SAME ANSWER, DIFFERENT LESSON: DEFENDING ON-POLICY DISTILLATION VIA SIGNAL DECOUPLING

Anonymous authors
Paper under double-blind review

ABSTRACT

On-policy distillation (OPD) poses a growing threat to proprietary large language models (LLMs), as a student can efficiently imitate the teacher under per-token supervision on its own trajectories. Yet existing anti-distillation defenses target teacher outputs, leaving OPD's student-conditioned supervision underexplored. In this work, we address OPD protection in two components: First, we formulate OPD protection as a constrained adversarial optimization problem that decouples the teacher's prediction decision from the supervision signal exposed to the student, allowing the latter to be weakened while preserving the former. Second, we propose *Selective Asymmetric Supervision Reshaping (SASR)* to realize this objective at the OPD scoring interface. SASR consists of two key designs: 1) selective intervention, which focuses protection on positions with large teacher-student disagreement, 2) asymmetric reshaping, which modifies supervision according to the direction of the teacher-student gap. A decision-preserving projection further constrains the reshaped distribution to retain the teacher decision while controlling supervision distortion. Experiments across different teacher–student capacity settings and multiple OPD variants show that SASR consistently suppresses student performance, reducing accuracy by up to 7.26 percentage points compared with the evaluated state-of-the-art defense. To our knowledge, SASR is the first defense to explicitly target student-conditioned supervision in OPD, addressing a previously underexplored protection surface.

1 INTRODUCTION

Large language models (LLMs) have demonstrated strong capabilities across a wide range of tasks, including sentiment analysis and code generation. Developing such capable models requires substantial training resources, making their model knowledge a valuable intellectual property. Knowledge distillation transfers knowledge and capabilities from strong teacher models to student models through either offline or online distillation. In particular, on-policy distillation (OPD) has emerged as a prominent paradigm by providing teacher supervision on student-generated trajectories (Agarwal et al., 2024; Gu et al., 2023; Yang et al., 2026; Li et al., 2026) as shown in Fig. 1. However, this rich supervision also exposes model knowledge to unauthorized distillation, raising an important challenge for reducing the distillation value of exposed supervisory signals while preserving the utility of the teacher output quality.

To pursue this joint objective, recent studies have explored defenses against distillation, i.e., anti-distillation, by modifying the outputs exposed by the teacher (Savani et al., 2025; Allouah et al., 2026; Hartman et al., 2026; Fang et al., 2026). These defense methods mainly intervene at two levels: text-based distillation and logit-based distillation. Text-based approaches operate on generated content, rewriting or obfuscating responses and reasoning traces to reduce their value for downstream imitation while maintaining their utility to users. For logit-based protection, Conditional Mutual Information-Based approach (Fang et al., 2026) defenses globally transform the full teacher logit vector through a learned transformation matrix to suppress distillation-relevant information. However, existing defenses do not address on-policy supervision at student-generated contexts. Crucially, output decisions and OPD learning signals are distinct: the former depend on the top-ranked token, while the latter depend on the underlying distribution. This distinction suggests that the learning signal can be reshaped while preserving the teacher output decision.

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.

Building on this observation, we first formulate protected OPD as a constrained adversarial problem. This formulation, in turn, motivates *Selective Asymmetric Supervision Reshaping (SASR)(Fig. 2)*, which reshapes teacher feedback according to its effect on student learning. Specifically, SASR operates through two core mechanisms: 1) *selective intervention*, which ranks token positions by teacher-student probability discrepancy and retains only a top fraction with the largest absolute differences, concentrating protection on positions where their predictive distributions diverge most; and 2) *asymmetric reshaping*, which adjusts the supervision at selected positions according to the direction of the learning signal, while constraining the resulting distribution to preserve the teacher's top-ranked token. Together, these designs directly weaken the supervision consumed by the OPD student while retaining the teacher prediction decision.

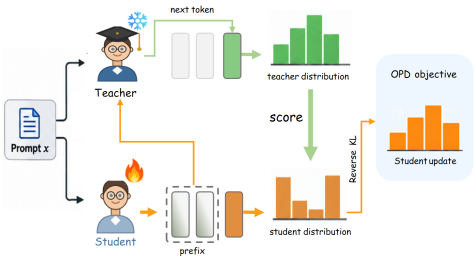

Figure 1: Illustration of OPD. The teacher scores student-sampled tokens and the resulting teacher–student gap guides updates to the student distribution.

The main contributions of this study are threefold:

- **Adversarial formulation.** We formulate anti-distillation in OPD as a constrained adversarial problem that explicitly separates the teacher output decision from the learning signal delivered to the student. This formulation characterizes how supervision can be reshaped within a decision-preserving region to reduce its effectiveness for student learning.
- **Selective OPD defense.** We propose *Selective Asymmetric Supervision Reshaping (SASR)*, which selectively targets high-discrepancy positions and asymmetrically reshapes their supervision according to the learning-signal direction, thereby reducing distillation effectiveness without altering the teacher decision or sacrificing utility.
- **Extensive empirical evaluation.** Extensive experiments across multiple OPD settings, representative distillation algorithms, and seven widely used datasets, demonstrate the effectiveness of SASR in suppressing unauthorized on-policy distillation.

2 Adversarial Formulation for Distillation Protection

We formulate distillation protection as a constrained adversarial problem over the teacher-student supervision interface and instantiate it in the OPD setting studied in this work. We first define the protection setting, then characterize teacher fidelity through decision-preserving and bounded-distortion constraints. Finally, we introduce a vanilla-relative adversarial objective that measures protection through its effect on the resulting distilled student.

2.1 Teacher–Student Protection Setting

To characterize the supervision process in on-policy distillation, we model the interaction between a student policy and a teacher providing token-level feedback on student-generated trajectories. Unlike distillation on fixed data, this interaction is inherently adaptive: each student update changes the policy, which in turn alters subsequent trajectories and the contexts on which the teacher is queried.

Given an input prompt $x \sim \mathcal{D}$, where $\mathcal{D}$ denotes the prompt distribution, let y denote a complete response sampled from the student policy. At step t, let $y_{<t} = (y_1, \ldots, y_{t-1})$ denote the generated prefix and define the current context as $c_t = (x, y_{<t})$. For a vocabulary $\mathcal{V}$, the vanilla teacher supervision interface is

$$R_T(c_t) = P_T(\cdot \mid c_t), \tag{1}$$

where $P_T(\cdot \mid c_t)$ denotes the teacher's conditional next-token distribution. Analogously, $P_{S,\theta}(\cdot \mid c_t)$ denotes the corresponding student distribution, where θ is student model parameters.

From the attacker's perspective, an admissible OPD strategy $a \in \mathcal{A}$ specifies how the exposed teacher feedback is converted into student updates, including the training loss, optimizer, and feedback processing. For a supervision interface R, let $S_a(R)$ denote the resulting student under strategy a; in particular, $S_a(R_T)$ represents the student distilled from the vanilla teacher interface.

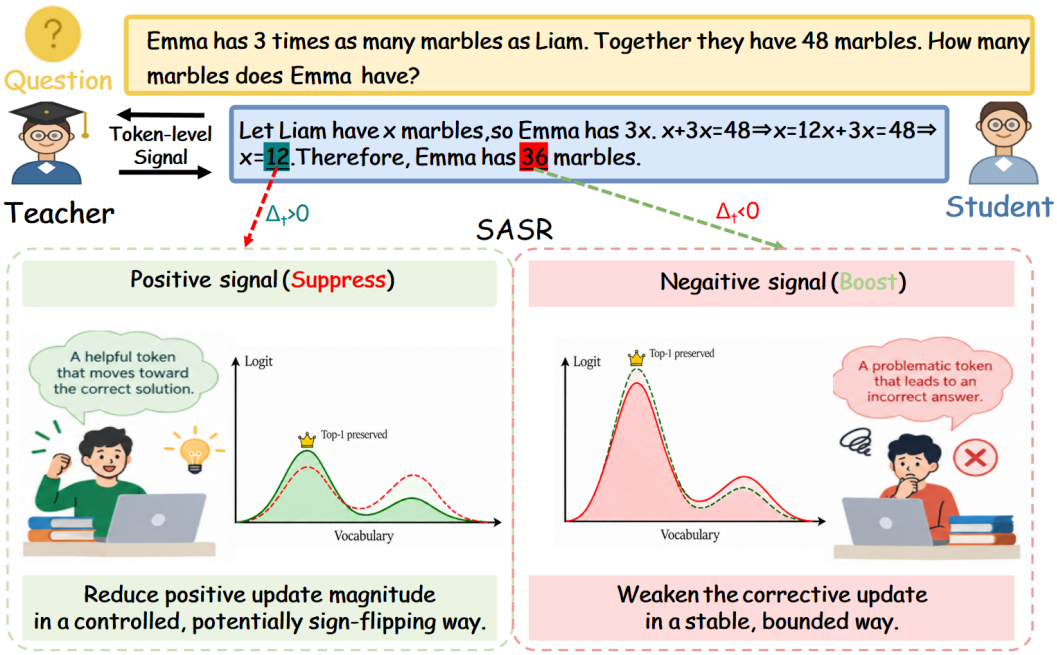

Figure 2: Overview of SASR. SASR operates on teacher supervision at student-generated contexts. It selects high-disagreement positions, asymmetrically reshapes the sampled-token supervision according to the sign of the teacher–student gap, and projects the modified distribution to preserve the vanilla teacher's top-1 decision. Positive signal is suppressed, while negative signal is weakened.

The defender replaces R_T with a protected interface $\widetilde{R}_T$ that weakens the supervision available to the student, while leaving the student model and update rule unchanged. To preserve teacher utility, $\widetilde{R}_T$ is further constrained to retain the teacher decision. Therefore, protection effectiveness is determined by how much student learning can be suppressed under this fidelity requirement.

2.2 Teacher Fidelity as a Constraint

To maintain the fidelity of the protected teacher, we constrain the interface along two dimensions: preservation of the teacher decision and bounded distortion of the returned supervision.

First, the protected interface should preserve the teacher top-1 decision at each queried context. Let $\mathcal{C}_q$ denote the set of contexts queried during OPD and $\mathcal{V}$ the model vocabulary. Following Eq. (1), we consider a protected interface $\widetilde{R}_T$ that returns a modified next-token distribution $\widetilde{P}_T(\cdot \mid c)$ for each $c \in \mathcal{C}_q$. Assuming a fixed tie-breaking rule for $\arg\max$, we define the decision-preserving set as

$$\mathcal{B}_{\text{top}}(R_T) = \{\widetilde{R}_T : \arg\max_{v \in \mathcal{V}} \widetilde{P}_T(v \mid c) = \arg\max_{v \in \mathcal{V}} P_T(v \mid c), \quad \forall c \in \mathcal{C}_q\}, \tag{2}$$

Here, $\mathcal{B}_{\text{top}}(R_T)$ contains all protected interfaces whose top-1 decision agrees with that of the vanilla teacher at every queried context.

Second, the protected supervision should remain close to the vanilla supervision at the distribution level. Using the queried context set $\mathcal{C}_q$ defined above, we measure the average deviation between the protected and vanilla teacher distributions by the KL divergence. Let $b \geq 0$ denote the target average KL-distortion budget used to control the strength of the protected supervision. We define

$$\mathcal{B}_b(R_T) = \{\widetilde{R}_T : \frac{1}{|\mathcal{C}_q|} \sum_{c \in \mathcal{C}_q} D_{\text{KL}}\left(\widetilde{P}_T(\cdot \mid c) \| P_T(\cdot \mid c)\right) \leq b\}. \tag{3}$$

where $D_{\text{KL}}\left(\widetilde{P}_T(\cdot \mid c) \| P_T(\cdot \mid c)\right)$ measures the distributional deviation introduced by the protected interface at context c, $|\mathcal{C}_q|$ denotes the number of queried contexts, and b specifies the maximum

allowable average deviation. $\mathcal{B}_b(R_T) = \mathcal{B}_{\text{top}}(R_T) \cap \mathcal{B}_b(R_T)$, it contains all protected interfaces whose average KL divergence from the vanilla teacher over the queried contexts does not exceed b.

Accordingly, these constraints define the admissible protection set $\mathcal{B}(R_T)$, within which the defender seeks to maximally reduce the utility of the distilled student.

2.3 Adversarial objective

With the admissible protection set defined, we next quantify how effectively a protected interface suppresses distillation. Since different OPD strategies $a \in \mathcal{A}$ may achieve different absolute student utility even under the same vanilla teacher, we evaluate each protected run relative to the vanilla run of the *same* strategy.

Let $U_{\text{atk}}(S)$ denote the attacker's held-out utility for student S, with larger values indicating a more useful distilled student. We define the vanilla-relative defense gain as

$$G_{\text{def}}(a, \widetilde{R}_T; R_T) = U_{\text{atk}}\left(S_a(R_T)\right) - U_{\text{atk}}\left(S_a(\widetilde{R}_T)\right), \tag{4}$$

where $S_a(R_T)$ and $S_a(\widetilde{R}_T)$ are the students obtained under the same distillation strategy a using the vanilla and protected teacher interfaces, respectively. Hence, $G_{\text{def}} > 0$ indicates a reduction in student utility caused by the protected supervision, while $G_{\text{def}} \approx 0$ indicates little measurable protection against strategy a.

Using G_{def} as the protection criterion, we cast the interaction between the protected teacher and the distillation strategy as a maximin optimization problem.

$$\widetilde{R}_T^* = \arg \max_{\widetilde{R}_T \in \mathcal{B}(R_T)} \inf_{a \in \mathcal{A}} G_{\text{def}}(a, \widetilde{R}_T; R_T). \tag{5}$$

The inner minimization selects the admissible distillation strategy that yields the smallest defense gain, corresponding to the attacker least affected by the protection.

The outer maximization then searches over $\mathcal{B}(R_T)$ for the protected interface that maximizes this worst-case vanilla-relative degradation.

This formulation separates the protection problem into two components: 1) The admissible set $\mathcal{B}(R_T)$ specifies how the teacher supervision may be altered under fidelity constraints, while G_{def} quantifies whether such changes actually reduce the utility acquired by the student. 2) The maximin structure further accounts for adaptation by requiring the protected interface to remain effective against the least-affected admissible distillation strategy.

3 Method

Building on the protection objective in Section 2, we introduce *Selective Asymmetric Supervision Reshaping (SASR)*, as illustrated in Fig. 2, to weaken the supervision available for OPD while preserving teacher fidelity. First, we characterize the teacher–student learning signal that governs the student update. Second, SASR identifies high-disagreement positions and reshape their supervision signals according to the direction of disagreement. To satisfy the teacher-fidelity constraints, the reshaped signal is realized as a protected distribution with preserved top-1 decision and bounded distortion.

3.1 Learning Signal in OPD

Standard OPD minimizes the reverse KL between the student and teacher on student-generated trajectories (Gu et al., 2023; Yang et al., 2026). Under the commonly used sampled-token approximation, the OPD update takes the form

$$\nabla_\theta \mathcal{J}_{\text{OPD}}(\theta) = \mathbb{E}_{x \sim \mathcal{D},\, y \sim P_{S,\theta}(\cdot|x)} \left[\sum_{t=1}^{T} \left(\log P_S(y_t \mid c_t) - \log P_T(y_t \mid c_t) \right) \nabla_\theta \log P_S(y_t \mid c_t) \right], \tag{6}$$

where the expectation is taken over prompts and student-generated trajectories, and θ denotes the student parameters.

Following prior token-level analyses of OPD (Shao et al., 2026), we denote the teacher–student log-probability gap for a student-sampled token y_t as

$$\Delta_t = \log P_T(y_t \mid c_t) - \log P_{S,\theta}(y_t \mid c_t), \tag{7}$$

Substituting Eq. (7) into Eq. (6) gives

$$\nabla_\theta \mathcal{J}_{\mathrm{OPD}}(\theta) = -\mathbb{E}_{x\sim\mathcal{D},\, y\sim P_{S,\theta}(\cdot|x)} \left[\sum_{t=1}^{T} \Delta_t \nabla_\theta \log P_{S,\theta}(y_t \mid c_t) \right], \tag{8}$$

Hence, Δ_t directly determines the signed sampled-token update, and SASR uses it to guide signal reshaping.

However, the same decision can correspond to different update signals. The following proposition shows that prediction-equivalent teacher distributions can induce opposite local update directions.

Proposition 1 (Prediction Equivalence with Opposite Updates)**.** *Consider a binary vocabulary* $(\{0,1\})$*. Let* $s = P_S(1) \in (1/2, 1)$ *and* $p = P_T(1) \in (0,1)$ *denote the student and teacher probabilities of token* 1*, respectively, yielding distributions* $(s, 1-s)$ *and* $(p, 1-p)$*. Let* $z = \log \frac{s}{1-s}$ *denote the student's binary logit. The reverse-KL derivative with respect to* z *is*

$$\frac{\mathrm{d}}{\mathrm{d}z} D_{\mathrm{KL}}\left((s,1-s)\|(p,1-p)\right) = s(1-s)\log\frac{s(1-p)}{p(1-s)}. \tag{9}$$

For any $s \in (1/2, 1)$, there exist $p_- < s < p_+$ with $p_-, p_+ \in (1/2, 1)$ such that both teachers predict token 1, while the corresponding derivatives have opposite signs. Since $s(1-s) > 0$, the derivative is negative when $p > s$ and positive when $p < s$. Under gradient descent, the cases p_+ and p_- therefore move the same student logit z in opposite directions, despite preserving the same teacher top-1 prediction.

Proposition 1 shows that prediction preservation does not uniquely determine the local OPD update signal, motivating the disagreement-aware shaping introduced next. The proof is provided in Appendix A.2.

3.2 Disagreement-Aware Signal Shaping

Equation (8) reveals two forms of heterogeneity in OPD supervision: 1) the magnitude of Δ_t varies across token positions, 2) its sign determines the update direction. As illustrated in Fig. 2, SASR selectively targets high-disagreement positions and reshapes positive and negative signals separately.

Token Selection. Building on TSD-KD's (Kim & Baek, 2026) use of teacher–student confidence discrepancy, SASR prioritizes tokens with large teacher-student gaps. Specifically, given a student-generated response $y = (y_1, \ldots, y_T)$, SASR computes the disagreement magnitude $|\Delta_t|$ at each valid response position t, where Δ_t is the teacher–student log-probability gap defined in Eq. (7). The response positions are then ranked in descending order of $|\Delta_t|$, prioritizing positions with larger teacher–student disagreement.

A selection rate $\rho \in [0,1]$ determines the fraction of response positions selected for intervention. SASR retains the top-ρ fraction in the ranking and encodes the selection with a binary mask $m_t \in \{0,1\}$, where $m_t = 1$ denotes a selected position and $m_t = 0$ otherwise. Selected positions are passed to the subsequent signal-shaping step, while the remaining positions retain the vanilla teacher supervision.

Asymmetric Signal Shaping. Selection determines where the supervision is modified, while the sign of Δ_t determines how it should be reshaped. Recall that $\Delta_t > 0$ indicates that the teacher assigns a higher probability to the sampled token than the student, whereas $\Delta_t < 0$ indicates the opposite. These two cases correspond to different directions of the OPD learning signal and are therefore treated separately.

For a selected position, SASR defines the target change to the teacher log-probability as

$$\delta_t = -\lambda m_t w_t \Delta_t, \qquad w_t = \begin{cases} w_+, & \Delta_t \geq 0, \\ w_-, & \Delta_t < 0, \end{cases} \tag{10}$$

where λ controls the overall intervention strength. w_+ and w_- control the relative strengths of the positive and negative branches, respectively. Since $m_t = 0$ for unselected positions, Eq. (10) leaves their supervision unchanged. The target teacher–student gap is

$$\Delta_t^{tar} = \Delta_t + \delta_t = (1 - \lambda m_t w_t)\, \Delta_t \tag{11}$$

For $\Delta_t > 0$, the target change decreases the teacher log-probability of the sampled token, reducing the positive teacher–student gap and potentially reversing its sign. For $\Delta_t < 0$, the target change increases the teacher log-probability, moving the negative gap toward zero or across it. The two branch weights therefore allow the strength of these effects to be controlled independently.

Eq. (10) specifies the desired supervision change before teacher fidelity is enforced. Because this target modification may violate the admissible set $\mathcal{B}(R_T)$ defined in Section 2, SASR next converts it into a feasible protected teacher distribution.

3.3 Fidelity-Constrained Realization

The target change in Eq. (10) specifies how SASR reshapes the teacher supervision before fidelity is enforced. An excessive change may alter the teacher decision or introduce excessive distributional distortion. SASR therefore realizes the target change through a decision-preserving projection, while using the KL budget b to calibrate the global intervention strength λ.

Let $p_t = P_T(y_t \mid c_t)$ denote the vanilla teacher probability of the sampled token, and let $v \in \mathcal{V}$ denote an arbitrary vocabulary token. Applying the target log-probability change δ_t gives the unprojected value $p_t \exp(\delta_t)$. SASR projects this value onto the feasible interval that preserves the vanilla teacher's local top-1 decision, yielding the realized probability $\widetilde{p}_t = \widetilde{P}_T(y_t \mid c_t)$. The protected distribution is then constructed as

$$\widetilde{P}_T(v \mid c_t) = \begin{cases} \widetilde{p}_t, & v = y_t, \\ P_T(v \mid c_t)\dfrac{1 - \widetilde{p}_t}{1 - p_t}, & v \in \mathcal{V} \setminus \{y_t\}. \end{cases} \tag{12}$$

Thus, SASR assigns the realized probability $\widetilde{p}_t$ to the sampled token y_t and redistributes the remaining probability mass proportionally over all other vocabulary tokens. When the target value is clipped by the projection, the realized supervision change can be smaller than the target change δ_t in Eq. (10).

The construction in Eq. (12) is normalized and preserves the vanilla teacher decision within the feasible interval. Its protected-to-vanilla KL reduces to a Bernoulli KL, used to select a fixed λ under the average distortion budget b. Detailed derivations are provided in Appendix A.3.

4 Experiments

4.1 Experimental Setup

Models. We evaluate SASR under two OPD scenarios. 1) **Strong-to-Weak**, where Qwen2.5-3B-Instruct serves as the teacher model and Qwen2.5-0.5B-Instruct as the student model (Qwen Team et al., 2024), evaluating protection under a substantial teacher–student capacity gap. 2) **Single-Teacher**, where Qwen3-4B-Non-Thinking-RL-Math-Step500 serves as the teacher and Qwen3-4B-Non-Thinking as the student (Yang et al., 2025), evaluating protection when teacher model and student model have matched model capacity. These two settings cover both cross-capacity and matched-capacity OPD.

Datasets. For training, we use DeepMath-103K (He et al., 2025) filtered to difficulty level 6, following the data configuration adopted in recent OPD studies (Yang et al., 2026; Li et al., 2026). The same training data are used for the vanilla and protected variants within each setting.

Table 1: Strong-to-weak distillation results on seven mathematical reasoning benchmarks, measured by Avg@8. Bold values indicate the strongest protection (lowest student accuracy) within each OPD method.

Method	Variant	AIME24	AIME25	MATH	AMC	Olymp.	Miner.	HMMT Feb	Avg↓
Teacher	–	4.17	0.83	61.02	30.63	26.22	23.81	0.83	21.07
Initial student	–	0.00	0.42	22.45	10.00	4.77	2.99	0.42	5.86
OPD	Vanilla	0.83	0.42	29.10	12.50	7.12	5.51	0.00	7.93
	+CMI	0.00	0.42	28.52	11.56	7.16	5.88	0.42	7.71
	+Ours	0.00	**0.00**	**24.75**	**8.44**	**6.01**	**4.37**	**0.00**	**6.22**
ExOPD	Vanilla	1.25	0.42	29.12	13.12	7.73	5.42	0.00	8.15
	+CMI	0.42	0.00	27.93	10.62	6.95	5.10	**0.00**	7.29
	+Ours	**0.00**	**0.00**	**20.67**	**8.44**	**4.60**	**3.54**	0.42	**5.38**
FiRe-OPD	Vanilla	0.83	0.00	29.88	10.94	7.47	5.01	0.83	7.85
	+CMI	0.42	**0.00**	27.68	11.88	6.77	5.06	0.00	7.40
	+Ours	**0.00**	0.83	**24.68**	**10.31**	**5.21**	**3.95**	0.00	**6.43**

We evaluate mathematical reasoning on seven benchmarks spanning different levels of difficulty: AIME 2024, AIME 2025, AMC 2023, MATH-500 (MATH) (Hendrycks et al., 2021), OlympiadBench (Olymp.) (He et al., 2024), Minerva Math (Miner.) (Lewkowycz et al., 2022) and HMMT February 2025 (Balunovic et al., 2025).

Baselines. We consider three OPD methods: standard OPD, ExOPD (Yang et al., 2026), and FiReOPD (Li et al., 2026), together with CMI (Fang et al., 2026) as a distillation-defense baseline. For each applicable method, we compare the original training procedure with its SASR-protected counterpart. SASR is inserted only at the teacher interface; the student model, optimization objective, and downstream processing of each OPD method are otherwise left unchanged.

Metrics. Following the OPD evaluation protocol, we sample eight responses per problem with temperature 1.0 and top-$p = 1.0$, and report Avg@ 8 accuracy, defined as the average correctness over the eight sampled responses. For ablations, we report the average rollout reward as a training-time measure of student response quality, and gradient cosine similarity (C_{grad}) to measure the alignment between the vanilla and perturbed optimization directions (Yu et al., 2020).

Implementation Details. We use setting-specific training configurations while keeping all optimization and sampling conditions identical between vanilla OPD and its SASR-protected counterpart. In the Strong-to-Weak setting, training uses 40 optimizer updates, a global batch size of 256, a learning rate of 1×10^{-6}, and a maximum response length of 4096 tokens. In the matched-capacity setting, we use 40 optimizer updates, a global batch size of 1024, the same learning rate, and a maximum response length of 16384 tokens. During rollout, responses are sampled with temperature 1.0 and top-$p = 1.0$. Unless otherwise specified, SASR uses a selection rate of $\rho = 0.20$, an intervention strength of $\lambda = 0.75$, and asymmetric branch weights $(w_+, w_-) = (0.5, 1)$; λ is selected according to the prescribed average KL-distortion budget b and remains fixed throughout training.

4.2 Results

Strong-to-Weak Distillation. Table 1 reports Avg@8 results on seven mathematical reasoning benchmarks, where lower student accuracy indicates stronger protection. The results show that: 1) Our SASR consistently achieves the strongest overall suppression across all three OPD methods, reducing the average accuracy from 7.93% to 6.22% for OPD, from 8.15% to 5.38% for ExOPD, and from 7.85% to 6.43% for FiRe-OPD. 2) The largest reduction occurs with ExOPD, where the average drops by 2.77 percentage points, including reductions from 29.12% to 20.67% on MATH-500. 3)

Table 2: Single-teacher distillation results, measured by Avg@8. Bold values indicate the strongest protection. ↓ means lower is better.

Method	Variant	AIME24	AIME25	MATH	AMC	Olymp.	Miner.	HMMT Feb	Avg ↓
Teacher	–	57.50	56.25	95.53	93.44	68.81	47.61	27.50	63.81
Initial student	–	23.75	21.67	83.60	66.88	51.74	39.80	10.42	42.55
OPD	Vanilla	57.08	56.67	95.18	**93.75**	68.90	47.93	30.00	64.22
	Ours	**51.67**	**52.92**	**94.78**	94.38	**67.88**	**46.28**	30.00	**62.56**
ExOPD	Vanilla	60.00	53.33	95.05	93.75	66.93	47.89	33.33	64.33
	Ours	**45.42**	**37.08**	**92.28**	**84.38**	**59.38**	**45.91**	**23.33**	**55.40**
FiRe-OPD	Vanilla	57.92	54.17	95.23	95.00	68.94	48.25	34.17	64.81
	Ours	**52.08**	**52.92**	**94.60**	**92.50**	**67.69**	**46.14**	**29.58**	**62.22**

Compared with CMI, SASR produces a larger accuracy reduction across all three OPD variants. Fig. 3 further visualizes this gap. Overall, SASR consistently reduces the gains obtained from different OPD variants.

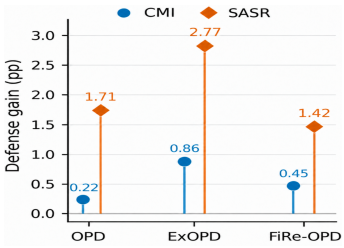

Figure 3: Defense Gain across OPD Variants.

Single-Teacher Distillation. As shown in Table 2, We evaluate SASR with several representative OPD methods, including standard OPD, ExOPD and FiRe-OPD. The results show that: 1) SASR consistently reduces Avg@8 across all three OPD methods, demonstrating that the protection effect persists in the matched-capacity setting. 2) The magnitude of suppression varies substantially across OPD methods. ExOPD shows the largest average reduction of 8.93 pp, whereas standard OPD and FiRe-OPD exhibit more moderate decreases. This variation suggests that different OPD objectives may have different sensitivities to supervision reshaping. 3) OPD and FiRe-OPD show more moderate but broadly consistent reductions. The only exception is AMC under standard OPD, where accuracy changes slightly from 93.75% to 94.38%, but its small magnitude does not alter the overall suppression trend.

Adaptive Recovery Attacks. We further evaluate three recovery strategies that attempt to reconstruct the original teacher supervision from SASR-protected feedback. *Full GFR* applies the inverse feedback transformation globally without identifying the protected positions; *SPFR* first estimates the sparsely modified positions and restores only the selected subset; and *KL-GFR* constrains the restoration strength with a KL budget to avoid overly aggressive inversion. Table 3 indicates that: 1) The three recovery strategies do not restore performance beyond passive SASR and all remain below vanilla OPD at 7.93%. 2) KL-GFR achieves the strongest recovery among the evaluated attacks, but still remains 0.51 pp below passive SASR, indicating that constraining the inverse transformation does not recover the original OPD supervision. 3) Full GFR causes the largest additional decrease (-0.96%), suggesting that globally inverting the protected feedback without locating the modified positions can further impair student learning.

Ablation Study. Table 4 analyzes the contribution of the main SASR components in the Strong-to-Weak setting using the fixed DeepMath dev set and Step-5 accuracy. The results show that: 1)Full SASR reduces accuracy (Acc) from 63.28% to 58.20% and achieves the lowest rollout reward (0.5805), indicating degradation not only at the final evaluation point but also throughout the short training trajectory. 2) The negative branch provides the strongest short-horizon suppression. Negative-only achieves the lowest accuracy of 55.86%, compared with 59.77% for Positive-only, suggesting that interventions on negative-gap positions play a major role in weakening distillation. 3) Structured reshaping is more important than simply maximizing gradient deviation. Random Selection is weaker than Full SASR (60.35% vs. 58.20%), supporting disagreement-based token selection. Meanwhile, gradient cosine similarity is not monotonic with accuracy: Full SASR retains

Table 3: Recovery attacks against SASR. Δ_{rec} denotes the change relative to passive SASR.

Method	Avg@8 (%)↑	Δ_{rec}↓
vanilla OPD	7.93	+1.71
SASR	**6.22**	**0.00**
+ Full GFR	5.26	-0.96
+ SPFR	5.68	-0.54
+ KL-GFR	5.71	-0.51

Table 4: Ablation study on component contributions (Strong-to-Weak setting).

Method	Acc↓	Reward	$\mathbf{C}_{\text{grad}}$
vanilla OPD	63.28%	0.6120	-
Full SASR	58.20%	**0.5805**	0.632
Positive-only	59.77%	0.5999	0.007
Negative-only	**55.86**%	0.5939	0.586
Random Selection	60.35%	0.6072	0.079

moderate alignment with clean OPD while still reducing both reward and accuracy, suggesting that effective suppression comes from structured supervision reshaping rather than arbitrary disruption of the optimization direction.

5 Related Work

On-Policy Distillation for LLMs. Knowledge distillation transfers teacher knowledge through soft probability targets or generated sequences (Hinton et al., 2015; Kim & Rush, 2016). For autoregressive language models, GKD reduces train–inference distribution mismatch by providing teacher feedback on student-generated sequences, while MiniLLM develops reverse-KL distillation for generative LLMs (Agarwal et al., 2024; Gu et al., 2023). Recent OPD methods further refine the supervision and optimization process. Standard OPD uses dense teacher feedback on student rollouts (Lu & Lab, 2025); ExOPD generalizes the objective through reward extrapolation (Yang et al., 2026); TSD-KD selectively applies teacher supervision according to token-level teacher–student confidence (Kim & Baek, 2026); AOPD applies different student-side optimization behaviors across token-level signal regions (Jia et al., 2026); and FiRe-OPD combines trajectory filtering with token-level reweighting (Li et al., 2026). Recent analyses also identify long-horizon instability and length inflation in OPD (Jang et al., 2026). In contrast to these approaches, SASR does not modify the downstream student optimization rule. It intervenes on the teacher supervision returned at student-generated contexts before the original OPD algorithm consumes that feedback.

Distillation Protection with LLMs. Prior work has explored limiting unauthorized knowledge transfer while retaining model utility. Prediction Poisoning perturbs prediction scores to disrupt model extraction (Orekondy et al., 2019), while Nasty Teacher and Safe Distillation Box modify teacher-side behavior to restrict unauthorized distillation (Ma et al., 2021; Ye et al., 2021). Recent LLM defenses increasingly focus on the information exposed through generation or prediction interfaces. Antidistillation Sampling modifies next-token sampling to produce less distillable reasoning traces (Savani et al., 2025), while Trace Rewriting and TraceGuard perturb generated reasoning traces while seeking to preserve their utility (Ma et al., 2026; Hartman et al., 2026). The Distillation Game studies the interaction between a utility-constrained teacher and adaptive students and derives a Product-of-Experts generation-time defense (Allouah et al., 2026). Most closely related to our setting, Fang et al. (2026) globally transform teacher logits to suppress distillation-relevant information. In contrast, SASR selectively reshapes student-conditioned OPD supervision while explicitly preserving the teacher decision.

6 Conclusion

We identify the OPD supervision interface as a distinct surface for model protection and casts its protection as a constrained adversarial optimization problem. Based on this formulation, we propose *Selective Asymmetric Supervision Reshaping (SASR)*, which reshapes teacher supervision within a prediction-preserving region to weaken student learning while retaining the teacher top-1 decision. Experiments across cross-capacity and matched-capacity settings demonstrate consistent suppression across different OPD variants, with ablations and recovery attacks further supporting the proposed design. These findings highlight supervision intervention as a promising direction for protecting teacher knowledge in on-policy distillation.

Ethics Statement

This work studies protection against unauthorized on-policy distillation and does not involve human subjects, private data, or personally identifiable information. The authors have reviewed and adhered to the ICLR Code of Ethics.

Reproducibility Statement

The main paper provides the experimental setup, model configurations, evaluation protocol, and implementation details needed to reproduce our results. Complete theoretical derivations are included in the appendix, while the supplementary materials provide the implementation code, configuration files, experiment scripts, and a concise PDF/webpage summary of the method and experimental results.

AI Use Statement

In this work, we used generative AI tools to assist with the interpretation and presentation of experimental results. Additionally, generative AI tools were used to improve the readability of the manuscript, identify and search for relevant literature, suggest the structure and organization of the paper, and provide suggestions for scientific figures and visual presentation. All AI-assisted content, including experimental interpretations, literature references, and manuscript revisions, was reviewed and verified by the authors. We take full responsibility for the final content of this work, including all text, claims, results, and artifacts produced with the aid of generative AI.

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, 2024. URL https://arxiv.org/abs/2306.13649.

Youssef Allouah, Mahdi Haghifam, Sanmi Koyejo, and Reza Shokri. The distillation game: Adaptive attacks & efficient defenses, 2026. URL https://arxiv.org/abs/2605.22737.

M. Balunovic, J. Dekoninck, I. Petrov, et al. MathArena: Evaluating LLMs on uncontaminated math competitions. *Conference on Neural Information Processing Systems*, 38, 2025.

Hao Fang, Tianyi Zhang, Tianqu Zhuang, Jiawei Kong, Kuofeng Gao, Bin Chen, Leqi Zheng, Shu-Tao Xia, and Ke Xu. Towards distillation-resistant large language models: An information-theoretic perspective, 2026. URL https://arxiv.org/abs/2602.03396.

Yuxian Gu, Li Dong, Furu Wei, and Minlie Huang. Minillm: On-policy distillation of large language models, 2023. URL https://arxiv.org/abs/2306.08543.

Max Hartman, Vidhata Jayaraman, Moulik Choraria, Yash Savani, and Lav R. Varshney. Hiding in plain sight: Detectability-aware antidistillation of reasoning models, 2026. URL https://arxiv.org/abs/2604.23238.

Chaoqun He, Renjie Luo, Yuzhuo Bai, Shengding Hu, Zhen Leng Thai, Junhao Shen, Jinyi Hu, Xu Han, Yujie Huang, Yuxiang Zhang, Jie Liu, Lei Qi, Zhiyuan Liu, and Maosong Sun. Olympiadbench: A challenging benchmark for promoting agi with olympiad-level bilingual multimodal scientific problems, 2024. URL https://arxiv.org/abs/2402.14008.

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. URL https://arxiv.org/abs/2504.11456.

Dan Hendrycks, Collin Burns, Saurav Kadavath, Akul Arora, Steven Basart, Eric Tang, Dawn Song, and Jacob Steinhardt. Measuring mathematical problem solving with the math dataset, 2021. URL https://arxiv.org/abs/2103.03874.

Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. Distilling the knowledge in a neural network, 2015. URL https://arxiv.org/abs/1503.02531.

Ijun Jang, Jewon Yeom, Juan Yeo, Hyunggyu Lim, and Taesup Kim. Stable on-policy distillation through adaptive target reformulation, 2026. URL https://arxiv.org/abs/2601.07155.

N. Jia, H. Yang, X. Ma, et al. Asymmetric On-Policy Distillation: Bridging Exploitation and Imitation at the Token Level. *arXiv preprint arXiv:2605.06387*, 2026.

M. Kim and S. Baek. Explain in your own words: Improving reasoning via token-selective dual knowledge distillation. In *International Conference on Learning Representations (ICLR)*, 2026. URL https://doi.org/10.48550/arXiv.2603.13260.

Yoon Kim and Alexander M. Rush. Sequence-level knowledge distillation, 2016. URL https://arxiv.org/abs/1606.07947.

Aitor Lewkowycz, Anders Andreassen, David Dohan, Ethan Dyer, Henryk Michalewski, Vinay Ramasesh, Ambrose Slone, Cem Anil, Imanol Schlag, Theo Gutman-Solo, Yuhuai Wu, Behnam Neyshabur, Guy Gur-Ari, and Vedant Misra. Solving quantitative reasoning problems with language models, 2022. URL https://arxiv.org/abs/2206.14858.

Yuying Li, Leqi Zheng, Yongzi Yu, Wenrui Zhou, Xuchang Zhong, Xing Hu, Jing Jin, Hangjie Yuan, and Tao Feng. Filter, then reweight: Rethinking optimization granularity in on-policy distillation, 2026. URL https://arxiv.org/abs/2606.02684.

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

Haoyu Ma, Tianlong Chen, Ting-Kuei Hu, Chenyu You, Xiaohui Xie, and Zhangyang Wang. Undistillable: Making a nasty teacher that cannot teach students, 2021. URL https://arxiv.org/abs/2105.07381.

Xinhang Ma, William Yeoh, Ning Zhang, and Yevgeniy Vorobeychik. Protecting language models against unauthorized distillation through trace rewriting. In Maria Liakata, Viviane P. Moreira, Jiajun Zhang, and David Jurgens (eds.), *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 11307–11324, San Diego, California, United States, July 2026. Association for Computational Linguistics. ISBN 979-8-89176-390-6. doi: 10.18653/v1/2026.acl-long.519. URL https://aclanthology.org/2026.acl-long.519/.

Tribhuvanesh Orekondy, Bernt Schiele, and Mario Fritz. Prediction Poisoning: Towards Defenses Against DNN Model Stealing Attacks. arXiv preprint arXiv:1906.10908, 2019. URL https://arxiv.org/abs/1906.10908.

Qwen Team, An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jingren Zhou, Junyang Lin, Kai Dang, Keming Lu, Keqin Bao, Kexin Yang, Le Yu, Mei Li, Mingfeng Xue, Pei Zhang, Qin Zhu, Rui Men, Runji Lin, Tianhao Li, Tianyi Tang, Tingyu Xia, Xingzhang Ren, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yu Wan, Yuqiong Liu, Zeyu Cui, Zhenru Zhang, and Zihan Qiu. Qwen2.5 Technical Report. arXiv preprint arXiv:2412.15115, 2024. URL https://arxiv.org/abs/2412.15115.

Yash Savani, Asher Trockman, Zhili Feng, Yixuan Even Xu, Avi Schwarzschild, Alexander Robey, Marc Finzi, and J. Zico Kolter. Antidistillation Sampling. arXiv preprint arXiv:2504.13146, 2025. URL https://arxiv.org/abs/2504.13146.

Bing Shao, Jiazheng Zhang, Long Ma, Yujiong Shen, Senjie Jin, Xin Guo, Yuming Yang, Mingxu Chai, Zhiheng Xi, Boyang Liu, et al. A token-level analysis of sampled-token reverse-kl on-policy distillation. *arXiv preprint arXiv:2608.25643*, 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, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jing Zhou, Jingren Zhou, Junyang Lin, Kai Dang, Keqin Bao, Kexin Yang, Le Yu, Lianghao Deng, Mei Li, Mingfeng Xue, Mingze Li, Pei Zhang, Peng Wang, Qin Zhu, Rui Men, Ruize Gao, Shixuan Liu, Shuang Luo, Tianhao Li, Tianyi Tang, Wenbiao Yin, Xingzhang Ren, Xinyu Wang, Xinyu Zhang, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yinger Zhang, Yu Wan, Yuqiong Liu, Zekun Wang, Zeyu Cui, Zhenru Zhang, Zhipeng Zhou, and Zihan Qiu. Qwen3 Technical Report. arXiv preprint arXiv:2505.09388, 2025. URL https://arxiv.org/abs/2505.09388.

Wenkai Yang, Weijie Liu, Ruobing Xie, Kai Yang, Saiyong Yang, and Yankai Lin. Learning beyond teacher: Generalized on-policy distillation with reward extrapolation, 2026. URL https://arxiv.org/abs/2602.12125.

Jingwen Ye, Yining Mao, Jie Song, Xinchao Wang, Cheng Jin, and Mingli Song. Safe Distillation Box. arXiv preprint arXiv:2112.03695, 2021. URL https://arxiv.org/abs/2112.03695.

Tianhe Yu, Saurabh Kumar, Abhishek Gupta, Sergey Levine, Karol Hausman, and Chelsea Finn. Gradient surgery for multi-task learning. In *Advances in Neural Information Processing Systems*, volume 33, pp. 5824–5836, 2020.

The appendix is structured as follows:

Section A. Proofs and Technical Derivations

- Section A.1: Fixed-Context Reverse-KL Gradient
- Section A.2: Proof of Proposition 1
- Section A.3: Fidelity-Constrained Realization

Section B. Additional Experimental Results

A Proofs and Technical Derivations

A.1 Fixed-Context Reverse-KL Gradient

We derive the signed token-level coefficient used in Eq. (8). For a fixed context c, the reverse-KL objective is

$$D_{\mathrm{KL}}\left(P_{S,\theta}(\cdot \mid c) \| P_T(\cdot \mid c)\right) = \sum_{v \in \mathcal{V}} P_{S,\theta}(v \mid c) \log \frac{P_{S,\theta}(v \mid c)}{P_T(v \mid c)}. \tag{13}$$

Treating the teacher distribution as fixed with respect to the student parameters and using $\sum_v \nabla_\theta P_{S,\theta}(v \mid c) = 0$ gives

$$\nabla_\theta D_{\mathrm{KL}} = -\mathbb{E}_{v \sim P_{S,\theta}(\cdot|c)} \left[\Delta(v, c) \nabla_\theta \log P_{S,\theta}(v \mid c)\right], \tag{14}$$

where

$$\Delta(v, c) = \log P_T(v \mid c) - \log P_{S,\theta}(v \mid c). \tag{15}$$

For a sampled token y_t at the student-generated context $c_t = (x, y_{<t})$, this gives

$$\Delta_t = \log P_T(y_t \mid c_t) - \log P_{S,\theta}(y_t \mid c_t), \tag{16}$$

with sampled update contribution

$$-\Delta_t \nabla_\theta \log P_{S,\theta}(y_t \mid c_t). \tag{17}$$

The derivation is conditional on a realized context and therefore characterizes the local reverse-KL update without differentiating through the student-dependent rollout distribution.

A.2 Proof of Proposition 1

Consider a binary vocabulary $\{0, 1\}$ and let

$$s = P_S(1) \in (1/2, 1), \qquad p = P_T(1) \in (0, 1), \tag{18}$$

with student logit

$$z = \log \frac{s}{1 - s}. \tag{19}$$

The reverse-KL divergence is

$$D_{\mathrm{KL}}\left((s, 1-s) \| (p, 1-p)\right) = s \log \frac{s}{p} + (1 - s) \log \frac{1 - s}{1 - p}. \tag{20}$$

Differentiating with respect to z yields

$$\frac{\partial D_{\mathrm{KL}}}{\partial z} = s(1 - s) \log \frac{s(1 - p)}{p(1 - s)}. \tag{21}$$

Since $s(1 - s) > 0$ and the log-odds function is strictly increasing,

$$p > s \Rightarrow \frac{\partial D_{\mathrm{KL}}}{\partial z} < 0, \qquad p < s \Rightarrow \frac{\partial D_{\mathrm{KL}}}{\partial z} > 0. \tag{22}$$

For any $s \in (1/2, 1)$, choose

$$0 < \varepsilon < \min\left\{ s - \frac{1}{2},\, 1 - s \right\}, \tag{23}$$

and define

$$p_+ = s + \varepsilon, \qquad p_- = s - \varepsilon. \tag{24}$$

Both teacher distributions have token 1 as their unique top-1 prediction, yet $p_+ > s$ and $p_- < s$, so Eq. (21) has opposite signs under the two teachers.

Therefore, teacher distributions with the same top-1 prediction can induce opposite local reverse-KL update directions. As $\varepsilon \to 0$, both distributions converge to the student distribution and the corresponding KL divergence vanishes. The result is existential: it shows that the teacher's top-1 prediction does not uniquely determine the local learning signal.

A.3 Fidelity-Constrained Realization

Define the strongest competing probability as $r_t = \max_{v \neq y_t} P_T(v \mid c_t)$. If y_t is the clean teacher's top-1 token, preserving it with margin $\eta > 0$ requires

$$\widetilde{p}_t \geq \frac{r_t + \eta(1 - p_t)}{1 - p_t + r_t}. \tag{25}$$

If y_t is not the clean top-1 token, preserving the clean winner requires

$$\widetilde{p}_t \leq \frac{r_t - \eta(1 - p_t)}{1 - p_t + r_t}. \tag{26}$$

SASR therefore projects the desired probability $q_t = p_t \exp(\delta_t)$ onto the applicable feasible interval defined by Eqs. (25) and (26). In implementation, the interval is additionally clipped to $[\epsilon, 1 - \epsilon]$ for numerical stability, and infeasible cases revert to the clean teacher probability.

Finally, proportional redistribution gives

$$D_{\mathrm{KL}}\left(\widetilde{P}_T(\cdot \mid c_t) \| P_T(\cdot \mid c_t)\right) = \widetilde{p}_t \log \frac{\widetilde{p}_t}{p_t} + (1 - \widetilde{p}_t) \log \frac{1 - \widetilde{p}_t}{1 - p_t}.$$

Thus, the full-vocabulary fidelity cost reduces exactly to the Bernoulli KL between the probability assigned to y_t and its complement.

B Additional Experimental Results

Evaluation Metric. We evaluate on seven mathematical reasoning benchmarks: AIME24, AIME25, AMC23, HMMT February, MATH-500, Minerva Math, and OlympiadBench, containing 30, 30, 40, 30, 500, 272, and 674 problems, respectively. Each problem is sampled eight times.

For dataset d with n_d problems, let n_d^{corr} denote the number of correct responses among all $8n_d$ generations. We compute

$$\mathrm{Avg@8}_d = 100 \times \frac{n_d^{\mathrm{corr}}}{8n_d}. \tag{27}$$

The overall Avg. is the arithmetic mean of the seven benchmark scores:

$$\mathrm{Avg.} = \frac{1}{7} \sum_{d=1}^{7} \mathrm{Avg@8}_d. \tag{28}$$

For the ablation study, we additionally report gradient cosine similarity, denoted by C_{grad}, to measure the alignment between the vanilla OPD update and the corresponding protected update:

$$C_{\mathrm{grad}} = \frac{\mathbf{g}_{\mathrm{van}}^\top \mathbf{g}_{\mathrm{prot}}}{\|\mathbf{g}_{\mathrm{van}}\|_2 \|\mathbf{g}_{\mathrm{prot}}\|_2}, \tag{29}$$

where $\mathbf{g}_{\mathrm{van}}$ and $\mathbf{g}_{\mathrm{prot}}$ denote the vanilla and protected optimization gradients, respectively. A larger C_{grad} indicates greater directional similarity to vanilla OPD.

Table 5: Detailed results of adaptive recovery attacks in the Strong-to-Weak setting. All values are Avg@8 accuracies (%).

Method	Variant	AIME24	AIME25	MATH	AMC	Olymp.	Miner.	HMMT Feb	Avg.
OPD	Vanilla	0.83	0.42	29.10	12.50	7.12	5.51	0.00	7.93
SASR	Passive	0.00	0.00	24.75	8.44	6.01	4.37	0.00	6.22
	+ Full GFR	**0.42**	0.00	22.08	5.63	4.86	2.99	**0.83**	5.26
	+ SPFR	0.00	**0.42**	22.33	**9.38**	4.75	2.48	0.42	5.68
	+ KL-GFR	0.00	0.00	**22.53**	7.50	**5.14**	**3.95**	**0.83**	**5.71**

Detailed Results of Recovery Attacks We provide benchmark-level results for the adaptive recovery attacks summarized in the main text. Full GFR applies recovery globally to the protected feedback, SPFR restricts recovery to positions estimated to have been modified, and KL-GFR constrains the global recovery magnitude using a KL budget.

The results show that: 1) The three recovery strategies do not restore performance beyond passive SASR, and their overall Avg. values remain below vanilla OPD. 2) KL-GFR achieves the strongest recovery among the evaluated attacks with an Avg. of 5.71%, but remains below passive SASR at 6.22%. These results show that the evaluated recovery strategies do not offset the protection provided by SASR.