FROM COVERAGE TO RELIABLE SUCCESS: RESAMPLED-EVIDENCE POLICY OPTIMIZATION FOR TEST-TIME REINFORCEMENT LEARNING

Anonymous authors
Paper under double-blind review

ABSTRACT

We study how label-free reinforcement learning can improve sampled accuracy (pass@1) while retaining broad solution coverage (pass@16) and repeated correctness. The first two metrics are our primary outcomes; pass^4, the probability that four independent responses to the same question are all correct, provides a supporting reliability assessment. A central challenge is that finite verification evidence can support different responses unequally, even after a supervision target has been selected. We propose Resampled-Evidence Policy Optimization (REPO), which retains coverage-oriented training while adapting each response's contribution to its evidence support. The student and teacher maintain separate parameter states; the teacher supplies judgments and is updated only by an exponential moving average (EMA) of the student parameters. By resampling cached verification judgments and reconstructing supervision on fixed student responses, REPO attenuates less reproducible decisions without additional verification decoding. For each response, agreement and conflict between replayed and original advantage signs determine a bounded update weight. This weight preserves the original direction while reducing the influence of decisions sensitive to the sampled evidence. Across five external benchmarks, REPO improves macro-averaged pass@1 over Co-rewarding by 2.52 and 1.60 percentage points at 4B and 8B, with pass^4 gains of 2.12 and 2.18 points. Macro-averaged pass@16 remains close to Base (+0.18/+1.06 points), despite larger gaps to Co-rewarding (+11.50/+12.00 points). These results indicate aggregate coverage retention, with losses on some individual benchmarks. Code is available at https://anonymous.4open.science/r/REPO-4F30.

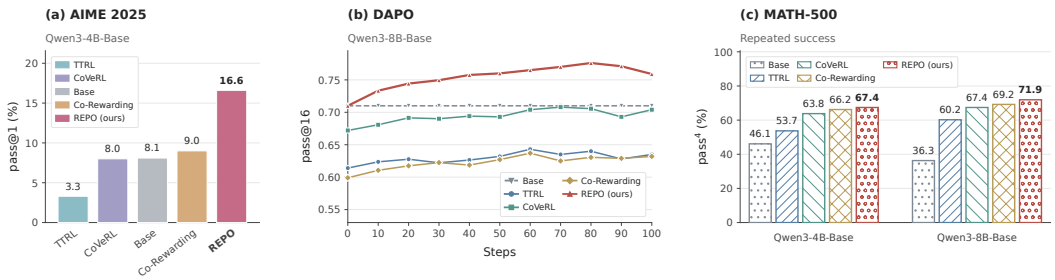


Figure 1: **Performance overview.** (a) AIME25 pass@1 at 4B. (b) DAPO pass@16 at 8B over training steps; the dashed line marks Base. (c) MATH-500 pass^4 at 4B and 8B.

1 INTRODUCTION

Test-time reinforcement learning (TTRL) adapts language models using unlabeled reasoning questions (Zuo et al., 2025). Unlike reinforcement learning with verifiable rewards (Shao et al., 2024; Yu et al., 2025a), it constructs supervision from model responses or judgments, making feedback quality central to adaptation (Zhang et al., 2025d;c). When no answer key is available during adaptation, the same mechanism that generates useful supervision can also reinforce systematic mistakes or ambiguous pseudo-targets. Complementary supervision and joint generation–verification improve this feedback (Zhang et al., 2025d; Pan et al., 2026b), while coverage-oriented objectives and entropy-based adaptation encourage alternative solutions (Chen et al., 2025; Cui et al., 2025; Liu et al., 2025a).

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.

These advances motivate better average accuracy and a greater chance of finding a correct answer within a decoding budget. Under answer-agreement rewards, a selected pseudo-target supervises a group of sampled responses, reinforcing matching answers and penalizing alternatives. Ambiguity in that target therefore propagates into multiple learning decisions. Dependable reasoning also requires repeated correctness. We therefore ask: *can label-free adaptation improve accuracy while retaining broad coverage and repeated-success reliability?*

We prioritize sampled accuracy ($\mathrm{pass@1}$) and solution coverage ($\mathrm{pass@16}$), which credits at least one correct answer within sixteen attempts (Chen et al., 2021). Repeated success (pass^4) provides a supporting reliability check: four independent responses to the same question must all be correct. It is the all-correct case of G-Pass@k for mathematical reasoning (Liu et al., 2024), also used in τ-bench (Yao et al., 2024).[1] Coverage can reward an occasional success, while mean accuracy does not fully describe the distribution of correctness across questions. Reporting these measures together separates finding a successful solution from producing correct responses consistently, making the practical trade-off visible rather than inferring it from one aggregate score. A policy that occasionally solves more questions may still answer previously solvable questions less consistently. We therefore evaluate the three metrics on the same sampled responses, so the reliability assessment accompanies the primary performance comparison. We seek accuracy gains while retaining broad coverage and repeated success, as illustrated by Figure 1a,c.

Achieving this balance requires examining how uncertain feedback becomes a policy update. Question-level acceptance assigns a target, but finite verification evidence may support the resulting response-level directions unequally. Uncertainty between two targets can reverse their supervision while leaving a third answer consistently penalized. A single question-level weight treats these contributions alike and may weaken supported penalties together with disputed reinforcement. The relevant evidence should therefore be assessed after target selection and advantage computation. Otherwise, confidence in a candidate need not describe the stability of the update direction assigned to each response. We assess uncertainty for each response's signed update, while retaining question-level selection as the decision about whether supervision is admitted. Section 2 examines these signals retrospectively within one REPO run, providing diagnostic context rather than independent evidence for weighting. Reproducible evidence need not be correct.

We propose Resampled-Evidence Policy Optimization (REPO), which retains coverage-oriented training and weights each response by its evidence support. The student produces reasoning responses and receives policy-gradient updates; a separately maintained teacher supplies verification judgments. To smooth the verifier's evolution as the student changes, we update the teacher with an exponential moving average (EMA) of the student parameters. Averaging successive parameter states limits abrupt changes to the verifier while allowing it to track the adapting student. Resampling cached teacher judgments reconstructs targets and advantages on fixed student responses, attenuating less reproducible directions without additional verification decoding. Each replay passes the resampled judgments through the same target selector and complete advantage computation. Agreement and conflict with the original advantage sign set a bounded multiplier, applied without changing the original target or reversing any update direction. This separates the decision to admit a question from the strength of its individual update signals, preserving supported contributions within otherwise ambiguous groups. Experiments on two Qwen3 backbones (Yang et al., 2025) assess five external benchmarks and a separate DAPO diagnostic. The full method improves external-macro accuracy and repeated success over Co-rewarding while retaining coverage close to Base on average. Component analyses assess evidence weighting and response-level allocation within a fixed training configuration.

Our contributions are threefold: (1) We frame label-free adaptation around improving $\mathrm{pass@1}$ while retaining broad $\mathrm{pass@16}$, using pass^4 as a supporting reliability check. **(2)** We introduce REPO to allocate response-level policy updates by replaying cached verification evidence, while retaining coverage-oriented training. **(3)** Experiments at 4B and 8B show external-macro accuracy gains with coverage largely retained relative to Base; component controls assess weighting within the shared training configuration.

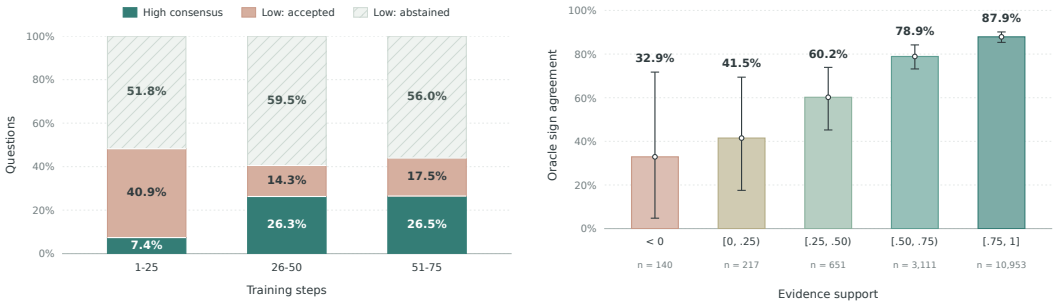
[1] The superscript denotes joint success, not $\mathrm{pass@4}$ or the fourth power of dataset-averaged $\mathrm{pass@1}$. Appendix E.2 gives the estimators.

2 MOTIVATION

Figure 2 retrospectively audits one completed REPO training trajectory. It characterizes supervision encountered under REPO, rather than independently testing whether weighting is needed. A matched *w/o EW* or baseline trajectory is not included. Ground-truth answers are used only for this diagnostic analysis, not to construct the training rewards.

2.1 QUESTION SELECTION IN THE OBSERVED TRAJECTORY

In this trajectory, low-consensus questions remain the majority throughout adaptation: their prevalence declines early and then stabilizes (Figure 2(a)). The share of all encountered questions that are low-consensus and accepted for updates also drops after the early phase and remains below its initial level, while a substantial fraction leads to abstention. These outcomes do not establish answer correctness: accepting a question assigns a training target without ensuring that every resulting response-level update is equally trustworthy. We next examine the supervision among accepted questions.



(a) Question composition during training

(b) Evidence support and update reliability

Figure 2: **Supervision diagnostics on a REPO trajectory.** (a) Low-consensus questions remain prevalent, while their update acceptance changes across training phases. (b) Among accepted low-consensus questions, higher evidence support is associated with greater agreement between the original and oracle advantage signs.

2.2 EVIDENCE SUPPORT AND ORACLE-SIGN AGREEMENT

Within accepted low-consensus questions, each response's advantage sign determines whether the policy is encouraged to increase or decrease its likelihood. We measure evidence support by resampling recorded teacher judgments and checking support for the original update direction. To assess reliability, we compute oracle advantages from reference-answer rewards using the same advantage computation, then compare signs after excluding responses for which either advantage is zero. Higher evidence support is associated with greater oracle sign agreement (Figure 2(b)), revealing variation in update reliability even after question-level acceptance. This diagnostic concerns update directions rather than answer accuracy, and does not establish a calibrated correctness probability. This association does not show that weighting caused the agreement pattern or that the same pattern holds without EW.

3 RESAMPLED-EVIDENCE POLICY OPTIMIZATION

REPO allocates policy updates according to the support that finite verification evidence provides for each response's update direction. Its central computation is to perturb the recorded evidence, propagate each perturbation through supervision, and measure which reinforcement or penalty decisions survive. The original target supplies the actual training signal; replay determines the strength of its individual contributions. Figure 3 summarizes this computation.

3.1 PROPAGATING EVIDENCE UNCERTAINTY TO UPDATE DIRECTIONS

Let π_θ and π_ϕ denote the student and EMA-teacher models, with parameters θ and ϕ. For an unlabeled question x, retain N_s student responses $Y = (y_i)_{i=1}^{N_s}$, where response y_i has extracted answer a_i. The candidate-answer set $\mathcal{A}_x$ is selected by voting frequency, and $n_x(a)$ counts voting responses with answer a. Each candidate a has J cached teacher judgments $V_a = (v_{aj})_{j=1}^{J}$, where

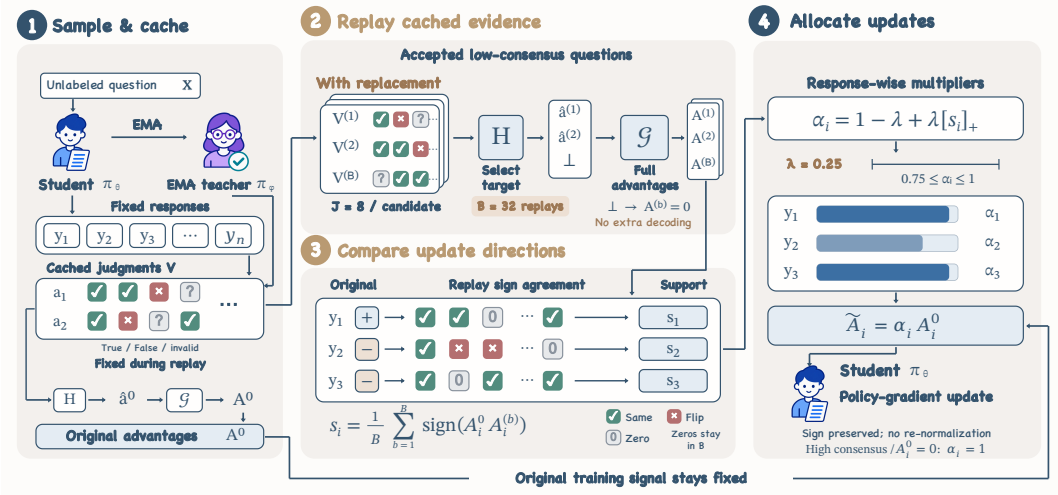


Figure 3: **Overview of REPO. Step 1:** Cache fixed student responses and teacher judgments. **Step 2:** Resample judgments in B replays and recompute targets and advantages without extra decoding. **Step 3:** Compute each response's net sign support s_i. **Step 4:** Map s_i to a bounded multiplier $\alpha_i \in [1-\lambda, 1]$ that rescales but never flips the original advantage.

v_{aj} is the jth verdict, parsed as True, False, or invalid. Write $V = (V_a)_{a \in \mathcal{A}_x}$ for the full cache. The selector $H(V; n_x)$ uses this cache and the voting counts to return a target answer or $\perp$, the abstention outcome. Section 3.4 specifies candidate screening and target selection.

We write $\mathbb{I}[\cdot]$ for the indicator function: it equals 1 when the bracketed condition holds. For a target answer a, the reward vector is $r(a) = (\mathbb{I}[a_i = a])_{i=1}^{N_s}$, so matching responses receive 1 and the others receive 0. Given the originally selected target $\hat{a}^0$, the advantage vector is $A^0 = \mathcal{G}(Y, r(\hat{a}^0), C(x))$. The superscript 0 marks quantities obtained from the original cache. Here $C(x)$ is the fraction of voting responses sharing the most frequent answer. $\mathcal{G}$ is the pass@K advantage operator (Chen et al., 2025): it compares each response's expected pass@K reward over size-K subsets containing it with the group's pass@K rate, normalized by the group standard deviation, so matching responses receive a positive advantage and nonmatching ones a negative advantage. When the standard deviation falls below a tolerance, as in groups where all or no responses match the target, $\mathcal{G}$ returns zero advantages. Its ith output A_i^0 sets the direction and magnitude of response i's update; a zero advantage contributes no policy signal and keeps unit weight.

For an accepted low-consensus question, replay keeps the responses, candidates, voting counts, consensus route, and teacher snapshot fixed. Only the cached judgments are resampled. Let $\widehat{P}_a(v)$ be the empirical probability of verdict $v \in \{\text{True}, \text{False}, \text{invalid}\}$ among candidate a's J judgments. In each of B replays, indexed by $b = 1, \ldots, B$, draw J verdicts from these observed frequencies:

$$\widehat{P}_a(v) = \frac{1}{J}\sum_{j=1}^{J} \mathbb{I}[v_{aj} = v], \qquad v_{aj}^{(b)} \sim \widehat{P}_a, \quad j = 1, \ldots, J. \tag{1}$$

Here $\sim$ means "sampled from." Conditional on the fixed cache, draws are independent across candidates, judgment positions, and replays. This is equivalent to sampling J entries with replacement from each V_a, including invalid judgments. Collect them as $V_a^{(b)} = (v_{aj}^{(b)})_{j=1}^{J}$ and $V^{(b)} = (V_a^{(b)})_{a \in \mathcal{A}_x}$. Each replay first selects a target from this resampled cache, then recomputes advantages on the original student responses:

$$\hat{a}^{(b)} = H(V^{(b)}; n_x), \qquad A^{(b)} = \begin{cases} \mathcal{G}(Y, r(\hat{a}^{(b)}), C(x)), & \hat{a}^{(b)} \neq \perp, \\ \mathbf{0}, & \hat{a}^{(b)} = \perp. \end{cases} \tag{2}$$

The superscript (b) identifies replay b: $\hat{a}^{(b)}$ is its selected target and $A^{(b)}$ is its vector of N_s response advantages. If the selector abstains, $\mathbf{0}$ assigns zero advantage to every response in that replay. The original question remains accepted; this replay contributes no support or conflict when estimating update reliability. Recomputing the full advantage vector propagates target changes into individual learning signals while keeping the responses fixed. Replays reuse cached judgments, and repeated targets can reuse their computed advantages, so no additional student or teacher decoding is needed.

3.2 MEASURING SUPPORT FOR REINFORCEMENT AND PENALTIES

A replay supports a response's update when it preserves the original advantage sign, and conflicts when it reverses that sign. We estimate their probability difference by

$$s_i = \frac{1}{B}\sum_{b=1}^{B}\Big(\mathbb{I}[A_i^0 A_i^{(b)} > 0] - \mathbb{I}[A_i^0 A_i^{(b)} < 0]\Big), \qquad w_i = [s_i]_+. \tag{3}$$

The score $s_i \in [-1, 1]$ measures net support; $[z]_+ = \max(0, z)$ gives $w_i \in [0, 1]$. A zero replay advantage contributes neither agreement nor conflict, but remains in the denominator B. Consequently, frequent abstention or zero advantages reduce support even without explicit sign reversals.

Different responses to the same question can receive different support. If replay switches the target between two answers, their matching responses may exchange reinforcement and penalty, while a response matching neither answer may remain consistently penalized. A single question weight cannot express this distinction. The score assesses the reproducibility of a signed learning decision, and is not a calibrated correctness probability. We use $B = 32$ cached-evidence replays to estimate support. Appendix E.3 derives its exact conditional expectation and finite-replay error bounds.

3.3 ALLOCATING THE POLICY UPDATE WITH BOUNDED WEIGHTS

We map support to a response multiplier α_i using weighting strength λ, then form the weighted advantage $\widetilde{A}_i$:

$$\alpha_i = (1 - \lambda) + \lambda w_i, \qquad \widetilde{A}_i = \alpha_i A_i^0, \qquad 0 \le \lambda < 1. \tag{4}$$

High-consensus questions use $\alpha_i = 1$. For weighted responses, $\alpha_i \in [1 - \lambda, 1]$ preserves the original sign and bounds the adjustment by $|\widetilde{A}_i - A_i^0| \le \lambda |A_i^0|$. Multipliers are applied after advantage computation, held fixed during differentiation, and not followed by re-centering or normalization. The selected target and coverage-oriented objective continue to define what the policy learns; response-level support controls the relative strength of those contributions.

With the default $\lambda = 0.25$, responses with nonpositive support retain 75% of their original magnitude, keeping updates anchored to the selected target. Appendix D.3.2 examines weighting strength, including the zero-floor boundary $\lambda = 1$.

Let $y_{i\ell}$ be token ℓ of response i and $y_{i,<\ell}$ its preceding tokens, giving context $h_{i\ell} = (x, y_{i,<\ell})$. The importance ratio $\rho_{i\ell}(\theta) = \pi_\theta(y_{i\ell} \mid h_{i\ell}) / \pi_{\theta_{\text{old}}}(y_{i\ell} \mid h_{i\ell})$ compares the current student with the snapshot θ_{old} that generated the responses. The dual-clipped surrogate $h_{\text{DC}}(\rho, A)$ evaluates the policy gain from a token probability ratio ρ and advantage A; Appendix E.1.1 gives its definition and scaling property. The student is optimized with

$$\mathcal{L}_x(\theta) = \underbrace{-\sum_{i,\ell} \nu_{i\ell} m_{i\ell} \alpha_i \, h_{\text{DC}}(\rho_{i\ell}(\theta), A_i^0)}_{\mathcal{L}_{S,x}(\theta; \widetilde{A})} + \mathcal{L}_{\text{reg},x}(\theta). \tag{5}$$

Here $\mathcal{L}_{S,x}$ is the student policy loss for question x, and $\mathcal{L}_{\text{reg},x}$ is its reference-policy regularization term. The mask $m_{i\ell}$ is 1 for a valid response token and 0 otherwise; $\nu_{i\ell}$ is the fixed coefficient from valid-token averaging and gradient accumulation. Gradients update only θ; cached judgments, targets, and replay weights are constants during differentiation. The teacher parameters ϕ are maintained by EMA. Reference-policy regularization is not evidence-weighted.

3.4 CONSTRUCTING THE TRAINING SIGNALS

Student rollouts use snapshot $\pi_{\theta_{\text{old}}}$; verification uses π_ϕ. The student uses pass@K advantages (Chen et al., 2025), where K is the training subset size. The teacher has no optimizer and receives no GRPO or other gradient updates. Confidence-based candidate screening has precedents in Li et al. (2026); the response-specific evidence allocation is defined in Sections 3.1–3.3.

From N_v voting responses, counts $n_x(a)$ determine the majority a_{maj} and frequency-ranked candidates $\mathcal{A}_x$. Our configuration uses a single threshold τ and two routes:

$$C(x) = \frac{n_x(a_{\text{maj}})}{N_v}, \qquad \hat{a}^0 = \begin{cases} a_{\text{maj}}, & C(x) \ge \tau, \\ H(V; n_x), & C(x) < \tau. \end{cases} \tag{6}$$

Table 1: **Sampled accuracy and coverage.** Entries are pass@1 / pass@16 (%). Bold marks the best label-free adapted model; shading identifies REPO. § Supervised GRPO uses ground-truth rewards.

Method	AIME24	AIME25	AMC	MATH-500	DAPO	GPQA
Qwen3-4B-Base						
Base	9.9 / 33.2	8.1 / 35.7	41.0 / 76.1	67.9 / 91.5	28.9 / 67.3	34.2 / 90.5
GRPO§	18.8 / 36.5	16.2 / 37.7	52.7 / 82.4	81.3 / 93.2	42.8 / 75.8	40.6 / 89.2
TTRL	10.8 / 14.9	3.3 / 17.6	42.2 / 60.0	72.8 / 85.3	31.2 / 55.1	35.6 / 74.9
Co-rewarding	13.8 / 22.6	9.0 / 23.2	48.5 / 64.6	76.1 / 86.9	38.0 / 59.5	**37.4** / 73.1
CoVerRL	11.2 / 27.4	8.0 / 34.1	45.8 / **78.7**	75.0 / 91.2	33.9 / 67.9	35.9 / 85.6
REPO	**16.0 / 33.8**	**16.6 / 37.6**	**49.4** / 76.6	**78.3 / 92.9**	**38.3 / 75.0**	37.1 / **87.0**
Qwen3-8B-Base						
Base	10.4 / 46.9	10.1 / 35.0	39.0 / 80.5	63.2 / 92.2	28.7 / 71.0	34.8 / 89.3
GRPO§	22.0 / 41.6	17.7 / 35.3	58.7 / 84.0	83.5 / 94.1	48.2 / 77.2	47.6 / 90.7
TTRL	13.6 / 28.4	9.9 / 24.3	50.6 / 72.9	77.0 / 87.7	38.7 / 64.3	39.1 / 79.6
Co-rewarding	16.7 / 30.7	10.6 / 19.8	**55.0** / 71.1	79.4 / 89.2	**43.5** / 63.7	40.1 / 78.4
CoVerRL	13.3 / 35.1	14.1 / 29.9	52.1 / 78.6	78.2 / 91.8	40.2 / 70.8	39.8 / 85.7
REPO	**17.9 / 42.8**	**15.6 / 40.0**	51.8 / **82.3**	**80.7 / 93.4**	42.0 / **75.9**	**43.8 / 90.7**

For the lower-consensus route, H considers candidates whose True-judgment fraction exceeds the acceptance threshold $t_a/J > \eta$, where $t_a = \sum_{j=1}^{J} \mathbb{I}[v_{aj} = \text{True}]$ counts cached True verdicts; False and invalid judgments remain in J. It selects the eligible candidate with the highest voting frequency, breaking ties by True count and then a fixed order. If no candidate qualifies, $H = \perp$ and the question contributes no student update. High-consensus questions use the majority target without evidence attenuation. Appendix C.2 gives the response and verification budgets.

Teacher-derived student rewards. Teacher verdicts determine which candidate may serve as the pseudo-label. Once $\hat{a}^0$ is selected, the student receives $r_i = \mathbb{I}[a_i = \hat{a}^0]$. These are answer-agreement rewards without reference answers. The teacher generates judgments but is not optimized against a separate verification-reward objective.

Teacher maintenance. We initialize separate student and teacher parameter copies from the same pretrained checkpoint, $\phi_0 = \theta_0$. Let T be the total number of outer iterations, $\theta_{t,\text{old}}$ the student snapshot at the start of iteration t, and β_t the fraction of the previous teacher parameters retained. Before generating judgments at iteration t, we update the teacher using

$$\phi_t = \beta_t \phi_{t-1} + (1 - \beta_t)\theta_{t,\text{old}}, \qquad \beta_t = \beta_{\text{end}} - \frac{\beta_{\text{end}} - \beta_{\text{start}}}{2}\left[1 + \cos\left(\frac{\pi(t-1)}{T-1}\right)\right], \quad (7)$$

Here β_{start} and β_{end} are the initial and final retention coefficients of the cosine schedule, with $t = 1, \ldots, T$. EMA refreshes the teacher once per outer batch; its parameters then stay fixed through verification, replay, and student optimization. EMA smooths the teacher across training, while replay measures sensitivity to finite judgments within a fixed snapshot.

4 Experiments

Our experiments address three questions. **RQ1:** Does REPO improve sampled accuracy while retaining solution coverage? **RQ2:** Does it also improve repeated success? **RQ3:** Which components of REPO account for these effects? Section 4.1 describes the experimental setup. Section 4.2 answers RQ1 at the macro, sampling-budget, and task levels. Section 4.3 answers RQ2 with pass^4 and correct-count profiles. Section 4.4 answers RQ3 with training-feedback diagnostics and matched component controls.

4.1 Experimental setup

We adapt Qwen3-4B-Base and Qwen3-8B-Base (Yang et al., 2025) on unlabeled DAPO-derived questions (Yu et al., 2025a) and evaluate on five external benchmarks: AIME24, AIME25, AMC, MATH-500, and GPQA-Diamond (Hendrycks et al., 2021; Rein et al., 2023). A held-out DAPO split serves as an in-distribution diagnostic and is excluded from macro averages. We compare with

Table 2: **Supporting reliability assessment.** Entries are pass^4 (%). Bold marks the best label-free adapted model; shading identifies REPO. § Supervised GRPO uses ground-truth rewards.

Method	AIME24	AIME25	AMC	MATH-500	DAPO	GPQA
Qwen3-4B-Base						
Base	2.1	0.7	20.2	46.1	9.5	9.3
GRPO§	8.8	8.2	36.5	70.6	24.6	16.1
TTRL	3.7	1.2	24.6	53.7	13.5	9.7
Co-rewarding	7.7	2.3	37.2	66.2	22.5	15.8
CoVerRL	5.5	0.4	30.5	63.8	16.1	13.4
REPO	**8.0**	**5.2**	**39.6**	**67.4**	**25.5**	**19.6**
Qwen3-8B-Base						
Base	1.1	0.9	14.4	36.3	8.3	8.6
GRPO§	10.4	9.7	42.2	74.6	30.3	22.6
TTRL	3.5	5.2	29.6	60.2	17.3	12.6
Co-rewarding	9.1	3.8	43.7	69.2	**31.5**	17.2
CoVerRL	4.2	5.4	35.1	67.4	22.0	15.1
REPO	**9.6**	**7.5**	**45.1**	**71.9**	31.2	**19.8**

Base and three label-free methods: TTRL (Zuo et al., 2025), Co-rewarding (Zhang et al., 2025d), and CoVerRL (Pan et al., 2026b). We also report supervised GRPO (Shao et al., 2024) as a reference that uses ground-truth rewards. For each question, we draw 32 responses from the final checkpoint and compute pass@1, pass@16, and pass^4 from the same samples. External-macro scores weight the five benchmarks equally. Reference answers are used only for evaluation. Appendix C details the baselines, training configuration, and evaluation protocol.

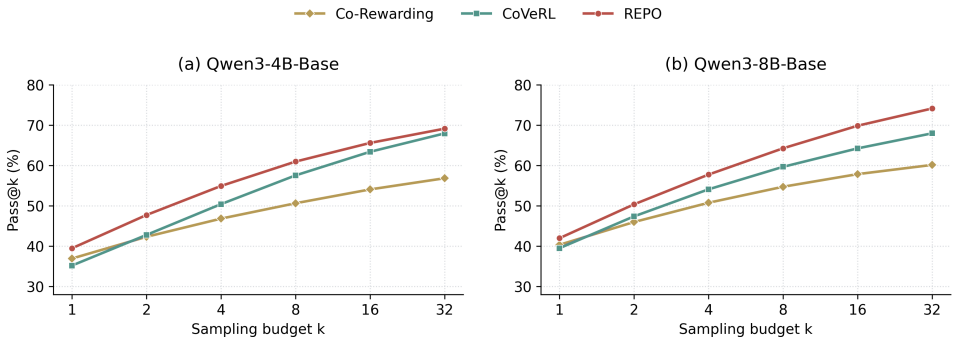


Figure 4: **Coverage across sampling budgets.** Pass@k for Qwen3-4B-Base (left) and Qwen3-8B-Base (right). Higher is better.

4.2 RQ1: Accuracy and Coverage

This section compares accuracy and coverage at three levels of granularity: macro averages (Table 1), sampling budgets (Figure 4), and individual tasks.

Macro-level results. REPO achieves the highest label-free external-macro pass@1 and pass@16 at both scales. Its gains over Co-rewarding are 2.52/11.50 percentage points (pp) at 4B and 1.60/12.00 pp at 8B. The larger coverage gap reflects a pattern shared by the label-free baselines: they improve accuracy over Base but lose coverage. For example, macro pass@16 falls by 11.32/10.94 pp for Co-rewarding and by 14.86/10.20 pp for TTRL at 4B/8B. In contrast, REPO raises accuracy over Base by 7.26/10.46 pp while keeping macro coverage close to Base (+0.18/+1.06 pp). Supervised GRPO remains more accurate, as expected given its ground-truth rewards.

Across sampling budgets. To check that the comparison does not depend on the two reporting budgets, Figure 4 traces external-macro pass@k for $k \in \{1, 2, 4, 8, 16, 32\}$. REPO exceeds CoVerRL and Co-rewarding at both scales throughout this range. Co-rewarding starts close to REPO at $k = 1$

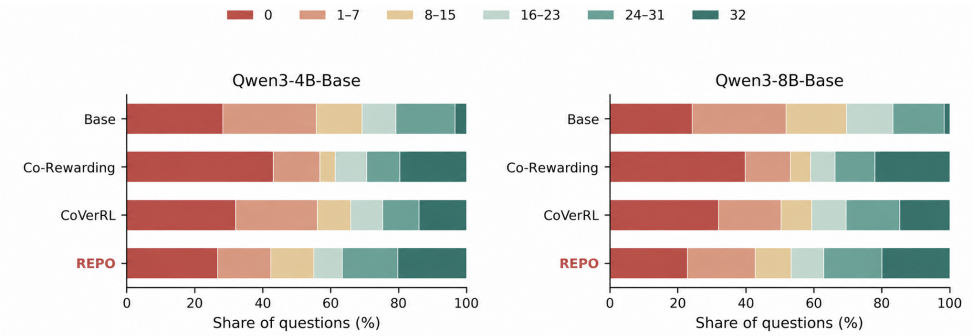


Figure 5: **Correct-count profiles.** Colors denote the number of correct responses among 32 samples per question for Qwen3-4B-Base and Qwen3-8B-Base; each bar represents 100% of questions.

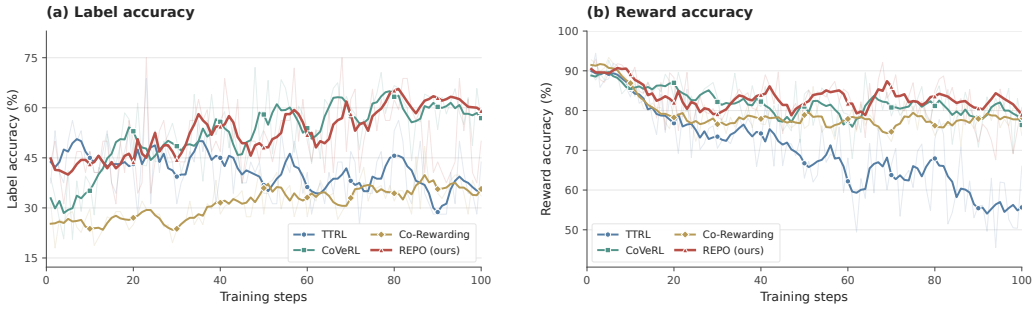


Figure 6: **Training-feedback quality on Qwen3-4B-Base.** (a) Label accuracy and (b) reward accuracy against reference answers, which are used only for measurement (Appendix C.4). Faint traces show per-step values; solid curves show five-step moving averages.

but flattens as k grows, so the 4B gap widens from 2.5 pp at $k = 1$ to 11.5 pp at $k = 16$. This pattern is consistent with a policy concentrating on a narrower set of answers, whereas the REPO curve keeps rising with additional samples.

Across tasks. REPO improves pass@16 over Co-rewarding on all five external benchmarks at both scales and improves pass@1 on four of five. The gains are largest on the hardest benchmark, AIME25: accuracy/coverage increase from 9.0/23.2% to 16.6/37.6% at 4B and from 10.6/19.8% to 15.6/40.0% at 8B. The results also include exceptions. Accuracy is lower than Co-rewarding on 4B GPQA (by 0.3 pp) and 8B AMC (by 3.2 pp). Coverage is lower than Base on 8B AIME24 (46.9% to 42.8%) and 4B GPQA (90.5% to 87.0%). On 8B DAPO, coverage rises over Co-rewarding from 63.7% to 75.9% while accuracy falls from 43.5% to 42.0%. Because accuracy and coverage can move in opposite directions on the same task, we report them separately and describe the overall result as aggregate coverage retention.

4.3 RQ2: Repeated Success

This section asks whether the gains in RQ1 come with consistent correctness. It first reports pass^4 (Table 2) and then uses correct-count profiles (Figure 5) to show how coverage and consistency relate.

Repeated-success results. pass^4 requires four independent responses to the same question to all be correct, so it measures consistency rather than occasional success. REPO exceeds every label-free baseline in pass^4 on all five external benchmarks at both scales. Its external-macro gains over Co-rewarding are 2.12/2.18 pp at 4B/8B. The largest benchmark-level gains are on 4B AIME25 (2.3% to 5.2%) and 4B GPQA (15.8% to 19.6%).

Correct-count profiles. Figure 5 groups questions by the number of correct responses among 32 samples. The zero-correct bin marks complete failure and limits coverage, intermediate bins mark partial success, and the all-correct bin marks full consistency. Compared with Base, Co-rewarding enlarges both extremes: more questions become all-correct, but more also become zero-correct. This explains how Co-rewarding combines competitive pass^4 with low coverage. REPO enlarges the high-correct bins to a similar extent while keeping the zero-correct share close to Base. It therefore improves consistency without giving up questions that were previously solved only occasionally.

Table 3: **Component ablations (%).** Scores use one decimal.

Variant	AMC			DAPO		
	pass@1	pass@16	pass^4	pass@1	pass@16	pass^4
Qwen3-4B-Base						
Base	41.0	76.1	20.2	28.9	67.3	9.5
w/ REPO	49.4	76.6	39.6	38.3	75.0	25.5
- w/o *EW*	48.4	76.2	32.1	37.5	74.8	19.5
- w/o *RA*	47.6	75.8	29.8	36.9	74.1	17.2
Qwen3-8B-Base						
Base	39.0	80.5	14.4	28.7	71.0	8.3
w/ REPO	51.8	82.3	45.1	42.0	75.9	31.2
- w/o *EW*	50.7	81.8	36.5	41.1	75.4	23.9
- w/o *RA*	49.9	81.1	33.7	40.2	74.5	21.5

4.4 RQ3: Training diagnostics and component ablations

This section examines where the gains come from. It first compares feedback quality during training (Figure 6). It then isolates evidence weighting and response-level allocation with matched controls (Table 3), and finally varies the weighting strength and allocation granularity (Table 5, Appendix D.3.2).

Training-feedback quality. Label accuracy measures how often the pseudo-label matches the reference answer. Reward accuracy measures how often a response's assigned reward matches its reference-based reward (definitions in Appendix C.4). For the 4B runs, REPO keeps label accuracy among the highest throughout training. TTRL's reward accuracy declines steadily, consistent with majority voting reinforcing its own mistakes.

Component ablations. Table 3 reports matched runs on AMC and DAPO at both scales (control definitions in Appendix C.4). *w/o EW* removes evidence weighting ($\lambda = 0$). *w/o RA* replaces the response-specific multipliers with one question-level value that preserves the total absolute token-advantage mass (Equation 8). Removing EW mainly lowers repeated success: pass^4 drops by 6.0–8.6 pp, while pass@1 and pass@16 change by at most 1.1 and 0.5 pp. *w/o RA* lowers pass^4 by 8.3–11.4 pp relative to REPO, despite using the same replay evidence. Averaging weakens well-supported updates together with disputed ones, so how attenuation is distributed within a question matters, not only how much attenuation is applied.

Weighting strength and granularity. Appendix D.3.2 varies how α_i is computed on Qwen3-4B-Base and reports external-macro scores (Table 5). Introducing evidence weighting ($\lambda = 0 \rightarrow 0.25$) raises pass^4 from 21.5% to 28.0%, while pass@1 and pass@16 change by only 1.0 and 0.4 pp. Stronger weighting gives similar repeated success (29.0% at $\lambda = 0.5$ and 28.3% with the zero floor $\lambda = 1$), so the gain is not specific to the default strength within this range. A question-level multiplier built from the same replays, using the fraction that reproduces the original target, reaches 24.0% pass^4, 4.0 pp below response-level sign support. Together with *w/o RA*, which averages REPO's own multipliers at matched token-advantage mass, this indicates that the benefit depends on how attenuation is distributed across responses, not only on how much is applied.

5 Conclusion

Across two model scales and five external benchmarks, REPO improves macro-averaged accuracy and repeated success over Co-rewarding while largely retaining Base coverage, with coverage losses on some individual tasks. It replays cached EMA-teacher judgments through target selection and advantage computation, then uses response-specific sign support to attenuate uncertain updates without additional verification decoding. Question-level selection decides whether to accept supervision, whereas response-level weights determine how strongly to retain each signed update. The EMA teacher smooths feedback across training, while evidence replay probes uncertainty within a fixed teacher snapshot. The component controls assess weighting within a shared training configuration; they do not partition the baseline gaps among all components. Reliability preservation remains empirical; reproducible supervision need not be correct.

AI USE STATEMENT

We used generative AI tools only to assist with the presentation of this manuscript: correcting grammar and spelling, improving the wording and readability of sentences, keeping terminology and notation consistent, and formatting the LaTeX source, tables, and references to the conference style.

REFERENCES

Xin-Qiang Cai, Wei Wang, Feng Liu, Tongliang Liu, Gang Niu, and Masashi Sugiyama. Reinforcement Learning with Verifiable yet Noisy Rewards under Imperfect Verifiers. *arXiv preprint arXiv:2510.00915*, 2025. URL https://arxiv.org/abs/2510.00915.

Mark Chen, Jerry Tworek, Heewoo Jun, Qiming Yuan, Henrique Ponde de Oliveira Pinto, Jared Kaplan, et al. Evaluating Large Language Models Trained on Code. *arXiv preprint arXiv:2107.03374*, 2021. URL https://arxiv.org/abs/2107.03374.

Zhipeng Chen, Xiaobo Qin, Youbin Wu, Yue Ling, Qinghao Ye, Wayne Xin Zhao, and Guang Shi. Pass@k Training for Adaptively Balancing Exploration and Exploitation of Large Reasoning Models. *arXiv preprint arXiv:2508.10751*, 2025. URL https://arxiv.org/abs/2508.10751.

Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, Christopher Hesse, and John Schulman. Training Verifiers to Solve Math Word Problems. *arXiv preprint arXiv:2110.14168*, 2021. URL https://arxiv.org/abs/2110.14168.

Ganqu Cui, Yuchen Zhang, Jiacheng Chen, Lifan Yuan, Zhi Wang, Yuxin Zuo, Haozhan Li, Yuchen Fan, Huayu Chen, Weize Chen, Zhiyuan Liu, Hao Peng, Lei Bai, Wanli Ouyang, Yu Cheng, Bowen Zhou, and Ning Ding. The Entropy Mechanism of Reinforcement Learning for Reasoning Language Models. *arXiv preprint arXiv:2505.22617*, 2025. URL https://arxiv.org/abs/2505.22617.

DeepSeek-AI, Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, et al. DeepSeek-R1: Incentivizing Reasoning Capability in LLMs via Reinforcement Learning. *arXiv preprint arXiv:2501.12948*, 2025. URL https://arxiv.org/abs/2501.12948v1.

Bodong Du, Xuanqi Huang, and Xiaomeng Li. Distribution-Aware Reward Estimation for Test-Time Reinforcement Learning. *arXiv preprint arXiv:2601.21804*, 2026. URL https://arxiv.org/abs/2601.21804.

Caglar Gulcehre, Tom Le Paine, Srivatsan Srinivasan, Ksenia Konyushkova, Lotte Weerts, Abhishek Sharma, Aditya Siddhant, Alex Ahern, Miaosen Wang, Chenjie Gu, Wolfgang Macherey, Arnaud Doucet, Orhan Firat, and Nando de Freitas. Reinforced Self-Training (ReST) for Language Modeling. *arXiv preprint arXiv:2308.08998*, 2023. URL https://arxiv.org/abs/2308.08998.

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. *arXiv preprint arXiv:2103.03874*, 2021. URL https://arxiv.org/abs/2103.03874.

Zican Hu, Shilin Zhang, Yafu Li, Jianhao Yan, Xuyang Hu, Leyang Cui, Xiaoye Qu, Chunlin Chen, Yu Cheng, and Zhi Wang. Diversity-Incentivized Exploration for Versatile Reasoning. *arXiv preprint arXiv:2509.26209*, 2025. URL https://arxiv.org/abs/2509.26209.

Chengsong Huang, Wenhao Yu, Xiaoyang Wang, Hongming Zhang, Zongxia Li, Ruosen Li, Jiaxin Huang, Haitao Mi, and Dong Yu. R-Zero: Self-Evolving Reasoning LLM from Zero Data. *arXiv preprint arXiv:2508.05004*, 2025. URL https://arxiv.org/abs/2508.05004.

Aviral Kumar, Vincent Zhuang, Rishabh Agarwal, Yi Su, John D. Co-Reyes, Avi Singh, et al. Training Language Models to Self-Correct via Reinforcement Learning. *arXiv preprint arXiv:2409.12917*, 2024. URL https://arxiv.org/abs/2409.12917.

Harrison Lee, Samrat Phatale, Hassan Mansoor, Thomas Mesnard, Johan Ferret, Kellie Ren Lu, Colton Bishop, Ethan Hall, Victor Carbune, Abhinav Rastogi, and Sushant Prakash. RLAIF vs. RLHF: Scaling Reinforcement Learning from Human Feedback with AI Feedback. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235, pp. 26874–26901. PMLR, 2024. URL https://proceedings.mlr.press/v235/lee24t.html.

Jiahui Li, Jianfeng Shan, Wenpei Chen, Shunyu Wu, Jian Lou, Wenjie Feng, Dan Li, and See-Kiong Ng. Exploiting Verification-Generation Gap: Test-Time Reinforcement Learning with Confidence-Conditioned Verification. *arXiv preprint arXiv:2606.03608*, 2026. URL https://arxiv.org/abs/2606.03608.

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. *arXiv preprint arXiv:2305.20050*, 2023. URL https://arxiv.org/abs/2305.20050.

Jia Liu, ChangYi He, YingQiao Lin, MingMin Yang, FeiYang Shen, and ShaoGuo Liu. ETTRL: Balancing Exploration and Exploitation in LLM Test-Time Reinforcement Learning Via Entropy Mechanism. *arXiv preprint arXiv:2508.11356*, 2025a. URL https://arxiv.org/abs/2508.11356.

Junnan Liu, Hongwei Liu, Linchen Xiao, Ziyi Wang, Kuikun Liu, Songyang Gao, Wenwei Zhang, Songyang Zhang, and Kai Chen. Are Your LLMs Capable of Stable Reasoning? *arXiv preprint arXiv:2412.13147*, 2024. URL https://arxiv.org/abs/2412.13147.

Mingjie Liu, Shizhe Diao, Ximing Lu, Jian Hu, Xin Dong, Yejin Choi, Jan Kautz, and Yi Dong. ProRL: Prolonged Reinforcement Learning Expands Reasoning Boundaries in Large Language Models. *arXiv preprint arXiv:2505.24864*, 2025b. URL https://arxiv.org/abs/2505.24864.

Zichen Liu, Changyu Chen, Wenjun Li, Penghui Qi, Tianyu Pang, Chao Du, Wee Sun Lee, and Min Lin. Understanding R1-Zero-Like Training: A Critical Perspective. *arXiv preprint arXiv:2503.20783*, 2025c. URL https://arxiv.org/abs/2503.20783.

Long Ouyang, Jeff Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, et al. Training language models to follow instructions with human feedback. *arXiv preprint arXiv:2203.02155*, 2022. URL https://arxiv.org/abs/2203.02155.

Licheng Pan, Haocheng Yang, Haoxuan Li, Yichen Sun, Yunsheng Lu, Shijian Wang, Lei Shen, Yuan Lu, Zhixuan Chu, and Hao Wang. Uncertainty-Aware Reward Modeling for Stable RLHF. *arXiv preprint arXiv:2606.19818*, 2026a. URL https://arxiv.org/abs/2606.19818.

Teng Pan, Yuchen Yan, Zixuan Wang, Ruiqing Zhang, Guiyang Hou, Wenqi Zhang, Weiming Lu, Jun Xiao, and Yongliang Shen. CoVerRL: Breaking the Consensus Trap in Label-Free Reasoning via Generator-Verifier Co-Evolution. *arXiv preprint arXiv:2603.17775*, 2026b. URL https://arxiv.org/abs/2603.17775.

Shubham Parashar, Shurui Gui, Xiner Li, Hongyi Ling, Sushil Vemuri, Blake Olson, Eric Li, Yu Zhang, James Caverlee, Dileep Kalathil, and Shuiwang Ji. Curriculum Reinforcement Learning from Easy to Hard Tasks Improves LLM Reasoning. *arXiv preprint arXiv:2506.06632*, 2025. URL https://arxiv.org/abs/2506.06632.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Stefano Ermon, Christopher D. Manning, and Chelsea Finn. Direct Preference Optimization: Your Language Model is Secretly a Reward Model. *arXiv preprint arXiv:2305.18290*, 2023. URL https://arxiv.org/abs/2305.18290.

David Rein, Betty Li Hou, Asa Cooper Stickland, Jackson Petty, Richard Yuanzhe Pang, Julien Dirani, Julian Michael, and Samuel R. Bowman. GPQA: A Graduate-Level Google-Proof Q&A Benchmark. *arXiv preprint arXiv:2311.12022*, 2023. URL https://arxiv.org/abs/2311.12022.

John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. Proximal Policy Optimization Algorithms. *arXiv preprint arXiv:1707.06347*, 2017. URL https://arxiv.org/abs/1707.06347.

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.

Guangming Sheng, Chi Zhang, Zilingfeng Ye, Xibin Wu, Wang Zhang, Ru Zhang, Yanghua Peng, Haibin Lin, and Chuan Wu. HybridFlow: A Flexible and Efficient RLHF Framework. *arXiv preprint arXiv:2409.19256*, 2024. URL https://arxiv.org/abs/2409.19256.

Avi Singh, John D. Co-Reyes, Rishabh Agarwal, Ankesh Anand, Piyush Patil, Xavier Garcia, et al. Beyond Human Data: Scaling Self-Training for Problem-Solving with Language Models. *Transactions on Machine Learning Research*, 2024. URL https://arxiv.org/abs/2312.06585.

Nisan Stiennon, Long Ouyang, Jeff Wu, Daniel M. Ziegler, Ryan Lowe, Chelsea Voss, Alec Radford, Dario Amodei, and Paul Christiano. Learning to summarize from human feedback. *arXiv preprint arXiv:2009.01325*, 2020. URL https://arxiv.org/abs/2009.01325.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, et al. Qwen3 Technical Report. *arXiv preprint arXiv:2505.09388*, 2025. URL https://arxiv.org/abs/2505.09388.

Shunyu Yao, Noah Shinn, Pedram Razavi, and Karthik Narasimhan. τ-bench: A Benchmark for Tool-Agent-User Interaction in Real-World Domains. *arXiv preprint arXiv:2406.12045*, 2024. URL https://arxiv.org/abs/2406.12045.

Deheng Ye, Zhao Liu, Mingfei Sun, Bei Shi, Peilin Zhao, Hao Wu, et al. Mastering Complex Control in MOBA Games with Deep Reinforcement Learning. *arXiv preprint arXiv:1912.09729*, 2019. URL https://arxiv.org/abs/1912.09729.

Qiying Yu, Zheng Zhang, Ruofei Zhu, Yufeng Yuan, Xiaochen Zuo, Yu Yue, et al. DAPO: An Open-Source LLM Reinforcement Learning System at Scale. *arXiv preprint arXiv:2503.14476*, 2025a. URL https://arxiv.org/abs/2503.14476.

Zhaoning Yu, Will Su, Leitian Tao, Haozhu Wang, Aashu Singh, Hanchao Yu, Jianyu Wang, Hongyang Gao, Weizhe Yuan, Jason Weston, Ping Yu, and Jing Xu. RESTRAIN: From Spurious Votes to Signals – Self-Driven RL with Self-Penalization. *arXiv preprint arXiv:2510.02172*, 2025b. URL https://arxiv.org/abs/2510.02172.

Weizhe Yuan, Richard Yuanzhe Pang, Kyunghyun Cho, Xian Li, Sainbayar Sukhbaatar, Jing Xu, and Jason Weston. Self-Rewarding Language Models. *arXiv preprint arXiv:2401.10020*, 2024. URL https://arxiv.org/abs/2401.10020.

Yang Yue, Zhiqi Chen, Rui Lu, Andrew Zhao, Zhaokai Wang, Yang Yue, Shiji Song, and Gao Huang. Does Reinforcement Learning Really Incentivize Reasoning Capacity in LLMs Beyond the Base Model? *arXiv preprint arXiv:2504.13837*, 2025. URL https://arxiv.org/abs/2504.13837.

Dan Zhang, Sining Zhoubian, Ziniu Hu, Yisong Yue, Yuxiao Dong, and Jie Tang. ReST-MCTS*: LLM Self-Training via Process Reward Guided Tree Search. *arXiv preprint arXiv:2406.03816*, 2024. URL https://arxiv.org/abs/2406.03816.

Qingyang Zhang, Haitao Wu, Changqing Zhang, Peilin Zhao, and Yatao Bian. Right Question is Already Half the Answer: Fully Unsupervised LLM Reasoning Incentivization. *arXiv preprint arXiv:2504.05812*, 2025a. URL https://arxiv.org/abs/2504.05812.

Shimao Zhang, Xiao Liu, Xin Zhang, Junxiao Liu, Zheheng Luo, Shujian Huang, and Yeyun Gong. Process-based Self-Rewarding Language Models. In *Findings of the Association for Computational Linguistics: ACL 2025*, pp. 18097–18110, 2025b. URL https://aclanthology.org/2025.findings-acl.930/.

Yanzhi Zhang, Zhaoxi Zhang, Haoxiang Guan, Yilin Cheng, Yitong Duan, Chen Wang, Yue Wang, Shuxin Zheng, and Jiyan He. No Free Lunch: Rethinking Internal Feedback for LLM Reasoning. *arXiv preprint arXiv:2506.17219*, 2025c. URL https://arxiv.org/abs/2506.17219.

Zizhuo Zhang, Jianing Zhu, Xinmu Ge, Zihua Zhao, Zhanke Zhou, Xuan Li, Xiao Feng, Jiangchao Yao, and Bo Han. Co-rewarding: Stable Self-supervised RL for Eliciting Reasoning in Large Language Models. *arXiv preprint arXiv:2508.00410*, 2025d. URL https://arxiv.org/abs/2508.00410.

Andrew Zhao, Yiran Wu, Yang Yue, Tong Wu, Quentin Xu, et al. Absolute Zero: Reinforced Self-play Reasoning with Zero Data. *arXiv preprint arXiv:2505.03335*, 2025a. URL https://arxiv.org/abs/2505.03335.

Xuandong Zhao, Zhewei Kang, Aosong Feng, Sergey Levine, and Dawn Song. Learning to Reason without External Rewards. *arXiv preprint arXiv:2505.19590*, 2025b. URL https://arxiv.org/abs/2505.19590.

Xinyu Zhu, Mengzhou Xia, Zhepei Wei, Wei-Lin Chen, Danqi Chen, and Yu Meng. The Surprising Effectiveness of Negative Reinforcement in LLM Reasoning. *arXiv preprint arXiv:2506.01347*, 2025. URL https://arxiv.org/abs/2506.01347.

Yuxin Zuo, Kaiyan Zhang, Li Sheng, Shang Qu, Ganqu Cui, Xuekai Zhu, Haozhan Li, Yuchen Zhang, Xinwei Long, Ermo Hua, Biqing Qi, Youbang Sun, Zhiyuan Ma, Lifan Yuan, Ning Ding, and Bowen Zhou. TTRL: Test-Time Reinforcement Learning. *arXiv preprint arXiv:2504.16084*, 2025. URL https://arxiv.org/abs/2504.16084.

Adam Zweiger, Jyothish Pari, Han Guo, Ekin Akyürek, Yoon Kim, and Pulkit Agrawal. Self-Adapting Language Models. *arXiv preprint arXiv:2506.10943*, 2025. URL https://arxiv.org/abs/2506.10943.

A Additional Related Works

A.1 Test-time reinforcement learning

Label-free reasoning RL constructs supervision from the model's own outputs: TTRL uses majority agreement (Zuo et al., 2025), EMPO uses semantic-answer concentration (Zhang et al., 2025a), and Intuitor uses self-certainty (Zhao et al., 2025b). RESTRAIN combines frequency weighting with self-penalization (Yu et al., 2025b), while DARE constructs rewards from answer frequencies and trajectory entropy (Du et al., 2026). Complementary supervision, joint generation–verification, and confidence-conditioned verification provide further target-construction mechanisms (Zhang et al., 2025d; Pan et al., 2026b; Li et al., 2026). Related self-rewarding methods construct response-level or step-wise preferences for preference optimization (Yuan et al., 2024; Zhang et al., 2025b), and self-evolving approaches differ in their feedback: Absolute Zero uses executable tasks (Zhao et al., 2025a), R-Zero uses uncertainty-driven curricula and majority-voted targets (Huang et al., 2025), and SEAL rewards self-edits through post-adaptation task performance (Zweiger et al., 2025). These are distinct supervision regimes, and internal feedback need not track correctness (Zhang et al., 2025c). Confidence-sensitive weighting also has direct precedents: ETTRL uses response entropy for sampling and advantage shaping (Liu et al., 2025a), and noisy-reward methods correct policy-gradient terms or use calibrated reward uncertainty (Cai et al., 2025; Pan et al., 2026a). REPO instead fixes the student responses and teacher snapshot, resamples cached judgments, and replays target selection and the complete advantage operator. Agreement and conflict with each original advantage sign determine bounded response-specific multipliers, which preserve supported reinforcement and penalties while attenuating disputed update directions within the coverage-oriented objective. This criterion measures update reproducibility, not correctness probability: the original target remains the training target, and reproducible verifier errors can remain fully weighted. For repeated success, we adopt the established G-Pass@k stable-reasoning criterion (Liu et al., 2024).

A.2 Reinforcement Learning for LLM Reasoning

LLM alignment learns rewards from human preferences (Stiennon et al., 2020; Ouyang et al., 2022), substitutes AI feedback (Lee et al., 2024), or optimizes preferences directly with DPO (Rafailov et al., 2023). For reasoning, GRPO, DAPO, and DeepSeek-R1 use verifiable rewards (Shao et al., 2024; Yu et al., 2025a; DeepSeek-AI et al., 2025); learned outcome and process verifiers provide complementary evaluation signals (Cobbe et al., 2021; Lightman et al., 2023). Reward-guided self-training reuses generated samples in ReST (Gulcehre et al., 2023), filters solutions in ReST-EM (Singh et al., 2024), and uses process-reward-guided search in ReST-MCTS* (Zhang et al., 2024). SCoRe trains multi-turn self-correction with correctness rewards (Kumar et al., 2024). Normalization choices (Liu et al., 2025c), task curricula (Parashar et al., 2025), diversity incentives (Hu et al., 2025), and negative reinforcement (Zhu et al., 2025) alter optimization or exploration. Findings on reasoning coverage depend on the training regime, motivating evaluation across sampling budgets (Yue et al., 2025; Liu et al., 2025b; Chen et al., 2025). For noisy feedback, PGBC and PGFC use verifier error rates for reward or signed-gradient corrections (Cai et al., 2025); UARM uses calibrated reward intervals to weight responses and modify GRPO normalization (Pan et al., 2026a). REPO uses neither error-rate estimates nor labeled calibration, applying positive multipliers after complete advantage computation without re-centering. Its contribution is the evidence-replay criterion for weighting, not response weighting itself or an unbiased clean-reward gradient. G-Pass@k and τ-bench provide the established repeated-success criterion used alongside our primary accuracy and coverage metrics (Liu et al., 2024; Yao et al., 2024).

B PSEUDOCODE OF REPO

Algorithm 1 summarizes training with separate student parameters θ and teacher parameters ϕ. The assignment arrow $\leftarrow$ sets a variable, and $\mathrm{stopgrad}$ holds its argument constant during differentiation. Each outer iteration refreshes the EMA teacher once from the student snapshot before teacher judgments are generated; ϕ then stays fixed throughout that batch. Student voting responses determine the candidate set and consensus route. High consensus assigns the majority target and unit multipliers; low consensus applies H to cached teacher judgments, and original abstention skips the question's student update. For accepted low-consensus questions, replays keep responses, candidates, voting counts, teacher parameters, and route fixed. Resampled abstention contributes zero advantages and remains in the denominator B. Multipliers are applied after the complete advantage computation, with no re-centering or normalization. Only the student receives gradient updates; teacher maintenance uses Equation 7. The learning-rate scheduler advances after the outer iteration.

Algorithm 1 Resampled-Evidence Policy Optimization (REPO)

Require: Unlabeled question set $\mathcal{D}$, student θ, teacher $\phi \leftarrow \theta$, EMA schedule β_t, and T iterations.
Require: Response budgets N_v, N_s, judgments J, replays B, thresholds τ, η, and strength λ.
for $t = 1, \ldots, T$ **do**
  Draw a question batch $\mathcal{X}$ from $\mathcal{D}$; set $\theta_{t,\mathrm{old}} \leftarrow \theta$ and refresh ϕ using Equation 7.
  for each question $x \in \mathcal{X}$ **do**
    Generate N_v voting responses from $\pi_{\theta_{t,\mathrm{old}}}$; retain $Y = (y_i)_{i=1}^{N_s}$ and extract a_i.
    Compute voting counts n_x; set $a_{\mathrm{maj}} \leftarrow \arg\max_a n_x(a)$ and $C(x) \leftarrow n_x(a_{\mathrm{maj}})/N_v$.
    Select frequency-ranked candidates $\mathcal{A}_x$; cache J judgments V_a per candidate from π_ϕ.
    $\hat{a}^0 \leftarrow a_{\mathrm{maj}}$ if $C(x) \geq \tau$, otherwise $H(V; n_x)$; initialize $\alpha_i \leftarrow 1$.
    if $\hat{a}^0 = \perp$ **then**
      continue ▷ Skip this question's student update.
    end if
    $A^0 \leftarrow \mathcal{G}(Y, r(\hat{a}^0), C(x))$ using the complete advantage operator.
    if $C(x) < \tau$ **then**
      Compute the verdict frequencies $\widehat{P}_a(v)$ for each candidate using Equation 1.
      Initialize $s_i \leftarrow 0$ for all i; keep Y, $\mathcal{A}_x$, n_x, and the consensus route fixed.
      for $b = 1, \ldots, B$ **do**
        Independently draw J verdicts from each $\widehat{P}_a$ to form the resampled cache $V^{(b)}$.
        $\hat{a}^{(b)} \leftarrow H(V^{(b)}; n_x)$.
        Recompute $A^{(b)}$ with $\mathcal{G}$; use $\mathbf{0}$ if $\hat{a}^{(b)} = \perp$ (Equation 2).
        $s_i \leftarrow s_i + B^{-1}\big(\mathbb{I}[A_i^0 A_i^{(b)} > 0] - \mathbb{I}[A_i^0 A_i^{(b)} < 0]\big)$ for all i.
      end for
      $w_i \leftarrow \max(0, s_i)$; $\alpha_i \leftarrow (1 - \lambda) + \lambda w_i$ for $A_i^0 \neq 0$.
    end if
    $\widetilde{A}_i \leftarrow \mathrm{stopgrad}(\alpha_i) A_i^0$ for all i, without re-centering or normalization.
    Update student θ with Equation 5; keep teacher ϕ fixed.
  end for
  Advance the learning-rate scheduler after the outer iteration.
end for
return θ.

C Implementation details

C.1 Baselines

All adapted methods start from the same Base checkpoint, train on the same unlabeled question pool, and share the evaluation protocol in Appendix C.3. Each baseline keeps its own reward construction and objective. Base uses the unadapted Qwen3 weights and serves as the reference point for coverage retention. TTRL (Zuo et al., 2025) rewards agreement with the majority-voted answer; Co-rewarding (Zhang et al., 2025d) uses complementary supervision; CoVerRL (Pan et al., 2026b) uses its official generator–verifier co-evolution objective on our question pool. All methods are evaluated on single-turn responses without verification or revision at test time. Supervised GRPO (Shao et al., 2024) replaces model-derived supervision with binary correctness rewards and is reported only as a reference; bold entries in all tables exclude it.

C.2 Training configuration

Table 4 summarizes the configuration, implemented with verl (Sheng et al., 2024), and some of our method designs are inspired by Zhang et al. (2025d); Li et al. (2026). Training uses 14,108 DAPO-derived questions whose reference answers are never used to construct rewards or select targets. High-consensus questions ($C(x) \geq \tau$) use the majority answer directly without evidence weighting. For low-consensus questions, the teacher judges the top five candidates; invalid judgments remain in the denominator J and therefore count against eligibility. The same advantage operator $\mathcal{G}$ is used for the original advantages and for every replay, so replay changes only the target. Algorithm 1 specifies the update and scheduling order.

Table 4: **REPO training and evaluation settings.** Component controls inherit this configuration except for their named intervention.

Component	Configuration
Student	Qwen3-4B-Base / Qwen3-8B-Base; separate student parameters θ; voting responses $N_v = 64$; retained responses $N_s = 32$; coverage subset $K = 4$
Teacher	Consensus threshold $\tau = 0.8$; judgments per candidate $J = 8$; low Consensus: top 5 candidates, temperature 1.0; top-$p = 0.85$; acceptance fraction strictly above $\eta = 0.6$
Teacher maintenance	Separate parameters ϕ, initialized with $\phi_0 = \theta_0$; EMA once per outer batch before verification; no teacher optimizer; cosine β_t: $0.99 \rightarrow 0.9999$ over $T = 100$ iterations
Evidence weighting	Replay budget $B = 32$; weighting strength $\lambda = 0.25$
Policy optimization	Learning rate 5×10^{-7}; PPO clip 0.2; dual clip 3.0; valid-token means with gradient accumulation; reference-policy regularization
Batching and schedule	32 questions per outer iteration; actor micro-batch per GPU: 4; 100 training steps; warmup ratio 0.03
Token limits	Student prompt: 1024; student response: 2048; teacher response: 2048
Evaluation	32 responses per question; temperature 0.6; top-$p = 0.95$; top-$k = 20$; metrics: pass@1, pass@16, pass^4

C.3 Evaluation protocol

Evaluation covers AIME24 (30 questions), AIME25 (30), AMC (83), MATH-500 (500), GPQA-Diamond (197), and DAPO (500). DAPO is an in-distribution diagnostic and is excluded from macro averages. For each question, we draw 32 responses from the final checkpoint; sampling requests are deterministic functions of the checkpoint, question ID, and sample ordinal. All metrics are computed from the same responses using the unbiased estimators in Appendix E.2. Dataset scores average per-question estimates, and external-macro scores weight the five external benchmarks equally. The

pass@k curves in Figure 4 use the same estimator for $k \in \{1, 2, 4, 8, 16, 32\}$. Figure 5 bins questions by their number of correct responses into 0, 1–7, 8–15, 16–23, 24–31, and 32. Label-free runs use ground truth only for evaluation.

C.4 Training diagnostics and controls

In Figure 6, label accuracy is the fraction of pseudo-labeled questions whose training target matches the reference answer, and reward accuracy is the fraction of training responses whose assigned reward matches the reward under the reference answer. Reference answers are used only for these measurements. All controls in Tables 3 and 5 inherit Table 4 and change only how α_i is computed for accepted low-consensus questions. *w/o EW* sets $\lambda = 0$; *w/o RA* replaces the response-specific multipliers with the question-level mean in Equation 8; the question-level variant uses the fraction of replays that reproduce the original target.

C.5 System Prompt

Figure 7 shows the instruction for reasoning and final-answer formatting.

System Prompt

Please reason step by step, and put your final answer within `\boxed{}`.

Figure 7: System prompt for reasoning and final-answer formatting.

C.6 Dataset Examples

Table 6 presents illustrative questions and reference answers from the six datasets.

D Additional Experimental Results

D.1 Majority-vote accuracy

Figure 8 reports majority-vote accuracy for Qwen3-4B-Base and Qwen3-8B-Base. Each question is answered by selecting the most frequent answer among 32 student generations. REPO reaches 49.0% and 50.1%, respectively, exceeding Co-rewarding (41.1% and 43.9%) by 7.9 and 6.2 percentage points and CoVerRL (42.7% and 46.6%) by 6.3 and 3.5 points. Supervised GRPO attains 49.4% and 54.1%, exceeding REPO by 0.4 and 4.0 points. These comparisons complement the primary accuracy and coverage metrics by evaluating the correctness of the single answer selected through majority voting.

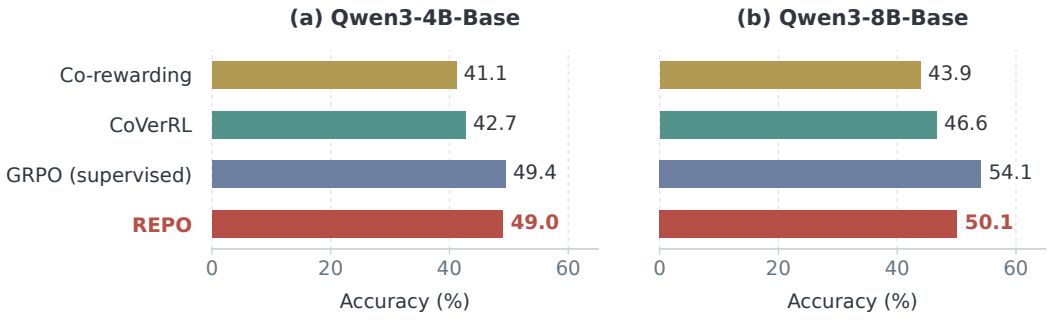


Figure 8: **Majority-vote accuracy.** Results for (a) Qwen3-4B-Base and (b) Qwen3-8B-Base, using 32 responses per question. Accuracy is an equal-weight macro average over the five external benchmarks, excluding DAPO. Red highlights REPO. GRPO uses supervised rewards.

D.2 Additional backbone: Qwen2.5-Math-7B

Figure 9 compares Qwen2.5-Math-7B-Base with its REPO-adapted counterpart on six datasets. REPO improves pass@1 and pass^4 on every dataset, with the largest gains on MATH-500 (+28.13 and +40.09 percentage points, respectively). The pass@16 changes are mixed: coverage increases on four datasets but decreases on AIME25 and GPQA by 3.23 and 3.88 points. Across the five external benchmarks, equal-weight macro averages improve by 13.88, 1.50, and 16.63 points for pass@1, pass@16, and pass^4, respectively; DAPO is excluded from these averages.

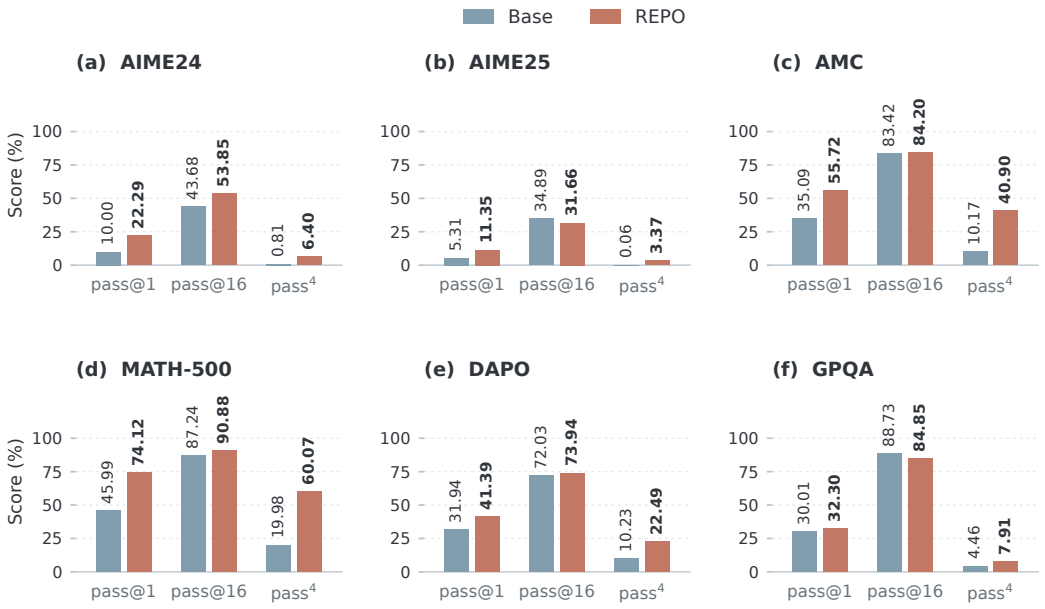


Figure 9: **Base and REPO on Qwen2.5-Math-7B.** Each panel compares pass@1, pass@16, and pass^4 on one dataset using grouped vertical bars. Gray-blue denotes Base and terracotta denotes REPO. All values are percentages, and all panels use the same axis scale. DAPO is shown separately from the five external benchmarks.

D.3 Controlled ablations

D.3.1 Evidence Weighting and Response-Level Allocation

We compare Base and REPO with two matched controls on AMC and DAPO at both model scales (Table 3). The *w/o EW* control sets $\lambda = 0$, while *w/o RA* replaces response-specific weights with a uniform multiplier. Both controls retain the teacher initialization and EMA schedule, two-route selector, pass@K advantages, and verification settings from Table 4. The first comparison measures the effect of evidence weighting within this configuration; the second examines response-level allocation at matched absolute token-advantage mass.

For question x, the uniform allocation control uses

$$\bar{\alpha}_x = \frac{\sum_i L_i \alpha_i |A_i^0|}{\sum_i L_i |A_i^0|}, \tag{8}$$

where L_i counts valid response tokens; set $\bar{\alpha}_x = 1$ for a zero denominator. This preserves absolute token-advantage mass while replacing the individual multipliers by one question-level value.

At 4B, removing EW gives pass^4 of 32.1% on AMC and 19.5% on DAPO, drops of 7.5 and 6.0 pp from REPO. Uniform allocation gives 29.8% and 17.2%, drops of 9.8 and 8.3 pp. At 8B, the corresponding scores are 36.5%/23.9% without EW and 33.7%/21.5% without RA, compared with 45.1%/31.2% for REPO. Across the four settings, removing EW reduces pass@1 by 0.8–1.1 pp, pass@16 by 0.2–0.5 pp, and pass^4 by 6.0–8.6 pp. Removing RA reduces the same metrics by 1.4–1.9,

0.8–1.4, and 8.3–11.4 pp, respectively. Coverage is largely retained across these controls, with the largest differences appearing in repeated success.

Uniform allocation redistributes attenuation by lowering response weights above the question average and raising those below it. Matching absolute token-advantage mass therefore permits changes in the aggregate update direction, because responses contribute different token gradients. The comparison examines how attenuation is allocated within a question, rather than a scalar interpolation between REPO and no weighting. The three evaluation metrics likewise capture different properties: pass^4 averages the fourth power of each question's correctness probability, making it sensitive to how correctness is distributed across questions alongside mean accuracy and solution coverage (Appendix E.2). These controls characterize response-level allocation under the common teacher and training configuration.

D.3.2 Weighting Strength and Allocation Granularity

We vary the weighting strength λ and compare REPO with a question-level confidence weight. All runs use Qwen3-4B-Base and the configuration in Table 4. Each run changes only how α_i is computed for accepted low-consensus questions. Setting $\lambda = 0$ corresponds to *w/o EW*, and $\lambda = 0.25$ is the default REPO setting. Setting $\lambda = 1$ extends Equation 4 to the boundary of its default range $0 \leq \lambda < 1$ and gives $\alpha_i = w_i$. This zero-floor variant drops responses with $s_i \leq 0$ and further attenuates those with $0 < s_i < 1$. The question-level variant assigns one shared multiplier to every response of question x: $\alpha_i = 1 - \lambda + \lambda\gamma_x$ with $\lambda = 0.25$, where $\gamma_x = B^{-1} \sum_{b=1}^{B} \mathbb{I}[\hat{a}^{(b)} = \hat{a}^0]$ is the fraction of replays that reproduce the original target. It uses the same replays as REPO to compare target-level confidence weighting with response-level sign-support weighting. Strictly negative-support responses account for about 0.9% of the valid responses audited among accepted low-consensus questions in Figure 2(b).

Table 5: **Weighting strength and allocation granularity.** Qwen3-4B-Base, external-macro %.

Variant	pass@1	pass@16	pass^4
$\lambda = 0$ (*w/o EW*)	38.5	65.2	21.5
$\lambda = 0.25$ (REPO)	39.5	65.6	28.0
$\lambda = 0.5$	39.6	65.5	29.0
$\lambda = 1$ (zero floor)	39.3	65.3	28.3
Question-level (γ_x)	38.9	65.3	24.0

Evidence weighting mainly affects repeated success (Table 5). Moving from $\lambda = 0$ to $\lambda = 0.25$ raises pass^4 by 6.5 pp, while pass@1 and pass@16 change by 1.0 and 0.4 pp. Across $\lambda \in \{0.25, 0.5, 1\}$, pass^4 ranges from 28.0 to 29.0, indicating similar gains over the unweighted setting. The default recovers most of the observed gain; $\lambda = 0.5$ has the highest point estimate, and zero-floor weighting reaches 28.3. Using the same replays, the question-level multiplier reaches 24.0 pass^4, 4.0 pp below the default response-level allocation. This comparison favors response-level sign support over question-level target-replay confidence. Across weighted variants, pass@1 spans 38.9–39.6 and pass@16 spans 65.3–65.6.

E Theoretical Analysis

E.1 Student Policy Objective

Only the student is optimized by policy gradients. The EMA teacher supplies fixed judgments during each outer batch and has no separate reinforcement-learning loss. Define the proximal policy optimization (PPO) surrogate $h(\rho, A) = \min\{\rho A, \mathrm{clip}(\rho, 1-\epsilon, 1+\epsilon)A\}$ (Schulman et al., 2017), where ρ is the current-to-rollout-policy token probability ratio, A is the fixed advantage, and $\epsilon > 0$ is the clipping radius. The operator $\mathrm{clip}(z, l, u)$ truncates z to the interval $[l, u]$. Rewards, advantages, replay weights, and old log-probabilities are held fixed during differentiation. Token aggregation and reference regularization are fixed across component controls. The teacher's EMA state, the student rollout snapshot, and any reference-policy state used for regularization have distinct roles; evidence weights act only on the student policy term.

E.1.1 Coverage-oriented advantages and dual clipping

The student uses PASS_GRPO (Chen et al., 2025). Let d of the N_s retained answers disagree with the target, and let K be the sampled subset size. The binomial coefficient $\binom{n}{k}$ counts size-k subsets of n items; it is zero when $0 \le n < k$. A uniformly sampled K-subset contains no target match with probability f. Let $\sigma = \sqrt{f(1-f)}$ be the standard deviation of this failure indicator. For $\sigma > 0$, matching and nonmatching responses receive advantages A_+ and A_-, respectively:

$$f = \frac{\binom{d}{K}}{\binom{N_s}{K}}, \qquad \sigma = \sqrt{f(1-f)}, \qquad A_+ = \frac{f}{\sigma}, \qquad A_- = \frac{f - \binom{d-1}{K-1}/\binom{N_s-1}{K-1}}{\sigma}. \tag{9}$$

Each advantage applies only when its response category is present. The advantage computation returns zero below its variance tolerance, including saturated mixed-answer groups; replay preserves this rule. Using dual-clipping coefficient $c_{\mathrm{DC}} > 1$ (Ye et al., 2019), the surrogate and aggregate student policy loss $\mathcal{L}_S$ are

$$h_{\mathrm{DC}}(\rho, A) = \begin{cases} \max\{h(\rho, A), c_{\mathrm{DC}} A\}, & A < 0, \\ h(\rho, A), & A \ge 0, \end{cases} \qquad \mathcal{L}_S = -\sum_{i,\ell} \nu_{i\ell} m_{i\ell} h_{\mathrm{DC}}(\rho_{i\ell}, \widetilde{A}_i). \tag{10}$$

Positive homogeneity of minimum and maximum gives $h_{\mathrm{DC}}(\rho, \alpha A) = \alpha h_{\mathrm{DC}}(\rho, A)$ for $\alpha \ge 0$, establishing Equation 5. This scales individual contributions without bounding relative changes in the aggregate gradient.

E.2 Accuracy, coverage, and repeated-success estimators

At a fixed checkpoint and decoding configuration, let p_x be the correctness probability for question x, and let k be the number of independent decoding attempts. Write $\mathbb{E}_x$ for the average over questions within a dataset. The population metrics are

$$\mathrm{pass@1} = \mathbb{E}_x[p_x], \qquad \mathrm{pass@}k = \mathbb{E}_x[1 - (1-p_x)^k], \qquad \mathrm{pass}^4 = \mathbb{E}_x[p_x^4]. \tag{11}$$

Pass@k measures at least one success (Chen et al., 2021). Requiring every draw to succeed gives the all-correct G-Pass@k criterion for mathematical reasoning (Liu et al., 2024), also used in τ-bench (Yao et al., 2024). In its thresholded form, G-Pass@$k_1 = \mathrm{pass}^k$, where the subscript 1 requires a success fraction of 1 (all k attempts correct); pass^4 sets $k = 4$. This established metric differs from pass@4 and $(\mathbb{E}_x[p_x])^4$.

For c_x correct responses among n independent draws, hats denote sample estimates and the subscript x identifies the question. The unbiased subset estimators are

$$\widehat{\mathrm{pass@1}}_x = \frac{c_x}{n}, \qquad \widehat{\mathrm{pass@}k}_x = 1 - \frac{\binom{n-c_x}{k}}{\binom{n}{k}}, \qquad \widehat{\mathrm{pass}^k}_x = \frac{\binom{c_x}{k}}{\binom{n}{k}}. \tag{12}$$

We use $n = 32$, $k = 16$ for coverage, and $k = 4$ for repeated success, with $\binom{u}{v} = 0$ for $0 \le u < v$. Each distinct k-subset is all-correct with probability p_x^k and all-incorrect with probability $(1-p_x)^k$; averaging these indicators proves unbiasedness. Overlapping subsets remain dependent and do not add independent questions. Dataset scores average per-question estimates; external-macro scores then average the five external dataset scores equally. Convexity of p^4 and concavity of $1-(1-p)^k$ give, for $\mu = \mathbb{E}_x[p_x]$,

$$\mu^4 \le \mathrm{pass}^4 \le \mu \le \mathrm{pass@}k \le 1 - (1-\mu)^k. \tag{13}$$

Thus policies with equal mean accuracy can differ in coverage and repeated success, motivating joint reporting without implying a universal tradeoff.

E.3 Closed-form support and weighting bias–variance

Conditional selection probabilities. Condition on an accepted low-consensus question, the teacher parameter snapshot, its retained responses, candidate set, voting counts, route, and cached judgments; $\mathcal{G}$ is deterministic given these quantities and a target. Let $p_a = t_a/J$ be candidate a's observed True-judgment frequency and $\kappa = \lfloor J\eta \rfloor + 1$ the minimum True count required for eligibility; $\lfloor z \rfloor$ is the greatest integer no larger than z. In one replay, the True count T_a follows a binomial distribution $\mathrm{Binomial}(J, p_a)$: it counts True outcomes in J independent draws, each with probability p_a. These counts are independent across candidates. Define the count probability $P_a(m) = \binom{J}{m} p_a^m (1-p_a)^{J-m}$

and cumulative probability $F_a(m) = \sum_{\ell=0}^{m} P_a(\ell)$ for $m = 0, \ldots, J$, with $F_a(m) = 0$ for $m < 0$ and $F_a(m) = 1$ for $m \geq J$. Thus $P_a(m)$ is the probability of exactly m True verdicts, whereas $F_a(m)$ is the probability of at most m; these describe counts, while $\widehat{P}_a(v)$ in Equation 1 describes a single verdict. For candidate a, let $\mathcal{H}_a$ contain candidates with larger original voting counts; let $\mathcal{E}_a^-$ and $\mathcal{E}_a^+$ contain equal-count candidates preceding and following a in the fixed final tie order. Write q_a for the probability that replay selects candidate a and $q_\perp$ for the probability of abstention:

$$q_a = \sum_{m=\kappa}^{J} P_a(m) \prod_{a' \in \mathcal{H}_a} F_{a'}(\kappa - 1) \prod_{a' \in \mathcal{E}_a^-} F_{a'}(m-1) \prod_{a' \in \mathcal{E}_a^+} F_{a'}(m), \tag{14}$$

$$q_\perp = \prod_{a \in \mathcal{A}_x} F_a(\kappa - 1). \tag{15}$$

Conditioning on $T_a = m$ proves the first expression: higher-frequency competitors must be ineligible, while equal-frequency competitors must lose the True-count and fixed-order comparison. Each product multiplies the relevant competitor probabilities; an empty product equals 1. Candidate independence gives these products. False and invalid outcomes are both included in $1 - p_a$.

For a response with $A_i^0 \neq 0$, let $A(a) = \mathcal{G}(Y, r(a), C(x))$ be the advantage vector for target a, with $A_i(a)$ its ith entry. Define the sign agreement $z_i(a) = \operatorname{sgn}(A_i^0 A_i(a))$, where $\operatorname{sgn}(z)$ is 1, 0, or -1 when z is positive, zero, or negative. Set $z_i(\perp) = 0$ for abstention. The mean support μ_i and single-replay variance v_i are

$$\mu_i = \sum_{a \in \mathcal{A}_x \cup \{\perp\}} q_a z_i(a), \qquad v_i = \sum_{a \in \mathcal{A}_x \cup \{\perp\}} q_a z_i(a)^2 - \mu_i^2. \tag{16}$$

These sums include every candidate target and abstention, with zero-advantage cases contributing zero sign agreement. They characterize the same empirical resampling distribution used during training.

Finite-replay error. Here $\mathbb{E}$ and Var denote expectation and variance over replay draws, conditional on the fixed cache and responses. Independent replay batches give $\mathbb{E}[s_i] = \mu_i$ and $\mathrm{Var}(s_i) = v_i/B$. Let $\alpha_i^* = 1 - \lambda + \lambda[\mu_i]_+$ be the infinite-replay multiplier. Convexity and the 1-Lipschitz property of the positive part yield

$$0 \leq \mathbb{E}[\alpha_i] - \alpha_i^* \leq \lambda\sqrt{v_i/B}, \qquad \mathbb{E}[(\alpha_i - \alpha_i^*)^2] \leq \lambda^2 v_i/B. \tag{17}$$

Thus clipping introduces a bounded upward bias relative to the infinite-replay multiplier. At a fixed parameter value, let g_i be response i's unweighted student surrogate-ascent gradient, including token-loss aggregation. With the multipliers held constant during differentiation, let $\widehat{G} = \sum_i \alpha_i g_i$ and $G_* = \sum_i \alpha_i^* g_i$ be the finite- and infinite-replay aggregate gradients. Writing $\|\cdot\|$ for the Euclidean norm, Minkowski's inequality gives

$$\mathbb{E}\|\widehat{G} - G_*\|^2 \leq \frac{\lambda^2}{B}\left(\sum_i \|g_i\|\sqrt{v_i}\right)^2. \tag{18}$$

This bound allows dependence among responses sharing the same replay.

Across-rollout reweighting. Fix the training history and current student and teacher snapshots, then vary accepted low-consensus questions, responses, and caches, assuming finite second moments. In this paragraph, $\mathbb{E}$ averages over this rollout randomness. Let $G_0 = \sum_i g_i$ be the unweighted aggregate gradient and $R = \sum_i (1 - \alpha_i^*) g_i$ the contribution removed by weighting. Then $G_* = G_0 - R$, hence

$$\mathbb{E}[G_*] - \mathbb{E}[G_0] = -\mathbb{E}[R], \tag{19}$$

$$\mathcal{V}(G_*) = \mathcal{V}(G_0) + \mathcal{V}(R) - 2C_g(G_0, R), \tag{20}$$

where $\mathcal{V}(U) = \mathbb{E}\|U - \mathbb{E}U\|^2$ is the total variance of a random gradient vector U, and $C_g(U, W) = \mathbb{E}[(U - \mathbb{E}U)^\top (W - \mathbb{E}W)]$ sums the coordinatewise covariances of random gradient vectors U and W. The superscript $\top$ denotes transpose, so the product is an inner product. Reweighting does not increase aggregate student-gradient variance exactly when $2C_g(G_0, R) \geq \mathcal{V}(R)$; strict reduction requires a strict inequality. The mean shift is relative to the unweighted update, not an unknown clean-reward gradient: reproducible teacher errors can remain fully weighted. Thus weighting can introduce bias and need not reduce variance; increasing B controls only the conditional replay

approximation. This comparison concerns updates from the same sampling distribution, not the variance of different training trajectories.

F LIMITATIONS

Our main evaluation covers two Qwen3 scales and five external benchmarks, with component controls on AMC and DAPO and a weighting-strength study at 4B across the external benchmarks. These controls characterize evidence weighting within the shared teacher, selector, and advantage configuration; whole-method comparisons also reflect those components. Evidence support measures update reproducibility under a fixed verification cache, so systematic teacher errors can remain highly supported. The trajectory audit describes supervision encountered during one REPO training run. The weighting-strength study covers the reported settings within this training configuration.

Table 6: Illustrative questions and reference answers. GPQA answer choices are ordered as displayed.

Type	Content
	AIME 2024
Question	Every morning Aya goes for a 9-kilometer-long walk and stops at a coffee shop afterwards. When she walks at a constant speed of s kilometers per hour, the walk takes her 4 hours, including t minutes spent in the coffee shop. When she walks $s+2$ kilometers per hour, the walk takes her 2 hours and 24 minutes, including t minutes spent in the coffee shop. Suppose Aya walks at $s+\frac{1}{2}$ kilometers per hour. Find the number of minutes the walk takes her, including the t minutes spent in the coffee shop.
Answer	$\boxed{204}$
	AIME 2025
Question	Find the sum of all integer bases $b > 9$ for which 17_b is a divisor of 97_b.
Answer	$\boxed{70}$
	AMC
Question	Let $\frac{m}{n}$ be the reduced form of $3 + \dfrac{1}{3 + \frac{1}{3+\frac{1}{3}}}$. Find $m+n$.
Answer	$\boxed{142}$
	MATH-500
Question	Convert the point $(0, 3)$ in rectangular coordinates to polar coordinates. Enter your answer in the form (r, θ), where $r > 0$ and $0 \le \theta < 2\pi$.
Answer	$\boxed{\left(3, \frac{\pi}{2}\right)}$
	GPQA
Question	Which of the following (effective) particles is not associated with a spontaneously-broken symmetry? **(A)** Phonon **(B)** Magnon **(C)** Pion **(D)** Skyrmion
Answer	$\boxed{\text{D}}$ — Skyrmion.
	DAPO
Question	How many ways are there to partition 7 students into groups of 2 or 3?
Answer	$\boxed{105}$