ϵ-POLICY OPTIMIZATION: TIMING-AWARE CONFIDENCE-BASED EXPLORATION CONTROL IN RLVR

Anonymous authors
Paper under double-blind review

ABSTRACT

Reinforcement learning with verifiable rewards (RLVR) improves language model reasoning, but reward-driven optimization can prematurely concentrate policies on successful behaviors, limiting exploration of alternative reasoning trajectories. Existing methods emphasize how to preserve exploration, with less attention to when it is needed. Since concentration also enables exploitation, exploration control must distinguish premature concentration from useful consolidation. We introduce confidence, defined through token probabilities, to connect policy concentration with token-level optimization dynamics. Our analysis identifies two mechanisms that increase concentration: reinforcing likely tokens and suppressing unlikely alternatives. Based on that, we propose ϵ-Policy Optimization (ϵ-PO), a timing-aware exploration control method that estimates exploration space from rollout confidence and intervenes only when it falls below a desired margin ϵ. It selectively suppresses token updates that contribute to concentration, otherwise retaining the original objective to minimize intervention. And annealing ϵ progressively relaxes intervention to balance exploration and exploitation throughout training. Across four backbone models and five reasoning benchmarks, ϵ-PO achieves higher average Pass@1 and Pass@16 than the evaluated baselines. Training dynamics and ablations support its effectiveness in delaying early concentration while allowing later exploitation.

1 INTRODUCTION

Reinforcement learning with verifiable rewards (RLVR) improves large language models' reasoning by optimizing automatically verifiable outcomes, enabling complex reasoning behaviors to emerge through trial-and-error (Shao et al., 2024; DeepSeek-AI et al., 2025). Group Relative Policy Optimization (GRPO) further improves this process using relative rewards among sampled responses (Shao et al., 2024). However, its reward-driven optimization can rapidly concentrate the policy on previously discovered successful behaviors (Cui et al., 2025; Chen et al., 2025). Such concentration is necessary for exploitation, but premature concentration may limit exploration of alternative reasoning trajectories before the solution space is sufficiently explored. Therefore, RLVR exploration control must address both how and when to preserve exploration.

Existing methods mainly focus on how to preserve exploration. DAPO expands the probability-ratio clipping region for tokens that contribute to entropy increase (Yu et al., 2025), while Clip-Cov (Cui et al., 2025) and entropy-based methods (Jiang et al., 2025; Wang et al., 2025a) identify or regularize uncertainty-reducing updates. Other approaches maintain reasoning diversity through confidence-aware advantage shaping (Xie et al., 2025; Li et al., 2026) or staged exploration (Li et al., 2025; Chen et al., 2025). These studies highlight the importance of exploration control in RLVR, but leaves when to preserve exploration remains largely unexplored. Since concentration is essential for successful policy optimization, timing-aware control must distinguish premature concentration from necessary consolidation.

To enable timing-aware control, we introduce confidence (token output probabilities) to connect policy concentration with optimization dynamics. At the distribution level, confidence measures concentration, with its complement representing the remaining exploration space. At the token level,

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.

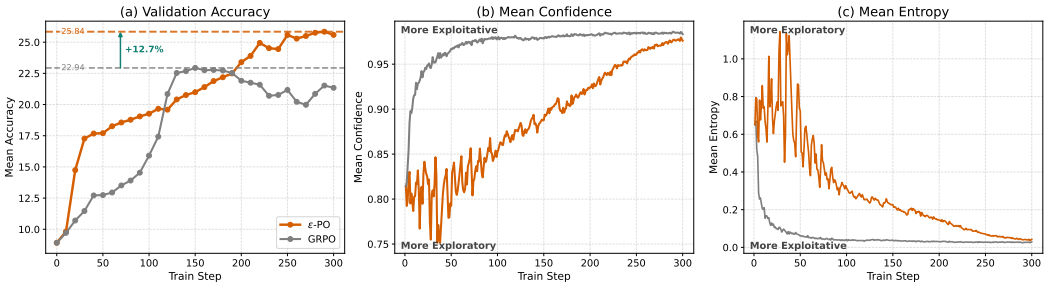


Figure 1: **Controlling confidence growth to balance exploration and exploitation on Qwen3-4B-Base.** (a) Validation accuracy, (b) mean confidence, and (c) mean entropy during training. GRPO rapidly concentrates and its validation accuracy later declines. ϵ-PO preserves lower confidence and higher entropy early, then gradually permits concentration while validation accuracy continues to improve. The 12.7% annotation in (a) denotes the relative improvement from 22.94% to 25.84% validation accuracy.

it is the probability of the sampled action optimized during RLVR, linking individual updates to concentration changes. We relate confidence to policy uncertainty and analyze its gradient-driven dynamics, showing that both reinforcing likely actions and suppressing unlikely alternatives can increase concentration. These findings explain how RLVR optimization changes exploration space and motivate selective control.

Based on these findings, we propose ϵ-Policy Optimization (ϵ-PO), a confidence-based timing-aware exploration control mechanism for RLVR. Following a minimal-intervention principle, it monitors rollout batch confidence and intervenes only when concentration exceeds the level allowed by an exploration margin ϵ. It restricts high-confidence, positive-advantage token updates that reinforce dominant behaviors, preserving negative correction and reinforcement of less certain behaviors. To allow later consolidation, we anneal ϵ during training, raising the confidence threshold and progressively relaxing intervention (Figure 1). Together, confidence-based triggering and margin annealing determine when intervention is activated and when it is relaxed. Experiments across four backbone models and five reasoning benchmarks show that ϵ-PO achieves higher average Pass@1 and Pass@16 than the evaluated baselines.

Our main contributions are as follows:

- We study how to balance exploration and exploitation in RLVR by controlling premature policy concentration. We link concentration to token-level GRPO dynamics through confidence, revealing how different updates affect exploration space and motivating selective control.
- We propose ϵ-PO, a lightweight timing-aware confidence-based method that monitors concentration, selectively restricts dominant-token reinforcement, and gradually relaxes intervention through exploration-margin annealing.
- Experiments across four backbone models and five reasoning benchmarks show that ϵ-PO achieves higher average Pass@1 and Pass@16 than the evaluated baselines. Training trajectories and ablations support its effectiveness in delaying early concentration while allowing later exploitation.

2 Preliminaries

RLVR and group relative policy optimization. For a prompt $x \sim \mathcal{D}$, policy π_θ generates $y = (a_1, \ldots, a_T)$ over vocabulary $\mathcal{V}$ of size V, with state $s_t = (x, a_{<t})$. RLVR maximizes the expected verifier reward $R(x, y)$. To optimize this objective, GRPO samples G responses per prompt from $\pi_{\theta_{\text{old}}}$ and assigns each token in response i the advantage $\hat{A}_i = (r_i - \mu_r)/(\sigma_r + \varepsilon_A)$ (Shao et al., 2024). Here μ_r, σ_r are the group reward mean and standard deviation, and $\varepsilon_A > 0$ stabilizes normalization. With importance ratio $\rho_{i,t}(\theta) = \pi_\theta(a_{i,t} \mid s_{i,t})/\pi_{\theta_{\text{old}}}(a_{i,t} \mid s_{i,t})$, the token surrogate

follows PPO clipping (Schulman et al., 2017):

$$\ell_{i,t}^{\mathrm{GRPO}}(\theta) = \min\Big\{\rho_{i,t}(\theta)\hat{A}_i, \mathrm{clip}(\rho_{i,t}(\theta), 1-\delta_{\mathrm{clip}}, 1+\delta_{\mathrm{clip}})\,\hat{A}_i\Big\}. \tag{1}$$

For response lengths T_i, the length-normalized objective is:

$$\mathcal{J}_{\mathrm{GRPO}}(\theta) = \mathbb{E}\left[\frac{1}{G}\sum_{i=1}^{G}\frac{1}{T_i}\sum_{t=1}^{T_i}\ell_{i,t}^{\mathrm{GRPO}}(\theta)\right]. \tag{2}$$

3 METHOD

ϵ-Policy Optimization (ϵ-PO) controls concentration through three decisions: when to intervene, which token updates to suppress, and how to relax intervention during training. It activates when average rollout confidence exceeds a scheduled threshold, then masks positive-advantage updates for tokens whose probabilities exceed the same threshold. Annealing the exploration margin raises this threshold, allowing increasing concentration as training progresses. We develop the confidence statistic and its local dynamics before presenting the masked objective and schedule.

3.1 CONFIDENCE AS A MEASURE OF CONCENTRATION

Token confidence. We use the probability of a sampled token to connect rollout monitoring with token-level optimization. For a token-level state s and a sampled action a, define

$$c_\theta(a \mid s) = \pi_\theta(a \mid s). \tag{3}$$

Token confidence measures how strongly the policy supports the sampled action. It is already available from the rollout policy for the GRPO importance ratio.

State confidence. The expected confidence of a policy sample measures the concentration of its next-token distribution:

$$C_\theta(s) = \mathbb{E}_{a\sim\pi_\theta(\cdot|s)}[c_\theta(a \mid s)] = \sum_{v\in\mathcal{V}}\pi_\theta(v \mid s)^2. \tag{4}$$

Equivalently, $C_\theta(s)$ is the probability that two independent samples from the next-token distribution produce the same token (Simpson, 1949). Its complement,

$$E_\theta(s) = 1 - C_\theta(s) = \Pr_{a,b\overset{\mathrm{iid}}{\sim}\pi_\theta(\cdot|s)}(a \neq b), \tag{5}$$

measures the probability that two independent next-token samples differ. We use this probability to quantify exploration space at a given state: higher confidence means that repeated samples are more likely to select the same token, leaving less room for alternative choices.

Connection to entropy. Confidence also provides a lower bound on Shannon entropy, defined as $H_\theta(s) = -\sum_{v\in\mathcal{V}}\pi_\theta(v \mid s)\log\pi_\theta(v \mid s)$ with natural logarithms.

Proposition 1 (Entropy–confidence relationship)**.** *For any categorical policy,*

$$H_\theta(s) \geq -\log C_\theta(s). \tag{6}$$

For any fixed state distribution, $\mathbb{E}_s[H_\theta(s)] \geq -\log\mathbb{E}_s[C_\theta(s)]$.

Equation (6) connects confidence to policy uncertainty: lower confidence yields a higher lower bound on Shannon entropy. Here, $-\log C_\theta(s)$ is order-2 Rényi entropy (Rényi, 1961). The bound follows from Jensen's inequality, with the proof provided in Appendix B.1.

3.2 CONFIDENCE DYNAMICS UNDER GRPO

To guide selective intervention, we analyze how GRPO token-level updates affect confidence. At a fixed state s, let $z_v = z_\theta(s)_v$ denote the policy's pre-softmax logit for token $v \in \mathcal{V}$, and write $c_v = \pi_\theta(v \mid s) = \exp(z_v)/\sum_w \exp(z_w)$ and $C_s = \sum_v c_v^2$. We analyze a local logit-space step, progressing from one coordinate to the complete batch update.

Lemma 1 (Sensitivity to a token logit)**.** *For a perturbation of z_v with the other logits fixed,*

$$\frac{\partial C_s}{\partial z_v} = 2c_v(c_v - C_s), \qquad \Delta C_{s,v} = 2c_v(c_v - C_s)\Delta z_v + O((\Delta z_v)^2). \tag{7}$$

Raising z_v increases confidence when $c_v > C_s$ and decreases it when $c_v < C_s$, to first order. Here, $\Delta C_{s,v}$ captures only the effect of changing z_v while holding the other logits fixed. The following theorem also considers the combined effect of all logit changes induced by a GRPO update.

Theorem 1 (Local confidence dynamics under GRPO)**.** *Consider an unclipped GRPO term at $\theta = \theta_{\text{old}}$, with sampled token v, fixed advantage A_v, and an infinitesimal gradient-ascent step in logit space. Its sampled-logit step is $\Delta z_v = \eta A_v(1 - c_v)$, where $\eta > 0$ is the step size. Then*

$$\Delta C_{s,v} = 2\eta c_v(1 - c_v)A_v(c_v - C_s) + O(\eta^2). \tag{8}$$

For $0 < c_v < 1$, its first-order term is positive exactly when $A_v(c_v - C_s) > 0$. The complete update changes all logits by $\Delta z_i = \eta A_v(\mathbf{1}[i = v] - c_i)$, yielding

$$\begin{aligned} \Phi_{s,v} &= c_v(c_v - C_s) - \left(\sum_i c_i^3 - C_s^2\right), \\ \Delta C_{s,\text{full}} &= 2\eta A_v \Phi_{s,v} + O(\eta^2). \end{aligned} \tag{9}$$

The sampled-coordinate term identifies two ways an update can contribute to increasing confidence:

1. **Reinforcing likely tokens:** $A_v > 0$ and $c_v > C_s$. Raising the sampled token's logit contributes to concentration by favoring an already likely choice.
2. **Suppressing unlikely tokens:** $A_v < 0$ and $c_v < C_s$. Lowering the sampled token's logit contributes to concentration by removing probability from an unlikely choice.

From token updates to batch confidence. The complete logit contributions add to give the first-order change in confidence over a fixed batch. Fix a rollout batch $\mathcal{B}$ of N token positions. For each position u, denote its state, sampled action, advantage, and positive loss weight by s_u, a_u, A_u, and w_u. Let $\mathcal{S}_\mathcal{B}$ be the set of visited states and $q_s = |\{u \in \mathcal{B} : s_u = s\}|/N$ their empirical frequencies. The exact batch confidence is $C_\mathcal{B} = \sum_{s \in \mathcal{S}_\mathcal{B}} q_s C_s$. With independently parameterized logits at distinct states, take a gradient-ascent step on $\mathcal{J}_\mathcal{B} = \sum_{u \in \mathcal{B}} w_u \ell_u^{\text{GRPO}}$ at the rollout policy, holding states, advantages, and weights fixed. Equation (9) gives

$$\begin{aligned} \Delta C_\mathcal{B} &= \eta \sum_{u \in \mathcal{B}} d_u + O(\eta^2), \\ d_u &= 2q_{s_u} w_u A_u \Phi_{s_u, a_u}. \end{aligned} \tag{10}$$

Each d_u is the contribution of one token loss to the batch confidence drift. The factors w_u and q_{s_u} separately account for loss normalization and state frequency. Masking a loss removes its contribution from this local sum. Appendix B.2 gives the derivation, including repeated states.

3.3 ϵ-POLICY OPTIMIZATION

Exploration margin. The exploration margin $\epsilon_k \in [0, 1]$ specifies a target minimum for exploration space at iteration k, corresponding to the confidence threshold

$$\tau_k = 1 - \epsilon_k. \tag{11}$$

At a single state, $C_\theta(s) > \tau_k$ is equivalent to a disagreement probability $E_\theta(s) < \epsilon_k$. We use this threshold to determine whether to activate intervention and which token updates to suppress.

Algorithm 1 ϵ-Policy Optimization

Require: Policy π_θ, iterations K, group size G, margin ϵ_0

for $k = 0, \ldots, K-1$ **do**
  Set $\theta_{\text{old}} \leftarrow \theta$; sample a batch of prompts
  Sample G responses per prompt from $\pi_{\theta_{\text{old}}}$
  Compute verifier rewards and group-relative advantages A_u
  Collect token confidences c_u and estimate $\bar{C}_k$ by Eq. (12)
  Set ϵ_k by Eq. (16) and $\tau_k \leftarrow 1 - \epsilon_k$
  Compute γ_k by Eq. (13) and masks by Eq. (14)
  Optimize Eq. (15), holding advantages and masks fixed
end for

Batch-level monitoring. We monitor concentration using sampled-token probabilities already collected for GRPO. Let $\mathcal{T}_k$ contain the sampled token indices $u = (j, i, t)$ for prompt j, response i, and position t, and let $N_k = |\mathcal{T}_k|$. Compute

$$c_u = c_{\theta_{\text{old}}}(a_u \mid s_u), \qquad \bar{C}_k = \frac{1}{N_k} \sum_{u \in \mathcal{T}_k} c_u. \tag{12}$$

For an action sampled on policy at state s_u, $\mathbb{E}[c_u \mid s_u] = C_{\theta_{\text{old}}}(s_u)$. Thus, $\bar{C}_k$ estimates average state confidence over visited token positions without computing the full vocabulary sum in Eq. (4). We activate intervention when

$$\gamma_k = \mathbf{1}[\bar{C}_k > \tau_k]. \tag{13}$$

Token-level intervention. When intervention is active, we suppress positive-advantage token terms whose token confidence exceeds the threshold. Writing $A_u = \hat{A}_{j,i}$, define the binary mask

$$m_u^{(\epsilon)} = 1 - \gamma_k \, \mathbf{1}[A_u > 0] \, \mathbf{1}[c_u > \tau_k]. \tag{14}$$

The mask targets positive reinforcement of already dominant tokens, one source of concentration identified in Section 3.2. Negative-advantage updates remain active, preserving learning from responses with negative advantages.

The mask is computed from the rollout batch and held fixed during its optimization. Applying the mask to the surrogate yields

$$\mathcal{J}_{\epsilon\text{-PO}}(\theta) = \mathbb{E}\left[\frac{1}{G} \sum_{i=1}^{G} \frac{1}{T_i} \sum_{t=1}^{T_i} m_{i,t}^{(\epsilon)} \ell_{i,t}^{\text{GRPO}}(\theta) \right], \tag{15}$$

An inactive trigger exactly recovers GRPO because $\gamma_k = 0$ makes every mask equal one. When intervention is active, masked token terms contribute no gradient; verifier rewards, advantages, and clipping remain unchanged. No auxiliary exploration loss is added.

Annealing the exploration margin. A fixed threshold may continue suppressing dominant-token reinforcement even when later consolidation is desirable. We therefore progressively relax intervention by linearly annealing the exploration margin:

$$\epsilon_k = \epsilon_0 \left(1 - \frac{k}{K}\right)_+, \tag{16}$$

where $\epsilon_0 \in [0, 1)$ is the initial margin, $(x)_+ = \max\{x, 0\}$, and K is the number of training iterations. A larger initial margin lowers $\tau_0 = 1 - \epsilon_0$, making batch activation easier and more positive-advantage tokens eligible for masking. As ϵ_k decreases, the threshold for batch activation and token masking rises, progressively relaxing intervention and allowing more dominant-token reinforcement. The schedule is prescribed by training progress; adaptation to observed concentration comes from the batch trigger. Appendix B.4 gives further details.

Connection to the confidence dynamics. When $\tau_k \geq 1/2$, every term selected for masking would make a positive first-order contribution to confidence under the local logit update. The practical rule compares c_u with τ_k, while the sampled-coordinate condition in Theorem 1 compares it with $C_{s_u} = C_{\theta_{\text{old}}}(s_u)$ at the rollout policy. For a selected token, these conditions are linked by

$$C_{s_u} \leq c_u^2 + (1 - c_u)^2 < c_u \qquad \text{when } \tfrac{1}{2} < c_u < 1. \tag{17}$$

Thus, $c_u > \tau_k \geq 1/2$ and $A_u > 0$ imply a positive sampled-coordinate contribution. In this regime, all other token probabilities at the same state are below C_{s_u}, so their downward logit updates also increase confidence to first order. Consequently, $\Phi_{s_u,a_u} > 0$ for the complete update, connecting the mask to the batch dynamics in Eq. (10).

Proposition 2 (Reduction of batch confidence drift)**.** *Consider the batch logit model of Eq.* (10) *for $\mathcal{B} = \mathcal{T}_k$, with finite logits and $\tau_k \geq 1/2$. Compare masked and unmasked gradient-ascent steps from the same rollout policy, with the same step size and fixed weights; hold the masks fixed. Let $D_{\text{GRPO}} = \sum_u d_u$, $D_{\epsilon\text{-}PO} = \sum_u m_u^{(\epsilon)} d_u$, and $\mathcal{M}_k = \{u : m_u^{(\epsilon)} = 0\}$. Every masked term has $d_u > 0$, and*

$$\begin{aligned} D_{\epsilon\text{-}PO} &= D_{\text{GRPO}} - \sum_{u \in \mathcal{M}_k} d_u \leq D_{\text{GRPO}}, \\ \Delta C_{\mathcal{B}}^{\epsilon\text{-}PO} - \Delta C_{\mathcal{B}}^{\text{GRPO}} &= -\eta \sum_{u \in \mathcal{M}_k} d_u + O(\eta^2). \end{aligned} \tag{18}$$

The first-order reduction is strict whenever $\mathcal{M}_k$ is nonempty.

Under the stated local logit model, masking reduces the first-order increase in batch confidence relative to unmasked GRPO. The threshold condition holds throughout the schedule when $\epsilon_0 \leq 1/2$. Appendix B.3 provides the proof.

Algorithm 1 summarizes ϵ-PO. Section 4 evaluates this design through confidence trajectories and ablations of token selection, batch activation, and annealing.

4 Experiments

We evaluate reasoning performance, examine concentration and masking during training, and assess the effects of the initial exploration margin, batch trigger, and margin annealing.

4.1 Experimental Setup

Models and training. We train Qwen3-1.7B-Base, Qwen3-4B-Base, Qwen3-8B-Base, and Llama-3.2-3B-Instruct (Yang et al., 2025; Meta, 2024) on DAPO-Math-17k using VERL (Yu et al., 2025; Sheng et al., 2024). We compare ϵ-PO with the starting checkpoint, GRPO (Shao et al., 2024), entropy regularization (Jiang et al., 2025), KL regularization, and Clip-Cov and KL-Cov (Cui et al., 2025) under a shared training and evaluation protocol; Appendix D.2 details the baseline methods and their hyperparameters. Training uses batches of 128 prompts with eight responses per prompt and a maximum response length of 8,192 tokens. Unless stated otherwise, ϵ-PO uses $\epsilon_0 = 0.20$ with linear annealing. Appendix D specifies optimizer settings and prompts.

Benchmarks and metrics. We evaluate mathematical reasoning on AIME 24, AIME 25, AMC 23, and MATH500, and scientific reasoning on GPQA-Diamond (Mathematical Association of America, 2026; Hendrycks et al., 2021; Lightman et al., 2023; Rein et al., 2023). We report Pass@1 and Pass@16 (%) (Chen et al., 2021), with an unweighted average over the five benchmarks. Evaluation uses temperature 0.6 and top-p 0.95, with 128 responses per problem for AIME and AMC and 32 for MATH500 and GPQA-Diamond. Both metrics are estimated from these samples as described in Appendix D.

4.2 Main Results

Higher average Pass@1 and Pass@16. ϵ-PO achieves the highest average Pass@1 and Pass@16 among the evaluated methods on all four backbones (Table 1). In order of increasing Qwen model size

Table 1: Reasoning performance across four backbones. P@1 and P@16 denote Pass@1 and Pass@16 (%); Average is the unweighted mean over five benchmarks, and GPQA denotes GPQA-Diamond. Base denotes the starting checkpoint before RLVR. Bold marks the best value in each metric column within each backbone.

Method	AIME 24		AIME 25		AMC 23		GPQA		MATH500		Average	
	P@1 ↑	P@16 ↑	P@1 ↑	P@16 ↑	P@1 ↑	P@16 ↑	P@1 ↑	P@16 ↑	P@1 ↑	P@16 ↑	P@1 ↑	P@16 ↑
Qwen3-1.7B-Base												
Base	3.85	19.71	2.72	17.69	30.24	71.98	21.40	83.54	54.64	86.74	22.57	55.93
GRPO	6.55	23.44	6.17	23.99	41.86	76.09	25.73	79.86	68.17	85.46	29.70	57.77
Entropy Reg.	9.53	27.46	4.82	21.21	43.89	77.95	**27.57**	**86.57**	67.31	88.01	30.62	60.24
KL Reg.	8.41	25.02	4.77	20.31	43.40	74.74	27.38	83.85	67.03	86.51	30.20	58.09
Clip-Cov	9.56	27.49	4.37	20.36	43.50	73.30	25.31	75.85	66.01	84.64	29.75	56.33
KL-Cov	7.55	26.54	5.34	23.56	41.04	75.23	25.04	82.28	69.07	88.72	29.61	59.27
ϵ-PO (ours)	**10.92**	**27.85**	**6.77**	**25.95**	**44.05**	**78.06**	26.93	85.47	**70.18**	**89.26**	**31.77**	**61.32**
Qwen3-4B-Base												
Base	10.10	33.88	7.73	30.16	45.24	85.46	28.87	83.51	70.06	92.68	32.40	65.14
GRPO	25.44	42.65	20.43	41.76	69.24	92.87	39.70	79.37	84.55	95.08	47.87	70.35
Entropy Reg.	19.75	39.94	20.10	36.40	54.23	87.18	39.95	78.93	81.06	93.62	43.02	67.21
KL Reg.	24.56	45.36	17.43	43.63	69.26	93.75	40.32	80.84	84.62	96.03	47.24	71.92
Clip-Cov	23.70	43.43	16.65	42.09	67.76	**94.39**	41.03	81.72	81.20	94.67	46.07	71.26
KL-Cov	26.56	**51.30**	21.46	43.66	**71.23**	93.69	40.66	85.03	85.55	**96.33**	49.09	74.00
ϵ-PO (ours)	**28.01**	50.03	**23.67**	**47.32**	70.74	94.16	**43.36**	**85.16**	**85.67**	95.18	**50.29**	**74.37**
Qwen3-8B-Base												
Base	12.81	41.03	11.25	32.33	54.06	86.71	35.73	87.02	73.96	93.50	37.56	68.12
GRPO	28.54	55.46	22.49	47.11	73.83	**95.05**	47.84	85.96	88.37	97.05	52.21	76.13
Entropy Reg.	28.93	59.32	21.43	43.83	75.10	94.20	46.61	86.28	88.06	96.23	52.03	75.97
KL Reg.	28.10	54.96	21.38	44.85	75.98	94.78	47.38	86.21	89.32	**97.42**	52.43	75.64
Clip-Cov	27.76	60.49	**25.70**	44.91	76.37	94.56	47.91	86.82	88.76	96.81	53.30	76.72
KL-Cov	31.30	58.20	23.93	40.62	75.98	94.61	48.36	**89.78**	**89.94**	97.15	53.90	76.07
ϵ-PO (ours)	**31.67**	**60.90**	23.91	**47.93**	**77.21**	95.00	**49.00**	88.08	89.66	96.99	**54.29**	**77.78**
Llama-3.2-3B-Instruct												
Base	3.61	21.34	0.25	3.60	16.54	66.22	13.19	67.72	34.16	**77.91**	13.55	47.36
GRPO	9.92	22.97	0.23	4.30	27.14	64.11	14.70	64.54	46.18	71.24	19.63	45.43
Entropy Reg.	10.52	27.07	0.81	6.87	33.42	65.11	21.79	66.15	46.19	72.65	22.55	47.57
KL Reg.	8.62	22.93	0.70	**8.19**	32.95	65.02	16.71	64.01	45.93	72.64	20.98	46.56
Clip-Cov	10.34	25.84	0.28	2.70	32.81	64.90	20.95	62.26	**49.33**	74.44	22.74	46.03
KL-Cov	9.89	23.74	0.78	4.85	35.72	66.66	12.14	63.21	46.60	72.32	21.03	46.16
ϵ-PO (ours)	**11.16**	**28.14**	**1.03**	4.41	**35.85**	**69.21**	**21.84**	**71.57**	48.61	71.50	**23.70**	**48.97**

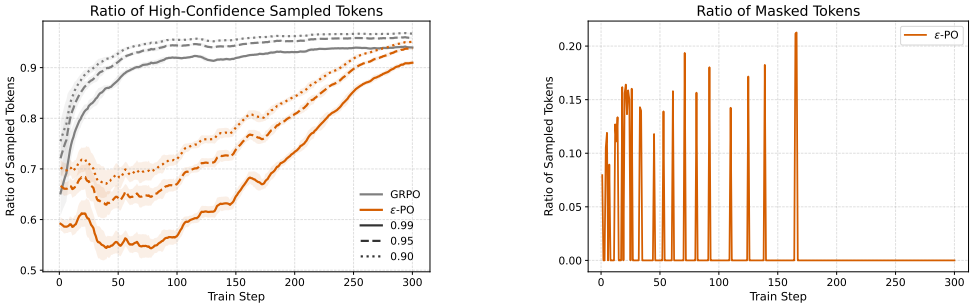
followed by Llama, its Pass@1 gains over GRPO are 2.07, 2.42, 2.08, and 4.07 percentage points; the corresponding Pass@16 gains are 3.55, 4.02, 1.65, and 3.54 points. Relative to the strongest baseline average for each backbone and metric, the gains range from 0.39 to 1.20 points for Pass@1 and from 0.37 to 1.40 points for Pass@16.

Consistent gains across benchmarks. ϵ-PO improves over GRPO on all 20 backbone–benchmark pairs for Pass@1 and 18 of 20 for Pass@16. Other methods still achieve the best scores on some individual benchmark metrics.

4.3 Concentration and Intervention Dynamics

Concentration is delayed, then resumes. Figure 1 compares validation accuracy with mean confidence and entropy during training. GRPO rapidly reaches high confidence and low entropy; its validation accuracy subsequently plateaus and declines. In contrast, ϵ-PO maintains lower early confidence and higher entropy, with validation accuracy continuing to improve later. Its confidence eventually rises and entropy falls, matching the intended progression from early exploration toward later exploitation.

Slower early growth in the fraction of high-confidence tokens. To examine the token distribution behind the aggregate curves, we measure the fraction of rollout tokens above a fixed confidence



(a) High-confidence token ratio

(b) Masked-token ratio

Figure 2: **Token concentration and intervention on Qwen3-4B-Base.** (a) Fractions of sampled tokens with $c_u > q$ for fixed diagnostic thresholds $q \in \{0.90, 0.95, 0.99\}$. (b) Fraction masked after applying the batch trigger $\bar{C}_k > \tau_k$ and token rule $A_u > 0$, $c_u > \tau_k$. Masking occurs at selected training steps and stops in the later part of training.

Table 2: Effect of the initial exploration margin ϵ_0 on reasoning performance. We report P@1 and P@16, denoting Pass@1 and Pass@16 (%), on Qwen3-4B-Base with linear annealing. Average is the unweighted mean over five benchmarks; GPQA denotes GPQA-Diamond. GRPO is included as the baseline. Bold marks the best value in each metric column.

Method	ϵ_0	AIME 24		AIME 25		AMC 23		GPQA		MATH500		Average	
		P@1 ↑	P@16 ↑	P@1 ↑	P@16 ↑	P@1 ↑	P@16 ↑	P@1 ↑	P@16 ↑	P@1 ↑	P@16 ↑	P@1 ↑	P@16 ↑
GRPO	–	25.44	42.65	20.43	41.76	69.24	92.87	39.70	79.37	84.55	95.08	47.87	70.35
ϵ-PO	0.10	26.30	48.69	22.37	45.24	70.92	**95.34**	42.96	82.62	83.81	**95.70**	49.27	73.52
	0.15	26.77	47.39	**23.80**	45.80	**71.15**	93.52	41.90	84.57	84.68	94.70	49.66	73.20
	0.20	**28.01**	**50.03**	23.67	**47.32**	70.74	94.16	**43.36**	**85.16**	**85.67**	95.18	**50.29**	**74.37**

threshold q:

$$R_{>q}(k) = \frac{1}{N_k} \sum_{u \in \mathcal{T}_k} \mathbf{1}[c_u > q], \tag{19}$$

where $\mathcal{T}_k$ is the set of sampled token positions at iteration k, $N_k = |\mathcal{T}_k|$, and c_u is the rollout probability of sampled token u. Figure 2(a) uses $q \in \{0.90, 0.95, 0.99\}$. Under GRPO, the fraction of sampled tokens above each threshold rises rapidly. Under ϵ-PO, these fractions remain lower early in training and increase later, consistent with the mean-confidence trajectories.

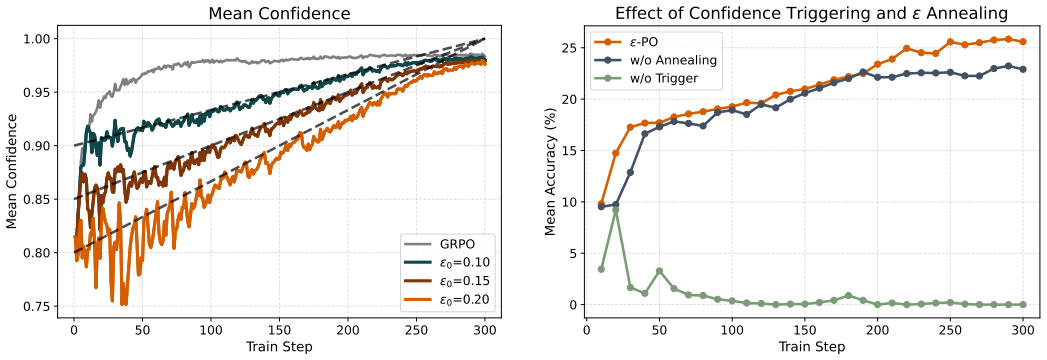
Masking stops in the later part of training. We measure the fraction of rollout tokens masked at each training step,

$$R_{\text{mask}}(k) = \frac{1}{N_k} \sum_{u \in \mathcal{T}_k} (1 - m_u^{(\epsilon)}). \tag{20}$$

In Figure 2(b), masking occurs at selected training steps and stops in the later part of training. Most nonzero masked-token fractions are below 20%, with one peak slightly above it.

4.4 Exploration Margin and Annealing

All tested initial margins improve average performance. We vary $\epsilon_0 \in \{0.10, 0.15, 0.20\}$ on Qwen3-4B-Base to test sensitivity to the initial exploration target (Table 2). These margins set initial confidence thresholds $\tau_0 \in \{0.90, 0.85, 0.80\}$. All three improve average Pass@1 over GRPO by 1.40–2.42 points and average Pass@16 by 2.85–4.02 points. The default $\epsilon_0 = 0.20$ gives the highest averages, although smaller margins lead some individual benchmark entries. Figure 3(a) shows lower early confidence under larger margins, followed by increasing confidence for every annealed schedule.



(a) Initial exploration margin

(b) Trigger and annealing

Figure 3: **Effects of the exploration margin, batch trigger, and annealing on Qwen3-4B-Base.** (a) Confidence trajectories under different ϵ_0 values; dashed lines show the scheduled thresholds $\tau_k = 1 - \epsilon_k$. (b) Validation accuracy when the batch trigger is removed or the threshold is fixed. Larger initial margins maintain lower early confidence, while removing the trigger or disabling annealing reduces validation accuracy.

Effects of batch triggering and margin annealing. We separately remove the batch trigger and disable annealing to assess their effects on validation accuracy (Figure 3(b)). Without the trigger, $\gamma_k = 1$ at every step while token selection remains conditional on advantage and confidence; validation accuracy falls close to zero in this run. Without annealing, the threshold stays fixed at $\tau_k = 1 - \epsilon_0$; validation accuracy improves initially but plateaus below the full method later in training. These results support conditioning intervention on observed confidence and progressively relaxing it during training.

5 Related Work

Concentration and selective updates. DAPO adjusts clipping and token aggregation to improve reasoning RL (Yu et al., 2025). Entropy analyses connect concentration to probability–logit covariance and local policy updates (Cui et al., 2025; Wang et al., 2026), motivating Clip-Cov and KL-Cov interventions. Confidence-guided methods instead reshape advantages using uncertainty or generative probabilities (Xie et al., 2025; Li et al., 2026). Our analysis links confidence to token-level drift and uses one scheduled threshold for both batch activation and selective suppression of dominant-token reinforcement.

Preserving exploration and controlling its timing. AEPO regulates entropy around a target (Wang et al., 2025a), while CURE and EEPO promote rollout exploration through branching or temporary unlearning (Li et al., 2025; Chen et al., 2025). These approaches provide mechanisms for preserving exploration. Our focus is when to suppress concentration-driving updates: ϵ-PO activates a token mask only when batch confidence exceeds a scheduled threshold, then progressively relaxes intervention through annealing to permit later exploitation. This couples intervention timing with selective update control while retaining the GRPO advantages and rollout procedure. Appendix A provides further comparisons.

6 Conclusion

We introduced ϵ-PO to control premature concentration in RLVR. Guided by confidence-based analysis, it conditionally masks high-confidence, positive-advantage token updates and anneals the exploration margin to permit later consolidation. Across four backbones and five reasoning benchmarks, ϵ-PO achieves the highest average Pass@1 and Pass@16 among the evaluated methods on every backbone. Training dynamics and ablations support delaying early concentration while allowing later exploitation.

AI USE STATEMENT

We used generative AI tools to implement the method, translate and polish the manuscript, assist in writing mathematical proofs, and generate plotting code for the figures. The authors reviewed the AI-assisted code, text, proofs, and figures and verified their correctness. We take responsibility for the final content of this work, including all AI-assisted text, claims, code, and figures.

REFERENCES

Liang Chen, Xueting Han, Qizhou Wang, Bo Han, Jing Bai, Hinrich Schütze, and Kam-Fai Wong. EEPO: Exploration-enhanced policy optimization via sample-then-forget. *arXiv preprint arXiv:2510.05837*, 2025. URL https://arxiv.org/abs/2510.05837.

Mark Chen, Jerry Tworek, Heewoo Jun, et al. Evaluating large language models trained on code. *arXiv preprint arXiv:2107.03374*, 2021. URL https://arxiv.org/abs/2107.03374.

Ganqu Cui, Yuchen Zhang, Jiacheng Chen, et al. The entropy mechanism of reinforcement learning for reasoning language models. *arXiv preprint arXiv:2505.22617*, 2025. URL https://arxiv.org/abs/2505.22617.

DeepSeek-AI, Daya Guo, Dejian Yang, et al. DeepSeek-R1: Incentivizing reasoning capability in LLMs via reinforcement learning. *arXiv preprint arXiv:2501.12948*, 2025. URL https://arxiv.org/abs/2501.12948.

Tuomas Haarnoja, Aurick Zhou, Kristian Hartikainen, George Tucker, Sehoon Ha, Jie Tan, Vikash Kumar, Henry Zhu, Abhishek Gupta, Pieter Abbeel, and Sergey Levine. Soft actor-critic algorithms and applications. *arXiv preprint arXiv:1812.05905*, 2018. URL https://arxiv.org/abs/1812.05905.

Zhezheng Hao, Hong Wang, Haoyang Liu, Jian Luo, Jiarui Yu, Hande Dong, Qiang Lin, Can Wang, and Jiawei Chen. Rethinking entropy interventions in RLVR: An entropy change perspective. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 2026. URL https://aclanthology.org/2026.acl-long.1436/.

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

Yuxian Jiang, Yafu Li, Guanxu Chen, Dongrui Liu, Yu Cheng, and Jing Shao. Rethinking entropy regularization in large reasoning models. *arXiv preprint arXiv:2509.25133*, 2025. URL https://arxiv.org/abs/2509.25133.

Pengyi Li, Elizaveta Goncharova, Andrey Kuznetsov, and Ivan Oseledets. Back to basics: Revisiting exploration in reinforcement learning for LLM reasoning via generative probabilities. *arXiv preprint arXiv:2602.05281*, 2026. URL https://arxiv.org/abs/2602.05281.

Qingbin Li, Rongkun Xue, Jie Wang, et al. CURE: Critical-token-guided re-concatenation for entropy-collapse prevention. *arXiv preprint arXiv:2508.11016*, 2025. URL https://arxiv.org/abs/2508.11016.

Hunter Lightman, Vineet Kosaraju, Yura Burda, Harri Edwards, Bowen Baker, Teddy Lee, Jan Leike, John Schulman, Ilya Sutskever, and Karl Cobbe. Let's verify step by step. *arXiv preprint arXiv:2305.20050*, 2023. URL https://arxiv.org/abs/2305.20050.

Mingjie Liu, Shizhe Diao, Ximing Lu, Jian Hu, Xin Dong, Yejin Choi, Jan Kautz, and Yi Dong. ProRL: Prolonged reinforcement learning expands reasoning boundaries in large language models. *arXiv preprint arXiv:2505.24864*, 2025a. URL https://arxiv.org/abs/2505.24864.

Zichen Liu, Changyu Chen, Wenjun Li, Penghui Qi, Tianyu Pang, Chao Du, Wee Sun Lee, and Min Lin. Understanding R1-Zero-like training: A critical perspective. *arXiv preprint arXiv:2503.20783*, 2025b. URL https://arxiv.org/abs/2503.20783.

Mathematical Association of America. American mathematics competitions, 2026. URL https://maa.org/student-programs/amc/. Competition information. Accessed September 24, 2026.

Meta. Llama 3.2 model card, 2024. URL https://huggingface.co/meta-llama/Llama-3.2-3B-Instruct.

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

Alfréd Rényi. On measures of entropy and information. In *Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Probability*, volume 1, pp. 547–561. University of California Press, 1961. URL https://static.renyi.hu/renyi_cikkek/1961_on_measures_of_entropy_and_information.pdf.

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.

Rulin Shao, Shuyue Stella Li, Rui Xin, Scott Geng, Yiping Wang, Sewoong Oh, Simon Shaolei Du, Nathan Lambert, Sewon Min, Ranjay Krishna, Yulia Tsvetkov, Hannaneh Hajishirzi, Pang Wei Koh, and Luke Zettlemoyer. Spurious rewards: Rethinking training signals in RLVR. *arXiv preprint arXiv:2506.10947*, 2025. URL https://arxiv.org/abs/2506.10947.

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

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

Edward H. Simpson. Measurement of diversity. *Nature*, 163:688, 1949. doi: 10.1038/163688a0. URL https://doi.org/10.1038/163688a0.

Chen Wang, Zhaochun Li, Jionghao Bai, Yuzhi Zhang, Shisheng Cui, Zhou Zhao, and Yue Wang. Arbitrary entropy policy optimization: Entropy is controllable in reinforcement finetuning. *arXiv preprint arXiv:2510.08141v1*, 2025a. URL https://arxiv.org/abs/2510.08141v1.

Shenzhi Wang, Le Yu, Chang Gao, et al. Beyond the 80/20 rule: High-entropy minority tokens drive effective reinforcement learning for LLM reasoning. *arXiv preprint arXiv:2506.01939*, 2025b. URL https://arxiv.org/abs/2506.01939.

Shumin Wang, Yuexiang Xie, Wenhao Zhang, Yuchang Sun, Yanxi Chen, Yaliang Li, and Yanyong Zhang. On the entropy dynamics in reinforcement fine-tuning of large language models. *arXiv preprint arXiv:2602.03392*, 2026. URL https://arxiv.org/abs/2602.03392.

Zhiheng Xi, Xin Guo, Yang Nan, Enyu Zhou, Junrui Shen, Wenxiang Chen, Jiaqi Liu, Jixuan Huang, Xun Deng, Zhihao Zhang, Honglin Guo, Zhikai Lei, Miao Zheng, Guoteng Wang, Peng Sun, Rui Zheng, Hang Yan, Tao Gui, Qi Zhang, and Xuanjing Huang. BAPO: Stabilizing off-policy reinforcement learning for LLMs via balanced policy optimization with adaptive clipping. In *International Conference on Learning Representations*, 2026. URL https://proceedings.iclr.cc/paper_files/paper/2026/hash/cca4df598d8066ac9e985852dd1d48b5-Abstract-Conference.html.

Can Xie, Ruotong Pan, Xiangyu Wu, Yunfei Zhang, Jiayi Fu, Tingting Gao, and Guorui Zhou. Unlocking exploration in RLVR: Uncertainty-aware advantage shaping for deeper reasoning. *arXiv preprint arXiv:2510.10649*, 2025. URL https://arxiv.org/abs/2510.10649.

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

Qiying Yu, Zheng Zhang, Ruofei Zhu, et al. DAPO: An open-source LLM reinforcement learning system at scale. *arXiv preprint arXiv:2503.14476*, 2025. URL https://arxiv.org/abs/2503.14476.

Yang Yue, Zhiqi Chen, Rui Lu, Andrew Zhao, Zhaokai Wang, Yang Yue, Shiji Song, and Gao Huang. Does reinforcement learning really incentivize reasoning capacity in LLMs beyond the base model? *arXiv preprint arXiv:2504.13837*, 2025. URL https://arxiv.org/abs/2504.13837.

Xiaoyun Zhang, Xiaojian Yuan, Di Huang, Wang You, Chen Hu, Jingqing Ruan, Ai Jian, Kejiang Chen, and Xing Hu. Revisiting entropy regularization: Adaptive coefficient unlocks its potential for LLM reinforcement learning. *arXiv preprint arXiv:2510.10959*, 2025. URL https://arxiv.org/abs/2510.10959.

Tianyu Zheng, Tianshun Xing, Qingshui Gu, Taoran Liang, Xingwei Qu, Xin Zhou, Yizhi Li, Zhoufutu Wen, Chenghua Lin, Wenhao Huang, Qian Liu, Ge Zhang, and Zejun Ma. First return, entropy-eliciting explore. *arXiv preprint arXiv:2507.07017*, 2025. URL https://arxiv.org/abs/2507.07017.

Xinyu Zhu, Mengzhou Xia, Zhepei Wei, Wei-Lin Chen, Danqi Chen, and Yu Meng. The surprising effectiveness of negative reinforcement in LLM reasoning. *arXiv preprint arXiv:2506.01347*, 2025. URL https://arxiv.org/abs/2506.01347.

A Supplementary Related Work

A.1 Reasoning Coverage and Optimization in RLVR

RLVR can improve the probability of a correct response without uniformly improving coverage across sampling budgets. Yue et al. (2025) find that, in their evaluated settings, gains at small k coexist with lower Pass@k than the base model at large k. ProRL reports broader coverage under prolonged training with KL control and reference-policy resets (Liu et al., 2025a). Together, these findings suggest that reasoning coverage depends on the training regime; contraction is not an inevitable consequence of RL. This motivates evaluating both single-sample accuracy and multi-sample coverage, as we do with Pass@1 and Pass@16, without treating either metric as a direct measure of semantic diversity.

The optimization objective and starting model also shape the effect of reward-based training. PPO introduced the clipped surrogate underlying GRPO (Schulman et al., 2017; Shao et al., 2024), and DeepSeek-R1 demonstrated reasoning-oriented RL at scale (DeepSeek-AI et al., 2025). DAPO changes clipping and token aggregation (Yu et al., 2025), while Dr. GRPO identifies normalization-induced optimization bias and modifies the objective to improve token efficiency (Liu et al., 2025b). Shao et al. (2025) further show that gains from uninformative rewards can depend strongly on the model family. These results motivate specifying the objective and evaluating multiple backbones. Our intervention operates on the token terms of GRPO and leaves its advantage estimator unchanged.

A.2 Entropy Dynamics and Adaptive Regularization

Entropy provides both an exploration objective and a diagnostic of policy concentration. In maximum-entropy RL, Soft Actor-Critic jointly optimizes return and entropy and can automatically tune the temperature coefficient (Haarnoja et al., 2018). For reasoning models, Cui et al. (2025) connect entropy change to probability–logit covariance, and Wang et al. (2026) derive local entropy-change expressions under logit and GRPO updates. Clip-Cov and KL-Cov restrict high-covariance updates (Cui et al., 2025), while STEER reweights tokens using estimated entropy changes (Hao et al., 2026). These methods establish a direct connection between concentration dynamics and selective update control.

Other approaches adapt regularization or the sampling distribution. SIREN selects actions and states for entropy regularization (Jiang et al., 2025); AER adapts regularization coefficients using task difficulty and entropy feedback (Zhang et al., 2025); and AEPO uses gradients on temperature-adjusted distributions to regulate entropy around a target (Wang et al., 2025a). BAPO adapts clipping bounds to balance positive- and negative-advantage contributions during off-policy training (Xi et al., 2026). Thus, adaptive exploration control is established prior work. The specific design of ϵ-PO couples a batch concentration trigger with a token-selection threshold, using confidence dynamics to identify positive updates to suppress and an annealed margin to release that suppression.

A.3 Confidence-Guided Updates and Structured Exploration

Token-level methods allocate learning signals according to uncertainty, confidence, or advantage sign. High-entropy minority-token training concentrates gradients on uncertain decisions (Wang et al., 2025b). UCAS combines response-level confidence modulation with token-level certainty penalties (Xie et al., 2025), whereas ProGRPO reweights advantages using prompt perplexity and answer confidence (Li et al., 2026). Zhu et al. (2025) show that penalizing incorrect responses can improve reasoning while preserving broader sampling coverage than positive-only reinforcement in their evaluated settings. Our method retains negative correction and reinforcement of less confident choices, while conditionally suppressing positive reinforcement of dominant tokens.

Trajectory-level methods change where or how alternatives are sampled. FR3E performs targeted rollouts from uncertain decision points to construct intermediate feedback (Zheng et al., 2025). CURE branches at high-entropy tokens before returning to standard sampling for exploitation (Li et al., 2025), and EEPO temporarily unlearns sampled responses between rollout stages to diversify subsequent samples (Chen et al., 2025). These mechanisms explicitly organize exploration across decisions or stages. ϵ-PO controls the training updates selected from ordinary GRPO rollouts, with no change to the rollout procedure.

A.4 Statistical Foundations of Confidence

The state-level statistic $C(p) = \sum_v p_v^2$ is a classical concentration measure associated with Simpson's diversity index (Simpson, 1949); it is also the probability that two independent samples agree. Its negative logarithm is order-2 Rényi entropy (Rényi, 1961), and $1 - C(p)$ measures next-token disagreement. These identities provide the statistical foundation for our terminology. Our contribution is the connection to local token-update dynamics and the resulting conditional intervention, rather than a new diversity index. The bound $H(p) \geq -\log C(p)$ does not imply that Shannon entropy and confidence rank every pair of distributions identically, or that next-token disagreement directly measures diversity of complete reasoning trajectories.

B Proofs and Method Details

B.1 Proof of Proposition 1

At a fixed state, write $p_v = \pi_\theta(v \mid s)$ and restrict expectations to the support $S = \{v : p_v > 0\}$. If $a \sim p$, convexity of $-\log$ gives

$$H(p) = \mathbb{E}_{a\sim p}[-\log p_a] \geq -\log \mathbb{E}_{a\sim p}[p_a] = -\log \sum_{v\in S} p_v^2 = -\log C(p). \tag{21}$$

Equality holds when the positive probabilities are equal, including a deterministic distribution as the one-element case. Averaging over any fixed state distribution and applying Jensen's inequality again yields

$$\mathbb{E}_s[H_\theta(s)] \geq \mathbb{E}_s[-\log C_\theta(s)] \geq -\log \mathbb{E}_s[C_\theta(s)]. \tag{22}$$

The bound is a population statement; replacing the expectation by a finite rollout estimate does not produce a certified entropy bound without accounting for estimation error.

For completeness, Cauchy–Schwarz gives $1 = (\sum_v p_v)^2 \leq V \sum_v p_v^2$, and $\sum_v p_v^2 \leq (\sum_v p_v)^2 = 1$. Hence $C \in [1/V, 1]$. If a and b are independent draws from p, then $\Pr(a = b) = \sum_v p_v^2$ and $\Pr(a \neq b) = 1 - C$. These identities justify the concentration and exploration-space interpretations in Section 3.1.

B.2 Proofs of Lemma 1 and Theorem 1

Logit sensitivity. At a fixed state s, let $c_i = \exp(z_i)/\sum_j \exp(z_j)$ and $C_s = C_\theta(s) = \sum_i c_i^2$. The softmax Jacobian is

$$\frac{\partial c_i}{\partial z_v} = c_i(\mathbf{1}[i = v] - c_v).$$

Therefore,

$$\begin{aligned} \frac{\partial C_s}{\partial z_v} &= 2\sum_i c_i^2(\mathbf{1}[i = v] - c_v) \\ &= 2c_v^2 - 2c_v \sum_i c_i^2 = 2c_v(c_v - C_s). \end{aligned} \tag{23}$$

A one-coordinate Taylor expansion proves Lemma 1. For a general vector perturbation, the corresponding identity is

$$\Delta C_s = 2\sum_i c_i(c_i - C_s)\Delta z_i + O(\|\Delta z\|^2). \tag{24}$$

GRPO gradient and clipping. For an active unclipped term, $\ell_v = \rho_v A_v$, where $\rho_v = c_v/c_{\mathrm{old},v}$. Holding the advantage and rollout probability fixed,

$$\frac{\partial \ell_v}{\partial z_i} = \rho_v A_v(\mathbf{1}[i = v] - c_i).$$

At the rollout policy $\rho_v = 1$, so the sampled-coordinate step is $\Delta z_v = \eta A_v(1 - c_v)$. Substitution in Lemma 1 proves the sampled-coordinate part of Theorem 1. More generally, if χ_v is one on an active ratio branch and zero on a flat clipped branch, the sampled-coordinate contribution is

$$\Delta C_{s,v} = 2\eta\chi_v\rho_v c_v(1 - c_v)A_v(c_v - C_s) + O(\eta^2). \tag{25}$$

We exclude nondifferentiable clipping boundaries from this local calculation. For finite logits, $0 < c_v < 1$, and the first-order sign is determined by $A_v(c_v - C_s)$ whenever the branch is active. Zero advantage, a flat branch, or $c_v = C_s$ gives a zero first-order term.

The complete update in logit space. For one unclipped term at the rollout policy, all coordinates change by $\Delta z_i = \eta A_v(\mathbf{1}[i = v] - c_i)$. Equation (24) gives

$$\begin{aligned}
\Delta C_{s,\text{full}} &= 2\eta A_v \sum_i c_i(c_i - C_s)(\mathbf{1}[i = v] - c_i) + O(\eta^2) \\
&= 2\eta A_v \left[c_v(c_v - C_s) - \sum_i c_i^2(c_i - C_s) \right] + O(\eta^2) \\
&= 2\eta A_v \left[c_v(c_v - C_s) - \left(\sum_i c_i^3 - C_s^2 \right) \right] + O(\eta^2) \\
&= 2\eta A_v \Phi_{s,v} + O(\eta^2).
\end{aligned}$$

This proves the full logit expression in Theorem 1. An active off-policy term multiplies this contribution by $\rho_v > 0$, while a flat clipped branch contributes zero. The majority-token regime used by ϵ-PO gives $\Phi_{s,v} > 0$, as shown in Appendix B.3.

Batch aggregation. Use the fixed batch, state frequencies, and objective weights defined in Section 3.2. Write $\mathbf{c}_s$ for the probability vector at state s. When distinct states have independent logit parameters, all terms visiting the same state act on the same vector, so the batch gradient step at the rollout policy is

$$\Delta \mathbf{z}_s = \eta \sum_{u \in \mathcal{B}: s_u = s} w_u A_u (\mathbf{e}_{a_u} - \mathbf{c}_s).$$

Taking a first-order expansion of the fixed-state average gives

$$\begin{aligned}
\Delta C_{\mathcal{B}} &= \sum_{s \in \mathcal{S}_{\mathcal{B}}} q_s \big(\nabla_{\mathbf{z}_s} C_s \big)^{\mathsf{T}} \Delta \mathbf{z}_s + O(\eta^2) \\
&= 2\eta \sum_{u \in \mathcal{B}} q_{s_u} w_u A_u \Phi_{s_u, a_u} + O(\eta^2) \\
&= \eta \sum_{u \in \mathcal{B}} d_u + O(\eta^2).
\end{aligned}$$

This proves Eq. (10). Grouping repeated states before summing includes the effect of each token loss on every occurrence of its state in the batch average. The weights w_u may encode response-wise length normalization or any other fixed positive aggregation scheme.

B.3 Proof of Proposition 2

Positivity of the full logit contribution. At a fixed state s, a masked token satisfies $A_v > 0$ and $c_v > \tau_k \geq 1/2$ at the rollout policy. Since the remaining probabilities sum to $1 - c_v$,

$$C_s = c_v^2 + \sum_{i \neq v} c_i^2 \leq c_v^2 + (1 - c_v)^2.$$

Moreover,

$$c_v - \left[c_v^2 + (1 - c_v)^2 \right] = (1 - c_v)(2c_v - 1) > 0.$$

Consequently $c_v > C_s$, giving a positive sampled-coordinate contribution. Every other probability $q = c_i$ satisfies $0 \leq q \leq 1 - c_v < 1/2$. Since $q^2 - q$ decreases on this interval,

$$C_s - q \geq c_v^2 + q^2 - q \geq c_v^2 + (1 - c_v)^2 - (1 - c_v) = c_v(2c_v - 1) > 0.$$

Thus $c_v > C_s$ and $c_i < C_s$ for $i \neq v$. Under positive-advantage logit ascent, $\Delta z_v = \eta A_v(1 - c_v) > 0$ and $\Delta z_i = -\eta A_v c_i \leq 0$. Each product in Eq. (24) is nonnegative, with a positive sampled-coordinate term. Equivalently,

$$\Phi_{s,v} = c_v(c_v - C_s)(1 - c_v) - \sum_{i \neq v} c_i^2(c_i - C_s) > 0.$$

Effect on the batch update. For every masked term u, the state frequency q_{s_u}, loss weight w_u, and advantage A_u are positive. The preceding argument therefore gives $d_u = 2q_{s_u} w_u A_u \Phi_{s_u,a_u} > 0$. Holding masks and normalization fixed replaces w_u by $m_u^{(\epsilon)} w_u$ in the batch derivation of Appendix B.2. Hence

$$D_{\epsilon\text{-PO}} = \sum_u m_u^{(\epsilon)} d_u = D_{\text{GRPO}} - \sum_{u \in \mathcal{M}_k} d_u \le D_{\text{GRPO}}.$$

Both updates start from the same logits and use the same step size, so their confidence changes satisfy

$$\Delta C_{\mathcal{B}}^{\epsilon\text{-PO}} - \Delta C_{\mathcal{B}}^{\text{GRPO}} = -\eta \sum_{u \in \mathcal{M}_k} d_u + O(\eta^2).$$

The leading-order reduction is strict when at least one term is masked. This proves Proposition 2 for the batch average, including repeated states.

A generic target below $1/2$ is insufficient: for probabilities $(0.40, 0.59, 0.01)$ and target 0.35, the first token exceeds the target but has $c_v = 0.40 < C_s = 0.5082$ and $\Phi_{s,v} < 0$. Positive reinforcement of this token decreases confidence even under the full logit update. This example falls outside the proposition's condition $\tau_k \ge 1/2$.

Scope for neural-policy training. For a neural policy and a fixed diagnostic state set $\mathcal{F}$, define $C_{\mathcal{F}}(\theta) = |\mathcal{F}|^{-1} \sum_{s \in \mathcal{F}} C_\theta(s)$. If $g_u = \nabla_\theta \ell_u$ and w_u is its aggregation weight, ordinary gradient ascent gives

$$\Delta C_{\mathcal{F}} = \eta \sum_u w_u \nabla_\theta C_{\mathcal{F}}^{\mathsf{T}} g_u + O(\eta^2).$$

Proposition 2 establishes a batch-level first-order reduction in the independent-state logit model. With shared neural parameters, each token loss can affect logits at other states, so the inner product above need not have the sign of the corresponding logit-space contribution. Adam and gradient clipping add optimizer interactions. Also, a mask computed from rollout probabilities can become stale after multiple updates.

B.4 Annealing and the Exploration–Exploitation Transition

The margin ϵ_k specifies the desired disagreement probability between two next-token samples. It is a control target, not a probability of replacing the policy's action with a random action. Decreasing this target allows increasing concentration as the policy learns.

At the terminal boundary of the linear schedule, $\epsilon_K = 0$ and $\tau_K = 1$. Since $\bar{C}_K \le 1$, the strict trigger is zero. Training iterations are indexed by $0, \dots, K-1$, so this endpoint describes the schedule's limit; the last scheduled update can still have a positive margin.

The sum of the scheduled margins over those K updates is

$$\sum_{k=0}^{K-1} \epsilon_k = \epsilon_0 \sum_{k=0}^{K-1} \left(1 - \frac{k}{K}\right) = \epsilon_0 \frac{K+1}{2}.$$

This quantity describes the schedule, not the realized amount of intervention. Masking also depends on the batch trigger, token probabilities, and advantages. In particular, a positive margin need not mask any token. The realized ratio in Eq. (20) is the appropriate empirical measure of intervention.

C Additional Intervention Analyses

C.1 Alternative Token-Intervention Strategies

Theorem 1 identifies both positive reinforcement above state confidence and negative reinforcement below it as sources of increasing sampled-coordinate confidence. We compare three practical rules under the same batch trigger and linear target schedule on Qwen3-4B-Base. With A_u the response advantage, let

$$B_u^+ = \mathbf{1}[A_u > 0]\mathbf{1}[c_u > \tau_k], \qquad B_u^- = \mathbf{1}[A_u < 0]\mathbf{1}[c_u < \epsilon_k]. \tag{26}$$

Table 3: Token-selection ablation on Qwen3-4B-Base with the same batch trigger and linear schedule. P@1 and P@16 denote Pass@1 and Pass@16 (%); Average is the unweighted mean over five benchmarks, and GPQA denotes GPQA-Diamond. The masks are defined in Eq. (27). Bold marks the best value in each metric column.

Intervention Strategy	AIME 24		AIME 25		AMC 23		GPQA		MATH500		Average	
	P@1 ↑	P@16 ↑	P@1 ↑	P@16 ↑	P@1 ↑	P@16 ↑	P@1 ↑	P@16 ↑	P@1 ↑	P@16 ↑	P@1 ↑	P@16 ↑
GRPO	25.44	42.65	20.43	41.76	69.24	92.87	39.70	79.37	84.55	95.08	47.87	70.35
Positive-high intervention (ϵ-PO)	**28.01**	**50.03**	**23.67**	**47.32**	**70.74**	**94.16**	**43.36**	85.16	85.67	95.18	**50.29**	74.37
Negative-low intervention	26.15	48.71	22.63	45.97	68.52	93.37	42.33	84.18	**86.22**	95.70	49.17	73.59
Bidirectional intervention	25.31	49.86	23.26	46.98	69.98	93.54	41.82	**86.17**	84.92	**96.12**	49.06	**74.53**

The three masks are

$$\begin{aligned} m_u^{(+)} &= 1 - \gamma_k B_u^+, \\ m_u^{(-)} &= 1 - \gamma_k B_u^-, \\ m_u^{(\pm)} &= 1 - \gamma_k \mathbf{1}[B_u^+ \vee B_u^-]. \end{aligned} \tag{27}$$

Each replaces the mask in Eq. (15); all other components are shared.

For all tested margins, $\tau_k \geq 0.80$, so every token selected by the positive rule satisfies the majority-token condition in Proposition 2. The rule therefore removes terms with positive local confidence drift under the proposition's logit-space assumptions. The negative threshold ϵ_k is a practical proxy: $c_u < \epsilon_k$ does not, by itself, prove $c_u < C_{s_u}$. Consequently the negative and bidirectional variants are not exact tests of all concentration-increasing updates. Unlike the majority-token implication for the positive rule, their state-level signs require full-distribution confidence measurements. Neither the aggregate trajectories nor this comparison measures token-level drift under the neural optimizer.

Positive high-confidence masking gives the best Pass@1. Table 3 shows that the positive rule used by ϵ-PO yields the highest average Pass@1 (50.29) and the strongest AIME 24 and AIME 25 results among these variants. The negative rule reaches averages of 49.17/73.59 for Pass@1/Pass@16 and wins MATH500 Pass@1. Bidirectional masking reaches 49.06/74.53 and has the best average Pass@16, exceeding the positive rule by 0.16 points. All three outperform GRPO's 47.87/70.35 averages. These point estimates support the positive rule as a useful selective control, while indicating that the best intervention depends on whether single-sample accuracy or broader coverage is emphasized.

D Experimental Setup and Prompts

D.1 Training and Evaluation Protocol

Table 4 summarizes the shared training and evaluation configuration. Unless otherwise specified, we exclude the KL divergence penalty and entropy regularization loss, and employ standard symmetric clipping. Evaluation reports Pass@1 and Pass@16. To reduce evaluation variance, we draw more samples for smaller and more difficult benchmarks: AIME 24, AIME 25, and AMC 23 use $N = 128$ sampled responses per problem, while MATH500 and GPQA-Diamond use $N = 32$ sampled responses per problem. Following standard pass@k estimation, Pass@16 is computed as

$$\text{Pass@}k = \mathbb{E}\left[1 - \frac{\binom{N-c}{k}}{\binom{N}{k}}\right], \tag{28}$$

where N is the number of sampled responses and c is the number of correct responses for a prompt; the numerator is zero when $N - c < k$. Pass@1 is the mean fraction c/N.

D.2 Baseline Methods

We compare ϵ-PO against the following baselines under the same training and evaluation protocol:

- Base: the starting checkpoint before our RLVR training, using a base checkpoint for Qwen3 and an instruction-tuned checkpoint for Llama-3.2-3B.

Table 4: Training and evaluation configurations.

Rollout Inference	
Global Batch Size	128
Rollout Num per Prompt	8
Temperature	1.0
Top-p	1.0
Top-k	-1
Max Prompt Length	2048
Max Response Length	8192
Training	
PPO Mini Batch Size	32
Advantage Estimation Type	GRPO
Optimizer	Adam
Learning Rate	10^{-6}
Weight Decay	0.1
(β_1, β_2)	$(0.9, 0.999)$
Gradient norm clipping	1.0
Learning Rate Scheduler	constant
Warmup Steps	0
Evaluation Setup	
Temperature	0.6
Top-p	0.95
Top-k	-1
Max Generation Length	8192

- GRPO: the clipped, group-relative policy objective in Section 2, without an entropy bonus or reference-policy KL penalty.
- Entropy Regularization: GRPO with an entropy bonus of coefficient 0.001.
- KL Regularization: GRPO with a reference-policy KL penalty of coefficient 0.001.
- Clip-Cov: covariance-based selection of token updates for clipping or gradient detachment (Cui et al., 2025), using $\omega_{\text{low}} = 1.0$, $\omega_{\text{high}} = 5.0$, and clipping ratio 0.0002.
- KL-Cov: a KL penalty on selected high-covariance token updates (Cui et al., 2025), using $k = 0.002$ and $\beta = 1$.

D.3 Computational Resources

All experiments were conducted using NVIDIA A100 80G GPUs. A single run for the 1.7B model takes approximately 8×20 GPU hours on $8\times$ A100 GPUs. A single run for the 4B model takes approximately 16×30 GPU hours on $16\times$ A100 GPUs. A single run for the 8B model takes approximately 16×35 GPU hours on $16\times$ A100 GPUs.

D.4 Prompts

Prompts for Training and Math Evaluation

{*problem statement*}

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

Prompts for GPQA-Diamond

{*problem statement*}
A. {*Choice_A*}
B. {*Choice_B*}
C. {*Choice_C*}
D. {*Choice_D*}

Please reason step by step, and put your final answer (only the choice letter) within \boxed{}.