student-generated trajectories encountered at inference time (Agarwal et al., 2024). On-policy distillation (OPD) addresses this mismatch by providing teacher supervision on responses sampled from the student's current policy, thereby transferring knowledge directly to the states the student actually encounters (Agarwal et al., 2024).

The effectiveness of this supervision depends on how teacher predictions are translated into student updates. Generalized OPD separates teacher-derived rewards from KL regularization, allowing their relative influence to be adjusted beyond the fixed balance imposed by the standard objective (Yang et al., 2026). Incorporating outcome-based reinforcement learning further requires coordinating dense teacher guidance with task-level feedback. This has motivated methods that regulate the magnitude and training schedule of distillation advantages (Ding et al., 2026), as well as methods that selectively use teacher likelihoods to reweight tokens in positive-advantage responses (Wang et al., 2026). Across these formulations, the teacher's next-token predictions provide the underlying fine-grained guidance, while the optimization design determines its strength, its interaction with task rewards, and the responses to which it is applied.

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

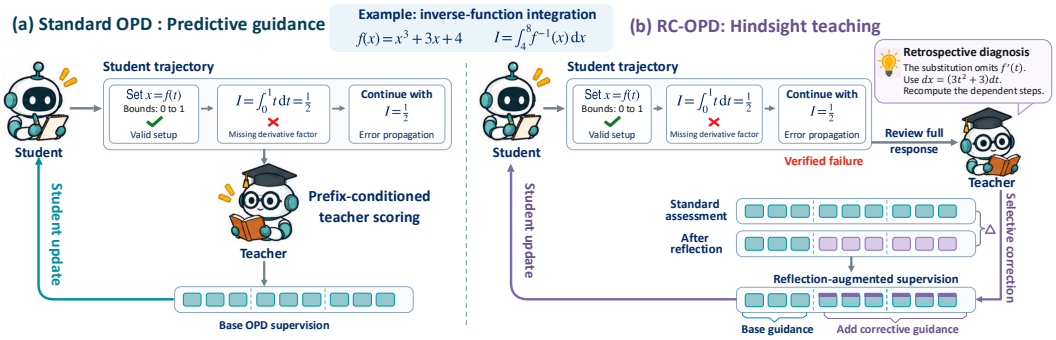


Figure 1: Teaching with hindsight through reflection-corrected on-policy distillation. (a) Standard OPD provides token-level supervision through prefix-conditioned teacher predictions. (b) RC-OPD incorporates retrospective diagnosis of the failed attempt to reassess teacher support for the same trajectory. Diagnosis-induced changes selectively refine base OPD supervision, allowing the student to learn from its original trajectory without requiring diagnostic feedback at inference.

The value of teacher guidance, however, depends on more than how densely it is provided or how strongly it is weighted. Targeting distillation toward problems on which the student fails but the teacher succeeds can improve the effectiveness of knowledge transfer (Han et al., 2026). An unsuccessful reasoning trajectory can nevertheless contain valid intermediate deductions alongside a consequential mistake, making the location of the error important for assessing the response (Zheng et al., 2025). A final-answer verifier reports the outcome without specifying which intermediate steps should be retained or revised, a distinction that motivates supervision of the reasoning process itself (Lightman et al., 2024). Standard OPD provides finer-grained guidance through teacher predictions conditioned on student-generated prefixes (Agarwal et al., 2024). Although this scoring can discourage incorrect continuations, it does not explicitly incorporate the teacher's retrospective diagnosis of the completed attempt. Such a diagnosis may provide additional guidance by relating individual reasoning choices to the failure of the overall response.

Natural-language critiques offer access to this diagnostic capability, with stronger models demonstrating an ability to critique weaker models' responses (Lin et al., 2024). Verbal feedback can guide iterative revision by identifying shortcomings and suggesting improvements (Madaan et al., 2023). Feedback-conditioned self-distillation further enables rich environmental feedback to be converted into dense supervision for policy learning (Hübotter et al., 2026). Model-generated critiques can likewise inform token-level supervision through a critique-conditioned self-teacher, allowing their benefits to be internalized without requiring critiques at inference time (Jin et al., 2026). Building on these findings, we focus on the additional supervision a critique provides relative to the teacher's existing predictions. A critique-conditioned teacher distribution reflects both its original predictive tendencies and the changes associated with the feedback. Using the teacher's original predictions as a reference makes these changes explicit along the same student trajectory. This motivates identifying useful corrections within those changes and integrating them with standard OPD to transfer the benefits of retrospective diagnosis into the student's reasoning.

To this end, we propose Reflection-Corrected On-Policy Distillation (RC-OPD), a framework for teaching with hindsight which translates the teacher's retrospective diagnosis of student errors into corrective supervision on the student's own reasoning trajectories. As shown in Figure 1, the student correctly sets up a substitution but omits the required derivative factor, causing the error to propagate through subsequent calculations. Reviewing the completed attempt allows the teacher to identify this omission and explain why the dependent calculations need to be revisited. RC-OPD incorporates this diagnostic guidance by having the same frozen teacher score the original student response with and without reflection. The token-level log-probability differences quantify how retrospective diagnosis changes the teacher's assessment of the sampled trajectory. Taking the teacher's standard predictions as a reference makes the additional guidance from reflection explicit and separately controllable, allowing failure-specific corrections to supplement the predictive supervision provided by OPD. Diagnostic validation screens the feedback, while bounded, selective correction controls the scope and strength of its token-level influence. This preserves the supervision already available through

standard OPD while allowing retrospective diagnosis to refine the learning signal. The training objective combines both sources of teacher guidance with outcome-based advantages to optimize the student on the original question and sampled response. Reflection enriches the training supervision without becoming part of the student's inference-time input.

In summary, our contributions are as follows:

- We introduce Reflection-Corrected On-Policy Distillation (RC-OPD), which complements the predictive supervision of standard OPD with retrospective teacher diagnosis of completed student failures. This enables students to learn from failure-specific corrective knowledge on their own trajectories while retaining feedback-free, single-pass inference.
- The proposed corrective distillation mechanism converts retrospective diagnosis of completed trajectories into localized token-level supervision on the student's original response. By selectively incorporating reflection-induced adjustments while retaining base OPD guidance, it integrates failure-specific correction with predictive teacher supervision and outcome-based learning.
- Extensive experiments across mathematical reasoning and code generation with multiple model scales demonstrate consistent improvements over strong reinforcement-learning and distillation baselines. Further analyses show how retrospective diagnosis, selective correction, and retained predictive supervision contribute to the effectiveness of RC-OPD.

2 Related Work

On-Policy Distillation and Combination with RLVR. On-policy distillation (OPD) transfers teacher knowledge on student-generated trajectories, reducing the train–inference mismatch of conventional distillation (Gu et al., 2024). Subsequent work improves the formulation and integration of teacher supervision with outcome-based optimization. GKD supports flexible divergence objectives and joint optimization of task rewards and teacher alignment (Agarwal et al., 2024), while G-OPD generalizes OPD through flexible reference models and reward scaling, with ExOPD strengthening teacher-derived rewards through extrapolation (Yang et al., 2026). OPD^2 instead isolates reasoning-specific teacher knowledge using log-probability differences between a reasoning-tuned teacher and its pretrained base model (Heo et al., 2026). Recent methods further combine dense teacher guidance with RLVR: Distilled RL reweights policy gradients using teacher–student likelihood ratios (Wang et al., 2026), SAF-OPD adaptively controls the magnitude and timing of OPD advantages before fusion with GRPO (Ding et al., 2026), and RSTG selectively introduces teacher guidance where RL groups provide insufficient learning signals (Han et al., 2026). Beyond token-local supervision, γOPD introduces temporal credit assignment and reward-compatible mixing to balance teacher guidance with outcome feedback (Liu et al., 2026). These methods primarily improve how predictive teacher distributions are extracted, weighted, or combined with rewards. In contrast, RC-OPD additionally exploits retrospective diagnoses of completed student attempts, converting reflection-induced changes in teacher assessment into targeted corrective supervision on the same original trajectories.

Learning from Feedback. Natural-language feedback can identify errors and guide iterative revision (Madaan et al., 2023; Shinn et al., 2023). RISE learns revision through iterative fine-tuning on model-generated attempts (Qu et al., 2024), whereas SCoRe develops self-correction through multi-turn reinforcement learning on self-generated correction trajectories (Kumar et al., 2025). Other approaches distill the benefits of feedback into policies that operate without it. SDPO uses a feedback-conditioned self-teacher to provide dense supervision on the original student trajectory (Hübotter et al., 2026). OPCD obtains critiques from weaker external models, filters them using revision outcomes and rubric-based assessment, and distills the resulting critique-conditioned strong-model predictions (Jin et al., 2026). SD-Zero develops a reviser conditioned on an initial response and its binary reward, then distills its predictions into the initial-response policy (He et al., 2026). Kaur et al. show that privileged-context distillation can suppress reconsideration in thinking models, particularly under long reasoning budgets (Kaur et al., 2026). RC-OPD retains the original OPD supervision from a frozen external teacher and incorporates reflection-induced changes as a separately screened and bounded correction on the student's original trajectory.

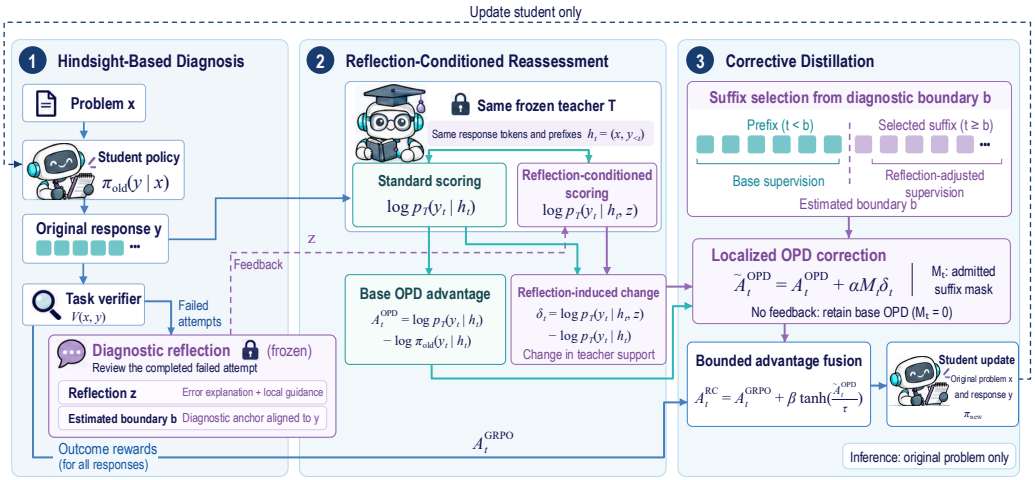


Figure 2: Overview of RC-OPD. For verifier-rejected student responses, a frozen teacher generates diagnostic feedback and an estimated intervention boundary. The teacher scores the same original response with and without this feedback, yielding token-level log-probability differences that adjust OPD supervision on the selected suffix while preserving base guidance elsewhere. The corrected distillation advantage is bounded and fused with the outcome-based GRPO advantage to update the student on the original problem–response pair.

3 METHODS

3.1 PRELIMINARIES

Let π_θ denote the student policy and p_T a frozen teacher. For $x \sim \mathcal{D}$, the rollout policy π_{old} samples a response group $\{y_i\}_{i=1}^G$. We write $\mathcal{B} = (x, \{y_i\}_{i=1}^G)$, $T_i = |y_i|$, and $h_{i,t} = (x, y_{i,<t})$, with $h_t = (x, y_{<t})$ for a generic response y.

On-Policy Distillation. OPD aligns student and teacher predictions on student-generated prefixes (Agarwal et al., 2024; Gu et al., 2024). We consider the token-local reverse-KL objective

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

With the rollout distribution fixed, the negative reverse-KL gradient at each prefix, evaluated at $\pi_\theta = \pi_{\text{old}}$, admits a score-function estimator weighted by $A_{i,t}^{\text{OPD}} = \log \frac{p_T(y_{i,t}|h_{i,t})}{\pi_{\text{old}}(y_{i,t}|h_{i,t})}$, which serves as the token-level distillation advantage.

Group Relative Policy Optimization. Given verifier rewards $r_i = R(x, y_i)$, GRPO computes a group-normalized advantage shared by all tokens of each response:

$$A_{i,t}^{\text{GRPO}} = \frac{r_i - \bar{r}}{s_r + \epsilon_r}, \qquad \rho_{i,t} = \frac{\pi_\theta(y_{i,t} \mid h_{i,t})}{\pi_{\text{old}}(y_{i,t} \mid h_{i,t})}, \tag{2}$$

where $\bar{r}$ and s_r are the group reward mean and standard deviation, $\epsilon_r > 0$ ensures numerical stability, and $\rho_{i,t}$ is the policy probability ratio. For fixed advantages $A = \{A_{i,t}\}$, the clipped surrogate is

$$\mathcal{J}_{\text{clip}}(\theta; A) = \mathbb{E}_{\mathcal{B}} \left[\frac{1}{G} \sum_{i=1}^{G} \frac{1}{T_i} \sum_{t=1}^{T_i} \min(\rho_{i,t} A_{i,t}, \text{clip}(\rho_{i,t}, 1-\epsilon_c, 1+\epsilon_c) A_{i,t}) \right], \tag{3}$$

where ϵ_c controls clipping. Without reference-policy KL penalty, $\mathcal{J}_{\text{GRPO}}(\theta) = \mathcal{J}_{\text{clip}}(\theta; A^{\text{GRPO}})$.

Joint Optimization through Advantage Fusion. In the sampled, unclipped surrogate, both signals weight the same policy ratio and combine into $A_{i,t}^{\text{joint}} = A_{i,t}^{\text{GRPO}} + \lambda A_{i,t}^{\text{OPD}}$, where $\lambda \geq 0$ controls distillation. Applying clipping after fusion yields

$$\mathcal{J}_{\text{joint}}(\theta) = \mathcal{J}_{\text{clip}}\left(\theta; A^{\text{joint}}\right). \tag{4}$$

Both advantages are computed from the collected trajectories and held fixed during policy updates. Setting $\lambda = 0$ recovers GRPO, while the distillation term adds token-dependent teacher supervision.

3.2 Hindsight-Based Diagnosis of Failed Reasoning

As shown in Figure 2, RC-OPD obtains diagnostic feedback by prompting a frozen teacher to assess completed student attempts that fail verification. Let $\mathcal{F}_x$ denote the indices of sampled responses reliably rejected by the task verifier. For each eligible attempt, the diagnostic prompt $\mathcal{P}_{\mathrm{diag}}(x, y_i)$ combines the original problem, the completed student response, and its verification outcome. For code generation, the context includes the submitted program and its execution summary. A single diagnostic generation produces a structured teacher response:

$$d_i \sim p_T(\cdot \mid \mathcal{P}_{\mathrm{diag}}(x, y_i)), \quad i \in \mathcal{F}_x. \tag{5}$$

The completed attempt provides hindsight context for interpreting earlier reasoning in light of subsequent deductions and self-corrections. The teacher is asked to identify a specific reasoning or implementation error that remains unresolved, explain its consequences, and provide a local corrective principle. This assessment accounts for whether an apparent mistake was later corrected or the associated reasoning was abandoned. When the supplied evidence does not support a specific diagnosis, the teacher may abstain. The core diagnostic instructions are provided in Appendix.

We construct the reflection z_i from the diagnostic explanation and corrective guidance in d_i. This text augments the teacher's scoring context, while the original student response remains the object of supervision. The next subsection evaluates the same response with and without z_i to measure reflection-induced changes in teacher predictions and defines how these changes contribute to corrective supervision. Baseline OPD and outcome supervision remain available for all sampled responses, including those without usable diagnostic feedback. Student optimization retains the original problem–response pair and its original outcome reward.

3.3 Reflection-Induced Correction

Standard OPD evaluates each sampled token using the teacher's predictions under the original problem and the preceding student response. RC-OPD augments this supervision with the retrospective information contained in z_i, while retaining the original student trajectory as the object of learning. To determine where this additional information applies, we align the diagnostic anchor with the original response and denote the resulting token boundary by b_i. Let $m_i \in \{0, 1\}$ indicate whether the feedback is admitted after diagnostic validation and teacher scoring. Admission requires a supported and usable diagnosis, a reliable alignment, and finite teacher scores over the selected suffix. We define the token-level correction mask as

$$M_{i,t} = \begin{cases} \mathbf{1}[t \geq b_i], & m_i = 1, \\ 0, & m_i = 0. \end{cases} \tag{6}$$

Thus, the reflection can modify supervision only from the diagnostic boundary onward. Earlier tokens and responses without admitted feedback retain their original OPD supervision.

We next quantify the information contributed by the reflection relative to the teacher's original predictions. For each selected token, the same frozen teacher evaluates the collected response under two contexts: the original context $h_{i,t} = (x, y_{i,<t})$ and the context augmented with z_i. Both evaluations use teacher forcing on the unchanged response, ensuring that the target token and its preceding student tokens are identical. The reflection-induced change in teacher support is defined as

$$\delta_{i,t} = \log \frac{p_T(y_{i,t} \mid h_{i,t}, z_i)}{p_T(y_{i,t} \mid h_{i,t})}, \quad M_{i,t} = 1, \tag{7}$$

with $\delta_{i,t} = 0$ whenever $M_{i,t} = 0$. A positive value indicates that the diagnostic context increases the probability assigned to the sampled token, whereas a negative value indicates reduced support. Accordingly, $\delta_{i,t}$ measures a change in the teacher's assessment rather than serving as an independent correctness label. The retained change adjusts the standard OPD advantage:

$$\widetilde{A}_{i,t}^{\mathrm{OPD}} = A_{i,t}^{\mathrm{OPD}} + \alpha M_{i,t} \delta_{i,t}, \quad \alpha \in [0, 1], \tag{8}$$

where α controls the strength of the reflection-conditioned correction. At selected positions, this update interpolates between the original and reflection-conditioned teacher log-ratio advantages. Setting $\alpha = 0$ recovers standard OPD, while $\alpha = 1$ adopts the reflection-conditioned teacher assessment on the selected suffix. Where the correction mask is zero, the advantage remains unchanged.

The reflection is used only to construct teacher-side supervision. The student continues to condition on the original problem and response prefix, and all teacher scores, masks, and advantage coefficients are held fixed during each policy update. Section 3.4 combines the corrected distillation advantage with the outcome advantage derived from verifiable rewards.

3.4 Corrective Distillation with Verifiable Rewards

The corrected distillation advantage distinguishes individual tokens within a student response, while the group-relative outcome advantage supplies a response-level learning signal. We combine them by applying a bounded transformation to the corrected teacher advantage:

$$A_{i,t}^{\mathrm{RC}} = A_{i,t}^{\mathrm{GRPO}} + \beta \, \tanh\left(\frac{\widetilde{A}_{i,t}^{\mathrm{OPD}}}{\tau}\right). \tag{9}$$

Here, $\beta \geq 0$ bounds the magnitude of the teacher contribution, and $\tau > 0$ controls its sensitivity and saturation scale, giving $|A_{i,t}^{\mathrm{RC}} - A_{i,t}^{\mathrm{GRPO}}| \leq \beta$. Near the origin, the teacher contribution is approximately $(\beta/\tau)\widetilde{A}_{i,t}^{\mathrm{OPD}}$; its sensitivity decreases as the adjusted advantage enters the saturation regime. The transformation acts on the entire corrected OPD advantage from Section 3.3. Thus, α controls how reflection revises teacher supervision, whereas β and τ control how that supervision enters the joint update. For fixed outcome supervision, baseline teacher scores, and admission decisions, the fused advantage is nondecreasing in the reflection-induced increment. Increased teacher support therefore raises or preserves the effective advantage, including on trajectories with a negative outcome advantage.

The reflection mask controls only the additional correction within $\widetilde{A}_{i,t}^{\mathrm{OPD}}$. Positions outside the selected suffix and responses without admitted feedback retain the transformed baseline OPD signal alongside outcome supervision. Setting $\alpha = 0$ recovers this baseline throughout the response, while setting $\beta = 0$ removes teacher supervision and recovers outcome-based GRPO under the same training surrogate. We optimize the student using the fused advantage in the shared policy surrogate. Writing $\mathcal{J}_{\mathrm{train}}$ for the training surrogate, including the rollout importance correction and dual clipping specified in the appendix, the final objective is

$$\mathcal{J}_{\mathrm{RC}}(\theta) = \mathcal{J}_{\mathrm{train}}\big(\theta; A^{\mathrm{RC}}\big). \tag{10}$$

Advantage fusion precedes policy clipping, so both supervision sources contribute to the same update coefficient. The fused advantages are held fixed during each policy update. Training uses the original student trajectories and original outcome rewards, with reflections entering through teacher supervision. At inference, the student answers the original problem without diagnostic feedback.

4 Experiments

4.1 Experimental Setup

Models and Datasets. We evaluate RC-OPD on mathematical reasoning and code generation with three student models: Qwen3-1.7B, Qwen3-4B, and Qwen3-8B. A frozen Qwen3-30B-A3B-Instruct-2507 model provides teacher supervision (Yang et al., 2025); all models operate in non-thinking mode. For training, we select approximately 57K problems with difficulty scores of at least 6 from DeepMath-103K (He et al., 2025) and use the 25K problems in Eurus-RL-Code (Cui et al., 2025) for code generation. Mathematical reasoning is evaluated on AIME 2024 (Mathematical Association of America, 2024), AIME 2025, HMMT February 2025, HMMT November 2025, and AMC 2023. Code generation is evaluated on HumanEval+, MBPP+ (Liu et al., 2023), and LiveCodeBench (v6, February May 2025) (Jain et al., 2025).

Baselines. We report the student checkpoint (Base) and the frozen teacher as reference points. The training baselines include GRPO-only (Shao et al., 2024), which learns from verifiable out-

Table 1: Mathematical reasoning results across five benchmarks. We report mean@16 accuracy and pass@16 for three student model sizes. **Best** and second-best are highlighted.

Method	AIME24		AIME25		HMMT-Feb		HMMT-Nov		AMC23		Average	
	Acc.	Pass	Acc.	Pass	Acc.	Pass	Acc.	Pass	Acc.	Pass	Acc.	Pass
Teacher	73.54	90.00	62.08	83.33	43.54	70.00	58.33	83.33	96.41	100.00	66.78	85.33
Student: Qwen3-1.7B												
Base	11.25	40.00	12.08	33.33	4.58	16.67	5.00	23.33	44.69	85.00	15.52	39.67
GRPO-only	36.04	63.33	30.42	56.67	18.12	33.33	13.96	36.67	75.47	92.50	34.80	56.50
OPD-only	39.17	56.67	29.79	**60.00**	17.08	26.67	17.50	33.33	72.81	95.00	35.27	54.33
ExOPD	36.25	63.33	33.33	50.00	15.21	30.00	15.62	40.00	76.41	95.00	35.36	55.67
Distilled RL	37.08	63.33	29.38	53.33	15.42	36.67	16.04	36.67	76.09	95.00	34.80	57.00
SAF-OPD	38.13	**70.00**	29.79	56.67	17.29	40.00	18.54	**43.33**	75.00	95.00	35.75	**61.00**
RC-OPD	**40.62**	66.67	**34.16**	56.67	**19.38**	**40.00**	**19.38**	40.00	**78.91**	**95.00**	**38.49**	59.67
Student: Qwen3-4B												
Base	25.42	56.67	22.92	56.67	11.88	30.00	8.33	30.00	65.47	95.00	26.80	53.67
GRPO-only	56.04	80.00	**53.12**	73.33	30.63	53.33	36.04	60.00	92.97	97.50	53.76	72.83
OPD-only	56.04	83.33	48.75	70.00	26.46	**56.67**	36.46	63.33	90.62	100.00	51.67	74.67
ExOPD	57.50	83.33	50.00	70.00	30.42	50.00	40.21	63.33	92.62	**100.00**	54.15	73.33
Distilled RL	57.50	83.33	50.42	70.00	29.17	53.33	38.54	63.33	92.81	97.50	53.69	73.50
SAF-OPD	58.54	83.33	50.42	70.00	**30.83**	53.33	38.75	60.00	92.81	97.50	54.27	72.83
RC-OPD	**64.17**	**83.33**	**53.75**	**76.67**	30.00	50.00	39.17	**66.67**	**94.22**	97.50	**56.26**	**74.83**
Student: Qwen3-8B												
Base	26.67	50.00	21.25	46.67	10.62	26.67	8.75	36.67	68.91	92.50	27.24	50.50
GRPO-only	62.29	83.33	51.46	73.33	28.54	46.67	37.50	60.00	89.69	100.00	53.90	72.67
OPD-only	59.58	83.33	49.38	70.00	29.58	53.33	39.79	**66.67**	93.44	100.00	54.35	74.67
ExOPD	62.50	83.33	51.25	76.67	32.08	56.67	44.79	63.33	**93.75**	**100.00**	56.87	76.00
Distilled RL	63.33	83.33	50.21	80.00	29.17	56.67	41.67	63.33	93.44	97.50	55.56	76.17
SAF-OPD	62.92	80.00	51.04	70.00	31.04	60.00	39.38	63.33	93.28	97.50	55.53	74.17
RC-OPD	**70.00**	**86.67**	**57.70**	**83.33**	**32.71**	56.67	**45.21**	63.33	**95.16**	97.50	**60.16**	**77.50**

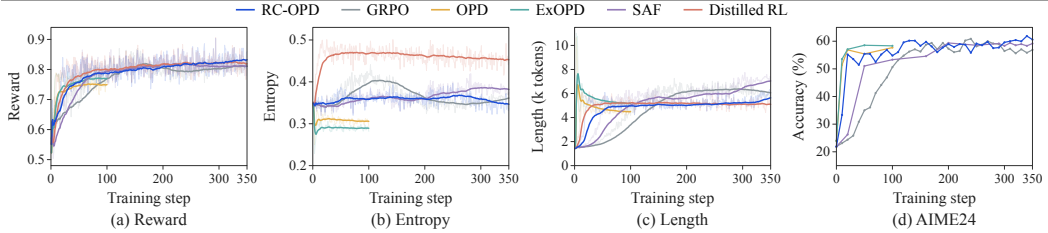


Figure 3: Training dynamics of RC-OPD and baseline methods during mathematical reasoning tasks.

come rewards; OPD-ONLY (Agarwal et al., 2024; Gu et al., 2024), which distills teacher predictions on student-generated trajectories; and EXOPD (Yang et al., 2026), which extrapolates the teacher-derived reward in generalized on-policy distillation. We also compare approaches that combine teacher guidance with outcome supervision. DISTILLED RL (Wang et al., 2026) reweights positive-advantage updates using clipped and normalized teacher-to-student likelihood ratios, while SAF-OPD (Ding et al., 2026) controls the magnitude and training-time strength of distillation advantages before fusing them with outcome advantages. GRPO+OPD provides a fixed-coefficient combination of the two advantages. All reported methods follow the same evaluation protocol.

Implementation and Evaluation. We implement RC-OPD with verl (Sheng et al., 2025) and generate student rollouts using vLLM. Student prompts are capped at 2,048 tokens during training; response limits are 16,384 tokens for mathematics and 8,192 for code in both training and evaluation. For each eligible failed response, the frozen teacher generates a critique in one online call using greedy decoding, capped at 1,024 tokens. We set $(\alpha, \beta, \tau) = (1, 0.1, 1)$. The teacher, initial checkpoints, and trained students follow the same benchmark-specific evaluation protocols. For mathematics, we sample 16 responses per problem and report mean@16 accuracy, with answers checked by Math-Verify. Code solutions are evaluated against benchmark test suites: HumanEval+ and MBPP+ use one greedily decoded solution per problem, while on LiveCodeBench we sample four programs per problem and report empirical pass@1. Further training and decoding settings are provided in the Appendix.

4.2 MAIN RESULTS

Mathematical Reasoning. As shown in Table 1, RC-OPD achieves the highest average mean@16 accuracy among student methods at all scales, reaching 38.49%, 56.26%, and 60.16% for Qwen3-

Table 2: Code generation results on three benchmarks. **Best** and second-best are highlighted.

Method	*Qwen3-1.7B*				*Qwen3-4B*				*Qwen3-8B*			
	HEval+	MBPP+	LCB	Avg.	HEval+	MBPP+	LCB	Avg.	HEval+	MBPP+	LCB	Avg.
Base	60.98	54.23	15.14	43.45	79.27	63.49	24.57	55.78	80.49	72.49	23.71	58.90
GRPO-only	65.24	53.70	17.57	45.50	80.49	68.78	32.14	60.47	81.10	72.22	28.85	60.72
OPD-only	70.73	58.73	25.86	51.77	78.66	69.31	30.71	59.56	84.15	71.96	33.29	63.13
GRPO+OPD	70.12	56.08	26.43	50.88	79.88	66.14	33.86	59.96	78.66	71.69	34.86	61.74
Distilled RL	69.97	57.47	27.43	51.62	84.60	69.91	38.29	64.27	84.15	73.41	38.86	65.47
SAF-OPD	70.73	57.14	26.29	51.39	82.93	70.63	34.43	62.66	81.10	71.69	37.43	63.41
RC-OPD	**71.19**	**59.20**	**28.71**	**53.03**	**86.13**	**70.97**	**38.29**	**65.13**	**87.20**	**74.07**	**42.86**	**68.04**

Table 3: Component ablations on mathematical reasoning with Qwen3-4B.

Method/Setting	AIME24	AIME25	HMMT-Feb	HMMT-Nov	AMC23	Avg.
RC-OPD (full)	64.17	53.75	30.00	39.17	94.22	56.26
⊢ w/o Reflection	62.29	51.46	29.17	37.71	93.44	54.81
⊢ w/o Localization	63.33	52.71	30.42	38.54	94.06	55.81
⊢ Random-span control	62.71	51.88	28.96	37.29	93.59	54.89
⊢ Residual-only guidance	61.25	50.42	28.54	36.67	93.28	54.03

1.7B, 4B, and 8B, respectively. These results exceed the strongest baseline at each scale–SAF-OPD for 1.7B and 4B, and ExOPD for 8B–by +2.74, +1.99, and +3.29 percentage points. Gains over methods already combining teacher guidance with outcome rewards suggest that retrospective diagnosis provides complementary supervision. Relative to OPD-only, RC-OPD improves every benchmark across all scales, with average gains of +3.22, +4.59, and +5.81 points, demonstrating consistent benefits across student capacities and tasks. At both 1.7B and 8B, it outperforms all competing student methods on all five benchmarks, with particularly large 8B gains on AIME24 and AIME25 (+6.67 and +6.24 points over the strongest per-benchmark baselines). RC-OPD also improves average pass@16 over OPD-only at all scales, reaching 59.67%, 74.83%, and 77.50%, and achieving the highest student averages at 4B and 8B, indicating broader problem coverage.

Code Generation. Results in Table 2 show that RC-OPD achieves the highest average score across all three model scales, outperforming the strongest average baseline, Distilled RL, by +1.41, +0.86, and +2.57 percentage points for Qwen3-1.7B, 4B, and 8B, respectively. The gains become particularly pronounced at 8B, where RC-OPD achieves the best performance on all three benchmarks, including a +4.00-point improvement on LCB over the strongest baseline. These results show that retrospective diagnosis also provides effective corrective supervision for program generation, extending the benefits of RC-OPD beyond mathematical reasoning.

Training Dynamics. Figure 3 shows that RC-OPD maintains stable optimization throughout training while achieving stronger performance as training progresses. This suggests that reflection-induced corrective supervision can be effectively incorporated into on-policy distillation and translated into improved reasoning performance without destabilizing the underlying learning process.

4.3 Ablation Studies and Analysis

We examine the contributions of reflection, region selection, and the base distillation signal on the mathematical reasoning benchmarks in Table 3, complemented by a visualization of reflection-induced changes in token-level distillation signals, illustrating the corrective supervision.

Effects of Reflection and Region Selection. The *w/o Reflection* variant removes the reflection-induced adjustment while retaining base OPD and outcome supervision, isolating the contribution of retrospective diagnosis beyond standard teacher predictions and task rewards. Average accuracy decreases from 56.26% to 54.81%, supporting the additional value of incorporating the teacher's diagnosis of student failures into the distillation signal. The *w/o Localization* variant extends reflection adjustments from the selected suffix to every valid token of an admitted response, while preserving the diagnostic procedure and localization-based admission criteria. This comparison assesses the effect of restricting corrective guidance to the selected suffix. The variant achieves 55.81%, exceeding the variant without reflection but falling below full RC-OPD by 0.45 points. These results suggest that reflection provides useful information even when applied throughout an admitted response, with further gains from restricting its influence to the selected suffix.

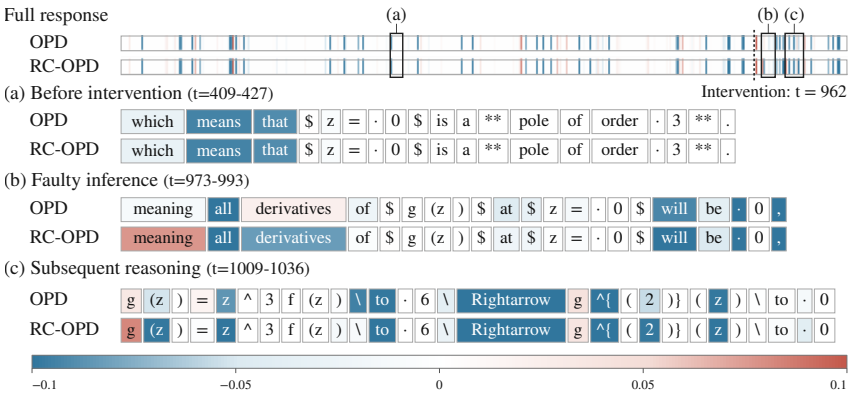


Figure 4: Token-level effects of reflection-induced correction. The upper strips show the same failed student response, with the dashed line marking the estimated intervention boundary. Enlarged excerpts illustrate (a) preservation of the base signal before intervention and reflection-induced adjustments at (b) the faulty inference and (c) subsequent reasoning.

Since extending the application region also increases the number of affected tokens, we introduce *Random-span control* to examine region selection at a matched token count. This variant retains the reflective feedback but applies the token-specific adjustments within a uniformly sampled contiguous window of the same length as the selected suffix. Average accuracy drops by 1.37% below full RC-OPD. This comparison supports the value of selecting intervention regions through diagnosis beyond simply limiting the number of tokens receiving reflection adjustments.

Role of the Base Distillation Signal. The *Residual-only guidance* variant removes the base OPD term while retaining the masked reflection increment and outcome supervision, assessing the contribution of ordinary distillation alongside localized corrective guidance. Standard teacher scores are still used to compute the reflection difference, but no longer contribute through a separate base OPD term. Tokens outside the selected region and responses without admitted feedback consequently receive only outcome supervision. This variant achieves 54.03%, a reduction of 2.23 percentage points, highlighting the value of preserving the broader teacher guidance supplied by base OPD. The result supports using reflection-induced changes to supplement the existing distillation signal.

Reflection-Induced Changes in Distillation Signals. We consider computing the residue of $f(z) = 1/(z - \sin z)$ at $z = 0$. Figure 4 compares the bounded token-level distillation signals of OPD and RC-OPD on the failed response. The student correctly identifies a third-order pole but incorrectly infers from $g(z) = z^3 f(z) \to 6$ that its higher-order derivatives vanish at the origin. After the teacher diagnoses this error, the signal for `derivatives` switches from $+0.022$ to -0.083 in panel (b), showing that retrospective diagnosis can reverse a contribution favored by standard OPD. The subsequent derivative-order token receives a stronger negative signal in panel (c), while neighboring tokens such as `meaning` receive positive adjustments, demonstrating token-specific changes in both direction and strength. Before the estimated intervention boundary, the base signal is retained by design, as shown in panel (a). Overall, RC-OPD incorporates failure-specific diagnostic information into the existing distillation signal while preserving the student's original trajectory.

5 Conclusion

In this work, we propose RC-OPD, a framework for teaching with hindsight that translates retrospective teacher diagnoses into corrective supervision on student-generated trajectories. A frozen teacher reviews failed responses and evaluates the same original tokens under standard and reflection-conditioned contexts. Their log-probability differences make the diagnostic contribution explicit relative to base OPD, allowing its influence to be controlled separately. RC-OPD selectively applies these adjustments over diagnosed reasoning suffixes while retaining standard guidance elsewhere, and bounds the combined distillation signal before integrating it with outcome-based advantages. This design connects the teacher's reasoning about a completed failure with targeted learning on the student's own attempt, preserving both predictive guidance and optimization toward task success. Experiments demonstrate improvements over strong reinforcement-learning and distillation baselines. Future work can further improve diagnostic reliability and training efficiency.

6 AI USE STATEMENT

Generative AI tools were used to assist with manuscript writing and refinement, including translation and language editing, literature search and the identification of relevant prior work, as well as code implementation and debugging. The authors independently verified all AI-assisted content and take full responsibility for the final manuscript, implementation, experimental results, scientific claims, and references.

7 REPRODUCIBILITY STATEMENT

We provide the complete formulation of RC-OPD in Section 3 and detailed training, reflection, and evaluation configurations in Appendix D. The diagnostic prompt, analytical properties, and additional qualitative analyses are provided in the appendices. Code will be released upon acceptance.

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.

Ganqu Cui, Lifan Yuan, Zefan Wang, Hanbin Wang, Wendi Li, Bingxiang He, Yuchen Fan, Tianyu Yu, Qixin Xu, Weize Chen, et al. Process reinforcement through implicit rewards. *arXiv preprint arXiv:2502.01456*, 2025.

DeepSeek-AI. DeepSeek-R1: Incentivizing reasoning capability in LLMs via reinforcement learning. *arXiv preprint arXiv:2501.12948*, 2025.

DeepSeek-AI, Anyi Xu, Bangcai Lin, Bing Xue, et al. Deepseek-v4: Towards highly efficient million-token context intelligence, 2026.

Yifan Ding, Xincheng Wei, Yoshua Y. Li, Ziheng Li, Yuquan Lu, Siyu Zhang, Dongsheng Ma, Rongxiang Weng, Xunliang Cai, and Yun Chen. SAF-OPD: Stable advantage fusion for on-policy distillation. *arXiv preprint arXiv:2607.29209*, 2026.

GLM-5-Team, :, Aohan Zeng, Xin Lv, Zhenyu Hou, et al. Glm-5: from vibe coding to agentic engineering, 2026.

Yuxian Gu, Li Dong, Furu Wei, and Minlie Huang. MiniLLM: Knowledge distillation of large language models. In *International Conference on Learning Representations*, 2024.

Zhuowen Han, Jinwei Xiao, Zhengxi Lu, Renren Jin, Zhiyuan Yao, Yuxin Liu, Hongyan Hao, Yueqing Sun, Yu Yang, Qi Gu, Xunliang Cai, and Deyi Xiong. Distill where you fail: Recovering learning signals of negative RL-groups from adaptive teacher guidance. *arXiv preprint arXiv:2608.00782*, 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.

Zhiwei He, Tian Liang, Jiahao Xu, Qiuzhi Liu, Xingyu Chen, Yue Wang, Linfeng Song, Dian Yu, Zhenwen Liang, Wenxuan Wang, Zhuosheng Zhang, Rui Wang, Zhaopeng Tu, Haitao Mi, and Dong Yu. DeepMath-103K: A large-scale, challenging, decontaminated, and verifiable mathematical dataset for advancing reasoning. *arXiv preprint arXiv:2504.11456*, 2025.

Byeongho Heo, Jaehui Hwang, Sangdoo Yun, and Dongyoon Han. On-policy delta distillation. *arXiv preprint arXiv:2607.15161*, 2026.

Geoffrey E. Hinton, Oriol Vinyals, and Jeffrey Dean. Distilling the knowledge in a neural network. *ArXiv*, abs/1503.02531, 2015.

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

Jingcheng Hu, Yinmin Zhang, Qi Han, Daxin Jiang, Xiangyu Zhang, and Heung-Yeung Shum. Open-reasoner-zero: An open source approach to scaling up reinforcement learning on the base model. In *Advances in Neural Information Processing Systems*, volume 38, 2025.

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

Naman Jain, Alex Gu, Wen-Ding Li, Fanjia Yan, Tianjun Zhang, Sida Wang, Armando Solar-Lezama, Koushik Sen, and Ion Stoica. Livecodebench: Holistic and contamination free evaluation of large language models for code. In *International Conference on Learning Representations*, pp. 58791–58831, 2025.

Can Jin, Jiakang Li, Rui Wu, Eddy Z. Zhang, and Dimitris N. Metaxas. Weak critics make strong learners: On-policy critique distillation for scalable oversight. *arXiv preprint arXiv:2606.00424*, 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.

Yoon Kim and Alexander M. Rush. Sequence-level knowledge distillation. *ArXiv*, abs/1606.07947, 2016.

Aviral Kumar, Vincent Zhuang, Rishabh Agarwal, Yi Su, John D Co-Reyes, Avi Singh, Kate Baumli, Shariq Iqbal, Colton Bishop, Rebecca Roelofs, Lei M Zhang, Kay McKinney, Disha Shrivastava, Cosmin Paduraru, George Tucker, Doina Precup, Feryal Behbahani, and Aleksandra Faust. Training language models to self-correct via reinforcement learning. In *International Conference on Learning Representations*, 2025.

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

Zicheng Lin, Zhibin Gou, Tian Liang, Ruilin Luo, Haowei Liu, and Yujiu Yang. CriticBench: Benchmarking LLMs for critique-correct reasoning. In *Findings of the Association for Computational Linguistics*, pp. 1552–1587, 2024.

Jiawei Liu, Chunqiu Steven Xia, Yuyao Wang, and Lingming Zhang. Is your code generated by ChatGPT really correct? rigorous evaluation of large language models for code generation. In *Advances in Neural Information Processing Systems*, volume 36, 2023.

Shiqi Liu, Zeyu He, Letian Tao, Guojian Zhan, Jiaxin Gao, Feihong Zhang, Jingliang Duan, Wei Xiong, Kehua Sheng, Bo Zhang, Yang Guan, and Shengbo Eben Li. Beyond token-local imitation: Reward-compatible temporal credit assignment for on-policy distillation. *arXiv preprint arXiv:2609.16937*, 2026.

Zichen Liu, Changyu Chen, Wenjun Li, Penghui Qi, Tianyu Pang, Chao Du, Wee Sun Lee, and Min Lin. Understanding R1-Zero-like training: A critical perspective. *arXiv preprint arXiv:2503.20783*, 2025.

Aman Madaan, Niket Tandon, Prakhar Gupta, Skyler Hallinan, Luyu Gao, Sarah Wiegreffe, Uri Alon, Nouha Dziri, Shrimai Prabhumoye, Yiming Yang, Shashank Gupta, Bodhisattwa Prasad Majumder, Katherine Hermann, Sean Welleck, Amir Yazdanbakhsh, and Peter Clark. Self-Refine: Iterative refinement with self-feedback. In *Advances in Neural Information Processing Systems*, volume 36, 2023.

Mathematical Association of America. American invitational mathematics examination (AIME) 2024, 2024.

OpenAI. GPT-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023.

Yuxiao Qu, Tianjun Zhang, Naman Garg, and Aviral Kumar. Recursive introspection: Teaching language model agents how to self-improve. In *Advances in Neural Information Processing Systems*, volume 37, 2024.

Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, Y. K. Li, Y. Wu, and Daya Guo. DeepSeekMath: Pushing the limits of mathematical reasoning in open language models. *arXiv preprint arXiv:2402.03300*, 2024.

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

Noah Shinn, Federico Cassano, Ashwin Gopinath, Karthik Narasimhan, and Shunyu Yao. Reflexion: Language agents with verbal reinforcement learning. In *Advances in Neural Information Processing Systems*, volume 36, 2023.

Kumar Shridhar, Alessandro Stolfo, and Mrinmaya Sachan. Distilling reasoning capabilities into smaller language models. In *Findings of the Association for Computational Linguistics*, pp. 7059–7073, 2023.

Kimi Team, Tongtong Bai, Yifan Bai, Yiping Bao, M. C., Jianfeng Cai, Xinyuan Cai, et al. Kimi k3: Open frontier intelligence, 2026.

Chen Wang, Zhaochun Li, Jionghao Bai, Yining Zhang, Hexuan Deng, Ge Lan, and Yue Wang. Distilled reinforcement learning for LLM post-training. *arXiv preprint arXiv:2607.17247*, 2026.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, Chujie Zheng, Dayiheng Liu, Fan Zhou, Fei Huang, Feng Hu, Hao Ge, Haoran Wei, Huan Lin, Jialong Tang, Jian Yang, 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.

Qiying Yu, Zheng Zhang, Ruofei Zhu, Yufeng Yuan, Xiaochen Zuo, Yu Yue, Weinan Dai, Tiantian Fan, Gaohong Liu, Lingjun Liu, Xin Liu, Haibin Lin, Zhiqi Lin, Bole Ma, Guangming Sheng, et al. DAPO: An open-source LLM reinforcement learning system at scale. *arXiv preprint arXiv:2503.14476*, 2025.

Weihao Zeng, Yuzhen Huang, Qian Liu, Wei Liu, Keqing He, Zejun Ma, and Junxian He. Simplerl-zoo: Investigating and taming zero reinforcement learning for open base models in the wild. In *Proceedings of the Second Conference on Language Modeling*, 2025.

Chujie Zheng, Zhenru Zhang, Beichen Zhang, Runji Lin, Keming Lu, Bowen Yu, Dayiheng Liu, Jingren Zhou, and Junyang Lin. ProcessBench: Identifying process errors in mathematical reasoning. In *Annual Meeting of the Association for Computational Linguistics*, pp. 1009–1024, 2025.

A More Related Work

A.1 Reinforcement Learning with Verifiable Rewards

Reinforcement learning with verifiable rewards (RLVR) optimizes language models using automatically checked outcomes, such as mathematical answers and program test results (DeepSeek-AI, 2025; Zeng et al., 2025; Hu et al., 2025; Cui et al., 2025). Group Relative Policy Optimization (GRPO) estimates advantages from groups of sampled responses without learning a separate value function (Shao et al., 2024). DeepSeek-R1-Zero demonstrates substantial reasoning improvements through large-scale reinforcement learning with rule-based rewards (DeepSeek-AI, 2025). DAPO improves exploration and training stability through decoupled clipping, dynamic sampling, and token-level loss aggregation (Yu et al., 2025). Dr. GRPO addresses length and difficulty biases arising from normalization in the GRPO objective (Liu et al., 2025). Under outcome supervision, tokens within a response share the same reward-derived advantage (Shao et al., 2024), leaving the contributions of individual reasoning decisions unspecified.

B Algorithm and Implementation Details

B.1 Diagnostic Reflection Prompt

We provide the diagnostic prompt used by the frozen teacher in Box 1. Given a problem and a verifier-rejected student attempt, the teacher examines the completed trajectory to identify an unresolved error and provide a local corrective principle. The prompt explicitly accounts for subsequent self-corrections and allows abstention when the available evidence does not support a specific error. It excludes final answers and complete replacement solutions.

Teacher Diagnostic Instructions

Problem: {`problem`}

Student Attempt: {`completed_student_attempt`}

Verification Outcome: The attempt was not accepted by the task verifier.

Diagnostic Instructions.
Examine the complete attempt, including later revisions and self-corrections. Identify a specific reasoning or implementation error that remains unresolved. Ground your assessment in the supplied problem, student attempt, and verification feedback. Explain how the identified error affects the solution and provide a local corrective principle. Do not flag a mistake that was subsequently corrected or abandoned. If the evidence does not support a specific unresolved error, abstain from providing corrective guidance.

Response Requirements.
Return a single structured assessment. Set `verdict` to `supported_error` when a specific unresolved error is supported. Otherwise, select `no_supported_error`, `ambiguous_problem`, or `suspected_format_issue`, as appropriate. For a supported error, explain the issue and its consequences in `diagnosis`, and state the local corrective principle in `correction`. For other verdicts, explain the reason for your assessment in `diagnosis` and leave `correction` empty. Do not provide the final answer, a complete replacement solution, or a fully rewritten program.

C Analytical Properties of RC-OPD

We analyze how reflection changes teacher supervision on student trajectories. The analysis connects localized reassessment to teacher likelihoods, bounds the resulting coefficient changes, and establishes a condition under which bounded fusion preserves the signs of binary-reward GRPO advantages. These properties hold without assuming that the diagnostic feedback is correct.

For a fixed problem x and response $y = (y_1, \ldots, y_L)$, let $h_t = (x, y_{<t})$. Write $p_t^0 = p_T(y_t \mid h_t)$, $p_t^z = p_T(y_t \mid h_t, z)$, and $q_t = \pi_{\text{old}}(y_t \mid h_t)$, where z is the diagnostic feedback. Assume that the probabilities being scored are positive, and let

$$D_t = \log p_t^0 - \log q_t, \quad \delta_t = \log p_t^z - \log p_t^0. \tag{11}$$

Let $M_t \in \{0,1\}$ indicate membership in an admitted correction suffix, and let g denote the response's outcome-based GRPO advantage. The fused token-level coefficient is

$$A_t(M) = g + \beta \tanh\left(\frac{D_t + \alpha M_t \delta_t}{\tau}\right), \quad 0 \le \alpha \le 1, \quad \beta \ge 0, \quad \tau > 0. \tag{12}$$

The correction is defined to be zero wherever $M_t = 0$, so reflection-conditioned scoring is not required at those positions. Unless stated otherwise, comparisons hold the trajectory, diagnostic feedback, outcome advantage, teacher scores, and hyperparameters fixed.

C.1 Interpretation of Reflection-Conditioned

At the experimental setting $\alpha = 1$, substituting the definition of δ_t gives

$$D_t + M_t \delta_t = \begin{cases} \log(p_t^0/q_t), & M_t = 0, \\ \log(p_t^z/q_t), & M_t = 1. \end{cases} \tag{13}$$

The residual replaces the teacher score with the reflection-conditioned score on the selected suffix, while retaining the base score elsewhere. Both scores supervise the same original student tokens. For a nonempty suffix beginning at token b, the raw residuals also satisfy

$$\sum_{t=b}^{L} \delta_t = \log \frac{p_T(y_{b:L} \mid x, y_{<b}, z)}{p_T(y_{b:L} \mid x, y_{<b})}. \tag{14}$$

This identity follows from the autoregressive chain rule, using the same fixed diagnostic feedback at every reflection-conditioned scoring position. If an EOS token is scored, it is included in both products. The sum therefore measures how the diagnosis changes teacher support for the observed suffix. It does not certify the correctness of that suffix or the preceding tokens. Because the diagnosis and boundary are obtained from the completed response, this scoring identity does not establish equivalence to distillation from a single fixed teacher policy; stopping gradients through the feedback does not remove its statistical dependence on the response.

C.2 Locality and Bounded Influence of Reflection

Compare Eq. 12 with the coefficient using the same outcome signal, base teacher, and compression but no reflection residual:

$$A_t(0) = g + \beta \tanh(D_t/\tau), \quad \Delta A_t = A_t(M) - A_t(0). \tag{15}$$

Proposition 1 (localized and bounded coefficient changes). For every valid response token

$$M_t = 0 \implies \Delta A_t = 0, \quad |\Delta A_t| \le \beta M_t \min\left\{2, \frac{\alpha|\delta_t|}{\tau}\right\}. \tag{16}$$

If $M_t = 1$, $\alpha > 0$, and $\beta > 0$, then $\operatorname{sgn}(\Delta A_t) = \operatorname{sgn}(\delta_t)$.

Proof. When $M_t = 0$, the two coefficients coincide. Otherwise, $|\tanh'(u)| \le 1$ gives $|\Delta A_t| \le \beta\alpha|\delta_t|/\tau$ by the mean value theorem. The range of $\tanh$ also gives $|\Delta A_t| \le 2\beta$. Taking the smaller bound establishes the inequality, and strict monotonicity of $\tanh$ gives the sign statement. □

Let $\mathcal{S} = \{t : M_t = 1\}$ and $f = |\mathcal{S}|/L$. For a nonempty selected suffix, averaging the two bounds in Proposition 1 separately yields

$$\frac{1}{L}\sum_{t=1}^{L} |\Delta A_t| \le \beta f \min\left\{2, \frac{\alpha}{\tau}\frac{1}{|\mathcal{S}|}\sum_{t\in\mathcal{S}} |\delta_t|\right\}. \tag{17}$$

When $\mathcal{S}$ is empty, the average perturbation is zero. The same argument applies to a minibatch using its total number of valid response tokens as the denominator. This bound separates the coverage of reflection from its reassessment magnitude within the selected region. In particular, an average over selected suffix tokens alone does not determine the average perturbation over all tokens. The bound also accounts for saturation: large raw scoring changes need not produce proportionally large coefficient changes.

Corollary (sensitivity to the suffix boundary). For the same response, diagnostic feedback, and admission decision, consider suffix masks M^b and $M^{\widetilde{b}}$ beginning at $b, \widetilde{b} \in \{1, \ldots, L+1\}$, where $L+1$ denotes an empty suffix. Then

$$\frac{1}{L}\sum_{t=1}^{L} |A_t(M^b) - A_t(M^{\widetilde{b}})| \le 2\beta \frac{|b - \widetilde{b}|}{L}. \tag{18}$$

For an admitted reflection, the masks differ at exactly $|b - \widetilde{b}|$ positions. The coefficients coincide elsewhere, and Proposition 1 bounds each changed position by 2β, proving the result. If the reflection is not admitted, both corrections are zero. This bound quantifies the effect of moving a token boundary while holding the diagnostic content fixed; regenerating the diagnosis can also change the scores at other positions.

C.3 Sign Preservation under Bounded Advantage Fusion

Reflection may reverse the distillation component even though its contribution to the fused coefficient is bounded. Since $\tanh$ preserves signs, for $\alpha > 0$, $\beta > 0$, $M_t = 1$, and $D_t \neq 0$, a strict reversal occurs exactly when

$$D_t(D_t + \alpha\delta_t) < 0 \quad \Longleftrightarrow \quad D_t\delta_t < 0 \text{ and } \alpha|\delta_t| > |D_t|. \tag{19}$$

This is the reversal measured by the distillation-direction statistic. The sign of the full coefficient also depends on the outcome advantage. In particular,

$$|A_t(M) - g| \le \beta, \qquad |g| > \beta \implies \operatorname{sgn}(A_t(M)) = \operatorname{sgn}(g). \tag{20}$$

The first inequality follows from $|\tanh u| \le 1$, and the second follows because a perturbation of magnitude at most β cannot reverse a coefficient whose magnitude exceeds β. For binary-reward GRPO, the following result gives an explicit sufficient condition on the teacher weight.

Proposition 2 (sign preservation in mixed-reward GRPO groups). Consider a group of $G \ge 2$ responses with binary rewards, containing $k \in \{1, \ldots, G-1\}$ successful responses. Suppose that GRPO uses sample-standard-deviation normalization with $\epsilon_r > 0$:

$$\bar{r} = \frac{k}{G}, \quad s_k = \sqrt{\frac{k(G-k)}{G(G-1)}}, \quad g_i = \frac{r_i - k/G}{s_k + \epsilon_r}. \tag{21}$$

Across all such groups, the smallest absolute outcome advantage is

$$m_G \equiv \min_{1 \le k \le G-1} \min_i |g_i| = \frac{1}{\sqrt{G} + G\epsilon_r}. \tag{22}$$

Consequently, if $\beta < m_G$, then every fused coefficient $A_{i,t}(M)$, formed using response i's advantage g_i, satisfies $\operatorname{sgn}(A_{i,t}(M)) = \operatorname{sgn}(g_i)$ regardless of the diagnostic content or suffix boundary.

Proof. Before adding the stabilizer, the absolute advantages for rewards one and zero are:

$$\frac{1 - k/G}{s_k} = \sqrt{\frac{(G-1)(G-k)}{Gk}}, \quad \frac{k/G}{s_k} = \sqrt{\frac{(G-1)k}{G(G-k)}}. \tag{23}$$

Both are at least $1/\sqrt{G}$, and $s_k \ge 1/\sqrt{G}$. Hence

$$|g_i| = \frac{|r_i - k/G|/s_k}{1 + \epsilon_r/s_k} \ge \frac{1/\sqrt{G}}{1 + \epsilon_r\sqrt{G}} = \frac{1}{\sqrt{G} + G\epsilon_r}. \tag{24}$$

Equality holds for a reward-zero response when $k = 1$, or a reward-one response when $k = G - 1$. This proves the exact minimum. Eq. 20 gives sign preservation whenever $\beta < m_G$. □

For $G = 8$ and $\epsilon_r = 10^{-6}$, $m_G \approx 0.3535524$. Under the stated normalization, the schedule $\beta \le 0.1$ therefore preserves outcome-advantage signs in mixed-reward groups, even when the distillation component reverses. Reflection can still change coefficient magnitudes and the relative supervision across tokens. When all rewards are equal, $g_i = 0$ because $\epsilon_r > 0$, and the teacher term determines

Table 4: Method-specific baseline settings. The Distilled RL ratio bound is applied before geometric normalization. SAF warm-up concerns the distillation weight, not the optimizer learning rate.

Method	Hyperparameter	Value
ExOPD	Extrapolation coefficient	1.25
Distilled RL	Reweighted responses	Positive GRPO advantage
	Pre-normalization ratio bound	3
	Within-response geometric normalization	Enabled
SAF-OPD	Token retention fraction	20%
	Tanh compression amplitude	0.1
	Relative KL-reduction threshold	0.2
	Maximum distillation warm-up steps	100
	Annealing enabled	Yes
	Configured `anneal_steps`	200

the fused coefficients. All-failure groups can thus receive nonzero teacher supervision despite a vanishing outcome signal, with reflection modifying supervision on admitted suffixes.

These results characterize supervision coefficients. The binary-reward analysis assumes the normalization in Eq. 21, without additional reward shaping or subsequent advantage rescaling. The sign-preservation condition applies generally to bounded teacher fusion, with reflection and suffix selection shaping this contribution in RC-OPD. Since tokens share model parameters, localized coefficient changes need not produce localized parameter updates.

D Detailed Experimental Setup

Models and Training Data. We use Qwen3-1.7B, Qwen3-4B, and Qwen3-8B as students and Qwen3-30B-A3B-Instruct-2507 as the frozen teacher. Separate student models are trained for mathematical reasoning and code generation, using the filtered DeepMath-103K split and the prepared Eurus-RL-Code split, respectively. Student and teacher generation use non-thinking mode. Mathematical rewards are binary outcomes from the shared answer verifier. For code generation, the reward is one if the generated program passes all required tests, and zero otherwise. The Base and Teacher entries evaluate the corresponding original checkpoints without additional task-specific training. We organize the training setups into two groups: GRPO-based methods, which use group-relative outcome advantages, and OPD-based methods, which optimize distillation objectives without an outcome-based RL term. GRPO-based methods may also incorporate distillation guidance. Optimization settings are described within each group below.

GRPO-based Methods. All methods in this group use outcome-based GRPO advantages. GRPO-only uses no teacher supervision, while GRPO+OPD adds the base token-level distillation advantage with fixed weights. In the Qwen3-4B mathematics runs, Distilled RL reweights only positive GRPO advantages using ratios of teacher to rollout-policy token probabilities, clipped before within-response geometric normalization. SAF-OPD fuses GRPO and OPD advantages through token selection, bounded compression, adaptive warm-up, and annealing. Table 4 reports the method-specific baseline settings for the mathematics runs.

We implement RC-OPD in `verl`, with `vLLM` for on-policy generation. We use full-parameter optimization and GRPO advantages with within-group standard-deviation normalization. Training and reflection settings are given in Tables 5 and 6, respectively. While teacher supervision is active, the frozen teacher jointly generates a diagnosis and an intervention boundary for each verifier-rejected attempt admitted by the reflection pipeline. The reflection correction is applied to the selected suffix of the original student response, while base OPD supervision is retained outside this suffix. The same-sign gate is disabled.

OPD-based Methods. This group comprises OPD-only and ExOPD. Both obtain their training advantages from teacher guidance without an outcome-based GRPO term. OPD-only trains on student-generated trajectories using the difference between frozen-teacher and rollout-policy token log-probabilities as the distillation advantage. ExOPD extrapolates the teacher's implicit reward relative to a frozen base policy before constructing the distillation advantage. Both methods use a

Table 5: Training configuration of RC-OPD.

Hyperparameter	Mathematics	Code generation
Prompts per iteration	128	128
Responses per prompt	8	8
Responses per iteration	1,024	1,024
Maximum student prompt length	2,048	2,048
Maximum student response length	16,384	8,192
Rollout temperature	1.0	1.0
Rollout top-p	1.0	1.0
Optimizer	AdamW	AdamW
Actor learning rate	10^{-6}	10^{-6}
Learning-rate schedule	Constant	Constant
Learning-rate warm-up steps	0	0
Optimization epochs per rollout batch	1	1
Policy-ratio clipping range	$[0.8, 1.2]$	$[0.8, 1.2]$
Maximum gradient norm	1.0	1.0
Loss aggregation	Token mean	Token mean
Entropy-bonus coefficient	0	0
Additional actor KL-loss coefficient	0	0

Table 6: Reflection and advantage-fusion settings for RC-OPD.

Hyperparameter	Value
Diagnosis and localization	Joint, single generation call
Diagnostic decoding	Greedy
Maximum diagnostic input length	16,384
Maximum diagnostic output length	1,024
Maximum teacher-scoring sequence length	32,768
Block-size parameter	1,200 characters
Reflection correction region	Selected suffix
Same-sign gating	Disabled
Reflection strength α	1.0
Compression temperature τ	1.0
Initial teacher weight β	0.1

learning rate of 10^{-5} and sample one response for each of 1,024 prompts per iteration; ExOPD uses an extrapolation coefficient of 1.25.

Evaluation Protocol. Mathematical evaluation covers AIME 2024, AIME 2025, HMMT February 2025, HMMT November 2025, and AMC 2023. Code evaluation covers HumanEval+, MBPP+, and LiveCodeBench. For each mathematical problem, we sample 16 responses, compute the fraction with correct final answers, and average this fraction across problems. For HumanEval+ and MBPP+, we generate one program per problem using greedy decoding and report the fraction of problems for which the program passes all required tests. For LiveCodeBench, we sample four programs per problem and report the fraction of problems for which at least one program passes all required tests. Sampling-based evaluation uses a temperature of 1.0 and top-p of 1.0. Mathematical answers are checked with the shared Math-Verify-based scorer, and generated programs are evaluated with the corresponding benchmark tests. Students receive only the benchmark input at evaluation time, without teacher feedback or additional revision rounds.

E Extended Ablations and Analyses

E.1 Sensitivity to Reflection Strength

We examine the sensitivity of RC-OPD to the reflection strength α, which controls the contribution of the reflection-induced correction relative to the base OPD signal. We vary α while keeping all other training settings fixed on Qwen3-4B mathematical reasoning. As shown in Table 7, introducing the reflection-induced correction consistently improves over $\alpha = 0$, while performance remains stable across nonzero strengths. We use $\alpha = 1$ in the main experiments.

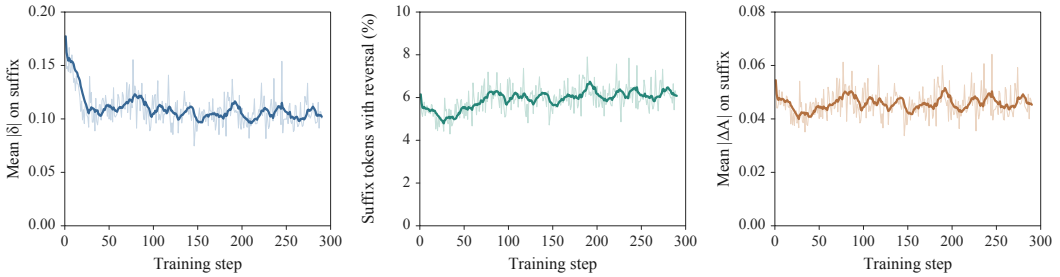
Table 7: Sensitivity to the reflection strength α on mathematical reasoning with Qwen3-4B.

α	AIME24	AIME25	HMMT-Feb	HMMT-Nov	AMC23	Avg.
0.00	62.29	51.46	29.17	37.71	93.44	54.81
0.25	62.08	52.29	29.79	38.12	93.28	55.11
0.50	63.54	52.08	29.38	38.75	93.91	55.53
1.00	64.17	53.75	30.00	39.17	94.22	**56.26**

E.2 Dynamics of Reflection-Induced Correction

We examine how reflection modifies token-level supervision during mathematical reasoning training with Qwen3-4B. Figure 5 reports three statistics over selected suffix token visits across training steps 1–290. Panel (a) measures the mean absolute reflection-induced change in teacher log-probabilities, $\mathbb{E}_{t\in\mathcal{S}}[|\delta_t|]$, where $\mathcal{S}$ denotes the selected suffix token visits. Panel (b) reports the percentage of these visits for which the distillation signal changes sign, i.e., $D_t(D_t + \alpha\delta_t) < 0$, where D_t is the log-probability gap between the base teacher and the rollout policy. This statistic captures sign reversals in the distillation component, rather than in the full RL-fused advantage. Panel (c) reports the normalized magnitude of the reflection-induced advantage correction, comparing advantages that share the same RL and base distillation components and differ only in the reflection correction. Light curves show per-step values, and dark curves show trailing 10-step means.

The mean reassessment magnitude decreases early in training and then fluctuates around a relatively stable level. Reflection nevertheless continues to change the sign of the distillation component on selected suffix tokens, with a mean per-step reversal rate of approximately 5.92% over the displayed interval. The normalized advantage correction also remains approximately stable, averaging 0.0462 over steps 51–150 and 0.0461 over steps 151–290. These results indicate that reflection continues to modify distillation directions and advantage magnitudes beyond the initial training stage. Since the selected suffix population varies during training, these statistics characterize aggregate supervision dynamics within this run; they do not directly measure localization accuracy, gradient contributions, or the correctness of individual corrections.



(a) Reflection reassessment (b) Guidance direction reversal (c) Normalized advantage correction

Figure 5: Dynamics of reflection-conditioned supervision during mathematical reasoning training. (a) Mean absolute reflection-induced change in teacher token log-probabilities. (b) Percentage of token visits exhibiting a sign reversal in the distillation component. (c) Mean normalized magnitude of the reflection-induced advantage correction.

F Qualitative Examples

We present two qualitative examples illustrating the effect of retrospective diagnosis on token-level distillation signals, with student excerpts and the corresponding teacher feedback. Following the notation in the main text, the OPD and RC-OPD rows compare the base distillation advantage $A^{\mathrm{OPD}}_{i,t}$ with its reflection-corrected counterpart $\widetilde{A}^{\mathrm{OPD}}_{i,t}$ on the same original student tokens. For the dis-

played tokens within the selected suffix, $\alpha = M_{i,t} = 1$, giving

$$A_{i,t}^{\mathrm{OPD}} = \log \frac{p_T(y_{i,t} \mid h_{i,t})}{\pi_{\mathrm{old}}(y_{i,t} \mid h_{i,t})},$$
$$\widetilde{A}_{i,t}^{\mathrm{OPD}} = A_{i,t}^{\mathrm{OPD}} + \delta_{i,t} = \log \frac{p_T(y_{i,t} \mid h_{i,t}, z_i)}{\pi_{\mathrm{old}}(y_{i,t} \mid h_{i,t})}.$$

The figures display these raw advantages before the bounded transformation and β scaling, excluding the outcome-based GRPO advantage.

Qualitative Examples

Case 1: An invalid inference from a function limit.

Problem. Find the residue of the function $f(z) = \frac{1}{z - \sin z}$ at the pole $z = 0$. Please reason step by step, and put your final answer within `\boxed{}`.

Student response (excerpts). So near $z = 0$, we have:

$$z - \sin z \sim \frac{z^3}{6}.$$

[Intervening text omitted.]
So the function behaves like $\frac{6}{z^3}$ near $z = 0$, which means that $z = 0$ is a **pole of order 3**.
[Intervening steps omitted.]
But instead of working with the full function, since we already know:

$$g(z) = \frac{z^3}{z - \sin z} \to 6 \quad \text{as } z \to 0$$

So $g(z) \to 6$, meaning all derivatives of $g(z)$ at $z = 0$ will be 0, except the zeroth derivative, which is 6.
[Intervening steps omitted.]
So:

$$\mathrm{Res}_{z=0} f(z) = \frac{1}{2!} \cdot 0 = \boxed{0}.$$

Teacher diagnosis. The assertion that all derivatives of $g(z) = \frac{z^3}{z - \sin z}$ at $z = 0$ are zero except the zeroth derivative is incorrect. Although $g(z) \to 6$ as $z \to 0$, this does not imply that higher-order derivatives vanish at 0. The function $g(z)$ is analytic at $z = 0$ (since the singularity is removable), and its Taylor expansion around 0 includes non-zero higher-order terms. Thus, $g^{(2)}(0)$ is not necessarily zero.

Correction guidance. Compute the second derivative of $g(z) = \frac{z^3}{z - \sin z}$ using its Taylor expansion around $z = 0$, or use series expansion of $z - \sin z$ to find the Laurent series of $f(z)$, and extract the coefficient of z^{-1}.

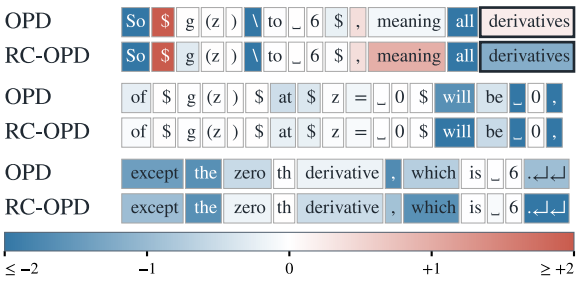


Figure 6: Base and reflection-corrected distillation advantages on the same 43-token student excerpt from Box 1.

The teacher identifies the invalid inference from the limit of $g(z)$ to its higher-order derivatives and recommends retaining the relevant Taylor terms. In Figure 6, the raw distillation advantage for `derivatives` changes from $+0.225$ to -1.201. This sign reversal shows that retrospective

diagnosis can reverse a locally positive distillation contribution within a faulty inference, providing corrective guidance beyond the teacher's standard prefix-conditioned predictions.

Qualitative Examples

Case 2: A missing condition for the divergence theorem.

Problem.
Calculate the flux of the vector field $\nabla \times A$ through the surface $S = \{z = 2(1 - x^2 - y^2),\ z \geq 0\}$, where $A = (y, z, x)$. Please reason step by step, and put your final answer within `\boxed{}`.

Student response (excerpts).
This surface is bounded above by $z = 2$, since:

$$z = 2(1 - x^2 - y^2) \leq 2 \quad (\text{when } x^2 + y^2 \geq 0)$$

So the surface is a closed surface (since it is bounded) and we can use the **Divergence Theorem** to compute the flux of $\nabla \times A$ over the surface S.
[Intervening steps omitted.]
But for any vector field A, we know that:

$$\nabla \cdot (\nabla \times A) = 0.$$

This is a **vector calculus identity**—the divergence of a curl is always zero.
So:

$$\iiint_V \nabla \cdot (\nabla \times A)\, dV = \iiint_V 0\, dV = 0.$$

[Intervening heading omitted.]
Therefore, the flux of $\nabla \times A$ through the surface S is: $\boxed{0}$.

Teacher diagnosis.
The surface S is not closed; it is the graph of $z = 2(1 - x^2 - y^2)$ for $z \geq 0$, which is a paraboloid cap bounded below by the disk $x^2 + y^2 \leq 1$ in the plane $z = 0$. It does not include the bottom disk, so it is an open surface. The Divergence Theorem applies only to closed surfaces enclosing a volume, which is not the case here.

Correction guidance.
The Divergence Theorem cannot be applied because the surface S is not closed. Instead, the flux of $\nabla \times A$ through S must be computed directly using Stokes' Theorem, by evaluating the line integral of A over the boundary curve of S, which is the circle $x^2 + y^2 = 1$ in the plane $z = 0$.

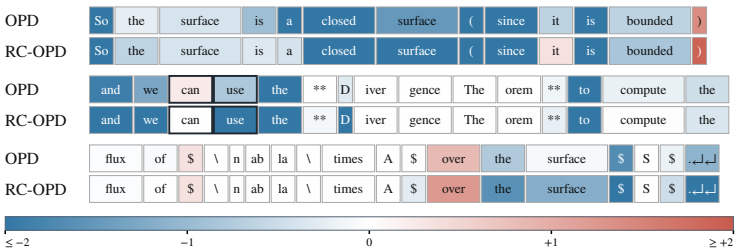


Figure 7: Base and reflection-corrected distillation advantages on the same 46-token student excerpt from Box 2.

The teacher identifies the missing closed-surface condition for applying the divergence theorem. In Figure 7, the raw distillation advantage for `use` decreases from -0.806 to -3.988. Here, retrospective diagnosis strengthens an already negative distillation contribution within the unsupported theorem-application claim, illustrating that reflection can refine the strength as well as the direction of teacher guidance. The original problem leaves surface orientation unspecified, but the diagnosed violation of the closed-surface condition is independent of that choice.