DELTA-GRPO: INTEGRATING OPD-STYLE POLICY SHIFT INTO REINFORCEMENT LEARNING

Anonymous authors
Paper under double-blind review

ABSTRACT

On-policy distillation (OPD) transfers teacher knowledge through dense token-level feedback on trajectories generated by the student. Recent work on policy-shift distillation draws supervision from the change between a teacher's pre- and post-reinforcement-learning (RL) policies, showing that RL-induced change can serve as a learning signal. Meanwhile, group relative policy optimization (GRPO) learns from outcome rewards by assigning a shared advantage to all tokens in a response, leaving accumulated token-level changes unused as explicit guidance. This motivates integrating OPD-style policy-shift signals into ongoing RL to complement response-level feedback. We propose DELTA-GRPO, which combines the dense guidance of OPD with the outcome-based optimization of GRPO. It uses the training policy's accumulated shift to guide token-level credit allocation within outcome-driven optimization. Experiments with Qwen3-4B and Qwen3-8B on eight mathematics benchmarks show that DELTA-GRPO outperforms the evaluated RL and OPD baselines. It improves macro accuracy over GRPO by 13.59 and 12.04 percentage points at the two model sizes, respectively, and exceeds SC-GRPO by 8.90 and 7.89 points.

1 INTRODUCTION

On-policy distillation (OPD) transfers a teacher's knowledge on trajectories generated by the student itself (Gu et al., 2024; Agarwal et al., 2024). The student samples a response, and the teacher supplies distributional feedback at the prefixes encountered along that response (Hinton et al., 2015; Tunstall et al., 2024). Training then adjusts the student's token predictions using this feedback. This procedure provides supervision on the states the student actually visits (Agarwal et al., 2024). For reasoning, the relevant property is the granularity of supervision, which is also central to process-supervision and self-improvement approaches (Uesato et al., 2022; Wang et al., 2024; Lightman et al., 2024; Zelikman et al., 2022; 2024): teacher feedback can vary across the tokens of a long solution, whereas a final-answer correctness judgment summarizes the response with a single outcome. OPD thus provides a way to learn from dense feedback on self-generated reasoning trajectories.

Policy-shift distillation changes what information is transferred from the teacher. ExOPD, the reward-extrapolation variant of generalized OPD, connects distillation to reinforcement learning (RL) by amplifying the implicit reward defined by the teacher-to-reference log-probability ratio (Yang et al., 2026). Direct-OPD uses the difference between a teacher's post-RL and pre-RL token log probabilities as supervision on student-generated trajectories (Feng et al., 2026). Evaluated at the same prefix, this difference indicates which token choices became more or less likely through RL and records the teacher's training-induced change. OPD^2 likewise uses the log-probability difference between a reasoning-tuned teacher and its base model as a distillation reward (Heo et al., 2026). Its objective transfers the teacher's reasoning-related changes to a student; we use the learner's own accumulated changes to redistribute outcome credit within each response. This motivates a further use of policy shift: an RL learner also accumulates such changes, which can supply dense guidance during its own training.

Outcome-based RL offers a complementary source of supervision by evaluating generated answers against task objectives (Christiano et al., 2017; Stiennon et al., 2020; Shao et al., 2024; Guo et al., 2025). Group relative policy optimization (GRPO) samples several responses to a prompt and normalizes their rewards within the group to obtain a response-level advantage (Schulman et al.,

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

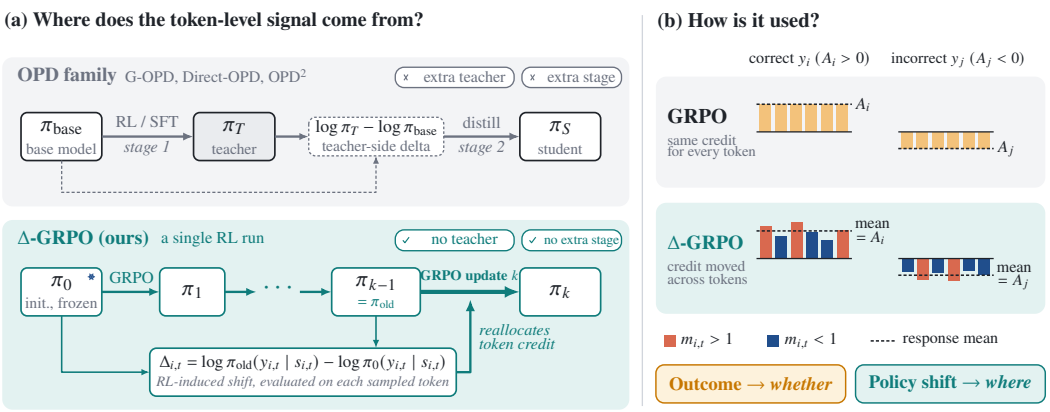


Figure 1: High-level view of DELTA-GRPO. OPD-style policy shift supplies dense token guidance, while GRPO supplies outcome-level direction; the combined update redistributes credit within each sampled response.

2016; Shao et al., 2024; Zheng et al., 2025). In its outcome-supervised form, the same advantage is assigned to every token in a response. It indicates whether the response should be reinforced or suppressed relative to its peers while leaving the roles of individual token choices unresolved. SC-GRPO conditions the model on its own verified solutions and uses the token-level KL divergence between the original and conditioned distributions to reweight GRPO updates (Shan et al., 2026). EP-GRPO combines entropy-based token weighting with outcome-anchored policy/reference divergence signals, normalizing the latter within cumulative-entropy progress buckets to construct token-level advantages (Yu et al., 2026). These approaches motivate examining how a dense signal's source and its integration with outcome feedback shape token credit allocation (Cui et al., 2026; 2025; Wang et al., 2025).

These developments motivate our central question: can OPD-style policy-shift supervision guide token-level credit allocation within GRPO? Distillation and RL can already be combined through teacher feedback and task rewards (Agarwal et al., 2024); here, we use the learner's own accumulated policy shift as the dense signal. At each update, the pre-update policy and its frozen initialization provide a checkpoint pair that can be evaluated on the current rollouts. The verifier supplies fresh outcome information, while this pair records how earlier learning changed conditional token preferences.

We propose DELTA-GRPO, which combines OPD-style policy-shift guidance with the outcome-based optimization of GRPO. Outcome advantages determine whether tokens are reinforced or suppressed, while historical shift modulates their relative update magnitudes within the same response. To implement this separation, we align the shift with the advantage sign and then center it within each response. The shift is measured as the sampled-token log-probability change from initialization to the current pre-update policy, with clipping and normalization to bound its scale. A bounded, detached multiplier converts the centered alignment into token-specific advantages. Centering gives the multiplier a mean of one, while its positive bound preserves the sign of the original advantage. The mechanism redistributes outcome credit within a response while preserving its arithmetic mean and reinforcement direction.

This construction makes the OPD–RL combination explicit in the optimization. The token advantage decomposes into the original GRPO advantage and a centered policy-shift term gated by the magnitude of that advantage (Section 3). The historical term is zero for responses with zero group-relative advantage. The method derives dense guidance from the learner's own checkpoint pair (Tunstall et al., 2024; Rafailov et al., 2023). Our analysis establishes mean and sign preservation and derives the local gradient contribution of the dense signal. Preserving mean credit still changes the policy gradient because tokens have different score gradients; this is how historical guidance affects learning within the outcome-driven update.

We evaluate DELTA-GRPO on Qwen3-4B and Qwen3-8B across eight mathematics benchmarks against both RL and OPD baselines. It improves macro accuracy over GRPO by 13.59 and 12.04

percentage points, respectively, and outperforms the strongest baseline, SC-GRPO, by 8.90 and 7.89 points. Component ablations further support the roles of outcome alignment and response centering.

Our contributions are threefold:

- We formulate the integration of OPD-style policy-shift supervision into outcome-based RL, using the learner's own accumulated policy changes to complement GRPO's response-level feedback with dense token-level guidance.
- We propose DELTA-GRPO, which combines outcome alignment and response centering to redistribute credit across tokens. We establish that its bounded modulation preserves the mean and sign of response advantages while introducing a policy-shift contribution to the local gradient.
- Experiments on Qwen3-4B and Qwen3-8B across eight mathematics benchmarks demonstrate improvements over the evaluated RL and OPD baselines. Ablations support the roles of outcome alignment and response centering.

2 RELATED WORK

On-policy distillation and policy-shift supervision. OPD trains a student using teacher feedback on student-generated trajectories, bringing distributional supervision to the states visited by the learner (Gu et al., 2024; Agarwal et al., 2024). This provides dense token-level feedback on reasoning states. ExOPD interprets OPD through a generalized RL objective with a teacher-to-reference log-probability reward and a separately weighted KL constraint. It extrapolates the implicit reward to strengthen supervision beyond standard distribution matching (Yang et al., 2026). Direct-OPD specifically transfers the log-probability shift between a teacher's post-RL and pre-RL checkpoints, using the change induced by RL to supervise student-generated trajectories (Feng et al., 2026). OPD^2 derives a delta reward from a reasoning-tuned teacher and its base model, centering the distillation signal over candidate tokens under the student's distribution at a given prefix (Heo et al., 2026). In DELTA-GRPO, centering operates across positions in an already sampled response to redistribute its current outcome advantage while preserving the response mean. These methods connect distillation to policy optimization through dense log-policy signals. Our work brings this perspective into ongoing outcome-based RL: the checkpoint pair consists of the learner's pre-update policy and its own initialization, and the resulting shift modulates GRPO credit inside the same optimization objective.

Outcome-based reinforcement learning. RL with verifiable rewards improves reasoning policies by assigning rewards to answers whose correctness can be checked automatically (Cobbe et al., 2021; Uesato et al., 2022; Lightman et al., 2024; Shao et al., 2024; Guo et al., 2025; Yu et al., 2025). A rollout contains a prompt, several sampled completions, and a scalar reward for each completion. The reward can combine correctness with task-specific shaping, such as a length term, so that learning reflects both solution quality and generation behavior. Policy-gradient updates inherit trust-region and clipped-ratio formulations from policy optimization (Schulman et al., 2015; 2017; Hu et al., 2025) and reinforce completions that outperform their group while reducing the probability of completions that underperform.

GRPO makes this procedure practical for language models by estimating the advantage from reward statistics inside each prompt group, eliminating a separately trained value model (Shao et al., 2024; Schulman et al., 2017). The resulting advantage gives every token in a response the same outcome-level coefficient. This creates a stable response-level learning signal and keeps the optimization tied to verifiable task feedback, while leaving the relative contribution of individual reasoning tokens unspecified. Token-credit methods build on this common objective by adding dense signals that distinguish positions within the sampled trajectory (Kazemnejad et al., 2025; Wang et al., 2024; Yuan et al., 2025); DELTA-GRPO uses the policy's accumulated shift for that purpose.

Token-level credit assignment in GRPO. SC-GRPO and EP-GRPO refine the granularity of outcome-based updates using different dense signals, following a broader line of process-supervision and token-credit work (Wang et al., 2024; Lightman et al., 2024; Shan et al., 2026; Yu et al., 2026). SC-GRPO conditions the model on its own verified solutions and measures token-level KL divergence between the original and solution-conditioned distributions. A bounded multiplicative weight

suppresses low-divergence token updates while preserving more of the update at high-divergence tokens (Shan et al., 2026). EP-GRPO combines entropy-gated outcome advantages with an outcome-anchored policy/reference log-ratio signal, normalized within cumulative-entropy progress buckets (Yu et al., 2026). Thus, SC-GRPO uses sensitivity to solution conditioning, whereas EP-GRPO combines uncertainty and policy divergence with progress alignment. Our signal measures signed change from initialization. Its distinctive use is response-local, mean-preserving redistribution: advantage-sign alignment and within-response centering produce positive, bounded multipliers with mean one. This preserves the original advantage's sign and arithmetic mean while adding historical guidance to the gradient. The combination of OPD-style policy shift and these outcome-credit constraints defines DELTA-GRPO's relationship to the existing token-credit approaches.

3 DELTA-GRPO

We propose DELTA-GRPO to integrate OPD-style policy-shift guidance into outcome-based GRPO. The method uses two complementary signals on the same sampled responses: a group-relative advantage from outcome rewards and a token-level record of change from the policy's initialization. We first define the rollout and reward setting, then construct the historical signal and use it to redistribute credit within each response. The resulting token advantages enter a single policy optimization objective.

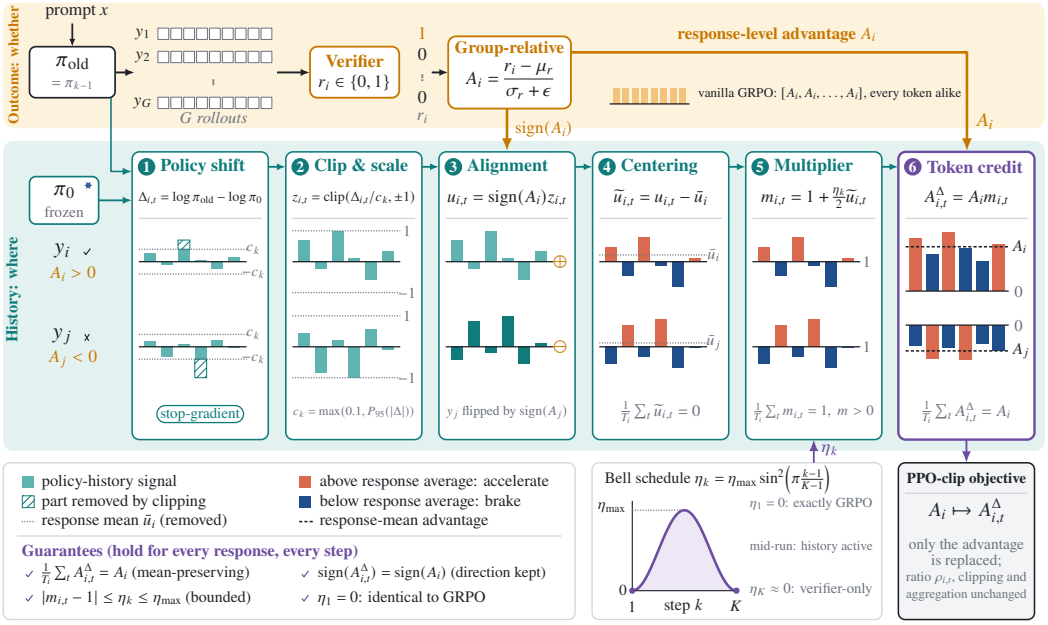


Figure 2: Mechanism used by DELTA-GRPO. The accumulated policy shift is aligned with the sign of the group advantage, centered over positions in the sampled response, and applied as a bounded mean-one multiplier.

3.1 PROBLEM FORMULATION AND ROLLOUT SAMPLING

Let x be a reasoning prompt and $\pi_\theta(y \mid x)$ the trainable policy. At update k, we sample G responses $\{y_i\}_{i=1}^{G}$ from the pre-update policy π_{old}. Response y_i contains T_i valid tokens, with prefix $s_{i,t} = (x, y_{i,<t})$. We retain a frozen initialization π_0 throughout training and synchronize π_{old} with the actor before each update. The two policies evaluate identical sampled tokens and prefixes.

Each response receives a scalar reward r_i based on answer correctness and the specified length penalty (Appendix C.1). Following GRPO (Shao et al., 2024), we compute the advantage within each

prompt group:

$$A_i = \frac{r_i - \mu_r}{\sigma_r + \epsilon}, \qquad \mu_r = \frac{1}{G}\sum_{j=1}^{G} r_j, \qquad \sigma_r^2 = \frac{1}{G-1}\sum_{j=1}^{G}(r_j - \mu_r)^2, \tag{1}$$

where $\epsilon = 10^{-6}$ and $G > 1$. A positive advantage reinforces a response relative to its group, while a negative advantage suppresses it. Its sign reflects the shaped relative reward. Standard outcome-based GRPO assigns A_i to all valid tokens of y_i. Our objective supplements this response-level signal with historical guidance while preserving its mean and sign.

3.2 Accumulated policy-shift signal

Policy-shift distillation extracts dense supervision from the difference between a trained teacher and its earlier checkpoint (Feng et al., 2026; Heo et al., 2026). We apply this idea to the learner itself. The initialization supplies the reference point, and the pre-update policy records the cumulative change produced by earlier RL updates. For each sampled token, we compute

$$\Delta_{i,t} = \mathrm{sg}[\log \pi_{\mathrm{old}}(y_{i,t} \mid s_{i,t}) - \log \pi_0(y_{i,t} \mid s_{i,t})], \quad z_{i,t} = \frac{\mathrm{clip}(\Delta_{i,t}, -c_k, c_k)}{c_k}, \tag{2}$$

where sg denotes stop-gradient and $c_k = \max\{0.1, P_{95}(|\Delta|)\}$. The percentile uses all valid response tokens in the collected global rollout batch; padding is excluded. The normalization bounds $z_{i,t}$ in $[-1, 1]$ while retaining the shift's sign. Positive shift means that preceding training increased the probability of the sampled token relative to initialization, evaluated on the same prefix. The quantile limits the effect of extreme changes and adapts the scale as history accumulates. The signal records the policy's cumulative change from initialization on the current rollouts. A fixed initialization makes the shift comparable across updates and lets the modulation accumulate learning history across the training trajectory. Appendix B.3 gives the connection to OPD's log-policy supervision.

3.3 Outcome-aligned credit redistribution

We use policy history to vary update magnitudes across a response while retaining the direction specified by its outcome advantage. To orient the historical signal, define $u_{i,t} = \mathrm{sign}(A_i) z_{i,t}$. For a positive-advantage response, historically promoted tokens have positive alignment. For a negative-advantage response, historically suppressed tokens have positive alignment. We center this alignment over the T_i valid tokens and construct

$$\bar{u}_i = \frac{1}{T_i}\sum_t u_{i,t}, \qquad m_{i,t} = \mathrm{sg}\Big[1 + \frac{\eta_k}{2}(u_{i,t} - \bar{u}_i)\Big], \qquad A^{\Delta}_{i,t} = \mathrm{sg}[A_i m_{i,t}]. \tag{3}$$

Here $\eta_k \in [0, 1)$ controls the strength of modulation. Centering removes the component shared by the response, making allocation depend on each token's alignment relative to the other sampled positions. This is an arithmetic mean over response positions; OPD^2 centers its signal over candidate tokens at a fixed prefix. A bounded, positive multiplier changes how strongly each token is reinforced or suppressed.

Expanding Equation 3 gives the central combination:

$$A^{\Delta}_{i,t} = \underbrace{A_i}_{\text{outcome-based RL}} + \underbrace{\frac{\eta_k}{2}|A_i|(z_{i,t} - \bar{z}_i)}_{\text{outcome-gated OPD-style policy-shift guidance}}. \tag{4}$$

Here $\bar{z}_i = T_i^{-1}\sum_t z_{i,t}$. The second term always follows the centered shift in signed advantage space, while its *relative effect on magnitude* depends on the sign of A_i. On a negative-advantage response, a positive centered shift makes the advantage less negative, braking suppression; a negative centered shift strengthens suppression. On a positive-advantage response, the same two cases strengthen and weaken reinforcement, respectively. This is how OPD-style guidance and RL outcomes interact in the update.

3.4 Optimization objective and properties

We optimize a single GRPO surrogate using the redistributed advantages. Let $N = \sum_i T_i$ denote the total number of valid tokens in the rollout batch, with i now indexing responses across prompt groups, and $\rho_{i,t}(\theta) = \pi_\theta(y_{i,t} \mid s_{i,t})/\pi_{\text{old}}(y_{i,t} \mid s_{i,t})$. We maximize the token-mean clipped surrogate

$$J_\Delta(\theta) = \frac{1}{N} \sum_{i,t} \min\left\{\rho_{i,t} A^\Delta_{i,t},\ \text{clip}(\rho_{i,t}, 1-\epsilon_-, 1+\epsilon_+) A^\Delta_{i,t}\right\}. \tag{5}$$

Here ϵ_- and ϵ_+ are the lower and upper importance-ratio clipping widths. J_{GRPO} denotes the same objective with $A^\Delta_{i,t}$ replaced by A_i. The quantities A_i, z, and m are detached, and all sums use valid response tokens. The normalization by N averages over tokens, so longer responses contribute more terms.

The token advantages satisfy the following conservation and boundedness properties.

Proposition 1 *For a nonempty response and $0 \le \eta_k < 1$, Equation 3 satisfies*

$$\frac{1}{T_i} \sum_t m_{i,t} = 1, \qquad \frac{1}{T_i} \sum_t A^\Delta_{i,t} = A_i, \qquad 1 - \eta_k \le m_{i,t} \le 1 + \eta_k. \tag{6}$$

Every nonzero token advantage has the sign of A_i, and $A_i = 0$ implies $A^\Delta_{i,t} = 0$ for all t.

The proof follows from zero-mean centering and $u_{i,t} \in [-1, 1]$ (Appendix B). These properties preserve the arithmetic mean of the response's credit while changing its parameter gradient through token-specific score gradients.

At $\theta = \theta_{\text{old}}$, the importance ratios equal one and clipping is inactive. Substituting Equation 4 gives the local gradient decomposition

$$\nabla J_\Delta = \nabla J_{\text{GRPO}} + \frac{1}{N} \sum_i \frac{\eta_k |A_i|}{2} \sum_t (z_{i,t} - \bar{z}_i) \nabla_\theta \log \pi_\theta(y_{i,t} \mid s_{i,t}). \tag{7}$$

The first term is the outcome-based GRPO gradient. The second is OPD-style historical guidance, scaled by the current outcome advantage magnitude. Its scalar coefficients sum to zero within each response, while token score gradients differ, producing a distinct optimization direction. The historical term is zero when $A_i = 0$; setting $\eta_k = 0$ recovers GRPO. Equation 5 defines the update for the full clipped-ratio regime.

3.5 Training schedule and implementation

We use a finite-horizon bell schedule,

$$\eta_k = \eta_{\max} \sin^2\left(\frac{\pi(k-1)}{K-1}\right), \qquad k = 1, \ldots, K. \tag{8}$$

Modulation starts at zero, peaks near the middle of training, and returns to zero at the final update. The initial policy has little accumulated shift, while the later phase reduces historical modulation as training ends. The maximum multiplier range is $[1 - \eta_{\max}, 1 + \eta_{\max}]$. Appendix C.1 gives the evaluated values of K and $\eta_{\max}$.

The policy-shift signal is computed from the initialization and the pre-update policy on the current rollout batch, then combined with the GRPO objective in Equation 5. Appendix A gives the complete procedure and Appendix C.1 records the training configuration. Inference uses the trained actor.

4 Experiments

We address five research questions, following the order of the analyses in Section 5.

RQ1: How does Delta-GRPO compare with RL and OPD baselines in accuracy and sampled-solution coverage?

RQ2: What do outcome alignment and response centering contribute to performance?

RQ3: How does performance vary across the evaluated modulation-strength settings?

RQ4: Do gains persist at equal expected output-token budgets, and what additional training cost does Delta-GRPO incur?

RQ5: How is token credit redistributed locally, and how do response length, token entropy, and reward-group activity evolve during training?

Table 1: Average accuracy (%) with the best scores in bold. Macro gains are relative to SC-GRPO, the strongest baseline. MATH: MATH500; A24/A25/A26: AIME 2024/2025/2026; Minerva: MinervaMath; Olymp.: OlympiadBench; HMMT25: HMMT Feb 2025.

(a) Qwen3-4B Macro gain: +8.90 pp

Method	MATH	A24	A25	AMC23	Minerva	Olymp.	HMMT25	A26	Macro
Base	84.42	21.04	18.02	66.41	40.03	53.49	11.15	17.40	38.99
GRPO	92.10	41.35	35.31	87.11	43.84	67.06	18.85	36.77	52.80
SC-GRPO	92.67	52.08	42.08	90.39	44.03	69.57	25.00	44.06	57.49
EP-GRPO	79.03	31.56	26.04	67.81	33.82	56.79	17.50	24.27	42.10
ExOPD	91.42	42.71	34.06	87.97	43.80	66.39	19.06	40.62	53.26
Direct-OPD	91.38	37.50	29.48	81.88	43.66	64.39	18.12	32.29	49.84
Delta-GRPO	**95.50**	**66.88**	**55.21**	**96.17**	**46.23**	**74.87**	**36.98**	**59.27**	**66.39**

(b) Qwen3-8B Macro gain: +7.89 pp

Method	MATH	A24	A25	AMC23	Minerva	Olymp.	HMMT25	A26	Macro
Base	84.05	24.69	19.58	68.12	40.67	54.62	9.79	16.04	39.70
GRPO	92.70	51.25	35.52	87.03	43.66	69.21	21.67	42.19	55.40
SC-GRPO	93.55	58.65	43.96	90.78	42.92	71.07	27.08	48.44	59.56
EP-GRPO	60.50	23.12	18.96	49.30	32.26	40.30	11.15	16.56	31.52
ExOPD	92.07	51.77	34.58	86.41	43.84	68.29	23.54	43.54	55.51
Direct-OPD	91.22	39.69	27.50	82.11	42.10	65.10	16.35	31.77	49.48
Delta-GRPO	**95.70**	**71.04**	**57.50**	**93.05**	**47.10**	**75.91**	**37.92**	**61.35**	**67.45**

4.1 Models and training configuration

We train Qwen3-4B and Qwen3-8B (Qwen et al., 2024; Yang et al., 2024; 2025) on 17,920 processed DAPO mathematics prompts (Yu et al., 2025). The principal GRPO and Delta-GRPO runs share the rollout recipe, initialization, input order, optimizer, and three training seeds. Full training settings and reward shaping are given in Appendix C.1; the implementation uses verl and vLLM (Sheng et al., 2025; Kwon et al., 2023).

4.2 Evaluation protocol and metrics

We evaluate on MATH500 (Hendrycks et al., 2021), AIME24, AIME25, AMC23, MinervaMath (Lewkowycz et al., 2022), OlympiadBench (He et al., 2024), HMMT Feb 2025, and AIME26 using three random seeds. We report average per-question sampled accuracy (avg@n) and macro accuracy, the equal-weight mean over the eight benchmarks. Sampling and answer-scoring details appear in Appendix C.2.

Accuracy measures the quality of individual samples. To measure coverage, we additionally report pass@k, estimated as $1 - \binom{n-c}{k} / \binom{n}{k}$ for a question with c correct responses among n samples (Chen et al., 2021). This estimates the probability of at least one correct response in a size-k subset of the sample bank. We report response lengths and logged training time separately to characterize the associated resource use. Accuracy captures the quality of one sampled solution, whereas pass@k captures whether a collection contains at least one solution. Reporting both separates individual-sample quality from solution coverage.

Table 2: Component ablations on Qwen3-4B.

Method	MATH500	AIME24	AIME25	AMC23	Minerva	Olympiad	HMMT25	AIME26	Macro
GRPO	92.10	41.35	35.31	87.11	43.84	67.06	18.85	36.77	52.80
No alignment	93.27	46.46	38.65	87.89	44.39	68.43	21.15	40.94	55.15
No centering	95.15	63.85	51.35	93.36	**46.60**	73.91	**38.65**	54.48	64.67
Delta-GRPO	**95.50**	**66.88**	**55.21**	**96.17**	46.23	**74.87**	36.98	**59.27**	**66.39**

4.3 Baselines and comparison protocol

We compare five methods spanning outcome-based RL and teacher-based OPD. For each baseline, we follow the training protocol and hyperparameters reported in its original paper. The RL baselines comprise GRPO, SC-GRPO with self-conditioned token credit (Shan et al., 2026), and EP-GRPO with entropy-progress guidance (Yu et al., 2026), while the OPD baselines are ExOPD and Direct-OPD (Yang et al., 2026; Feng et al., 2026). The GRPO pair controls the intervention most directly; implementation and rollout-budget details for all baselines are recorded in Appendix C.3.

5 Results and Analysis

5.1 Main results and solution coverage

Delta-GRPO obtains the highest average accuracy on every benchmark at both sizes in Table 1. Its 4B macro accuracy is 66.39%, compared with 52.80% for GRPO and 57.49% for SC-GRPO. At 8B, the corresponding values are 67.45%, 55.40%, and 59.56%. The gains over GRPO are 13.59 and 12.04 percentage points, with paired 95% question-bootstrap intervals of [11.17, 16.09] and [10.10, 14.00]. The gains over SC-GRPO are 8.90 [7.03, 10.89] and 7.89 [6.20, 9.62] points. Removing any one benchmark leaves a positive macro gain over GRPO: 11.88–15.19 points at 4B and 10.62–13.33 at 8B. The improvement therefore spans the benchmark suite and appears at both model sizes, with the same policy-shift mechanism applied throughout.

The OPD baselines test using a frozen teacher's dense signal in a separate student run. ExOPD reaches 53.26% and 55.51% macro accuracy, and Direct-OPD reaches 49.84% and 49.48%. The stronger results of Delta-GRPO support combining history with ongoing outcome feedback under the evaluated recipes; their protocol and compute details are documented in Appendix C.3.

The accuracy gains also extend to sampled-solution coverage. On AIME24/25, Delta-GRPO has higher point estimates than GRPO at all tested budgets $k \in \{1, 2, 4, 8, 16, 32\}$. At 4B, empirical pass@32 improves from 73.33/70.00% to 90.00/90.00%; at 8B, it improves from 86.67/66.67% to 90.00/90.00%. AIME sets contain 30 questions each, so the 8B AIME24 endpoint difference is one additional observed solved question. Appendix D.1 supplies the curves and uncertainty. These measurements distinguish single-sample quality from coverage within a finite sample bank.

5.2 Alignment and centering ablations

Table 2 tests the two operations that connect policy history to outcome credit. The no-alignment variant centers z without multiplying by $\mathrm{sign}(A_i)$. For fixed inputs, it agrees with the full method on positive advantages and reverses the additive redistribution on negative advantages. Its 4B macro accuracy is 55.15%, compared with 66.39% for the full method, a paired difference of 11.24 [9.10, 13.49] points. The result supports orienting historical information relative to the outcome update.

Removing centering yields 64.67%, a drop of 1.72 [0.56, 2.89] points. The full method improves six of eight benchmarks, while the uncentered variant is slightly better on MinervaMath and HMMT25. At step 70, the standard deviation of response-mean multipliers is 4.1×10^{-8} with centering and 0.00175 without it. In the uncentered diagnostic, response-mean multipliers correlate with advantage at 0.762 and with length at -0.300. Centering removes this response-wide shift by construction. Table 3 compares the modulation-strength settings.

Table 3: Accuracy (%) across modulation-strength settings on Qwen3-4B.

$\eta_{\max}$	MATH500	AIME24	AIME25	AMC23	Minerva	Olympiad	HMMT25	AIME26	Macro
0.01	92.92	44.58	32.92	87.66	45.86	67.84	18.12	38.65	53.57
0.03	94.58	54.58	42.19	91.56	**46.65**	69.99	27.19	47.60	59.29
0.05	**95.50**	**66.88**	**55.21**	**96.17**	46.23	**74.87**	**36.98**	**59.27**	**66.39**

5.3 Exploratory modulation strength

All three modulation-strength settings use the same training configuration and 140 updates. Their macro accuracies are 53.57%, 59.29%, and 66.39%, respectively. The 0.05 setting improves all eight benchmarks over 0.01 and seven over 0.03, with MinervaMath slightly lower (46.23% versus 46.65%).

5.4 Equal-token-budget gains and compute cost

At matched expected output-token budgets, Delta-GRPO gains 15.35 points on 4B AIME24, 14.32 on 4B AIME25, 2.30 on 8B AIME24, and 20.42 on 8B AIME25. Paired intervals include zero for 4B AIME25 and 8B AIME24. Training costs 827.2 seconds per Delta-GRPO step versus 531.5 for GRPO in the matched 4B logs, a 55.6% increase. This additional per-step computation is accompanied by substantial gains in accuracy, solution coverage, and update efficiency, giving a favorable quality–compute trade-off in the matched comparison.

5.5 Training dynamics and local credit allocation

The frozen step-100 audit covers 64 active responses (32 with positive and 32 with negative advantages). The top 10% of tokens ranked by $|m - 1|$ carry 36–37% of the total absolute multiplier deviation. The lowest normalized concentration statistic, $(\sum_t m_t)^2/(T_i \sum_t m_t^2)$, is 0.999973493: local deviations coexist with near-uniform multiplier mass. Centering preserves mean advantage, and zero-advantage responses have $m = 1$ and $A^{\Delta}_{i,t} = 0$. Distinct token score gradients allow these local reallocations to change the update direction while preserving the mean response advantage.

Table 4: Training dynamics for matched Qwen3-4B runs.

Metric	GRPO	Delta-GRPO
Initial correctness (%)	44.14	44.14
Step 20 online MATH500 (%)	83.90	84.33
Step 70 mean response length	3,931.5	6,183.4
Step 140 token entropy	0.3317	0.0618

In the matched Qwen3-4B run, it reaches 92.03% online MATH500 accuracy at step 60, exceeding GRPO's step-140 accuracy of 91.80%. Delta-GRPO subsequently reaches 93.98% at step 140. The same rollout supplies outcome feedback and policy-shift guidance, making the movement of credit within each response directly measurable. Figure 5 and Appendix E provide the corresponding trajectories and detailed analysis.

6 Discussion and Conclusion

Delta-GRPO combines OPD-style policy-shift guidance with GRPO outcome supervision, turning accumulated policy changes into dense token-level guidance. Outcome alignment and response centering redistribute token credit while preserving response-level direction and mean credit. Results and ablations across two Qwen3 sizes and eight mathematics benchmarks support this integration, with gains in accuracy and sampled-solution coverage. Matched training trajectories reach comparable validation accuracy in fewer updates and achieve higher final performance. Gains persist after expected-token normalization at additional training cost. The unified objective keeps outcome rewards as the source of response direction while reusing policy history to refine token priorities. This provides a direct bridge between dense OPD supervision and practical outcome-based RL. The resulting formulation also offers a simple interface for future token-credit signals within GRPO.

AI USE STATEMENT

In this study, large language models (LLMs) were used to aid and polish the writing, support literature retrieval and discovery, and assist with research execution, including code writing and mathematical exposition. All AI-assisted material was reviewed and verified by the authors, who retain full responsibility for the content, originality, and accuracy of this manuscript.

REFERENCES

Rishabh Agarwal, Nino Vieillard, Yongchao Zhou, Piotr Stanczyk, Sabela Ramos Garea, Matthieu Geist, and Olivier Bachem. On-policy distillation of language models: Learning from self-generated mistakes. In B. Kim, Y. Yue, S. Chaudhuri, K. Fragkiadaki, M. Khan, and Y. Sun (eds.), *International Conference on Learning Representations*, volume 2024, pp. 21246–21263, 2024. URL https://proceedings.iclr.cc/paper_files/paper/2024/hash/5be69a584901a26c521c2b51e40a4c20-Abstract-Conference.html.

Mark Chen, Jerry Tworek, Heewoo Jun, Qiming Yuan, Henrique Ponde de Oliveira Pinto, Jared Kaplan, Harri Edwards, Yuri Burda, Nicholas Joseph, Greg Brockman, Alex Ray, Raul Puri, Gretchen Krueger, Michael Petrov, Heidy Khlaaf, Girish Sastry, Pamela Mishkin, Brooke Chan, Scott Gray, Nick Ryder, Mikhail Pavlov, Alethea Power, Lukasz Kaiser, Mohammad Bavarian, Clemens Winter, Philippe Tillet, Felipe Petroski Such, Dave Cummings, Matthias Plappert, Fotios Chantzis, Elizabeth Barnes, Ariel Herbert-Voss, William Hebgen Guss, Alex Nichol, Alex Paino, Nikolas Tezak, Jie Tang, Igor Babuschkin, Suchir Balaji, Shantanu Jain, William Saunders, Christopher Hesse, Andrew N. Carr, Jan Leike, Josh Achiam, Vedant Misra, Evan Morikawa, Alec Radford, Matthew Knight, Miles Brundage, Mira Murati, Katie Mayer, Peter Welinder, Bob McGrew, Dario Amodei, Sam McCandlish, Ilya Sutskever, and Wojciech Zaremba. Evaluating Large Language Models Trained on Code. arXiv preprint arXiv:2107.03374, 2021. URL https://arxiv.org/abs/2107.03374.

Paul F Christiano, Jan Leike, Tom Brown, Miljan Martic, Shane Legg, and Dario Amodei. Deep reinforcement learning from human preferences. In I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett (eds.), *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017. URL https://proceedings.neurips.cc/paper/2017/hash/d5e2c0adad503c91f91df240d0cd4e49-Abstract.html.

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.

Ganqu Cui, Lifan Yuan, Zefan Wang, Hanbin Wang, Yuchen Zhang, Jiacheng Chen, Wendi Li, Bingxiang He, Yuchen Fan, Tianyu Yu, Qixin Xu, Weize Chen, Jiarui Yuan, Huayu Chen, Kaiyan Zhang, Xingtai Lv, Shuo Wang, Yuan Yao, Xu Han, Hao Peng, Yu Cheng, Zhiyuan Liu, Maosong Sun, Bowen Zhou, and Ning Ding. Process reinforcement through implicit rewards. *Transactions on Machine Learning Research*, 2026. ISSN 2835-8856. URL https://openreview.net/forum?id=9SkkifLopZ.

Shiyuan Feng, Huan-ang Gao, Haohan Chi, Hanlin Wu, Zhilong Zhang, Zheng Jiang, Bingxiang He, Wei-Ying Ma, Ya-Qin Zhang, and Hao Zhou. Weak-to-Strong Generalization via Direct On-Policy Distillation. arXiv preprint arXiv:2607.05394, 2026. URL https://arxiv.org/abs/2607.05394.

Yuxian Gu, Li Dong, Furu Wei, and Minlie Huang. Minillm: Knowledge distillation of large language models. In B. Kim, Y. Yue, S. Chaudhuri, K. Fragkiadaki, M. Khan, and Y. Sun

(eds.), *International Conference on Learning Representations*, volume 2024, pp. 32694–32717, 2024. URL https://proceedings.iclr.cc/paper_files/paper/2024/hash/8ac015d409635f196f9e3e9dcfb9a94e-Abstract-Conference.html.

Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Peiyi Wang, Qihao Zhu, Runxin Xu, Ruoyu Zhang, Shirong Ma, Xiao Bi, Xiaokang Zhang, Xingkai Yu, Yu Wu, Z. F. Wu, Zhibin Gou, Zhihong Shao, Zhuoshu Li, Ziyi Gao, Aixin Liu, Bing Xue, Bingxuan Wang, Bochao Wu, Bei Feng, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chong Ruan, Damai Dai, Deli Chen, Dongjie Ji, Erhang Li, Fangyun Lin, Fucong Dai, Fuli Luo, Guangbo Hao, Guanting Chen, Guowei Li, H. Zhang, Hanwei Xu, Honghui Ding, Huazuo Gao, Hui Qu, Hui Li, Jianzhong Guo, Jiashi Li, Jingchang Chen, Jingyang Yuan, Jinhao Tu, Junjie Qiu, Junlong Li, J. L. Cai, Jiaqi Ni, Jian Liang, Jin Chen, Kai Dong, Kai Hu, Kaichao You, Kaige Gao, Kang Guan, Kexin Huang, Kuai Yu, Lean Wang, Lecong Zhang, Liang Zhao, Litong Wang, Liyue Zhang, Lei Xu, Leyi Xia, Mingchuan Zhang, Minghua Zhang, Minghui Tang, Mingxu Zhou, Meng Li, Miaojun Wang, Mingming Li, Ning Tian, Panpan Huang, Peng Zhang, Qiancheng Wang, Qinyu Chen, Qiushi Du, Ruiqi Ge, Ruisong Zhang, Ruizhe Pan, Runji Wang, R. J. Chen, R. L. Jin, Ruyi Chen, Shanghao Lu, Shangyan Zhou, Shanhuang Chen, Shengfeng Ye, Shiyu Wang, Shuiping Yu, Shunfeng Zhou, Shuting Pan, S. S. Li, Shuang Zhou, Shaoqing Wu, Tao Yun, Tian Pei, Tianyu Sun, T. Wang, Wangding Zeng, Wen Liu, Wenfeng Liang, Wenjun Gao, Wenqin Yu, Wentao Zhang, W. L. Xiao, Wei An, Xiaodong Liu, Xiaohan Wang, Xiaokang Chen, Xiaotao Nie, Xin Cheng, Xin Liu, Xin Xie, Xingchao Liu, Xinyu Yang, Xinyuan Li, Xuecheng Su, Xuheng Lin, X. Q. Li, Xiangyue Jin, Xiaojin Shen, Xiaosha Chen, Xiaowen Sun, Xiaoxiang Wang, Xinnan Song, Xinyi Zhou, Xianzu Wang, Xinxia Shan, Y. K. Li, Y. Q. Wang, Y. X. Wei, Yang Zhang, Yanhong Xu, Yao Li, Yao Zhao, Yaofeng Sun, Yaohui Wang, Yi Yu, Yichao Zhang, Yifan Shi, Yiliang Xiong, Ying He, Yishi Piao, Yisong Wang, Yixuan Tan, Yiyang Ma, Yiyuan Liu, Yongqiang Guo, Yuan Ou, Yuduan Wang, Yue Gong, Yuheng Zou, Yujia He, Yunfan Xiong, Yuxiang Luo, Yuxiang You, Yuxuan Liu, Yuyang Zhou, Y. X. Zhu, Yanping Huang, Yaohui Li, Yi Zheng, Yuchen Zhu, Yunxian Ma, Ying Tang, Yukun Zha, Yuting Yan, Z. Z. Ren, Zehui Ren, Zhangli Sha, Zhe Fu, Zhean Xu, Zhenda Xie, Zhengyan Zhang, Zhewen Hao, Zhicheng Ma, Zhigang Yan, Zhiyu Wu, Zihui Gu, Zijia Zhu, Zijun Liu, Zilin Li, Ziwei Xie, Ziyang Song, Zizheng Pan, Zhen Huang, Zhipeng Xu, Zhongyu Zhang, and Zhen Zhang. DeepSeek-R1 incentivizes reasoning in LLMs through reinforcement learning. *Nature*, 645(8081):633–638, 2025. ISSN 1476-4687. doi: 10.1038/s41586-025-09422-z. URL https://doi.org/10.1038/s41586-025-09422-z.

Chaoqun He, Renjie Luo, Yuzhuo Bai, Shengding Hu, Zhen 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 Lun-Wei Ku, Andre Martins, and Vivek Srikumar (eds.), *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 3828–3850, Bangkok, Thailand, August 2024. Association for Computational Linguistics. 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 J. Vanschoren and S. Yeung (eds.), *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.

Byeongho Heo, Jaehui Hwang, Sangdoo Yun, and Dongyoon Han. On-Policy Delta Distillation. arXiv preprint arXiv:2607.15161, 2026. URL https://arxiv.org/abs/2607.15161.

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.

Jian Hu, Jason Klein Liu, Haotian Xu, and Wei Shen. REINFORCE++: Stabilizing Critic-Free Policy Optimization with Global Advantage Normalization. arXiv preprint arXiv:2501.03262, 2025. URL https://arxiv.org/abs/2501.03262.

Amirhossein Kazemnejad, Milad Aghajohari, Eva Portelance, Alessandro Sordoni, Siva Reddy, Aaron Courville, and Nicolas Le Roux. VinePPO: Refining credit assignment in RL training of LLMs. In Aarti Singh, Maryam Fazel, Daniel Hsu, Simon Lacoste-Julien, Felix Berkenkamp, Tegan Maharaj, Kiri Wagstaff, and Jerry Zhu (eds.), *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 29557–29590. PMLR, 13–19 Jul 2025. URL https://proceedings.mlr.press/v267/kazemnejad25a.html.

Woosuk Kwon, Zhuohan Li, Siyuan Zhuang, Ying Sheng, Lianmin Zheng, Cody Hao Yu, Joseph Gonzalez, Hao Zhang, and Ion Stoica. Efficient memory management for large language model serving with pagedattention. In *Proceedings of the 29th Symposium on Operating Systems Principles*, SOSP '23, pp. 611–626. ACM, 2023. doi: 10.1145/3600006.3613165. URL https://doi.org/10.1145/3600006.3613165.

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 S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh (eds.), *Advances in Neural Information Processing Systems*, volume 35, pp. 3843–3857. Curran Associates, Inc., 2022. doi: 10.52202/068431-0278. URL https://papers.nips.cc/paper_files/paper/2022/hash/18abbeef8cfe9203fdf9053c9c4fe191-Abstract-Conference.html.

Hunter Lightman, Vineet Kosaraju, Yuri Burda, Harrison Edwards, Bowen Baker, Teddy Lee, Jan Leike, John Schulman, Ilya Sutskever, and Karl Cobbe. Let's verify step by step. In B. Kim, Y. Yue, S. Chaudhuri, K. Fragkiadaki, M. Khan, and Y. Sun (eds.), *International Conference on Learning Representations*, volume 2024, pp. 39578–39601, 2024. URL https://proceedings.iclr.cc/paper_files/paper/2024/hash/aca97732e30bcf1303bc22ac3924fd16-Abstract-Conference.html.

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

Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D Manning, Stefano Ermon, and Chelsea Finn. Direct preference optimization: Your language model is secretly a reward model. In A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine (eds.), *Advances in Neural Information Processing Systems*, volume 36, pp. 53728–53741. Curran Associates, Inc., 2023. doi: 10.52202/075280-2338. URL https://proceedings.neurips.cc/paper_files/paper/2023/hash/a85b405ed65c6477a4fe8302b5e06ce7-Abstract-Conference.html.

John Schulman, Sergey Levine, Pieter Abbeel, Michael Jordan, and Philipp Moritz. Trust region policy optimization. In Francis Bach and David Blei (eds.), *Proceedings of the 32nd International Conference on Machine Learning*, volume 37 of *Proceedings of Machine Learning Research*, pp. 1889–1897, Lille, France, 07–09 Jul 2015. PMLR. URL https://proceedings.mlr.press/v37/schulman15.html.

John Schulman, Philipp Moritz, Sergey Levine, Michael Jordan, and Pieter Abbeel. High-Dimensional Continuous Control Using Generalized Advantage Estimation. In *International Conference on Learning Representations*, 2016. URL https://iclr.cc/archive/www/doku.php%3Fid%3Diclr2016%3Amain.html.

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.

Yingyu Shan, Yuhang Guo, Zihao Cheng, Zeming Liu, Xiangrong Zhu, Xinyi Wang, Jiashu Yao, Wei Lin, Hongru Wang, and Heyan Huang. Learning from Own Solutions: Self-Conditioned Credit Assignment for Reinforcement Learning with Verifiable Rewards. arXiv preprint arXiv:2606.18810, 2026. URL https://arxiv.org/abs/2606.18810.

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. In *Proceedings of the Twentieth European Conference on Computer Systems*, EuroSys '25, pp. 1279–1297. ACM, 2025. doi: 10.1145/3689031.3696075. URL https://doi.org/10.1145/3689031.3696075.

Nisan Stiennon, Long Ouyang, Jeffrey Wu, Daniel Ziegler, Ryan Lowe, Chelsea Voss, Alec Radford, Dario Amodei, and Paul F Christiano. Learning to summarize with human feedback. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin (eds.), *Advances in Neural Information Processing Systems*, volume 33, pp. 3008–3021. Curran Associates, Inc., 2020. URL https://proceedings.neurips.cc/paper/2020/hash/1f89885d556929e98d3ef9b86448f951-Abstract.html.

Lewis Tunstall, Edward Emanuel Beeching, Nathan Lambert, Nazneen Rajani, Kashif Rasul, Younes Belkada, Shengyi Huang, Leandro von Werra, Clémentine Fourrier, Nathan Habib, Nathan Sarrazin, Omar Sanseviero, Alexander M. Rush, and Thomas Wolf. Zephyr: Direct Distillation of LM Alignment. In *Conference on Language Modeling*, 2024. URL https://openreview.net/forum?id=aKkAwZB6JV.

Jonathan Uesato, Nate Kushman, Ramana Kumar, Francis Song, Noah Siegel, Lisa Wang, Antonia Creswell, Geoffrey Irving, and Irina Higgins. Solving math word problems with process- and outcome-based feedback. arXiv preprint arXiv:2211.14275, 2022. URL https://arxiv.org/abs/2211.14275.

Peiyi Wang, Lei Li, Zhihong Shao, Runxin Xu, Damai Dai, Yifei Li, Deli Chen, Yu Wu, and Zhifang Sui. Math-shepherd: Verify and reinforce LLMs step-by-step without human annotations. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar (eds.), *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 9426–9439, Bangkok, Thailand, August 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.acl-long.510. URL https://aclanthology.org/2024.acl-long.510/.

Shenzhi Wang, Le Yu, Chang Gao, Chujie Zheng, Shixuan Liu, Rui Lu, Kai Dang, Xionghui Chen, Jianxin Yang, Zhenru Zhang, Yuqiong Liu, An Yang, Andrew Zhao, Yang Yue, Shiji Song, Bowen Yu, Gao Huang, and Junyang Lin. Beyond the 80/20 rule: High-entropy minority tokens drive effective reinforcement learning for llm reasoning. In D. Belgrave, C. Zhang, H. Lin, R. Pascanu, P. Koniusz, M. Ghassemi, and N. Chen (eds.), *Advances in Neural Information Processing Systems*, volume 38, Main Conference, pp. 115452–115486. Curran Associates, Inc., 2025. doi: 10.52202/085713-3850. URL https://proceedings.neurips.cc/paper_files/paper/2025/hash/a797c2d2e0c1fdabf4d1ab8cd0b465c6-Abstract-Conference.html.

An Yang, Beichen Zhang, Binyuan Hui, Bofei Gao, Bowen Yu, Chengpeng Li, Dayiheng Liu, Jianhong Tu, Jingren Zhou, Junyang Lin, Keming Lu, Mingfeng Xue, Runji Lin, Tianyu Liu, Xingzhang Ren, and Zhenru Zhang. Qwen2.5-Math Technical Report: Toward Mathematical Expert Model via Self-Improvement. arXiv preprint arXiv:2409.12122, 2024. URL https://arxiv.org/abs/2409.12122.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, Chujie Zheng, Dayiheng Liu, Fan Zhou, Fei Huang, Feng Hu, Hao Ge, Haoran Wei, Huan Lin, Jialong Tang, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jing Zhou, Jingren Zhou, Junyang Lin, Kai Dang, Keqin Bao, Kexin Yang, Le Yu, Lianghao Deng, Mei Li, Mingfeng Xue, Mingze Li, Pei Zhang, Peng Wang, Qin

Zhu, Rui Men, Ruize Gao, Shixuan Liu, Shuang Luo, Tianhao Li, Tianyi Tang, Wenbiao Yin, Xingzhang Ren, Xinyu Wang, Xinyu Zhang, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yinger Zhang, Yu Wan, Yuqiong Liu, Zekun Wang, Zeyu Cui, Zhenru Zhang, Zhipeng Zhou, and Zihan Qiu. Qwen3 Technical Report. arXiv preprint arXiv:2505.09388, 2025. URL https://arxiv.org/abs/2505.09388.

Wenkai Yang, Weijie Liu, Ruobing Xie, Kai Yang, Saiyong Yang, and Yankai Lin. Learning beyond Teacher: Generalized On-Policy Distillation with Reward Extrapolation. arXiv preprint arXiv:2602.12125, 2026. 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 D. Belgrave, C. Zhang, H. Lin, R. Pascanu, P. Koniusz, M. Ghassemi, and N. Chen (eds.), *Advances in Neural Information Processing Systems*, volume 38, Main Conference, pp. 113222–113244. Curran Associates, Inc., 2025. doi: 10.52202/085713-3775. URL https://proceedings.neurips.cc/paper_files/paper/2025/hash/a4277440d50f1f15d2cb4c14f7e0c0d2-Abstract-Conference.html.

Song Yu, Li Li, Wenwen Zhao, and Zhisheng Yang. EP-GRPO: Entropy-Progress Aligned Group Relative Policy Optimization with Implicit Process Guidance. arXiv preprint arXiv:2605.04960, 2026. URL https://arxiv.org/abs/2605.04960.

Lifan Yuan, Wendi Li, Huayu Chen, Ganqu Cui, Ning Ding, Kaiyan Zhang, Bowen Zhou, Zhiyuan Liu, and Hao Peng. Free process rewards without process labels. In Aarti Singh, Maryam Fazel, Daniel Hsu, Simon Lacoste-Julien, Felix Berkenkamp, Tegan Maharaj, Kiri Wagstaff, and Jerry Zhu (eds.), *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 73511–73525. PMLR, 13–19 Jul 2025. URL https://proceedings.mlr.press/v267/yuan25c.html.

Eric Zelikman, Yuhuai Wu, Jesse Mu, and Noah Goodman. Star: Bootstrapping reasoning with reasoning. In S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh (eds.), *Advances in Neural Information Processing Systems*, volume 35, pp. 15476–15488. Curran Associates, Inc., 2022. doi: 10.52202/068431-1126. URL https://proceedings.neurips.cc/paper_files/paper/2022/hash/639a9a172c044fbb64175b5fad42e9a5-Abstract-Conference.html.

Eric Zelikman, Georges Raif Harik, Yijia Shao, Varuna Jayasiri, Nick Haber, and Noah D. Goodman. Quiet-STaR: Language Models Can Teach Themselves to Think Before Speaking. In *Conference on Language Modeling*, 2024. URL https://openreview.net/forum?id=oRXPiSOGH9.

Chujie Zheng, Shixuan Liu, Mingze Li, Xiong-Hui Chen, Bowen Yu, Chang Gao, Kai Dang, Yuqiong Liu, Rui Men, An Yang, Jingren Zhou, and Junyang Lin. Group Sequence Policy Optimization. arXiv preprint arXiv:2507.18071, 2025. URL https://arxiv.org/abs/2507.18071.

A ALGORITHM

Table 5: Delta-GRPO training procedure. All sums and quantiles exclude padding.

Initialize actor $\pi_\theta \leftarrow \pi_0$ and retain frozen initialization weights. Set $K = 140$, $G = 8$, $\eta_{\max} = 0.05$, and clip floor 0.1.
For update $k = 1, \ldots, K$:

1. Synchronize the frozen pre-update snapshot π_{old} with the actor.
2. Sample a prompt batch and G responses per prompt. Record valid-token masks.
3. Compute correctness and soft length rewards; obtain A_i within each prompt group.
4. Evaluate sampled-token log probabilities with the two frozen policies in FP32.
5. Compute $\Delta = \log \pi_{\text{old}} - \log \pi_0$ and the global valid-token P95 scale.
6. Clip and normalize to $z \in [-1, 1]$; set $u_{i,t} = \text{sign}(A_i) z_{i,t}$.
7. Set η_k by Equation 8, then $m_{i,t} = 1 + \eta_k(u_{i,t} - \bar{u}_i)/2$.
8. Detach $A^{\Delta}_{i,t} = A_i m_{i,t}$ and update the actor with Equation 5.

Return the step-K actor. Inference uses this actor.

The history measures cumulative policy change from initialization. The stored old policy has the actor's pre-update parameters. The frozen computations provide the dense training signal; the answer verifier supplies the current outcome reward. Response lengths, group advantages, and normalization statistics are recomputed from each collected rollout batch.

B PROPERTIES AND GRADIENT INTERPRETATION

B.1 PROOF OF PROPOSITION 1

For valid response tokens, $|z_{i,t}| \le 1$ and hence $|u_{i,t}| \le 1$. The response mean also lies in $[-1, 1]$, so $u_{i,t} - \bar{u}_i \in [-2, 2]$. Therefore $1 - \eta_k \le m_{i,t} \le 1 + \eta_k$. Centering gives $\sum_t (u_{i,t} - \bar{u}_i) = 0$, yielding $\sum_t m_{i,t} = T_i$ and $\sum_t A_i m_{i,t} = T_i A_i$. For $\eta_k < 1$, all multipliers are strictly positive and thus preserve nonzero advantage signs. For $A_i = 0$, $u_{i,t} = 0$, $m_{i,t} = 1$, and $A^{\Delta}_{i,t} = 0$. These identities are exact in real arithmetic and hold to floating-point precision in the implementation. □

At $\eta_k = 0$, the multiplier is one. If the normalized history is constant across a response, centering likewise removes it completely. These two cases reduce the response's update to the original GRPO advantage. For example, with $z = (-1, -1/3, 1/3, 1)$ and $\eta = 0.05$, a positive unit advantage produces advantages $(0.975, 0.991\overline{6}, 1.008\overline{3}, 1.025)$. A negative unit advantage produces $(-1.025, -1.008\overline{3}, -0.991\overline{6}, -0.975)$. Both average to the original advantage.

B.2 GRADIENT EFFECT OF RESPONSE CENTERING

For fixed rollouts, define the linear token-mean surrogate $J^{\text{lin}}_{\Delta} = N^{-1} \sum_{i,t} \rho_{i,t} A^{\Delta}_{i,t}$. At $\theta = \theta_{\text{old}}$, all ratios equal one. Mean preservation makes $J^{\text{lin}}_{\Delta} = J^{\text{lin}}_{\text{GRPO}}$ at this point. The derivatives of the ratios satisfy $\nabla \rho_{i,t} = \nabla \log \pi_\theta(y_{i,t} \mid s_{i,t})$. Substituting Equation 4 gives Equation 7. Since unit ratios lie in the interior of the clipping interval, the same gradient identity holds for the clipped surrogate at that point.

For a single response normalized by T_i, let $g_{i,t} = \nabla \log \pi_\theta(y_{i,t} \mid s_{i,t})$ and $\bar{g}_i = T_i^{-1} \sum_t g_{i,t}$. Its perturbation is

$$\delta g_i = \frac{\eta_k |A_i|}{2T_i} \sum_t (z_{i,t} - \bar{z}_i)(g_{i,t} - \bar{g}_i). \tag{9}$$

Subtracting $\bar{g}_i$ leaves the sum unchanged. The perturbation is therefore proportional to an empirical, within-response scalar–vector covariance. It can vanish for particular score-gradient patterns even when individual weights change. The covariance captures the gradient change under mean-preserving redistribution.

The triangle inequality gives the local bound

$$\|\delta g_i\| \le \frac{\eta_k |A_i|}{2T_i} \sum_t |z_{i,t} - \bar{z}_i| \, \|g_{i,t}\|. \tag{10}$$

This bounds the absolute local gradient perturbation. The objective weights responses by valid token count through the outer normalization $1/N$.

For arbitrary positive m, the clipped token term obeys $\min\{\rho A m, \mathrm{clip}(\rho) A m\} = m \min\{\rho A, \mathrm{clip}(\rho) A\}$. Thus positivity preserves the per-token clipping branch for fixed A and ρ. Across tokens, different ratios and score gradients still alter the aggregate objective and gradient. The multiplier m is detached during optimization.

B.3 Connection to OPD

For a fixed prefix s, let π_θ be the student and π_T a fixed teacher. Reverse-KL distillation maximizes

$$-D_{\mathrm{KL}}(\pi_\theta(\cdot \mid s) \| \pi_T(\cdot \mid s)) = \mathbb{E}_{a \sim \pi_\theta}[\log \pi_T(a \mid s) - \log \pi_\theta(a \mid s)] \, . \tag{11}$$

Direct-OPD constructs supervision from the teacher's post-RL and pre-RL checkpoints:

$$d_T(s, a) = \log \pi_T^{\mathrm{post}}(a \mid s) - \log \pi_T^{\mathrm{pre}}(a \mid s). \tag{12}$$

This checkpoint difference motivates the accumulated history signal in Equation 2, where the learner's own pre-update policy and initialization supply the pair. At a fixed state, differentiating the reverse-KL objective in Equation 11 yields

$$\nabla_\theta[-D_{\mathrm{KL}}(\pi_\theta \| \pi_T)] = \mathbb{E}_{a \sim \pi_\theta}[(\log \pi_T(a \mid s) - \log \pi_\theta(a \mid s)) \nabla_\theta \log \pi_\theta(a \mid s)] \, , \tag{13}$$

because $\sum_a \nabla_\theta \pi_\theta(a \mid s) = 0$. This shows how a log-policy difference appears as dense feedback in an OPD update. Direct-OPD uses a different log difference, the teacher's pre-/post-RL shift. Delta-GRPO adopts that information source with an internal checkpoint pair, then clips, centers, and gates it by $|A_i|$. Its local gradient is given by Equation 7. The response mean is computed from sampled actions, and the state distribution evolves with the RL policy.

C Training and Evaluation Details

C.1 Matched training configurations

Table 6: Principal GRPO and Delta-GRPO training settings. The archived 4B pair additionally verifies identical source, initialization, and input order.

Setting	GRPO	Delta-GRPO
Model family	Qwen3-4B / Qwen3-8B	Same
Processed training prompts	17,920	Same
Prompt batch / rollout count	128 / 8	Same
Updates / inner epochs	140 / 1	Same
PPO mini-batch	128 prompts	Same
Training seeds	Three independent seeds	Same
Learning rate / warmup	10^{-6} / 20 steps	Same
Ratio clip (low / high)	0.2 / 0.28	Same
Response cap	8,192	Same
Sampling temperature / top-p	1.0 / 1.0	Same
Correctness reward	$+1/-1$	Same
Soft length penalty	6,144–8,192 tokens	Same
KL reward / KL loss / entropy coefficient	0 / 0 / 0	Same
Loss aggregation	Valid-token mean	Same
Actor micro-batch / sequence parallel	1 / 2	Same
Distributed strategy	FSDP2	Same
History modulation	Disabled	Centered mean-one
$\eta_{\max}$ / clip floor	0 / 0.1	0.05 / 0.1

Prompt mini-batch size is expanded by the rollout count before data-parallel partitioning. Equal prompt-batch and mini-batch settings yield one optimizer update per collected rollout batch. The actor's single-update branch forms its ratio denominator from detached current log probabilities, and the recorded clipping fractions are zero. The history computation is a separate frozen pair-forward, with reference-parameter offloading disabled. The infrastructure uses an NPU-adapted verl implementation. The reported step times are measured on this NPU setup.

The soft length reward is zero up to 6,144 tokens and decreases linearly to -1 at 8,192, where generation is capped. Shaped reward is correctness plus this penalty. Advantages reflect shaped rewards relative to the group: a correct response can have a negative relative advantage, and an incorrect response can have a positive one. Equation 1 uses the sample standard deviation (denominator $G-1$) and numerical stabilizer $\epsilon = 10^{-6}$. Reported endpoint scores average the three training seeds per method; the seeds provide independent trajectories under the shared recipe.

C.2 Evaluation identity, scoring, and metrics

Evaluation uses the same endpoint decoding protocol across methods with a 16,384-token response cap, temperature 0.6, top-p 0.95, and disabled top-k. All methods use the same endpoint sampling procedure. MATH500, MinervaMath, and OlympiadBench use eight responses per question; the remaining benchmarks use 32. The response-bank composition is listed below.

Table 7: Endpoint evaluation composition. All methods use the same questions and response counts.

Benchmark	Questions	Samples per question	Responses
MATH500	500	8	4,000
AIME24	30	32	960
AIME25	30	32	960
AMC23	40	32	1,280
MinervaMath	272	8	2,176
OlympiadBench	674	8	5,392
HMMT Feb 2025	30	32	960
AIME26	30	32	960
Total	1,606	Mixed	16,688

The evaluated competition sets use the supplied processed mathematics questions. Predictions are mapped to exact processed prompts before computing benchmark metrics.

The shared scorer uses `math_verify` to parse mathematical expressions and LaTeX according to the boxed-answer convention and checks equivalence against the target answer. Extraction errors, exceptions, and timeouts receive zero. A response reaching the cap can still be scored correct if an extractable correct answer is present.

For benchmark b with Q_b questions, n_b responses per question, and binary scores $r_{q,j}^{\mathrm{eval}}$, reported average accuracy is

$$\widehat{a}_b = \frac{100}{Q_b} \sum_{q=1}^{Q_b} \frac{1}{n_b} \sum_{j=1}^{n_b} r_{q,j}^{\mathrm{eval}}, \qquad \widehat{a}_{\mathrm{macro}} = \frac{1}{8} \sum_{b=1}^{8} \widehat{a}_b. \tag{14}$$

Average accuracy estimates the success rate of an individual sampled answer. Pass@k estimates the probability of at least one correct answer among k samples.

C.3 Baseline adaptations, budgets, and compute

SC-GRPO and EP-GRPO are implementation adaptations of their published mechanisms. ExOPD and Direct-OPD use frozen same-size step-140 GRPO teachers, one response per prompt, and an additional teacher-training stage. The principal pair matches prompts, updates, initialization, and the three training seeds. Delta-GRPO additionally performs two frozen FP32 forwards; in the 4B experiment, mean step time rises from 531.5 seconds for GRPO to 827.2 seconds, a 55.6% increase. SC-GRPO conditions the same actor on verified solutions and derives token credit from

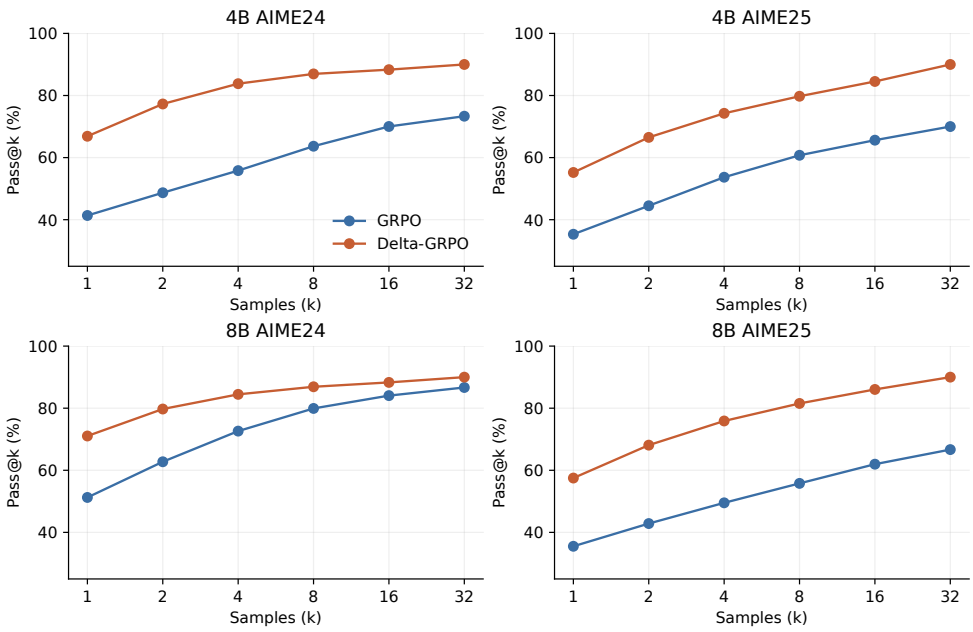


Figure 3: AIME sampled-solution coverage from the same 32 responses per question. Each panel contains 30 questions. Curves use the all-subset estimator.

$D_{\mathrm{KL}}(\pi_{\mathrm{conditioned}} \| \pi_{\mathrm{unconditioned}})$ (Shan et al., 2026). The implementation uses $\alpha = 0.1$, verified-solution weighting for groups with $2 \leq n_{\mathrm{correct}} < 8$, and an active-batch P75 normalization scale with floor 10^{-4}. For groups with no correct solution, a random reference response supplies a diversity signal; the reference itself is excluded from the diversity calculation. We use a mathematics-specific conditioning prompt.

EP-GRPO uses entropy-progress weighting with $\lambda = 0.1$, $\eta = 0.2$, ten progress buckets, and $\gamma = 5$ (Yu et al., 2026). Its alignment anchor uses unshaped correctness. The 8B run has a high cap-hit rate under this configuration.

ExOPD uses the same-size GRPO step-140 teacher and the sampled-token reward $1.25 \log \pi_T - 0.25 \log \pi_0 - \log \pi_\theta$, clipped to $[-20, 20]$ (Yang et al., 2026). Its rollout count is one response per prompt and its learning rate is 10^{-5}. Direct-OPD uses the teacher's post-/pre-RL shift with a top-16 approximation and one response per prompt (Feng et al., 2026). Both teacher-based methods use same-size models and the common training dataset. Teacher training is additional to the student budget. The comparisons use 140 student updates and the common endpoint decoding protocol.

The no-alignment ablation uses $u = z$ before centering. Its additive term is $(\eta_k/2)A_i(z_{i,t} - \bar{z}_i)$, which reverses the full method's term on negative advantages. The no-centering ablation uses $m = 1 + \eta_k u/2$, retaining signed alignment but omitting response-mean subtraction. These controls and the full method use 4B models and the 140-update training protocol.

D Additional Experimental Results

D.1 Sampled Coverage

The endpoint sample banks yield higher Delta-GRPO point estimates at all supported k for all eight benchmarks in both sizes. For 8B AIME24, the paired interval for the gain is [11.98, 27.81] points at $k = 1$ but $[-6.67, 13.33]$ at $k = 32$. Tables 8 and 9 report endpoint coverage on all eight benchmarks using exact question identities.

Observed solved sets also describe how coverage changes. For 4B AIME24, 22 questions are solved by both methods, five only by Delta-GRPO, none only by GRPO, and three by neither. For AIME25 the counts are 20, seven, one, and two. At 8B they are 25, two, one, and two on AIME24, and 19, eight, one, and two on AIME25. These refer to at least one success in the fixed 32-response banks.

Table 8: Qwen3-4B empirical pass@k (%) on all eight benchmarks. The header specifies k for each benchmark. Bold marks the best score, including ties.

Method	MATH500 pass@8	AIME24 pass@32	AIME25 pass@32	AMC23 pass@32	Minerva pass@8	Olympiad pass@8	HMMT25 pass@32	AIME26 pass@32
Base	95.60	53.33	46.67	92.50	50.74	72.85	26.67	43.33
GRPO	97.60	73.33	70.00	97.50	55.15	78.93	43.33	63.33
SC-GRPO	98.60	83.33	80.00	97.50	54.41	81.60	53.33	76.67
EP-GRPO	97.00	70.00	73.33	97.50	52.21	78.64	43.33	60.00
ExOPD	97.40	76.67	66.67	97.50	52.21	78.19	50.00	60.00
Direct-OPD	96.60	76.67	70.00	97.50	54.41	77.15	53.33	66.67
Delta-GRPO	**99.00**	**90.00**	**90.00**	**100.00**	**55.88**	**85.46**	**70.00**	**86.67**

Table 9: Qwen3-8B empirical pass@k (%) on all eight benchmarks. The header specifies k for each benchmark. Bold marks the best score, including ties.

Method	MATH500 pass@8	AIME24 pass@32	AIME25 pass@32	AMC23 pass@32	Minerva pass@8	Olympiad pass@8	HMMT25 pass@32	AIME26 pass@32
Base	94.80	56.67	53.33	95.00	49.26	73.74	26.67	53.33
GRPO	97.80	86.67	66.67	97.50	52.94	81.31	50.00	73.33
SC-GRPO	**98.60**	80.00	76.67	97.50	52.94	83.98	66.67	83.33
EP-GRPO	93.00	63.33	60.00	95.00	51.47	74.93	33.33	50.00
ExOPD	98.20	80.00	73.33	95.00	52.94	80.86	60.00	73.33
Direct-OPD	97.00	76.67	66.67	**100.00**	53.31	79.08	50.00	66.67
Delta-GRPO	**98.60**	**90.00**	**90.00**	**100.00**	**55.51**	**86.35**	**73.33**	**86.67**

D.2 Equal Expected Output-Token Budgets

Set $B = k_g L_g$ for GRPO mean output length L_g and integer budget k_g. For Delta mean length L_d, define $k_d = B/L_d$, $\ell = \lfloor k_d \rfloor$, and $\alpha = k_d - \ell$. Draw $\ell + 1$ responses with probability α and ℓ otherwise. Its expected token cost is B, using the benchmark-average length estimate. Estimated coverage is $(1-\alpha)\widehat{\text{pass}}@\ell + \alpha\widehat{\text{pass}}@(\ell+1)$. The analysis covers sample counts supported by the retained response banks. At budgets where $k_d < 1$, this rule includes skipping some questions, with pass@0 defined as zero.

Table 10: Coverage at the expected token cost of 32 GRPO responses. Paired 95% intervals re-estimate lengths and the sample allocation in each question-bootstrap draw. Values are percentages or percentage-point differences.

Size	Set	Delta $\mathbb{E}[k]$	GRPO	Delta	Gain [95% CI]
4B	AIME24	19.39	73.33	88.69	15.35 [3.33, 28.83]
4B	AIME25	15.54	70.00	84.32	14.32 [−1.02, 30.31]
8B	AIME24	22.09	86.67	88.97	2.30 [−7.57, 12.36]
8B	AIME25	19.25	66.67	87.08	20.42 [3.17, 38.10]

Sample allocation uses test-set mean response lengths, and coverage is evaluated from answer correctness. The matched resource is expected generated-token count. At the lowest budgets, random skipping can reduce Delta-GRPO coverage.

E Mechanism Diagnostics

E.1 Training dynamics and normalization

Figure 5 shows the retained matched 4B training trajectories. Both runs start with correctness 44.14%, mean response length 1,971.4, and token entropy 0.3532. At step 20, online MATH500 accuracy is 83.90% for GRPO and 84.33% for Delta-GRPO; at step 70, mean training response

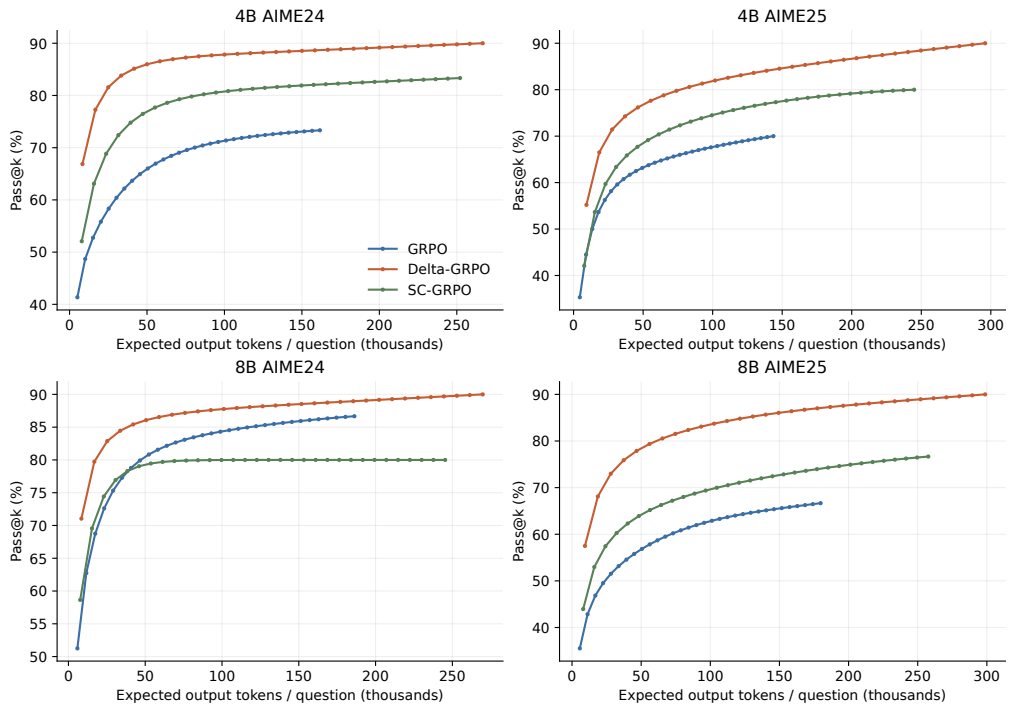


Figure 4: Coverage versus expected generated tokens under the retained response banks. The analysis matches average output-token cost, including failures, and interpolates randomized sample counts.

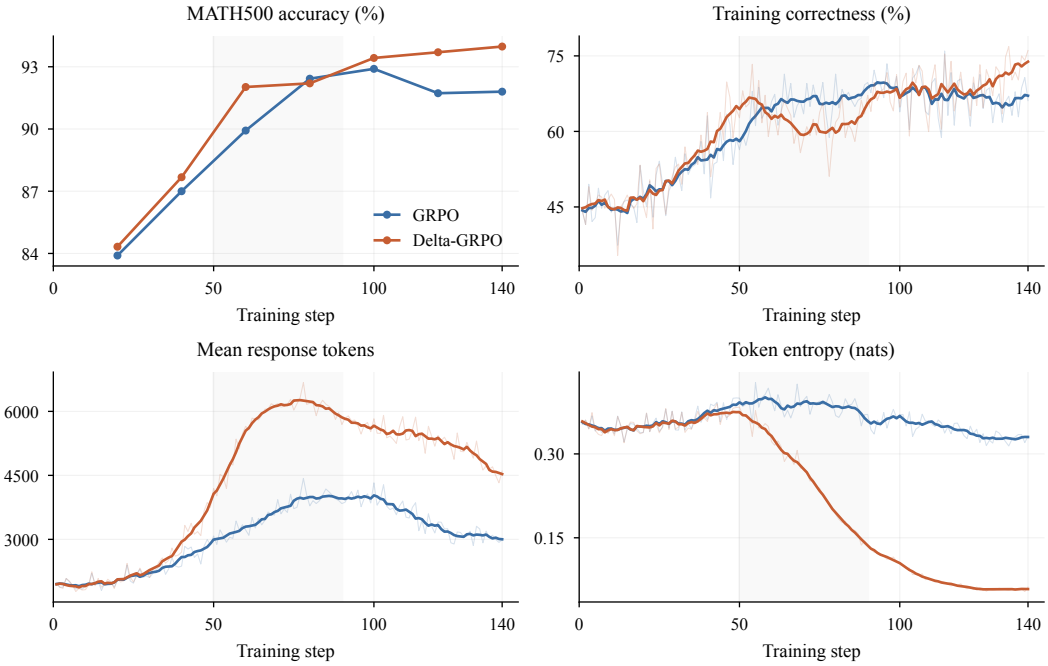


Figure 5: Matched Qwen3-4B training trajectories. Raw measurements are faint and seven-step moving averages are dark. Online validation uses the historical 8,192-token setting. Accuracy improvements accompany longer generations and lower token entropy.

lengths are 3,931.5 and 6,183.4. By step 140, token entropy is 0.3317 for GRPO and 0.0618 for Delta-GRPO. Token entropy measures next-token uncertainty on training states, whereas pass@k measures question-level success under evaluation decoding.

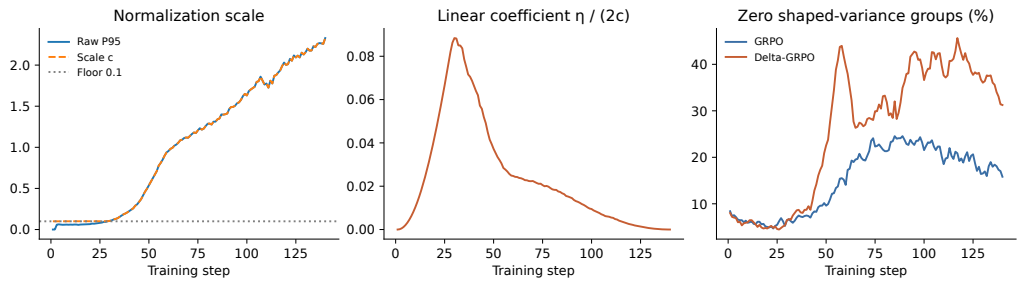


Figure 6: Dynamic normalization and reward-group activity. The right panel smooths the fraction of groups with zero shaped-reward variance over seven steps.

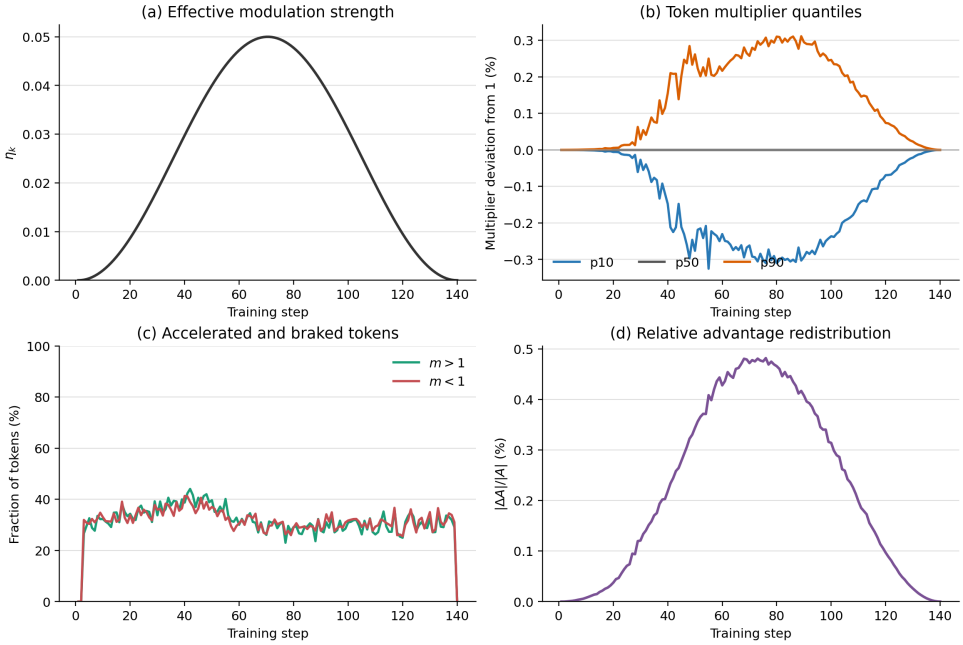


Figure 7: Multiplier quantiles, sign fractions, and relative advantage changes across the 140-update 4B run.

The P95 floor is active through step 28. The scale c_k reaches 1.116 at step 70 and 2.329 at step 140. For unclipped shift, the effective coefficient is $\eta_k/(2c_k)$; it peaks at 0.0884 at step 30 and equals 0.0224 at step 70. At steps 70 and 140, zero shaped-reward variance occurs in 36.72% and 23.44% of Delta groups, versus 21.88% and 9.38% for GRPO. A zero-variance group assigns identical shaped rewards to its responses, yielding $A_i = 0$ throughout the group. Groups with varying rewards supply the relative advantages used by GRPO and Delta-GRPO.

E.2 Multiplier Size and Response Centering

In the frozen audit, the lowest normalized concentration measure $(\sum_t m_t)^2/(T_i \sum_t m_t^2)$ is 0.999973493. This weight-concentration statistic indicates near-uniform multiplier mass. On active responses, the top 10% of tokens carry approximately 36–37% of the absolute perturbation $|m - 1|$. The deviations have local structure within a nearly uniform distribution of total multiplier mass.

E.3 Illustrative Token Windows

Four self-check windows average multiplier changes of +0.345%, +0.407%, +0.375%, and +0.012%, while their final-answer windows average -0.320%, -0.251%, -0.163%, and -0.233%. Incorrect

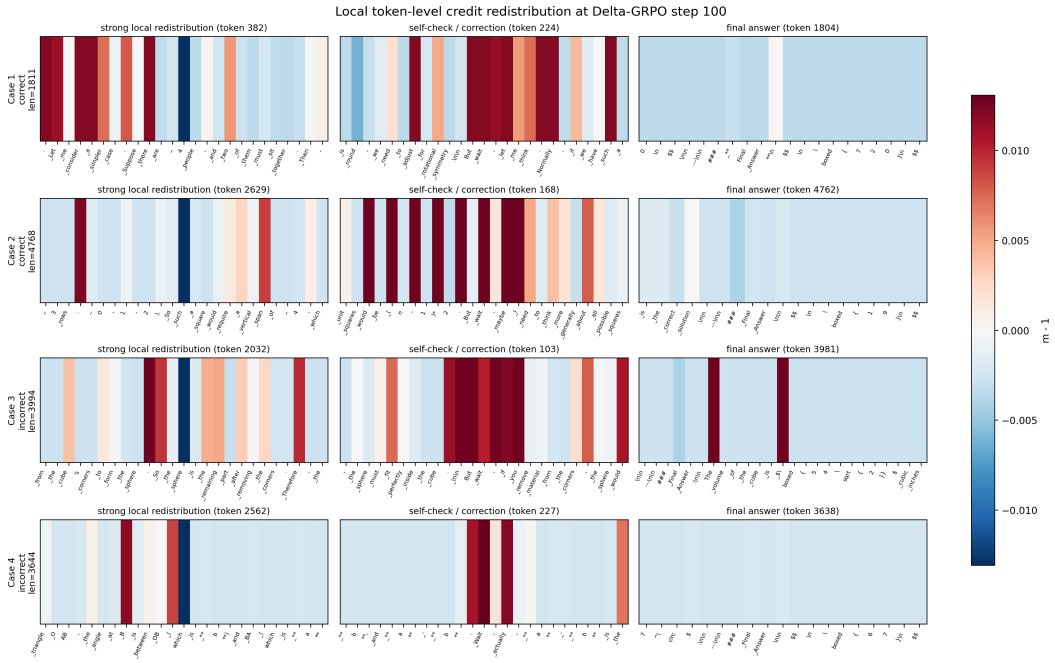


Figure 8: Illustrative local windows from four untruncated responses on distinct questions, two correct and two incorrect. Red denotes multipliers above one and blue below one. The windows illustrate local multiplier patterns.

trajectories also contain amplified self-checking windows. These examples illustrate allocation according to historical alignment.