WHEN AND WHERE TO TRUST THE TEACHER: UNIFYING ON-POLICY DISTILLATION AND GRPO THROUGH ENTROPY-CALIBRATED CREDIT ASSIGNMENT

Anonymous authors
Paper under double-blind review

ABSTRACT

Reinforcement learning with verifiable rewards (RLVR) supervises mathematical reasoning through final-answer correctness, but provides little guidance on individual tokens. On-policy distillation (OPD) supplies dense feedback on student-generated responses, yet teacher preference need not reflect correctness. Recent hybrids combine OPD and verifier-derived advantages or reweight task credit using teacher ratios. However, teacher guidance enters after verifier-based group normalization, and token reweighting need not preserve the total task credit assigned to each response. We introduce Unified Entropy-Calibrated Credit Redistribution for GRPO (UECR-GRPO), which integrates verifier and teacher signals within a single GRPO-style update at both the response and token levels. *Path-Utility Unification* (PUU) combines verifier reward and a teacher-to-anchor path log-ratio in a single KL-regularized objective. Its on-policy implementation uses a length-normalized teacher score and combines both rewards before group normalization and PPO clipping, allowing teacher evidence to influence the response ranking. *Entropy-Calibrated Redistribution* (ECR) then uses the signed teacher–old-policy token gap to redistribute the verifier-derived component. Full-vocabulary teacher entropy attenuates uncertain guidance, while a response-wise zero-sum projection preserves the total task credit and its token-wise sign before clipping. Across five mathematical reasoning benchmarks, UECR-GRPO achieves average $\mathrm{Avg@12}$ accuracies of 17.21% and 65.09% with Qwen3-1.7B and Qwen3-4B students, respectively, exceeding the strongest baseline at each scale by 0.89 and 0.56 percentage points.

1 INTRODUCTION

Reinforcement learning with verifiable rewards has become a practical route for eliciting mathematical reasoning from language models (Shao et al., 2024; Guo et al., 2025; Yu et al., 2025). Group Relative Policy Optimization (GRPO) is particularly attractive because it replaces a learned value model with comparisons among multiple responses to the same problem (Shao et al., 2024). The verifier checks the final answer against an external correctness criterion, but provides coarse credit. One response-level advantage is normally broadcast to every generated token, so a decisive algebraic step, a harmless stylistic token, and the first local error all receive the same advantage. Moreover, when responses within a group receive identical binary rewards, the group-relative task signal vanishes.

On-policy distillation (OPD) provides complementary information (Agarwal et al., 2024; Yang et al., 2026c). A stronger teacher scores the student's own prefixes and therefore supplies dense feedback exactly on the states the student visits. Such feedback can distinguish trajectories even when their verifier rewards are tied. Yet the teacher may prefer an ultimately incorrect path, discourage an unfamiliar but valid derivation, or transfer style rather than mathematical substance. The verifier therefore indicates *whether* a response succeeds but not *where*; the teacher provides local preference but does not define terminal correctness.

Figure 1 shows that teacher evidence adds resolution without reliably defining response quality. Across checkpoints, teacher scores distinguish 84.3–98.4% of verifier-degenerate groups under the fixed threshold. However, teacher-induced ordering conflicts with verifier correctness on 39.1–50.8% of correct–incorrect response pairs. Among 347 judgeable verifier-tied pairs, teacher preference agrees with blinded Opus 4.8 process-quality judgments in only 56.8% of cases. Teacher evidence is

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.

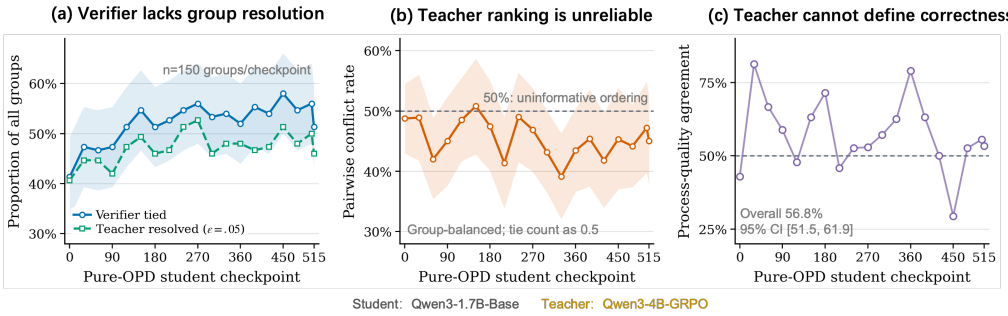


Figure 1: Signal complementarity across 19 Pure-OPD checkpoints. At each checkpoint, the student generates eight responses for each of 150 prompts. (a) reports verifier-degenerate groups and the subset whose teacher-score range exceeds $\epsilon_{\text{gap}} = 0.05$. (b) reports the group-balanced conflict rate over correct–incorrect pairs, with teacher-score ties counted as 0.5. (c) reports agreement between teacher preference and a blinded independent process-quality judge on verifier-tied pairs.

thus informative as a policy-relative signal, but is neither a replacement for terminal verification nor a reliable process-quality label by itself. This motivates retaining the verifier as the task-defining signal while using the teacher only to supply additional resolution.

Recent methods combine GRPO and teacher supervision beyond a naive sum of independent losses. Distilled RL multiplicatively reweights each positive, normalized GRPO advantage using teacher-to-student token ratios (Wang et al., 2026). In particular, when identical verifier rewards produce a zero GRPO advantage, the reweighted update remains zero. ATOD instead adds an annealed, turn-weighted token-level OPD advantage to an already normalized GRPO advantage (Tan et al., 2026), so its teacher signal remains active when verifier rewards are tied. However, teacher evidence enters after group normalization as an additive token-level objective rather than as part of a joint response utility. Moreover, its token weighting is not constrained to preserve the response-average update magnitude. The separately weighted OPD gradient may reinforce or oppose the verifier-derived gradient, and ATOD does not explicitly control this directional interaction. This leaves open how to incorporate teacher evidence into group-relative comparison and localize verifier-derived credit without altering response-level scale, and whether the resulting update better aligns with a held-out verifier gradient.

To address these questions, we propose *Unified Entropy-Calibrated Credit Redistribution for GRPO* (UECR-GRPO). Its trajectory-level component, *Path-Utility Unification* (PUU), incorporates teacher evidence into group-relative comparison. Viewing each response as an autoregressive path, the sum of teacher-to-anchor token log-ratios gives an exact path log-density ratio. We treat this quantity as an implicit teacher reward and combine it with verifier utility in a single KL-regularized objective. The resulting Gibbs optimum defines a target distribution shaped jointly by task reward, teacher preference, and an anchor policy. In the on-policy implementation, verifier reward and a length-normalized teacher score are combined before group normalization, allowing teacher evidence to influence the sampled response ranking before a clipped GRPO-style update. Verifier utility remains explicit, and an offline sensitivity audit shows that our chosen setting lies well within the empirically observed safe region for correct–incorrect ordering reported in Appendix F.4.

However, trajectory unification alone does not resolve task-credit localization without uncontrolled response-level scaling, because PUU still assigns one scalar advantage to each response. We therefore introduce *Entropy-Calibrated Redistribution* (ECR). The signed log-probability gap between the teacher and the old policy provides a direction for adjusting verifier-derived credit, while the teacher's full-vocabulary entropy measures predictive uncertainty and attenuates the guidance under diffuse next-token distributions. A response-wise zero-sum projection redistributes the task component while preserving its total credit and token-wise sign before clipping. This allows teacher-preferred tokens in failed responses to receive less negative task credit without becoming positive imitation targets, separating local credit assignment from response-level scaling.

Our contributions are:

- We formulate PUU as a joint KL-regularized objective combining verifier utility and teacher-to-anchor improvement, and derive its Gibbs-optimal path distribution. On policy, PUU combines both signals before group normalization, allowing teacher evidence to influence response ranking.

- We introduce ECR to redistribute verifier-derived credit using an entropy-attenuated teacher preference signal. A zero-sum projection preserves each response's total task credit and token-wise sign before clipping, separating token credit assignment from response-level scaling.
- Experiments on mathematical reasoning benchmarks with 1.7B and 4B students show average accuracy gains of 0.89 and 0.56 percentage points over the strongest baseline, respectively.

2 Related Work

RLVR and group-relative optimization. PPO uses a clipped policy-gradient surrogate (Schulman et al., 2017), while GRPO derives advantages from within-prompt reward statistics (Shao et al., 2024), whose normalization affects the weighting of successful and failed samples (Mroueh, 2025). DAPO improves clipping, sampling, token-level loss reduction, and length handling (Yu et al., 2025), and GSPO introduces sequence-level importance ratios and clipping (Zheng et al., 2025). These methods refine policy optimization, but token credit remains determined by sequence-level rewards.

On-policy distillation and joint RL–KD. Generalized Knowledge Distillation trains on student-generated sequences to reduce train–inference mismatch (Agarwal et al., 2024). KDRL combines reverse-KL distillation with rule-based rewards (Xu et al., 2025), while G-OPD formulates OPD as dense KL-regularized RL and decouples implicit-reward strength from regularization (Yang et al., 2026c). Reward-gated and sign-gated variants use verifier information to decide when teacher supervision applies (Akhondzadeh et al., 2026; Xu et al., 2026). PUU instead forms a shared response utility before computing the group-relative advantage.

Teacher-guided credit assignment. Distilled RL uses clipped, geometrically normalized teacher-to-old-policy ratios to reweight positive GRPO advantages, reverting to ordinary RL otherwise (Wang et al., 2026). ECR instead handles both advantage signs, calibrates guidance by teacher entropy, and preserves the arithmetic sum of verifier-derived credit rather than the product of token weights.

ATOD combines annealed token-level OPD and GRPO in one clipped update, with T-DUR reweighting turns using teacher–student disagreement and student uncertainty (Tan et al., 2026). Unlike PUU, its teacher evidence enters only after group normalization and therefore does not affect the response rankings or group statistics used to construct the GRPO advantage.

RLSD scales updates by teacher–student differences (Yang et al., 2026a), RLCSD contrasts correct- and wrong-hint teachers (Pan et al., 2026), and StepOPSD redistributes supervision over action segments (Zhang et al., 2026). TASPO uses mean-preserving privileged-information weights at the action level (Yang et al., 2026b), SGCD constructs sibling-based stepwise references (Ding et al., 2026), and SC-GRPO multiplicatively weights GRPO with self-conditioned KL (Shan et al., 2026). ECR instead uses entropy-calibrated teacher gaps for token-level redistribution while preserving each response's verifier-credit total.

Teacher uncertainty. Entropy-Aware OPD addresses unstable reverse KL in high-entropy teacher regions using forward KL (Jin et al., 2026). ATOD's T-DUR uses student sampled-token surprisal to strengthen OPD supervision (Tan et al., 2026). ECR gates local adjustments with full-vocabulary teacher entropy and recovers the broadcast PUU advantage as confidence vanishes.

KL-regularized control and inference. KL control represents a policy as a controlled change of measure from anchor dynamics (Todorov, 2006). Control-as-inference yields reward-tilted trajectory distributions and connects optimal control with variational inference (Levine, 2018; Theodorou et al., 2010). PUU applies this formulation to joint verifier and teacher utility, replacing the exact path log-ratio sum with a token mean before group-relative normalization.

3 Preliminaries and Problem Formulation

Setting and notation. Given a prompt x, a student policy generates a response $y = (y_1, \ldots, y_L)$. The behavior policy π_{old} collects the current batch, π_θ is the student being updated, and π_{T} is a frozen teacher. A separate frozen π_{ref}, when enabled, regularizes the actor update. For response i and token t, $s_{i,t} = (x, y_{i,<t})$, and $m_{i,t}$ selects actor-controlled response tokens while excluding prompts and padding. We use P_0 for the abstract anchor path measure in the theoretical analysis. In the on-policy algorithm we instantiate $P_0 = P_{\text{old}}$.

3.1 GRPO: TASK-VALID BUT SEQUENCE-LEVEL SUPERVISION

For each prompt, π_{old} samples G responses. Given verifier reward R_i^{task}, GRPO forms

$$A_i^{\text{GRPO}} = \frac{R_i^{\text{task}} - \text{Mean}_{j \le G} R_j^{\text{task}}}{\text{Std}_{j \le G}(R_j^{\text{task}}) + \epsilon}, \qquad y_i \sim \pi_{\text{old},T_r}(\cdot \mid x). \tag{1}$$

Here, T_r denotes the rollout temperature. GRPO avoids a value model but broadcasts one advantage across each response, preventing local credit assignment. When group reward variance is zero, its verifier-derived update vanishes entirely.

3.2 OPD: DENSE PREFERENCE ON STUDENT-VISITED STATES

At state $s_{i,t}$, define the raw-temperature teacher-to-old-policy gap

$$\delta_{i,t} = \log \pi_{\text{T},T=1}(y_{i,t} \mid s_{i,t}) - \log \pi_{\text{old},T=1}(y_{i,t} \mid s_{i,t}). \tag{2}$$

The gap measures the teacher's preference for each sampled token relative to the behavior policy. Here, π_{old} denotes the policy that generated the responses and remains fixed during the batch update, while π_θ is optimized. Dense preference is not process correctness. A positive gap only says that the teacher assigns more probability to the realized token than the old student does. It does not establish that the token is mathematically valid, causally important, or part of a successful final response.

3.3 THREE WAYS OF COMBINING REWARD AND TEACHER EVIDENCE

The relevant methods differ mainly in *when* teacher information enters the update.

1. **Independent objectives.** A naive hybrid adds a clipped GRPO surrogate and an independently constructed OPD surrogate. Because clipping is nonlinear, adding the two clipped losses is not equivalent to combining their evidence first. Teacher evidence cannot alter the reward ranking used to form the GRPO advantage.
2. **Post-normalization integration.** ATOD combines an OPD token advantage with an already normalized GRPO advantage before one clipped actor update (Tan et al., 2026), while Distilled RL multiplies the normalized GRPO advantage by a normalized teacher ratio (Wang et al., 2026). These are integrated actor updates rather than independent-loss baselines, but the verifier ranking is fixed before teacher information enters.
3. **Pre-normalization trajectory unification.** We instead combine verifier reward and teacher evidence into a single response utility before computing group-relative advantages. This response-level interpretation motivates the path-space formulation below. Token credit is then handled separately by redistributing the verifier-derived component within each response.

4 UNIFIED TRAJECTORY UTILITY AND CONSTRAINED TOKEN CREDIT

Overview. The method follows an ordered two-layer design. PUU first combines verifier reward and a response-level teacher improvement before group normalization, yielding one trajectory advantage. ECR then modifies only the placement of the verifier-derived component, using teacher uncertainty and a zero-sum constraint. Separating these levels lets the experiments ask two distinct questions: whether pre-normalization trajectory unification helps, and whether constrained localization adds value beyond that unified objective.

4.1 PUU: PATH-UTILITY UNIFICATION

Let P_0 and P_T be the anchor and teacher path distributions for a fixed prompt. Autoregressive factorization gives

$$\sum_{t=1}^{L} \log \frac{\pi_T(y_t \mid x, y_{<t})}{\pi_0(y_t \mid x, y_{<t})} = \log \frac{P_T(y \mid x)}{P_0(y \mid x)}. \tag{3}$$

Thus the summed OPD signal is an exact path log-density ratio rather than an unrelated auxiliary loss.

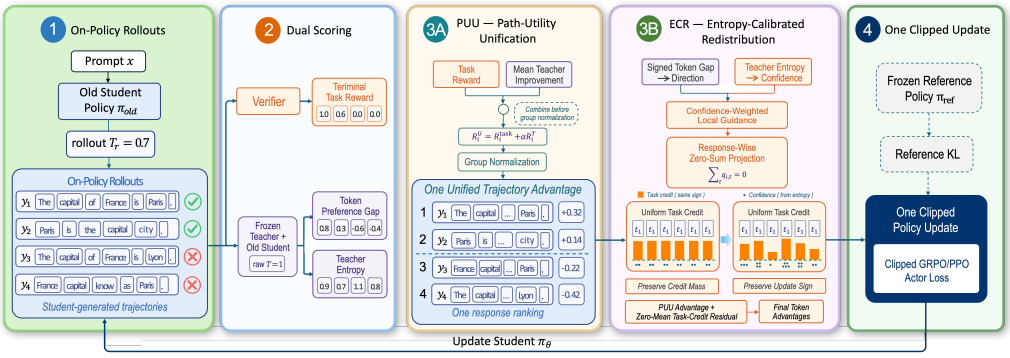


Figure 2: Method overview. PUU combines verifier reward and teacher improvement before group normalization. ECR then redistributes only the verifier-derived component using a signed teacher gap, teacher entropy, and a response-wise zero-sum projection.

For a candidate path distribution Q, define

$$J(Q) = \mathbb{E}_{y\sim Q}\left[R_{\text{task}}(x,y) + \alpha \log \frac{P_T(y \mid x)}{P_0(y \mid x)}\right] - \beta \mathrm{KL}(Q(\cdot \mid x) \,\|\, P_0(\cdot \mid x)). \tag{4}$$

The verifier specifies task utility, the log-ratio measures teacher improvement over the anchor, and the KL term controls departure from that anchor. The unique maximizer is

$$Q^*(y \mid x) \propto \exp\left(\frac{R_{\text{task}}(x,y)}{\beta}\right) P_T(y \mid x)^{\alpha/\beta} P_0(y \mid x)^{1-\alpha/\beta}. \tag{5}$$

Setting $\alpha = 0$ recovers reward-regularized RL; setting $R_{\text{task}} = 0$ and $\alpha = \beta$ recovers P_T; and $R_{\text{task}} = 0, \alpha > \beta$ yields reward extrapolation beyond the teacher in the teacher-over-anchor direction. These are properties of the distributional target. They do not imply that the finite-sample clipped algorithm is identical to each corresponding training procedure.

4.1.1 On-policy group-relative realization

Direct normalization over all language-model paths is intractable. At each iteration, we set $P_0 = P_{\text{old}}$, sample G responses from π_{old}, and score their realized tokens with π_T and π_{old}. The teacher term thus measures improvement over the behavior policy on the sampled responses. The exact path reward in Eq. (3) is a token sum whose magnitude grows mechanically with response length. We therefore use the length-normalized surrogate

$$R_i^T = \frac{\sum_t m_{i,t}\delta_{i,t}}{\sum_t m_{i,t}}, \qquad R_i^U = R_i^{\text{task}} + \alpha R_i^T. \tag{6}$$

For variable-length responses, R_i^T is no longer an exact path log-density ratio. It is a practical surrogate that prevents response length from implicitly changing the scale of the teacher contribution.

PUU forms the joint utility before computing the group-relative advantage

$$\mu_U = \frac{1}{G}\sum_{j=1}^{G} R_j^U, \qquad \sigma_U = \mathrm{Std}_{j=1}^{G}(R_j^U), \qquad A_i^U = \frac{R_i^U - \mu_U}{\sigma_U + \epsilon}. \tag{7}$$

Teacher evidence can therefore change the ordering among sampled responses, while the combined utilities also determine the group mean and scale used by the clipped update. In post-normalization hybrids, the verifier-derived ordering has already been fixed when teacher evidence is introduced.

To prepare for token-level redistribution, we decompose this single unified advantage into task and teacher components that share the same scale

$$A_i^{\text{task}|U} = \frac{R_i^{\text{task}} - \mu_{\text{task}}}{\sigma_U + \epsilon}, \qquad A_i^{T|U} = \frac{R_i^T - \mu_T}{\sigma_U + \epsilon}. \tag{8}$$

Since centering is linear, substituting Eq. (6) into Eq. (7) gives $A_i^U = A_i^{\text{task}|U} + \alpha A_i^{T|U}$ exactly. These are additive components of one advantage rather than two independently normalized advantages. Their

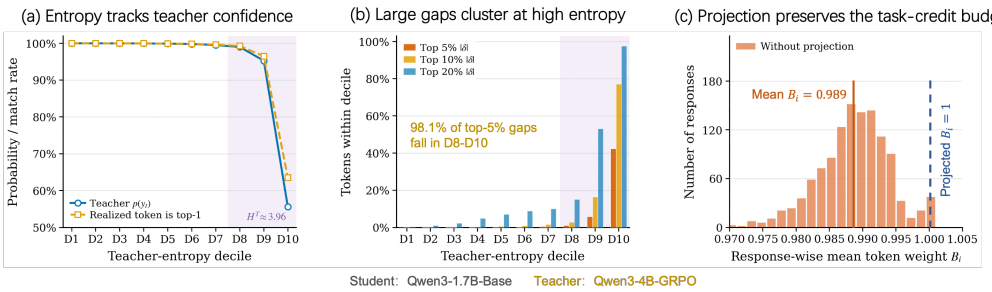


Figure 3: Initialization-time token-credit diagnosis. (a) relates teacher entropy to the probability and top-1 rank of the realized token. (b) shows where tokens with large absolute teacher–old-policy gaps occur across entropy deciles. (c) reports the response-wise mean token weight before projection, while zero-sum projection fixes this mean at one.

shared denominator preserves the intended coefficient α, whereas separate standardization would make its effective value vary across groups. For verifier-degenerate groups, the unified advantage approximately reduces to a group-standardized response-level OPD score. This decomposition allows ECR to redistribute verifier-derived credit while leaving the teacher component unchanged.

4.2 ECR: Entropy-Calibrated Redistribution

Figure 3 examines token-level teacher signals at initialization. We sample eight responses for each of 300 training prompts from the 1.7B student, yielding 2,400 responses, and evaluate every realized token under the frozen 4B teacher and the behavior policy. When tokens are grouped by teacher entropy, high-entropy prefixes assign lower probability to the realized token and contain most large teacher–old-policy log-probability gaps. Without response-wise centering, using these local signals as token weights also moves their arithmetic mean away from one. This changes the total verifier-derived credit instead of only redistributing it across tokens. The results motivate separate mechanisms for calibrating teacher uncertainty and preserving task credit.

4.2.1 Direction and Confidence

PUU ranks responses but still broadcasts one scalar to all their tokens. Let $s_i = \mathrm{sign}(A_i^{\mathrm{task}|U})$. We construct a bounded local direction from the teacher–old-policy gap and calibrate it using the teacher's uncertainty

$$d_{i,t} = \tanh\left(\frac{s_i \delta_{i,t}}{2\tau_\delta}\right), \qquad c_{i,t} = \exp\left(-\frac{H_{i,t}^T}{\tau_H}\right). \tag{9}$$

Here, $H_{i,t}^T$ is the frozen teacher's full-vocabulary entropy at the current prefix. The positive scale τ_δ controls how quickly the direction saturates as the gap grows, while τ_H controls the strength of entropy attenuation. A larger gap increases $|d_{i,t}|$, whereas higher teacher entropy decreases $c_{i,t}$.

For $A_i^{\mathrm{task}|U} > 0$, teacher-preferred tokens have positive local direction. When $A_i^{\mathrm{task}|U} < 0$, s_i reverses this direction, so teacher-preferred tokens can receive less negative task credit without becoming unconditional imitation targets. Entropy affects only the strength of this proposal and does not determine its direction or correctness.

Table 1: ECR credit directions.

	$\delta_{i,t} > 0$	$\delta_{i,t} < 0$
$A_i^{\mathrm{task}\|U} > 0$	↑	↓
$A_i^{\mathrm{task}\|U} < 0$	↑	↓

4.2.2 Zero-sum response-wise projection

Applying the local signal directly as a token weight would change both the allocation and the total amount of task credit as shown in Figure 3. We instead subtract its confidence-weighted response mean

$$\mu_i^c = \frac{\sum_t m_{i,t} c_{i,t} d_{i,t}}{\sum_t m_{i,t} c_{i,t}}, \qquad q_{i,t} = \tfrac{1}{2} m_{i,t} c_{i,t} (d_{i,t} - \mu_i^c), \qquad w_{i,t} = 1 + \rho q_{i,t}, \quad 0 \le \rho < 1. \tag{10}$$

Table 2: Main mathematical reasoning results for Qwen3 students. Block headers specify the student and frozen teacher. All values are Avg@12 accuracy (%) under a fixed 12-sample decoding setup. The average weights all five benchmarks equally. A dash denotes an unevaluated setting.

Method	Avg@12 accuracy (%)					
	AIME24	AIME25	AMC23	HMMT25 Feb	HMMT25 Nov	Avg.
Qwen3-1.7B-Base student / *Qwen3-4B-GRPO teacher*						
Initial student (Yang et al., 2025)	1.53	1.75	12.02	0.00	1.94	3.45
Vanilla-GRPO (Shao et al., 2024; Mroueh, 2025)	7.78	6.39	37.50	0.28	**5.28**	11.45
Vanilla-PG-OPD (Agarwal et al., 2024)	9.17	7.22	39.69	0.28	4.72	12.22
Naive-GRPO+PG-OPD	7.50	6.94	38.54	0.28	3.89	11.43
Distilled RL (Wang et al., 2026)	14.44	9.31	48.23	4.44	4.17	16.12
ATOD-aligned (Tan et al., 2026)	14.72	9.17	50.21	4.17	3.33	16.32
UECR-GRPO	**15.24** +0.52	**9.59** +0.28	**52.04** +1.83	**4.72** +0.28	4.44−0.84	**17.21** +0.89
Qwen3-4B student / *Qwen3-8B-Math-GRPO teacher*						
Initial student (Yang et al., 2025)	20.00	18.33	64.58	–	–	–
Vanilla-GRPO (Shao et al., 2024)	66.67	57.22	94.79	36.67	44.44	59.96
Vanilla-PG-OPD (Agarwal et al., 2024)	65.56	53.89	94.17	31.94	42.50	57.61
Naive-GRPO+PG-OPD	65.56	55.83	93.54	33.61	44.72	58.65
Distilled RL (Wang et al., 2026)	68.33	**67.50**	97.08	40.83	**48.89**	64.53
ATOD-aligned (Tan et al., 2026)	71.11	63.61	**97.92**	40.28	48.06	64.20
UECR-GRPO	**71.39** +0.28	66.94−0.56	97.78−0.14	**41.01** +0.18	48.33−0.56	**65.09** +0.56

Thus, $q_{i,t}$ measures each token's direction relative to the confidence-weighted response baseline. Tokens above this baseline receive larger task weights and those below it receive smaller weights, while ρ controls the overall redistribution strength.

For every non-empty response, $\sum_{t:m_{i,t}=1} q_{i,t} = 0$, so the arithmetic mean of $w_{i,t}$ over valid tokens remains exactly one. Moreover, $|q_{i,t}| \leq 1$ and $\rho < 1$ keep all weights positive. ECR therefore preserves both the response-wise additive mass and token-wise sign of the task component, although the full PUU advantage may still have a different sign.

4.2.3 Final actor advantage

The final advantage adds the zero-mean redistribution residual to the PUU advantage

$$A_{i,t}^{\text{final}} = A_i^U + \rho A_i^{\text{task}|U} q_{i,t} = A_i^{\text{task}|U}(1 + \rho q_{i,t}) + \alpha A_i^{T|U}. \tag{11}$$

This form makes clear that ECR redistributes only the verifier-derived component, while the teacher component introduced by PUU remains unchanged. The actor minimizes

$$\mathcal{L} = \mathcal{L}_{\text{PG}}(A^{\text{final}}) + \beta_{\text{ref}} \mathcal{L}_{\text{low-var-KL}}. \tag{12}$$

The reference KL enters only through the actor objective and is not included in either response reward. When $\rho = 0$, the method reduces exactly to PUU. The same reduction occurs when confidence is negligible or all local directions are identical, since the projected residual then vanishes.

These properties address the two observations in Figure 3. Entropy calibration limits the influence of large teacher gaps at uncertain prefixes, while the zero-sum projection redistributes token credit without changing the response-wise task-credit budget.

5 Experiments

We assess downstream mathematical-reasoning accuracy and investigate whether PUU improves trajectory ranking, ECR improves token-level credit assignment, and their combination yields updates better aligned with held-out verifier gradients. Main comparisons use a Qwen3-1.7B student with a Qwen3-4B-GRPO teacher and a Qwen3-4B student with a Qwen3-8B-Math-GRPO teacher. Component ablations use only the Qwen3-1.7B student and Qwen3-4B-GRPO teacher.

5.1 Experimental Setup

Models and training. The 1.7B setting trains Qwen3-1.7B-Base for 515 steps on difficulty-5–7 DeepMath-103K problems (He et al., 2025), with a frozen Qwen3-4B-GRPO teacher (Yang et al.,

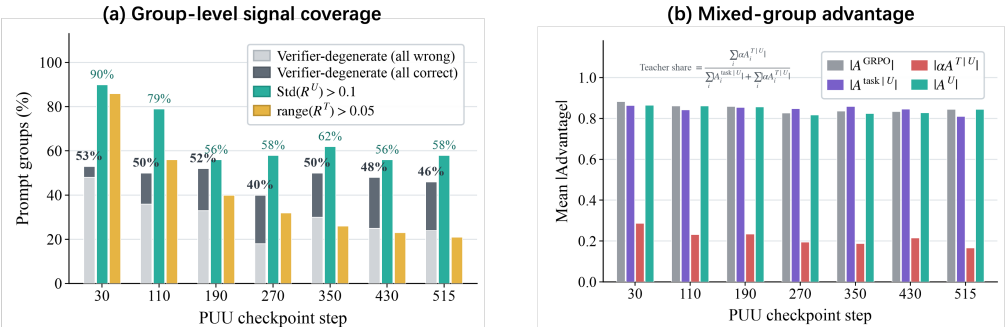


Figure 4: PUU signal analysis on AMC 2023. Twelve responses to each of 40 problems form pseudo-groups of eight and four. (a) Fractions of verifier-degenerate groups and groups with $\mathrm{Std}(R^U) > 0.1$ or $\mathrm{range}(R^T) > 0.05$. (b) Mean absolute PUU advantage components in mixed groups.

2025). The 4B setting pairs a Qwen3-4B student with a frozen Qwen3-8B-Math-GRPO teacher and trains for 160 steps on a sampled difficulty-6–8 subset. Both use non-thinking mode, a global prompt batch size of 126, and $G = 8$. Within each scale, methods share data, decoding, optimizer, clipping, and reference-KL settings. Rollouts and PPO importance ratios use $T_r = 0.7$; teacher and old-policy scores in the teacher gap use raw $T = 1$. Appendices F and A detail the remaining settings and actor-loss conventions.

Baselines. We compare Vanilla GRPO (Shao et al., 2024; Mroueh, 2025), Vanilla PG-OPD (Agarwal et al., 2024), a naive sum of independently clipped GRPO and PG-OPD losses, Distilled RL (Wang et al., 2026), and an aligned single-response adaptation of ATOD (Tan et al., 2026). All teacher-assisted methods use the same frozen scorer within each model scale.

Evaluation. We report $\mathrm{Avg@12}$ accuracy on AIME 2024, AIME 2025, AMC 2023, and the February and November 2025 HMMT contests under a fixed 12-sample decoding setup. The average weights all five benchmarks equally.

5.2 Main Results

Results. Table 2 shows that our method achieves the highest five-benchmark average at both model scales. For the 1.7B student, it leads on four benchmarks, exceeding the strongest baseline by 0.52, 0.28, 1.83, and 0.28 points on AIME24, AIME25, AMC23, and HMMT25-Feb, respectively. Vanilla GRPO leads on HMMT25-Nov by 0.84 points. Our average reaches 17.21%, surpassing ATOD-aligned by 0.89 points. For the 4B student, our method leads on AIME24 and HMMT25-Feb by 0.28 and 0.18 points, while trailing the best baseline on AIME25, AMC23, and HMMT25-Nov by 0.56, 0.14, and 0.56 points. Its average of 65.09% nevertheless exceeds Distilled RL by 0.56 points, showing that the aggregate gain does not imply uniform improvement. Training dynamics for both model scales are provided in Appendix F. In the 1.7B runs (Figure 8), our method achieves higher evaluation accuracy than pure OPD and the naive hybrid without ATOD's large initial gradient norm; the corresponding 4B curves appear in Figure 9.

5.3 Ablation Studies

We isolate trajectory construction from token redistribution with the Qwen3-1.7B student and frozen Qwen3-4B-GRPO teacher (Table 3). All training variants share the data, initialization, rollout, and optimizer settings, with $\alpha = 1.0$.

Trajectory-level normalization. With ECR disabled, Task only uses verifier reward, Separate norm. standardizes task and teacher advantages independently, and PUU normalizes their joint response utility. Figure 4 shows that verifier rewards are tied in 40–53% of pseudo-groups, while the unified utility retains nontrivial variation in 56–90% of all groups. In mixed groups, the teacher component contributes 17–25% of the summed absolute component magnitude, adding resolution without dominating the task signal in this audit. These are signal diagnostics, not gradient or performance contributions; see Appendix F.3. A fixed-rollout α sweep finds no observed correct–incorrect ordering inversions at the chosen setting, without establishing a general guarantee (Appendix F.4).

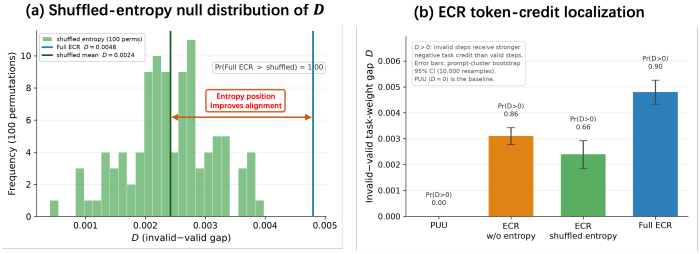


Figure 5: Offline ECR token-credit localization. (a) Invalid-minus-valid weight gaps under 100 within-response entropy shuffles. (b) Gaps for PUU and three ECR variants with 95% prompt-cluster bootstrap intervals. Labels indicate $\Pr(D_i > 0)$.

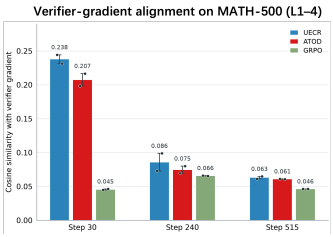


Figure 6: Cosine similarity between each method's actor-loss gradient and a held-out verifier-only gradient on MATH-500.

Table 3: Internal ablations on Qwen3-1.7B with a frozen Qwen3-4B-GRPO teacher. The first row group varies trajectory construction with ECR disabled; the second fixes the PUU trajectory advantage and varies ECR token-credit components. The + *PUU* row (Avg. 20.28) is the no-ECR baseline that the Token-credit rows build upon (denoted *PUU only* +). ✓/× mark which components are active.

Variant	Trajectory comp.			ECR comp.			Avg@12 (%)			
	Task	Teacher	Joint norm.	Gap	Entropy	Proj.	AIME24	AIME25	AMC23	Avg.
Trajectory-level *ECR disabled*										
Task only	✓	×	×	×	×	×	7.78	6.39	37.50	17.22
+ Separate norm.	✓	✓	×	×	×	×	8.50	7.13	38.94	18.19
+ PUU	✓	✓	✓	×	×	×	12.32	8.02	40.50	20.28
Token-credit *PUU baseline + ECR variants*										
PUU only + w/o entropy	✓	✓	✓	✓	×	✓	12.64	8.24	42.46	21.11
PUU only + w/o projection	✓	✓	✓	✓	✓	×	15.04	9.17	51.32	25.17
Full UECR-GRPO	✓	✓	✓	✓	✓	✓	15.24	9.59	52.04	25.62

Token-level redistribution. Holding PUU fixed, the lower block of Table 3 shows that Full ECR outperforms both component removals. Figure 5 provides a separate offline diagnosis of where ECR places task credit. On failed responses from mixed verifier groups, blinded Opus 4.8 labels identify valid and erroneous reasoning steps. Full ECR yields a positive invalid-minus-valid task-weight gap and exceeds all 100 controls that shuffle entropy positions within each response. Thus, ECR relatively protects judge-valid reasoning while concentrating negative task credit on identified errors, and its zero-sum projection preserves the response-wise budget. This establishes the intended localization behavior rather than a causal accuracy gain; the protocol and statistics appear in Appendix F.5.

Verifier-gradient alignment. Signal variation and token localization do not establish whether an update is task-directed. Motivated by work on auxiliary-gradient agreement and gradient interference (Lin et al., 2019; Yu et al., 2020), we compare each candidate actor-loss gradient with a verifier-only gradient computed on disjoint prompts. At fixed parameters and gradient norm, a higher cosine indicates greater first-order agreement with descent on the held-out verifier objective. Figure 6 shows that UECR-GRPO has the highest two-fold mean at all three displayed steps, suggesting better local directional agreement than ATOD and Vanilla GRPO in this audit. Different training trajectories and subsequent optimizer dynamics are not isolated, so this does not establish a causal accuracy gain. Appendix F.6 details the protocol.

6 Conclusion

UECR-GRPO separates trajectory ranking from token credit assignment. PUU combines verifier and teacher utility before group normalization, while ECR uses signed teacher gaps and entropy to redistribute verifier-derived credit without changing each response's task-credit budget. Experiments across five benchmarks and two Qwen3 model scales show higher average accuracy than the compared training objectives. Offline diagnostics support the intended group-level resolution and token localization, and find stronger mean alignment with held-out verifier gradients.

REFERENCES

Rishabh Agarwal, Nino Vieillard, Yongchao Zhou, Piotr Stanczyk, Sabela Ramos Garea, Matthieu Geist, and Olivier Bachem. On-policy distillation of language models: Learning from self-generated mistakes. In *International Conference on Learning Representations*, volume 2024, pp. 21246–21263, 2024.

Mohammad Sadegh Akhondzadeh, Vijay Lingam, Atula Tejaswi, Chanakya Ekbote, Sujay Sanghavi, and Aleksandar Bojchevski. Reward-gated on-policy distillation. *arXiv preprint arXiv:2607.04037*, 2026.

Tianyu Ding, Jianhong Xin, and Juan Pablo De la Cruz Weinstein. Keep policy gradient in charge: Sibling-guided credit distillation for long-horizon tool-use agents. *arXiv preprint arXiv:2606.12634*, 2026.

Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, et al. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*, 2025.

Zhiwei He, Tian Liang, Jiahao Xu, et al. Deepmath-103k: A large-scale, challenging, decontaminated, and verifiable mathematical dataset for advancing reasoning. *arXiv preprint arXiv:2504.11456*, 2025.

Shengyi Huang, Michael Noukhovitch, Arian Hosseini, Kashif Rasul, Weixun Wang, and Lewis Tunstall. The N+ implementation details of RLHF with PPO: A case study on TL;DR summarization, 2024. URL https://arxiv.org/abs/2403.17031.

Woogyeol Jin, Taywon Min, Yongjin Yang, Dennis Wei, Yi Zhou, Swanand Ravindra Kadhe, Nathalie Baracaldo, and Kimin Lee. Entropy-aware on-policy distillation of language models. In *International Conference on Machine Learning*, 2026.

Woosuk Kwon, Zhuohan Li, Siyuan Zhuang, Ying Sheng, Lianmin Zheng, Cody Hao Yu, Joseph E. 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*, 2023. URL https://arxiv.org/abs/2309.06180.

Sergey Levine. Reinforcement learning and control as probabilistic inference: Tutorial and review. *arXiv preprint arXiv:1805.00909*, 2018.

Hunter Lightman, Vineet Kosaraju, Yuri Burda, Harrison Edwards, Bowen Baker, Teddy Lee, Jan Leike, John Schulman, Ilya Sutskever, and Karl Cobbe. Let's verify step by step. In *International Conference on Learning Representations*, 2024. URL https://proceedings.iclr.cc/paper_files/paper/2024/hash/aca97732e30bcf1303bc22ac3924fd16-Abstract-Conference.html.

Xingyu Lin, Harjatin Baweja, George Kantor, and David Held. Adaptive auxiliary task weighting for reinforcement learning. In *Advances in Neural Information Processing Systems*, volume 32, 2019. URL https://proceedings.neurips.cc/paper/2019/hash/0e900ad84f63618452210ab8baae0218-Abstract.html.

Youssef Mroueh. Reinforcement learning with verifiable rewards: Grpo's effective loss, dynamics, and success amplification. *arXiv preprint arXiv:2503.06639*, 2025.

Leyi Pan, Shuchang Tao, Yunpeng Zhai, Lingzhe Zhang, Zhaoyang Liu, Bolin Ding, Aiwei Liu, and Lijie Wen. Rlcsd: Reinforcement learning with contrastive on-policy self-distillation. *arXiv preprint arXiv:2606.11709*, 2026.

Jan Peters, Katharina Mulling, and Yasemin Altun. Relative entropy policy search. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 24, pp. 1607–1612, 2010. doi: 10.1609/aaai.v24i1.7727.

Belinda Phipson and Gordon K. Smyth. Permutation P-values should never be zero: Calculating exact P-values when permutations are randomly drawn. *Statistical Applications in Genetics and Molecular Biology*, 9(1), 2010. doi: 10.2202/1544-6115.1585.

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 *Advances in Neural Information Processing Systems*, volume 36, 2023. URL https://proceedings.neurips.cc/paper_files/paper/2023/hash/a85b405ed65c6477a4fe8302b5e06ce7-Abstract-Conference.html.

John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*, 2017.

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.

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.

Qitai Tan, Mo Li, Zefang Zong, Yipeng Shi, Yang Li, and Peng Chen. Atod: Annealed turn-aware on-policy distillation for multi-turn agentic tasks. *arXiv preprint arXiv:2606.27814*, 2026.

Evangelos Theodorou, Jonas Buchli, and Stefan Schaal. Learning policy improvements with path integrals. In *Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics*, volume 9 of *Proceedings of Machine Learning Research*, pp. 828–835. PMLR, 2010.

Emanuel Todorov. Linearly-solvable markov decision problems. In *Advances in Neural Information Processing Systems*, volume 19, pp. 1369–1376. MIT Press, 2006.

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, 2022. URL https://arxiv.org/abs/2211.14275.

Chen Wang, Zhaochun Li, Jionghao Bai, Yining Zhang, Hexuan Deng, Ge Lan, and Yue Wang. Distilled reinforcement learning for llm post-training. *arXiv preprint arXiv:2607.17247*, 2026.

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

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

Hongling Xu, Qi Zhu, Heyuan Deng, Jinpeng Li, Lu Hou, Yasheng Wang, Lifeng Shang, Ruifeng Xu, and Fei Mi. Kdrl: Post-training reasoning llms via unified knowledge distillation and reinforcement learning. *arXiv preprint arXiv:2506.02208*, 2025.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, et al. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*, 2025.

Chenxu Yang, Chuanyu Qin, Qingyi Si, Minghui Chen, Naibin Gu, Dingyu Yao, Zheng Lin, Weiping Wang, Jiaqi Wang, and Nan Duan. Self-distilled rlvr. *arXiv preprint arXiv:2604.03128*, 2026a.

Jingxiao Yang, Wangjie Gan, Yingxuan Zhuang, Wenqi Zhang, Jintao Chen, and Xuhong Zhang. Reconciling process supervision with outcome-based credit in agentic policy optimization, 2026b. URL https://arxiv.org/abs/2608.31077.

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*, 2026c.

Qiying Yu, Zheng Zhang, Ruofei Zhu, Yufeng Yuan, Xiaochen Zuo, Yu Yue, Weinan Dai, Tiantian Fan, Gaohong Liu, Lingjun Liu, et al. Dapo: An open-source llm reinforcement learning system at scale. *Advances in Neural Information Processing Systems*, 38:113222–113244, 2025.

Tianhe Yu, Saurabh Kumar, Abhishek Gupta, Sergey Levine, Karol Hausman, and Chelsea Finn. Gradient surgery for multi-task learning. In *Advances in Neural Information Processing Systems*, volume 33, 2020. URL https://proceedings.neurips.cc/paper/2020/hash/3fe78a8acf5fda99de95303940a2420c-Abstract.html.

Yanfei Zhang, Xu Lin, and Chenglin Wu. Stepopsd: Step-aware online preference distillation for agent reinforcement learning. *arXiv preprint arXiv:2605.27140*, 2026.

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.

AI USE STATEMENT

Generative AI tools were used to improve the clarity and readability of the manuscript and to assist with writing and debugging research code. All AI-assisted text was reviewed and revised by the authors, and all AI-assisted code was inspected and tested for correctness. The authors take full responsibility for the final content of this work, including any text, claims, and artifacts produced with the assistance of generative AI.

A CLIPPED ACTOR UPDATE AND TEMPERATURE SEMANTICS

Let $m_{i,t}$ select valid response tokens and let $A_{i,t}$ be any detached token advantage. The current-to-old ratio and clipped policy surrogate are

$$\begin{aligned} r_{i,t}(\theta) &= \exp(\log \pi_{\theta,T_r}(y_{i,t} \mid s_{i,t}) - \log \pi_{\text{old},T_r}(y_{i,t} \mid s_{i,t}))\,, \\ \mathcal{L}_{\text{PG}}(A) &= -\operatorname{Reduce}_{m_{i,t}=1} \min[r_{i,t}(\theta)A_{i,t}, \operatorname{clip}(r_{i,t}(\theta), 1-\epsilon_\ell, 1+\epsilon_h)A_{i,t}]\,. \end{aligned} \tag{13}$$

Both actor log probabilities use the rollout convention $T_r = 0.7$. The teacher gap instead compares π_T and π_{old} at raw $T = 1$. Prompt tokens, response padding, and synthetic data-parallel padding rows are excluded. The reference-policy KL, when enabled, is added exactly once to the actor loss and is not included in task reward, teacher reward, or the PUU group statistics.

B PATH-SPACE TARGET

B.1 SETUP AND SUPPORT ASSUMPTIONS

Fix a prompt x and suppress it where unambiguous. Let $\mathcal{Y}$ be the set of finite response paths. The anchor P_0, teacher P_T, and candidate distribution Q are causal autoregressive path measures. We assume $P_T(y) > 0$ and $Q(y) > 0$ only where $P_0(y) > 0$, $\beta > 0$, and a finite partition function. These absolute-continuity conditions ensure that the density ratios and KL divergences are defined.

Lemma 1 (Token log-ratios telescope to a path log-ratio)**.** *For any $y = (y_1, \ldots, y_L)$,*

$$\sum_{t=1}^{L} \log \frac{\pi_T(y_t \mid x, y_{<t})}{\pi_0(y_t \mid x, y_{<t})} = \log \frac{P_T(y \mid x)}{P_0(y \mid x)}.$$

Proof. Autoregressive factorization gives $P_T(y \mid x) = \prod_t \pi_T(y_t \mid x, y_{<t})$ and the analogous identity for P_0. Dividing the products and taking a logarithm gives the result. □

Define

$$U(y) = R_{\text{task}}(y) + \alpha \log \frac{P_T(y)}{P_0(y)}$$

and consider

$$\max_Q \{\mathbb{E}_{y\sim Q}[U(y)] - \beta \mathrm{KL}(Q\|P_0)\}\,. \tag{14}$$

KL-constrained policy improvement is related to relative-entropy policy search (Peters et al., 2010); a policy log-ratio also admits an implicit-reward interpretation in preference optimization (Rafailov et al., 2023).

Proposition 1 (Gibbs-optimal path distribution)**.** *Let*

$$Z = \mathbb{E}_{y\sim P_0}\left[\exp\left(\frac{U(y)}{\beta}\right)\right].$$

The unique maximizer of Eq. (14) *is*

$$Q^*(y) = \frac{1}{Z} P_0(y) \exp\left(\frac{U(y)}{\beta}\right) = \frac{1}{Z} \exp\left(\frac{R_{\text{task}}(y)}{\beta}\right) P_T(y)^{\alpha/\beta} P_0(y)^{1-\alpha/\beta}.$$

Proof. The first expression is normalized by Z, and

$$\log \frac{Q^*(y)}{P_0(y)} = \frac{U(y)}{\beta} - \log Z.$$

For any feasible Q,

$$\mathrm{KL}(Q\|Q^*) = \mathbb{E}_Q\left[\log \frac{Q(y)}{P_0(y)} - \frac{U(y)}{\beta} + \log Z\right],$$
$$\mathbb{E}_Q[U] - \beta \mathrm{KL}(Q\|P_0) = \beta \log Z - \beta \mathrm{KL}(Q\|Q^*).$$

Non-negativity of KL gives a unique optimum at $Q = Q^*$ almost everywhere. The second form follows by substituting the verifier and teacher terms of U. □

Corollary 1 (Target-level reductions)**.** *The optimum has the following reductions:*

1. *If $\alpha = 0$, then $Q^* \propto P_0 \exp(R_{\mathrm{task}}/\beta)$, the standard KL-regularized reward tilt.*
2. *If $R_{\mathrm{task}} = 0$ and $\alpha = \beta$, then $Q^* = P_T$.*
3. *If $R_{\mathrm{task}} = 0$ and $\alpha > \beta$, then $Q^* \propto P_T^{\alpha/\beta} P_0^{1-\alpha/\beta}$, an extrapolation in the teacher-over-anchor reward direction.*
4. *With both terms active, task reward and teacher preference tilt the same target.*

Relationship to the implementation. The theorem is an organizing target, not a claim that every implementation detail exactly optimizes Eq. (14). First, the identity in Lemma 1 uses a token sum, whereas the algorithm uses a masked mean to control reward–length correlation. For variable-length responses this changes the utility. Second, the iteration-wise anchor is P_{old}, held fixed while a batch is scored and updated. Third, PPO clipping is the practical proximal mechanism; an optional long-horizon reference KL is an additional actor regularizer and is not part of the teacher reward.

C Exact Decomposition of the PUU Advantage

Within one prompt group, let

$$R_i^U = R_i^{\mathrm{task}} + \alpha R_i^T, \quad \mu_U = G^{-1} \sum_i R_i^U, \quad \sigma_U = \mathrm{Std}_{i=1}^G(R_i^U).$$

Define

$$A_i^{\mathrm{task}|U} = \frac{R_i^{\mathrm{task}} - \mu_{\mathrm{task}}}{\sigma_U + \epsilon}, \qquad A_i^{T|U} = \frac{R_i^T - \mu_T}{\sigma_U + \epsilon}, \qquad A_i^U = \frac{R_i^U - \mu_U}{\sigma_U + \epsilon}.$$

Proposition 2 (Advantage decomposition)**.** *Using the shared PUU denominator,*

$$A_i^U = A_i^{\mathrm{task}|U} + \alpha A_i^{T|U}.$$

Proof. Linearity gives $\mu_U = \mu_{\mathrm{task}} + \alpha \mu_T$, hence

$$R_i^U - \mu_U = (R_i^{\mathrm{task}} - \mu_{\mathrm{task}}) + \alpha(R_i^T - \mu_T).$$

Division by the common denominator proves the identity. □

If the two rewards were separately standardized by different standard deviations, this identity would fail and the effective teacher coefficient would vary with each group's empirical scales.

D Constrained Token-Credit Projection

For one non-empty response, let the raw local proposal be $\frac{1}{2} c_{i,t} d_{i,t}$. We seek its closest zero-sum adjustment:

$$\min_{\{q_{i,t}\}} \sum_{t:m_{i,t}=1} \frac{\left(q_{i,t} - \frac{1}{2} c_{i,t} d_{i,t}\right)^2}{c_{i,t}} \quad \text{subject to} \quad \sum_{t:m_{i,t}=1} q_{i,t} = 0. \tag{15}$$

Because $c_{i,t} = \exp(-H_{i,t}^T/\tau_H) > 0$ for finite entropy, the objective is strictly convex.

Proposition 3 (Closed-form projection)**.** *The unique solution is*

$$q_{i,t} = \frac{1}{2} c_{i,t}(d_{i,t} - \mu_i^c), \qquad \mu_i^c = \frac{\sum_t m_{i,t} c_{i,t} d_{i,t}}{\sum_t m_{i,t} c_{i,t}}.$$

Proof. Introduce a multiplier λ for the zero-sum constraint. Stationarity at a valid token gives

$$\frac{2}{c_{i,t}} \left(q_{i,t} - \tfrac{1}{2} c_{i,t} d_{i,t}\right) + \lambda = 0,$$

so $q_{i,t} = \frac{1}{2} c_{i,t}(d_{i,t} - \lambda)$. The constraint yields $\lambda = \mu_i^c$. Strict convexity gives uniqueness. □

Proposition 4 (Additive task-budget and sign preservation)**.** *For every non-empty response and* $0 \le \rho < 1$,

$$\frac{1}{L_i} \sum_{t: m_{i,t}=1} A_i^{\text{task}|U}(1 + \rho q_{i,t}) = A_i^{\text{task}|U},$$

and every nonzero token-wise task component retains the sign of $A_i^{\text{task}|U}$.

Proof. The definition of μ_i^c implies $\sum_{t: m_{i,t}=1} q_{i,t} = 0$. Since $d_{i,t} \in [-1, 1]$, its confidence-weighted mean $\mu_i^c \in [-1, 1]$. Together with $c_{i,t} \in (0, 1]$, this gives $|q_{i,t}| \le 1$, hence $1 - \rho \le 1 + \rho q_{i,t} \le 1 + \rho$. The weight is positive and has arithmetic mean one, proving both claims. □

The proposition concerns only the verifier-derived component before clipping. It does not fix the full PUU advantage sign or preserve the gradient norm and parameter update after clipping.

D.1 Entropy computation and limiting cases

For raw teacher logits $z_{i,t}^T$, entropy is evaluated in float32 as

$$H_{i,t}^T = \log \sum_v \exp z_{i,t,v}^T - \sum_v p_{i,t,v}^T z_{i,t,v}^T, \qquad p_{i,t}^T = \text{softmax}(z_{i,t}^T).$$

As $H_{i,t}^T$ grows, $c_{i,t}$ decreases and the local proposal is attenuated. If $\rho = 0$, confidence vanishes at every valid token, or all directions within a response are identical, the redistribution term vanishes and every token receives the original PUU advantage.

E Direct Comparison with Closest Methods

E.1 Distilled RL

For a response-level GRPO advantage A_i, Distilled RL defines

$$\rho_{i,t}^T = \frac{\pi_T(y_{i,t} \mid s_{i,t})}{\pi_{\text{old}}(y_{i,t} \mid s_{i,t})}, \qquad \bar{\rho}_{i,t}^T = \text{clip}(\rho_{i,t}^T, \epsilon_T^{-1}, \epsilon_T),$$

$$\widetilde{\rho}_{i,t}^T = \frac{\bar{\rho}_{i,t}^T}{\exp\left(L_i^{-1} \sum_s \log \bar{\rho}_{i,s}^T\right)}, \qquad w_{i,t}^{\text{DRL}} = \begin{cases} \widetilde{\rho}_{i,t}^T, & A_i > 0, \\ 1, & A_i \le 0. \end{cases}$$

It then inserts $w_{i,t}^{\text{DRL}} A_i$ into the clipped policy surrogate (Wang et al., 2026). The teacher ratio is thus applied after verifier group normalization and cannot alter response ranking. Its geometric normalization guarantees $(\prod_t \widetilde{\rho}_{i,t}^T)^{1/L_i} = 1$, not $L_i^{-1} \sum_t \widetilde{\rho}_{i,t}^T = 1$. Our no-projection audit measures the practical difference between these invariants under token-mean reduction.

E.2 ATOD

ATOD constructs

$$A_{i,t}^{\text{ATOD}} = \kappa(s)\, w_{k(t)} \Delta \log p_{i,t} + \rho(s) A_i^{\text{GRPO}},$$

where $\kappa(s)$ decreases, $\rho(s)$ increases, and $w_{k(t)}$ is the T-DUR weight for the containing interaction turn (Tan et al., 2026). T-DUR combines normalized turn-level teacher–student disagreement with student sampled-token surprisal and weights only the OPD term. The sum is optimized through one clipped actor surrogate. ATOD is therefore not an independent-loss method. It differs from PUU because the GRPO advantage is normalized before combination and the teacher term is a token-level OPD advantage rather than part of the response utility used for group ranking.

Table 4: Conceptual comparison with the closest baselines. "Exact additive budget" refers to the arithmetic token mean of the verifier-derived component before PPO clipping.

Method	Teacher in ranking	One clipped update	Negative task credit	Entropy attenuation	Exact credit budget
Naive GRPO+OPD	No	No	OPD independent	No	No
ATOD	No	Yes	Through additive OPD	No	No
Distilled RL	No	Yes	Reset to GRPO	No	No
PUU	Yes	Yes	Yes, at trajectory level	No	Broadcast
UECR-GRPO	Yes	Yes	Yes, signed protection	Yes	Yes

F Additional Experimental Protocol

F.1 Signal diagnosis across training

For each checkpoint and response, the verifier returns $R_i^{\text{task}} \in \{0, 1\}$, while the teacher score is

$$R_i^T = \frac{\sum_t m_{i,t} \left(\log \pi_{T,T=1}(y_{i,t} \mid s_{i,t}) - \log \pi_{\text{old},T=1}(y_{i,t} \mid s_{i,t})\right)}{\sum_t m_{i,t}}.$$

A verifier-degenerate group contains identical binary rewards and hence zero within-group task advantage. We report (i) the fraction of such groups, (ii) the fraction in which R_i^T has nonzero variation, (iii) the teacher–verifier ranking-conflict rate in non-degenerate groups, and (iv) a blinded model audit on verifier-tied pairs. The initialization tie rate is 41.3%, as shown in Figure 1.

F.2 Token-credit diagnosis at initialization

The mechanism audit uses the shared initial 1.7B student and frozen 4B teacher before training. Eight responses are sampled for each of 300 prompts, giving 2,400 trajectories and 21.55M valid response tokens. No optimizer update or test benchmark is involved. For each valid token,

$$\bar{d}_{i,t} = \tanh\left(\frac{\delta_{i,t}}{2\tau_\delta}\right), \quad d_{i,t} = \text{sign}(A_i^{\text{task}|U})\bar{d}_{i,t}, \quad c_{i,t} = \exp(-H_{i,t}^T/\tau_H).$$

Entropy strata are determined once from global step-0 quantiles. Tied near-deterministic values may collapse repeated boundaries, so these are called strata rather than strict deciles. We report token-weighted and response-balanced summaries with prompt-cluster bootstrap.

Teacher concentration is measured by realized-token probability and realized-token/top-1 match, neither of which is called token correctness. Large-gap concentration is reported at top 5%, 10%, and 20% thresholds. The primary audit finds that 82.4% of top-10% gaps occur in the highest entropy quartile and that calibration reduces their mean magnitude by 37%.

For projected and unprojected weights, define

$$B_i = \frac{\sum_t m_{i,t} w_{i,t}}{\sum_t m_{i,t}}.$$

The projection gives $B_i = 1$ analytically and a maximum observed numerical error of 2.22×10^{-16}. The unprojected variant ranges from 0.98 to 1.02. These quantities measure task-credit mass before clipping; they do not measure the resulting gradient norm or parameter update.

F.3 Offline PUU group-signal audit

Rollouts and scoring. We evaluate all 40 AMC 2023 problems using Qwen3-1.7B checkpoints from PUU training. Each checkpoint samples 12 responses per problem with temperature 0.6, top-$p = 0.95$, top-$k = 20$, and seed 42. The verifier assigns $R_i^{\text{task}} \in \{0, 1\}$ by exact integer match with the reference answer. The frozen Qwen3-4B-GRPO teacher and Qwen3-1.7B-Base anchor, denoted

by π_T and π_B, score the same realized response tokens at raw temperature 1. The offline teacher score is

$$R_i^T = \frac{1}{L_i} \sum_{t=1}^{L_i} \mathrm{clip}(\log \pi_T(y_{i,t} \mid s_{i,t}) - \log \pi_B(y_{i,t} \mid s_{i,t}), -5, 5)\,, \qquad R_i^U = R_i^{\mathrm{task}} + R_i^T, \quad (16)$$

which corresponds to $\alpha = 1$ in this audit. For each problem, a fixed seed partitions the 12 responses into a group of eight and a retained group of four, producing 80 pseudo-groups per checkpoint. These groups are used only for the diagnostic and do not alter training.

Group-level signal coverage. Panel (a) of Figure 4 uses checkpoints 30, 110, 190, 270, 350, 430, and 515. A group is verifier-degenerate when all its task rewards are identical. We report its all-wrong and all-correct portions, the fraction of all groups with $\mathrm{Std}(R^U) > 0.1$, and the fraction with $\max_i R_i^T - \min_i R_i^T > 0.05$. The latter two statistics are computed over all groups rather than conditioned on verifier degeneracy. Across the selected checkpoints, verifier-degenerate groups account for 40–53% of the total, while nontrivial unified-utility variation appears in 56–90%. The teacher-range statistic decreases from 86% at step 30 to 21% at step 515, showing that its additional response-level variation is strongest early in training.

Mixed-group advantage composition. Panel (b) uses checkpoints 30, 60, 90, 120, 180, 210, and 230 and retains only groups containing both verifier outcomes. We compare the verifier-only GRPO advantage with the exact PUU decomposition

$$A_i^U = A_i^{\mathrm{task}|U} + \alpha A_i^{T|U},$$

where the two components share the unified denominator. Each plotted value is the response-mean absolute magnitude at that checkpoint. We additionally compute

$$\mathrm{TeacherShare} = \frac{\sum_i |\alpha A_i^{T|U}|}{\sum_i |A_i^{\mathrm{task}|U}| + \sum_i |\alpha A_i^{T|U}|}.$$

The teacher share ranges from 17% to 25%, and $|A^U|$ remains close in mean magnitude to $|A^{\mathrm{GRPO}}|$. Thus the teacher term is visible without overwhelming the task component under this construction.

This is a signal-geometry audit rather than a training ablation. It uses pseudo-groups and a fixed base-model anchor, whereas the practical on-policy update uses the behavior policy as its current anchor. The plotted component magnitudes therefore do not estimate parameter-gradient or accuracy contributions; the controlled training variants are reported in Table 3.

F.4 Offline sensitivity of verifier-consistent ordering to α

Scope and fixed rollouts. We perform a fixed-rollout sensitivity audit of the PUU teacher coefficient α; this is not a training-performance ablation. The audit uses 17 Qwen3-1.7B checkpoints from UECR-GRPO training (steps 30, 60, 90, through 480, and 515) with the frozen Qwen3-4B-GRPO teacher. At each checkpoint, we reuse previously generated responses on the held-out AMC 2023 benchmark: 24 samples per problem through step 240 and 12 samples thereafter, for 40 problems per checkpoint. The resulting data contain 451 mixed verifier groups and 22,811 correct–incorrect response pairs. These are benchmark groups rather than training-prompt groups, and their sizes differ from the training group size $G = 8$.

Checkpoint-relative teacher score. For each response, the teacher and the student checkpoint that generated that response are evaluated at raw temperature 1. We compute

$$R_i^T = \frac{1}{L_i} \sum_{t=1}^{L_i} \mathrm{clip}(\log \pi_T(y_{i,t} \mid s_{i,t}) - \log \pi_{\mathrm{old},i}(y_{i,t} \mid s_{i,t}), -5, 5)\,, \quad (17)$$

where $\pi_{\mathrm{old},i}$ is the generating checkpoint, not a fixed reference. The masked token mean, per-token clipping, and exclusion of prompt and padding tokens match the training reward configuration. Log probabilities are obtained once with vLLM (Kwon et al., 2023) prompt scoring and cached before the CPU-only α sweep. Scoring uses a maximum model length of 8,192 tokens. Responses beyond this limit, accounting for approximately 10–24% depending on the checkpoint, are head-truncated; their R_i^T therefore uses only the available response tokens. Truncation and non-finite-token counts are recorded rather than silently discarded.

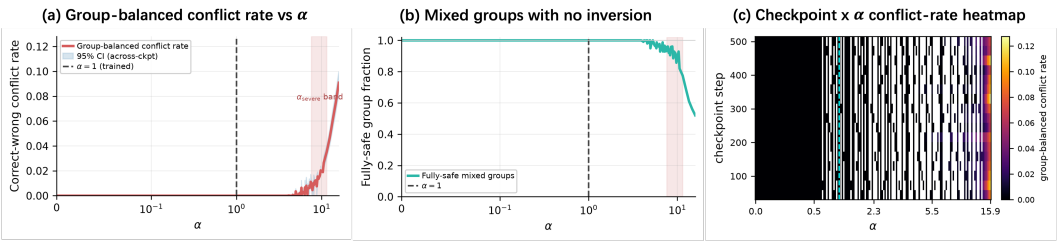


Figure 7: Fixed-rollout sensitivity of verifier-consistent ordering to α on AMC 2023. The audit covers 17 UECR-GRPO-trained Qwen3-1.7B checkpoints, 451 mixed verifier groups, and 22,811 correct–incorrect pairs. (a) Group-balanced strict conflict rate, averaged across checkpoints, with a 95% problem-cluster bootstrap interval. (b) Fraction of mixed groups in which every correct response remains above every incorrect response. (c) Checkpoint-wise group-balanced conflict rate. Dashed lines mark the trained setting $\alpha = 1$; the shaded band spans the checkpoint-wise coefficients 7.6–11.5 at which at least 10% of mixed groups first contain an inversion. This is an offline ordering audit, not a retraining or performance comparison.

Ordering metrics. For a mixed group g, let $\mathcal{C}_g$ and $\mathcal{W}_g$ denote its correct and incorrect responses. At coefficient α,

$$U_c(\alpha) = 1 + \alpha R_c^T, \qquad U_w(\alpha) = \alpha R_w^T. \tag{18}$$

Group normalization is order preserving, so comparing these utilities is equivalent to comparing their PUU advantages. With tie tolerance 10^{-6}, the primary conflict statistic first averages strict inversions within each group and then gives every mixed group equal weight:

$$C_g(\alpha) = \frac{1}{|\mathcal{C}_g||\mathcal{W}_g|} \sum_{c \in \mathcal{C}_g} \sum_{w \in \mathcal{W}_g} \mathbf{1}\left[U_w(\alpha) > U_c(\alpha) + 10^{-6}\right]. \tag{19}$$

We also report a pair-micro average as a secondary statistic. A group is fully safe at α when $M_g(\alpha) = \min_{c \in \mathcal{C}_g} U_c(\alpha) - \max_{w \in \mathcal{W}_g} U_w(\alpha) > 10^{-6}$. Its first analytical tie point is

$$\alpha_g^* = \left(\max_{w \in \mathcal{W}_g} R_w^T - \min_{c \in \mathcal{C}_g} R_c^T \right)^{-1} \tag{20}$$

when the parenthesized difference is positive, and $+\infty$ otherwise.

We separately measure whether the sign of $A_i^U(\alpha)$ differs from that of the centered task reward, using population standard deviation and $\epsilon = 10^{-6}$. This trajectory-sign statistic is not an ordering statistic: a correct response can remain above every incorrect response yet fall below the unified group mean. All-correct and all-wrong groups are excluded from cross-label inversion metrics; for them we report only whether the teacher score supplies nontrivial within-group variation.

Sweep and uncertainty. The pre-specified sweep combines explicit operating points with a dense grid over $[0, 8]$. A second grid places ten points below the first coefficient at which at least 10% of mixed groups at a checkpoint contain an inversion. We use 2,000 bootstrap replicates with seed 42, clustered by AMC problem, for 95% intervals. The analytical thresholds are checked against the grid transitions, and $\alpha = 0$ recovers unit correct–incorrect margins and zero conflicts. For a zero-event estimate at $\alpha = 1$, we additionally compute the corresponding 95% rule-of-three upper bound from the eligible sample count.

Results and limitations. At $\alpha = 1$, we observe no strict correct–incorrect ordering inversion at any of the 17 checkpoints: both the group-balanced conflict rate and the any-inversion group rate are zero over the 451 eligible mixed groups and 22,811 pairs. The distinct trajectory-sign statistic ranges from 0 to 3.1% across checkpoints. The most restrictive observed group has $\alpha_g^* \approx 3.88$; the checkpoint-wise 5th percentile of α_g^*, corresponding to 95% empirically safe groups, ranges from 5.3 to 9.3, while the median ranges from 12.6 to 21.3. Between 0 and 7.7% of groups at a checkpoint have $\alpha_g^* = +\infty$. Thus $\alpha = 1$ empirically preserves verifier-consistent correct–incorrect ordering on these observed rollouts, but this does not constitute a universal guarantee.

The audit is limited to held-out AMC 2023 rollouts rather than the training prompt distribution; it uses benchmark sample groups of size 12 or 24 rather than training groups of eight; long responses

are truncated at 8,192 tokens; and $\pi_{\text{old},i}$ is the checkpoint that generated each response rather than the transient old policy inside a training iteration. Finally, the audit tests cross-label ordering only. It does not establish that teacher-induced ordering within the correct or incorrect class reflects process quality, nor does it estimate the performance that retraining at another α would produce.

F.5 Offline ECR credit-localization audit

Data and selection. This audit uses the shared initialization, with Qwen3-1.7B-Base as both the student and old policy and the frozen Qwen3-4B-GRPO model as teacher. Prompts are drawn from the difficulty-5–7 portion of DeepMath-103K after shuffling with seed 42. The student samples eight responses per prompt at temperature 0.7, top-$p = 1$, and a 16,384-token response limit. A rule-based verifier compares the extracted $\boxed{\cdots}$ answer with the reference answer. Teacher and old-policy log probabilities are evaluated at raw temperature 1, and teacher entropy uses the full softmax distribution in float32.

We retain real eight-response groups containing both rewards, select only incorrect responses with $A_i^{\text{task}|U} < 0$, and remove empty or reasoning-free outputs. Selection does not use teacher scores, entropy, ECR weights, or anticipated labels. This produces 200 candidate responses for blinded annotation. A deterministic parser divides each response at line, enumeration, and sentence boundaries. Opus 4.8 sees only the problem, reference answer, and numbered reasoning steps, and labels each step as valid, invalid, uncertain, or not reasoning. It also marks at most one first substantive error. Method signals and model identities remain hidden. Uncertain and non-reasoning steps are excluded from the primary statistic.

Prior work contrasts outcome feedback with supervision of individual reasoning steps (Uesato et al., 2022; Lightman et al., 2024) and constructs process rewards without human step annotations (Wang et al., 2024). Here, step judgments audit ECR's token weights; they are not training rewards.

Counterfactual weights. For every retained token, we reproduce the training implementation in Eq. (10). Four weights are computed on the same responses: PUU sets $w_{i,t} = 1$; the no-entropy variant sets $c_{i,t} = 1$ and recomputes the projection; the shuffled control independently permutes $H_{i,t}^T$ across valid tokens within each response 100 times, preserving its multiset while breaking positional alignment; and Full ECR uses the original entropy and complete zero-sum projection. Each shuffle recomputes $c_{i,t}$, μ_i^c, $q_{i,t}$, and $w_{i,t}$.

For response i, the primary localization statistic is

$$D_i = \operatorname*{mean}_{t \in \mathcal{I}_i} w_{i,t} - \operatorname*{mean}_{t \in \mathcal{V}_i} w_{i,t}, \tag{21}$$

where $\mathcal{I}_i$ and $\mathcal{V}_i$ contain tokens in judge-invalid and judge-valid steps. Because the analyzed task advantage is negative, $D_i > 0$ means that invalid steps receive stronger negative task credit than valid steps. We macro-average D_i over responses and obtain 95% confidence intervals by resampling prompts 10,000 times with seed 42. The reported $\Pr(D_i > 0)$ is the fraction of eligible responses with a positive per-response gap, rather than a token-level significance test. A secondary statistic compares the first erroneous step with valid steps preceding it. All inference remains clustered by prompt.

Results and scope. Full ECR gives $D = 0.0048$ and lies to the right of all 100 entropy-position permutations, whose mean gap is 0.0024. Zero exceedances among 100 shuffles are a finite permutation comparison, not a zero p-value (Phipson & Smyth, 2010). PUU is exactly zero because it broadcasts one task weight across the response. The positive Full-ECR gap shows that its redistribution agrees on average with an independent audit of step quality, while the permutation result indicates that entropy contributes positional information beyond its marginal distribution. The projection preserves a unit response-wise mean weight and hence reallocates rather than increases the verifier-derived budget. This audit tests semantic alignment of the realized weights; it neither estimates gradient contribution nor proves a causal improvement in downstream accuracy.

F.6 Offline verifier-gradient alignment audit

We use the 366 difficulty-1–4 MATH-500 problems and checkpoints at steps 30, 240, and 515 from the aligned training runs, including UECR-GRPO and ATOD. Each frozen checkpoint generates

eight responses per problem using temperature 0.6, top-$p = 0.95$, top-$k = 20$, and a 2,048-token response limit. Correctness is determined from the extracted final $\boxed{\cdots}$ answer. Empty and truncated responses are excluded.

We split the problems into two disjoint folds of 183 problems. For each split direction, $A \to B$ and $B \to A$, we reconstruct response-masked, clipped-surrogate gradients for UECR-GRPO, ATOD, and Vanilla GRPO on the candidate fold. On the other fold, we compute a verifier-only GRPO gradient from binary correctness rewards. Each candidate gradient and its verifier reference are evaluated at the same frozen student parameters; teacher-derived quantities are detached, and no optimizer step is taken. Following the motivation of Lin et al. (2019) and Yu et al. (2020), we measure their directional agreement as

$$C_m^{A \to B} = \frac{\langle g_m^A, g_V^B \rangle}{\|g_m^A\|_2 \|g_V^B\|_2}, \qquad m \in \{\mathrm{UECR}, \mathrm{ATOD}, \mathrm{GRPO}\}, \tag{22}$$

and analogously for $B \to A$. Figure 6 reports the mean and standard deviation of these two split directions.

UECR-GRPO shows the highest mean cosine at each displayed step, with a smaller margin over ATOD late in training. Cosine similarity removes gradient-norm differences but does not reproduce an optimizer update. Moreover, two folds do not establish statistical significance, and comparisons involving different training trajectories do not isolate the effect of the update rule at identical parameters. This audit therefore tests local alignment with a held-out verifier objective, not causal improvement in final accuracy.

F.7 Aligned configurations

For the 1.7B setting, we train a Qwen3-1.7B student for 515 steps on difficulty-5–7 problems from DeepMath-103K, using a frozen Qwen3-4B-GRPO teacher. Rollouts use temperature 0.7, top-$p = 0.95$, top-$k = 20$, a 2,048-token prompt limit, and a 16,384-token response limit. Optimization uses a learning rate of 10^{-6}, one PPO epoch per batch, and symmetric clipping of 0.2. Reference-policy regularization is applied through the actor-loss `low_var_kl` with coefficient 0.001, while reward-level KL is disabled. UECR-GRPO uses $\alpha = 1$, $\rho = 0.5$, $\tau_\delta = 1.0$, and $\tau_H = 2.0$. Realized-token log probabilities and full-vocabulary entropy are obtained from the same teacher logits at raw $T = 1$ in a single chunked scoring pass.

The Qwen3-4B student uses a frozen Qwen3-8B-Math-GRPO teacher and a sampled subset of DeepMath-103K problems with difficulty levels 6–8. The global prompt batch size is 126, with $G = 8$. Both models use non-thinking mode in training and evaluation.

Data hygiene and checkpoint selection. The draft protocol screens training manifests against AIME 2024, AIME 2025, and AMC 2023 using normalized exact match, n-gram overlap, and embedding retrieval followed by manual inspection. The screening results and actual checkpoint-selection rule still require confirmation from the completed runs.

Training dynamics. Figures 8 and 9 show the training dynamics of the aligned 1.7B and 4B student runs, respectively.

F.8 Remaining sensitivity and additional controls

Appendix F.4 reports the completed fixed-rollout sensitivity audit for α. Retraining-based sensitivity to α, ρ, τ_δ, and τ_H remains future work; the offline audit does not estimate the performance of those counterfactual training runs.

The following controls complement Table 3:

1. set $\rho = 0$ to recover PUU exactly;
2. set $c_{i,t} = 1$ while retaining the projection;
3. set $\mu_i^c = 0$ to remove the additive constraint;
4. apply $w_{i,t}$ to the full A_i^U, testing whether the teacher sequence component should remain unmodulated;
5. set $\alpha = 0$ while retaining ECR;
6. remove the signed gap to test whether confidence alone supplies a useful direction.

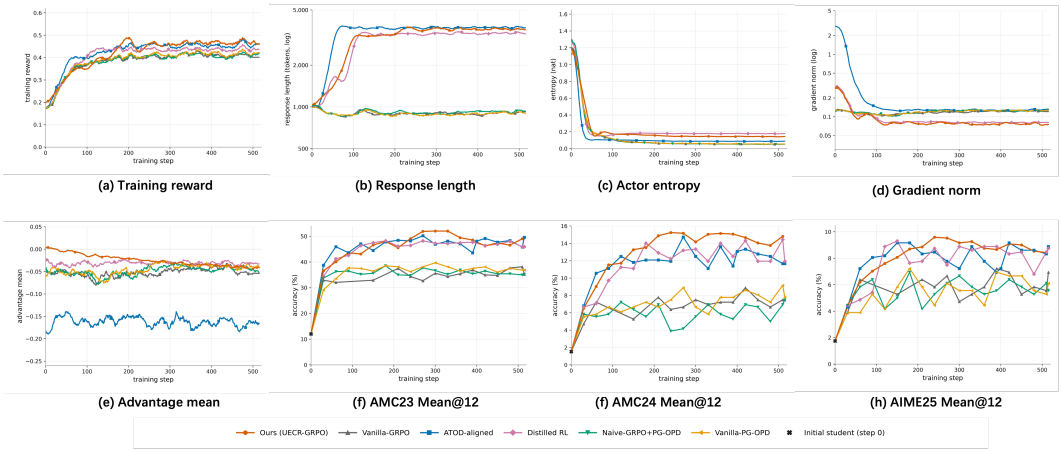


Figure 8: Training dynamics for the aligned runs with the Qwen3-1.7B student and Qwen3-4B-GRPO teacher, including training reward, response length, actor entropy, gradient norm, advantage statistics, and fixed-decoding validation accuracy.

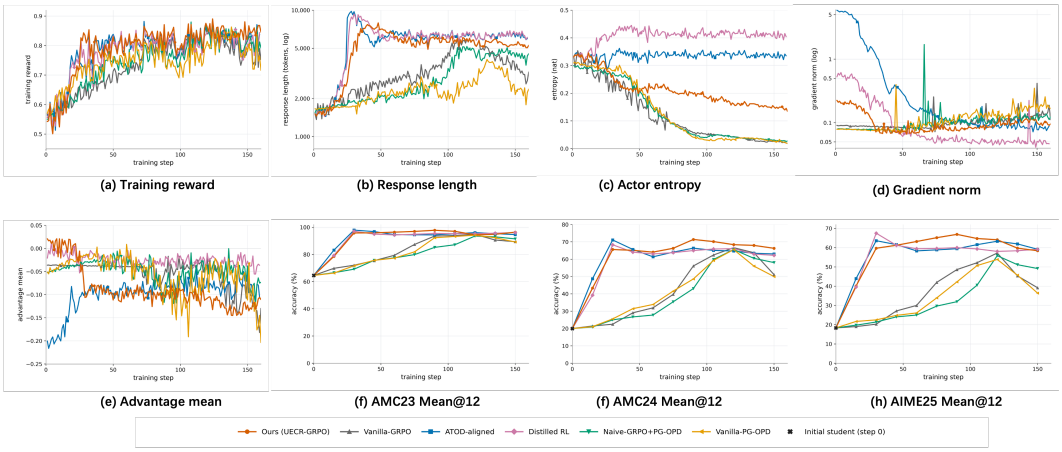


Figure 9: Training dynamics for the aligned runs with the Qwen3-4B student and Qwen3-8B-Math-GRPO teacher, including training reward, response length, actor entropy, gradient norm, advantage statistics, and fixed-decoding validation accuracy.

F.9 Efficiency and batch-fact auditing

We report wall-clock, rollout time, scorer time, tokens/s, GPU0–7 peak memory, and scorer overhead relative to Vanilla GRPO. Logs contain unique prompts, generated/rewarded/loss trajectories, valid response tokens, global steps, and optimizer steps. Seven-way data-parallel padding is removed before response-wise normalization and restored only as masked rows. A real empty response is an error. Any run that changes the intended global batch, duplicates prompts, drops trajectories, or miscounts microbatches as global steps fails the audit.

F.10 Regression and smoke tests

Before full training, the implementation verifies that $\rho = 0$ reproduces the PUU loss and gradient, projection invariants hold on all valid masks, raw-$T = 1$ teacher and old-policy scores align token by token, and reference KL enters once. Disabling teacher scoring must recover native GRPO without loading the teacher. CPU unit tests precede 1-, 5-, and 10-step GPU runs.

G REPRODUCIBILITY CHECKLIST

PPO-style results depend on implementation conventions (Huang et al., 2024); we therefore record the scoring, objective, and effective-batch details needed to reproduce the comparisons.

Artifacts. Archive checkpoint, tokenizer, chat template, manifest, source commit, and resolved-configuration hashes. Record reward and group-normalization conventions, decoding parameters, answer extraction, checkpoint selection, and per-problem outputs.

Shared telemetry. For every method, log reward and advantage statistics, zero-variance groups, PPO ratio/clip/KL statistics, response length and truncation, prompt, trajectory, and token counts, and peak memory. Teacher metrics are disabled, not zero, for Vanilla GRPO.

Method-specific telemetry. Additionally log δ, H^T, c, d, μ_i^c, q, w, teacher reward, decomposed PUU advantages, response-wise budget error, and $\max_i |A_i^U - (A_i^{\mathrm{task}|U} + \alpha A_i^{T|U})|$. State numerical tolerances before training and treat violations as run failures.