DOR^2: DUAL ON-POLICY SELF-DISTILLATION FOR REFLECTION AND REASONING

Anonymous authors
Paper under double-blind review

ABSTRACT

On-policy self-distillation (OPSD) improves language model reasoning by transferring privileged information from a self-teacher to an unprivileged student. Recent reflection-guided approaches leverage self-generated reflections as privileged information, allowing models to learn from their own reasoning errors. However, these methods primarily optimize the reasoning policy to internalize reflections, while leaving the generation of useful reflections largely unoptimized. We propose **DOR**^{2}, a dual on-policy self-distillation framework that jointly improves reflection generation and reasoning. Our key insight is that self-generated reflections are themselves on-policy sequences and can therefore be directly improved through privileged self-distillation. To assess reflection quality, we compare how reflection conditioning changes the likelihood of the model's own generated response and use its verified correctness to determine whether this change is desirable: a useful reflection should preserve or increase support for a correct response, while reducing support for an incorrect one. The resulting qualitative feedback serves as privileged information for an evaluation-conditioned reflection teacher, whose token-level distribution is distilled into the unprivileged reflector alongside conventional reflection-guided reasoning distillation. Across a suite of reasoning benchmarks, DOR^2 consistently outperforms standard OPSD and reflection-guided OPSD baselines across multiple model scales, demonstrating the benefit of explicitly optimizing both reflection generation and reasoning.

1 INTRODUCTION

On-policy distillation (OPD) (Agarwal et al., 2024; Gu et al., 2024; Ko et al., 2024; Lu & Thinking Machines Lab, 2025) has emerged as a promising approach to post-training large language models (LLMs), particularly for reasoning tasks. Unlike conventional distillation on fixed teacher-generated responses, OPD provides teacher supervision on trajectories sampled from the student, reducing the mismatch between training and inference. On-policy self-distillation (OPSD) (Zhao et al., 2026a; Shenfeld et al., 2026) further removes the need for a separate teacher model: the same LLM serves as both student and teacher, while the teacher receives privileged information unavailable to the student. Distilling the resulting privileged teacher distributions provides dense token-level supervision beyond sparse outcome rewards, and recent work has explored privileged information ranging from verified solutions to rich task feedback (Hübotter et al., 2026).

Self-reflection provides a particularly appealing form of privileged information because it can offer trajectory-specific guidance derived from the model's own reasoning behavior. Building on earlier work on self-generated rationales and feedback (Zelikman et al., 2022; Madaan et al., 2023; Shinn et al., 2023), ROSD (Zhao et al., 2026b) uses corrective reflections to guide on-policy self-distillation, while SRPO (Liu et al., 2026a) converts trajectory feedback into reflection-based policy supervision. However, these methods primarily optimize how the reasoning policy *uses* reflection, while the generation of useful reflections itself receives no dedicated supervision. This creates a critical asymmetry: reflection quality directly affects the privileged teacher, yet the reflector may produce inaccurate or uninformative guidance, and a large reflection-induced teacher-student shift does not necessarily indicate a beneficial correction (Kaur et al., 2026).

Our key insight is that a self-generated reflection is itself an on-policy sequence and can therefore be treated as an additional self-distillation target. This enables a second teacher-student pair over

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.

reflection generation: the unprivileged reflector generates the reflection, while a privileged reflection teacher additionally observes a post-hoc evaluation of its quality. The remaining challenge is to determine whether the reflection induces a *task-aligned* change rather than merely a large distributional change. We formalize this principle through *rationalization gain* and instantiate it with a correctness-aware likelihood comparison. We score the student's generated response before and after reflection conditioning, and use its verified correctness to determine the desirable direction of change. A useful reflection should preserve or increase support for a correct response, while reducing support for an incorrect one. This evaluation requires no additional reflection-conditioned response rollout and is converted into qualitative feedback for reflection self-distillation.

Building on this insight, we propose **DOR**^{2}, *Dual On-Policy Self-Distillation for Reflection and Reasoning*. Our framework jointly optimizes two complementary distillation processes within a single shared model. *Reasoning distillation* internalizes reflection-conditioned guidance into the reasoning policy, while *reflection distillation* uses post-hoc quality feedback to condition a privileged reflection teacher whose token-level distribution supervises the original reflector. Both processes operate on the model's own generated trajectories and update the shared parameters, providing dedicated supervision for both learning from reflection and generating useful reflections.

We evaluate DOR^2 on four scientific question-answering tasks and one tool-use task across Qwen3-1.7B, Qwen3-4B, and Qwen3-8B. DOR^2 achieves the highest average mean@8 at all three model scales. Ablations further show that correctness-aware reflection evaluation outperforms alternative feedback constructions, while random or constant quality labels substantially reduce performance, highlighting the importance of explicitly optimizing reflection generation with informative, sample-specific feedback.

Our contributions are summarized as follows:

- We identify reflection generation as an under-optimized component of reflection-guided OPSD and introduce DOR^2, a dual on-policy self-distillation framework that explicitly trains both reflection generation and reasoning within a single shared model.
- We develop a task-aligned reflection-evaluation principle based on rationalization gain and instantiate it as a correctness-aware likelihood comparison on the model's own generated response, enabling qualitative reflection feedback without an additional response rollout.
- We conduct extensive experiments across five tasks and three model scales, where DOR^2 consistently outperforms strong baselines; ablations further validate the importance of informative reflection-quality feedback and the dual-distillation design.

2 Related Work

2.1 On-Policy Distillation and Privileged Information

On-policy learning reduces the mismatch between training trajectories and those induced by the learned policy, a principle established in interactive imitation learning (Ross et al., 2011). For autoregressive language models, GKD (Agarwal et al., 2024) provides teacher supervision on student-generated sequences and supports multiple distribution-matching objectives. MiniLLM (Gu et al., 2024) studies reverse-KL-based distillation, while DistiLLM (Ko et al., 2024) develops alternative divergence objectives and more efficient use of student-generated data. Lu & Thinking Machines Lab (2025) further highlights OPD as a language-model post-training paradigm that combines on-policy trajectories with dense teacher supervision. OPSD removes the need for a separate teacher by assigning different information to different roles of the same model. The Self-Distilled Reasoner (Zhao et al., 2026a), SDFT (Shenfeld et al., 2026), and SDPO (Hübotter et al., 2026) instantiate this idea using verified solutions, demonstrations, and rich feedback, respectively. Recent work further studies the reliability and stability of privileged self-teaching, including reward-anchored self-distillation (Yang et al., 2026a), uncertainty-aware token weighting (Ke et al., 2026), position-dependent teacher reliability (Liu et al., 2026b), and stabilized feedback-augmented self-distillation (Yang et al., 2026b); privileged conditioning can also degrade long-form reasoning in some settings (Kaur et al., 2026). Our work instead focuses on improving the quality of self-generated privileged information by explicitly training reflection generation.

2.2 Self-Reflection and Reflection-Guided Learning

Self-reflection enables language models to evaluate and revise their own outputs. Early approaches use generated rationales, feedback, or verbal reflections to bootstrap reasoning and iterative refinement (Zelikman et al., 2022; Madaan et al., 2023; Shinn et al., 2023), although self-correction without reliable feedback can be ineffective (Huang et al., 2024). Training-based approaches such as RISE (Qu et al., 2024), SCoRe (Kumar et al., 2025), and effective-reflection training (Wang et al., 2026) explicitly improve reflective or corrective behavior, while complementary work explores process-level supervision and self-generated evaluation signals (Lightman et al., 2024; Wang et al., 2024; Yuan et al., 2024). More closely related to our work, ROSD (Zhao et al., 2026b) constructs error-focused reflections to guide on-policy self-distillation, while SRPO (Liu et al., 2026a) uses reflection patches to provide dense policy-training signals. These methods primarily leverage reflection to improve the reasoning policy; although shared-parameter updates may also affect reflection generation, they do not explicitly distill post-hoc reflection-quality feedback into the reflector. DOR^2 instead treats reflection generation itself as an OPSD problem. By evaluating a reflection through its effect on ground-truth answer likelihood, our method constructs a privileged reflection teacher and transfers its predictions to the unprivileged reflector, jointly improving reflection generation and reasoning.

3 Preliminaries

3.1 On-Policy Self-Distillation

Let $x \sim \mathcal{D}$ denote an input problem and $y = (y_1, \ldots, y_T)$ a response generated by an autoregressive language model π_θ, with $\pi_\theta(y \mid x) = \prod_{t=1}^{T} \pi_\theta(y_t \mid h_t)$ and $h_t \triangleq (x, y_{<t})$. On-policy self-distillation (OPSD) uses the same underlying model as both student and teacher. The student predicts from the original history, $\pi_\theta^S(\cdot \mid h_t) \triangleq \pi_\theta(\cdot \mid h_t)$, while the teacher additionally observes privileged information I, $\pi_{\bar{\theta}}^T(\cdot \mid h_t, I) \triangleq \pi_{\bar{\theta}}(\cdot \mid h_t, I)$. Here, $\bar{\theta} = \mathrm{sg}(\theta)$ denotes a frozen snapshot of the current model and $\mathrm{sg}(\cdot)$ is the stop-gradient operator. The OPSD objective is [1]

$$\mathcal{L}_{\mathrm{OPSD}}(\theta) = \mathbb{E}\left[\sum_{t=1}^{T} D_{\mathrm{JSD}}\left(\mathrm{sg}\left[\pi_{\bar{\theta}}^T(\cdot \mid h_t, I)\right] \,\middle\|\, \pi_\theta^S(\cdot \mid h_t)\right)\right], \tag{1}$$

where D_{JSD} denotes the Jensen-Shannon divergence. The key property of OPSD is that the teacher provides token-level supervision on the student's own on-policy prefixes, rather than requiring a separately generated teacher response. Privileged conditioning allows the self-teacher to provide information unavailable to the student, which the student learns to internalize through distillation.

3.2 Reflection as Privileged Information

Recent reflection-guided approaches use feedback on the model's own trajectories to construct trajectory-specific privileged information (Zhao et al., 2026b; Liu et al., 2026a). For example, ROSD (Zhao et al., 2026b) generates corrective reflection from an on-policy reasoning trajectory and uses it to condition the self-teacher, while error localization can further restrict distillation to the affected portion of the response. Compared with generic reference-solution conditioning, reflection provides guidance that is more closely tied to the model's own reasoning behavior.

Let c^R denote the reflection-generation context, which may contain the problem, the on-policy response, and available training-time feedback. The same model acts as the reflector and samples $r \sim \rho_{\bar{\theta}}^S(\cdot \mid c^R)$, where $\rho_{\bar{\theta}}^S$ denotes its reflection-generation role. The reflection and reasoning roles share parameters but differ in their prompting and conditioning contexts. The generated reflection is then supplied as privileged information to the reasoning teacher, $\pi_{\bar{\theta}}^T(\cdot \mid h_t, r)$, whose token-level distribution supervises the unprivileged reasoning student through OPSD.

[1] Alternative self-teacher updates include EMA teachers (Zhao et al., 2026a; Yang et al., 2026b) and periodic snapshots (Yang et al., 2026a), while alternative divergence choices includes forward and reverse KL (Agarwal et al., 2024)

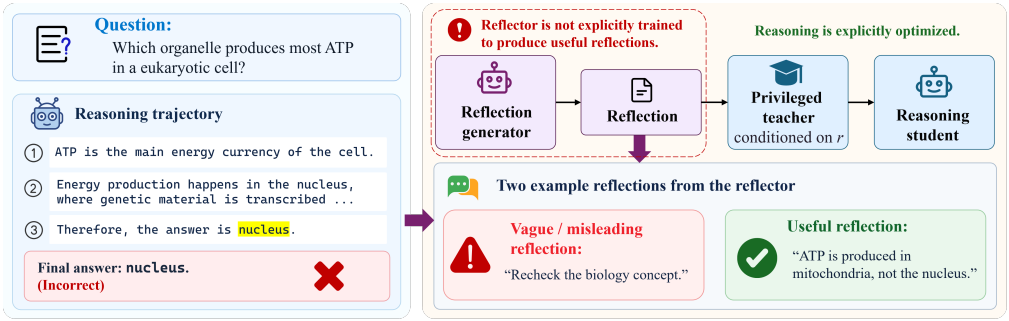

Figure 1: Reflection-guided OPSD optimizes reasoning from reflection, but not the generation of useful reflections that shape the privileged teacher.

Motivation. Reflection-guided OPSD explicitly trains the reasoning policy to learn from reflection, but provides no dedicated objective for learning to generate useful reflections. As illustrated in Figure 1, the reflection directly shapes the privileged teacher and thus the resulting reasoning supervision, yet the reflector itself is not explicitly optimized for reflection quality. Our key observation is that reflection is itself an on-policy sequence and can therefore be treated as an additional self-distillation target. This motivates a second teacher-student pair over reflection generation, where the privileged reflection teacher observes a post-hoc quality evaluation unavailable to the original reflector. The remaining challenge is to determine whether reflection conditioning changes the model in a task-aligned direction. DOR^2 addresses this with a correctness-aware likelihood comparison on the student's generated response, which provides qualitative reflection-quality feedback without requiring an additional response rollout.

4 Method

DOR^2 jointly optimizes two conditional behaviors of a shared language model: reasoning from the original problem and reflecting on its own reasoning trajectories. The central challenge is to determine whether a generated reflection provides useful guidance and transfer this quality information back into the reflector. As illustrated in Figure 2, DOR^2 augments conventional reflection-guided reasoning distillation with a second self-distillation process dedicated to reflection generation. We first formulate a task-aligned notion of reflection utility, instantiate it as a practical correctness-aware evaluation criterion, and then introduce the dual self-distillation objectives.

4.1 Task-Aligned Reflection Utility

4.1.1 Rationalization Gain

Privileged information is useful only when the distributional change it induces is aligned with the task objective. A large teacher-student discrepancy is therefore insufficient: privileged conditioning may substantially alter the model distribution while shifting probability toward undesirable responses. We formalize this distinction through *rationalization gain*.

Let $p^\star(\cdot \mid x)$ denote a desired distribution over response sequences y for problem x. For privileged information I, we define its rationalization gain as the reduction in divergence from this desired distribution:

$$\mathcal{G}^\star(I;x) \triangleq D_{\mathrm{KL}}\big(p^\star(\cdot \mid x) \,\big\|\, \pi_\theta^S(\cdot \mid x)\big) - D_{\mathrm{KL}}\big(p^\star(\cdot \mid x) \,\big\|\, \pi_\theta^T(\cdot \mid x, I)\big)\,. \tag{2}$$

We use forward KL here as an analytical measure of task-aligned utility, independently of the divergence used for token-level distillation; its log-ratio decomposition will enable the practical reflection evaluation developed below. Thus, $\mathcal{G}^\star(I;x) > 0$ means that privileged conditioning moves the model closer to the desired response distribution, $\mathcal{G}^\star(I;x) = 0$ leaves the divergence unchanged, and $\mathcal{G}^\star(I;x) < 0$ moves it farther away. Expanding the two KL divergences gives

$$\mathcal{G}^\star(I;x) = \mathbb{E}_{y\sim p^\star(\cdot|x)}\left[\log \frac{\pi_\theta^T(y \mid x, I)}{\pi_\theta^S(y \mid x)}\right]. \tag{3}$$

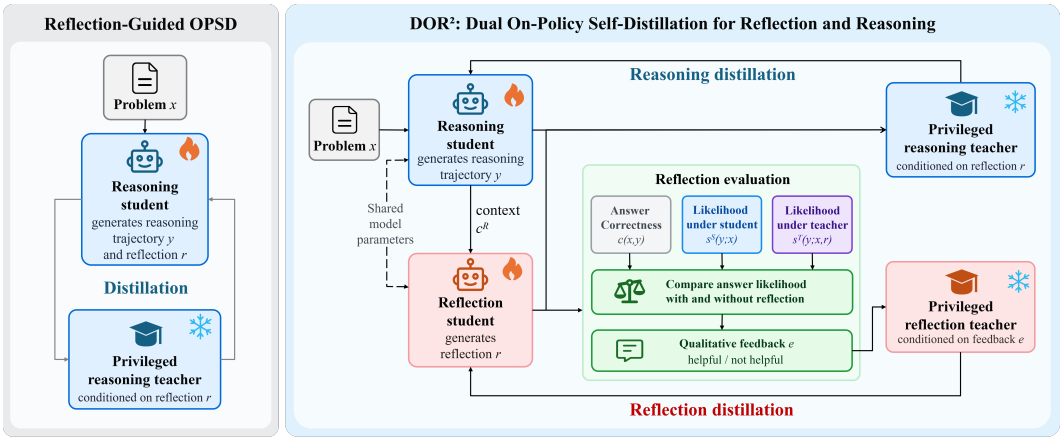

Figure 2: **Reflection-guided OPSD versus DOR^2**. Existing methods distill reflection-conditioned guidance only into the reasoning policy, whereas DOR^2 additionally distills post-hoc reflection-quality feedback into the reflector.

Equation 3 shows that useful privileged information should increase relative probability on responses favored by the desired distribution. What matters is therefore the *direction* of the induced change with respect to desirable outputs, rather than the magnitude of the teacher-student discrepancy alone. The full derivation is given in Appendix A.3, with a complementary Bayesian interpretation in Appendix A.1.

4.1.2 Correctness-Aware Reflection Evaluation

Equation 2 provides an ideal criterion, but the desired distribution $p^\star$ is generally unavailable. Moreover, many reasoning benchmarks provide only compact reference answers, such as option labels, which offer little semantic information for evaluating how a reflection changes the model's reasoning. We therefore evaluate each reflection through its effect on the student's own on-policy response, using verified correctness to determine the desirable direction of change.

Let $y = (y_1, \ldots, y_T)$ denote the sampled on-policy response, r its corresponding reflection, and $c(x, y) \in \{0, 1\}$ its verified correctness. We score the same response under the unprivileged student and the reflection-conditioned teacher:

$$s^S(y; x) = \sum_{t=1}^{T} \log \pi_\theta^S(y_t \mid x, y_{<t}), \qquad s^T(y; x, r) = \sum_{t=1}^{T} \log \pi_\theta^T(y_t \mid x, r, y_{<t}). \tag{4}$$

Both likelihoods are computed using teacher forcing (Bengio et al., 2015; Lamb et al., 2016): each model scores the fixed response y token by token while conditioning on the observed prefix $y_{<t}$. We define the reflection-induced likelihood shift as

$$\Delta_y(r; x) = s^T(y; x, r) - s^S(y; x). \tag{5}$$

Because both terms score the same sequence, length normalization does not affect the sign of Δ_y and is unnecessary for the binary criterion below.

The desirable direction depends on the correctness of the original response. A useful reflection should preserve or increase support for a correct response, but reduce support for an incorrect one. We therefore define

$$b(r; x, y) = \begin{cases} \mathbb{I}[\Delta_y(r; x) \geq 0], & c(x, y) = 1, \\ \mathbb{I}[\Delta_y(r; x) < 0], & c(x, y) = 0. \end{cases} \tag{6}$$

Here, $b(r; x, y) = 1$ indicates that reflection conditioning changes support for the original response in the direction implied by its verified correctness. The treatment of $\Delta_y = 0$ is intentionally asymmetric: unchanged support is acceptable for a correct response but provides no corrective effect for

an incorrect one. For a verified-correct response, $\Delta_y(r;x)$ equals the rationalization gain obtained by taking the point-mass target $p^\star = \delta_y$. For a verified-incorrect response, $-\Delta_y(r;x)$ is instead a verifier-guided directional proxy: suppressing a known incorrect response is desirable, although it does not guarantee that the displaced probability mass moves to a correct response. This criterion therefore operationalizes the task-aligned principle of Eq. 2 using only two likelihood evaluations of the same on-policy response, without an additional reflection-conditioned rollout. [2]

4.2 Dual On-Policy Self-Distillation

4.2.1 Evaluation-Conditioned Reflection Self-Distillation

DOR^2 uses the reflection-quality label in Eq. 6 as privileged information for reflection generation. Let c^R denote the reflection-generation context and $r = (r_1, \ldots, r_M) \sim \rho_{\bar{\theta}}^S(\cdot \mid c^R)$ an on-policy reflection, where M is the number of reflection tokens. The reflection generator and reasoning policy are conditional roles of the same underlying model and share all parameters.

After evaluating r, we verbalize its binary quality label as $e \triangleq \textsc{Verbalize}(b(r;x,y))$, where e is a textual indication of whether the reflection induces a desirable change in support for the original response. It contains neither the numerical likelihoods nor their difference and becomes available only after the completed reflection has been evaluated.

At reflection position j, the unprivileged reflection student conditions on the original reflection context and prefix, whereas the privileged reflection teacher additionally observes e:

$$\rho_\theta^S(\cdot \mid c^R, r_{<j}) \triangleq \pi_\theta(\cdot \mid c^R, r_{<j}), \qquad \rho_{\bar{\theta}}^T(\cdot \mid c^R, e, r_{<j}) \triangleq \pi_{\bar{\theta}}(\cdot \mid c^R, e, r_{<j}). \tag{7}$$

We distill the privileged teacher distribution into the unprivileged reflector along the original on-policy reflection trajectory:

$$\mathcal{L}_{\text{refl}}(\theta) = \mathbb{E}\left[\frac{1}{M}\sum_{j=1}^{M} D_{\text{JSD}}\left(\text{sg}\big[\rho_{\bar{\theta}}^T(\cdot \mid c^R, e, r_{<j})\big] \,\big\|\, \rho_\theta^S(\cdot \mid c^R, r_{<j})\right)\right]. \tag{8}$$

Student and teacher are evaluated on the same on-policy reflection prefixes, while the teacher additionally receives the post-hoc quality feedback. This provides dense token-level supervision for reflection generation without requiring a separately generated "improved reflection". Rather than treating $b(r;x,y)$ as a sequence-level reward, DOR^2 uses it to construct a privileged teacher distribution that can be internalized through OPSD.

4.2.2 Reflection-Guided Reasoning Distillation

In parallel, DOR^2 performs reflection-guided reasoning distillation on the original on-policy response y. Let $h_t = (x, y_{<t})$ denote the reasoning history and $\mathcal{S}(y)$ the set of positions selected for distillation, which may cover the full trajectory or, when error localization is available, only the error-affected region. For nonempty $\mathcal{S}(y)$, we optimize

$$\mathcal{L}_{\text{pol}}(\theta) = \mathbb{E}\left[\frac{1}{|\mathcal{S}(y)|}\sum_{t\in\mathcal{S}(y)} D_{\text{JSD}}\left(\text{sg}\big[\pi_{\bar{\theta}}^T(\cdot \mid h_t, r)\big] \,\big\|\, \pi_\theta^S(\cdot \mid h_t)\right)\right]. \tag{9}$$

A trajectory with no selected positions contributes no reasoning-distillation term. The two objectives serve complementary roles: $\mathcal{L}_{\text{pol}}$ internalizes reflection-conditioned guidance into the reasoning policy, whereas $\mathcal{L}_{\text{refl}}$ internalizes post-hoc reflection-quality feedback into the reflector.

4.2.3 Unified Dual-Distillation Objective

DOR^2 jointly optimizes reasoning and reflection generation through

$$\mathcal{L}_{\text{DOR}^2}(\theta) = (1-\lambda)\mathcal{L}_{\text{pol}}(\theta) + \lambda\mathcal{L}_{\text{refl}}(\theta), \qquad 0 \le \lambda \le 1. \tag{10}$$

[2] When semantically informative verified reference answers are available, rationalization gain also admits an exact reference-answer likelihood formulation; see Appendix B.

The coefficient λ controls the contribution of reflection self-distillation: $\lambda = 0$ recovers reasoning distillation alone, while $\lambda = 1$ retains only reflection distillation. Intermediate values balance the two token-normalized objectives. Both losses update the shared parameters θ, while all privileged teacher distributions are treated as stop-gradient targets parameterized by $\bar{\theta}$ during optimization. The complete training procedure is summarized in Algorithm 1 in Appendix C. The correctness-based reflection verdict and its verbalized feedback are used only during training and are not required at inference time.

5 EXPERIMENTS

5.1 EXPERIMENTAL SETUP

Datasets. We train and evaluate DOR^2 on four scientific question-answering tasks from SciKnowEval (Feng et al., 2024): Biology, Chemistry, Physics, and Material, as well as the tool-use benchmark ToolAlpaca (Tang et al., 2023). The SciKnowEval tasks cover diverse scientific reasoning domains, while ToolAlpaca extends the evaluation beyond question answering to tool-use tasks.

Baselines. We compare DOR^2 with the base model, GRPO, SDPO, and ROSD. **Base model** denotes the backbone model before post-training. **GRPO** (Khatri et al., 2026) represents outcome-level reinforcement learning with verifiable rewards. **SDPO** (Hübotter et al., 2026) represents standard on-policy self-distillation with privileged feedback, while **ROSD** (Zhao et al., 2026b) is the closest reflection-guided baseline, using corrective reflections to localize errors and guide reasoning distillation. In contrast, **DOR^2** additionally provides dedicated self-distillation for reflection generation through post-hoc reflection-quality feedback.

Implementation Details. We use Qwen3-1.7B, Qwen3-4B, and Qwen3-8B as backbone models to evaluate performance across model scales. During training, we sample eight on-policy responses per input and use task-specific verifiers to determine final-answer correctness. The distributional-divergence mixing coefficient is set to 0.5 for both reasoning and reflection distillation, and the reflection-distillation weight is set to $\lambda = 0.15$. At evaluation, we independently sample eight responses per question and report mean@8 as the primary metric, together with pass@8. Mean@8 measures average correctness across the eight samples, while pass@8 measures whether at least one sampled response is correct, capturing multi-sample solution coverage. We additionally report majority@8 in ablations, which evaluates the correctness of the majority-voted answer among the eight sampled responses. Additional implementation details and experimental settings are provided in Appendix D.

5.2 MAIN RESULTS

In-domain Results. Table 1 reports the in-domain results. DOR^2 achieves the best overall performance at all three model scales, reaching 70.24%, 71.81%, and 74.69% on Qwen3-1.7B, Qwen3-4B, and Qwen3-8B, respectively. These results improve over the strongest baseline by 6.07, 1.67, and 3.09 percentage points. At the task level, DOR^2 attains the highest mean@8 on all five tasks with Qwen3-1.7B, leads on all four scientific QA tasks with Qwen3-4B while matching the best ToolAlpaca result, and remains strongest on four of five tasks with Qwen3-8B.

The comparison with ROSD is particularly informative because both methods use reflection-guided reasoning distillation, while DOR^2 additionally optimizes reflection generation through post-hoc quality feedback. DOR^2 improves mean@8 over ROSD in all model-task comparisons, indicating that reflection generation itself is a meaningful optimization target rather than merely an auxiliary source of privileged information. The improvement is largest for Qwen3-1.7B, suggesting greater headroom for improving reflection quality at smaller model scales, while the consistent gains on larger models show that the benefit persists as model capacity increases.

Out-of-domain Generalization. DOR^2 also maintains competitive cross-domain generalization. Averaged across all source and target domains, it achieves 47.59% mean@8, outperforming SDPO (46.61%) and ROSD (46.45%) while remaining comparable to GRPO (47.92%). This suggests that

Table 1: In-domain results across model scales (%). Each task reports pass@8 and mean@8. Avg reports the unweighted five-task average of each metric. **Bold** and underlined values indicate the highest and second-highest distinct values in each column within each model scale.

Method	Biology		Chemistry		Physics		Material		ToolAlpaca		Avg	
	pass@8	mean@8	pass@8	mean@8	pass@8	mean@8	pass@8	mean@8	pass@8	mean@8	pass@8	mean@8
Qwen3-1.7B												
Base Model	66.39	32.00	69.77	40.77	71.40	50.62	70.46	49.34	53.48	52.22	66.30	44.99
GRPO	62.13	46.58	72.64	70.89	69.03	56.88	80.69	70.61	55.64	54.85	68.03	59.96
SDPO	67.10	49.25	78.98	74.76	60.80	56.88	66.21	61.30	66.49	58.27	67.92	60.09
ROSD	54.00	54.00	80.34	68.05	76.07	67.66	73.90	69.02	66.87	62.13	70.24	64.17
DOR^2 (Ours)	**67.75**	**60.00**	**84.82**	**81.73**	**77.62**	**69.69**	**82.59**	**74.87**	**68.45**	**64.89**	**76.25**	**70.24**
Qwen3-4B												
Base Model	58.04	31.75	69.27	43.21	74.10	61.41	72.67	60.77	60.29	58.46	66.87	51.12
GRPO	61.72	56.00	86.47	76.01	83.44	71.25	72.32	71.14	65.64	60.29	73.92	66.94
SDPO	58.00	57.25	76.87	74.05	76.77	70.16	75.29	67.42	68.42	**63.42**	71.07	66.46
ROSD	**78.45**	58.00	85.12	78.99	87.05	75.63	**80.20**	75.40	**70.10**	62.68	**80.18**	70.14
DOR^2 (Ours)	70.68	**58.50**	**86.64**	**81.43**	**87.67**	**77.50**	80.11	**78.19**	67.47	**63.42**	78.51	**71.81**
Qwen3-8B												
Base Model	62.36	30.00	72.15	42.74	81.15	60.47	68.43	59.71	64.75	58.09	69.77	50.20
GRPO	**77.07**	60.00	83.16	78.04	**94.45**	74.69	80.23	74.34	70.59	**68.57**	**81.10**	71.13
SDPO	74.46	55.00	84.86	80.30	85.49	77.80	80.44	77.39	70.11	66.18	79.07	71.33
ROSD	68.61	61.50	89.10	80.00	83.18	74.84	81.66	76.20	70.97	65.44	78.70	71.60
DOR^2 (Ours)	69.28	**62.25**	**89.57**	**83.45**	89.11	**78.13**	**84.59**	**81.78**	**72.01**	67.83	80.91	**74.69**

the substantial in-domain gains of DOR^2 do not come at the expense of cross-domain transfer. Full cross-domain results are provided in Appendix D.3.1.

5.3 Reflection Quality Analysis

Since DOR^2 explicitly optimizes reflection generation, we further examine how reflection quality evolves during training. For both ROSD and DOR^2, we apply the same correctness-aware evaluation criterion and define the *helpful rate* at each training step as the proportion of generated reflections classified as helpful.

Figure 3 shows that both methods rapidly learn to produce helpful reflections, but DOR^2 consistently maintains a higher helpful rate during the later stages of training on both Biology and Material. The gap is particularly clear on Biology, where DOR^2 remains close to saturation while ROSD exhibits larger degradation. Because both methods are evaluated using the same criterion, these results support the effectiveness of dedicated reflection self-distillation in improving the reflector toward more task-aligned and consistently useful reflections.

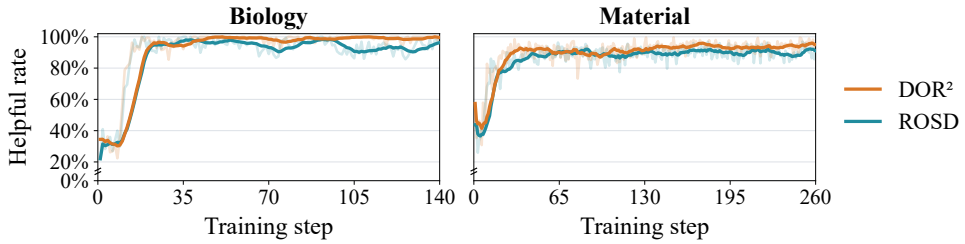

Figure 3: Helpful reflection rate during training on Biology (left) and Material (right).

5.4 Ablation Study

We conduct ablations on Qwen3-1.7B Biology to examine the construction of reflection-quality feedback, and the information carried by the quality labels, while keeping all other settings fixed. Additional ablations on the balance between reasoning and reflection distillation are provided in Appendix D.3.3.

Reflection Evaluation and Privileged Feedback. We compare three ways of evaluating reflections and constructing privileged feedback. The full DOR^2 design combines verified correctness

Table 2: Ablation on reflection evaluation and privileged feedback (%). **Bold** and underlined values indicate the best and second-best distinct values in each column.

Setting	pass@8	mean@8	majority@8
Likelihood-based evaluation (full DOR^2)	**67.75**	**60.00**	**59.71**
Single-rewrite verification	58.00	57.50	57.84
Numerical log-probability feedback	53.32	52.25	52.01

Table 3: Ablation on reflection-quality labels (%). **Bold** and underlined values indicate the best and second-best distinct values in each column.

Setting	pass@8	mean@8	majority@8
Quality-based labels (full DOR^2)	**67.75**	**60.00**	**59.71**
50% random labels	60.00	54.75	54.00
Always helpful	55.30	53.25	53.50
Always unhelpful	54.00	53.00	53.51

with the reflection-induced likelihood shift $\Delta_y(r; x)$ to obtain a binary quality judgment. *Single-rewrite verification* instead evaluates whether one reflection-conditioned rewritten rollout corrects the original response, while *numerical feedback* directly supplies the likelihood shift to the privileged reflection teacher.

Table 2 shows that the full correctness-aware evaluation performs best, achieving 60.00 mean@8 compared with 57.50 for single-rewrite verification and 52.25 for numerical feedback. The weaker performance of single-rewrite verification is consistent with its sampling dependence: a useful reflection may improve the response distribution without changing one sampled outcome. More notably, directly exposing the numerical likelihood shift is substantially less effective, indicating that a finer-grained scalar signal is not necessarily more useful to the privileged teacher. Converting the likelihood change into an explicit, task-aligned qualitative judgment provides a clearer supervision signal for reflection self-distillation.

Role of Reflection Quality Labels. We next isolate the information carried by the reflection-quality labels while keeping the training objective, sample selection, and prompt structure unchanged. We replace the original labels with uniformly random labels, always-helpful labels, or always-unhelpful labels.

As shown in Table 3, replacing the quality-aware labels consistently degrades performance. Random labels reduce mean@8 from 60.00 to 54.75, while always-helpful and always-unhelpful labels further reduce it to 53.25 and 53.00. The drop under random labels shows that the benefit does not arise merely from adding an auxiliary reflection-distillation objective; the privileged feedback must be aligned with the quality of each reflection. The similar performance of the two constant-label variants further shows that generic positive or negative feedback is insufficient. Effective reflection self-distillation therefore depends on *sample-specific quality information* that distinguishes useful reflections from unhelpful ones.

6 Conclusion

We introduced DOR^2 to address a key asymmetry in reflection-guided OPSD: existing methods train models to learn from reflections, but do not explicitly train them to generate useful reflections. Our central insight is that self-generated reflections are themselves on-policy sequences and can therefore be optimized through privileged self-distillation. Building on this observation, DOR^2 couples reflection self-distillation with reflection-guided reasoning distillation in a shared model, using correctness-aware post-hoc feedback to supervise the reflector. Experiments across reasoning and tool-use tasks consistently show that explicitly improving reflection generation, in addition to learning from reflection, leads to stronger reasoning performance.

AI USE STATEMENT

In this work, we used generative AI tools to assist with selected aspects of code development and with drafting, revising, and polishing the manuscript. These tools were used only as supporting aids and did not autonomously design, conduct, or evaluate the research. We did not use generative AI to generate human-subject data or replace expert judgment; human-subject data collection is not applicable to this work. All AI-assisted code was reviewed and tested by the authors before use, and all AI-assisted text was revised by the authors, who independently verified the scientific claims and citations. The authors take full responsibility for the final content of the paper, including all text, claims, code, and other artifacts produced with the assistance of generative AI.

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*, 2024.

Samy Bengio, Oriol Vinyals, Navdeep Jaitly, and Noam Shazeer. Scheduled sampling for sequence prediction with recurrent neural networks. In *Advances in Neural Information Processing Systems*, volume 28, 2015.

Kehua Feng, Keyan Ding, Weijie Wang, Xiang Zhuang, Zeyuan Wang, Ming Qin, Yu Zhao, Jianhua Yao, Qiang Zhang, and Huajun Chen. SciKnowEval: Evaluating multi-level scientific knowledge of large language models. *arXiv Preprint*, 2024. URL https://arxiv.org/abs/2406.09098.

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

Jie Huang, Xinyun Chen, Swaroop Mishra, Huaixiu Steven Zheng, Adams Wei Yu, Xinying Song, and Denny Zhou. Large language models cannot self-correct reasoning yet. In *International Conference on Learning Representations*, 2024.

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. In *International Conference on Machine Learning*, 2026.

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

Junlong Ke, Zichen Wen, Weijia Li, Conghui He, and Linfeng Zhang. Respecting self-uncertainty in on-policy self-distillation for efficient llm reasoning. *arXiv preprint arXiv:2605.13255*, 2026.

Devvrit Khatri, Lovish Madaan, Rishabh Tiwari, Rachit Bansal, Venkata Sai Surya Subramanyam Duvvuri, Manzil Zaheer, Inderjit Dhillon, David Brandfonbrener, and Rishabh Agarwal. The art of scaling reinforcement learning compute for LLMs. In *International Conference on Learning Representations*, 2026.

Jongwoo Ko, Sungnyun Kim, Tianyi Chen, and Se-Young Yun. Distillm: Towards streamlined distillation for large language models. In *Proceedings of the 41st International Conference on Machine Learning*, pp. 24872–24895, 2024.

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.

Alex Lamb, Anirudh Goyal, Ying Zhang, Saizheng Zhang, Aaron Courville, and Yoshua Bengio. Professor forcing: A new algorithm for training recurrent networks. In *Advances in Neural Information Processing Systems*, volume 29, 2016.

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.

Jialong Liu, Yuling Shi, Ning Yang, Xiaodong Gu, and Zuchao Li. Srpo: Self-reflective policy optimization for long-horizon reasoning. In *International Conference on Machine Learning*, 2026a.

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

Kevin Lu and Thinking Machines Lab. On-policy distillation. *Thinking Machines Lab: Connectionism*, 2025. doi: 10.64434/tml.20251026. URL https://thinkingmachines.ai/blog/on-policy-distillation/.

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.

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*, 2024.

Stephane Ross, Geoffrey Gordon, and Drew Bagnell. A reduction of imitation learning and structured prediction to no-regret online learning. In *Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics*, pp. 627–635, 2011.

Idan Shenfeld, Mehul Damani, Jonas Hübotter, and Pulkit Agrawal. Self-distillation enables continual learning. In *International Conference on Machine Learning*, 2026.

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.

Qiaoyu Tang, Ziliang Deng, Hongyu Lin, Xianpei Han, Qiao Liang, Boxi Cao, and Le Sun. ToolAlpaca: Generalized tool learning for language models with 3000 simulated cases. *arXiv Preprint*, 2023. URL https://arxiv.org/abs/2306.05301.

Hanbin Wang, Jingwei Song, Jinpeng Li, Qi Zhu, Fei Mi, Ganqu Cui, Yasheng Wang, and Lifeng Shang. Teaching large reasoning models effective reflection. *arXiv preprint arXiv:2601.12720*, 2026.

Peiyi Wang, Lei Li, Zhihong Shao, Runxin Xu, Damai Dai, Yifei Li, Deli Chen, Yu Wu, and Zhifang Sui. Math-shepherd: Verify and reinforce llms step-by-step without human annotations. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics*, pp. 9426–9439, 2024.

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.

Fan Yang, Rui Meng, and Yuxin Wen. Why does feedback-augmented self-distillation fail to improve retrieval-interleaved search agents? *arXiv preprint arXiv:2607.17558*, 2026b.

Weizhe Yuan, Richard Yuanzhe Pang, Kyunghyun Cho, Xian Li, Sainbayar Sukhbaatar, Jing Xu, and Jason Weston. Self-rewarding language models. In *International Conference on Machine Learning*, 2024.

Eric Zelikman, Yuhuai Wu, Jesse Mu, and Noah D. Goodman. Star: Bootstrapping reasoning with reasoning. In *Advances in Neural Information Processing Systems*, volume 35, 2022.

Siyan Zhao, Zhihui Xie, Mengchen Liu, Jing Huang, Guan Pang, Feiyu Chen, and Aditya Grover. Self-distilled reasoner: On-policy self-distillation for large language models. In *International Conference on Machine Learning*, 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 Additional Theoretical Analysis

A.1 A Bayesian Interpretation of Privileged Conditioning

We provide a complementary probabilistic interpretation of privileged conditioning. Let $y \sim \pi_{\bar{\theta}}^S(\cdot \mid x)$ denote an on-policy response, and let privileged information I be constructed from information associated with that response. The induced joint distribution can be written as

$$p_{\bar{\theta}}(y, I \mid x) = \pi_{\bar{\theta}}^S(y \mid x) p_{\bar{\theta}}(I \mid x, y). \tag{11}$$

Marginalizing over responses gives

$$p_{\bar{\theta}}(I \mid x) = \mathbb{E}_{y \sim \pi_{\bar{\theta}}^S(\cdot \mid x)} \left[p_{\bar{\theta}}(I \mid x, y) \right]. \tag{12}$$

For $p_{\bar{\theta}}(I \mid x) > 0$, Bayes' rule yields

$$p_{\bar{\theta}}(y \mid x, I) = \underbrace{\pi_{\bar{\theta}}^S(y \mid x)}_{\text{base distribution}} \underbrace{\frac{p_{\bar{\theta}}(I \mid x, y)}{p_{\bar{\theta}}(I \mid x)}}_{\text{privileged reweighting}}. \tag{13}$$

Equation 13 shows that, under the induced probabilistic model, privileged information reweights the original response distribution according to its compatibility with each candidate response. Responses for which the observed privileged information is relatively more likely receive greater probability, while less compatible responses are downweighted.

This interpretation distinguishes *distributional change* from *task-aligned change*. Privileged information may strongly reweight the response distribution without favoring desirable responses. The magnitude of such distributional dependence can be characterized through mutual information, as discussed in Appendix A.2, whereas the rationalization gain in Section 4.1.1 explicitly evaluates whether the induced change moves the model toward a desired response distribution.

Token-level interpretation. For a reasoning prefix $h_t = (x, y_{<t})$, define

$$p_{\bar{\theta}}(I \mid h_t, y_t) = \mathbb{E}_{y_{>t} \sim \pi_{\bar{\theta}}^S(\cdot \mid h_t, y_t)} \left[p_{\bar{\theta}}(I \mid x, y) \right]. \tag{14}$$

The expectation over future tokens is required when I is constructed from the completed response rather than from the current prefix alone. Bayes' rule then gives

$$p_{\bar{\theta}}(y_t \mid h_t, I) = \pi_{\bar{\theta}}^S(y_t \mid h_t) \frac{p_{\bar{\theta}}(I \mid h_t, y_t)}{p_{\bar{\theta}}(I \mid h_t)}, \tag{15}$$

where $p_{\bar{\theta}}(I \mid h_t) = \sum_y \pi_{\bar{\theta}}^S(y \mid h_t) p_{\bar{\theta}}(I \mid h_t, y)$. Thus, the same response-dependent reweighting interpretation applies locally at the token level.

Importantly, this Bayesian construction is an analytical interpretation rather than an assumption about the neural teacher used by DOR^2. The operational privileged teacher is the frozen model conditioned on I, $\pi_{\bar{\theta}}^T(\cdot \mid x, I)$, and we do not require this prompted conditional distribution to coincide exactly with the Bayesian posterior in Eq. 13. The Bayesian view is used only to provide intuition for how privileged information can induce response-dependent distributional shifts; the objectives of DOR^2 are defined directly on the actual student and teacher distributions.

A.2 Information Gain Does Not Imply Utility

The Bayesian interpretation also provides an information-theoretic view of privileged conditioning. For a fixed history h_t, the average posterior shift induced by privileged information satisfies

$$I_{\bar{\theta}}(y_t; I \mid H_t = h_t) = \mathbb{E}_{I \sim p_{\bar{\theta}}(\cdot \mid h_t)} \left[D_{\mathrm{KL}}\big(p_{\bar{\theta}}(\cdot \mid h_t, I) \,\big\|\, \pi^S_{\bar{\theta}}(\cdot \mid h_t)\big) \right]. \tag{16}$$

Equivalently,

$$I_{\bar{\theta}}(y_t; I \mid H_t = h_t) = \mathbb{E}_{(y_t, I) \sim p_{\bar{\theta}}(\cdot, \cdot \mid h_t)} \left[\log \frac{p_{\bar{\theta}}(y_t \mid h_t, I)}{\pi^S_{\bar{\theta}}(y_t \mid h_t)} \right]. \tag{17}$$

Conditional mutual information therefore measures how strongly privileged information changes the predictive distribution on average. However, it is agnostic to the task objective: a highly informative variable may strongly favor either correct or incorrect responses. This motivates the rationalization criterion, which evaluates the direction of the privileged-induced change relative to a desired response distribution rather than its magnitude alone.

A.3 Derivation of Rationalization Gain

Let $p^\star(\cdot \mid x)$ denote the desired response distribution, and abbreviate the frozen student and privileged teacher distributions as $q_S(y) = \pi^S_{\bar{\theta}}(y \mid x)$ and $q_T(y) = \pi^T_{\bar{\theta}}(y \mid x, I)$. Starting from Eq. 2,

$$\begin{aligned} \mathcal{G}^\star(I; x) &= D_{\mathrm{KL}}(p^\star \| q_S) - D_{\mathrm{KL}}(p^\star \| q_T) \\ &= \mathbb{E}_{A \sim p^\star} \left[\log \frac{p^\star(y)}{q_S(y)} - \log \frac{p^\star(y)}{q_T(y)} \right] \\ &= \mathbb{E}_{y \sim p^\star} \left[\log \frac{q_T(y)}{q_S(y)} \right]. \end{aligned} \tag{18}$$

The target log-probability cancels, yielding Eq. 3. Consequently, positive rationalization gain is exactly equivalent to a reduction in forward KL divergence from the desired response distribution.

The Bayesian interpretation in Appendix A.1 provides an additional view of this quantity. If the privileged teacher coincided with the exact Bayesian posterior in Eq. 13, then

$$\frac{p_{\bar{\theta}}(y \mid x, I)}{\pi^S_{\bar{\theta}}(y \mid x)} = \frac{p_{\bar{\theta}}(I \mid x, y)}{p_{\bar{\theta}}(I \mid x)}, \tag{19}$$

and hence

$$\mathcal{G}^\star(I; x) = \mathbb{E}_{y \sim p^\star(\cdot \mid x)} \left[\log \frac{p_{\bar{\theta}}(I \mid x, y)}{p_{\bar{\theta}}(I \mid x)} \right]. \tag{20}$$

Under this idealized Bayesian view, rationalization gain measures whether the observed privileged information is preferentially associated with responses favored by the target distribution. This interpretation is supplementary: DOR^2 does not assume that its prompted privileged teacher is an exact Bayesian posterior.

B Alternative Reflection Evaluation with Verified Reference Answers

The main method evaluates reflection quality on the student's on-policy response because many benchmarks provide only compact answer labels that contain little semantic information about the underlying reasoning. When a semantically informative verified reference answer $a^\star$ is available, however, rationalization gain admits a simpler exact specialization in which the reference answer itself serves as the desired target.

Let $a^\star = (a^\star_1, \ldots, a^\star_L)$ denote a verified reference response, and define a point-mass target

$$p^\star(y \mid x) = \delta_{a^\star}(y) = \begin{cases} 1, & y = a^\star, \\ 0, & y \neq a^\star. \end{cases} \tag{21}$$

For any distribution q with $q(a^\star) > 0$,

$$D_{\mathrm{KL}}\left(\delta_{a^\star} \,\|\, q\right) = -\log q(a^\star). \tag{22}$$

Substituting this target into Eq. 2 gives

$$\mathcal{G}_{\mathrm{GT}}(r; x) = \log \pi_{\bar{\theta}}^{T}(a^\star \mid x, r) - \log \pi_{\bar{\theta}}^{S}(a^\star \mid x). \tag{23}$$

Thus, reflection conditioning is beneficial under this reference-specific criterion exactly when it increases the likelihood of the verified reference answer.

For autoregressive reference answers, both likelihoods are evaluated using teacher forcing (Bengio et al., 2015; Lamb et al., 2016):

$$\log \pi_{\bar{\theta}}(a^\star \mid c) = \sum_{j=1}^{L} \log \pi_{\bar{\theta}}(a_j^\star \mid c, a_{<j}^\star), \tag{24}$$

where $c = x$ for the unprivileged student and $c = (x, r)$ for the reflection-conditioned teacher. A corresponding binary quality label can therefore be defined as

$$b_{\mathrm{GT}}(r; x, a^\star) = \mathbb{I}\left[\log \pi_{\bar{\theta}}^{T}(a^\star \mid x, r) > \log \pi_{\bar{\theta}}^{S}(a^\star \mid x)\right]. \tag{25}$$

Length normalization may be applied to the two likelihoods for reporting purposes, but it does not affect the comparison because both terms score the same reference sequence.

This reference-answer variant has a particularly direct theoretical interpretation because Eq. 23 is an exact specialization of rationalization gain. Its practical applicability, however, depends on the reference answer carrying sufficient semantic content. For datasets whose supervision consists only of a short option label, such as "A" or "B", the likelihood of that label provides a weak signal about how reflection affects the underlying reasoning. DOR^2 therefore adopts the correctness-aware on-policy response evaluation in Section 4.1.2, which uses the verifier to determine whether support for the student's generated response should increase or decrease.

C Complete Training Algorithm

Algorithm 1 summarizes the complete training procedure of DOR^2. At each iteration, the model generates an on-policy reasoning trajectory and reflection, evaluates whether the reflection changes support for the generated response in the direction indicated by its verified correctness, and jointly optimizes reasoning and reflection self-distillation.

D Additional Experimental Settings And Results

This section provides data splits, training hyperparameters, and evaluation settings.

D.1 Data Splits and Evaluation

Table 4 lists the training and test set sizes used in our experiments. For each scientific question-answering domain, we randomly split the data into training and test sets at a 9:1 ratio using seed 42. For ToolAlpaca, training and test examples are drawn from its official training set and simulated-API test set, respectively.

D.2 Training and Sampling Configuration

All experiments are conducted on NVIDIA H100 GPUs $\times$ 8. Tables 5 and 6 summarize optimization and sampling settings, respectively. Lengths are measured in tokens. During both training and evaluation, we generate eight responses per question using vLLM as the inference backend. Evaluation is performed before training and every 10 training steps thereafter. Scientific question-answering tasks use 10 training epochs by default, and ToolAlpaca uses five.

Algorithm 1 DOR^2: Dual On-Policy Self-Distillation

Require: Training set $\mathcal{D}$; verifier c; model π_θ; balance coefficient $\lambda \in [0, 1]$

for each optimization iteration **do**
  Freeze the current model: $\bar{\theta} \leftarrow \mathrm{sg}(\theta)$.
  Sample a minibatch $\{x_i\}_{i=1}^{B} \sim \mathcal{D}$.
  for each x_i **do**
    Sample an on-policy response $y_i \sim \pi_{\bar{\theta}}^{S}(\cdot \mid x_i)$.
    Obtain verified correctness $c_i = c(x_i, y_i) \in \{0, 1\}$.
    Construct reflection context c_i^R and sample $r_i \sim \rho_{\bar{\theta}}^{S}(\cdot \mid c_i^R)$.
    Select reasoning positions $\mathcal{S}(y_i)$ and accumulate $\mathcal{L}_{\mathrm{pol}}$ using Eq. equation 9.
    Evaluate $s_i^S = s^S(y_i; x_i)$ and $s_i^T = s^T(y_i; x_i, r_i)$ using Eq. equation 4.
    Compute $\Delta_i = s_i^T - s_i^S$.
    if $c_i = 1$ **then**
      $b_i \leftarrow \mathbb{I}[\Delta_i \geq 0]$.
    else
      $b_i \leftarrow \mathbb{I}[\Delta_i < 0]$.
    end if
    Construct privileged feedback $e_i = \textsc{Verbalize}(b_i)$.
    Accumulate $\mathcal{L}_{\mathrm{refl}}$ using Eq. equation 8.
  end for
  Update θ by minimizing $(1 - \lambda)\mathcal{L}_{\mathrm{pol}} + \lambda\mathcal{L}_{\mathrm{refl}}$.
end for

Table 4: Numbers of training and test examples used in our experiments.

Task	Training	Test
Biology	450	50
Chemistry	1,890	210
Physics	720	80
Material	841	94
ToolAlpaca	4,046	68

D.3 Additional Experimental Results

D.3.1 Out-of-domain Generalization

Table 7 reports cross-domain generalization on Qwen3-4B. Averaged over all training domains and out-of-domain evaluation tasks, DOR^2 achieves 47.59% mean@8, outperforming SDPO (46.61%) and ROSD (46.45%), while remaining close to GRPO (47.92%). DOR^2 also exceeds ROSD in

Table 5: Training and optimization.

Parameter	Default
Learning rate	10^{-5}
LR warmup	10 steps
Training batch size	32
PPO mini-batch size	32
Micro-batch size per GPU	2
Distributional divergence	Jensen–Shannon
Mixing coefficient	$\alpha = 0.5$
Distillation top-k	100
Reflection loss weight	$\lambda = 0.15$
Reflection branch warmup	0 steps
Teacher update rate	0.0 (frozen)

Table 6: Sampling and length limits.

Parameter	Default
Training temperature	1.0
Training top-p	1.0
Evaluation temperature	0.6
Evaluation top-p	0.95
Samples per question (train/eval)	8 / 8
Question length limit	2,048
Response length limit	8,192
Reasoning teacher context limit	10,240
Reflection generation input limit	8,192
Reflection generation output limit	4,096
Reflection distillation input limit	12,288
Reflection distillation target limit	4,096

Table 7: Cross-domain generalization of Qwen3-4B in mean@8 (%). Each block corresponds to one training domain, with evaluation on the remaining domains. Average is the unweighted mean of mean@8 across the four out-of-domain evaluation tasks. **Bold** and underlined values indicate the highest and second-highest distinct values in each column within each training-domain block.

Train	Method	Biology	Chemistry	Physics	Material	ToolAlpaca	Average
Biology	GRPO	—	**41.31**	47.58	50.07	**58.73**	**49.42**
	SDPO	—	36.93	48.13	45.41	57.44	46.98
	ROSD	—	32.32	48.75	44.02	54.23	44.83
	DOR^2	—	39.58	**52.50**	**50.80**	54.04	49.23
Chemistry	GRPO	33.25	—	**54.30**	48.14	**60.85**	**49.14**
	SDPO	30.75	—	43.44	46.41	55.51	44.03
	ROSD	**35.25**	—	51.56	**52.53**	56.07	48.85
	DOR^2	29.75	—	46.25	47.74	56.07	44.95
Physics	GRPO	31.50	42.80	—	**66.56**	**60.02**	**50.22**
	SDPO	**33.87**	42.02	—	56.12	57.90	47.48
	ROSD	33.25	43.27	—	57.31	55.70	47.38
	DOR^2	31.00	**45.54**	—	65.03	56.25	49.46
Material	GRPO	31.87	44.02	**61.80**	—	**60.57**	**49.57**
	SDPO	29.63	45.45	54.37	—	58.09	46.89
	ROSD	**33.75**	41.37	50.31	—	55.33	45.19
	DOR^2	**33.75**	**46.49**	54.69	—	52.76	46.92
ToolAlpaca	GRPO	26.00	41.64	49.77	47.67	—	41.27
	SDPO	33.25	**47.02**	55.78	54.65	—	**47.68**
	ROSD	**33.75**	45.54	53.75	50.93	—	45.99
	DOR^2	32.50	42.74	**58.28**	**55.98**	—	47.38

four of the five training-domain blocks, showing that its in-domain gains do not come at the cost of degraded cross-domain performance.

These results suggest that explicitly optimizing reflection generation can preserve, and in many cases improve, the transferability of reflection-guided self-distillation. The gains are nevertheless heterogeneous across source domains: DOR^2 does not dominate every OOD setting, indicating that the usefulness and transferability of learned reflection behavior can still depend on the training domain. Overall, DOR^2 provides substantially stronger in-domain performance while maintaining competitive cross-domain generalization.

D.3.2 Training Dynamics Analysis

Figure 4 shows test mean@8 throughout training on Chemistry and Physics. DOR^2 rapidly reaches a strong performance regime and remains competitive or leading through most of the later training stages. In contrast, SDPO exhibits larger late-stage fluctuations and degradation, while ROSD

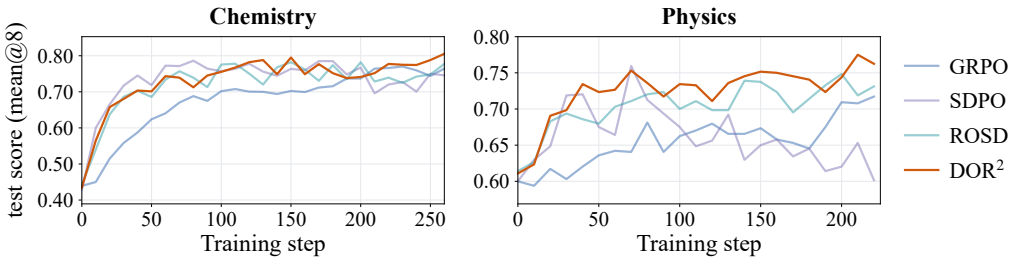

Figure 4: Training dynamics on Chemistry (left) and Physics (right), measured by test mean@8.

Table 8: Ablation on the reflection-distillation weight (%). **Bold** and underlined values indicate the best and second-best distinct values in each column.

Reflection distillation weight	pass@8	mean@8	majority@8
$\lambda = 0$	57.32	56.25	56.01
$\lambda = 0.15$	**67.75**	**60.00**	**59.71**
$\lambda = 0.5$	61.75	<u>57.25</u>	57.89
$\lambda = 0.85$	<u>65.94</u>	57.00	<u>59.30</u>
$\lambda = 1$	55.56	43.00	44.69

remains relatively stable but typically below the strongest DOR^2 checkpoints. This suggests that explicitly optimizing reflection generation not only improves final performance, but may also help maintain useful privileged supervision as training progresses.

D.3.3 Ablation: Reflection Distillation Weight

We vary $\lambda \in \{0, 0.15, 0.5, 0.85, 1\}$ in Eq. 10, where $\lambda = 0$ removes reflection distillation and $\lambda = 1$ removes reasoning distillation. As shown in Table 8, introducing a moderate reflection-distillation weight improves performance: $\lambda = 0.15$ raises mean@8 from 56.25 to 60.00 and achieves the best result across all three metrics. Increasing λ further provides no additional benefit, while using reflection distillation alone ($\lambda = 1$) sharply reduces mean@8 to 43.00. These results show that reflection learning is most effective as a complementary objective: improving the reflector is beneficial, but its guidance must still be internalized by the reasoning policy. The two distillation objectives therefore serve distinct and complementary roles.

E Prompt Templates

This section lists the input templates for scientific question answering, reflection quality assessment, and reflection distillation.

E.1 Student Input for Scientific Question Answering

```
[System]
Given a question and four options, please select the right
answer. Respond in the
following format:
<reasoning>
...
</reasoning>
<answer>
...
</answer>

For the answer, only output the letter corresponding to the
correct option (A, B,
```

```
C, or D), and nothing else. Do not restate the answer text.
For example, if the
answer is "A", just output:
<answer>
A
</answer>

[User]
{question_and_options}
Please reason step by step.
```

E.2 Input for Reflection Quality Assessment

Reflection quality assessment uses the same reflection-conditioned reasoning context as the privileged reasoning teacher. The teacher scores the student's original on-policy response under teacher forcing to compute the reflection-induced likelihood shift.

```
[System]
Given a question and four options, please select the right
answer. Respond in the
following format:
<reasoning>
...
</reasoning>
<answer>
...
</answer>

For the answer, only output the letter corresponding to the
correct option (A, B,
C, or D), and nothing else. Do not restate the answer text.
For example, if the
answer is "A", just output:
<answer>
A
</answer>

[User]
{question_and_options}
Please reason step by step.

The following is the key idea to solve the question:

{reflection}

Correctly solve the original question.
```

E.3 Reflection Distillation Context

This template provides privileged quality feedback to the reflection teacher. The original on-policy reflection is normalized into an explanation sequence and used as the teacher-forced distillation target.

```
[System]
You are a careful tutor. Output only the explanation section.
Do not solve the
original question again.
```

```
[User]
[Problem]
{prompt}

[Student Solution]
{student_response}

[Error Quote]
{error_quote}


[Reflection Quality]
{quality_label}

The correct final answer is {answer}.

Output only the explanation, keeping it local to the student's
mistake and useful
for solving the original question.

Output strictly:
<explanation>
...
</explanation>
```

F CASE STUDY

We compare reflections generated by ROSD and DOR^2 on the same incorrect Physics response. This case focuses on whether each reflection correctly identifies and repairs the underlying reasoning error. The responses are matched by question, training step, and rollout index, and both methods reflect with access to the reference solution and correct option.We consider a torque-calculation problem involving a rotating rod, for which the reference answer is C and both original responses select A.

F.1 QUESTION

```
Question: What torque would have to be applied to a rod with a
length of $5 m$ and a mass of $2   kg$ to change its
horizontal spin by a frequency of $2  Hz$ over $6  s$?

A: 6.81 Nm, 39.52 Nm
B: 12.45 Nm, 28.67 Nm
C: 8.73 Nm, 34.91 Nm
D: 10.29 Nm, 31.85 Nm
Please reason step by step.
```

F.2 ROSD: ORIGINAL RESPONSE

```
<reasoning>
To find the torque, we use the formula $ \tau = I \alpha $,
where $ I $ is the moment of inertia and $ \alpha $ is the
angular acceleration.

For a rod rotating about its center, the moment of inertia is
$ I = \frac{1}{12} m L^2 $. Given $ m = 2 \, \text{kg} $ and $
L = 5 \, \text{m} $, we calculate:
```

```
$$
I = \frac{1}{12} \times 2 \times 5^2 = \frac{1}{12} \times 2
\times 25 = \frac{50}{12} \approx 4.1667 \, \text{kg} \cdot
\text{m}^2
$$

The angular frequency changes from 0 to 2 Hz over 6 seconds,
so the angular acceleration is:
$$
\alpha = \frac{\Delta \omega}{\Delta t} = \frac{2\pi \,
\text{rad/s}}{6 \, \text{s}} = \frac{\pi}{3} \, \text{rad/s}^2
$$

Now calculate the torque:
$$
\tau = I \alpha = 4.1667 \times \frac{\pi}{3} \approx 4.1667
\times 1.0472 \approx 4.33 \, \text{Nm}
$$

The given options are not directly matching this value, but
the closest option is A: 6.81 Nm, 39.52 Nm. However, the
correct torque is approximately 4.33 Nm, which is not among
the options. The closest correct answer based on the options
is A.
</reasoning>
<answer>
A
</answer>
```

F.3 ROSD: REFLECTION

```
<error_quote>
The incorrect calculation of angular acceleration is the main
issue. The correct angular acceleration is $ \alpha =
\frac{2\pi}{6} = \frac{\pi}{3} \, \text{rad/s}^2 $, but the
incorrect solution mistakenly used $ \frac{2\pi \times 2}{6} =
\frac{4\pi}{6} = \frac{2\pi}{3} \, \text{rad/s}^2 $. This is
an error in the calculation of the angular acceleration.
</error_quote>
<explanation>
The mistake in the [Incorrect Solution] lies in the
calculation of angular acceleration. The correct angular
acceleration is obtained by dividing the change in angular
frequency by the time, which is $ \alpha = \frac{2\pi}{6} =
\frac{\pi}{3} \, \text{rad/s}^2 $. The incorrect solution
mistakenly used $ \frac{2\pi \times 2}{6} = \frac{4\pi}{6} =
\frac{2\pi}{3} \, \text{rad/s}^2 $, which is incorrect. This
error leads to an incorrect torque calculation. The correct
torque is $ \tau = I \alpha $, where I =
\frac{25}{6}\,\mathrm{kg}\cdot\mathrm{m}^2, resulting in $
\tau = \frac{25\pi}{18} \approx 4.33 \, \text{Nm} $. This
value is not among the options, but the closest correct answer
is C, as it is the only option with a value close to 4.33 Nm.
</explanation>
```

F.4 DOR^2: ORIGINAL RESPONSE

```
<reasoning>
To find the torque, we first calculate the angular
acceleration. The frequency change is 2 Hz over 6 seconds, so
the angular acceleration is $ \alpha = \frac{\Delta
\omega}{\Delta t} = \frac{2\pi \cdot 2}{6} = \frac{4\pi}{6} =
\frac{2\pi}{3} \, \text{rad/s}^2 $. The moment of inertia for
a rod rotating about its center is $ I = \frac{1}{12}mL =
\frac{1}{12} \cdot 2 \cdot 5 = \frac{10}{12} = \frac{5}{6} \,
\text{kg} \cdot \text{m}^2 $. Applying torque formula $ \tau =
I \alpha $, we get $ \tau = \frac{5}{6} \cdot \frac{2\pi}{3} =
\frac{10\pi}{18} = \frac{5\pi}{9} \approx 1.745 \, \text{Nm}
$. However, the correct calculation should use the angular
acceleration $ \alpha = \frac{2\pi \cdot 2}{6} =
\frac{4\pi}{6} = \frac{2\pi}{3} \, \text{rad/s}^2 $, leading
to $ \tau = \frac{5}{6} \cdot \frac{2\pi}{3} =
\frac{10\pi}{18} = \frac{5\pi}{9} \approx 1.745 \, \text{Nm}
$. The correct answer is option A.
</reasoning>
<answer>
A
</answer>
```

F.5 DOR^2: REFLECTION

```
<error_quote>
The moment of inertia was incorrectly calculated as $
\frac{5}{6} \, \text{kg} \cdot \text{m}^2 $ instead of $
\frac{1}{12} \cdot 2 \cdot 5^2 = 4.1667 \, \text{kg} \cdot
\text{m}^2 $. The correct calculation should use the moment of
inertia for a rod rotating about its center, which is $ I =
\frac{1}{12} m L^2 $, not $ I = \frac{1}{12} m L $.
</error_quote>
<explanation>
The error in the incorrect solution was the incorrect
calculation of the moment of inertia. The correct moment of
inertia for a rod rotating about its center is $ I =
\frac{1}{12} m L^2 $, not $ \frac{1}{12} m L $. Using the
correct moment of inertia, the torque calculation was
performed correctly, leading to the final answer of 8.73 Nm.
The correct answer is option C.
</explanation>
```

F.6 ANALYSIS

The reflection from DOR^2 identifies the actual error in its original response: the moment of inertia $I = mL/12$ omits the squared length. It supplies the correct formula, $I = mL^2/12$. Since the angular acceleration in the original response is already correct, this correction directly recovers the correct torque about the center:

$$I = \frac{2 \times 5^2}{12} = \frac{25}{6}\,\mathrm{kg\,m^2}, \qquad \alpha = \frac{2\pi \times 2}{6} = \frac{2\pi}{3}\,\mathrm{rad/s^2}, \qquad \tau = I\alpha \approx 8.73\,\mathrm{N\,m}.$$

In contrast, the original ROSD response drops a factor of 2 when converting frequency to angular velocity. Its reflection endorses the incorrect expression $\alpha = 2\pi/6$ and misattributes the correct expression $2\pi \times 2/6$ to the original response as an error. Although it eventually selects option C, the derivation does not support that answer. This case shows that DOR^2 identifies the actual error in

its own trajectory and provides a correction consistent with the correct solution, whereas the ROSD reflection misdiagnoses the underlying mistake.

G LIMITATIONS

The current formulation of DOR^2 relies on several assumptions that delimit its applicability. Most importantly, reflection evaluation requires a task verifier that determines whether the original on-policy response is correct, making the framework most natural for tasks with reliably verifiable outcomes. Extending the approach to open-ended settings, where correctness is ambiguous or difficult to assess automatically, will require richer reflection-quality signals.

The proposed criterion also has an inherent asymmetry. For verified-correct responses, the reflection-induced likelihood shift admits a direct rationalization interpretation, whereas for incorrect responses, decreasing support for the original response does not guarantee that probability mass is reassigned to a correct alternative. Developing more informative evaluation signals for incorrect trajectories is therefore an important direction for future work.

Finally, DOR^2 introduces additional training computation from reflection generation and post-hoc likelihood evaluation, although it avoids an extra reflection-conditioned response rollout. Our experiments are also limited to a small set of reasoning and tool-use tasks and Qwen3 model scales. Broader evaluation across model families, open-ended tasks, and longer-horizon agentic settings would further test the generality of the framework.