# REWARD-ALIGNED REWEIGHTING FOR ON-POLICY DISTILLATION

**Anonymous authors**
Paper under double-blind review

## ABSTRACT

On-policy distillation (OPD) has become a standard stage of large language model post-training: a student learns from a stronger teacher on trajectories generated by its own policy. Yet on-policy feedback does not make every teacher correction equally useful. Standard OPD weights token-level distillation terms uniformly, implicitly treating local teacher preference as a proxy for correction utility. A decision's task value, however, depends on how the student completes the subsequent reasoning. This mismatch can cause imitation to suppress viable student strategies or reinforce paths the student cannot reliably execute. Verified trajectory outcomes provide complementary evidence about continuation quality, but do not directly identify the utility of individual decisions. This motivates using outcome evidence to guide the relative strength of teacher corrections. We introduce Reward-Aligned Reweighting for On-Policy Distillation ($R^2$-OPD), which continuously reallocates teacher supervision using outcome agreement and the magnitude of teacher–student disagreement. It gives reward-aligned corrections greater relative influence while retaining dense feedback, moving beyond uniform imitation and hard filtering. Our analysis formalizes the mismatch between local teacher preference and student continuation value, and establishes sufficient conditions for reallocation to improve first-order task progress over uniform OPD. Across seven mathematical reasoning benchmarks, $R^2$-OPD achieves the highest average accuracy among the compared training methods in both cross-size and same-size distillation. It outperforms standard OPD on all seven benchmarks, with average gains of $+3.5$ and $+2.4$ for 1.7B and 4B students, respectively. An extension to code generation yields an average gain of $+1.6$ over standard OPD. Together, these results highlight outcome-guided supervision allocation as an effective means of translating dense teacher feedback into stronger student performance across model scales and task domains. Code is available at `https://anonymous.4open.science/r/r2-opd-62F4/`.

## 1 INTRODUCTION

On-policy distillation (OPD) trains a student language model with dense teacher feedback on trajectories generated by the student itself. By querying the teacher at student-visited prefixes, it reduces the mismatch between training contexts and inference-time behavior (Agarwal et al., 2024; Gu et al., 2024; Lu & Thinking Machines Lab, 2025). However, aligning the contexts in which supervision is provided does not ensure that every teacher correction helps the student complete the task. Effective distillation therefore requires considering not only *where* teacher feedback is obtained, but also *how strongly* each correction should influence learning.

Standard sampled-token OPD uses teacher–student log-probability gaps as token-level correction coefficients, without outcome-dependent reweighting. These coefficients encode *local teacher preference*, whereas task success depends on *student continuation value*: the expected reward after a decision when the student, rather than the teacher, generates the remaining response. This distinction gives rise to two potential failure modes. **Over-correction on successful trajectories:** suppressing teacher-disfavored decisions can erode valid student strategies that differ from the teacher's (Yuan et al., 2023). **Over-trust on failed trajectories:** reinforcing teacher-preferred decisions can encourage reasoning paths that the student cannot reliably complete. In Section 3, we formalize this preference–value mismatch: moving a local next-token distribution toward the teacher can reduce



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.

reverse KL while decreasing expected task reward when the student's continuation policy is held fixed. Closer agreement with the teacher is therefore not, by itself, sufficient evidence of useful supervision.

Verified trajectory outcomes provide complementary evidence about continuation quality, but do not identify the utility of individual decisions. Existing approaches use outcome feedback to combine learning objectives or select teacher supervision (Li et al., 2026a; Xu et al., 2026a; Lin et al., 2026). However, deciding whether to retain a correction does not resolve how strongly it should influence learning, and hard filtering discards feedback whose utility remains uncertain. This motivates a different use of outcomes: *guiding correction strength rather than treating outcome agreement as a definitive token-value label*.

We introduce Reward-Aligned Reweighting for On-Policy Distillation (**R$^2$-OPD**), a continuous reweighting method that combines outcome agreement with the magnitude of teacher–student disagreement. It gives reward-aligned corrections greater relative influence while retaining feedback that disagrees with the outcome. The reweighting preserves correction signs and each response's total absolute coefficient mass, reallocating rather than simply reducing teacher supervision. Our analysis establishes sufficient reward–utility agreement conditions under which this reallocation improves first-order task progress relative to uniform OPD. R$^2$-OPD uses outcomes to guide imitation without requiring a separate reinforcement-learning objective or a learned value estimator.

We evaluate R$^2$-OPD in same-size and cross-size distillation settings studied in prior OPD work (Lin et al., 2026; Yang et al., 2026b). Across seven mathematical reasoning benchmarks, R$^2$-OPD achieves the highest average accuracy among the compared methods in both settings, with gains of $+3.5$ and $+2.4$ over standard OPD for 1.7B and 4B students, respectively. An extension to code generation yields an average gain of $+1.6$ over standard OPD. Learning curves show early performance gains, while ablations support the joint use of outcome alignment, disagreement magnitude, and coefficient-mass normalization.

**Contributions.** **(1) Preference–value analysis.** We formalize the mismatch between local teacher preference and student continuation value, and establish sufficient conditions for supervision reallocation to improve first-order task progress over standard OPD. **(2) Reward-aligned supervision allocation.** We introduce R$^2$-OPD, a continuous, magnitude-sensitive reweighting method that uses verified outcomes while preserving nonzero teacher feedback, correction signs, and per-response absolute coefficient mass. **(3) Empirical validation across tasks and scales.** Experiments on seven mathematical reasoning benchmarks and two code-generation benchmarks demonstrate gains over standard OPD, with ablations examining the roles of the key reweighting components.

## 2 Preliminaries

### 2.1 Notation and Distillation Objective

Let $q \sim \mathcal{D}$ denote a prompt and $o = (o_1, \ldots, o_{|o|})$ a response, with $s_t = (q, o_{<t})$ and $\pi(o \mid q) = \prod_{t=1}^{|o|} \pi(o_t \mid s_t)$. We consider a trainable student $\pi_\theta$ and a frozen teacher $\pi_T$ over a shared finite vocabulary $\mathcal{V}$. At each iteration, rollouts are generated by a fixed student snapshot $\pi_{\bar{\theta}}$. For the identities below, we assume a differentiable student, positive teacher and student probabilities, and a bounded generation horizon. Following reverse-KL distillation (Gu et al., 2024), we define

$$\begin{aligned} K_t(\theta) &= D_{\mathrm{KL}}(\pi_\theta(\cdot \mid s_t) \,\|\, \pi_T(\cdot \mid s_t)) \\ &= \sum_{v \in \mathcal{V}} \pi_\theta(v \mid s_t) \log \frac{\pi_\theta(v \mid s_t)}{\pi_T(v \mid s_t)}. \end{aligned} \tag{1}$$

### 2.2 On-Policy Distillation

OPD supervises the student on prefixes visited by its own rollouts (Agarwal et al., 2024). Given $o \sim \pi_{\bar{\theta}}(\cdot \mid q)$, the statewise reverse-KL objective is

$$\mathcal{L}_{\mathrm{OPD}}(\theta; \bar{\theta}) = \mathbb{E}_{\substack{q \sim \mathcal{D} \\ o \sim \pi_{\bar{\theta}}(\cdot \mid q)}} \left[ \sum_{t=1}^{|o|} K_t(\theta) \right]. \tag{2}$$

The rollout distribution and sampled prefixes are held fixed when differentiating with respect to $\theta$. Evaluating the full-vocabulary KL requires a vocabulary-wide comparison at every visited prefix, which can be expensive for large vocabularies and long responses. We therefore focus on sampled-token OPD, which uses teacher log-probabilities at student-sampled tokens (Lu & Thinking Machines Lab, 2025). For $o_t \sim \pi_{\bar{\theta}}(\cdot \mid s_t)$, define the *distillation advantage*

$$A_t = \log \frac{\pi_T(o_t \mid s_t)}{\pi_{\bar{\theta}}(o_t \mid s_t)}. \tag{3}$$

Conditional on $s_t$, this gives $\mathbb{E}[-A_t \mid s_t] = K_t(\bar{\theta})$, so $-A_t$ is an unbiased single-sample estimate of the local reverse KL at the rollout parameters. For analysis, we use the detached log-probability surrogate

$$\mathcal{L}_{\mathrm{OPD}}^{\mathrm{sample}}(\theta; \bar{\theta}) = -\mathbb{E}_{\substack{q \sim \mathcal{D} \\ o \sim \pi_{\bar{\theta}}(\cdot|q)}} \left[ \sum_{t=1}^{|o|} \mathrm{sg}[A_t] \log \pi_\theta(o_t \mid s_t) \right], \tag{4}$$

where $\mathrm{sg}[\cdot]$ denotes the stop-gradient operator. At $\theta = \bar{\theta}$, its expected gradient matches that of the fixed-rollout reverse-KL objective:

$$\nabla_\theta \mathcal{L}_{\mathrm{OPD}}^{\mathrm{sample}}(\theta; \bar{\theta})\Big|_{\theta=\bar{\theta}} = \nabla_\theta \mathcal{L}_{\mathrm{OPD}}(\theta; \bar{\theta})\big|_{\theta=\bar{\theta}}. \tag{5}$$

This is a first-order identity at the rollout parameters, not an equality of loss values or a derivative through the state-visitation distribution. In an individual surrogate term, positive and negative $A_t$ contribute to reinforcing and suppressing the sampled token, respectively, while $|A_t|$ measures the sampled-token log-probability discrepancy. Standard sampled-token OPD uses these coefficients without additional outcome-dependent reweighting: they encode local teacher preference rather than verified trajectory outcomes.

### 2.3 Reinforcement Learning with Verifiable Rewards

We consider tasks with a binary verifier $R(q, o) \in \{0, 1\}$. Reinforcement learning with verifiable rewards (RLVR) seeks to maximize the expected task reward

$$J(\theta) = \mathbb{E}_{\substack{q \sim \mathcal{D} \\ o \sim \pi_\theta(\cdot|q)}} [R(q, o)]. \tag{6}$$

Reinforcement learning with verifiable rewards provides a trajectory-level signal indicating whether a sampled response satisfies the task verifier. We encode this outcome as

$$z(q, o) = 2R(q, o) - 1 \in \{-1, +1\}, \tag{7}$$

where $+1$ denotes success and $-1$ denotes failure. The outcome is shared by all tokens in the response and does not identify the utility of any individual decision. This differs from the token-specific distillation signal $A_t$, which measures the teacher's preference for the sampled token at position $t$.

## 3 Teacher Preference versus Student Continuation Value

The over-correction and over-trust described in the Introduction raise a question: when does local teacher imitation improve the student's expected task reward? We isolate this question through a local distributional intervention using the distillation advantage $A_t$ defined in Eq. (3).

**Local imitation versus student value.** Fix a visited prefix $s = s_t$. The task value of the next action depends on how the student completes the response:

$$Q_R^{\bar{\theta}}(s, a) = \mathbb{E}_{o_{>t} \sim \pi_{\bar{\theta}}(\cdot|s,a)} [R(q, o_{<t} a o_{>t})]. \tag{8}$$

The continuation follows the rollout student, not the teacher. To isolate the effect of local imitation, we interpolate only the next-action distribution toward the teacher:

$$\pi_\epsilon(a \mid s) \propto \pi_{\bar{\theta}}(a \mid s)^{1-\epsilon} \pi_T(a \mid s)^\epsilon, \qquad 0 \le \epsilon \le 1. \tag{9}$$

The distribution is normalized over actions, while all subsequent decisions remain governed by $\pi_{\bar{\theta}}$. The continuation values $Q_R^{\bar{\theta}}(s, a)$ remain fixed along the interpolation. Let $D_s(\epsilon)$ denote the reverse KL in Eq. (1) evaluated at $\pi_\epsilon(\cdot \mid s)$, and let $F_s(\epsilon) = \mathbb{E}_{a \sim \pi_\epsilon(\cdot|s)}[Q_R^{\bar{\theta}}(s, a)]$ denote the corresponding conditional expected reward. Moments with subscript $s$ are taken over $o_t \sim \pi_{\bar{\theta}}(\cdot \mid s)$.

**Proposition 1** (Local imitation need not improve student value)**.** *For a finite action set with positive teacher and student probabilities, the interpolation in Eq. (9) satisfies*

$$D_s'(0) = -\operatorname{Var}_s[A_t], \tag{10}$$

$$F_s'(0) = \operatorname{Cov}_s\Big(A_t, Q_R^{\bar{\theta}}(s, o_t)\Big). \tag{11}$$

*If the covariance is negative, every sufficiently small positive $\epsilon$ strictly decreases both the local reverse KL and the student's conditional expected reward.*

The proof is given in Appendix G. The proposition separates distributional agreement from task utility: local reverse KL decreases to first order whenever $A_t$ is nonconstant, whereas the reward change depends on how teacher preference covaries with student continuation value. To relate this distinction to the two failure modes, write

$$F_s'(0) = \mathbb{E}_s\big[(A_t - \mathbb{E}_s[A_t])(Q_R^{\bar{\theta}}(s, o_t) - F_s(0))\big]. \tag{12}$$

Under this interpolation, the centered preference $A_t - \mathbb{E}_s[A_t]$ determines the direction of first-order probability change. The value-based forms of over-correction and over-trust are therefore shifts away from actions with above-average student continuation value and toward those with below-average value, respectively, with $F_s(0)$ as the reference. Both contribute negatively to Eq. (12), although contributions from other actions can offset them. This is a local probability-space analysis, not a characterization of an arbitrary parameter-space update.

**Implications for supervision allocation.** Trajectory outcomes provide evidence about student continuation value: under student-generated continuations, $\mathbb{E}[z(q, o) \mid s_t = s, o_t = a] = 2Q_R^{\bar{\theta}}(s, a) - 1$. Although disagreement $z(q, o)A_t < 0$ does not establish that an individual correction is harmful, outcome feedback complements local teacher preference with evidence about the student's ability to complete the task. This complementarity motivates R$^2$-OPD, which uses reward alignment to guide the relative strength of teacher corrections.

## 4 Reward-Aligned Reweighting

To realize this allocation principle, we combine continuous token-level gating with per-response normalization. The gate adjusts correction strength using reward alignment and disagreement magnitude, while normalization preserves the original absolute coefficient mass within each response.

### 4.1 Reward-Aligned Distillation Advantages

For a complete response $o$, let $A_t$ and $z(q, o)$ denote the distillation advantage and signed outcome defined in Section 2. A correction is *reward-aligned* if $z(q, o)A_t > 0$ and *reward-inconsistent* if $z(q, o)A_t < 0$. We combine alignment and disagreement magnitude through a response-local scale and a sigmoid gate:

$$\nu_o = \max\big\{\operatorname{median}_{1 \le j \le |o|} |A_j|, \epsilon\big\}, \qquad \epsilon > 0, \tag{13}$$

$$g_t = \sigma\left(\beta \frac{z(q, o)A_t}{\nu_o}\right), \qquad \beta \ge 0, \tag{14}$$

where $\sigma(u) = (1 + e^{-u})^{-1}$. Scaling by $\nu_o$ expresses disagreement relative to its typical within-response magnitude, while $\beta$ controls the sharpness of the weighting. For $\beta > 0$, aligned corrections receive $g_t > 1/2$ and inconsistent corrections receive $g_t < 1/2$. When $\beta|A_t|/\nu_o$ is small, the gate remains close to $1/2$, producing a gradual adjustment. For finite $\beta$, the gate remains strictly positive.

Since gating alone reduces absolute coefficient mass, we normalize each response to preserve $M_o = \sum_t |A_t|$. For $M_o > 0$, the *reweighted distillation advantage* is:

$$Z_o = \frac{M_o}{\sum_j g_j |A_j|}, \qquad A_t^{\mathrm{R2}} = Z_o g_t A_t. \tag{15}$$

The common factor $Z_o$ restores the original coefficient budget without changing the relative gate multipliers across positions.

**Algorithm 1** One step of R$^2$-OPD.

**Require:** Student $\pi_\theta$, teacher $\pi_T$, prompts $\mathcal{D}$, verifier $R$; sharpness $\beta \geq 0$, scale floor $\epsilon > 0$.

```
Freeze θ̄ ← θ; sample {q_i}_{i=1}^B ~ D.
Sample o_i ~ π_θ̄(· | q_i); collect B = {(q_i, o_i)}_{i=1}^B.
for each (q, o) ∈ B with |o| > 0 do
    Query π_T to compute A_t; obtain z(q, o) from R.        ▷ token-wise teacher feedback
    A_t^R2 ← A_t; M_o ← Σ_t |A_t|.                          ▷ vanilla fallback
    if o is complete and M_o > 0 then
        Compute ν_o; g_t ← σ(β z(q, o) A_t / ν_o).          ▷ soft alignment
        Z_o ← M_o / (Σ_j g_j |A_j|); A_t^R2 ← Z_o g_t A_t.  ▷ mass-preserving reweighting
Update θ via the OPD surrogate with sg[A_t^R2].
```

### 4.2 Training Objective

Substituting $A_t^{\mathrm{R2}}$ into Eq. (4) gives the token-summed analysis surrogate

$$\mathcal{L}_{\mathrm{R2}}(\theta; \bar{\theta}) = -\mathbb{E}\left[\sum_t \mathrm{sg}[A_t^{\mathrm{R2}}] \log \pi_\theta(o_t \mid s_t)\right]. \tag{16}$$

Advantages and reweighting factors are detached; gradients flow only through $\log \pi_\theta(o_t \mid s_t)$. Reward modulates existing token-level contributions without an additional policy-gradient term. Algorithm 1 summarizes reweighting; the implemented loss, response-wise reduction, and gradient correspondence are detailed in Appendix A.2.

### 4.3 Structural Properties

The multiplicative reweighting admits the following exact decomposition into a teacher-preference component and an outcome-dependent modulation.

**Proposition 2** (Exact teacher–outcome decomposition)**.** *For a complete response with finite advantages, $M_o > 0$, and finite $\beta \geq 0$,*

$$A_t^{\mathrm{R2}} = \underbrace{\frac{Z_o}{2} A_t}_{\substack{\textit{Term A} \\ \textit{teacher preference}}} + \underbrace{\frac{Z_o}{2} z(q,o) |A_t| \tanh\left(\frac{\beta |A_t|}{2\nu_o}\right)}_{\substack{\textit{Term B} \\ \textit{reward-aligned modulation}}}. \tag{17}$$

Term A preserves the teacher's signed correction under a common response-level scale. Term B is nonnegative on successful responses and nonpositive on failed responses. Relative to Term A, it attenuates suppression on successful responses and reinforcement on failed responses, while strengthening reward-aligned corrections. This provides a coefficient-level interpretation of how the reweighting addresses over-correction and over-trust. The $\tanh$ factor controls the modulation strength through the normalized disagreement.

For nonzero $A_t$ and finite $\beta$, Term B has strictly smaller magnitude than Term A, so their sum preserves the original coefficient sign. The two terms form a decomposition of a single distillation coefficient. Term B is an outcome-dependent component of this coefficient, rather than a separate or generally unbiased task-advantage estimator. Appendix H.1 provides the proof.

We quantify supervision allocation by coefficient mass. For $M_o > 0$, define the *reward-inconsistent mass share*

$$\kappa_o(\beta) = \frac{\sum_{t: z(q,o) A_t < 0} |A_t^{\mathrm{R2}}|}{\sum_t |A_t^{\mathrm{R2}}|}. \tag{18}$$

**Proposition 3** (Mass-preserving reward-aligned allocation)**.** *Fix a complete response with finite advantages and $M_o > 0$. For every finite $\beta \geq 0$, reweighting retains all nonzero corrections and their signs, with*

$$\sum_t |A_t^{\mathrm{R2}}| = M_o, \qquad \kappa_o'(\beta) < 0 \text{ if } 0 < \kappa_o(0) < 1. \tag{19}$$

*The derivative at zero is right-sided. At $\beta = 0$, $A_t^{\mathrm{R2}} = A_t$. If aligned mass is positive, the limit $\beta \to \infty$ is hard reward-aligned filtering renormalized to mass $M_o$.*

When both groups carry mass, increasing $\beta$ monotonically shifts a fixed coefficient budget toward reward-aligned feedback. This does not imply monotonic changes for individual token coefficients. As $\beta \to \infty$, the update converges to mass-normalized hard reward-aligned gating. Preserving coefficient mass does not preserve the parameter-gradient norm, since token score vectors may reinforce or cancel one another. The proof and boundary cases are given in Appendix H.2.

### 4.4 When Reallocation Improves Task Progress

The preceding results describe supervision allocation, but not its task utility. We therefore connect the allocation to first-order progress on the expected reward $J(\theta) = \mathbb{E}_{q\sim\mathcal{D},\, o\sim\pi_\theta}[R(q,o)]$. Let $g_R = \nabla J(\bar{\theta})$ and $\psi_t = \nabla_\theta \log \pi_\theta(o_t \mid s_t)|_{\theta=\bar{\theta}}$. For a fixed complete response with $M_o > 0$, define

$$p_o(t) = \frac{|A_t|}{M_o}, \qquad u_t = \operatorname{sgn}(A_t)\langle g_R, \psi_t\rangle. \tag{20}$$

Here, $p_o(t)$ is the original absolute-coefficient mass share at position $t$, not an action probability, and $u_t$ is the task-gradient projection per unit mass in the teacher-directed score direction. Its sign need not coincide with reward alignment. The corresponding per-response surrogate directions are $d_{\mathrm{V}}(o) = \sum_t A_t\psi_t$ and $d_{\mathrm{R2}}(o) = \sum_t A_t^{\mathrm{R2}}\psi_t$.

**Proposition 4** (Conditional task-progress identity)**.** *For the fixed response and directions above,*

$$\langle g_R, d_{\mathrm{R2}}(o) - d_{\mathrm{V}}(o)\rangle = \frac{M_o \operatorname{Cov}_{p_o}(g,u)}{\mathbb{E}_{p_o}[g]}. \tag{21}$$

Since $\mathbb{E}_{p_o}[g] > 0$, reweighting improves the response-level first-order task projection over standard OPD if and only if $\operatorname{Cov}_{p_o}(g,u) > 0$. Thus, larger gates must be associated with higher task-projection utility under the original mass distribution. Reducing reward-inconsistent mass alone does not guarantee task progress. Neither $g_R$ nor $u_t$ is a training input; both are used only for analysis. Appendix H.3 proves the identity and gives a sufficient reward–utility agreement condition. Appendices H.4 and H.5 distinguish this local criterion from finite-step reward ascent and surrogate-gradient compatibility.

## 5 Experiments

We assess R$^2$-OPD across student sizes and task domains, with controlled ablations to examine the role of reward-aligned supervision allocation.

### 5.1 Experimental Settings

**Models and training data.** We initialize students from Qwen3-4B-Base and Qwen3-1.7B-Base (Yang et al., 2025). The teacher, denoted Qwen3-4B-GRPO, is obtained by applying GRPO (Shao et al., 2024) to Qwen3-4B-Non-Thinking. The same frozen checkpoint is shared across all mathematical distillation methods and student sizes. Both teacher RL and student OPD use DAPO-Math-17k (Yu et al., 2025). We use a fixed $\beta = 0.001$ across all R$^2$-OPD experiments, without experiment-specific tuning. Full training details are provided in Appendix A.

**Evaluation.** We report avg@16 on AIME24/25 (MAA, n.d.a), and avg@4 on AMC23/24 (MAA, n.d.b), MATH500 (Hendrycks et al., 2021; Lightman et al., 2024), MinervaMath (Lewkowycz et al., 2022), and OlympiadBench (He et al., 2024). For the main 1.7B cross-size mathematical comparison, each method uses three independent training seeds with the same initial student and fixed teacher checkpoints, and results are reported as mean $\pm$ sample standard deviation. Given the computational budget, all remaining experiments use one training run per method or variant. The two initial students and fixed teacher are each evaluated once under the same benchmark-specific sampling protocol. Evaluation details are provided in Appendix A.3.

**Baselines.** We compare sampled-token OPD (Lu & Thinking Machines Lab, 2025), OPDVR (Lin et al., 2026), SG-OPD (Xu et al., 2026a), and ExOPD (Yang et al., 2026b). We also include Top-64 OPD, our top-$k$ implementation control of reverse-KL distillation (Gu et al., 2024), and OPD +

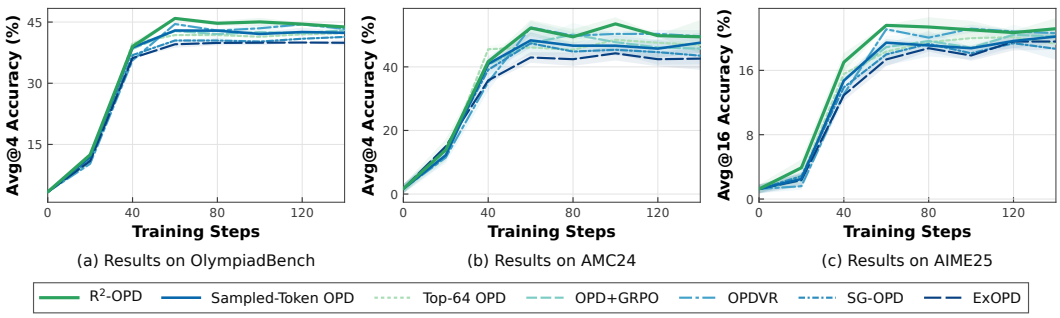



Figure 1: Evaluation accuracy on OlympiadBench, AMC24, and AIME25 during cross-size distillation from Qwen3-4B-GRPO to Qwen3-1.7B-Base. Curves show the mean, and shaded bands indicate ± one sample standard deviation across three independent training seeds.

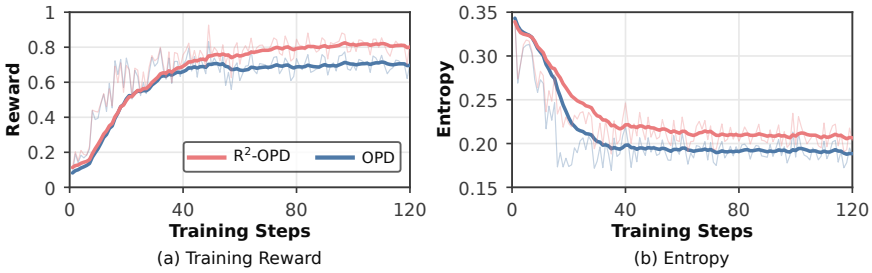
Table 1: Accuracy (%) on mathematical reasoning benchmarks under cross-size and same-size distillation. Within each distillation setting, **bold** and underlined scores indicate the best and second-best results in each column, respectively.

| Method | AIME24 | AIME25 | AMC23 | AMC24 | MATH500 | Minerva | Olympiad | Avg. |
|---|---|---|---|---|---|---|---|---|
| *Reference models* | | | | | | | | |
| Qwen3-1.7B-Base (Student) | 1.9 | 1.7 | 4.4 | 2.2 | 5.8 | 3.1 | 3.7 | 3.3 |
| Qwen3-4B-Base (Student) | 3.1 | 1.4 | 11.9 | 3.3 | 8.4 | 4.4 | 4.8 | 5.3 |
| Qwen3-4B-GRPO (Teacher) | 30.1 | 26.5 | 62.5 | 51.1 | 77.2 | 31.0 | 46.1 | 46.4 |
| *Qwen3-4B-GRPO → Qwen3-1.7B-Base* | | | | | | | | |
| Sampled-Token OPD | $25.5_{\pm 1.1}$ | $20.2_{\pm 0.8}$ | $60.0_{\pm 2.7}$ | $48.3_{\pm 2.4}$ | $74.7_{\pm 0.2}$ | $27.6_{\pm 0.9}$ | $42.9_{\pm 0.5}$ | $42.7_{\pm 0.6}$ |
| Top-64 OPD | $25.0_{\pm 1.3}$ | $20.1_{\pm 0.9}$ | $60.6_{\pm 1.0}$ | $48.5_{\pm 2.7}$ | $74.5_{\pm 1.2}$ | $27.1_{\pm 0.4}$ | $42.3_{\pm 0.0}$ | $42.6_{\pm 0.4}$ |
| OPD + GRPO | $26.1_{\pm 1.0}$ | $\underline{21.2_{\pm 1.2}}$ | $60.6_{\pm 3.4}$ | $50.2_{\pm 3.6}$ | $75.1_{\pm 0.3}$ | $\mathbf{30.6_{\pm 0.6}}$ | $43.0_{\pm 0.4}$ | $43.8_{\pm 0.9}$ |
| OPDVR | $\mathbf{29.8_{\pm 1.0}}$ | $20.4_{\pm 0.5}$ | $\mathbf{63.3_{\pm 3.0}}$ | $\underline{52.4_{\pm 2.0}}$ | $\underline{76.3_{\pm 1.3}}$ | $28.0_{\pm 1.0}$ | $\underline{44.5_{\pm 0.3}}$ | $\underline{45.0_{\pm 0.3}}$ |
| SG-OPD | $23.6_{\pm 0.2}$ | $19.4_{\pm 0.6}$ | $58.4_{\pm 2.4}$ | $47.4_{\pm 3.6}$ | $73.7_{\pm 0.2}$ | $27.8_{\pm 0.9}$ | $41.4_{\pm 0.3}$ | $41.7_{\pm 0.2}$ |
| ExOPD | $24.9_{\pm 1.1}$ | $19.6_{\pm 1.1}$ | $59.0_{\pm 3.4}$ | $44.3_{\pm 2.2}$ | $71.8_{\pm 0.2}$ | $27.9_{\pm 0.1}$ | $40.0_{\pm 0.0}$ | $41.1_{\pm 0.1}$ |
| **R²-OPD** | $\underline{29.4_{\pm 0.8}}$ | $\mathbf{24.6_{\pm 0.9}}$ | $\underline{63.0_{\pm 2.0}}$ | $\mathbf{53.5_{\pm 2.1}}$ | $\mathbf{77.6_{\pm 0.8}}$ | $\underline{29.1_{\pm 0.6}}$ | $\mathbf{45.9_{\pm 0.3}}$ | $\mathbf{46.2_{\pm 0.4}}$ |
| *Qwen3-4B-GRPO → Qwen3-4B-Base* | | | | | | | | |
| Sampled-Token OPD | 32.9 | 26.8 | 65.5 | 53.8 | 80.6 | 32.1 | 47.3 | 48.4 |
| Top-64 OPD | 32.6 | 27.0 | 65.2 | 53.2 | 80.1 | 31.6 | 46.9 | 48.1 |
| OPD + GRPO | 33.3 | 26.5 | 66.5 | $\underline{57.3}$ | 82.3 | 33.1 | 48.1 | 49.6 |
| OPDVR | 33.5 | 26.9 | 67.6 | 55.2 | **84.6** | 32.5 | $\underline{48.4}$ | 49.8 |
| SG-OPD | 33.1 | 26.6 | 67.9 | 54.9 | 81.8 | 31.3 | 47.6 | 49.0 |
| ExOPD | $\underline{34.7}$ | $\underline{27.6}$ | **68.9** | 54.5 | 82.1 | **33.9** | 47.5 | $\underline{49.9}$ |
| **R²-OPD** | **35.1** | **28.1** | $\underline{68.3}$ | **57.8** | $\underline{83.6}$ | $\underline{33.5}$ | **48.9** | **50.8** |

GRPO, an additive GRPO–OPD control (Shao et al., 2024; Lu & Thinking Machines Lab, 2025). The initial students and fixed teacher are reference models, not additional trained baselines.

### 5.2 Main Results

As shown in Table 1, R$^2$-OPD achieves the highest average accuracy in both distillation settings and outperforms standard OPD on all seven benchmarks. It reaches $46.2 \pm 0.4\%$ with the 1.7B student and $50.8\%$ with the 4B student, improving over standard OPD with average gains of $+3.5$ and $+2.4$, respectively. It also surpasses the strongest competing baselines, OPDVR and ExOPD, by $+1.2$ and $+0.9$, respectively. The gains are especially pronounced on challenging competition-mathematics benchmarks, reaching $+4.4$ on AIME25, $+5.2$ on AMC24, and $+3.0$ on OlympiadBench in the 1.7B setting. Figure 1 shows that these gains emerge early: a clear mean-accuracy advantage on AIME25 is already visible around step 40. Across all plotted benchmarks, R$^2$-OPD approaches near-peak accuracy by roughly step 60 and generally maintains higher mean accuracy thereafter, indicating both faster early learning and sustained performance gains.




(a) Training Reward (b) Entropy

Figure 2: Training reward and student policy entropy for R$^2$-OPD and standard OPD. Thin lines show raw observations; thick lines show smoothed trends.

### 5.3 Extension to Code Generation

We further evaluate R$^2$-OPD on code generation by distilling an separate RL-trained Qwen3-4B-Non-Thinking teacher into a Qwen3-1.7B-Base student. Both teacher RL and student OPD use TACO (Li et al., 2023), and we report avg@4 on HumanEval+ (Liu et al., 2023) and LiveCodeBench (Jain et al., 2025). As shown in Table 2, R$^2$-OPD outperforms vanilla OPD and OPDVR (Lin et al., 2026) on both benchmarks, achieving an average avg@4 of 61.2% with gains of +1.6 and +1.0, respectively. These results extend the observed benefits beyond mathematical reasoning.

Table 2: Code-generation avg@4 (%) under cross-size distillation from Qwen3-4B-GRPO to Qwen3-1.7B-Base. **Bold** and underlined scores indicate the best and second-best results among student-training methods.

| Method | HumanEval+ | LiveCodeBench | Avg. |
|---|---|---|---|
| Student | 64.0 | 21.1 | 42.6 |
| Teacher | 78.0 | 46.5 | 62.3 |
| Vanilla OPD | 75.4 | 43.7 | 59.6 |
| OPDVR | 76.2 | 44.2 | 60.2 |
| **R$^2$-OPD** | **77.3** | **45.1** | **61.2** |

### 5.4 Ablation Studies

Table 3 examines reward alignment, disagreement magnitude, and per-response coefficient-mass normalization, using one training run per variant. Reversed alignment negates $z(q, o)A_t$ in the gate without changing coefficient signs, yielding the largest average accuracy drop of 3.8. The sharpness sweep retains strong performance across the tested values from $10^{-4}$ to $10^{-1}$ (Appendix C). Together, these observations support reward-consistent allocation without requiring narrowly tuned gate sharpness. The outcome-permuted control shuffles labels across complete responses while preserving per-response mass and the marginal outcome distribution. Its 1.7 drop supports the value of correctly matched outcome feedback beyond mass preservation. The sign-only gate removes magnitude dependence while retaining the original advantages, reducing average accuracy by 1.6; the magnitude-only gate removes outcome alignment and incurs a larger decline of 2.4. These controls favor the full outcome- and magnitude-dependent configuration under the shared settings. Removing mass normalization while retaining the response-local scale reduces accuracy by 0.9 points, supporting normalization in this configuration. Adding GRPO improves all three benchmarks with an average gain of +0.8, suggesting complementarity with direct reward optimization. Control definitions and the $\beta$ sensitivity analysis are provided in Appendices B and C, respectively.

Table 3: Accuracy (%) for ablation studies and GRPO composition. $\Delta_{\text{avg}}$ denotes the change in mean accuracy across the three benchmarks relative to full R$^2$-OPD.

| Variant | AIME24 | MATH500 | Minerva | $\Delta_{\text{avg}}$ |
|---|---|---|---|---|
| **R$^2$-OPD (full)** | 30.1 | 77.8 | 29.4 | – |
| Reversed alignment | 25.4 | 73.5 | 26.9 | −3.8 |
| Outcome-permuted | 27.9 | 75.9 | 28.5 | −1.7 |
| Sign-only gate | 27.3 | 76.6 | 28.7 | −1.6 |
| Magnitude-only gate | 26.1 | 75.8 | 28.2 | −2.4 |
| No mass-norm | 28.1 | 76.7 | 29.8 | −0.9 |
| R$^2$-OPD + GRPO | 31.5 | 78.0 | 30.2 | +0.8 |

### 5.5 Training Dynamics

To examine how reward-aligned supervision allocation affects the distillation process, we track training reward and student policy entropy. Figure 2 compares R$^2$-OPD with standard sampled-token

OPD, denoted as OPD in the figure. Both methods improve training reward, but $R^2$-OPD maintains a higher smoothed reward trajectory over most of the plotted interval, with a separation that persists into late training. The raw observations fluctuate and overlap, so this comparison describes the overall trend rather than an advantage at every update. Policy entropy decreases for both methods, while $R^2$-OPD retains higher entropy during the middle and later stages. Thus, its higher reward trajectory coexists with higher policy entropy, rather than a greater reduction in entropy.

## 6 Related Work

**On-policy distillation of language models.** Knowledge distillation transfers a teacher distribution into a smaller student (Hinton et al., 2015). For autoregressive language models, MiniLLM adopts the mode-seeking reverse KL (Gu et al., 2024), GKD reduces train–test mismatch by querying the teacher on student-generated prefixes (Agarwal et al., 2024), and DistiLLM studies divergence choice and adaptive reuse of student rollouts (Ko et al., 2024). A common scalable formulation of on-policy distillation casts the detached teacher–student log-probability gap as a per-token reward in a policy-gradient-style objective (Lu & Thinking Machines Lab, 2025). Such objectives have been adopted in large-scale reasoning post-training (Yang et al., 2025; Xiao et al., 2026; DeepSeek-AI, 2026); see Song & Zheng (2026) for a broader survey. Recent work studies OPD optimization and failure modes (Li et al., 2026c; Fu et al., 2026), improves sampled-token estimation or target shaping (Oh et al., 2026; Jang et al., 2026), and rebalances training data (Hou et al., 2026). G-OPD generalizes OPD as dense KL-constrained RL; its ExOPD variant amplifies teacher–reference log-ratio rewards relative to KL regularization (Yang et al., 2026b).

**Selective token supervision in OPD.** Complementary work asks whether OPD should supervise every visited position equally. TSD-KD applies direct distribution matching to selected tokens and weaker preference feedback elsewhere (Kim & Baek, 2026). Other methods select or reweight tokens using teacher–student disagreement, uncertainty, or support compatibility (Xu et al., 2026b; Wang et al., 2026b; Xing et al., 2026; Jin et al., 2026); asymmetric or hierarchical schemes distinguish positive- and negative-advantage regions or combine trajectory filtering with token weighting (Jia et al., 2026; Li et al., 2026d). EGRSD incorporates outcome information by coupling reward-grounded update directions with teacher-derived magnitudes and confidence gating (Ke et al., 2026).

**Outcome-guided distillation and RLVR.** RLVR complements dense OPD feedback with outcome-level rewards (Shao et al., 2024; Guo et al., 2025). Existing methods combine or sequence these objectives (Xu et al., 2025; Xiao et al., 2026; Li et al., 2026a), route samples between them (Li et al., 2026b), or incorporate teacher information and verifier feedback into RL and self-distillation (Wang et al., 2026a; Yang et al., 2026a; Cai et al., 2026). Closely related, SG-OPD combines token-level sign-consistency routing with phased sampling of verified teacher responses (Xu et al., 2026a); and OPDVR masks outcome-inconsistent corrections (Lin et al., 2026). Related diagnostics examine outcome-dependent local supervision (Ma, 2026). $R^2$-OPD continuously reweights the original distillation coefficients using outcome alignment and disagreement magnitude.

## 7 Conclusion

We revisit uniform supervision in on-policy distillation through the mismatch between *local teacher preference* and *student continuation value*. Our analysis shows that local imitation can reduce reverse KL while lowering expected reward under a fixed student continuation policy, providing a unified account of over-correction and over-trust. This motivates $R^2$-OPD, which uses trajectory outcomes as evidence for supervision allocation and continuously reweights teacher corrections using reward alignment and disagreement magnitude. The method preserves dense feedback, correction signs, and per-response absolute coefficient mass. We establish sufficient reward–utility agreement conditions under which reallocation improves first-order task progress over uniform OPD. Experiments across student sizes and task domains demonstrate consistent gains over standard OPD in mathematical reasoning and code generation, while ablations support the core design. Together, these findings highlight that effective OPD depends not only on *where* teacher feedback is collected, but also on *how its influence is allocated* using evidence of student task success.

## Reproducibility Statement

Section 4 and Algorithm 1 specify the reweighting rule and the training procedure; Appendix A reports the training configurations, computing environment, and evaluation protocols, and Appendix B defines the ablations. Assumptions and derivations appear in Appendices F and G, and the task-progress analysis in Appendix H.3. A reference implementation—the $R^2$-OPD coefficient, its ablations, the outcome-permuted control, the compared baselines, and the tests that pin them—is available at `https://anonymous.4open.science/r/r2-opd-62F4/`; it is extracted verbatim from the code that produced our runs, checked against it numerically, and depends only on PyTorch.

## AI Use Statement

Large language models assisted with literature search, manuscript drafting and editing, checks of mathematical derivations, and LaTeX formatting. The authors designed and conducted the experiments, reviewed the AI-assisted content, and take full responsibility for the methodology, theoretical claims, empirical results, and final manuscript.

## References


Rishabh Agarwal, Nino Vieillard, Yongchao Zhou, Piotr Stanczyk, Sabela Ramos, Matthieu Geist, and Olivier Bachem. On-policy distillation of language models: Learning from self-generated mistakes. In *International Conference on Learning Representations*, 2024. URL https://proceedings.iclr.cc/paper_files/paper/2024/hash/5be69a584901a26c521c2b51e40a4c20-Abstract-Conference.html.

Qiye Cai, Yichuan Ma, Peiji Li, Yongkang Chen, Qipeng Guo, Yicheng Zou, Linyang Li, Xiaocheng Feng, and Bing Qin. $H^2$sd: Hybrid hindsight self-distillation, 2026. URL https://arxiv.org/abs/2607.18955.

DeepSeek-AI. Deepseek-v4: Towards highly efficient million-token context intelligence. arXiv preprint arXiv:2606.19348, 2026. URL https://arxiv.org/abs/2606.19348.

Yuqian Fu, Haohuan Huang, Kaiwen Jiang, Jiacai Liu, Zhuo Jiang, Yuanheng Zhu, and Dongbin Zhao. Revisiting on-policy distillation: Empirical failure modes and simple fixes. arXiv preprint arXiv:2603.25562, 2026. URL https://arxiv.org/abs/2603.25562.

Yuxian Gu, Li Dong, Furu Wei, and Minlie Huang. Minillm: Knowledge distillation of large language models. In *International Conference on Learning Representations*, 2024. URL https://proceedings.iclr.cc/paper_files/paper/2024/hash/8ac015d409635f196f9e3e9dcfb9a94e-Abstract-Conference.html.

Daya Guo, Dejian Yang, Haowei Zhang, et al. DeepSeek-R1 incentivizes reasoning in LLMs through reinforcement learning. *Nature*, 645(8081):633–638, 2025. doi: 10.1038/s41586-025-09422-z. URL https://www.nature.com/articles/s41586-025-09422-z.

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. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 3828–3850. Association for Computational Linguistics, 2024. doi: 10.18653/v1/2024.acl-long.211. URL https://aclanthology.org/2024.acl-long.211/.

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. In *Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks*, volume 1, 2021. URL https://datasets-benchmarks-proceedings.neurips.cc/paper/2021/hash/be83ab3ecd0db773eb2dc1b0a17836a1-Abstract-round2.html.

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

Wenjin Hou, Shangpin Peng, Weinong Wang, Zheng Ruan, Yue Zhang, Zhenglin Zhou, Mingqi Gao, Yifei Chen, Kaiqi Wang, Hongming Yang, Chengquan Zhang, Zhuotao Tian, Han Hu, Yi Yang, Fei Wu, and Hehe Fan. Uni-opd: Unifying on-policy distillation with a dual-perspective recipe. arXiv preprint arXiv:2605.03677, 2026. URL https://arxiv.org/abs/2605.03677.

Naman Jain, King Han, Alex Gu, Wen-Ding Li, Fanjia Yan, Tianjun Zhang, Sida I. Wang, Armando Solar-Lezama, Koushik Sen, and Ion Stoica. Livecodebench: Holistic and contamination free evaluation of large language models for code. In *International Conference on Learning Representations*, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/hash/94074dd5a072d28ff75a76dabed43767-Abstract-Conference.html.

Ijun Jang, Jewon Yeom, Juan Yeo, Hyunggyu Lim, and Taesup Kim. Stable on-policy distillation through adaptive target reformulation. In *Findings of the Association for Computational Linguistics: ACL 2026*, pp. 42217–42227. Association for Computational Linguistics, 2026. doi: 10.18653/v1/2026.findings-acl.2094. URL https://aclanthology.org/2026.findings-acl.2094/.

Nan Jia, Haojin Yang, Xing Ma, Jiesong Lian, Shuailiang Zhang, Weipeng Zhang, Ke Zeng, Xunliang Cai, and Zequn Sun. Asymmetric on-policy distillation: Bridging exploitation and imitation at the token level, 2026. URL https://arxiv.org/abs/2605.06387.

Woogyeol Jin, Taywon Min, Yongjin Yang, Dennis Wei, Yi Zhou, Swanand Ravindra Kadhe, Nathalie Baracaldo, and Kimin Lee. Entropy-aware on-policy distillation of language models. In *Proceedings of the 43rd International Conference on Machine Learning*, 2026. URL https://arxiv.org/abs/2603.07079.

Junlong Ke, Zichen Wen, Weijia Li, Conghui He, and Linfeng Zhang. Respecting self-uncertainty in on-policy self-distillation for efficient llm reasoning, 2026. URL https://arxiv.org/abs/2605.13255.

Minsang Kim and Seung Jun Baek. Explain in your own words: Improving reasoning via token-selective dual knowledge distillation. In *International Conference on Learning Representations*, 2026. URL https://proceedings.iclr.cc/paper_files/paper/2026/hash/a89331e2ecbbd9d9618bc94717a91155-Abstract-Conference.html.

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*, volume 235 of *Proceedings of Machine Learning Research*, pp. 24872–24895. PMLR, 2024. URL https://proceedings.mlr.press/v235/ko24c.html.

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. In *Advances in Neural Information Processing Systems*, volume 35, 2022. doi: 10.52202/068431-0278. URL https://proceedings.neurips.cc/paper_files/paper/2022/hash/18abbeef8cfe9203fdf9053c9c4fe191-Abstract.html.

Boyan Li, Bingsen Chen, Chenghao Yang, Ping Nie, Chen Zhao, and Xi Ye. Sequential beats joint: On the interplay between on-policy distillation and rlvr, 2026a. URL https://arxiv.org/abs/2609.04108.

Gengsheng Li, Tianyu Yang, Junfeng Fang, Mingyang Song, Mao Zheng, Haiyun Guo, Dan Zhang, Jinqiao Wang, and Tat-Seng Chua. Unifying group-relative and self-distillation policy optimization via sample routing, 2026b. URL https://arxiv.org/abs/2604.02288.

Rongao Li, Jie Fu, Bo-Wen Zhang, Tao Huang, Zhihong Sun, Chen Lyu, Guang Liu, Zhi Jin, and Ge Li. Taco: Topics in algorithmic code generation dataset, 2023. URL https://arxiv.org/abs/2312.14852.

Yaxuan Li, Yuxin Zuo, Bingxiang He, Jinqian Zhang, Chaojun Xiao, Cheng Qian, Tianyu Yu, Huanang Gao, Wenkai Yang, Zhiyuan Liu, and Ning Ding. Rethinking on-policy distillation of large language models: Phenomenology, mechanism, and recipe. arXiv preprint arXiv:2604.13016, 2026c. URL https://arxiv.org/abs/2604.13016. Accepted at the ICML 2026 FoGen Workshop.

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, 2026d. URL https://arxiv.org/abs/2606.02684.

Hunter Lightman, Vineet Kosaraju, Yura Burda, Harri 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. URL https://proceedings.iclr.cc/paper_files/paper/2024/hash/aca97732e30bcf1303bc22ac3924fd16-Abstract-Conference.html.

Wenze Lin, Jiale Zhao, Xitai Jiang, Songde Rao, Yining Li, Shenzhi Wang, Bingxiang He, and Gao Huang. On-policy distillation with verifiable reward, 2026. URL https://arxiv.org/abs/2608.24696.

Jiawei Liu, Chunqiu Steven Xia, Yuyao Wang, and Lingming Zhang. Is your code generated by chatgpt really correct? rigorous evaluation of large language models for code generation. In *Advances in Neural Information Processing Systems*, volume 36, 2023. doi: 10.52202/075280-0943. URL https://proceedings.neurips.cc/paper_files/paper/2023/hash/43e9d647ccd3e4b7b5baab53f0368686-Abstract.html.

Kevin Lu and Thinking Machines Lab. On-policy distillation. Thinking Machines Lab: Connectionism, 2025. URL https://thinkingmachines.ai/blog/on-policy-distillation/. Blog post.

Guoqing Ma. Outcome-confounded local supervision in on-policy distillation, 2026. URL https://arxiv.org/abs/2607.23731.

MAA. MAA invitational competitions: American invitational mathematics examination (AIME). Mathematical Association of America website, n.d.a. URL https://maa.org/maa-invitational-competitions/. Accessed: 2026-09-15.

MAA. American mathematics competitions. Mathematical Association of America website, n.d.b. URL https://maa.org/student-programs/amc/. Accessed: 2026-09-15.

Minjae Oh, Sangjun Song, Gyubin Choi, Yunho Choi, and Yohan Jo. Kl for a kl: On-policy distillation with control variate baseline, 2026. URL https://arxiv.org/abs/2605.07865.

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

Mingyang Song and Mao Zheng. A survey of On-Policy Distillation for large language models. arXiv preprint arXiv:2604.00626, 2026. URL https://arxiv.org/abs/2604.00626.

Chen Wang, Zhaochun Li, Jionghao Bai, Yining Zhang, Hexuan Deng, Ge Lan, and Yue Wang. Distilled reinforcement learning for llm post-training, 2026a. URL https://arxiv.org/abs/2607.17247.

Yuanyi Wang, Su Lu, Yanggan Gu, Pengkai Wang, Yifan Yang, Zhaoyi Yan, Congkai Xie, Jianmin Wu, and Hongxia Yang. Nnot all disagreement is learnable: Token teachability in on-policy distillation. arXiv preprint arXiv:2605.26844, 2026b. URL https://arxiv.org/abs/2605.26844.

Bangjun Xiao et al. MiMo-V2-Flash technical report. arXiv preprint arXiv:2601.02780, 2026. URL https://arxiv.org/abs/2601.02780.

Xingrun Xing, Haoqing Wang, Boyan Gao, Ziheng Li, and Yehui Tang. Trust region on-policy distillation, 2026. URL https://arxiv.org/abs/2606.01249.

Haoran Xu, Hongyu Wang, Yifei Gao, Jiaze Li, Xiaofeng Zhang, and Xiaosong Yuan. Sg-opd: Sign-gated on-policy distillation via sign-consistency gating and phased teacher sampling. arXiv preprint arXiv:2606.09304, 2026a. URL https://arxiv.org/abs/2606.09304.

Hongling Xu, Qi Zhu, Heyuan Deng, Jinpeng Li, Lu Hou, Yasheng Wang, Lifeng Shang, Ruifeng Xu, and Fei Mi. Kdrl: Post-training reasoning llms via unified knowledge distillation and reinforcement learning, 2025. URL https://arxiv.org/abs/2506.02208.

Yuanda Xu, Hejian Sang, Zhengze Zhou, Ran He, Zhipeng Wang, and Alborz Geramifard. Tip: Token importance in on-policy distillation. arXiv preprint arXiv:2604.14084, 2026b. URL https://arxiv.org/abs/2604.14084.

An Yang et al. Qwen3 technical report. arXiv preprint arXiv:2505.09388, 2025. URL https://arxiv.org/abs/2505.09388.

Chenxu Yang, Chuanyu Qin, Qingyi Si, Minghui Chen, Naibin Gu, Dingyu Yao, Zheng Lin, Weiping Wang, Jiaqi Wang, and Nan Duan. Self-distilled rlvr, 2026a. URL https://arxiv.org/abs/2604.03128.

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

Qiying Yu, Zheng Zhang, Ruofei Zhu, Yufeng Yuan, Xiaochen Zuo, Yu Yue, Weinan Dai, Tiantian Fan, Gaohong Liu, Juncai Liu, Lingjun Liu, Xin Liu, Haibin Lin, Zhiqi Lin, Bole Ma, Guangming Sheng, Yuxuan Tong, Chi Zhang, Mofan Zhang, Ru Zhang, Wang Zhang, Hang Zhu, Jinhua Zhu, Jiaze Chen, Jiangjie Chen, Chengyi Wang, Hongli Yu, Yuxuan Song, Xiangpeng Wei, Hao Zhou, Jingjing Liu, Wei-Ying Ma, Ya-Qin Zhang, Lin Yan, Yonghui Wu, and Mingxuan Wang. Dapo: An open-source llm reinforcement learning system at scale. In *Advances in Neural Information Processing Systems*, volume 38, 2025. doi: 10.52202/085713-3775. URL https://proceedings.neurips.cc/paper_files/paper/2025/hash/a4277440d50f1f15d2cb4c14f7e0c0d2-Abstract-Conference.html.

Zheng Yuan, Hongyi Yuan, Chengpeng Li, Guanting Dong, Keming Lu, Chuanqi Tan, Chang Zhou, and Jingren Zhou. Scaling relationship on learning mathematical reasoning with large language models, 2023. URL https://arxiv.org/abs/2308.01825.

# A Implementation Details

Table 4: Training settings and mathematical evaluation.

| Parameter | Value |
|---|---|
| *Optimization* | |
| Optimizer | AdamW |
| Learning rate ($\eta$) | $1 \times 10^{-6}$ |
| Schedule / warmup | Constant / none |
| AdamW $(\beta_1, \beta_2)$ | (0.9, 0.99) |
| AdamW $\epsilon$ | $10^{-8}$ |
| Weight decay | 0.1 |
| Global $\ell_2$-norm clip | 1.0 |
| Accumulation and all-reduce precision | FP32 |
| *Distillation loss* | |
| Clipping $(\epsilon_{\text{low}}, \epsilon_{\text{high}})$ | (0.2, 0.2) |
| OPD multiplier ($c_{\text{OPD}}$) | 1 |
| Scale floor ($\epsilon$) | $2^{-23}$ |
| Dual clipping | Disabled |
| Entropy / reference-KL coefficients | 0 / 0 |
| *Training and rollout* | |
| Prompts per rollout batch | 64 |
| Responses per prompt | 4 |
| Optimizer batch (responses) | 256 |
| Micro-batch tokens per device | $\leq$ 10,240 |
| Rollout temperature | 1.0 |
| Rollout top-$p$ / top-$k$ | 1 / disabled |
| Maximum prompt length | 2,048 tokens |
| Maximum response length | 8,192 tokens |
| *Mathematical evaluation* | |
| Samples per AIME24/25 problem | 16 |
| Samples per other problem | 4 |
| Temperature | 1.0 |
| Top-$p$ | 0.95 |

## A.1 Training Configuration

Mathematical reasoning and code generation share the optimization and rollout settings in Table 4. Unless otherwise specified, controlled comparisons use the same prompt budget.

**Hardware and software.** All experiments use a single machine with eight NVIDIA H800-80GB GPUs interconnected via 400 GB/s NVLink. The software environment comprises Ubuntu 22.04.5, Python 3.11.15, PyTorch 2.11.0, CUDA 12.9, NCCL 2.26.2, Ray 2.55.1, SGLang 0.5.12.post1, Transformers 5.6.0, and FlashAttention 2.7.4.post1.

## A.2 Policy Loss and Optimization Protocol

**Policy loss and reduction.** For a batch $\mathcal{B}$ of $N$ responses, let $m_{i,t}$ mark valid response tokens, $n_i = \sum_t m_{i,t}$, $\widetilde{n}_i = \max\{n_i, 1\}$, and $\widehat{A}_{i,t} = \text{sg}[c_{\text{OPD}} A^{\text{R2}}_{i,t}]$. With a detached snapshot denominator,

define $\rho_{i,t}(\theta) = \pi_\theta(o_{i,t} \mid s_{i,t})/\pi_{\bar{\theta}}(o_{i,t} \mid s_{i,t})$. The implemented loss is

$$\widehat{\mathcal{L}}_{\text{impl}}(\theta; \mathcal{B}) = \frac{1}{N}\sum_{i=1}^{N}\frac{1}{\widetilde{n}_i}\sum_t m_{i,t}\,\ell_{\text{clip}}(\rho_{i,t}(\theta), \widehat{A}_{i,t}),$$
$$\ell_{\text{clip}}(\rho, a) = -\min\{\rho a, \text{clip}(\rho, 1-\epsilon_{\text{low}}, 1+\epsilon_{\text{high}})a\}\,, \tag{22}$$

with numerical settings in Table 4. This averages per-response token means rather than pooling tokens across responses. Masks include generated tokens and terminal EOS when present, exclude prompts and padding, and retain truncated responses; empty masks contribute zero. Distributed accumulation preserves the weights $1/(N\widetilde{n}_i)$. No additional advantage whitening, centering, coefficient clipping, or weight cap is applied. Truncated responses retain standard OPD coefficients, and zero-mass responses retain zero coefficients. Pure distillation uses outcomes only through the gate, without an additional reward-derived advantage.

**Rollout and update schedule.** Each fresh rollout batch forms one optimizer batch, with gradients accumulated across token-packed micro-batches before a single optimizer step, without replay or additional optimization epochs. Student weights are synchronized before sampling and remain unchanged until the update. The trainer recomputes detached snapshot log probabilities for both $A_{i,t}$ and the ratio denominator; teacher scores and reweighting coefficients remain fixed throughout accumulation. Engine–trainer log-probability discrepancies are logged without importance-sampling correction. With consistent trainer evaluations, $\rho_{i,t}(\bar{\theta}) = 1$, so clipping is inactive and the policy-loss gradient matches the detached log-probability surrogate with the same response-wise reduction. Its relation to the token-summed analysis is given in Appendix I.

### A.3 Evaluation and Reporting

Using the benchmark-specific sample count $k_{\mathcal{E}}$ and sampling settings in Table 4, we report

$$\text{Acc}^{(s)}(\mathcal{E}) = \frac{100}{|\mathcal{E}|k_{\mathcal{E}}}\sum_{q\in\mathcal{E}}\sum_{j=1}^{k_{\mathcal{E}}} R(q, o_{q,j}^{(s)}). \tag{23}$$

For the main mathematical comparison, we report mean $\pm$ sample standard deviation across three training seeds for 1.7B cross-size distillation and single-run results for 4B same-size distillation. Reference checkpoints are evaluated once. The seven-benchmark macro-average is computed within each run before cross-seed aggregation.

## B Ablation Experiments

Unless otherwise specified, the controls in Table 3 use the same training and evaluation settings as full R$^2$-OPD.

### B.1 Ablation Definition

**Response-level normalization.** For a complete response with $M_o = \sum_t |A_t| > 0$, each modified gate $g_t^{(v)}$ uses a recomputed normalizer:

$$Z_o^{(v)} = \frac{M_o}{\sum_j g_j^{(v)}|A_j|}, \qquad A_t^{(v)} = Z_o^{(v)} g_t^{(v)} A_t. \tag{24}$$

This preserves coefficient mass under each gate modification; no mass-norm is the exception. All advantages and reweighting factors are detached. Zero-mass responses retain zero coefficients with $Z_o = 1$; truncated responses retain standard OPD coefficients. Below, $z = z(q, o)$ and $\nu_o$ follows Eq. (13).

**Reversed alignment.** Negating the alignment signal gives

$$g_t^{\text{rev}} = \sigma\left(-\beta\frac{zA_t}{\nu_o}\right). \tag{25}$$

This reverses the gate's relative emphasis on aligned and inconsistent corrections without changing $A_t$ or verifier labels.

**Outcome-permuted.** Within each rollout batch, we permute outcome labels across complete responses while retaining their advantages, scales, and token masks. For a permutation $\varpi$ of these responses, the gate is

$$g_{i,t}^{\text{perm}} = \sigma\left(\beta \frac{z_{\varpi(i)} A_{i,t}}{\nu_{o_i}}\right). \tag{26}$$

The normalizer is recomputed using Eq. (24), preserving per-response coefficient mass and the marginal outcome distribution. Truncated responses retain standard OPD coefficients.

**Sign-only gate.** We remove magnitude dependence from the gate by assigning one fixed level per alignment group:

$$g_t^{\text{sign}} = \begin{cases} 3/4, & zA_t > 0, \\ 1/4, & zA_t < 0, \\ 1/2, & zA_t = 0, \end{cases} \tag{27}$$

so the gate reads neither $\beta$ nor $\nu_o$, while the original $A_t$ remains in Eq. (**??**). Let $a$ denote the aligned share of the response's coefficient mass. Normalization gives multipliers $Z_o^{\text{sign}} g_t^{\text{sign}} = 3/(1+2a)$ on aligned positions and $1/(1+2a)$ on inconsistent ones, so aligned corrections receive exactly three times the multiplier of inconsistent ones, within $[1, 3]$ and $[1/3, 1]$ respectively. This control removes magnitude dependence at a fixed reallocation ratio; it does not match the full gate's reweighting strength, which for two positions of equal magnitude and opposite alignment is exactly $\exp(\beta|A_t|/\nu_o)$, about 1.001 at a median-magnitude position under the shared $\beta = 10^{-3}$.

**Magnitude-only gate.** We retain disagreement magnitude without outcome alignment:

$$g_t^{\text{mag}} = \sigma\left(\beta \frac{|A_t|}{\nu_o}\right). \tag{28}$$

Weighting depends only on normalized disagreement magnitude; the final coefficients retain their original signs.

**No mass-norm.** Setting $Z_o = 1$ while retaining the full gate and scale gives

$$A_t^{\text{no-norm}} = g_t A_t. \tag{29}$$

This removes coefficient-mass restoration, not the scale $\nu_o$. It changes coefficient scale and is therefore not a direction-only comparison.

**Composition with GRPO.** We replace the OPD component in OPD + GRPO with the full $R^2$-OPD coefficient:

$$A_t^{\text{joint}} = A_t^{\text{GRPO}} + \lambda_{\text{OPD}} A_t^{\text{R2}}, \qquad \lambda_{\text{OPD}} = 1. \tag{30}$$

The GRPO advantage is computed unchanged and excluded from the $R^2$ gate and response-level coefficient-mass normalization. The combined detached advantage is inserted into the clipped policy surrogate using the mean of per-response token means, as specified in Appendix A.2. All OPD coefficients are detached. This control tests the complementarity between direct reward optimization and reward-aligned distillation.

### B.2 Outcome-Branch Controls

Table 5 compares full $R^2$-OPD with reweighting restricted to successful or failed responses, using one training run per variant. For complete responses, the controls are defined as follows.

**Success-only.** We reweight successful responses and retain standard OPD on failures:

$$A_t^{\text{success}} = \begin{cases} A_t^{\text{R2}}, & z = +1, \\ A_t, & z = -1. \end{cases} \tag{31}$$

Relative to full $R^2$-OPD, this removes reweighting on failed responses without discarding their teacher supervision.

Table 5: Accuracy (%) on seven mathematical reasoning benchmarks for outcome-branch controls. $\Delta_{\text{avg}}$ reports the change in the seven-benchmark average relative to full R$^2$-OPD.

| Variant | AIME24 | AIME25 | AMC23 | AMC24 | MATH500 | Minerva | Olympiad | $\Delta_{\text{avg}}$ |
|---|---|---|---|---|---|---|---|---|
| R$^2$-OPD (full) | 30.1 | 24.5 | 65.0 | 54.4 | 77.8 | 29.4 | 45.9 | – |
| Success-only | 31.3 | 22.3 | 63.7 | 50.0 | 76.3 | 29.1 | 45.8 | -1.2 |
| Failure-only | 28.1 | 20.0 | 63.7 | 52.2 | 75.1 | 28.7 | 43.7 | -2.2 |

**Failure-only.** We reweight failed responses and retain standard OPD on successes:

$$A_t^{\text{failure}} = \begin{cases} A_t, & z = +1, \\ A_t^{\text{R2}}, & z = -1. \end{cases} \tag{32}$$

Relative to full R$^2$-OPD, this removes reweighting on successful responses without discarding their teacher supervision.

Both controls use the full gate and per-response normalization on the selected branch, without rescaling for branch frequency. Truncated responses retain standard OPD coefficients. The full reference uses the same experimental setting as the branch controls.

**Results and interpretation.** Full R$^2$-OPD achieves the highest seven-benchmark average of 46.7, exceeding Success-only and Failure-only by 1.2 and 2.2, respectively. The larger decline of Failure-only suggests that removing reweighting on successful responses is more detrimental in this setting. Together, these results support reweighting both outcome branches, consistent with their complementary roles in addressing over-correction and over-trust. The comparisons reflect branch-level effects under the observed outcome distribution, not causal token-level utility.

## C Hyperparameter Sensitivity

We examine gate sharpness in the Qwen3-4B-GRPO → Qwen3-1.7B-Base setting through an accuracy sweep, a fixed-rollout coefficient analysis, and training-time mass profiles. All settings retain the full reweighting rule, with the response-level normalizer recomputed for each $\beta$. Accuracy results use one training run per setting; the default $\beta = 10^{-3}$ is shared across the main experiments without experiment-specific tuning.

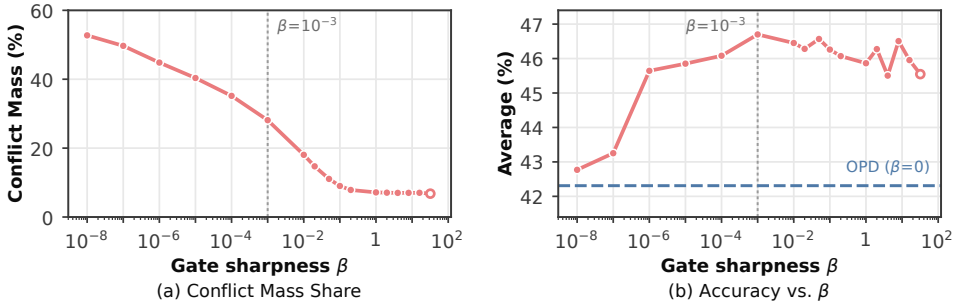



Figure 3: Gate-sharpness analysis. (a) Mean response-level inconsistent coefficient-mass share on a fixed rollout batch, holding advantages, outcomes, response-local scales, and token masks constant across $\beta$. Responses are equally weighted, including identity-fallback responses. (b) Average mathematical accuracy from one training run per setting. Vertical dotted lines mark $\beta = 10^{-3}$; the horizontal dashed line in (b) denotes standard OPD, corresponding to $\beta = 0$.

### C.1 Sensitivity of Task Performance

Figure 3(b) shows that accuracy varies non-monotonically with gate sharpness. It increases from approximately 42.8% at $\beta = 10^{-8}$ to 46.7% at $\beta = 10^{-3}$, the highest observed score in the sweep. The sampled values between $10^{-4}$ and $10^{-1}$ achieve roughly 46.1–46.7%, whereas those with $\beta \geq 1$

yield approximately 45.5–46.5%. All tested settings exceed the displayed OPD reference. Thus, strong performance extends beyond a single sharpness value, but stronger gating offers no consistent improvement over the default.

### C.2 Reward-Inconsistent Mass Allocation

For a response with $M_o > 0$, $\kappa_o(\beta)$ is the fraction of its absolute coefficient mass assigned to positions with $z(q,o)A_t < 0$ (Eq. (18)). For a rollout batch $\mathcal{B}$ of $N = |\mathcal{B}|$ responses, we report $\bar{\kappa}_{\mathcal{B}}(\beta) = N^{-1}\sum_{o\in\mathcal{B}} \kappa_o(\beta)$, using equal response weights $\omega_o = 1/N$ rather than pooling coefficient mass across responses. Truncated responses routed to the vanilla fallback remain in the average. Their original coefficients give $\kappa_o(\beta) = \kappa_o(0)$ for every $\beta$; their outcomes are used for this diagnostic but not for reweighting.

**Fixed-rollout reallocation.** Figure 3(a) applies the reweighting rule and its fallback to identical rollouts. The mean inconsistent share decreases from approximately 53% at $\beta = 10^{-8}$ to 28% at the default, approaching 7% at large sharpness. The default thus lowers the mean response-level share by about 25 percentage points with batch membership and original coefficients fixed. Unchanged fallback contributions remain part of the residual share.

**Paired training-time analysis.** Figure 4 compares original and applied coefficients on each R$^2$-OPD rollout batch. R$^2$-OPD (natural) uses the original $A_t$, corresponding to $\bar{\kappa}_{\mathcal{B}}(0)$ on that batch, not a separate training run. Its trajectory closely tracks independently trained OPD, indicating similar mean pre-reweighting conflict shares. From approximately step 50 onward, the natural share remains near 50–55%, whereas the applied share is 25–30%. The paired curves use identical responses, advantages, outcomes, masks, and equal response weights; fallback responses contribute identically to both curves. Their gap therefore isolates the direct reallocation effect on the mean within-response share, which is approximately halved while per-response coefficient mass is preserved.

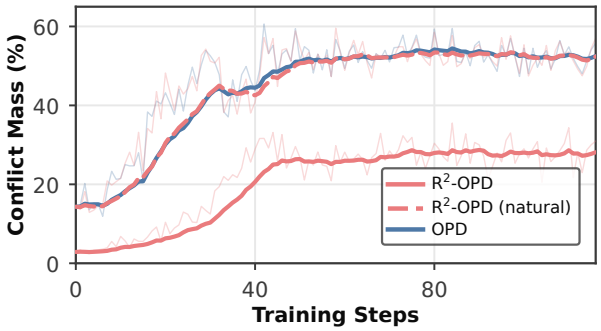



Figure 4: Mean response-level inconsistent coefficient-mass share during 4B-to-1.7B distillation, averaged equally over all batch responses, including identity-fallback responses. R$^2$-OPD (red solid) uses the applied coefficients at $\beta = 10^{-3}$; R$^2$-OPD (natural, red dashed) uses the original coefficients on the same rollouts. OPD (blue solid) denotes an independent standard-OPD run. Thin traces show raw observations; thick curves show smoothed trends.

**Effective weighting and coefficient redistribution.** For $M_o > 0$, let $p_o(t) = |A_t|/M_o$. On reweighted responses, the effective multiplier at nonzero-advantage positions is

$$w_t(\beta) = \frac{A_t^{\mathrm{R2}}(\beta)}{A_t} = \frac{g_t(\beta)}{\mathbb{E}_{p_o}[g(\beta)]}. \tag{33}$$

On identity-fallback responses, $w_t(\beta) = 1$. Thus, effective reweighting depends on normalized disagreement and its coefficient-mass distribution, not on $\beta$ alone. Near-uniform gating at a median-magnitude token does not imply near-identity weighting over the response.

Using the same full-batch denominator, define

$$D_{\mathcal{B}}(\beta) = \frac{1}{N}\sum_{o\in\mathcal{B}} d_o(\beta), \qquad d_o(\beta) = \mathbb{E}_{p_o}[|w_t(\beta) - 1|]\,. \tag{34}$$

Fallback responses have $d_o = 0$ but remain counted in $N$. If their batch fraction is $f_{\mathcal{B}} < 1$, then $D_{\mathcal{B}} = (1 - f_{\mathcal{B}})\bar{d}_{\text{nonfb}}$, where $\bar{d}_{\text{nonfb}}$ is the mean displacement over the remaining responses. Mass conservation holds on both reweighted and fallback responses, so the triangle inequality gives

$$D_{\mathcal{B}}(\beta) \geq 2\,|\bar{\kappa}_{\mathcal{B}}(\beta) - \bar{\kappa}_{\mathcal{B}}(0)|\,. \tag{35}$$

Proposition 3 gives nonincreasing shares on reweighted responses, while fallback shares are constant. Hence $\bar{\kappa}_{\mathcal{B}}(0) \geq \bar{\kappa}_{\mathcal{B}}(10^{-8})$, and the reported shares imply $D_{\mathcal{B}}(10^{-3}) \gtrsim 0.50$. This is a lower bound on the mean response-normalized coefficient displacement across the full batch, including unchanged fallback responses, not on a pooled mass ratio or parameter-gradient displacement. It does not identify the mass-weighted tail of normalized disagreement or attribute accuracy gains to particular token ranges.

**Implications for supervision allocation.** The fixed-rollout analysis shows substantial within-response reallocation at the default, while the accuracy sweep shows that stronger suppression of inconsistent mass does not consistently improve performance. This distinction is consistent with Eq. (21): first-order task progress depends on the association between gate weights and correction utility, not on inconsistent mass alone. The results support calibrating correction strength using outcome evidence, without treating maximal outcome selectivity as the goal.

# D Additional Training Dynamics

## D.1 Response Length

Figure 5 compares training-time response lengths for R$^2$-OPD and standard sampled-token OPD. Both methods exhibit rapid length growth during early training, reaching approximately 5.5k tokens by step 40. Their trajectories remain closely matched thereafter, including similar fluctuations during later training. The performance gains therefore coexist with comparable response lengths, providing evidence against longer training rollouts as their primary explanation.

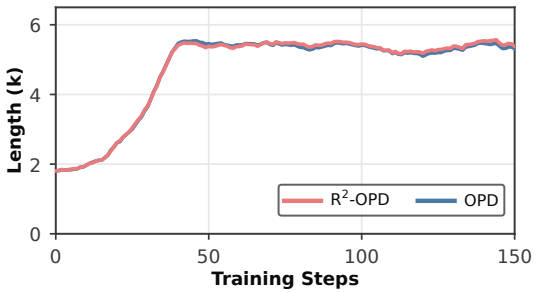



Figure 5: Training-time response length for R$^2$-OPD and standard sampled-token OPD, denoted as OPD. Length is measured in thousands of tokens.

## D.2 Evaluation Trajectories Across All Benchmarks

Figure 6 extends the learning curves in Figure 1 to all seven mathematical benchmarks in the 1.7B cross-size setting. Most methods improve rapidly within the first 40–60 training steps and change more gradually thereafter. R$^2$-OPD shows early and sustained gains on several benchmarks, with particularly visible separation on AIME24, AIME25, AMC24, and OlympiadBench. These benchmark-wise trajectories complement the reported results in Table 1 and reveal variation in the magnitude of the gains.

# E Computational Cost

**Model computation.** For matched rollout batches and update schedules, R$^2$-OPD uses the same student generation, teacher scoring, and student optimization as standard sampled-token OPD. Reweighting introduces no additional model passes, trainable parameters, or inference-time components. Neither teacher-generated responses nor auxiliary value estimation is required.

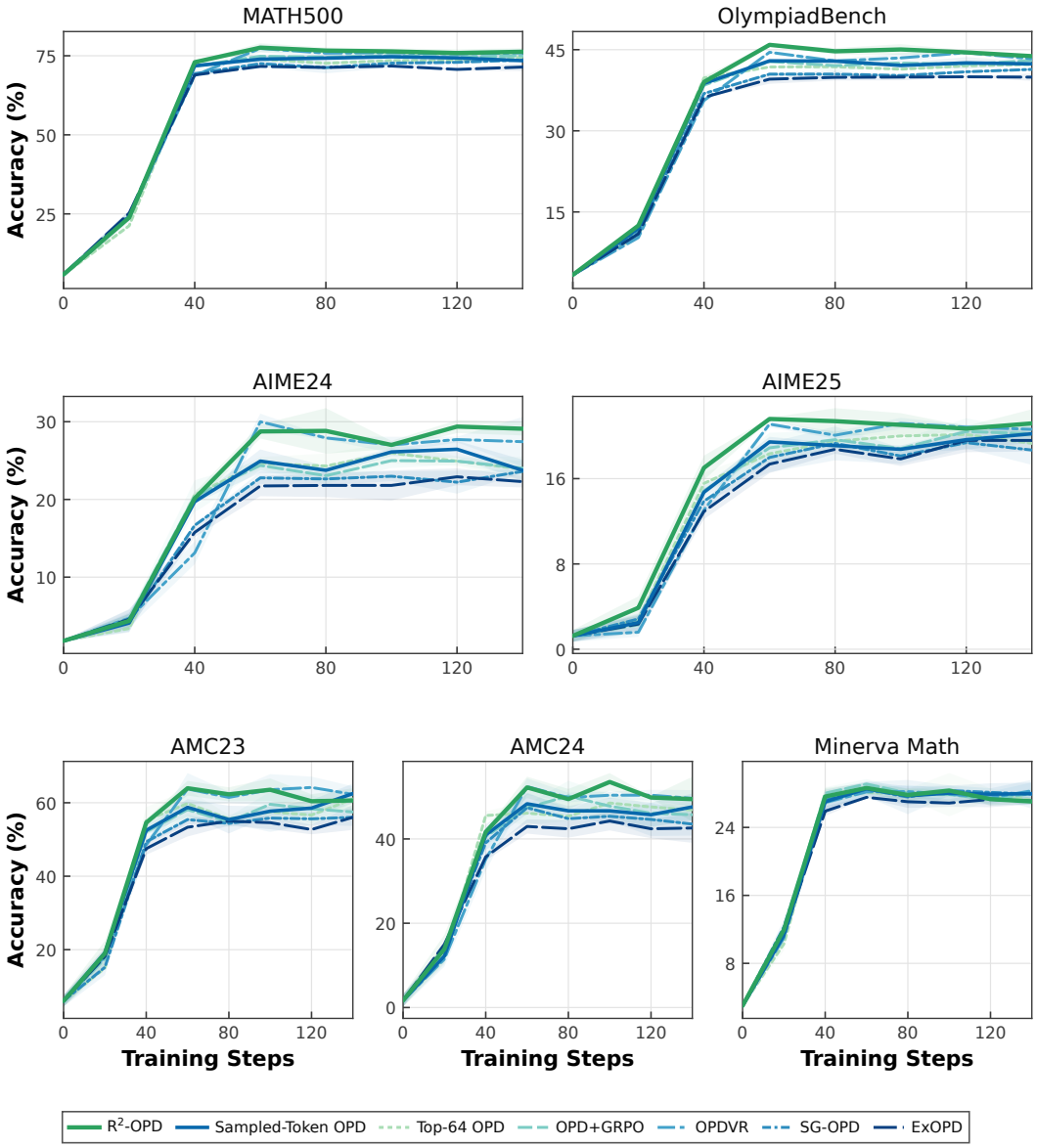



Figure 6: Evaluation accuracy across seven mathematical reasoning benchmarks during cross-size distillation from Qwen3-4B-GRPO to Qwen3-1.7B-Base. Curves show the mean and shaded bands indicate $\pm$ one sample standard deviation across three independent training seeds.

**Reweighting overhead.** For a batch containing $N$ valid response tokens, let $\mathcal{I}$ index complete positive-mass responses and $n_i$ denote their lengths. The additional coefficient computation has cost

$$C_{\mathrm{rw}} = \sum_{i \in \mathcal{I}} C_{\mathrm{med}}(n_i) + O(N), \tag{36}$$

where $C_{\mathrm{med}}$ denotes the implementation-dependent median cost; gate evaluation and mass normalization are linear in token count. Reweighting can be implemented with $O(N)$ auxiliary scalar storage, excluding median workspace, without additional vocabulary-sized tensors. All reweighting operations are detached from backpropagation.

**Verification and end-to-end cost.** Outcome verification adds task-dependent cost when the baseline does not already score the same rollouts; otherwise, verifier outputs can be reused. Online teacher scoring belongs to distillation cost, whereas teacher preparation and benchmark evaluation are separate costs. End-to-end runtime also depends on realized response lengths and implementation, so unchanged model-pass counts alone do not establish wall-clock equivalence or negligible overhead.

# F Gradient Analysis of Sampled-Token OPD

This appendix derives Eq. (5) and clarifies the coefficient-level meaning of reward alignment. We adopt the assumptions of Section 2 and assume the regularity required to interchange differentiation and expectation. Rollout distributions and sampled coefficients are fixed during differentiation. The analysis concerns the unclipped surrogate, not subsequent optimizer transformations.

## F.1 Statewise Reverse-KL Gradient

At a fixed prefix $s$, write

$$K_s(\theta) = \sum_{a \in \mathcal{V}} \pi_\theta(a \mid s) \log \frac{\pi_\theta(a \mid s)}{\pi_T(a \mid s)}. \tag{37}$$

Differentiation gives

$$\begin{aligned} \nabla_\theta K_s(\theta) &= \sum_a \nabla_\theta \pi_\theta(a \mid s) \log \frac{\pi_\theta(a \mid s)}{\pi_T(a \mid s)} + \sum_a \nabla_\theta \pi_\theta(a \mid s) \\ &= \mathbb{E}_{a \sim \pi_\theta(\cdot|s)} \left[ \log \frac{\pi_\theta(a \mid s)}{\pi_T(a \mid s)} \nabla_\theta \log \pi_\theta(a \mid s) \right], \end{aligned} \tag{38}$$

where $\sum_a \nabla_\theta \pi_\theta(a \mid s) = 0$ by normalization. Define $A_\theta(s,a) = \log \pi_T(a \mid s) - \log \pi_\theta(a \mid s)$, so that $A_t = A_{\bar{\theta}}(s_t, o_t)$ in Eq. (3). Then

$$\nabla_\theta K_s(\theta) = -\mathbb{E}_{a \sim \pi_\theta(\cdot|s)} [A_\theta(s,a) \nabla_\theta \log \pi_\theta(a \mid s)]. \tag{39}$$

Equation (39) gives an unbiased single-action gradient estimator. Separately,

$$\mathbb{E}_{a \sim \pi_\theta(\cdot|s)}[A_\theta(s,a)] = -K_s(\theta) \le 0. \tag{40}$$

A nonpositive mean coefficient does not imply uniform suppression: the update depends on its association with the action-dependent score function, whose expectation is zero.

## F.2 First-Order Equivalence of the Sampled Surrogate

Let $\mathbb{E}_{\bar{\theta}}$ denote expectation over $q \sim \mathcal{D}$ and $o \sim \pi_{\bar{\theta}}(\cdot \mid q)$. The main-text surrogate is

$$\mathcal{L}_{\mathrm{OPD}}^{\mathrm{sample}}(\theta; \bar{\theta}) = -\mathbb{E}_{\bar{\theta}} \left[ \sum_{t=1}^{|o|} \mathrm{sg}[A_t] \log \pi_\theta(o_t \mid s_t) \right]. \tag{41}$$

**Proposition 5** (Rollout-point gradient equivalence)**.** *With the rollout distribution fixed,*

$$\nabla_\theta \mathcal{L}_{\mathrm{OPD}}^{\mathrm{sample}}(\theta; \bar{\theta})\Big|_{\theta=\bar{\theta}} = \nabla_\theta \mathcal{L}_{\mathrm{OPD}}(\theta; \bar{\theta})\big|_{\theta=\bar{\theta}}. \tag{42}$$

*Proof.* Let $\psi_t = \nabla_\theta \log \pi_\theta(o_t \mid s_t)|_{\theta=\bar{\theta}}$. Since $A_t$ is detached, the left-hand side equals $-\mathbb{E}_{\bar{\theta}}[\sum_t A_t \psi_t]$. Conditioning on each visited prefix and applying Eq. (39) gives

$$-\mathbb{E}_{o_t \sim \pi_{\bar{\theta}}(\cdot|s_t)}[A_t \psi_t \mid s_t] = \nabla_\theta K_t(\theta)|_{\theta=\bar{\theta}}\,. \tag{43}$$

Summing over response positions and taking the rollout expectation therefore yields

$$\begin{aligned} -\mathbb{E}_{\bar{\theta}}\left[\sum_t A_t \psi_t\right] &= \mathbb{E}_{\bar{\theta}}\left[\sum_t \nabla_\theta K_t(\theta)|_{\theta=\bar{\theta}}\right] \\ &= \nabla_\theta \mathcal{L}_{\mathrm{OPD}}(\theta;\bar{\theta})\big|_{\theta=\bar{\theta}}\,, \end{aligned} \tag{44}$$

which proves Eq. (5). □

The identity establishes gradient equivalence at the rollout parameters, not equality of loss values or differentiation through state visitation. It also requires the sampling distribution to agree with the $\pi_{\bar{\theta}}$ used in $A_t$. For comparison, the importance-ratio surrogate uses $r_t(\theta) = \pi_\theta(o_t \mid s_t)/\pi_{\bar{\theta}}(o_t \mid s_t)$ in place of $\log \pi_\theta(o_t \mid s_t)$. It has the same rollout-point gradient because $\nabla_\theta r_t(\theta)|_{\theta=\bar{\theta}} = \psi_t$, but the two surrogates generally differ away from $\bar{\theta}$.

### F.3 Outcome-Aligned Coefficient Decomposition

For a fixed complete response, let $z = z(q, o) \in \{-1, +1\}$ and let $m_t \in \{0, 1\}$ mark valid response tokens. Define the aligned and inconsistent components

$$a_t^{\mathrm{ali}} = m_t A_t \mathbb{I}[zA_t > 0], \qquad a_t^{\mathrm{inc}} = m_t A_t \mathbb{I}[zA_t < 0]. \tag{45}$$

The two events partition all nonzero coefficients, giving

$$m_t A_t = a_t^{\mathrm{ali}} + a_t^{\mathrm{inc}}. \tag{46}$$

For $A_t \neq 0$, $\mathbb{I}[zA_t > 0] = (1 + z\,\mathrm{sgn}(A_t))/2$. Using $A_t\,\mathrm{sgn}(A_t) = |A_t|$ yields

$$a_t^{\mathrm{ali}} = \frac{m_t}{2}(A_t + z|A_t|), \qquad a_t^{\mathrm{inc}} = \frac{m_t}{2}(A_t - z|A_t|). \tag{47}$$

Both identities also hold when $A_t = 0$.

For $\ell_o(w;\theta) = -\sum_t \mathrm{sg}[w_t] \log \pi_\theta(o_t \mid s_t)$, linearity in detached coefficients gives

$$\begin{aligned} \ell_o(mA;\theta) &= \ell_o(a^{\mathrm{ali}};\theta) + \ell_o(a^{\mathrm{inc}};\theta), \\ \nabla_\theta \ell_o(mA;\theta) &= \nabla_\theta \ell_o(a^{\mathrm{ali}};\theta) + \nabla_\theta \ell_o(a^{\mathrm{inc}};\theta). \end{aligned} \tag{48}$$

These identities extend to batch averages under the same linear reduction, without separately normalizing the two components. This partition describes the dense coefficients; it is not the R$^2$-OPD update rule. At finite sharpness, R$^2$-OPD retains nonzero corrections in both groups through positive reweighting; truncated responses retain standard OPD.

### F.4 Token-Level Alignment and Its Scope

At $\bar{\theta}$, let $\psi_t = \nabla_\theta \log \pi_\theta(o_t \mid s_t)|_{\theta=\bar{\theta}}$ and define

$$d_t^{\mathrm{out}} = z\psi_t, \qquad d_t^{\mathrm{ali}} = a_t^{\mathrm{ali}} \psi_t, \qquad d_t^{\mathrm{inc}} = a_t^{\mathrm{inc}} \psi_t. \tag{49}$$

The coefficient partition implies

$$\begin{aligned} \langle d_t^{\mathrm{ali}}, d_t^{\mathrm{out}} \rangle &= m_t \mathbb{I}[zA_t > 0] |A_t| \|\psi_t\|_2^2 \geq 0, \\ \langle d_t^{\mathrm{inc}}, d_t^{\mathrm{out}} \rangle &= -m_t \mathbb{I}[zA_t < 0] |A_t| \|\psi_t\|_2^2 \leq 0. \end{aligned} \tag{50}$$

The inequalities are strict on the corresponding active component when $\psi_t \neq 0$. They characterize alignment with an outcome-signed direction for the same token, not with the expected task gradient $g_R$. Summing in the direct-sum space of token directions preserves these signs; aggregation in the shared parameter space introduces cross-token inner products and need not preserve them. Consequently, token-level outcome alignment alone does not imply reward ascent. The task-progress criterion in Eq. (21) additionally depends on correction utility.

**Trajectory centering.** For $N_o = \sum_t m_t > 0$, define $\bar{A}_o = N_o^{-1} \sum_t m_t A_t$ and $\widetilde{A}_t = A_t - \bar{A}_o$. Then $\sum_t m_t \widetilde{A}_t = 0$, so nonconstant valid advantages acquire both positive and negative centered values. Routing by $z\widetilde{A}_t$ therefore need not agree with routing by $zA_t$; R$^2$-OPD uses the latter. This response-level centering is also distinct from the conditional action centering $A_t - \mathbb{E}_s[A_t]$ in Eq. (12).

# G Teacher Preference and Student Continuation Value: Further Analysis

This appendix proves Proposition 1, relates its covariance criterion to over-correction and over-trust, and clarifies the outcome evidence motivating supervision allocation in Section 3.

## G.1 Setup and Normalized Local Perturbation

Fix a prefix $s = s_t = (q, o_{<t})$ and a finite action set $\mathcal{V}_s$ on which $\pi_{\bar{\theta}}(\cdot \mid s)$ and $\pi_T(\cdot \mid s)$ are normalized and strictly positive. A finite generation horizon and a common termination convention ensure that the binary reward is defined, so $Q_R^{\bar{\theta}}(s,a) \in [0,1]$. Only the next-action distribution at $s$ changes; the continuation policy remains $\pi_{\bar{\theta}}$, leaving $Q_R^{\bar{\theta}}(s,a)$ fixed. Unless stated otherwise, moments indexed by $s$ are taken over $o_t \sim \pi_{\bar{\theta}}(\cdot \mid s)$, with $A_t$ defined in Eq. (3).

For $0 \le \epsilon \le 1$, the normalized interpolation in Eq. (9) is

$$Z_s(\epsilon) = \sum_{a \in \mathcal{V}_s} \pi_{\bar{\theta}}(a \mid s)^{1-\epsilon} \pi_T(a \mid s)^{\epsilon}, \tag{51}$$

$$\pi_\epsilon(a \mid s) = \frac{\pi_{\bar{\theta}}(a \mid s)^{1-\epsilon} \pi_T(a \mid s)^{\epsilon}}{Z_s(\epsilon)}. \tag{52}$$

Thus $Z_s(0) = Z_s(1) = 1$, $\pi_0 = \pi_{\bar{\theta}}$, and $\pi_1 = \pi_T$ at the fixed prefix. Finiteness and strict positivity imply smoothness in a neighborhood of $\epsilon = 0$. Differentiating gives

$$\frac{\partial}{\partial \epsilon} \log \pi_\epsilon(a \mid s) = \log \frac{\pi_T(a \mid s)}{\pi_{\bar{\theta}}(a \mid s)} - \frac{Z_s'(\epsilon)}{Z_s(\epsilon)}. \tag{53}$$

At $\epsilon = 0$,

$$\frac{Z_s'(0)}{Z_s(0)} = \sum_a \pi_{\bar{\theta}}(a \mid s) \log \frac{\pi_T(a \mid s)}{\pi_{\bar{\theta}}(a \mid s)} = \mathbb{E}_s[A_t]. \tag{54}$$

Writing $\dot{\pi}_\epsilon(a \mid s) = \partial_\epsilon \pi_\epsilon(a \mid s)$, we obtain

$$\dot{\pi}_0(a \mid s) = \pi_{\bar{\theta}}(a \mid s) \left( \log \frac{\pi_T(a \mid s)}{\pi_{\bar{\theta}}(a \mid s)} - \mathbb{E}_s[A_t] \right), \tag{55}$$

with $\sum_a \dot{\pi}_0(a \mid s) = 0$. This normalized probability-space intervention need not coincide with gradient descent in shared model parameters.

## G.2 Proof of Proposition 1

*Proof.* The local reverse KL is

$$D_s(\epsilon) = \sum_a \pi_\epsilon(a \mid s) \log \frac{\pi_\epsilon(a \mid s)}{\pi_T(a \mid s)}.$$

Differentiating and using $\sum_a \dot{\pi}_\epsilon(a \mid s) = 0$ yields

$$\begin{aligned} D_s'(\epsilon) &= \sum_a \dot{\pi}_\epsilon(a \mid s) \log \frac{\pi_\epsilon(a \mid s)}{\pi_T(a \mid s)} + \sum_a \dot{\pi}_\epsilon(a \mid s) \\ &= \sum_a \dot{\pi}_\epsilon(a \mid s) \log \frac{\pi_\epsilon(a \mid s)}{\pi_T(a \mid s)}. \end{aligned} \tag{56}$$

At $\epsilon = 0$, the log ratio is $-A_t$ for $o_t = a$. Substituting Eq. (55) gives

$$\begin{aligned} D_s'(0) &= -\mathbb{E}_s[(A_t - \mathbb{E}_s[A_t])A_t] \\ &= -\operatorname{Var}_s[A_t], \end{aligned} \tag{57}$$

which proves Eq. (10).

Since the continuation values are fixed, differentiating $F_s(\epsilon) = \sum_a \pi_\epsilon(a \mid s) Q_R^{\bar{\theta}}(s, a)$ yields

$$\begin{aligned} F_s'(0) &= \sum_a \dot{\pi}_0(a \mid s) Q_R^{\bar{\theta}}(s, a) \\ &= \mathbb{E}_s\Big[(A_t - \mathbb{E}_s[A_t]) Q_R^{\bar{\theta}}(s, o_t)\Big] \\ &= \operatorname{Cov}_s\Big(A_t, Q_R^{\bar{\theta}}(s, o_t)\Big), \end{aligned} \tag{58}$$

proving Eq. (11).

Negative covariance implies that $A_t$ is nonconstant and hence $\operatorname{Var}_s[A_t] > 0$. Smoothness gives

$$D_s(\epsilon) - D_s(0) = -\epsilon \operatorname{Var}_s[A_t] + O(\epsilon^2), \tag{59}$$

$$F_s(\epsilon) - F_s(0) = \epsilon \operatorname{Cov}_s\Big(A_t, Q_R^{\bar{\theta}}(s, o_t)\Big) + O(\epsilon^2). \tag{60}$$

Both first-order coefficients are strictly negative, so there exists $\epsilon_0 > 0$ such that both quantities decrease for every $0 < \epsilon < \epsilon_0$. □

### G.3 Centered Preferences and the Two Failure Modes

Since $F_s(0) = \mathbb{E}_s[Q_R^{\bar{\theta}}(s, o_t)]$ and $\mathbb{E}_s[A_t - \mathbb{E}_s[A_t]] = 0$, the reward derivative also admits the centered form

$$F_s'(0) = \mathbb{E}_s\Big[(A_t - \mathbb{E}_s[A_t])(Q_R^{\bar{\theta}}(s, o_t) - F_s(0))\Big]. \tag{61}$$

Under the local interpolation, the value-based forms of the two failure modes are

$$\text{Over-correction:} \quad A_t - \mathbb{E}_s[A_t] < 0, \quad Q_R^{\bar{\theta}}(s, o_t) > F_s(0), \tag{62}$$

$$\text{Over-trust:} \quad A_t - \mathbb{E}_s[A_t] > 0, \quad Q_R^{\bar{\theta}}(s, o_t) < F_s(0). \tag{63}$$

The former shifts probability away from actions with above-average student value; the latter shifts it toward actions with below-average value. Both contribute negatively to Eq. (61), but the net derivative depends on the probability-weighted contributions of all actions.

The distinction between raw and centered preference is essential. The sign of $A_t$ describes an individual surrogate coefficient, whereas normalized probability changes depend on $A_t - \mathbb{E}_s[A_t]$. Because

$$\mathbb{E}_s[A_t] = -K_t(\bar{\theta}) \le 0, \tag{64}$$

a negative $A_t$ need not imply decreasing action probability along the interpolation. Neither sign alone determines an aggregate parameter update, which also depends on the remaining token contributions and parameter sharing.

### G.4 What Trajectory Outcomes Reveal

The signed outcome in Eq. (7) provides evidence about continuation quality. Conditional on the prefix and next action,

$$\mathbb{E}[z(q, o) \mid s_t = s, o_t = a] = 2Q_R^{\bar{\theta}}(s, a) - 1, \tag{65}$$

where the suffix follows $\pi_{\bar{\theta}}$. For moments involving $z(q, o)$, the conditional rollout law includes both the next action and the student-generated suffix. Since $A_t$ depends only on the fixed prefix and next action, the tower property gives

$$\begin{aligned} \operatorname{Cov}_s(A_t, z(q, o)) &= \mathbb{E}_s[(A_t - \mathbb{E}_s[A_t]) z(q, o)] \\ &= \mathbb{E}_s\Big[(A_t - \mathbb{E}_s[A_t])\big(2Q_R^{\bar{\theta}}(s, o_t) - 1\big)\Big] \\ &= 2 \operatorname{Cov}_s\Big(A_t, Q_R^{\bar{\theta}}(s, o_t)\Big) = 2F_s'(0). \end{aligned} \tag{66}$$

This expectation-level identity links outcome evidence to the local value criterion without identifying the utility of a single correction. In particular, the sign of $z(q,o)A_t$ need not match the sign of an action's contribution to Eq. (61): $z$ compares a sampled reward with a fixed threshold, not with $F_s(0)$, and contains continuation sampling noise. Excluding outcomes from the definition of $A_t$ therefore does not imply statistical independence between $A_t$ and $z(q,o)$.

R$^2$-OPD combines this outcome evidence with the sign and magnitude of $A_t$ to reallocate teacher supervision. Positive weights preserve coefficient signs and nonzero feedback, while per-response normalization conserves absolute coefficient mass. Proposition 1 motivates this allocation principle but does not identify optimal weights, establish the sigmoid as task-optimal, or guarantee positive reward change under reweighting. The method's allocation properties and empirical performance remain distinct from this local diagnostic.

### G.5 An Analytic Two-Route Counterexample

Consider a finite decision process in which two actions at a fixed prefix select routes $a_1$ and $a_2$, followed by a terminal success/failure action. The probabilities below specify an illustrative construction, not empirical measurements from a language model.

| Route | $\pi_{\bar{\theta}}(a \mid s)$ | $\pi_T(a \mid s)$ | $Q_R^{\bar{\theta}}(s,a)$ |
|---|---|---|---|
| $a_1$ | 0.2 | 0.8 | 0.1 |
| $a_2$ | 0.8 | 0.2 | 0.8 |

The final column gives the student's continuation success probabilities. Let the teacher succeed with probabilities $0.95$ and $0.90$ after selecting $a_1$ and $a_2$, respectively. Its total success probability is $0.8(0.95) + 0.2(0.90) = 0.94$, whereas the student's is

$$F_s(0) = 0.2(0.1) + 0.8(0.8) = 0.66.$$

The distillation advantages are $\log 4$ and $-\log 4$, with mean $-0.6 \log 4$ and centered values $1.6 \log 4$ and $-0.4 \log 4$. The corresponding student values relative to $F_s(0)$ are $-0.56$ and $0.14$. Thus, $a_1$ illustrates over-trust and $a_2$ over-correction in the value-based sense above, with

$$D_s'(0) = -0.64(\log 4)^2 < 0, \tag{67}$$

$$F_s'(0) = 0.2(0.8)(2 \log 4)(0.1 - 0.8) = -0.224 \log 4 < 0. \tag{68}$$

At $\epsilon = 1$, route selection exactly matches the teacher, so $D_s(1) = 0$, compared with $D_s(0) = 0.6 \log 4$. Keeping the student's continuation policy fixed nevertheless gives

$$F_s(1) = 0.8(0.1) + 0.2(0.8) = 0.24.$$

The example isolates partial imitation: adopting the teacher's route selection does not transfer its ability to complete those routes. It establishes the possibility of the preference–value mismatch, not its prevalence in language models or a guarantee that R$^2$-OPD resolves this particular decision process.

## H Core Proofs and Their Scope

For the coefficient-space results, fix a nonempty complete response $o$, its finite advantages $A_t$, outcome $z = z(q,o) \in \{-1, +1\}$, and scale $\nu_o > 0$. These quantities remain fixed as $\beta$ varies. Unless stated otherwise, $M_o = \sum_t |A_t| > 0$ and $0 \le \beta < \infty$. All sums are over valid response tokens, and normalization is exact.

### H.1 Proof of Proposition 2

*Proof.* Substituting the sigmoid gate into Eq. (15) gives

$$A_t^{\mathrm{R2}} = Z_o g_t A_t = Z_o A_t \sigma\left(\frac{\beta z A_t}{\nu_o}\right). \tag{69}$$

Using the identity

$$\sigma(x) = \frac{e^x}{1+e^x} = \frac{1}{2}\left(1 + \tanh\frac{x}{2}\right), \tag{70}$$

we obtain

$$A_t^{\mathrm{R2}} = \frac{Z_o}{2}A_t + \frac{Z_o}{2}A_t \tanh\left(\frac{\beta z A_t}{2\nu_o}\right). \tag{71}$$

For $A_t \neq 0$, oddness of $\tanh$ and $A_t \operatorname{sgn}(A_t) = |A_t|$ imply

$$A_t \tanh\left(\frac{\beta z A_t}{2\nu_o}\right) = z|A_t| \tanh\left(\frac{\beta |A_t|}{2\nu_o}\right). \tag{72}$$

Both sides vanish when $A_t = 0$. Substitution into Eq. (71) proves Eq. (17). The identity holds pointwise in $\beta$, including the dependence of $Z_o$ on $\beta$; no small-sharpness approximation is required. □

For nonzero $A_t$, the magnitude ratio of Term B to Term A is $\tanh(\beta|A_t|/(2\nu_o)) < 1$, so the coefficient retains its sign. This does not determine the token's probability change after a shared-parameter batch update.

### H.2 Proof of Proposition 3

*Proof.* Write $b_t = |A_t|$, $a_t = zA_t/\nu_o$, and $D_o = \sum_t g_t b_t$, so $g_t = \sigma(\beta a_t)$ and $Z_o = M_o/D_o$.

**Sign and mass preservation.** For finite $\beta$, $0 < g_t < 1$ implies $D_o > 0$ and $Z_o > 0$. Thus, multiplication by $Z_o g_t$ preserves nonzero coefficients and their signs, with

$$\sum_t |A_t^{\mathrm{R2}}| = Z_o \sum_t g_t b_t = \frac{M_o}{D_o} D_o = M_o. \tag{73}$$

**Monotone reallocation.** Let $\mathcal{A}_o = \{t : a_t > 0\}$ and $\mathcal{C}_o = \{t : a_t < 0\}$; zero-advantage positions carry no mass. Define the unnormalized group masses

$$B_{\mathrm{A}} = \sum_{t\in\mathcal{A}_o} b_t g_t, \qquad B_{\mathrm{C}} = \sum_{t\in\mathcal{C}_o} b_t g_t. \tag{74}$$

The common normalizer cancels in the inconsistent mass share:

$$\kappa_o(\beta) = \frac{Z_o B_{\mathrm{C}}}{Z_o(B_{\mathrm{A}} + B_{\mathrm{C}})} = \frac{B_{\mathrm{C}}}{B_{\mathrm{A}} + B_{\mathrm{C}}}. \tag{75}$$

With $a_t$ fixed,

$$g_t'(\beta) = a_t g_t (1 - g_t). \tag{76}$$

If both groups carry positive mass, then $B_{\mathrm{A}}, B_{\mathrm{C}} > 0$, $B_{\mathrm{A}}' > 0$, and $B_{\mathrm{C}}' < 0$. The quotient rule therefore yields

$$\kappa_o'(\beta) = \frac{B_{\mathrm{C}}' B_{\mathrm{A}} - B_{\mathrm{C}} B_{\mathrm{A}}'}{(B_{\mathrm{A}} + B_{\mathrm{C}})^2} < 0. \tag{77}$$

At $\beta = 0$, the derivative is right-sided. Since total mass remains $M_o$, inconsistent mass $M_o\kappa_o$ decreases while aligned mass $M_o(1-\kappa_o)$ increases.

**Endpoints.** At $\beta = 0$, $g_t = 1/2$ and $Z_o = 2$, giving $A_t^{\mathrm{R2}} = A_t$. As $\beta \to \infty$, $g_t$ tends to one on $\mathcal{A}_o$ and zero on $\mathcal{C}_o$. If $M_o^{\mathrm{ali}} = \sum_{t\in\mathcal{A}_o} b_t > 0$, then $D_o \to M_o^{\mathrm{ali}}$ and $Z_o \to M_o/M_o^{\mathrm{ali}}$. Hence

$$\lim_{\beta\to\infty} A_t^{\mathrm{R2}} = \frac{M_o}{M_o^{\mathrm{ali}}} A_t \mathbb{I}[t \in \mathcal{A}_o]. \tag{78}$$

Zero advantages remain zero at both endpoints. □

**Boundary cases and scope.** If all nonzero advantages are aligned or all are inconsistent, $\kappa_o$ is identically zero or one, respectively. The all-inconsistent case is excluded from Eq. (78): for $z = 1$ and $A = (-1, -1)$, identical gates give $Z_o = 1/g$, so the coefficients remain $(-1, -1)$ for every finite $\beta$. For $M_o = 0$, the coefficients are zero and $\kappa_o$ is undefined. Monotonicity applies to group mass, not individual coefficients, and the hard endpoint is not generally equivalent to unnormalized masking. Recovering the full vanilla update at $\beta = 0$ additionally requires identical data, masks, loss reduction, optimizer state, and settings.

### H.3 Proof of Proposition 4

Assume $J$ is differentiable at $\bar{\theta}$ and define $g_R = \nabla J(\bar{\theta})$ and $\psi_t = \nabla_\theta \log \pi_\theta(o_t \mid s_t)|_{\theta=\bar{\theta}}$. Rollout expectations are finite and admit differentiation under the expectation, with the rollout distribution held fixed. Differentiating the detached log-probability surrogate in Eq. (16) gives

$$-\nabla_\theta \mathcal{L}_{\mathrm{R2}}(\theta; \bar{\theta})\big|_{\theta=\bar{\theta}} = \mathbb{E}\left[\sum_t A_t^{\mathrm{R2}} \psi_t\right] = \mathbb{E}[d_{\mathrm{R2}}(o)]. \tag{79}$$

Replacing $A_t^{\mathrm{R2}}$ by $A_t$ gives $\mathbb{E}[d_{\mathrm{V}}(o)]$.

For comparison, the probability ratio $\rho_t(\theta) = \pi_\theta(o_t \mid s_t)/\pi_{\bar{\theta}}(o_t \mid s_t)$ satisfies

$$\nabla_\theta \rho_t(\theta) = \rho_t(\theta) \nabla_\theta \log \pi_\theta(o_t \mid s_t), \qquad \rho_t(\bar{\theta}) = 1. \tag{80}$$

Replacing the log probability by this ratio gives the same gradient at $\bar{\theta}$, but not generally away from the rollout parameters.

*Proof.* For the fixed complete response, recall $p_o(t) = |A_t|/M_o$ and $u_t = \mathrm{sgn}(A_t)\langle g_R, \psi_t \rangle$. Here $p_o$ is a distribution over token positions. Its mean gate satisfies

$$\mu_g := \mathbb{E}_{p_o}[g] = \frac{\sum_t |A_t| g_t}{M_o} = \frac{1}{Z_o} > 0. \tag{81}$$

Using $A_t = |A_t| \,\mathrm{sgn}(A_t)$, the standard OPD projection is

$$\langle g_R, d_{\mathrm{V}}(o) \rangle = \sum_t A_t \langle g_R, \psi_t \rangle = M_o \mathbb{E}_{p_o}[u]. \tag{82}$$

Similarly,

$$\langle g_R, d_{\mathrm{R2}}(o) \rangle = Z_o \sum_t g_t |A_t| u_t = \frac{M_o \mathbb{E}_{p_o}[gu]}{\mu_g}. \tag{83}$$

Subtracting yields

$$\begin{aligned} \langle g_R, d_{\mathrm{R2}}(o) - d_{\mathrm{V}}(o) \rangle &= \frac{M_o}{\mu_g} \left( \mathbb{E}_{p_o}[gu] - \mu_g \mathbb{E}_{p_o}[u] \right) \\ &= \frac{M_o \,\mathrm{Cov}_{p_o}(g, u)}{\mathbb{E}_{p_o}[g]}, \end{aligned} \tag{84}$$

which proves Eq. (21). Zero-advantage positions have zero $p_o$ mass. □

**A sufficient reward–utility agreement condition.** Suppose $\mathcal{A}_o$ and $\mathcal{C}_o$ both carry positive mass, and let $\alpha = p_o(\mathcal{A}_o) \in (0, 1)$. For $G \in \{\mathcal{A}_o, \mathcal{C}_o\}$, define $\mu_{g,G} = \mathbb{E}_{p_o}[g \mid G]$, $\mu_{u,G} = \mathbb{E}_{p_o}[u \mid G]$, and $c_G = \mathrm{Cov}_{p_o}(g, u \mid G)$. Conditioning on the two groups gives

$$\begin{aligned} \mathbb{E}_{p_o}[gu] &= \alpha(c_{\mathcal{A}} + \mu_{g,\mathcal{A}} \mu_{u,\mathcal{A}}) + (1-\alpha)(c_{\mathcal{C}} + \mu_{g,\mathcal{C}} \mu_{u,\mathcal{C}}), \\ \mathbb{E}_{p_o}[g] &= \alpha \mu_{g,\mathcal{A}} + (1-\alpha) \mu_{g,\mathcal{C}}, \\ \mathbb{E}_{p_o}[u] &= \alpha \mu_{u,\mathcal{A}} + (1-\alpha) \mu_{u,\mathcal{C}}. \end{aligned}$$

Subtracting the product of the last two expressions from the first yields

$$\begin{aligned} \mathrm{Cov}_{p_o}(g, u) = \alpha c_{\mathcal{A}} + (1-\alpha) c_{\mathcal{C}} \\ + \alpha(1-\alpha)(\mu_{g,\mathcal{A}} - \mu_{g,\mathcal{C}})(\mu_{u,\mathcal{A}} - \mu_{u,\mathcal{C}}). \end{aligned} \tag{85}$$

For $\beta > 0$, the gate satisfies $\mu_{g,\mathcal{A}} > 1/2 > \mu_{g,\mathcal{C}}$. Consequently, $\mu_{u,\mathcal{A}} > \mu_{u,\mathcal{C}}$ and $\alpha c_{\mathcal{A}} + (1-\alpha) c_{\mathcal{C}} \geq 0$ suffice for a strictly positive per-response projection gain. More generally, the between-group term must outweigh any negative within-group contribution. These are additional utility conditions, not consequences of outcome alignment alone.

### H.4 Relative Progress and Reward Ascent

Let $d_{\mathrm{V}} = \mathbb{E}[d_{\mathrm{V}}(o)]$ and $d_{\mathrm{R2}} = \mathbb{E}[d_{\mathrm{R2}}(o)]$ use the same rollout distribution and reduction, and define

$$\Delta_R = \langle g_R, d_{\mathrm{R2}} - d_{\mathrm{V}} \rangle, \qquad P_{\mathrm{V}} = \langle g_R, d_{\mathrm{V}} \rangle. \tag{86}$$

By linearity, $\Delta_R = \mathbb{E}[\Delta_o]$, where $\Delta_o$ is the right-hand side of Eq. (21) for complete positive-mass responses and zero for zero-mass or identity-fallback responses.

Suppose $\nabla J$ is $L_J$-Lipschitz, with $L_J > 0$, on a neighborhood containing both update segments. For either fixed direction $d$ and step size $\eta > 0$,

$$J(\bar{\theta} + \eta d) - J(\bar{\theta}) - \eta \langle g_R, d \rangle = \int_0^{\eta} \langle \nabla J(\bar{\theta} + td) - g_R, d \rangle \, dt. \tag{87}$$

Cauchy–Schwarz and Lipschitz continuity bound the absolute remainder by $L_J \eta^2 \|d\|_2^2/2$. Applying the lower bound to $d_{\mathrm{R2}}$ and the upper bound to $d_{\mathrm{V}}$ gives

$$\begin{aligned} &J(\bar{\theta} + \eta d_{\mathrm{R2}}) - J(\bar{\theta} + \eta d_{\mathrm{V}}) \\ &\quad \geq \eta \Delta_R - \frac{L_J \eta^2}{2} \left( \|d_{\mathrm{R2}}\|_2^2 + \|d_{\mathrm{V}}\|_2^2 \right). \end{aligned} \tag{88}$$

If $\Delta_R > 0$, relative reward improvement follows whenever both update segments remain in the stated neighborhood and

$$0 < \eta < \frac{2\Delta_R}{L_J \left( \|d_{\mathrm{R2}}\|_2^2 + \|d_{\mathrm{V}}\|_2^2 \right)}.$$

Reward ascent from the original policy is a separate condition. Since $\langle g_R, d_{\mathrm{R2}} \rangle = P_{\mathrm{V}} + \Delta_R$,

$$J(\bar{\theta} + \eta d_{\mathrm{R2}}) - J(\bar{\theta}) \geq \eta (P_{\mathrm{V}} + \Delta_R) - \frac{L_J \eta^2}{2} \|d_{\mathrm{R2}}\|_2^2. \tag{89}$$

If $P_{\mathrm{V}} + \Delta_R > 0$, this guarantees local ascent for

$$0 < \eta < \frac{2(P_{\mathrm{V}} + \Delta_R)}{L_J \|d_{\mathrm{R2}}\|_2^2}.$$

In particular, overcoming a negative baseline projection $P_{\mathrm{V}} < 0$ at first order requires $\Delta_R > -P_{\mathrm{V}}$, not merely $\Delta_R > 0$. These bounds concern the specified deterministic directions, not arbitrary stochastic or adaptive-optimizer updates.

### H.5 Gradient Compatibility and Task Progress

On a fixed batch of complete responses, define the aligned OPD loss by restricting Eq. (4) to $z(q, o)A_t > 0$. Let $G_{\mathrm{ali}} = \nabla \mathcal{L}_{\mathrm{ali}}(\bar{\theta})$ and $G_{\mathrm{R2}} = \nabla \mathcal{L}_{\mathrm{R2}}(\bar{\theta})$ use the same reduction. If $\mathcal{L}_{\mathrm{ali}}$ is $L_{\mathrm{ali}}$-smooth along the update segment, then

$$\begin{aligned} &\mathcal{L}_{\mathrm{ali}}(\bar{\theta} - \eta G_{\mathrm{R2}}) - \mathcal{L}_{\mathrm{ali}}(\bar{\theta}) \\ &\quad \leq -\eta \langle G_{\mathrm{ali}}, G_{\mathrm{R2}} \rangle + \frac{L_{\mathrm{ali}} \eta^2}{2} \|G_{\mathrm{R2}}\|_2^2. \end{aligned} \tag{90}$$

A positive inner product therefore permits aligned-surrogate descent for a sufficiently small step, but does not establish reward ascent. Likewise, coefficient-mass preservation does not fix update magnitude:

$$\|d_{\mathrm{R2}}(o)\|_2^2 = \sum_{i,j} A_i^{\mathrm{R2}} A_j^{\mathrm{R2}} \langle \psi_i, \psi_j \rangle. \tag{91}$$

The cross-token inner products determine how gradient contributions reinforce or cancel one another.

## I Implementation Conventions

**Fallback and masking.** Responses terminated by the length limit retain $A_t^{\mathrm{R2}} = A_t$; their outcomes are not used for reweighting. This fallback preserves standard OPD supervision rather than discarding responses. Prompt and padding positions are excluded from the scale, mass, and loss; empty responses contribute no token loss. Complete zero-mass responses retain zero coefficients with $Z_o = 1$.

**Relation to the analyzed surrogate.** For a fixed batch, all coefficients, masks, response lengths, and snapshot log probabilities are detached. Under consistent trainer evaluations, $\rho_{i,t}(\bar{\theta}) = 1$ and $\nabla_\theta \rho_{i,t}(\theta)|_{\bar{\theta}} = \psi_{i,t}$. Since the ratio lies inside the clipping interval, Eq. (22) has the rollout-point gradient

$$-\nabla_\theta \widehat{\mathcal{L}}_{\text{impl}}\Big|_{\bar{\theta}} = \frac{1}{N}\sum_{i=1}^{N}\frac{1}{\widetilde{n}_i}\sum_t m_{i,t} A^{\text{R2}}_{i,t}\psi_{i,t}. \tag{92}$$

This equals the gradient direction of the detached log-probability surrogate with the same per-response reduction. The equivalence is exact under the stated forward-evaluation assumption; separate floating-point passes may introduce numerical deviations.

For $X \in \{\text{V}, \text{R2}\}$, let $\widehat{d}^{\text{resp}}_X = N^{-1}\sum_i d_X(o_i)/\widetilde{n}_i$, with the per-response directions defined in Section 4.4 and zero direction for empty responses. Applying Eq. (21) response by response gives

$$\left\langle g_R, \widehat{d}^{\text{resp}}_{\text{R2}} - \widehat{d}^{\text{resp}}_{\text{V}} \right\rangle = \frac{1}{N}\sum_{i\in\mathcal{I}}\frac{M_{o_i}}{n_i}\frac{\text{Cov}_{p_{o_i}}(g,u)}{\mathbb{E}_{p_{o_i}}[g]}, \tag{93}$$

where $\mathcal{I}$ indexes complete, positive-mass responses; identity-fallback responses have zero difference. Thus, response averaging preserves the within-response allocation identity while introducing inverse-length weights in batch aggregation. Population directions accordingly use $\mathbb{E}[d_X(o)/\widetilde{n}_o]$ rather than $\mathbb{E}[d_X(o)]$.

The reverse-KL gradient identity in Eq. (5) concerns the token-summed analysis surrogate. It does not automatically extend to inverse-length weighting, since realized response length depends on sampled actions. The coefficient identities and the fixed-batch gradient relation above remain valid independently of that expectation-level identity. These relations describe the policy-loss gradient before global norm clipping and AdamW; finite-step guarantees for direct gradient updates are not guarantees for arbitrary adaptive-optimizer displacements.

**Exact normalization.** The scale floor in $\nu_o$ preserves the coefficient identities, whereas adding a positive constant to the mass denominator does not. For $M_o > 0$, $D_o = \sum_t g_t|A_t|$, and $Z_o = M_o/(D_o + \epsilon_{\text{den}})$ with $\epsilon_{\text{den}} > 0$,

$$\sum_t |A^{\text{R2}}_t| = \frac{M_o}{D_o + \epsilon_{\text{den}}}\sum_t g_t|A_t| = \frac{M_o D_o}{D_o + \epsilon_{\text{den}}} < M_o.$$

This modification breaks exact mass conservation and the vanilla endpoint at $\beta = 0$. Strict gate positivity holds in exact arithmetic; floating-point saturation and underflow require separate numerical care.