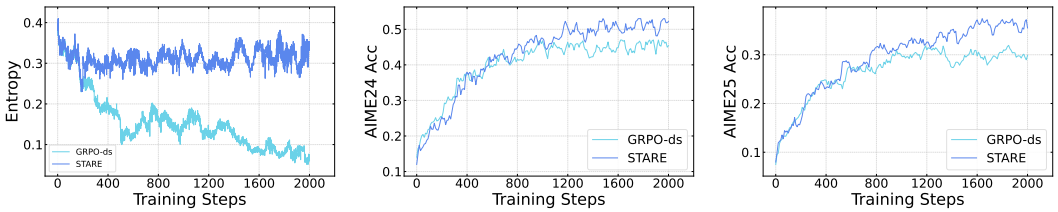
STARE: Surprisal-Guided Token-Level Advantage Reweighting for Policy Entropy Stability

Haipeng Luo[1,3] **Qingfeng Sun**[2] **Songli Wu**[1] **Can Xu**[2†] **Wenfeng Deng**[3†]
Han Hu[2] **Yansong Tang**[1]
[1]Shenzhen International Graduate School, Tsinghua University
[2]Tencent Hunyuan [3]Pengcheng Laboratory
{luohp24@mails., tang.yansong@sz.}tsinghua.edu.cn
{victorqsun,leocaxu}@tencent.com dengwf@pcl.ac.cn

Abstract

Reinforcement Learning with Verifiable Rewards algorithms like GRPO have emerged as the dominant post-training paradigm for complex reasoning in LLMs, yet commonly suffer from policy entropy collapse during training. We conduct a first-order gradient analysis of token-level entropy dynamics under GRPO and identify a token-level credit assignment mismatch: the per-token entropy variation decomposes into the product of the trajectory-level advantage and an entropy sensitivity function over the next-token distribution, yielding an advantage-surprisal four-quadrant structure and a near-criticality property. Motivated by it, we propose STARE (Surprisal-guided Token-level Advantage Reweighting for policy Entropy stability), which identifies entropy-critical token subsets via batch-internal surprisal quantiles, selectively reweights their effective advantages, and incorporates a target-entropy closed-loop gate for stable entropy regulation. Across model scales from 1.5B to 32B and three task families (Short CoT, Long CoT, and Multi-Turn Tool Use), STARE sustains stable RL training over thousands of steps while maintaining policy entropy within the target band. On AIME24 and AIME25, STARE outperforms DAPO and other competitive baselines by 4%–8% in average accuracy, with reflection tokens and response length growing in tandem, indicating sustained exploration–exploitation balance that further unlocks RL training potential. Code is available at https://github.com/hp-luo/STARE.



(a) Training Entropy (b) AIME24 Acc (c) AIME25 Acc

Figure 1: Training dynamics of STARE vs. GRPO-ds on Qwen2.5-7B-Base (cold-start SFT from Retool 2K) in the tool-use agent scenario: policy entropy, AIME24 accuracy, and AIME25 accuracy.

† Corresponding authors.

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

1 Introduction

Reinforcement Learning with Verifiable Rewards (RLVR) has emerged as the dominant post-training paradigm for eliciting complex reasoning in LLMs [1–5]. Among RLVR algorithms, Group Relative Policy Optimization (GRPO) dispenses with the value network and uses group-normalized rewards for advantage estimation; it is widely adopted in mathematical reasoning and code generation, and effectively induces emergent behaviors such as long-CoT reasoning and self-reflection [6, 7].

As RL training progresses, however, GRPO-style algorithms commonly suffer from policy entropy collapse: entropy decays rapidly, output diversity vanishes, and within-group rollouts homogenize, degrading advantage estimation and capping trainable steps [8, 9]. Existing mitigations fall into three categories. (i) Adjusting clipping thresholds for the importance-sampling ratio (e.g., DAPO's clip-higher) protects low-probability tokens but is largely inactive in the on-policy regime [10–13]. (ii) Asymmetric trajectory-level weighting between positive and negative rollouts provides only coarse-grained control [14–20]. (iii) Entropy-aware advantage reshaping or entropy regularization couples token-level entropy into the advantage but overamplifies high-entropy tokens and remains hyperparameter-sensitive [21–26]. These approaches operate at trajectory or sample granularity and lack a principled account of the collapse mechanism, leaving two questions open: which tokens drive entropy decay under GRPO, and how strong an intervention suffices to reverse it.

We conduct a first-order gradient analysis of token-level entropy dynamics under GRPO and identify a credit assignment mismatch: although the trajectory-level advantage $\hat{A}_i$ is shared across all tokens, the per-token entropy contribution decomposes into the product of $\hat{A}_i$ and a local entropy sensitivity Φ over the next-token distribution (Section 3). This yields an advantage–surprisal four-quadrant view: within positive-advantage trajectories, low-surprisal tokens dominate sampling and drive entropy-decreasing updates, whereas rare entropy-raising high-surprisal tokens are structurally diluted (a mirror-image asymmetry holds for negative-advantage trajectories). We further establish a near-criticality property: a mild weight perturbation suffices to flip the sign of entropy evolution, and beyond the critical threshold modulates the magnitude rather than the direction of the shift.

Motivated by these, we propose STARE (Surprisal-guided Token-level Advantage Reweighting for policy Entropy stability), a minimally invasive token-level credit-rebalancing mechanism within the clipped surrogate of GRPO. STARE selects an entropy-critical token subset via batch-internal surprisal quantiles, amplifying the effective advantage of positive-advantage high-surprisal tokens and attenuating that of negative-advantage ones. A target-entropy closed-loop gate governs the intervention: reweighting activates when the batch-averaged entropy $\bar{H}_k$ drops below H_{tgt} and reverts to GRPO once entropy recovers, yielding stable, low-intrusion entropy regulation (See Figure 1).

We validate STARE across multiple model scales and task regimes. STARE sustains over 5k RL training steps on 7B models and 1.5k steps on 14B/32B models, with the policy entropy held within the target band throughout, as shown in Figure 1. Across Short CoT, Long CoT, and multi-turn tool-use, STARE outperforms DAPO on AIME24/25 by 4%–8% in average accuracy, with reflection-related tokens and response length growing in tandem. The main contributions are: (i) From a first-order entropy-dynamics analysis, we expose the token-level credit assignment mismatch in GRPO and establish a near-criticality property: a mild weight perturbation flips the direction of entropy evolution while remaining robust to the specific weight value. (ii) We propose a surprisal-based advantage reweighting mechanism coupled with a target-entropy closed-loop constraint, achieving stable policy-entropy regulation through a minimal modification to the GRPO objective and sustaining RL training over thousands of steps. (iii) We validate STARE across 1.5B–32B model scales and Short CoT, Long CoT, and Tool Use regimes, where it maintains stable policy entropy and substantially outperforms DAPO and other competitive baselines. Related work is deferred to Appendix A.

2 Preliminaries

GRPO. Given prompt x, the old policy $\pi_{\theta_{\text{old}}}$ samples G responses with rewards $\{r_i\}_{i=1}^{G}$; the group-normalized advantage is $\hat{A}_i = \big(r_i - \text{mean}(\{r_j\})\big)/\,\text{std}(\{r_j\})$.The clipped surrogate objective is:

$$\mathcal{J}_{\text{GRPO}}(\theta) = \frac{1}{N}\sum_{i=1}^{B}\sum_{t=1}^{T_i} \min\Big(\rho_{i,t}(\theta)\,\hat{A}_i,\ \text{clip}\big(\rho_{i,t}(\theta),\,1{-}\epsilon,\,1{+}\epsilon\big)\,\hat{A}_i\Big),$$

where $\rho_{i,t}(\theta) \triangleq \pi_\theta(o_{i,t} \mid x_i, o_{i,<t}) / \pi_{\theta_{\text{old}}}(o_{i,t} \mid x_i, o_{i,<t})$ is the per-token importance ratio, B is the number of responses, $N = \sum_{i=1}^{B} T_i$ is the total token count, and $\beta = 0$ throughout (no KL penalty). Let $c = (x, o_{<t})$ denote context and $\mathcal{V}$ the vocabulary. The next-token distribution is parameterized as $\pi_v \triangleq \pi_\theta(v \mid c) = \exp(z_v) / \sum_{v'} \exp(z_{v'})$, with softmax Jacobian $\partial \pi_{v'} / \partial z_v = \pi_{v'}(\delta_{v'v} - \pi_v)$.

Token Surprisal, Entropy, and Logit updates. The *token surprisal* is $\mathfrak{s}_v \triangleq -\ln \pi_v$ and the position-level *policy entropy* is $H \triangleq -\sum_v \pi_v \ln \pi_v = \mathbb{E}_\pi[\mathfrak{s}]$[27–31]; the batch mean $\bar{H} \triangleq N^{-1} \sum_{i,t} H_{i,t}$ typically decreases during RL fine-tuning (*entropy collapse*). In the unclipped regime, the GRPO gradient at a position where token a was sampled yields the logit update $\Delta z_v = \eta \hat{A} (\delta_{va} - \pi_v)$ for all $v \in \mathcal{V}$, where $\eta > 0$ is an infinitesimal step size (Appendix D.2).

Lemma 2.1 (Entropy gradient w.r.t. logits: surprisal-deviation form)**.** *For any $v \in \mathcal{V}$, $\frac{\partial H}{\partial z_v} = \pi_v\big(\mathfrak{s}_v - H\big)$. Appendix D.3 provides the derivation.*

Raising z_v increases entropy when token v is rarer than average ($\mathfrak{s}_v > H$) and decreases it otherwise. It relates entropy to individual logits and underlies the token-level entropy dynamics derived below.

3 Theoretical Analysis

We develop a theoretical framework for analyzing entropy evolution during GRPO training, proceeding from token-level gradient analysis (Section 3.1) through an advantage–surprisal decomposition (Section 3.2) and a batch-level near-criticality result (Section 3.3) to cross-step feedback dynamics (Section 3.4). Complete proofs are in Appendices D–F.

3.1 First-Order Gradient Analysis of Token-Level Policy Entropy

Consider the next-token distribution $\pi(\cdot \mid c)$ with entropy H. Let a denote the sampled token, with probability $p = \pi(a \mid c)$ and surprisal $\mathfrak{s}_a = -\ln p$. Define $S_2 \triangleq \sum_{v \in \mathcal{V}} \pi_v^2\big(\ln \pi_v + H\big)$ and $\Phi(p) \triangleq p\big(\ln p + H\big) - S_2$. We call Φ the **entropy sensitivity function**: it measures the signed excess of the sampled token's probability-weighted surprisal deviation over the distributional baseline S_2.

Theorem 3.1 (Token-level entropy variation)**.** *In the unclipped regime of GRPO, let $\hat{A}$ denote the advantage at this position, and let η be the step size along the GRPO policy-gradient direction. Then $\frac{dH}{d\eta}\Big|_{\eta=0} = -\hat{A}\,\Phi(p)$.*

The proof (Appendix D.4) follows by taking the inner product of $\partial H / \partial z_v$ (Lemma 2.1) with the GRPO logit update. The result decomposes the instantaneous entropy effect into the advantage $\hat{A}$ and $\Phi(p)$, governing whether probability redistribution concentrates or disperses the distribution. And their product determines the sign and magnitude of entropy variation at each position.

3.2 Advantage–Surprisal Four-Quadrant Decomposition

Determining $\operatorname{sign}(dH/d\eta)$ reduces to analyzing $\operatorname{sign}(\Phi(p))$. Two properties, proved in Appendix E, underpin the analysis: $S_2 > 0$ for any non-uniform distribution (Lemma E.1), and $H > S_2$ for any non-degenerate distribution (Lemma E.2). These imply $\Phi(0^+) = -S_2 < 0$ and $\Phi(1) = H - S_2 > 0$. Since $\Phi'(p) = \ln p + H + 1$, the function is strictly increasing on $(e^{-(H+1)}, 1]$, yielding:

Proposition 3.2 (Uniqueness of the Critical Surprisal Threshold)**.** *For any non-uniform, non-degenerate π, there exists a unique $p^* \in (e^{-H}, 1)$ with $\mathfrak{s}^* \triangleq -\ln p^* \in (0, H)$ such that $\Phi(p^*) = 0$ and when $\Phi(p) > 0 \iff p > p^* \iff \mathfrak{s}_a < \mathfrak{s}^*$. The proof is in Appendix E.3.*

The threshold $\mathfrak{s}^*$ partitions the vocabulary into a **low-surprisal region** ($\mathfrak{s}_a < \mathfrak{s}^*$) and a **high-surprisal region** ($\mathfrak{s}_a > \mathfrak{s}^*$). Substituting into Theorem 3.1 yields the four-quadrant structure.

Corollary 3.3 (Four-Quadrant Decomposition)**.** *The sign of $dH/d\eta$ is determined by $(\operatorname{sign} \hat{A},\ \mathbf{1}[\mathfrak{s}_a < \mathfrak{s}^*])$: (i) reinforcing low-surprisal tokens ($\hat{A} > 0, \mathfrak{s}_a < \mathfrak{s}^*$) reduces entropy; (ii) reinforcing high-surprisal tokens ($\hat{A} > 0, \mathfrak{s}_a > \mathfrak{s}^*$) increases it; (iii) suppressing low-surprisal tokens ($\hat{A} < 0, \mathfrak{s}_a < \mathfrak{s}^*$) increases it; (iv) suppressing high-surprisal tokens ($\hat{A} < 0, \mathfrak{s}_a > \mathfrak{s}^*$) reduces it. Each GRPO step thus propagates four token-level entropy signals with opposing signs. The proof is in Appendix E.4.*

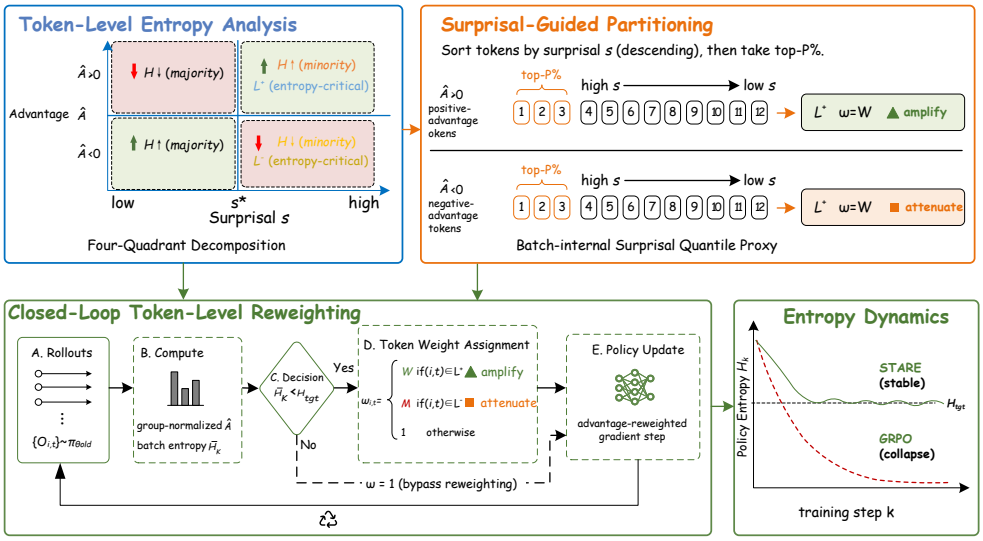


Figure 2: Overview of STARE. Guided by a four-quadrant decomposition of token-level entropy dynamics (top-left) and a batch-internal surprisal-quantile proxy that identifies entropy-critical tokens (top-right), STARE applies target-entropy-gated advantage reweighting in GRPO (bottom-left), stabilizing policy entropy where vanilla GRPO collapses (bottom-right).

Asymmetric entropy contributions. Since rollouts are sampled from π_θ, low-surprisal tokens are drawn more frequently than high-surprisal ones at each decoding position. Within positive-advantage trajectories, entropy-decreasing tokens (low surprisal, $\Phi > 0$) therefore constitute the statistical majority. Because GRPO assigns a single trajectory-level $\hat{A}_i$ to all tokens, it cannot distinguish these opposing entropy effects: the reinforced low-surprisal majority systematically drives the distribution toward concentration, while the high-surprisal minority that could preserve diversity contributes limited entropy-increasing effects. An analogous asymmetry governs the negative-advantage subset, revealing a fundamental gradient-level mechanism underlying entropy collapse in GRPO.

3.3 Batch-Level Entropy Decomposition and Near-Criticality

The preceding analysis establishes that the entropy-contribution asymmetry persists within both advantage subsets. A natural quantitative question arises: what reweighting of the entropy-increasing minority suffices to reverse the batch-level net entropy gradient?

Theorem 3.4 (Entropy Neutrality Identity)**.** *For any conditional distribution* π, $\mathbb{E}_{a\sim\pi}\big[\Phi(a)\big] = \sum_{v\in\mathcal{V}} \pi_v\, \Phi(\pi_v) = 0$. *The proof is in Appendix E.6.*

Token-level advantage reweighting. Define $\mathcal{L}^+ = \{(i,t) : \hat{A}_i > 0,\ \mathfrak{s}_{i,t} > \mathfrak{s}^*_{i,t}\}$. Scaling the effective advantage of each token in $\mathcal{L}^+$ by a multiplicative factor $W \geq 1$, while retaining unit weight at all remaining positions, yields:

Proposition 3.5 (Entropy Gradient under Token-Level Reweighting)**.** $\frac{d\bar{H}}{d\eta}\Big|_W = -\frac{1}{N}\big[\Lambda - (W-1)\Gamma\big]$, *where* $\Lambda \triangleq \sum_{i,t} \hat{A}_i\, \Phi_{i,t}$, $\Gamma \triangleq \sum_{(i,t)\in\mathcal{L}^+} \hat{A}_i\, |\Phi_{i,t}| > 0$, *and the critical weight is* $W^* = 1 + \Lambda/\Gamma$.

The proof is in Appendix E.7. Two structural properties jointly ensure that W^* remains near unity (**near-critical regime**). First, the entropy neutrality identity guarantees that Λ is a *weak residual* arising solely from the statistical dependence between trajectory-level advantages and token-level entropy sensitivities; under the assumptions formalized in Appendix E.9, $|\Lambda|/\Sigma_{\text{abs}} = O(T^{-1})$ where $\Sigma_{\text{abs}} = \sum_{i,t} |\hat{A}_i|\,|\Phi_{i,t}|$. Second, *high-surprisal tokens carry amplified entropy sensitivity* (Appendix E.8): $\mathfrak{s} \geq H \Rightarrow |\Phi(p)| \geq S_2$, whereas $\mathfrak{s} < \mathfrak{s}^* \Rightarrow |\Phi(p)| \leq H - S_2$. Although $\mathcal{L}^+$ is a sampling minority, the amplified per-token $|\Phi|$ values ensure that Γ remains appreciable.

Corollary 3.6 (Near-Criticality)**.** *When the sequence length T and the batch size are both sufficiently large, $W^* - 1 = \Lambda/\Gamma = O(T^{-1})$ (Appendix E.11). So beyond the critical threshold, the specific value of W principally controls the magnitude rather than the sign of the per-step entropy shift.*

3.4 Cross-Step Entropy Dynamics

Let $\Delta\bar{H}_k = -N^{-1}[\Lambda_k - (W-1)\,\Gamma_k]$ denote the batch-level entropy variation at step k. Under $W = 1$, the sampling asymmetry implies $\Lambda_k > 0$ in expectation, so $\Delta\bar{H}_k < 0$. The resulting entropy reduction further concentrates π_θ, lowering the sampling frequency of high-surprisal tokens in subsequent batches, shrinking Γ_{k+1}, and raising W^*_{k+1}, forming a **self-reinforcing loop of entropy collapse**. Conversely, when $W > W^*_k$, entropy increases disperse the distribution, enlarge Γ_{k+1}, and lower W^*_{k+1}, forming a symmetric loop of **entropy recovery** (formalized in Appendix F). The near-criticality result (corollary 3.6) establishes that a modest token-level weight adjustment suffices to alter the macroscopic entropy trajectory. This theoretical insight directly motivates the algorithmic design proposed in the next section: by selectively modulating the effective advantage weights of a targeted subset of tokens within the GRPO policy gradient, one can restore a sustainable dynamic equilibrium between the entropy-increasing and entropy-decreasing gradient forces.

4 Method

The theoretical analysis in Section 3 reveals that GRPO's shared trajectory-level advantages induce a token-level credit assignment mismatch: high-frequency low-surprisal tokens dominate gradient aggregation while sparse high-surprisal tokens with critical entropy effects are under-represented. Motivated by it, we propose **S**urprisal-Guided **T**oken-Level **A**dvantage **R**eweighting For Policy **E**ntropy Stability **(STARE)**: a surprisal-based credit rebalancing mechanism that operates within the clipped surrogate of GRPO, assigns differentiated weights to entropy-critical tokens, and incorporates target-entropy closed-loop feedback for stable entropy regulation on training, as shown in Figure 2.

4.1 Entropy-Critical Token Partitioning via High-Surprisal Quantiles

Let $\mathcal{T}^+ = \{(i,t) : \hat{A}_i > 0\}$ and $\mathcal{T}^- = \{(i,t) : \hat{A}_i < 0\}$ denote the positive- and negative-advantage token sets. Computing the exact critical threshold $\mathfrak{s}^*_{i,t}$ (Proposition 3.2) at every position requires the full conditional distribution, incurring prohibitive overhead. Instead, STARE employs a simple, stable, and theoretically motivated **batch-internal surprisal-quantile proxy**. Concretely, within $\mathcal{T}^+$ and $\mathcal{T}^-$ separately, tokens are ranked in descending order of surprisal $\mathfrak{s}_{i,t} = -\ln \pi_\theta(o_{i,t} \mid x_i, o_{i,<t})$, and the top $P\%$ tokens are selected to form the entropy-critical sets below:

$$\mathcal{L}^{\pm} = \big\{(i,t) \in \mathcal{T}^{\pm} : \mathfrak{s}_{i,t} \geq Q_P\big(\{\mathfrak{s}_{j,s}\}_{(j,s)\in\mathcal{T}^{\pm}}\big)\big\}.$$

By Corollary 3.3, $\mathcal{L}^+$ approximately denotes the entropy-increasing tokens among positive-advantage responses, and $\mathcal{L}^-$ indicates the entropy-decreasing tokens in negative-advantage responses. The fixed proportion P directly controls the intervention scale, obviating per-position threshold computation.

4.2 Advantage-Conditioned Token-Level Credit Rebalancing

We augment the GRPO objective with positive token-level weights $\omega_{i,t} > 0$:

$$\mathcal{J}_{\mathrm{STARE}}(\theta) = \frac{1}{N}\sum_{i,t} \omega_{i,t}\ \min\Big(\rho_{i,t}(\theta)\,\hat{A}_i,\ \mathrm{clip}\big(\rho_{i,t}(\theta),\,1-\epsilon,\,1+\epsilon\big)\,\hat{A}_i\Big).$$

Setting $\omega_{i,t} \equiv 1$ recovers STARE to standard GRPO. Because $\omega_{i,t} > 0$, STARE preserves all token-level gradient directions: tokens with positive advantage remain reinforced, while those with negative advantage remain suppressed. STARE therefore acts as an advantage-conditioned credit-rebalancing mechanism, selectively rescaling relative magnitudes along the surprisal dimension.

Variant I: One-Sided Entropy Amplification ($\hat{A}_i > 0$, High-Surprisal tokens, denoted as O1).

$$\omega^{(\mathrm{V1})}_{i,t} = \begin{cases} W, & (i,t) \in \mathcal{L}^+, \\ 1, & \text{otherwise}, \end{cases} \qquad W > 1.$$

Tokens in $\mathcal{L}^+$ simultaneously carry positive advantage and entropy-increasing effect; amplifying their weights directly strengthens the minority that GRPO systematically underweights. The resulting batch-level net entropy shift is $\Lambda_{\mathrm{V1}} = \Lambda - (W-1)\,\Gamma^+$, where $\Gamma^+ = \sum_{(i,t)\in\mathcal{L}^+} |\hat{A}_i\,\Phi_{i,t}| > 0$. And under the near-criticality condition (Corollary 3.6), a moderate $W > 1$ usually suffices to reverse the sign of the batch net entropy shift. We provide the Algorithmic 1 pseudocode in the Appendix C.

Variant II: Dual-Sided Entropy Regulation. Extending Variant I, this variant additionally attenuates the weights of tokens in $\mathcal{L}^-$ ($\hat{A}_i < 0$, High-Surprisal tokens, denoted as C2):

$$\omega_{i,t}^{(\mathrm{V2})} = \begin{cases} W, & (i,t) \in \mathcal{L}^+, \\ M, & (i,t) \in \mathcal{L}^-, \\ 1, & \text{otherwise}, \end{cases} \qquad W > 1,\ 0 < M < 1.$$

Tokens in $\mathcal{L}^-$ reside in the high-surprisal tail of negative-advantage responses; and large negative-advantage updates on these tokens redistribute mass from the tail toward the peak, exacerbating concentration. Attenuating their weights alleviates this entropy-decreasing pressure. Letting $\Gamma^- = \sum_{(i,t)\in\mathcal{L}^-} |\hat{A}_i\,\Phi_{i,t}|$, the batch net entropy shift becomes $\Lambda_{\mathrm{V2}} = \Lambda - (W-1)\,\Gamma^+ - (1-M)\,\Gamma^-$.

Two-sided regulation simultaneously amplifies the entropy-increasing signal and attenuates the entropy-decreasing signal to adjust policy entropy. Definitions and analysis for all four single-polarity (O1,O2,O3,O4) and four combined operations (C1,C2,C3,C4) are deferred to Appendices G and H.

4.3 Closed-Loop Regulation via Target-Entropy Gating

A purely open-loop reweighting strategy risks overshooting from entropy collapse into uncontrolled divergence. To achieve stable regulation, we introduce a batch-level target entropy H_{tgt} and employ the current batch mean entropy $\bar{H}_k$ as a closed-loop feedback signal via a binary gate $g_k = \mathbf{1}[\bar{H}_k < H_{\mathrm{tgt}}]$ and express the weights in unified form below. When $\bar{H}_k < H_{\mathrm{tgt}}$, the gate activates and

$$\omega_{i,t} = \begin{cases} 1 + g_k\,(W-1), & (i,t) \in \mathcal{L}^+, \\ 1 - g_k\,(1-M), & (i,t) \in \mathcal{L}^- \text{ (two-sided only)}, \\ 1, & \text{otherwise}. \end{cases}$$

STARE strengthens entropy-increasing signals; when $\bar{H}_k \geq H_{\mathrm{tgt}}$, all weights revert to unity, automatically recovering standard GRPO. This drives entropy toward H_{tgt} via bounded oscillation. Finer-grained sample-level and token-level closed-loop variants are presented in Appendix G.6.

4.4 Static and Adaptive Weighting Schedules

Fixed weights. Near-criticality (Corollary 3.6) implies that the required reweight perturbation is typically modest: beyond the critical point, the specific value principally controls the magnitude rather than the direction of the per-step entropy shift, reducing sensitivity to hyperparameter choices. Fixed W and M therefore suffice in most settings.

Adaptive weights. As training progresses, intervention strength may vary across phases with distinct distributional. STARE also supports adaptive weight updates driven by the target-entropy signal:

$$W_{k+1} = \mathrm{clip}\Big(W_k + \alpha\ \mathrm{sgn}\left(H_{\mathrm{tgt}} - \bar{H}_k\right),\ [1,\ W_{\max}]\Big),$$

$$M_{k+1} = \mathrm{clip}\Big(M_k - \alpha\ \mathrm{sgn}\left(H_{\mathrm{tgt}} - \bar{H}_k\right),\ [M_{\min},\ 1]\Big).$$

When $\bar{H}_k < H_{\mathrm{tgt}}$, W increases and M decreases, intensifying the intervention; otherwise both relax toward GRPO. The constraints $W \geq 1$ and $M \leq 1$ ensure graceful degradation. Setting $\alpha = 0$ recovers the fixed-weight. Default values α=0.01, $W_{\max}$=1.5, $M_{\min}$=0.5 yield robust performance.

Default configuration. All main experiments adopt Variant I (O1) with batch-level target-entropy gating and fixed weights. This minimal configuration suffices to stabilize entropy and improve performance. Ablations on two-sided regulation and adaptive weights are provided in Appendix H.

5 Experiments

5.1 Experimental Setup

Models and scenarios. We evaluate STARE across three scenarios. In the Short CoT scenario, we use Qwen2.5-Math-7B-Base with a maximum decoding length of 4k, Qwen2.5-14B-Instruct with 8k, and Qwen2.5-32B-Base with 8k[33, 44]. In the Long CoT scenario, we employ DeepSeek-R1-Distill-Qwen-1.5B and Qwen3-8B-Base with 16k to elicit deep reasoning and self-reflection[1, 4]. In the tool-use, we first perform cold-start SFT on Qwen2.5-7B-Base using Retool 2K data[45], then conduct RL training with 8k response length.

Training. We use a learning rate of 1×10^{-6} and a batch size of 64 samples with 8 rollouts per sample, and on-policy updates with a single gradient step per batch. Decoding adopts Top-$p = 1.0$ and temperature $T = 1.0$. The STARE hyperparameters are set to $W = 1.1$, $M = 0.9$, $H_{\text{tgt}} = 0.3$, and $P\% = 10\%$, with batch-level gating and fixed weights as the default configuration. The training corpus consists of 100k samples deduplicated and sampled from open-source RL datasets including DeepScaler, Skywork-o1, Polaris, and DAPO[10, 24, 46, 47]. We construct an enhanced GRPO baseline, denoted GRPO-ds, which removes the KL penalty and incorporates dynamic sampling with token-level loss[10]. We report two variants: STARE-O1 amplifies only $\mathcal{L}_q^+$, whereas STARE-C2 additionally attenuates $\mathcal{L}_q^-$ over O1. Ablation studies adopt STARE-O1 as the default configuration.

Evaluation. We evaluate on six mathematical benchmarks: AIME24, AIME25, AMC23, MATH-500, Minerva Math, and OlympiadBench[48–51], with Top-$p = 0.95$ and $T = 0.7$ at inference. AIME24/25 and AMC23 are evaluated 32 times and the other benchmarks 4 times, and all results report average accuracy.

Table 1: Performance comparison of STARE and competitive RL algorithms on six math benchmarks across 1.5B–32B scales and three scenarios (Acc avg@N); † denotes results from prior works.

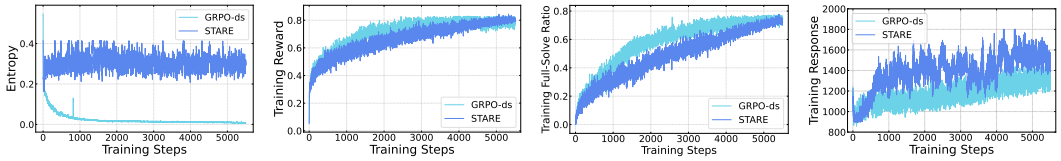
Method	AIME24	AIME25	AMC	MATH	Minerva	Olympiad	Avg.
In the Short CoT Scenario							
RL from the Qwen2.5-Math-7B-Base							
W-REINF† [14]	31.2	10.1	58.1	76.2	34.2	38.9	41.6
GRPO† [6]	34.0	9.3	58.6	79.9	38.1	42.8	43.8
EntroReg† [21]	34.6	12.6	61.1	82.7	43.1	42.8	46.2
DAPO† [10]	35.7	17.1	61.0	82.6	42.8	43.9	47.2
80/20 Rule† [16]	35.8	15.1	63.7	84.1	43.0	45.5	47.9
EntroAdv† [21]	36.8	16.3	63.8	83.5	42.7	44.5	47.9
STEER† [32]	36.9	16.2	72.2	82.4	41.7	43.3	49.1
GRPO-ds	37.1	17.7	75.3	82.7	39.4	42.6	49.1
KL-Cov† [22]	38.9	13.8	59.2	81.5	40.9	44.2	46.4
EAPO† [20]	39.8	17.2	62.1	83.7	43.9	45.1	48.6
STARE-O1	44.2	23.8	83.4	86.1	44.2	44.7	54.4
STARE-C2	42.9	24.2	84.1	85.8	44.7	45.3	54.5
RL from the Qwen2.5-14B-Instruct							
Base-14B† [33]	12.1	11.7	-	-	-	-	-
GRPO† [6]	22.5	17.6	-	-	-	-	-
GRPO-Clip$_\nu$† [17]	23.4	21.4	-	-	-	-	-
GRPO-ds	24.2	21.9	69.1	80.6	37.2	43.4	46.1
STARE-O1	30.8	27.1	77.5	85.4	40.5	50.9	52.0
STARE-C2	31.5	28.3	76.3	86.1	40.2	51.4	52.3
RL from the Qwen2.5-32B-Base							
GRPO† [6]	28.5	22.5	-	86.6	44.9	60.3	-
80/20 Rule† [16]	32.5	28.5	-	89.4	45.6	57.6	-
GSPO† [34]	33.3	22.3	-	87.6	48.5	55.6	-
Lp-Reg† [26]	38.1	27.1	-	90.0	46.32	61.2	-
DAPO† [10]	38.3	29.8	-	87.6	48.5	55.6	-
GRPO-ds	38.5	28.8	85.3	85.6	44.6	54.0	56.1
STARE-O1	43.3	34.1	87.3	90.4	48.8	60.1	60.7
STARE-C2	42.9	35.7	88.8	90.6	49.3	60.9	61.4
In the Long CoT Scenario							
RL from the DeepSeek-R1-Distill-Qwen-1.5B							
CE-GPPO† [35]	29.6	23.5	73.5	76.3	26.6	43.9	45.6
Base-1.5B† [1]	32.5	24.3	69.4	85.7	37.5	55.2	50.7
EntroReg† [21]	34.6	26.4	70.2	86.8	37.6	56.2	51.9
CISPO† [36]	34.8	25.8	76.9	76.8	26.5	45.8	48.4
W-REINF† [14]	35.0	25.0	69.8	87.5	37.8	55.9	51.8
CE-GPPO† [35]	35.1	27.7	82.5	76.7	27.8	45.6	49.2
ASPO† [37]	36.4	28.3	83.1	74.6	26.0	44.9	48.9
GRPO† [6]	38.1	27.1	71.4	88.9	39.4	58.7	53.9
DAPO† [10]	40.9	28.6	73.6	89.9	39.1	58.4	55.1
DGPO† [13]	43.3	32.8	86.0	77.9	28.2	48.0	52.7
KL-Cov† [22]	43.9	30.1	75.0	90.0	40.5	59.9	56.5
EntroAdv† [21]	44.0	30.4	73.9	90.2	40.4	59.1	56.3
EAPO† [20]	45.1	30.1	75.5	91.1	39.7	60.8	57.0
GRPO-ds	50.4	37.4	85.9	88.3	49.2	64.1	62.5
JustRL† [38]	52.6	38.8	91.0	91.7	51.5	68.0	65.6
STARE-O1	53.8	41.5	89.5	90.4	52.6	67.8	65.9
STARE-C2	53.1	40.5	91.3	92.1	51.8	68.8	66.3
RL from the Qwen3-8B-Base							
GRPO† [6]	31.3	24.7	75.2	88.9	55.9	61.5	56.2
80/20 Rule† [16]	31.3	27.5	79.9	89.9	54.8	62.5	57.6
STAPO† [39]	33.4	28.7	79.9	90.4	57.2	63.0	58.8
DAPO† [10]	34.2	26.1	-	84.5	-	-	-
Lp-Reg† [26]	35.9	25.8	-	87.4	-	-	-
A3PO† [15]	37.8	30.4	-	91.3	-	-	-
GRPO-ds	39.5	30.8	80.8	88.6	52.3	59.0	58.5
STARE-O1	43.9	34.7	85.3	90.6	55.6	61.8	62.0
STARE-C2	44.3	32.6	86.1	91.2	56.9	62.3	62.2
In the Tool-Use Agent Scenario							
RL from the Qwen2.5-7B-Base							
Base-7B-TIR† [33]	1.7	0.6	10.8	18.0	-	6.2	-
ToRL† [40]	40.2	27.9	75.0	82.2	-	49.9	-
Effective TIR† [41]	42.3	29.2	74.2	86.4	-	-	-
ZeroTIR† [42]	46.7	30.0	-	85.2		-	-
GRPO-ds	46.8	32.4	75.9	81.4	38.3	48.8	53.9
SimpleTIR† [43]	50.5	30.9	79.1	88.4	-	54.8	-
STARE-O1	53.2	37.5	84.9	86.8	41.9	52.3	59.4
STARE-C2	52.8	38.1	86.9	87.2	43.7	53.6	60.4

5.2 Main Results

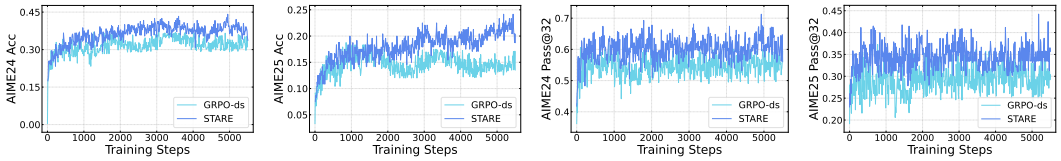
Table 1 presents the full performance comparison, where results marked with † are cited from prior work such as EAPO, STEER, GRPO-Clip$_\nu$, Lp-Reg, DGPO, JustRL, STAPO, A3PO, and SimpleTIR[13, 15, 17, 20, 26, 32, 38, 39, 43]. STARE consistently delivers substantial gains across the six math benchmarks at scales ranging from 1.5B to 32B and three scenarios.

Short-CoT Scenario. At the 7B scale, STARE-O1 attains an average accuracy of 54.4%, outperforming STEER (49.1%, +5.3%) and GRPO-ds (49.1%, +5.3%), while reaching 44.2% and 23.8% on AIME24 and AIME25, corresponding to improvements of roughly 10% and 7% over DAPO. At the 14B scale, STARE-O1 (52.0%) surpasses GRPO-ds (46.1%) by 5.9%. At the 32B scale, STARE-O1 (60.7%) exceeds GRPO-ds (56.1%) by 4.6%. **Tool-Use Scenario.** STARE-O1 achieves an average of 59.4%, improving over GRPO-ds (53.9%) by 5.5% and outperforms SimpleTIR, while reaching 53.2% and 37.5% on AIME24 and AIME25. STARE-C2 further lifts the average to 60.4%.

Long-CoT Scenario. At the 1.5B scale, STARE-O1 reaches an average of 65.9%, substantially exceeding EAPO (57.0%, +8.9%) and DAPO (55.1%, +10.8%), while also outperforming JustRL on AIME24 and AIME25. At the 8B scale, STARE-O1 (62.0%) surpasses STAPO (58.8%) and GRPO-ds (59.0%), and STARE-C2 further raises the average to 62.2%.

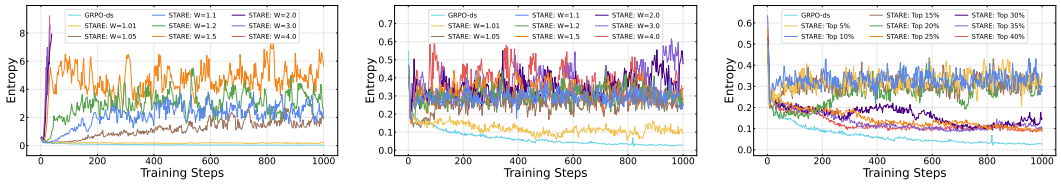


(a) Training Entropy (b) Training Reward (c) Train Full-solve Ratio (d) Train Response Length



(e) AIME24 Acc (f) AIME25 Acc (g) AIME24 Pass@32 (h) AIME25 Pass@32

Figure 3: Comparison of key training metrics between STARE and GRPO-ds on Qwen2.5-Math-7B-Base model in the Short CoT scenario over 5k RL steps.



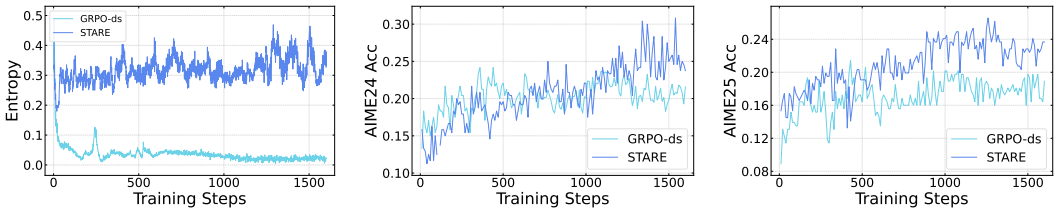
(a) Entropy evolution under varying W without target-entropy gating (b) Entropy evolution under varying W with target-entropy gating (c) Entropy evolution under varying Top-P ratio (W : 1.1, H_{tgt} : 0.3)

Figure 4: STARE policy entropy evolution under ablation on the varying reweighting factor W, target-entropy gating, and the high-surprisal selection ratio P on Qwen2.5-Math-7B-Base.

Key findings. Three principal observations emerge. **(i)** On the challenging AIME24 and AIME25 benchmarks, STARE improves over GRPO-ds by 4%–8% in average accuracy, with gains of 3%–6% across all six benchmarks. **(ii)** Across different thinking scenarios and model scales from 1.5B to 32B, STARE consistently outperforms the majority of competitive RL improvement methods, confirming its effectiveness and robustness. **(iii)** The additional gains of STARE-C2 further indicate that dual-sided regulation, which simultaneously strengthens entropy-increasing signals and attenuates entropy-decreasing ones, can yield an even more favorable exploration-exploitation balance.

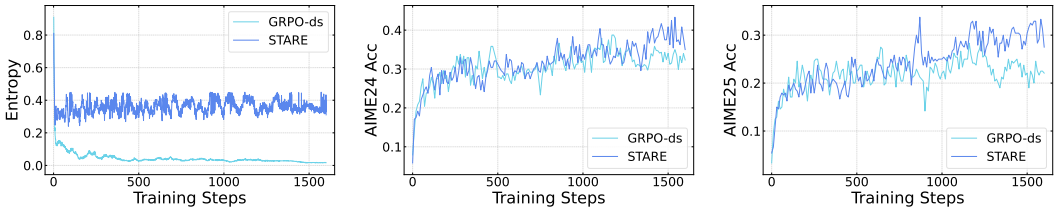
5.3 Cognitive Analysis

STARE vs. GRPO-ds: Training Dynamics across Scales and Scenarios. To validate STARE in long-horizon RL, we run 5000 training steps on Qwen2.5-Math-7B-Base under the Short CoT scenario. Figure 3 compares STARE with GRPO-ds on key metrics, and Figures 1, 5, 6, 7, 8 further verify its effectiveness across 1.5B–32B scales and scenarios. **Entropy stability and performance evolution.** GRPO-ds exhibits entropy collapse over steps 0–1000, with policy entropy approaching zero (Figure 3(a)), consistent with Section 3; correspondingly, its AIME24/25 accuracy peaks around step 1000 and saturates thereafter (Figure 3(e)-(f)), indicating premature convergence. In contrast, STARE stabilizes entropy near $H_{\text{tgt}} = 0.3$ via token-level reweighting and closed-loop gating, with accuracy continuing to rise beyond step 1000 and peaking at 5000, thereby unlocking long-horizon RL potential. **Exploration-exploitation balance.** STARE's Pass@32 consistently exceeds GRPO-ds throughout training (Figure 3(g)-(h)), preserving output diversity and mitigating mode-seeking; whereas GRPO-ds's reward and Full-Solve Ratio plateau early (Figure 3(b)-(c)), STARE keeps growing, and its sustained response-length increase (Figure 3(d)) reflects deeper reasoning. **Cross-scale and cross-scenario generalization.** Figures 5-1 reveal a consistent pattern across Short CoT (14B, 32B), Long CoT (R1-Distill-Qwen-1.5B, Qwen3-8B-Base), and Tool-Use (7B): GRPO-ds suffers collapse with performance saturation, while STARE maintains stable entropy and continuous accuracy gains. Notably, on DeepSeek-R1-Distill-Qwen-1.5B, GRPO-ds's entropy decays below 0.2 by step 5000, whereas STARE rapidly restores entropy to the target band starting around step 3500,



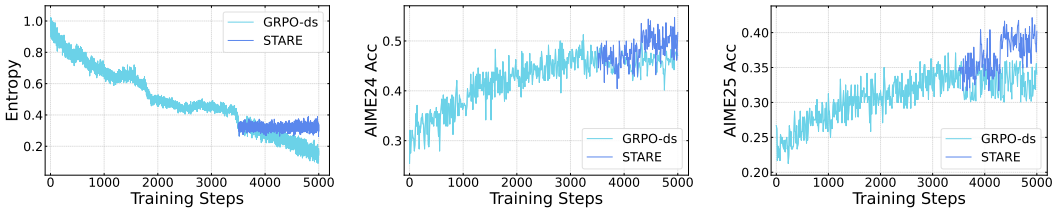
(a) Training Entropy (b) AIME24 Acc (c) AIME25 Acc

Figure 5: Training dynamics of STARE vs. GRPO-ds on Qwen2.5-14B-Instruct in the Short CoT scenario: policy entropy, AIME24 accuracy, and AIME25 accuracy.



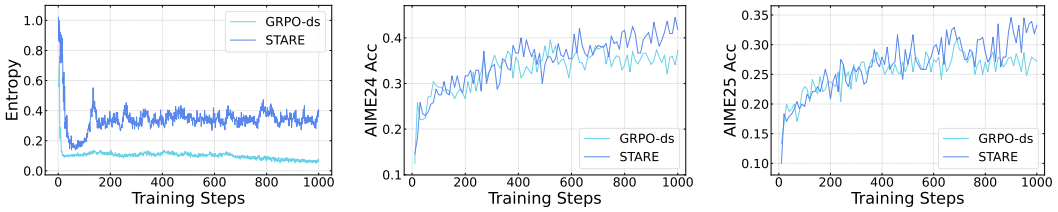
(a) Training Entropy (b) AIME24 Acc (c) AIME25 Acc

Figure 6: Training dynamics of STARE vs. GRPO-ds on Qwen2.5-32B-Base in the Short CoT scenario: policy entropy, AIME24 accuracy, and AIME25 accuracy.



(a) Training Entropy (b) AIME24 Acc (c) AIME25 Acc

Figure 7: Training dynamics of STARE vs. GRPO-ds on DeepSeek-R1-Distill-Qwen-1.5B in the Long CoT scenario: policy entropy, AIME24 accuracy, and AIME25 accuracy.



(a) Training Entropy (b) AIME24 Acc (c) AIME25 Acc

Figure 8: Training dynamics of STARE vs. GRPO-ds on Qwen3-8B-Base in the Long-CoT scenario: policy entropy, AIME24 accuracy, and AIME25 accuracy.

with AIME24/25 accuracy improving in tandem (Figure 7), showing its intervention-recovery capability and robustness across scales and scenarios. Further details are provided in Appendix B.1

Ablation on Key Hyperparameters and Target-Entropy Gating. We ablate W, P, and the target-entropy gate on Qwen2.5-Math-7B-Base (Figure 4). Under open-loop reweighting (without the target-entropy gate; Figure 4(a)), $W = 1.01$ already mitigates the entropy decay of GRPO-ds, $W \geq 1.05$ yields steady growth, and $W \geq 2.0$ triggers divergenc, corroborating the near-criticality property (Corollary 3.6) that beyond the critical threshold W controls magnitude rather than direction. Open-loop reweighting, however, stabilizes entropy at an excessively high level, inducing over-exploration that hampers overall training(Appendix B.5). With the closed-loop gate ($H_{\text{tgt}} = 0.3$),

Figure 4(b) shows that all $W \in [1.05, 1.5]$ steer entropy into the target band with bounded oscillation, confirming closed-loop stability and substantially reducing sensitivity to W. Fixing $W = 1.1$ and $H_{\text{tgt}} = 0.3$, Figure 4 (c) further shows that entropy stays within the target band for $P \in [5\%, 20\%]$ and remains confined to $[0.1, 0.2]$ even at $P = 40\%$, effectively preventing collapse and indicating a broad operating range. Overall, STARE is robust to W and P, and the target-entropy gate stably constrains the policy entropy. More detailed entropy evolution and performance comparisons are reported in Appendix B.5 (Table 3, Figure 12), with additional analysis provided in Appendix B.4.

Validation of the High-Surprisal Quantile Proxy. Figure 10 validates the batch-internal top-$P\%$ surprisal-quantile proxy. Panel (a) shows that the fraction of selected tokens falling within the theoretical entropy-increasing region ($\mathfrak{s}_{i,t} > \mathfrak{s}^*_{i,t}$) rises steadily from $\sim$60% to $\sim$95% during training, indicating strong consistency with the critical threshold $\mathfrak{s}^*$ (Proposition 3.2); panel (b) further confirms that the cumulative net entropy contribution of $\mathcal{L}_q^+$ remains positive and monotonically increasing, consistent with Corollary 3.3. Thus the proxy reliably identifies entropy-critical tokens while obviating per-position solutions of $\Phi(p^*) = 0$; see Appendix B.2 for further details.

Effects of Single-Polarity and Combined Operations in STARE. To validate the token-level reweighting mechanism under the four-quadrant decomposition, we ablate four single-polarity (O1–O4) and four combined (C1–C4) operations, where O1/O3 amplify entropy-increasing signals, O2/O4 attenuate entropy-decreasing ones, and C1–C4 jointly intervene on two quadrants. As shown in Table 2 and Figure 11, while GRPO-ds suffers rapid entropy decay, all eight variants effectively mitigate this collapse and substantially outperform GRPO-ds AIME24/25 across all four entropy-critical quadrants. Among them, O1 (amplifying $\mathcal{L}_q^+$) and C2 (amplifying $\mathcal{L}_q^+$ while attenuating $\mathcal{L}_q^-$) perform best at $44.2\%/23.8\%$ and $42.5\%/24.2\%$, so we adopt STARE-O1 as the default configuration and STARE-C2 as the dual-sided variant. Further details are provided in Appendix B.3.

Table 2: The results of single-polarity and combined STARE operations on AIME24 and AIME25.

Model	AIME24	AIME25
GRPO-ds	37.1	17.7
STARE-O1	44.2	23.8
STARE-O2	40.5	20.3
STARE-O3	39.6	21.6
STARE-O4	42.1	19.9
STARE-C1	43.1	23.5
STARE-C2	42.5	24.2
STARE-C3	39.9	20.8
STARE-C4	41.7	22.6

Emergent Reflection Behaviors. To examine how STARE elicits deep reasoning, we inspect the tokens selected for advantage amplification during RL training. As exemplified on Qwen2.5-32B-Base, the word cloud in Figure 9 shows reweighted tokens concentrating on uncertainty and self-correction vocabulary, such as *should be*, *but*, *instead*, and *verification*, confirming that the batch-internal surprisal-quantile proxy effectively identifies rare forking tokens [16] with exploratory semantics. A complementary count across six reflection categories (Figure 13) further shows STARE markedly surpasses GRPO-ds with the largest margins on reflection and self-correction, jointly demonstrating that STARE activates deep exploration and delivers consistent gains through token-level credit rebalancing. The complete analysis is deferred to Appendix B.6.



Figure 9: Word cloud of tokens selected by STARE for advantage reweighting.

Further ablations are deferred to the appendices: fixed vs. adaptive weighting(B.7), target-entropy threshold(B.8), gating granularity(B.9), off-policy training(B.10), fixed-threshold reweight(B.11).

6 Conclusion

We present STARE, a surprisal-guided token-level advantage reweighting mechanism that mitigates policy entropy collapse in GRPO-style RLVR. A first-order analysis of entropy dynamics reveals an advantage–surprisal four-quadrant structure and a near-criticality property. Guided by this insight, STARE reweights entropy-critical tokens via batch-internal surprisal quantiles, coupled with a target-entropy closed-loop gate for stable, minimally intrusive regulation. Across 1.5B–32B models and three task scenarios, STARE sustains thousands of stable RL steps and outperforms DAPO and other competitive baselines by 4%–8% on AIME24/25, offering a principled foundation for entropy-aware credit assignment in long-horizon RL post-training of LLMs.

Acknowledgments and Disclosure of Funding

This work was supported by the Major Key Project of Pengcheng Laboratory (Grant No. PCL2025A12 and PCL2026A01) and the CIE-Tencent Ph.D. Student Research Incentive Program (Tencent Hunyuan Special Fund).

References

[1] 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.

[2] Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023.

[3] Kimi Team, Angang Du, Bofei Gao, Bowei Xing, Changjiu Jiang, Cheng Chen, Cheng Li, Chenjun Xiao, Chenzhuang Du, Chonghua Liao, Chuning Tang, Congcong Wang, Dehao Zhang, Enming Yuan, Enzhe Lu, Fengxiang Tang, Flood Sung, Guangda Wei, Guokun Lai, Haiqing Guo, Han Zhu, Hao Ding, Hao Hu, Hao Yang, Hao Zhang, Haotian Yao, Haotian Zhao, Haoyu Lu, Haoze Li, Haozhen Yu, Hongcheng Gao, Huabin Zheng, Huan Yuan, Jia Chen, Jianhang Guo, Jianlin Su, Jianzhou Wang, Jie Zhao, Jin Zhang, Jingyuan Liu, Junjie Yan, Junyan Wu, Lidong Shi, Ling Ye, Longhui Yu, Mengnan Dong, Neo Zhang, Ningchen Ma, Qiwei Pan, Qucheng Gong, Shaowei Liu, Shengling Ma, Shupeng Wei, Sihan Cao, Siying Huang, Tao Jiang, Weihao Gao, Weimin Xiong, Weiran He, Weixiao Huang, Wenhao Wu, Wenyang He, Xianghui Wei, Xianqing Jia, Xingzhe Wu, Xinran Xu, Xinxing Zu, Xinyu Zhou, Xuehai Pan, Y. Charles, Yang Li, Yangyang Hu, Yangyang Liu, Yanru Chen, Yejie Wang, Yibo Liu, Yidao Qin, Yifeng Liu, Ying Yang, Yiping Bao, Yulun Du, Yuxin Wu, Yuzhi Wang, Zaida Zhou, Zhaoji Wang, Zhaowei Li, Zhen Zhu, Zheng Zhang, Zhexu Wang, Zhilin Yang, Zhiqi Huang, Zihao Huang, Ziyao Xu, and Zonghan Yang. Kimi k1.5: Scaling reinforcement learning with llms, 2025.

[4] Qwen Team. Qwen3 technical report, 2025.

[5] AI Anthropic. System card: Claude opus 4 & claude sonnet 4. *Claude-4 Model Card*, 2025.

[6] 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, 2024.

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

[8] 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? *CoRR*, abs/2504.13837, 2025.

[9] Sebastian Farquhar, Jannik Kossen, Lorenz Kuhn, and Yarin Gal. Detecting hallucinations in large language models using semantic entropy. *Nat.*, 630(8017):625–630, 2024.

[10] Qiying Yu, Zheng Zhang, Ruofei Zhu, Yufeng Yuan, Xiaochen Zuo, Yu Yue, Tiantian Fan, Gaohong Liu, Lingjun Liu, Xin Liu, Haibin Lin, Zhiqi Lin, Bole Ma, Guangming Sheng, Yuxuan Tong, Chi Zhang, Mofan Zhang, Wang Zhang, Hang Zhu, Jinhua Zhu, Jiaze Chen, Jiangjie Chen, Chengyi Wang, Hongli Yu, Weinan Dai, Yuxuan Song, Xiangpeng Wei, Hao Zhou, Jingjing Liu, Wei-Ying Ma, Ya-Qin Zhang, Lin Yan, Mu Qiao, Yonghui Wu, and Mingxuan Wang. Dapo: An open-source llm reinforcement learning system at scale, 2025.

[11] Kun Chen, Peng Shi, Fanfan Liu, Haibo Qiu, Zhixiong Zeng, Siqi Yang, and Wenji Mao. Flexible entropy control in rlvr with gradient-preserving perspective. *arXiv preprint arXiv:2602.09782*, 2026.

[12] Zhiheng Xi, Xin Guo, Yang Nan, Enyu Zhou, Junrui Shen, Wenxiang Chen, Jiaqi Liu, Jixuan Huang, Zhihao Zhang, Honglin Guo, et al. Bapo: Stabilizing off-policy reinforcement learning for llms via balanced policy optimization with adaptive clipping. *arXiv preprint arXiv:2510.18927*, 2025.

[13] Xiaoliang Fu, Jiaye Lin, Yangyi Fang, Chaowen Hu, Cong Qin, Zekai Shao, Binbin Zheng, Lu Pan, and Ke Zeng. From $\log \pi$ to π: Taming divergence in soft clipping via bilateral decoupled decay of probability gradient weight, 2026.

[14] Xinyu Zhu, Mengzhou Xia, Zhepei Wei, Wei-Lin Chen, Danqi Chen, and Yu Meng. The surprising effectiveness of negative reinforcement in llm reasoning. *ArXiv*, abs/2506.01347, 2025.

[15] Xinyu Tang, Yuliang Zhan, Zhixun Li, Wayne Xin Zhao, Zhenduo Zhang, Zujie Wen, Zhiqiang Zhang, and Jun Zhou. Rethinking sample polarity in reinforcement learning with verifiable rewards. *arXiv preprint arXiv:2512.21625*, 2025.

[16] Shenzhi Wang, Le Yu, Chang Gao, Chujie Zheng, Shixuan Liu, Rui Lu, Kai Dang, Xionghui Chen, Jianxin Yang, Zhenru Zhang, Yuqiong Liu, An Yang, Andrew Zhao, Yang Yue, Shiji Song, Bowen Yu, Gao Huang, and Junyang Lin. Beyond the 80/20 rule: High-entropy minority tokens drive effective reinforcement learning for LLM reasoning. *CoRR*, abs/2506.01939, 2025.

[17] 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.

[18] Zhihe Yang, Xufang Luo, Zilong Wang, Dongqi Han, Zhiyuan He, Dongsheng Li, and Yunjian Xu. Do not let low-probability tokens over-dominate in RL for llms. *CoRR*, abs/2505.12929, 2025.

[19] Andre Wang He, Daniel Fried, and Sean Welleck. Rewarding the unlikely: Lifting grpo beyond distribution sharpening. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 25559–25571, 2025.

[20] Yuhang He, Haodong Wu, Siyi Liu, Hongyu Ge, Hange Zhou, Keyi Wu, Zhuo Zheng, Qihong Lin, Zixin Zhong, and Yongqi Zhang. Rethinking token-level credit assignment in rlvr: A polarity-entropy analysis. *arXiv preprint arXiv:2604.11056*, 2026.

[21] Daixuan Cheng, Shaohan Huang, Xuekai Zhu, Bo Dai, Wayne Xin Zhao, Zhenliang Zhang, and Furu Wei. Reasoning with exploration: An entropy perspective. *CoRR*, abs/2506.14758, 2025.

[22] Ganqu Cui, Yuchen Zhang, Jiacheng Chen, Lifan Yuan, Zhi Wang, Yuxin Zuo, Haozhan Li, Yuchen Fan, Huayu Chen, Weize Chen, Zhiyuan Liu, Hao Peng, Lei Bai, Wanli Ouyang, Yu Cheng, Bowen Zhou, and Ning Ding. The entropy mechanism of reinforcement learning for reasoning language models. *CoRR*, abs/2505.22617, 2025.

[23] Wujiang Xu, Wentian Zhao, Zhenting Wang, Yu-Jhe Li, Can Jin, Mingyu Jin, Kai Mei, Kun Wang, and Dimitris N. Metaxas. EPO: entropy-regularized policy optimization for LLM agents reinforcement learning. *CoRR*, abs/2509.22576, 2025.

[24] Jujie He, Jiacai Liu, Chris Yuhao Liu, Rui Yan, Chaojie Wang, Peng Cheng, Xiaoyu Zhang, Fuxiang Zhang, Jiacheng Xu, Wei Shen, et al. Skywork open reasoner 1 technical report. *arXiv preprint arXiv:2505.22312*, 2025.

[25] Kai Yang, Xin Xu, Yangkun Chen, Weijie Liu, Jiafei Lyu, Zichuan Lin, Deheng Ye, and Saiyong Yang. Entropic: Towards stable long-term training of llms via entropy stabilization with proportional-integral control. *arXiv preprint arXiv:2511.15248*, 2025.

[26] Guanhua Huang, Tingqiang Xu, Mingze Wang, Qi Yi, Xue Gong, Siheng Li, Ruibin Xiong, Kejiao Li, Yuhao Jiang, and Bo Zhou. Low-probability tokens sustain exploration in reinforcement learning with verifiable reward. *arXiv preprint arXiv:2510.03222*, 2025.

[27] Claude Elwood Shannon. A mathematical theory of communication. *The Bell system technical journal*, 27(3):379–423, 1948.

[28] Byung-Doh Oh, Shisen Yue, and William Schuler. Frequency explains the inverse correlation of large language models' size, training data amount, and surprisal's fit to reading times, 2024.

[29] Wenhao Zeng, Yaoning Wang, Chao Hu, Yuling Shi, Chengcheng Wan, Hongyu Zhang, and Xiaodong Gu. Pruning the unsurprising: Efficient llm reasoning via first-token surprisal, 2026.

[30] Byung-Doh Oh and William Schuler. Transformer-based language model surprisal predicts human reading times best with about two billion training tokens. In *Findings of the association for computational linguistics: EMNLP 2023*, pages 1915–1921, 2023.

[31] Nathaniel J Smith and Roger Levy. The effect of word predictability on reading time is logarithmic. *Cognition*, 128(3):302–319, 2013.

[32] 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. *arXiv preprint arXiv:2510.10150*, 2025.

[33] Qwen, :, An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jingren Zhou, Junyang Lin, Kai Dang, Keming Lu, Keqin Bao, Kexin Yang, Le Yu, Mei Li, Mingfeng Xue, Pei Zhang, Qin Zhu, Rui Men, Runji Lin, Tianhao Li, Tianyi Tang, Tingyu Xia, Xingzhang Ren, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yu Wan, Yuqiong Liu, Zeyu Cui, Zhenru Zhang, and Zihan Qiu. Qwen2.5 technical report, 2025.

[34] Chujie Zheng, Shixuan Liu, Mingze Li, Xiong-Hui Chen, Bowen Yu, Chang Gao, Kai Dang, Yuqiong Liu, Rui Men, An Yang, et al. Group sequence policy optimization. *arXiv preprint arXiv:2507.18071*, 2025.

[35] Zhenpeng Su, Leiyu Pan, Minxuan Lv, Yuntao Li, Wenping Hu, Fuzheng Zhang, Kun Gai, and Guorui Zhou. Ce-gppo: Coordinating entropy via gradient-preserving clipping policy optimization in reinforcement learning, 2026.

[36] Aili Chen, Aonian Li, Bangwei Gong, Binyang Jiang, Bo Fei, Bo Yang, Boji Shan, Changqing Yu, Chao Wang, Cheng Zhu, et al. Minimax-m1: Scaling test-time compute efficiently with lightning attention. *arXiv preprint arXiv:2506.13585*, 2025.

[37] Jiakang Wang, Runze Liu, Lei Lin, Wenping Hu, Xiu Li, Fuzheng Zhang, Guorui Zhou, and Kun Gai. Aspo: Asymmetric importance sampling policy optimization, 2025.

[38] Bingxiang He, Zekai Qu, Zeyuan Liu, Yinghao Chen, Yuxin Zuo, Cheng Qian, Kaiyan Zhang, Weize Chen, Chaojun Xiao, Ganqu Cui, Ning Ding, and Zhiyuan Liu. Justrl: Scaling a 1.5b llm with a simple rl recipe, 2025.

[39] Shiqi Liu, Zeyu He, Guojian Zhan, Letian Tao, Zhilong Zheng, Jiang Wu, Yinuo Wang, Yang Guan, Kehua Sheng, Bo Zhang, et al. Stapo: Stabilizing reinforcement learning for llms by silencing rare spurious tokens. *arXiv preprint arXiv:2602.15620*, 2026.

[40] Xuefeng Li, Haoyang Zou, and Pengfei Liu. Torl: Scaling tool-integrated rl, 2025.

[41] Fei Bai, Yingqian Min, Beichen Zhang, Zhipeng Chen, Wayne Xin Zhao, Lei Fang, Zheng Liu, Zhongyuan Wang, and Ji-Rong Wen. Towards effective code-integrated reasoning, 2025.

[42] Xinji Mai, Haotian Xu, Zhong-Zhi Li, Weinong Wang, Jian Hu, Yingying Zhang, Wenqiang Zhang, et al. Agent rl scaling law: Agent rl with spontaneous code execution for mathematical problem solving. *arXiv preprint arXiv:2505.07773*, 2025.

[43] Zhenghai Xue, Longtao Zheng, Qian Liu, Yingru Li, Xiaosen Zheng, Zejun Ma, and Bo An. Simpletir: End-to-end reinforcement learning for multi-turn tool-integrated reasoning. *arXiv preprint arXiv:2509.02479*, 2025.

[44] An Yang, Beichen Zhang, Binyuan Hui, Bofei Gao, Bowen Yu, Chengpeng Li, Dayiheng Liu, Jianhong Tu, Jingren Zhou, Junyang Lin, Keming Lu, Mingfeng Xue, Runji Lin, Tianyu Liu, Xingzhang Ren, and Zhenru Zhang. Qwen2.5-math technical report: Toward mathematical expert model via self-improvement, 2024.

[45] Jiazhan Feng, Shijue Huang, Xingwei Qu, Ge Zhang, Yujia Qin, Baoquan Zhong, Chengquan Jiang, Jinxin Chi, and Wanjun Zhong. Retool: Reinforcement learning for strategic tool use in llms, 2025.

[46] Sijun Tan, Michael Luo, Justin Wong, Colin Cai, Xiaoxiang Shi, William Yuan Tang, Manan Roongta, Tianjun Zhang, Li Erran Li, Raluca Ada Popa, and Ion Stoica. Deepscaler: Effective RL scaling of reasoning models via iterative context lengthening, 2026.

[47] Chenxin An, Zhihui Xie, Xiaonan Li, Lei Li, Jun Zhang, Shansan Gong, Ming Zhong, Jingjing Xu, Xipeng Qiu, Mingxuan Wang, et al. Polaris: A post-training recipe for scaling reinforcement learning on advanced reasoning models, 2025. *URL https://hkunlp. github. io/blog/2025/Polaris*, 2025.

[48] Ivan Moshkov, Darragh Hanley, Ivan Sorokin, Shubham Toshniwal, Christof Henkel, Benedikt Schifferer, Wei Du, and Igor Gitman. Aimo-2 winning solution: Building state-of-the-art mathematical reasoning models with openmathreasoning dataset, 2025.

[49] 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, 2021.

[50] Aitor Lewkowycz, Anders Andreassen, David Dohan, Ethan Dyer, Henryk Michalewski, Vinay V. Ramasesh, Ambrose Slone, Cem Anil, Imanol Schlag, Theo Gutman-Solo, Yuhuai Wu, Behnam Neyshabur, Guy Gur-Ari, and Vedant Misra. Solving quantitative reasoning problems with language models. In *Advances in Neural Information Processing Systems 35: Annual Conference on Neural Information Processing Systems 2022, NeurIPS 2022, New Orleans, LA, USA, November 28 - December 9, 2022*, 2022.

[51] Chaoqun He, Renjie Luo, Yuzhuo Bai, Shengding Hu, Zhen Leng Thai, Junhao Shen, Jinyi Hu, Xu Han, Yujie Huang, Yuxiang Zhang, Jie Liu, Lei Qi, Zhiyuan Liu, and Maosong Sun. Olympiadbench: A challenging benchmark for promoting AGI with olympiad-level bilingual multimodal scientific problems. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2024, Bangkok, Thailand, August 11-16, 2024*, pages 3828–3850, 2024.

[52] Amirhossein Kazemnejad, Milad Aghajohari, Eva Portelance, Alessandro Sordoni, Siva Reddy, Aaron C. Courville, and Nicolas Le Roux. Vineppo: Refining credit assignment in RL training of llms. In *Forty-second International Conference on Machine Learning, ICML 2025, Vancouver, BC, Canada, July 13-19, 2025*, 2025.

[53] Leo Gao, John Schulman, and Jacob Hilton. Scaling laws for reward model overoptimization. In *International Conference on Machine Learning, ICML 2023, 23-29 July 2023, Honolulu, Hawaii, USA*, pages 10835–10866, 2023.

[54] Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35:24824–24837, 2022.

[55] Edward Y. Chang, Yuxuan Tong, Morry Niu, Graham Neubig, and Xiang Yue. Demystifying long chain-of-thought reasoning in llms. *CoRR*, abs/2502.03373, 2025.

[56] Yu Yue, Yufeng Yuan, Qiying Yu, Xiaochen Zuo, Ruofei Zhu, Wenyuan Xu, Jiaze Chen, Chengyi Wang, TianTian Fan, Zhengyin Du, et al. Vapo: Efficient and reliable reinforcement learning for advanced reasoning tasks. *arXiv preprint arXiv:2504.05118*, 2025.

[57] Renren Jin, Pengzhi Gao, Yuqi Ren, Zhuowen Han, Tongxuan Zhang, Wuwei Huang, Wei Liu, Jian Luan, and Deyi Xiong. Revisiting entropy in reinforcement learning for large reasoning models. *arXiv preprint arXiv:2511.05993*, 2025.

[58] Xingyu Chen, Jiahao Xu, Tian Liang, Zhiwei He, Jianhui Pang, Dian Yu, Linfeng Song, Qiuzhi Liu, Mengfei Zhou, Zhuosheng Zhang, Rui Wang, Zhaopeng Tu, Haitao Mi, and Dong Yu. Do NOT think that much for 2+3=? on the overthinking of long reasoning models. In *Forty-second International Conference on Machine Learning, ICML 2025, Vancouver, BC, Canada, July 13-19, 2025*. OpenReview.net, 2025.

[59] Tuomas Haarnoja, Aurick Zhou, Pieter Abbeel, and Sergey Levine. Soft actor-critic: Off-policy maximum entropy deep reinforcement learning with a stochastic actor. In *Proceedings of the 35th International Conference on Machine Learning, ICML 2018, Stockholmsmässan, Stockholm, Sweden, July 10-15, 2018*, pages 1856–1865, 2018.

[60] Wenlong Deng, Yi Ren, Muchen Li, Danica J. Sutherland, Xiaoxiao Li, and Christos Thrampoulidis. On the effect of negative gradient in group relative deep reinforcement optimization. *CoRR*, abs/2505.18830, 2025.

[61] Zhisheng Chen, Bingfan Zeng, Bangde Cao, Zhengwei Xie, Yuxuan Li, Jinhan Li, Zheng Lu, Xiangchen Guan, Zikai Xiao, Rui Qian, et al. Memprism: Task-conditioned relational memory views for long-horizon agents. *arXiv preprint arXiv:2608.06745*, 2026.

[62] Bingfan Zeng, Zhisheng Chen, Chenbo Sang, Zhengwei Xie, Jinpeng Wang, Xiangchen Guan, Rui Qian, Zheng Lu, and Jingwei Song. Memowm: How world models change what agents need to remember. *arXiv preprint arXiv:2610.10778*, 2026.

[63] Michael Luo, Sijun Tan, Roy Huang, Ameen Patel, Alpay Ariyak, Qingyang Wu, Xiaoxiang Shi, Rachel Xin, Colin Cai, Maurice Weber, Ce Zhang, Li Erran Li, Raluca Ada Popa, and Ion Stoica. Deepcoder: A fully open-source 14b coder at o3-mini level. https://pretty-radio-b75.notion.site/DeepCoder-A-Fully-Open-Source-14B-Coder-at-O3-mini-Level-1cf81902c14680b3bee5eb349a512a51, 2025. Notion Blog.

[64] Naman Jain, King Han, Alex Gu, Wen-Ding Li, Fanjia Yan, Tianjun Zhang, Sida Wang, Armando Solar-Lezama, Koushik Sen, and Ion Stoica. Livecodebench: Holistic and contamination free evaluation of large language models for code, 2024.

Appendix

The appendices are organized as follows. Appendix A reviews related work on reinforcement learning with verifiable rewards, existing entropy-collapse mitigation strategies, and emerging token-level perspectives, situating STARE within these literatures and clarifying its conceptual and methodological distinctions from prior approaches. Appendix B reports additional empirical results that complement the main paper, including a detailed comparison between STARE and GRPO-ds across diverse scenarios and model scales from 1.5B to 32B (Appendix B.1), an empirical validation of the batch-internal high-surprisal quantile proxy against the theoretical critical threshold (Appendix B.2), ablations of all four single-polarity and four combined token-level reweighting operations (Appendix B.3), ablations on the key hyperparameters W and P together with the target-entropy gate (Appendices B.4 and B.5), an analysis of emergent reflection behaviors elicited by STARE (Appendix B.6), a comparison between fixed and adaptive reweighting schedules (Appendix B.7), ablations on the target-entropy threshold and on the gate granularity (Appendices B.8 and B.9), a validation of STARE under off-policy training (Appendix B.10), and a comparison with fixed-threshold low-probability token reweighting (Appendix B.11). It further reports a progressive ablation from standard GRPO to STARE (Appendix B.12), controlled comparisons with entropy-regulation baselines (Appendix B.13), Pass@32 across scales and scenarios (Appendix B.14), robustness to noisy model-judged rewards (Appendix B.15), code-generation results (Appendix B.16), computational overhead (Appendix B.17). Appendix C presents the complete pseudocode of the default STARE-O1 procedure, integrating surprisal-guided entropy-critical token selection, fixed-weight token-level advantage reweighting, and batch-level closed-loop target-entropy gating into a unified algorithm. Appendix D establishes the basic differential identities used in Sections 2 and 3.1 and proves Theorem 3.1. Appendix E proves Proposition 3.2, Corollary 3.3, Theorem 3.4, Proposition 3.5, and Theorem 3.6, provides the detailed asymmetric entropy contribution analysis, and states explicitly the statistical assumptions underlying the near-criticality result. Appendix F formalizes the cross-step entropy dynamics discussed in Section 3.4. Appendix G presents all single-polarity operations referenced in the main text, together with sample-level and token-level closed-loop extensions. Appendix H presents all combined operations and the adaptive weighting scheme. Finally, Appendix I discusses the limitations of our study and its broader societal impacts.

To keep the notation unambiguous, Appendices D–F use the exact theoretical sets from Section 3,

$$\widetilde{\mathcal{L}}^{+} \triangleq \{(i,t) : \hat{A}_i > 0,\ \mathfrak{s}_{i,t} > \mathfrak{s}^*_{i,t}\}, \qquad \widetilde{\mathcal{L}}^{-} \triangleq \{(i,t) : \hat{A}_i < 0,\ \mathfrak{s}_{i,t} > \mathfrak{s}^*_{i,t}\},$$

namely, the sets of token positions whose trajectories have positive (resp. negative) advantage and whose token surprisal exceeds the position-dependent critical threshold. By contrast, Appendices G and H use the quantile-based proxy sets employed in the practical implementation of Section 4, namely $\mathcal{L}_q^+$ and $\mathcal{L}_q^-$.

A Related Work

Reinforcement Learning with Verifiable Rewards(RLVR). RLVR has emerged as the dominant post-training paradigm for enhancing the reasoning capability of LLMs, as it leverages verifiable signals to provide precise outcome-level rewards while avoiding the overfitting risks inherent in learned reward models[7, 52, 53]. GRPO removes the value network and estimates the advantage through group-normalized rewards, demonstrating strong effectiveness in mathematical reasoning, code generation, and tool-use tasks, and further eliciting emergent behaviors such as long chain-of-thought reasoning and self-reflection[1, 6, 54, 55]. Subsequent studies extend the GRPO framework along several dimensions, including advantage estimation, loss aggregation, and sampling strategies, among which DAPO has become a representative baseline through a combination of asymmetric clipping, dynamic sampling, and token-level loss normalization[10, 11, 21, 24, 56]. As training proceeds over more optimization steps, however, GRPO-style algorithms commonly suffer from policy entropy collapse, in which the entropy decays rapidly, the output diversity vanishes, within-group rollouts become homogeneous, and the number of trainable steps is ultimately capped[8, 9, 57, 58]. Existing studies have confirmed the prevalence of this phenomenon and have established empirical correlations between policy entropy and downstream performance; however, a fine-grained theoretical characterization of the underlying token-level gradient causes of entropy collapse remains absent.

Entropy Collapse Mitigation Methods.Existing mitigation strategies fall into three categories. The first category protects low-probability tokens by adjusting the clipping thresholds of the importance-

sampling ratio, as exemplified by the clip-higher mechanism in DAPO and by subsequent variants such as differentiated clipping and smooth gating; the influence of these mechanisms on entropy is asymmetric and difficult to control precisely, and in the on-policy regime where the sampling ratio remains close to one, clipping is rarely activated, leaving the actual regulatory capacity limited[10–12, 56, 59]. The second category applies trajectory-level differentiated weighting between positive and negative samples, including asymmetric importance sampling and the upweighting of rare correct rollouts; since these methods operate at the trajectory granularity, they cannot distinguish the opposing entropy effects of different tokens within the same trajectory[14–16, 25, 60]. The third category couples token-level entropy information into the advantage or loss through entropy-induced advantages[21, 22, 24, 26], explicit entropy regularization, or token filtering based on entropy variation; entropy rewards tend to overamplify the signal of high-entropy tokens and induce oscillations, regularization remains highly sensitive to the choice of its coefficient, and entropy-variation-based methods either rely on hard-to-estimate information about unsampled tokens or impose oversimplified binary partitions. In addition, raising the sampling temperature only delays rather than prevents entropy collapse. Overall, existing methods either operate at an excessively coarse granularity or lack a principled understanding of the underlying collapse mechanism.

Token-Level Perspectives and STARE. Recent studies have begun to focus on the differentiated contributions of individual tokens[11, 14–17]. Some works demonstrate that a small subset of high-entropy tokens dominates the effective learning signal in RLVR and that performing gradient updates only on this subset already yields efficient performance gains[16]; another line of work identifies critical tokens at decision branching points along reasoning chains and encourages exploration at these positions[11, 15, 21, 25]. These findings echo the analysis presented in this paper. Most existing methods, however, remain heuristic in nature and either fail to keep the policy entropy stable and controllable or lack comprehensive validation across multiple scenarios, model scales, and long-horizon agent training settings. Fine-grained token-level importance estimation is also critical for agent memory management and long-context compression[61, 62]. Starting from a first-order analysis of token-level entropy dynamics, STARE establishes an advantage-surprisal four-quadrant decomposition that exposes the credit assignment mismatch under shared trajectory-level advantages, together with a near-criticality property; building on this analysis, STARE identifies entropy-critical tokens through a batch-internal surprisal-quantile proxy, selectively reweights their effective advantages, and incorporates a target-entropy closed-loop gate, thereby achieving principled token-level regulation of the policy entropy through a minimally invasive modification to the GRPO objective. STARE is comprehensively validated across model scales ranging from 1.5B to 32B and across three task families covering Short-CoT, Long-CoT, and Tool-Use scenarios, where it consistently delivers stable performance improvements and further unlocks the optimization potential of long-horizon RL training.

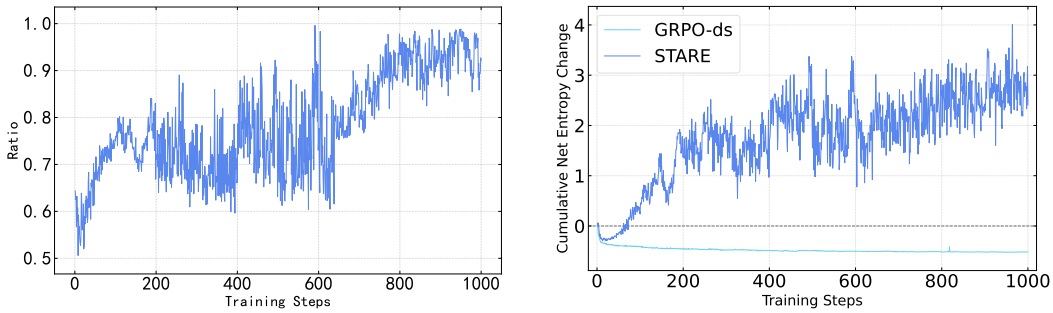
B Additional Experiments and Analysis on STARE

B.1 STARE vs. GRPO-ds: Detailed Comparison across Diverse Scenarios and Model Scales in RL Training

To validate the effectiveness of STARE in long-horizon RL training, we conduct RL training of 5000 steps on Qwen2.5-Math-7B-Base under the Short CoT scenario. Figure 3 compares STARE against GRPO-ds on key metrics, while Figures 1, 5, 6, 7, 8 further verify the effectiveness of STARE across model scales ranging from 1.5B to 32B and across diverse task scenarios.

Entropy stability and performance evolution. In Figure 3(a), GRPO-ds exhibits entropy collapse during the early training phase (0–1000 steps), in which the policy entropy decreases sharply and approaches zero, which is consistent with the theoretical analysis presented in Section 3. Meanwhile, the accuracy of GRPO-ds on AIME24 and AIME25 peaks around step 1000 and subsequently saturates, fluctuating without further improvement (Figure 3(e)-(f)), thereby indicating premature convergence of the policy distribution. In contrast, STARE stabilizes the policy entropy near $H_{\text{tgt}} = 0.3$ through token-level advantage reweighting and closed-loop entropy gating, and the accuracy of STARE continues to improve beyond step 1000, reaching the optimum at step 5000, which extends the number of trainable steps and unlocks the optimization potential of long-horizon RL training.

Exploration-exploitation balance. In Figure 3(g)-(h), the Pass@32 of STARE consistently surpasses that of GRPO-ds throughout training, which indicates that the policy retains sufficient output diversity



(a) Fraction of top-$P\%$ high-surprisal tokens in $\mathcal{L}_q^+$ that fall within the theoretical entropy-increasing region ($s_{i,t} > s^*_{i,t}$) across training.

(b) Cumulative net entropy contribution of $\mathcal{L}_q^+$ under STARE versus GRPO-ds.

Figure 10: Validation of the batch-internal surprisal-quantile proxy on Qwen2.5-Math-7B-Base over 1000 RL training steps, examining its alignment with the theoretical critical threshold and the cumulative net entropy contribution of the selected token subset.

and effectively mitigates mode-seeking behavior. In Figure 3(b)-(c), the training reward and Full-Solve Ratio of GRPO-ds rise rapidly during the early phase and subsequently plateau, whereas STARE maintains a steady upward trajectory throughout training. Meanwhile, the sustained growth of response length under STARE (Figure 3(d)) suggests that the model addresses complex problems by extending the reasoning depth.

Cross-scale and cross-scenario generalization. Figures 1, 5, 6, 7, 8 systematically present the behavior of STARE under broader configurations. Across the Short CoT scenario(14B, 32B), the Long CoT scenario(R1-Distill-Qwen-1.5B, Qwen3-8B-Base), and the tool-use scenario(7B), the results exhibit a consistent pattern: GRPO-ds suffers from entropy collapse accompanied by performance saturation, whereas STARE maintains entropy stability and delivers continuous accuracy improvements. Notably, on DeepSeek-R1-Distill-Qwen-1.5B, GRPO-ds undergoes gradual entropy decay that drops below 0.2 by step 5000, whereas under STARE the policy entropy rapidly recovers to the target band and remains stable starting around step 3500, with the accuracy on AIME24 and AIME25 improving in tandem (Figure 7), which validates the intervention-recovery capability of the proposed method.

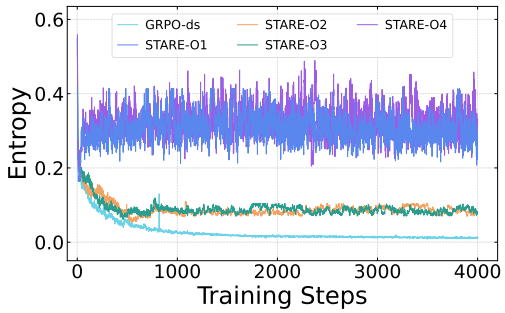
In summary, STARE maintains stable policy entropy and delivers consistent performance gains across model scales ranging from 1.5B to 32B and across the three task scenarios, which confirms the effectiveness and robustness of the proposed method.

B.2 Validation of the High-Surprisal Quantile Proxy.

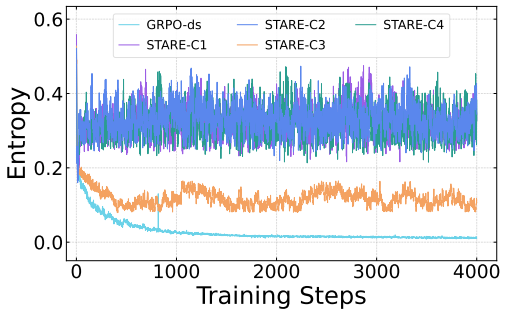
To verify the effectiveness of the batch-internal top-$P\%$ surprisal-quantile proxy in STARE, Figure 10 evaluates this proxy from two complementary perspectives. Figure 10(a) reports the fraction of tokens selected by the top-$P\%$ high-surprisal criterion that fall within the theoretical entropy-increasing region ($\mathfrak{s}_{i,t} > \mathfrak{s}^*_{i,t}$). This fraction rises steadily from approximately 60% in the early training phase to around 95%, indicating strong consistency between the surprisal-quantile proxy and the theoretical critical threshold $\mathfrak{s}^*$ (Proposition 3.2). Figure 10(b) further shows that the cumulative net entropy contribution of this subset remains positive and monotonically increasing throughout training, thereby confirming that $\mathcal{L}_q^+$ consistently produces a net entropy-increasing effect, as predicted by Corollary 3.3. These results demonstrate that the batch-internal surprisal-quantile mechanism effectively approximates the theoretical threshold $\mathfrak{s}^*$ and accurately identifies entropy-critical tokens while eliminating the need to solve $\Phi(p^*) = 0$ at each position, thereby validating both the feasibility and the effectiveness of the proposed proxy strategy.

B.3 Detailed Effects of Single-Polarity and Combined Operations in STARE.

To validate the effectiveness of the token-level reweighting mechanism under the four-quadrant decomposition framework, we ablate all four single-polarity operations (O1, O2, O3, O4) and all four



(a) Entropy evolution under single-polarity operations O1-O4, each targeting one entropy-critical quadrant.



(b) Entropy evolution under combined operations C1-C4, each jointly regulating two entropy-critical quadrants.

Figure 11: Policy entropy trajectories of all eight STARE variants versus GRPO-ds on Qwen2.5-Math-7B-Base over 4000 RL training steps, covering four single-polarity operations (O1-O4) and four combined operations (C1-C4) derived from the advantage-surprisal four-quadrant decomposition.

combined operations (C1, C2, C3, C4). Table 2 reports the accuracy of each operation on AIME24 and AIME25, and Figure 11 presents the corresponding policy entropy evolution curves. As shown in Figure 11, GRPO-ds suffers from rapid entropy decay toward zero in the early training phase, whereas all eight STARE variants effectively mitigate entropy collapse. Specifically, O1 and O3 strengthen entropy-increasing signals through weight amplification, O2 and O4 suppress entropy-decreasing signals through weight attenuation, and C1–C4 combine interventions from two quadrants simultaneously. Table 2 further shows that all STARE variants substantially outperform GRPO-ds, confirming that token-level reweighting is effective across all four entropy-critical quadrants. Among all variants, O1 (amplifying $\mathcal{L}_q^+$) and C2 (amplifying $\mathcal{L}_q^+$ while attenuating $\mathcal{L}_q^-$) achieve the strongest performance, reaching $44.2\%/23.8\%$ and $42.5\%/24.2\%$ on AIME24/AIME25 respectively. We therefore adopt STARE-O1 as the default configuration and STARE-C2 as the dual-sided variant in our work.

B.4 Detailed Ablation on Key Hyperparameters and Target-Entropy Gating

To validate the mechanisms underlying the key hyperparameters in STARE (W and P) and the target-entropy gating, we conduct ablation experiments on Qwen2.5-Math-7B-Base (Figure 4). Figure 4(a) examines the reweighting factor W under an open-loop setting (without the target-entropy gate). A minimal perturbation of $W = 1.01$ is sufficient to mitigate the entropy decay observed in GRPO-ds; $W \geq 1.05$ yields steady entropy growth, whereas $W \geq 2.0$ drives the entropy upward too aggressively and triggers divergence. This behavior corroborates the near-criticality property (Corollary 3.6): beyond the critical threshold, W governs the magnitude rather than the direction of the entropy shift. The open-loop regime, however, tends to stabilize the entropy at an excessively high level and thereby induces over-exploration. Once the target-entropy closed-loop gate is introduced with $H_{\text{tgt}} = 0.3$, Figure 4(b) shows that values of W across the range $[1.05, 1.5]$ steer the policy entropy into the target band and maintain bounded oscillation around it, which confirms the stability of the closed-loop mechanism and substantially reduces the sensitivity to W. Figure 4(c) further investigates the high-surprisal token selection ratio P. With $W = 1.1$ and $H_{\text{tgt}} = 0.3$ held fixed, the policy entropy remains within the target band for $P \in [5\%, 20\%]$, and even at $P = 40\%$ it is confined to $[0.1, 0.2]$, effectively preventing collapse and indicating that the surprisal-quantile selection admits a broad operating range. Overall, STARE exhibits strong robustness to the configurations of W and P, and the target-entropy gate stably constrains the policy entropy. More detailed entropy evolution and performance comparisons are reported in Appendix B.5, Table 3, and Figure 12

B.5 Effectiveness of Target-Entropy Closed-Loop Gating

To validate the necessity of the closed-loop target-entropy gate, we fix $W = 1.1$ and $P = 10\%$ on Qwen2.5-Math-7B-Base, and compare three configurations over 1000 RL training steps: GRPO-ds, open-loop STARE, and closed-loop STARE with $H_{\text{tgt}} = 0.3$. As shown in Figure 12, open-loop

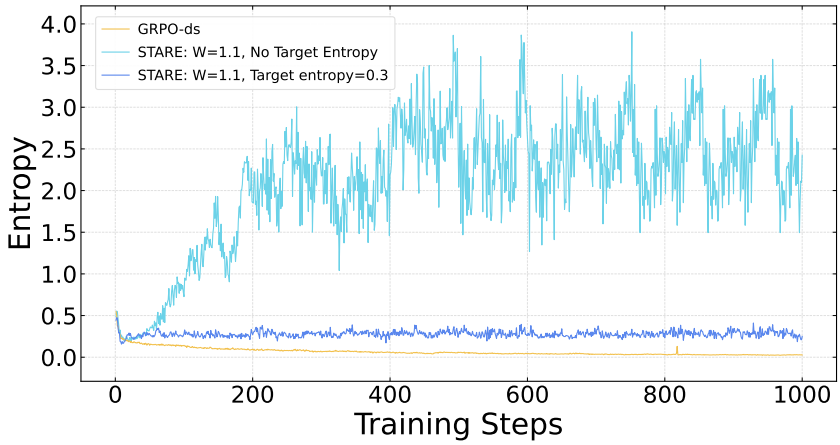


Figure 12: Effect of the target-entropy closed-loop gate on policy entropy regulation for STARE (W=1.1, P=10%) on Qwen2.5-Math-7B-Base over 1000 RL training steps. The closed-loop gate effectively confines the policy entropy to bounded oscillation around the target level H_{tgt}=0.3, preventing both entropy collapse and over-exploration observed under open-loop reweighting and standard GRPO-ds, respectively.

Table 3: Ablation of the target-entropy gate for STARE with $W = 1.1$ and P=10 on AIME24 and AIME25.

Model	**AIME24**	**AIME25**
GRPO-ds	35.2	17.3
STARE without target-entropy gate	36.7	18.9
STARE with target-entropy gate	38.0	20.0

STARE alleviates entropy collapse but drives the policy entropy upward to an excessively high level, which induces over-exploration; in contrast, closed-loop STARE confines the policy entropy to bounded oscillation around the target band. Table 3 further reports that open-loop STARE attains $36.7\%/18.9\%$ on AIME24/AIME25, already surpassing GRPO-ds, while closed-loop STARE further improves the accuracy to $38.0\%/20.0\%$, with both variants substantially outperforming the GRPO-ds baseline at $35.2\%/17.3\%$. These results confirm that token-level advantage reweighting in STARE mitigates entropy collapse and delivers consistent gains, and that the closed-loop gate stabilizes the policy entropy within a reasonable band to balance exploration and exploitation, thereby yielding additional performance improvements.

B.6 Details about Emergent Reflection Behaviors.

To investigate how STARE elicits deep reasoning, we analyze the reflection behaviors that emerge during RL training. Taking Qwen2.5-32B-Base as a representative example, we randomly sample 50 training steps shared by both STARE and GRPO-ds. This analysis is intended as a coarse, qualitative diagnostic rather than a principled measurement of reasoning depth: we apply heuristic regular-expression matching based on manually-observed patterns to obtain an approximate count of reflection-related tokens, grouped into six categories: contrast (*but, however*), reflection (*wait, reconsider*), self-correction (*made a mistake, correct*), hesitation (*perhaps, possibly, maybe*), backtracking (*recalculate, redo, go back*), and summary revision (*summary, ultimately*). Figure 13 reports reflection-related token counts per 1k samples; STARE substantially exceeds GRPO-ds across all six categories, with markedly larger margins on reflection and self-correction, indicating that the resulting policy retains stronger exploration and greater output diversity. Figure 9 further reports the frequency distribution of tokens selected by STARE for advantage amplification; the word cloud reveals that the reweighted tokens concentrate on vocabulary expressing uncertainty and self-correction, such as *should be*, *but*, *instead*, and *verification*, confirming that the batch-internal surprisal-quantile selection effectively identifies rare forking tokens carrying exploratory semantics. Taken together,

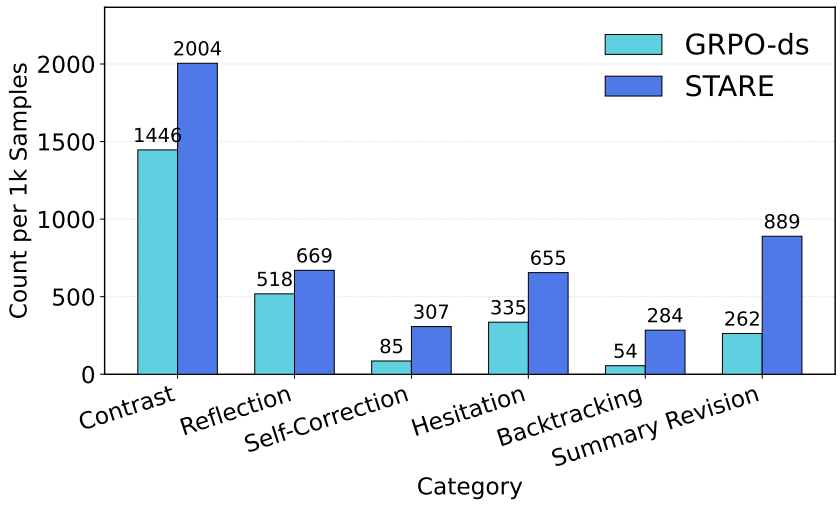


Figure 13: Reflection-related token counts per 1k samples for STARE vs. GRPO-ds on Qwen2.5-32B-Base.

the emergent growth of reflection tokens and the semantic bias of the reweighted token distribution demonstrate that STARE activates the deep exploration capability of the model and delivers consistent performance gains through token-level credit rebalancing.

Table 4: Ablation of fixed and adaptive reweighting coefficients for STARE in 1000 RL training steps on AIME24 and AIME25.

Model	**AIME24**	**AIME25**
GRPO-ds	35.2	17.3
STARE: Fixed $W = 1.1$	38.0	20.0
STARE: Adaptive W, $W_{\max} = 1.5$, $\alpha = 0.01$	37.7	19.5
STARE: Adaptive W, $W_{\max} = 1.5$, $\alpha = 0.02$	37.4	19.1
STARE: Adaptive W, $W_{\max} = 2.0$, $\alpha = 0.01$	36.9	18.6
STARE: Adaptive W, $W_{\max} = 2.0$, $\alpha = 0.02$	37.1	18.1

B.7 Fixed vs. Adaptive Weights

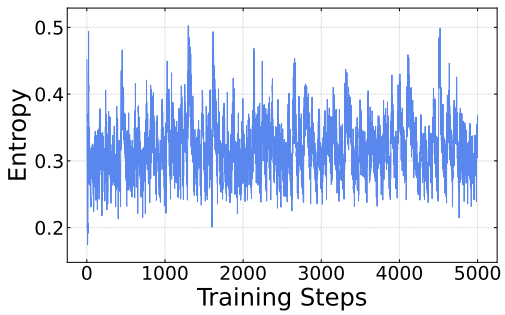
Table 4 compares fixed and adaptive weighting schemes on Qwen2.5-Math-7B-Base after 1000 RL training steps. All STARE configurations substantially outperform the GRPO-ds baseline, confirming the robustness of the token-level reweighting mechanism. The fixed weight $W = 1.1$ yields the strongest performance, reaching 38.0% on AIME24 and 20.0% on AIME25. This outcome aligns with the near-criticality property (Corollary 3.6): once the reweighting factor exceeds the critical threshold, its specific value primarily controls the magnitude rather than the direction of the per-step entropy shift, so that a moderate fixed value suffices for stable regulation. Among the adaptive variants, the configuration with $W_{\max} = 1.5$ and $\alpha = 0.01$ performs best and nearly matches the fixed-weight result, whereas enlarging either $W_{\max}$ or α leads to a slight degradation. We therefore adopt the fixed $W = 1.1$ as the default configuration, with $W_{\max} = 1.5$ and $\alpha = 0.01$ recommended when an adaptive schedule is preferred.

B.8 Ablation on Target Entropy Threshold

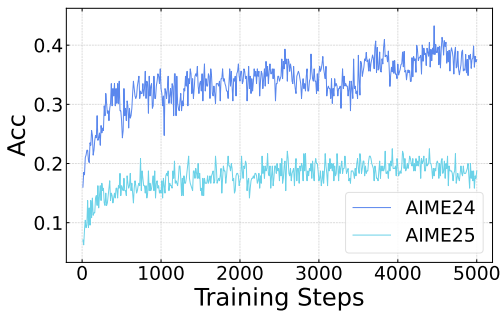
Table 5 reports the performance of STARE under different target entropy values $H_{\text{tgt}} \in \{0.1, 0.2, 0.3, 0.4\}$ on Qwen2.5-Math-7B-Base after 4000 RL training steps. All configurations substantially outperform the GRPO-ds baseline, confirming the general effectiveness of the closed-loop entropy regulation mechanism. Performance attains its optimum at $H_{\text{tgt}} = 0.3$, reaching 44.2% on AIME24 and 23.8% on AIME25, with gains of 7.1% and 6.1% over the baseline respectively.

Table 5: Ablation of target entropy threshlod for STARE in 4000 RL training steps on AIME24 and AIME25.

Model	AIME24	AIME25
GRPO-ds	37.1	17.7
STARE: $H_{\text{target}} = 0.1$	40.4	20.5
STARE: $H_{\text{target}} = 0.2$	43.2	23.1
STARE: $H_{\text{target}} = 0.3$	44.2	23.8
STARE: $H_{\text{target}} = 0.4$	42.8	21.6



(a) Policy entropy evolution of STARE under off-policy training (four gradient updates per batch).



(b) AIME24 and AIME25 accuracy of STARE under off-policy training.

Figure 14: Off-policy training dynamics of STARE on Qwen2.5-Math-7B-Base with four gradient updates per batch over 5,000 RL training steps, demonstrating that STARE preserves entropy stability and delivers consistent accuracy improvements beyond the on-policy setting.

The low target entropy restricts the exploration space, whereas the high value easily drives the policy distribution toward over-exploration, and both regimes yield suboptimal outcomes. These results indicate that H_{tgt} effectively governs the exploration-exploitation balance, and we therefore adopt $H_{\text{tgt}} = 0.3$ as the default configuration in the main experiments.

Table 6: Ablation of target-entropy gate granularity for STARE in 1000 RL training steps on AIME24 and AIME25.

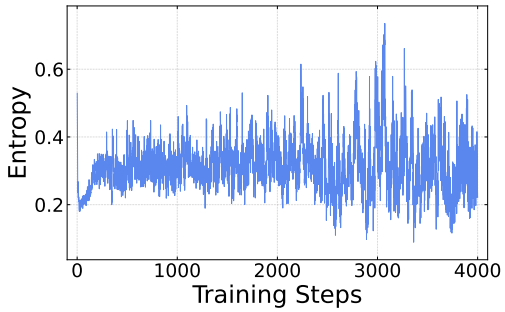
Model	AIME24	AIME25
GRPO-ds	35.2	17.3
STARE (token-level target-entropy gate)	36.9	19.1
STARE (sample-level target-entropy gate)	37.6	19.3
STARE (batch-level target-entropy gate)	38.0	20.0

B.9 Ablation on Target-Entropy Gate Granularity

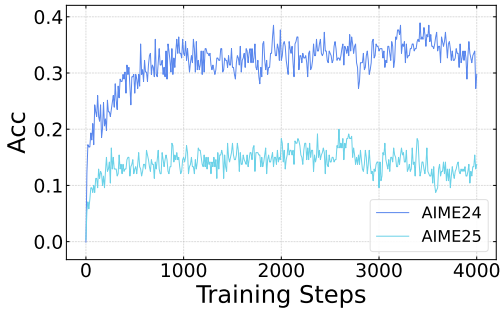
On Qwen2.5-Math-7B-Base, we compare three closed-loop gating granularities: a token-level gate that evaluates $H_{i,t} < H_{\text{tgt}}$ at each position, a sample-level gate that conditions activation on the per-sample average entropy $\bar{H}_i$, and a batch-level gate that makes a unified decision based on the batch-average entropy $\bar{H}_k$. As reported in Table 6, all three granularities substantially outperform the GRPO-ds baseline (35.2%/17.3%), confirming the general effectiveness of the closed-loop mechanism. The batch-level gate yields the best performance (38.0%/20.0%), surpassing the token-level (36.9%/19.1%) and sample-level (37.6%/19.3%) variants. Since local entropy estimates at the token and sample levels exhibit higher variance, they tend to induce frequent switching of the gating signal and undermine the stability of the regulation. We therefore adopt the batch-level gate as the default setting in our main experiments.

Table 7: Comparison of off-policy and on-policy STARE in 4000 RL training steps on AIME24 and AIME25.

Model	AIME24	AIME25
GRPO-ds	37.1	17.7
STARE with off-policy	43.8	22.1
STARE with on-policy	44.2	23.8



(a) Policy entropy evolution under fixed-threshold low-probability token reweighting.



(b) AIME24 and AIME25 accuracy under fixed-threshold low-probability token reweighting.

Figure 15: Comparison between fixed-threshold low-probability token reweighting ($p < 0.1$) and GRPO-ds on Qwen2.5-Math-7B-Base over 4000 RL training steps, demonstrating the inferior entropy regulation and benchmark performance of static probability-based selection relative to STARE's batch-internal surprisal-quantile proxy.

B.10 Validation of STARE under Off-Policy Training

To assess the applicability of STARE in the off-policy setting, we conduct 4000 steps of RL training on Qwen2.5-Math-7B-Base with four gradient updates per batch. Figure 14 shows that STARE still maintains the policy entropy within the target band, while the accuracy improves steadily throughout training. As reported in Table 7, off-policy STARE attains $43.8\%/22.1\%$ on AIME24/AIME25, exceeding GRPO-ds ($37.1\%/17.7\%$) by 6.7% and 4.4% respectively, and falling only marginally short of on-policy STARE ($44.2\%/23.8\%$). These results confirm that STARE remains effective under off-policy training, demonstrating the robustness and generality of the proposed method.

B.11 STARE vs. Fixed-Threshold Low-Probability Token Reweighting

STARE identifies entropy-critical tokens through a batch-internal surprisal-quantile proxy. A more direct alternative is to apply uniform weight amplification to all tokens whose sampling probability falls below a fixed threshold such as $p < 0.1$. Under an identical configuration on Qwen2.5-Math-7B-Base with 4000 RL training steps, $W = 1.1$, and $H_{\text{tgt}} = 0.3$, we compare these two token-selection strategies. As reported in Table 8, fixed-threshold reweighting yields only marginal improvements over GRPO-ds, with gains of 1.8% on AIME24 and 2.0% on AIME25, whereas STARE achieves substantially larger gains of 7.1% and 6.1% respectively, confirming the superiority of the batch-internal top-$P\%$ surprisal-quantile proxy in identifying entropy-critical tokens. Moreover, the quantile proxy of STARE adaptively tracks the current policy distribution at every training step, and as shown in Figure 10(a), its agreement with the theoretical threshold $\mathfrak{s}^*$ rises from approximately 60% to over 95% throughout training, ensuring high-precision identification of entropy-critical tokens. These results jointly demonstrate the effectiveness and reliability of the batch-internal surprisal-quantile proxy adopted by STARE.

B.12 Progressive Ablation from Standard GRPO to STARE

The main comparison with GRPO-ds isolates the effect of adding STARE to an enhanced GRPO recipe. To examine whether the gains depend on that recipe, we also apply STARE-O1 directly to standard GRPO and conduct a progressive ablation of the modifications that constitute GRPO-

Table 8: Comparison with Fixed-Threshold Low-Probability Token Reweighting on AIME24 and AIME25.

Method	AIME24	AIME25
GRPO-ds	37.1	17.7
Fixed-Threshold Reweighting ($p < 0.1$)	38.9	19.7
STARE	44.2	23.8

Table 9: STARE-O1 on standard GRPO and a progressive ablation from standard GRPO to GRPO-ds and STARE-O1, using Qwen2.5-Math-7B-Base in the Short CoT scenario. Results are average accuracy (%).

Configuration	AIME24	AIME25
Standard GRPO	33.1	10.5
Standard GRPO + STARE-O1	**40.5**	**19.2**
Standard GRPO	33.1	10.5
+ Remove KL regularization	34.0	11.9
+ Token-level loss	35.4	14.1
+ Dynamic sampling (GRPO-ds)	37.1	17.7
+ STARE-O1	**44.2**	**23.8**

ds. All comparisons in this ablation use Qwen2.5-Math-7B-Base in the Short CoT scenario, with matched training configurations within each comparison. We report average accuracy on AIME24 and AIME25.

STARE on standard GRPO. Table 9, panel (a), compares standard GRPO with the same baseline augmented by STARE-O1, which adds surprisal-guided token-level advantage reweighting and the target-entropy gate. STARE-O1 improves AIME24 from 33.1% to 40.5% and AIME25 from 10.5% to 19.2%, yielding gains of 7.4% and 8.7%, respectively. Thus, the proposed intervention remains effective without relying on the additional modifications in GRPO-ds.

Progressive construction of GRPO-ds. Panel (b) first removes KL regularization, then adopts token-level loss aggregation, and finally introduces dynamic sampling. Along this configuration sequence, AIME24 improves from 33.1% to 34.0%, 35.4%, and 37.1%, while AIME25 improves from 10.5% to 11.9%, 14.1%, and 17.7%. Adding STARE-O1 to the resulting GRPO-ds baseline further raises accuracy to $44.2\%/23.8\%$, an additional gain of $7.1\%/6.1\%$.

Table 10: Controlled comparison on AIME24 and AIME25 with Qwen2.5-Math-7B-Base, DAPO-17K, and a shared off-policy budget of 800 rollout steps and 3200 gradient updates. Results are Avg@32 accuracy (%).

Method	AIME24	AIME25
DAPO	36.7	16.5
STEER	38.6	19.2
EAPO	39.2	18.7
STARE	**42.7**	**21.8**

B.13 Controlled Comparison with Entropy-Regulation Baselines

To complement the results cited from prior work in Table 1, we reproduce DAPO, STEER, and EAPO and compare them with STARE under a common training and evaluation protocol. This comparison controls the training data, initialization, response-length limit, update budget, and shared optimization settings, which can differ across the published results.

Experimental protocol. All methods start from Qwen2.5-Math-7B-Base and train on DAPO-17K in the Short CoT scenario with a maximum output length of 3k tokens. We use a generation batch size of 512, a training batch size of 128, and eight rollouts per prompt. Each rollout step is followed by

Table 11: Pass@32 comparison between STARE-O1 and GRPO-ds on AIME24 and AIME25 across different model scales and task scenarios.

Method	AIME24 Pass@32	AIME25 Pass@32	Avg.
In the Short CoT Scenario			
RL from the Qwen2.5-Math-7B-Base			
GRPO-ds	60.9	34.0	47.5
STARE-O1	68.4	38.7	**53.6 (+6.1)**
RL from the Qwen2.5-14B-Instruct			
GRPO-ds	61.8	51.9	56.9
STARE-O1	67.5	58.2	**62.9 (+6.0)**
RL from the Qwen2.5-32B-Base			
GRPO-ds	67.4	54.4	60.9
STARE-O1	80.7	68.8	**74.8 (+13.9)**
In the Long CoT Scenario			
RL from the DeepSeek-R1-Distill-Qwen-1.5B			
GRPO-ds	74.3	54.6	64.5
STARE-O1	80.5	59.2	**69.9 (+5.4)**
RL from the Qwen3-8B-Base			
GRPO-ds	70.3	49.9	60.1
STARE-O1	75.6	55.6	**65.6 (+5.5)**
In the Tool-Use Agent Scenario			
RL from the Qwen2.5-7B-Base			
GRPO-ds	75.9	54.5	65.2
STARE-O1	81.0	59.9	**70.5 (+5.3)**

four gradient updates. Training lasts 800 rollout steps, corresponding to 3200 gradient updates. The shared settings use token-level loss aggregation, no KL regularization, a learning rate of 1×10^{-6}, and a training temperature of 1.0. These shared settings are combined with the respective methods' entropy-regulation mechanisms. We evaluate every ten rollout steps on AIME24 and AIME25 at temperature 0.7, using 32 independent samples per problem to report Avg@32, and report the best-performing checkpoint for each method.

Results. As shown in Table 10, STARE achieves $42.7\%/21.8\%$ on AIME24/AIME25. It exceeds DAPO by $6.0\%/5.3\%$, STEER by $4.1\%/2.6\%$, and EAPO by $3.5\%/3.1\%$. These results provide evidence for STARE's effectiveness in off-policy setting, in addition to the on-policy comparisons.

Relation to clip-higher. DAPO includes asymmetric clipping, which raises the upper importance-ratio threshold to preserve learning signals that would otherwise be clipped. At the start of a single-step on-policy update, $\rho_{i,t} = 1$ and both the standard and raised thresholds leave the local surrogate gradient unchanged. STARE can already modify that gradient through token-level weights. With multiple updates on a rollout batch, importance ratios can depart from one and the clipping thresholds can affect subsequent gradients. The controlled off-policy comparison above therefore evaluates DAPO in a setting where its clipping mechanism can operate.

B.14 Pass@32 across Model Scales and Task Scenarios

Table 11 reports Pass@32 on AIME24 and AIME25 for STARE-O1 and GRPO-ds across the six model - scenario configurations evaluated in the main paper. The 7B Short CoT training curves are shown in Figure 3(g)–(h).

STARE-O1 improves Pass@32 on both benchmarks in every configuration. In the 7B Short CoT setting, it increases AIME24 Pass@32 from 60.9% to 68.4% and AIME25 Pass@32 from 34.0% to 38.7%. The largest average improvement occurs at 32B, where Pass@32 increases from 60.9% to 74.8% across the two benchmarks. Improvements also persist in Long CoT and Tool-Use settings.

Table 12: AIME24 and AIME25 Avg@32 accuracy (%) after 3000 RL steps on DAPO-17K with Qwen3-32B-Thinking providing binary rewards without ground-truth answers.

Model / method	AIME24	AIME25
Qwen2.5-Math-7B-Base	10.0	5.1
GRPO-ds	29.5	14.7
STARE-O1	**34.1**	**18.9**

Table 13: Average code-generation accuracy (%) on LiveCodeBench v5, evaluated on problems released from August 1, 2024 to February 1, 2025. RL methods use DeepSeek-R1-Distill-Qwen-1.5B and 500 training steps on DeepCoder-Preview.

Model / method	LiveCodeBench v5
DeepSeek-R1-Distill-Qwen-1.5B	16.9
GRPO-ds	24.5
STARE-O1	**28.7**

Together with the entropy trajectories and average-accuracy results, these observations support the conclusion that STARE preserves useful exploration and increases multi-sample solution coverage.

B.15 Robustness to Noisy Model-Judged Rewards

STARE acts on trajectory-level advantages and therefore depends on the quality of the reward signal. To test its behavior when rewards are imperfect, we replace the rule-based training verifier with Qwen3-32B-Thinking as a judge. The judge assigns a binary reward, either 0 or 1, to each generated response without access to the ground-truth answer. This evaluates robustness to noisy model-judged outcome rewards while keeping the binary reward structure.

Reward agreement. We sample 500 problems from DAPO-17K and generate eight responses per problem using Qwen2.5-Math-7B-Instruct, yielding 4000 problem–response pairs. We compare the judge's binary decisions with ground-truth answer matching and obtain an agreement rate of 77.8%.

Training results. We initialize both GRPO-ds and STARE-O1 from Qwen2.5-Math-7B-Base and train on DAPO-17K for 3000 RL steps using the judge's rewards. Table 12 reports AIME24 and AIME25 Avg@32 accuracy. STARE-O1 reaches $34.1\%/18.9\%$, exceeding GRPO-ds at $29.5\%/14.7\%$ by $4.6\%/4.2\%$. It also improves substantially over the initial model at $10.0\%/5.1\%$. Thus, STARE retains a performance advantage under this imperfect reward source.

B.16 Generalization to Code Generation

To evaluate STARE beyond mathematical reasoning, we conduct a code-generation experiment following the DeepCoder training framework [63]. We initialize from DeepSeek-R1-Distill-Qwen-1.5B and use the approximately 24k-problem DeepCoder-Preview training dataset, whose problems are paired with executable test cases. During training, we sample eight responses per prompt and assign a binary reward: a generated program receives reward 1 if it passes all test cases and 0 otherwise. Both GRPO-ds and STARE-O1 are trained for 500 steps, with a maximum sequence length of 32k tokens.

We evaluate on the LiveCodeBench v5 code-generation split covering problems released from August 1, 2024 to February 1, 2025 [64], with a maximum evaluation sequence length of 64k tokens. Table 13 reports average accuracy. STARE-O1 improves the initial model's accuracy from 16.9% to 28.7% and exceeds GRPO-ds at 24.5% by 4.2%. This result provides evidence that surprisal-guided advantage reweighting also benefits code generation with execution-based rewards.

B.17 Computational Overhead and Training Efficiency

STARE reuses sampled-token log probabilities to compute surprisal and adds entropy reduction, quantile selection, and element-wise reweighting, without additional model forward/backward passes,

Table 14: Training efficiency of GRPO-ds and STARE on Qwen2.5-Math-7B in the Short CoT setting with 16 GPUs, averaged over 100 consecutive training steps.

Metric	**GRPO-ds**	**STARE**
Average response length	1167 tokens	1244 tokens
Average time per step	74 s	81 s
Training steps per hour	49	44
Token-generation throughput per GPU	1514 tokens/s	1476 tokens/s
Average GPU utilization	82%	79%

rollouts, or gradient updates. For N valid tokens, these operations cost $O(N|\mathcal{V}|)$ for entropy computation, $O(N_+ \log N_+)$ for sorting positive-advantage tokens in STARE-O1, and $O(N)$ for masking and reweighting.

We compare STARE with GRPO-ds on Qwen2.5-Math-7B in the Short CoT setting using 16 GPUs, a training batch size of 64, a generation batch size of 192, and eight responses per prompt. Both methods use dynamic sampling. Table 14 reports averages over 100 consecutive training steps. STARE increases step time by 9.5% and reduces token-generation throughput by 2.5%, while average response length increases by 6.6%. Thus, the measured overhead reflects both additional operations and longer generated responses. At inference, STARE requires no additional module.

C Algorithm: Main STARE Procedure

Algorithm 1 STARE-O1: Surprisal-Guided Token-Level Advantage Reweighting with Fixed Weights and Batch-Level Closed-Loop Target-Entropy Gating

Require: Initial policy π_{θ_0}; prompt distribution $\mathcal{D}$; batch size B; group size G; PPO clip range ϵ; reweighting factor $W > 1$; top-surprisal ratio $P \in (0, 1)$; target entropy H_{tgt}; learning rate η; total steps K

Ensure: Trained policy π_θ

Initialize $\theta \leftarrow \theta_0,\ \theta_{\text{old}} \leftarrow \theta_0$

for $k = 1, 2, \ldots, K$ **do**

 # Stage 1: Rollout and group-normalized advantage estimation

Sample a batch of prompts $\{x_i\}_{i=1}^{B} \sim \mathcal{D}$

for $i = 1, \ldots, B$ **do**

Roll out G responses $\{o_{i,j}\}_{j=1}^{G} \sim \pi_{\theta_{\text{old}}}(\cdot \mid x_i)$ and obtain verifier rewards $\{r_{i,j}\}_{j=1}^{G}$

$\hat{A}_{i,j} \leftarrow \big(r_{i,j} - \text{mean}(\{r_{i,\cdot}\})\big)/\text{std}(\{r_{i,\cdot}\})$ ▷ shared trajectory-level advantage

end for

Flatten all rollouts into a token-level batch with total length $N = \sum_{i,j} T_{i,j}$; broadcast $\hat{A}_i$ to every token (i, t)

 # Stage 2: Batch-level closed-loop entropy gating

$H_{i,t} \leftarrow -\sum_{v \in \mathcal{V}} \pi_\theta(v \mid c_{i,t}) \ln \pi_\theta(v \mid c_{i,t})$ for every position (i, t)

$\bar{H}_k \leftarrow \frac{1}{N} \sum_{(i,t)} H_{i,t}$

$g_k \leftarrow \mathbb{1}\big[\bar{H}_k < H_{\text{tgt}}\big]$ ▷ activate intervention only when entropy is below target

 # Stage 3: Surprisal-guided entropy-critical token selection (only when gated on)

if $g_k = 1$ **then**

$s_{i,t} \leftarrow -\ln \pi_\theta(o_{i,t} \mid x_i, o_{i,<t})$ for every token (i, t)

$\mathcal{T}^+ \leftarrow \{(i, t) \ :\ \hat{A}_i > 0\}$ ▷ positive-advantage token set

Sort $\mathcal{T}^+$ in *descending* order of $s_{i,t}$; let $K^+ \leftarrow \lceil P \cdot |\mathcal{T}^+| \rceil$

$\mathcal{L}_q^+ \leftarrow$ the first K^+ entries of the sorted list ▷ top-$P\%$ high-surprisal positive-advantage tokens

else

$\mathcal{L}_q^+ \leftarrow \emptyset$

end if

 # Stage 4: Token-level advantage reweighting (Variant I, one-sided amplification)

for each token (i, t) in the batch **do**

$$\omega_{i,t} \leftarrow \begin{cases} 1 + g_k\,(W - 1), & (i, t) \in \mathcal{L}_q^+ \\ 1, & \text{otherwise} \end{cases}$$

end for

 # Stage 5: STARE clipped-surrogate policy update

$$\rho_{i,t}(\theta) \leftarrow \frac{\pi_\theta(o_{i,t} \mid x_i, o_{i,<t})}{\pi_{\theta_{\text{old}}}(o_{i,t} \mid x_i, o_{i,<t})}$$

$$\mathcal{J}_{\text{STARE}}(\theta) \leftarrow \frac{1}{N} \sum_{(i,t)} \omega_{i,t} \min\Big(\rho_{i,t}(\theta)\,\hat{A}_i,\ \text{clip}\big(\rho_{i,t}(\theta),\, 1 - \epsilon,\, 1 + \epsilon\big)\hat{A}_i\Big)$$

$\theta \leftarrow \theta + \eta\, \nabla_\theta\, \mathcal{J}_{\text{STARE}}(\theta)$ ▷ single on-policy gradient step

$\theta_{\text{old}} \leftarrow \theta$

end for

return π_θ

D Basic Derivations for Sections 2 and 3.1

D.1 Softmax Jacobian Derivation

Proposition D.1 (Softmax Jacobian)**.** *Let*

$$\pi_\theta(v \mid c) = \frac{\exp(z_v)}{\sum_{u\in\mathcal{V}} \exp(z_u)}.$$

Then, for any $v, v' \in \mathcal{V}$*,*

$$\frac{\partial \pi_{v'}}{\partial z_v} = \pi_{v'}(\delta_{v'v} - \pi_v).$$

Proof. Let

$$Z \triangleq \sum_{u\in\mathcal{V}} \exp(z_u), \qquad \pi_{v'} = \frac{\exp(z_{v'})}{Z}.$$

Differentiating with respect to z_v yields

$$\frac{\partial \pi_{v'}}{\partial z_v} = \frac{\delta_{v'v}\exp(z_{v'})\,Z - \exp(z_{v'})\exp(z_v)}{Z^2}.$$

Factoring out $\exp(z_{v'})$ from the numerator gives

$$\frac{\partial \pi_{v'}}{\partial z_v} = \frac{\exp(z_{v'})}{Z}\left(\delta_{v'v} - \frac{\exp(z_v)}{Z}\right).$$

Using

$$\frac{\exp(z_{v'})}{Z} = \pi_{v'}, \qquad \frac{\exp(z_v)}{Z} = \pi_v,$$

we obtain

$$\frac{\partial \pi_{v'}}{\partial z_v} = \pi_{v'}(\delta_{v'v} - \pi_v).$$

This proves the result. □

D.2 Token-Level Logit Update in the Unclipped GRPO Regime

Proposition D.2 (Token-level logit update in the unclipped GRPO regime)**.** *In the unclipped regime of GRPO, at a given decoding position, the gradient of the local surrogate objective with respect to the logit vector is aligned with*

$$\hat{A}\,\nabla_z \log \pi_\theta(a \mid c).$$

Absorbing the positive proportionality constant into an infinitesimal step size $\eta > 0$ *gives the equivalent logit update*

$$\Delta z_v = \eta\,\hat{A}\,(\delta_{va} - \pi_v), \qquad v \in \mathcal{V}.$$

Proof. When clipping is inactive, the gradient direction of the local surrogate objective with respect to the model parameters is proportional to

$$\hat{A}\,\nabla_\theta \log \pi_\theta(a \mid c).$$

By the chain rule, an equivalent ascent direction in logit space is

$$\frac{\partial}{\partial z_v}\left[\hat{A}\log \pi_\theta(a \mid c)\right] = \hat{A}\,\frac{\partial}{\partial z_v}\log \pi_a.$$

Moreover,

$$\frac{\partial}{\partial z_v}\log \pi_a = \frac{1}{\pi_a}\frac{\partial \pi_a}{\partial z_v}.$$

Substituting Proposition D.1 gives

$$\frac{\partial \pi_a}{\partial z_v} = \pi_a(\delta_{va} - \pi_v),$$

and therefore

$$\frac{\partial}{\partial z_v} \log \pi_a = \frac{1}{\pi_a} \cdot \pi_a(\delta_{va} - \pi_v) = \delta_{va} - \pi_v.$$

Hence

$$\frac{\partial}{\partial z_v} \big[\hat{A} \log \pi_a\big] = \hat{A}(\delta_{va} - \pi_v).$$

Taking an infinitesimal gradient ascent step yields

$$\Delta z_v = \eta\, \hat{A}(\delta_{va} - \pi_v).$$

This proves the result. □

D.3 Lemma 2.1 (Entropy Gradient with Respect to Logits: Surprisal-Deviation Form)

For any $v \in \mathcal{V}$,

$$\frac{\partial H}{\partial z_v} = \pi_v(\mathfrak{s}_v - H), \qquad \mathfrak{s}_v \triangleq -\ln \pi_v.$$

Proof. Fix the context c and suppress it in the notation. By the definition of the Shannon entropy of the conditional next-token distribution,

$$H = -\sum_{u \in \mathcal{V}} \pi_u \ln \pi_u.$$

Differentiating with respect to z_v gives

$$\frac{\partial H}{\partial z_v} = -\sum_{u \in \mathcal{V}} \frac{\partial}{\partial z_v}\big(\pi_u \ln \pi_u\big).$$

Applying the product rule to each term,

$$\frac{\partial}{\partial z_v}(\pi_u \ln \pi_u) = \frac{\partial \pi_u}{\partial z_v} \ln \pi_u + \pi_u \cdot \frac{1}{\pi_u} \frac{\partial \pi_u}{\partial z_v} = \frac{\partial \pi_u}{\partial z_v}(\ln \pi_u + 1).$$

Thus

$$\frac{\partial H}{\partial z_v} = -\sum_{u \in \mathcal{V}} \frac{\partial \pi_u}{\partial z_v}(\ln \pi_u + 1).$$

Using Proposition D.1 yields

$$\frac{\partial H}{\partial z_v} = -\sum_{u \in \mathcal{V}} \pi_u(\delta_{uv} - \pi_v)(\ln \pi_u + 1).$$

Separating the two terms gives

$$\frac{\partial H}{\partial z_v} = -\sum_{u \in \mathcal{V}} \pi_u \delta_{uv}(\ln \pi_u + 1) + \sum_{u \in \mathcal{V}} \pi_u \pi_v (\ln \pi_u + 1).$$

The first term reduces to

$$-\sum_{u \in \mathcal{V}} \pi_u \delta_{uv}(\ln \pi_u + 1) = -\pi_v(\ln \pi_v + 1),$$

and the second term becomes

$$\sum_{u \in \mathcal{V}} \pi_u \pi_v (\ln \pi_u + 1) = \pi_v \sum_{u \in \mathcal{V}} \pi_u (\ln \pi_u + 1).$$

Since

$$\sum_{u \in \mathcal{V}} \pi_u \ln \pi_u = -H, \qquad \sum_{u \in \mathcal{V}} \pi_u = 1,$$

we have

$$\sum_{u \in \mathcal{V}} \pi_u (\ln \pi_u + 1) = -H + 1.$$

Substituting back yields

$$\frac{\partial H}{\partial z_v} = -\pi_v(\ln \pi_v + 1) + \pi_v(-H + 1) = -\pi_v \ln \pi_v - \pi_v H.$$

Using $\mathfrak{s}_v = -\ln \pi_v$, we obtain

$$-\pi_v \ln \pi_v = \pi_v \mathfrak{s}_v,$$

and therefore

$$\frac{\partial H}{\partial z_v} = \pi_v(\mathfrak{s}_v - H).$$

This proves the result. □

D.4 Theorem 3.1 (Token-Level Entropy Variation)

In the unclipped regime of GRPO, let $\hat{A}$ denote the trajectory-level normalized advantage assigned to the current token position, and let a denote the token sampled at the current decoding position, with conditional probability

$$p = \pi(a \mid c), \qquad \mathfrak{s}_a = -\ln p.$$

Define

$$S_2 \triangleq \sum_{v \in \mathcal{V}} \pi_v^2(\ln \pi_v + H),$$

and

$$\Phi(p) \triangleq p(\ln p + H) - S_2.$$

Then the first-order directional derivative of the conditional policy entropy along the GRPO policy-gradient direction satisfies

$$\left.\frac{dH}{d\eta}\right|_{\eta=0} = -\hat{A}\,\Phi(p).$$

Proof. By Proposition D.2, the logit velocity along the GRPO policy-gradient direction is

$$\left.\frac{dz_v}{d\eta}\right|_{\eta=0} = \hat{A}(\delta_{va} - \pi_v).$$

Therefore, by the definition of the directional derivative,

$$\left.\frac{dH}{d\eta}\right|_{\eta=0} = \sum_{v \in \mathcal{V}} \frac{\partial H}{\partial z_v} \left.\frac{dz_v}{d\eta}\right|_{\eta=0}.$$

Substituting Lemma 2.1 gives

$$\left.\frac{dH}{d\eta}\right|_{\eta=0} = \sum_{v \in \mathcal{V}} \pi_v(\mathfrak{s}_v - H)\,\hat{A}(\delta_{va} - \pi_v).$$

Factoring out $\hat{A}$ yields

$$\left.\frac{dH}{d\eta}\right|_{\eta=0} = \hat{A} \sum_{v \in \mathcal{V}} \pi_v(\mathfrak{s}_v - H)(\delta_{va} - \pi_v).$$

Splitting the sum gives

$$\left.\frac{dH}{d\eta}\right|_{\eta=0} = \hat{A}\left[\sum_{v \in \mathcal{V}} \pi_v(\mathfrak{s}_v - H)\delta_{va} - \sum_{v \in \mathcal{V}} \pi_v^2(\mathfrak{s}_v - H)\right].$$

Only the term $v = a$ contributes to the first sum, so

$$\sum_{v \in \mathcal{V}} \pi_v(\mathfrak{s}_v - H)\delta_{va} = \pi_a(\mathfrak{s}_a - H) = p(\mathfrak{s}_a - H).$$

Using $\mathfrak{s}_v = -\ln \pi_v$, the second term becomes

$$-\sum_{v \in \mathcal{V}} \pi_v^2(\mathfrak{s}_v - H) = \sum_{v \in \mathcal{V}} \pi_v^2(-\mathfrak{s}_v + H) = \sum_{v \in \mathcal{V}} \pi_v^2(\ln \pi_v + H) = S_2.$$

Hence

$$\left.\frac{dH}{d\eta}\right|_{\eta=0} = \hat{A}\big[p(\mathfrak{s}_a - H) + S_2\big].$$

Substituting $\mathfrak{s}_a = -\ln p$ yields

$$p(\mathfrak{s}_a - H) + S_2 = p(-\ln p - H) + S_2 = -\big[p(\ln p + H) - S_2\big] = -\Phi(p).$$

Therefore,

$$\left.\frac{dH}{d\eta}\right|_{\eta=0} = -\hat{A}\,\Phi(p).$$

This proves the result. □

E Complete Proofs for Sections 3.2 and 3.3 and Near-Criticality Analysis

E.1 Positivity of S_2 under Non-Uniform Distributions

Lemma E.1 (Positivity of S_2 under non-uniform distributions)**.** *If the conditional distribution π is non-uniform, then*

$$S_2 = \sum_{v\in\mathcal{V}} \pi_v^2(\ln \pi_v + H) > 0.$$

If π is uniform, then $S_2 = 0$.

Proof. Let $V \sim \pi$ and define

$$X \triangleq \pi_V.$$

Then

$$\begin{aligned}\mathbb{E}[\ln X + H] &= \sum_{v\in\mathcal{V}} \pi_v(\ln \pi_v + H)\\ &= \sum_v \pi_v \ln \pi_v + H\sum_v \pi_v\\ &= -H + H\\ &= 0.\end{aligned}$$

Moreover,

$$\begin{aligned}S_2 &= \sum_v \pi_v^2(\ln \pi_v + H)\\ &= \mathbb{E}\big[X(\ln X + H)\big].\end{aligned}$$

Since $\mathbb{E}[\ln X + H] = 0$, we obtain

$$\begin{aligned}S_2 &= \mathbb{E}\big[X(\ln X + H)\big]\\ &\quad - \mathbb{E}[X]\mathbb{E}[\ln X + H]\\ &= \mathrm{Cov}(X, \ln X + H)\\ &= \mathrm{Cov}(X, \ln X).\end{aligned}$$

Using the symmetric form of covariance, and letting X' be an independent copy of X, we have

$$\mathrm{Cov}(X, \ln X) = \frac{1}{2}\mathbb{E}\big[(X - X')(\ln X - \ln X')\big].$$

Therefore,

$$S_2 = \frac{1}{2}\sum_{u,v\in\mathcal{V}} \pi_u\pi_v(\pi_u - \pi_v)(\ln \pi_u - \ln \pi_v).$$

Because the logarithm is strictly increasing, for any $a, b > 0$,

$$(a - b)(\ln a - \ln b) \geq 0,$$

with equality if and only if $a = b$. Hence every term in the summation above is nonnegative.

If π is non-uniform, there exist u, v such that $\pi_u \neq \pi_v$. In the softmax setting considered in this paper, all token probabilities are strictly positive, so the corresponding weight $\pi_u \pi_v$ is also strictly positive, and the associated summand is strictly positive. Thus the full sum is strictly positive, which gives $S_2 > 0$.

If π is uniform, then $\pi_u = \pi_v$ for all u, v, so every summand is zero and $S_2 = 0$. □

E.2 $H > S_2$ under Non-Degenerate Distributions

Lemma E.2 ($H > S_2$ under non-degenerate distributions)**.** *If the distribution π is non-degenerate, namely it is not a point mass on a single token, then*

$$H > S_2.$$

Proof. By definition,

$$\begin{aligned} S_2 &= \sum_{v \in \mathcal{V}} \pi_v^2 (\ln \pi_v + H) \\ &= \sum_v \pi_v^2 \ln \pi_v + H \sum_v \pi_v^2. \end{aligned}$$

Hence

$$H - S_2 = H - \sum_v \pi_v^2 \ln \pi_v - H \sum_v \pi_v^2.$$

Collecting the terms that contain H yields

$$H - S_2 = H\Big(1 - \sum_v \pi_v^2\Big) - \sum_v \pi_v^2 \ln \pi_v.$$

Equivalently,

$$H - S_2 = H\Big(1 - \sum_v \pi_v^2\Big) + \sum_v \pi_v^2 (-\ln \pi_v).$$

We now inspect the two terms. Since π is non-degenerate, at least two tokens have positive probability. Therefore,

$$\sum_v \pi_v^2 < \Big(\sum_v \pi_v\Big)^2 = 1,$$

which implies

$$1 - \sum_v \pi_v^2 > 0.$$

The Shannon entropy of a non-degenerate distribution is strictly positive, so

$$H\Big(1 - \sum_v \pi_v^2\Big) > 0.$$

For the second term, $0 < \pi_v \leq 1$ in the softmax setting, and hence

$$-\ln \pi_v \geq 0.$$

Since $\pi_v^2 \geq 0$, each term satisfies

$$\pi_v^2 (-\ln \pi_v) \geq 0.$$

Thus

$$\sum_v \pi_v^2 (-\ln \pi_v) \geq 0.$$

Combining a strictly positive term with a nonnegative term gives

$$H - S_2 > 0,$$

and therefore

$$H > S_2.$$

□

E.3 Proposition 3.2 (Uniqueness of the Critical Surprisal Threshold)

For any non-uniform and non-degenerate distribution π, there exists a unique

$$p^* \in (e^{-H}, 1), \qquad \mathfrak{s}^* \triangleq -\ln p^* \in (0, H),$$

such that

$$\Phi(p^*) = 0,$$

and

$$\Phi(p) > 0 \iff p > p^* \iff \mathfrak{s}_a < \mathfrak{s}^*.$$

Proof. By definition,

$$\Phi(p) = p(\ln p + H) - S_2, \qquad p \in (0, 1].$$

We first determine the endpoint signs. As $p \to 0^+$, the standard limit $p \ln p \to 0$ gives

$$\Phi(0^+) = -S_2 < 0,$$

where the inequality follows from Lemma E.1. At $p = 1$,

$$\Phi(1) = H - S_2 > 0,$$

where the inequality follows from Lemma E.2.

Next, we show that Φ is strictly increasing on $[e^{-H}, 1]$. Differentiating gives

$$\begin{aligned}\Phi'(p) &= \frac{d}{dp}\left[p(\ln p + H) - S_2\right] \\ &= \ln p + H + 1.\end{aligned}$$

For any $p \in [e^{-H}, 1]$, we have $\ln p \geq -H$, and hence

$$\Phi'(p) \geq -H + H + 1 = 1 > 0.$$

Therefore, Φ is strictly increasing on $[e^{-H}, 1]$.

Moreover,

$$\begin{aligned}\Phi(e^{-H}) &= e^{-H}(-H + H) - S_2 \\ &= -S_2 < 0,\end{aligned}$$

whereas

$$\Phi(1) = H - S_2 > 0.$$

By continuity and the intermediate value theorem, there exists at least one

$$p^* \in (e^{-H}, 1)$$

such that $\Phi(p^*) = 0$. Since Φ is strictly increasing on this interval, the zero is unique.

Define

$$\mathfrak{s}^* \triangleq -\ln p^*.$$

Because $p^* \in (e^{-H}, 1)$, applying the negative logarithm gives

$$0 < -\ln p^* < H,$$

or equivalently,

$$\mathfrak{s}^* \in (0, H).$$

Finally, the strict monotonicity of Φ implies

$$p > p^* \iff \Phi(p) > 0.$$

Since $-\ln(\cdot)$ is strictly decreasing on $(0, 1]$,

$$p > p^* \iff -\ln p < -\ln p^* \iff \mathfrak{s}_a < \mathfrak{s}^*.$$

Combining the two equivalences yields

$$\Phi(p) > 0 \iff p > p^* \iff \mathfrak{s}_a < \mathfrak{s}^*.$$

□

E.4 Corollary 3.3 (Four-Quadrant Decomposition)

The sign of the first-order entropy variation at a single token position is jointly determined by $(\operatorname{sign} \hat{A}, \mathbf{1}[\mathfrak{s}_a < \mathfrak{s}^*])$. When $\hat{A} > 0$ and $\mathfrak{s}_a < \mathfrak{s}^*$, entropy decreases. When $\hat{A} > 0$ and $\mathfrak{s}_a > \mathfrak{s}^*$, entropy increases. When $\hat{A} < 0$ and $\mathfrak{s}_a < \mathfrak{s}^*$, entropy increases. When $\hat{A} < 0$ and $\mathfrak{s}_a > \mathfrak{s}^*$, entropy decreases.

Proof. By Theorem 3.1,

$$\left.\frac{dH}{d\eta}\right|_{\eta=0} = -\hat{A}\Phi(p).$$

Thus the sign of the first-order entropy variation is exactly the sign of $-\hat{A}\Phi(p)$. Proposition 3.2 gives

$$\mathfrak{s}_a < \mathfrak{s}^* \iff \Phi(p) > 0, \qquad \mathfrak{s}_a > \mathfrak{s}^* \iff \Phi(p) < 0.$$

If $\hat{A} > 0$ and $\mathfrak{s}_a < \mathfrak{s}^*$, then $\Phi(p) > 0$, so $-\hat{A}\Phi(p) < 0$ and entropy decreases. If $\hat{A} > 0$ and $\mathfrak{s}_a > \mathfrak{s}^*$, then $\Phi(p) < 0$, so $-\hat{A}\Phi(p) > 0$ and entropy increases. If $\hat{A} < 0$ and $\mathfrak{s}_a < \mathfrak{s}^*$, then $\Phi(p) > 0$, so $-\hat{A}\Phi(p) > 0$ and entropy increases. If $\hat{A} < 0$ and $\mathfrak{s}_a > \mathfrak{s}^*$, then $\Phi(p) < 0$, so $-\hat{A}\Phi(p) < 0$ and entropy decreases. At the boundary $\mathfrak{s}_a = \mathfrak{s}^*$, the first-order entropy variation is zero. □

E.5 Asymmetric Entropy Contributions under Shared Trajectory-Level Advantages

This subsection provides the detailed quantitative analysis of the asymmetric entropy contributions summarized in Section 3.2.

Consider the subset of trajectories assigned positive advantages ($\hat{A}_i > 0$). Because rollouts are sampled from the current policy π_θ, the probability of drawing a low-surprisal token at each decoding position is higher than that of drawing a high-surprisal token. Within positive-advantage trajectories, entropy-decreasing tokens (low surprisal, $\Phi > 0$) therefore constitute the statistical majority by sampling frequency, while entropy-increasing tokens (high surprisal, $\Phi < 0$) remain a statistical minority. Denoting $p_{i,t} = \pi_\theta(o_{i,t} \mid x_i, o_{i,<t})$ and $\mathfrak{s}_{i,t} = -\ln p_{i,t}$, the net first-order entropy contribution from positive-advantage samples decomposes as

$$\left.\frac{d\bar{H}^+}{d\eta}\right|_{\eta=0} = \frac{1}{N}\Bigg[\underbrace{-\sum_{\substack{i,t:\,\hat{A}_i>0 \\ \mathfrak{s}_{i,t}<\mathfrak{s}^*_{i,t}}} \hat{A}_i\,\Phi_{i,t}}_{\text{entropy-decreasing majority } (<0)} + \underbrace{\sum_{\substack{i,t:\,\hat{A}_i>0 \\ \mathfrak{s}_{i,t}>\mathfrak{s}^*_{i,t}}} \hat{A}_i\,|\Phi_{i,t}|}_{\text{entropy-increasing minority } (>0)} \Bigg].$$

During GRPO training, all tokens within a trajectory share a single trajectory-level advantage $\hat{A}_i$, leaving the algorithm unable to distinguish the opposing entropy effects of these two token categories. This observation reveals a fundamental gradient-level mechanism underlying entropy collapse in GRPO. Under the shared trajectory-level advantage, the reinforced low-surprisal majority systematically drives the distribution toward concentration, with entropy-decreasing contributions dominating in expectation. In contrast, the high-surprisal minority that could preserve diversity contributes limited entropy-increasing effects. Upon aggregation over the batch, the net outcome is a systematic reduction in entropy. An analogous but mirror-image asymmetry governs the negative-advantage subset.

E.6 Theorem 3.4 (Entropy Neutrality Identity)

For any conditional distribution π,

$$\mathbb{E}_{a\sim\pi}[\Phi(a)] = \sum_{v\in\mathcal{V}} \pi_v \Phi(\pi_v) = 0.$$

Proof. Here $\Phi(a)$ denotes the value obtained by first sampling $a \sim \pi$ and then evaluating $\Phi(p)$ at $p = \pi(a)$. Therefore,

$$\mathbb{E}_{a\sim\pi}[\Phi(a)] = \sum_{v\in\mathcal{V}} \pi_v \Phi(\pi_v).$$

Substituting the definition of Φ gives

$$\sum_{v\in\mathcal{V}} \pi_v \Phi(\pi_v) = \sum_v \pi_v \big[\pi_v(\ln \pi_v + H) - S_2\big].$$

Expanding the right-hand side yields

$$\sum_v \pi_v^2(\ln \pi_v + H) - S_2 \sum_v \pi_v.$$

By the definition of S_2,

$$\sum_v \pi_v^2(\ln \pi_v + H) = S_2,$$

and by normalization,

$$\sum_v \pi_v = 1.$$

Thus

$$\sum_{v\in\mathcal{V}} \pi_v \Phi(\pi_v) = S_2 - S_2 = 0,$$

which proves

$$\mathbb{E}_{a\sim\pi}[\Phi(a)] = 0.$$

□

E.7 Proposition 3.5 (Entropy Gradient under Token-Level Reweighting)

Let

$$\mathcal{L}^+ = \{(i,t) : \hat{A}_i > 0,\ \mathfrak{s}_{i,t} > \mathfrak{s}^*_{i,t}\}.$$

Suppose the effective advantage of every token in $\mathcal{L}^+$ is multiplied by a factor $W \geq 1$, while all other token positions keep unit weight. Then the first-order variation of the batch-averaged entropy satisfies

$$\left.\frac{d\bar{H}}{d\eta}\right|_W = -\frac{1}{N}\Big[\Lambda - (W-1)\Gamma\Big],$$

where

$$\Lambda \triangleq \sum_{i,t} \hat{A}_i \Phi_{i,t}, \qquad \Gamma \triangleq \sum_{(i,t)\in\mathcal{L}^+} \hat{A}_i |\Phi_{i,t}| > 0.$$

The critical weight is

$$W^* = 1 + \frac{\Lambda}{\Gamma}.$$

The batch-level net entropy variation is positive when $W > W^*$ and negative when $W < W^*$.

Proof. Under token-level reweighting, the policy-gradient contribution at position (i,t) is linearly scaled by a positive weight $\omega_{i,t}$. Since the first-order entropy derivative in Theorem 3.1 is linear in the policy-gradient direction, the corresponding entropy contribution becomes

$$-\omega_{i,t}\hat{A}_i\Phi_{i,t}.$$

Hence the batch-averaged first-order entropy variation is

$$\left.\frac{d\bar{H}}{d\eta}\right|_\omega = -\frac{1}{N}\sum_{i,t} \omega_{i,t}\hat{A}_i\Phi_{i,t}.$$

In the present reweighting scheme,

$$\omega_{i,t} = \begin{cases} W, & (i,t)\in\mathcal{L}^+, \\ 1, & \text{otherwise.} \end{cases}$$

Therefore,

$$\left.\frac{d\bar{H}}{d\eta}\right|_W = -\frac{1}{N}\left[\sum_{(i,t)\notin\mathcal{L}^+} \hat{A}_i\Phi_{i,t} + W\sum_{(i,t)\in\mathcal{L}^+} \hat{A}_i\Phi_{i,t}\right].$$

Equivalently,

$$\left.\frac{d\bar{H}}{d\eta}\right|_W = -\frac{1}{N}\left[\sum_{i,t}\hat{A}_i\Phi_{i,t} + (W-1)\sum_{(i,t)\in\mathcal{L}^+}\hat{A}_i\Phi_{i,t}\right].$$

Define

$$\Lambda \triangleq \sum_{i,t}\hat{A}_i\Phi_{i,t}.$$

For any $(i,t)\in\mathcal{L}^+$, we have $\hat{A}_i > 0$ and $\mathfrak{s}_{i,t} > \mathfrak{s}^*_{i,t}$. By Proposition 3.2, $\Phi_{i,t} < 0$. Hence

$$\hat{A}_i\Phi_{i,t} = -\hat{A}_i|\Phi_{i,t}|.$$

It follows that

$$\sum_{(i,t)\in\mathcal{L}^+}\hat{A}_i\Phi_{i,t} = -\sum_{(i,t)\in\mathcal{L}^+}\hat{A}_i|\Phi_{i,t}|.$$

Define

$$\Gamma \triangleq \sum_{(i,t)\in\mathcal{L}^+}\hat{A}_i|\Phi_{i,t}|.$$

When $\mathcal{L}^+$ is nonempty, $\Gamma > 0$ because $\hat{A}_i > 0$ and $|\Phi_{i,t}| > 0$ on this set. The expression above becomes

$$\left.\frac{d\bar{H}}{d\eta}\right|_W = -\frac{1}{N}\Big[\Lambda - (W-1)\Gamma\Big].$$

Let

$$W^* \triangleq 1 + \frac{\Lambda}{\Gamma}.$$

Then

$$\Lambda - (W-1)\Gamma = \Gamma(W^* - W).$$

Since $\Gamma > 0$,

$$\begin{aligned}\operatorname{sign}\left(\left.\frac{d\bar{H}}{d\eta}\right|_W\right) &= -\operatorname{sign}(W^* - W)\\ &= \operatorname{sign}(W - W^*).\end{aligned}$$

Thus $d\bar{H}/d\eta\big|_W > 0$ when $W > W^*$, $d\bar{H}/d\eta\big|_W < 0$ when $W < W^*$, and the first-order entropy variation is zero when $W = W^*$. □

E.8 Two Auxiliary Bounds: Lower Bound in the High-Surprisal Region and Upper Bound in the Low-Surprisal Region

E.8.1 Pointwise Lower Bound in the High-Surprisal Region

Lemma E.3 (Pointwise lower bound in the high-surprisal region)**.** *If a token satisfies*

$$\mathfrak{s} \geq H,$$

then

$$|\Phi(p)| \geq S_2.$$

Proof. The condition $\mathfrak{s} \geq H$ is equivalent to

$$-\ln p \geq H \iff \ln p + H \leq 0.$$

Since $p > 0$, multiplying both sides by p gives

$$p(\ln p + H) \leq 0.$$

Therefore,

$$\Phi(p) = p(\ln p + H) - S_2 \leq -S_2 < 0.$$

Consequently,

$$|\Phi(p)| = -\Phi(p) = S_2 - p(\ln p + H) \geq S_2.$$

□

E.8.2 Pointwise Upper Bound in the Low-Surprisal Region

Lemma E.4 (Pointwise upper bound in the low-surprisal region)**.** *If a token satisfies*

$$\mathfrak{s} < \mathfrak{s}^*,$$

then

$$|\Phi(p)| \le H - S_2.$$

Proof. By Proposition 3.2,

$$\mathfrak{s} < \mathfrak{s}^* \iff \Phi(p) > 0.$$

Therefore,

$$|\Phi(p)| = \Phi(p) = p(\ln p + H) - S_2.$$

Since $0 < p \le 1$, we have $\ln p \le 0$, and hence

$$p \ln p \le 0.$$

Also, $pH \le H$. Thus

$$p(\ln p + H) = p \ln p + pH \le H.$$

It follows that

$$|\Phi(p)| \le H - S_2.$$

□

E.9 Statistical Assumptions for the Near-Criticality Analysis

E.9.1 Single-Token Credit Dilution

Assumption E.1 (Single-token credit dilution)**.** *For long sequences and sufficiently large batches, define the conditional advantage function at any position (i,t) as*

$$g_{i,t}(v) \triangleq \mathbb{E}[\hat{A}_i \mid c_{i,t},\ o_{i,t} = v].$$

There exists a constant $C_g > 0$, independent of the sequence length T, such that for every position and every $u, v \in \mathcal{V}$,

$$|g_{i,t}(u) - g_{i,t}(v)| \le \frac{C_g}{T}.$$

This assumption states that, along a trajectory of length T, the marginal effect of any single token on the trajectory-level advantage is diluted at rate $1/T$.

E.9.2 Non-Degenerate Absolute Scale of the Advantage

Assumption E.2 (Non-degenerate absolute scale of the advantage)**.** *There exists a constant $a_- > 0$ such that, for every position (i,t) and every candidate token v,*

$$\mathbb{E}\big[|\hat{A}_i| \mid c_{i,t},\ o_{i,t} = v\big] \ge a_-.$$

This assumption states that the conditional absolute scale of the normalized advantage remains $O(1)$ rather than degenerating to zero.

E.9.3 Non-Vanishing Mass of a Strong High-Surprisal Positive-Advantage Subset

Assumption E.3 (Non-vanishing mass of a strong high-surprisal positive-advantage subset)**.** *Define*

$$\mathcal{H}^+ \triangleq \{(i,t) : \hat{A}_i > 0,\ \mathfrak{s}_{i,t} \ge H_{i,t}\}.$$

There exist constants $\rho_H > 0$ and $c_H > 0$ such that, for sufficiently large batches, with high probability,

$$|\mathcal{H}^+| \ge \rho_H N,$$

and

$$\frac{1}{|\mathcal{H}^+|} \sum_{(i,t)\in\mathcal{H}^+} \hat{A}_i S_{2,i,t} \ge c_H.$$

E.9.4 Linear Scale of Total Absolute Entropy Sensitivity

Assumption E.4 (Linear scale of total absolute entropy sensitivity)**.** *Let*

$$\Sigma_{\mathrm{abs}} \triangleq \sum_{i,t} |\hat{A}_i|\,|\Phi_{i,t}|.$$

There exists a constant $c_\Sigma < \infty$ such that, for sufficiently large batches, with high probability,

$$\Sigma_{\mathrm{abs}} \le c_\Sigma N.$$

E.10 Λ Is a Weak Residual

Lemma E.5 (Λ is a weak residual)**.** *Under Assumptions E.1 and E.2, and under standard large-batch concentration with uniformly bounded second moments,*

$$\frac{|\Lambda|}{\Sigma_{\mathrm{abs}}} = O(T^{-1}), \qquad \Lambda \triangleq \sum_{i,t} \hat{A}_i \Phi_{i,t}.$$

Proof. The proof first establishes the $O(T^{-1})$ scaling at the population level and then transfers it to empirical batch quantities by concentration.

Fix a position (i,t) and condition on $c_{i,t} = c$. Define

$$\Phi_c(v) \triangleq \pi_\theta(v \mid c)\big(\ln \pi_\theta(v \mid c) + H(c)\big) - S_2(c).$$

When the sampled token is $o_{i,t}$,

$$\Phi_{i,t} = \Phi_c(o_{i,t}).$$

By the law of total expectation,

$$\mathbb{E}[\hat{A}_i \Phi_{i,t}] = \mathbb{E}\Big[\mathbb{E}[\hat{A}_i \Phi_{i,t} \mid c_{i,t}]\Big].$$

Given $c_{i,t} = c$, the token $o_{i,t}$ is sampled from $\pi_\theta(\cdot \mid c)$. Thus

$$\mathbb{E}[\hat{A}_i \Phi_{i,t} \mid c_{i,t} = c] = \sum_{v \in \mathcal{V}} \pi_\theta(v \mid c)\mathbb{E}[\hat{A}_i \mid c_{i,t} = c,\ o_{i,t} = v]\Phi_c(v).$$

Let

$$g_c(v) \triangleq \mathbb{E}[\hat{A}_i \mid c_{i,t} = c,\ o_{i,t} = v].$$

Then

$$\mathbb{E}[\hat{A}_i \Phi_{i,t} \mid c_{i,t} = c] = \sum_v \pi_\theta(v \mid c) g_c(v) \Phi_c(v).$$

We now use the Entropy Neutrality Identity to remove the mean component of g_c. Define

$$\bar{g}_c \triangleq \sum_v \pi_\theta(v \mid c) g_c(v).$$

By Theorem 3.4,

$$\sum_v \pi_\theta(v \mid c)\Phi_c(v) = 0.$$

Therefore,

$$\sum_v \pi_\theta(v \mid c) g_c(v) \Phi_c(v) = \sum_v \pi_\theta(v \mid c)\big(g_c(v) - \bar{g}_c\big)\Phi_c(v),$$

which implies

$$\mathbb{E}[\hat{A}_i \Phi_{i,t} \mid c_{i,t} = c] = \sum_v \pi_\theta(v \mid c)\big(g_c(v) - \bar{g}_c\big)\Phi_c(v).$$

Since $\bar{g}_c$ is a $\pi_\theta(\cdot \mid c)$-weighted average of $\{g_c(v)\}_{v \in \mathcal{V}}$, for every v,

$$|g_c(v) - \bar{g}_c| \le \max_{u,v'} |g_c(u) - g_c(v')|.$$

Assumption E.1 gives

$$|g_c(v) - \bar{g}_c| \le \frac{C_g}{T}.$$

Consequently,

$$\left|\mathbb{E}[\hat{A}_i \Phi_{i,t} \mid c_{i,t} = c]\right| \le \frac{C_g}{T} \sum_v \pi_\theta(v \mid c) |\Phi_c(v)|.$$

Define

$$\Psi(c) \triangleq \sum_v \pi_\theta(v \mid c) |\Phi_c(v)|.$$

Then

$$\left|\mathbb{E}[\hat{A}_i \Phi_{i,t} \mid c_{i,t} = c]\right| \le \frac{C_g}{T} \Psi(c).$$

Summing over all token positions yields

$$|\mathbb{E}[\Lambda]| = \left|\sum_{i,t} \mathbb{E}[\hat{A}_i \Phi_{i,t}]\right| \le \sum_{i,t} \left|\mathbb{E}[\hat{A}_i \Phi_{i,t}]\right| \le \frac{C_g}{T} \sum_{i,t} \mathbb{E}[\Psi(c_{i,t})].$$

We next lower bound $\mathbb{E}[\Sigma_{\text{abs}}]$. Conditional on $c_{i,t} = c$,

$$\mathbb{E}\big[|\hat{A}_i|\,|\Phi_{i,t}| \mid c_{i,t} = c\big] = \sum_v \pi_\theta(v \mid c) \mathbb{E}\big[|\hat{A}_i| \mid c_{i,t} = c,\ o_{i,t} = v\big] |\Phi_c(v)|.$$

By Assumption E.2,

$$\mathbb{E}\big[|\hat{A}_i| \mid c_{i,t} = c,\ o_{i,t} = v\big] \ge a_-.$$

Therefore,

$$\mathbb{E}\big[|\hat{A}_i|\,|\Phi_{i,t}| \mid c_{i,t} = c\big] \ge a_- \sum_v \pi_\theta(v \mid c) |\Phi_c(v)| = a_- \Psi(c).$$

Summing over positions and taking expectations gives

$$\mathbb{E}[\Sigma_{\text{abs}}] = \sum_{i,t} \mathbb{E}\big[|\hat{A}_i|\,|\Phi_{i,t}|\big] \ge a_- \sum_{i,t} \mathbb{E}[\Psi(c_{i,t})].$$

Combining the upper bound on $|\mathbb{E}[\Lambda]|$ and the lower bound on $\mathbb{E}[\Sigma_{\text{abs}}]$ gives

$$\frac{|\mathbb{E}[\Lambda]|}{\mathbb{E}[\Sigma_{\text{abs}}]} \le \frac{C_g}{a_- T}.$$

Thus, at the population level,

$$\frac{|\mathbb{E}[\Lambda]|}{\mathbb{E}[\Sigma_{\text{abs}}]} = O(T^{-1}).$$

For sufficiently large batches, and assuming uniformly bounded second moments of the corresponding summands, Λ and Σ_{abs} concentrate around their population values at a scale that does not change the dependence on T. Hence, with high probability,

$$\frac{|\Lambda|}{\Sigma_{\text{abs}}} = O(T^{-1}).$$

□

E.11 Theorem 3.6 (Near-Criticality)

If Assumptions E.1–E.4 hold and both the sequence length T and the batch size are sufficiently large, then

$$W^* - 1 = \frac{\Lambda}{\Gamma} = O(T^{-1}).$$

Proof. By Proposition 3.5,

$$W^* - 1 = \frac{\Lambda}{\Gamma}.$$

It remains to control the numerator and denominator.

For the numerator, Lemma E.5 gives

$$\frac{|\Lambda|}{\Sigma_{\text{abs}}} = O(T^{-1}).$$

For the denominator, we show that Γ is lower bounded by a constant-order fraction of the total absolute entropy sensitivity. First, $\mathcal{H}^+ \subset \mathcal{L}^+$. Indeed, Proposition 3.2 gives, at every position,

$$\mathfrak{s}^*_{i,t} < H_{i,t}.$$

Therefore, any token satisfying $\mathfrak{s}_{i,t} \geq H_{i,t}$ also satisfies

$$\mathfrak{s}_{i,t} > \mathfrak{s}^*_{i,t}.$$

Together with $\hat{A}_i > 0$, this implies

$$\mathcal{H}^+ \subset \mathcal{L}^+.$$

By the definition of Γ and the inclusion above,

$$\Gamma = \sum_{(i,t)\in\mathcal{L}^+} \hat{A}_i |\Phi_{i,t}| \geq \sum_{(i,t)\in\mathcal{H}^+} \hat{A}_i |\Phi_{i,t}|.$$

On $\mathcal{H}^+$, the condition $\mathfrak{s}_{i,t} \geq H_{i,t}$ holds. Lemma E.3 therefore gives

$$|\Phi_{i,t}| \geq S_{2,i,t}.$$

Hence

$$\Gamma \geq \sum_{(i,t)\in\mathcal{H}^+} \hat{A}_i S_{2,i,t}.$$

By Assumption E.3,

$$\frac{1}{|\mathcal{H}^+|} \sum_{(i,t)\in\mathcal{H}^+} \hat{A}_i S_{2,i,t} \geq c_H, \qquad |\mathcal{H}^+| \geq \rho_H N.$$

Therefore,

$$\Gamma \geq c_H |\mathcal{H}^+| \geq c_H \rho_H N.$$

Assumption E.4 gives

$$\Sigma_{\text{abs}} \leq c_\Sigma N.$$

Consequently,

$$\frac{\Sigma_{\text{abs}}}{\Gamma} \leq \frac{c_\Sigma N}{c_H \rho_H N} = \frac{c_\Sigma}{c_H \rho_H} = O(1).$$

We now decompose the critical offset as

$$\frac{\Lambda}{\Gamma} = \frac{\Lambda}{\Sigma_{\text{abs}}} \cdot \frac{\Sigma_{\text{abs}}}{\Gamma}.$$

Taking absolute values gives

$$\left|\frac{\Lambda}{\Gamma}\right| = \frac{|\Lambda|}{\Sigma_{\text{abs}}} \cdot \frac{\Sigma_{\text{abs}}}{\Gamma}.$$

The first factor is $O(T^{-1})$ by Lemma E.5, and the second factor is $O(1)$ by the bound above. Hence

$$\left|\frac{\Lambda}{\Gamma}\right| = O(T^{-1}).$$

Equivalently,

$$W^* - 1 = \frac{\Lambda}{\Gamma} = O(T^{-1}).$$

□

In the entropy-collapse regime studied in the main text, baseline GRPO corresponds to $\Lambda > 0$. Hence $W^* > 1$, and the critical weight exceeds one only by an $O(T^{-1})$ offset. The proof lower bounds Γ using the stronger subset $\mathcal{H}^+$ rather than the full entropy-increasing positive-advantage set $\mathcal{L}^+$, so the argument is conservative. In actual training, additional contributions from $\mathcal{L}^+ \setminus \mathcal{H}^+$ further enlarge Γ and make the critical offset closer to zero.

F Formalization of the Cross-Step Entropy Dynamics in Section 3.4

This appendix casts the dynamical discussion in Section 3.4 as a discrete-time system under a mean-field closure. The purpose is to formalize the qualitative mechanism described in the main text. It is not intended as an unconditional global convergence theorem for general deep neural network training.

F.1 Mean-Field Approximation

Assume that, within the training interval of interest, the batch-level quantities Λ_k and Γ_k are primarily determined by the current batch-averaged policy entropy $\bar{H}_k$. More precisely, suppose that there exist continuous functions $\Lambda(h)$ and $\Gamma(h)$ such that

$$\Lambda_k \approx \Lambda(\bar{H}_k), \qquad \Gamma_k \approx \Gamma(\bar{H}_k), \qquad \Gamma(h) > 0.$$

The associated critical-weight curve is defined as

$$W^*(h) \triangleq 1 + \frac{\Lambda(h)}{\Gamma(h)}.$$

The statement in the main text that flatter policy distributions require a lower critical weight to sustain entropy growth is formalized through the following monotonicity condition:

$$W^*(h) \text{ is strictly decreasing in } h \text{ on the relevant interval.}$$

F.2 Sign of the One-Step Batch Entropy Change under a Fixed Weight

Proposition F.1 (Sign of the one-step batch entropy change under a fixed weight)**.** *For a fixed token-level reweighting factor W, the one-step change in batch-averaged policy entropy satisfies*

$$\Delta\bar{H}_k = -\frac{1}{N}\Big[\Lambda(\bar{H}_k) - (W-1)\Gamma(\bar{H}_k)\Big] = \frac{\Gamma(\bar{H}_k)}{N}\Big[W - W^*(\bar{H}_k)\Big].$$

Consequently,

$$\operatorname{sign}(\Delta\bar{H}_k) = \operatorname{sign}\big(W - W^*(\bar{H}_k)\big).$$

Proof. By the mean-field form of the batch-level entropy dynamics,

$$\Delta\bar{H}_k = -\frac{1}{N}\Big[\Lambda(\bar{H}_k) - (W-1)\Gamma(\bar{H}_k)\Big].$$

The bracketed term can be rewritten as

$$\Lambda(\bar{H}_k) - (W-1)\Gamma(\bar{H}_k) = \Gamma(\bar{H}_k)\left[\frac{\Lambda(\bar{H}_k)}{\Gamma(\bar{H}_k)} - (W-1)\right].$$

Since

$$W^*(\bar{H}_k) = 1 + \frac{\Lambda(\bar{H}_k)}{\Gamma(\bar{H}_k)},$$

we have

$$\frac{\Lambda(\bar{H}_k)}{\Gamma(\bar{H}_k)} - (W-1) = W^*(\bar{H}_k) - W.$$

Substituting this identity gives

$$\Delta\bar{H}_k = -\frac{\Gamma(\bar{H}_k)}{N}\big[W^*(\bar{H}_k) - W\big] = \frac{\Gamma(\bar{H}_k)}{N}\big[W - W^*(\bar{H}_k)\big].$$

Because $N > 0$ and $\Gamma(\bar{H}_k) > 0$, the sign of $\Delta\bar{H}_k$ is entirely determined by $W - W^*(\bar{H}_k)$. This proves the proposition. □

F.3 Open-Loop Self-Reinforcing Entropy Collapse and Recovery

Corollary F.2 (Open-loop self-reinforcing entropy collapse and recovery)**.** *Under the mean-field approximation above, the open-loop dynamics exhibit two symmetric feedback regimes. If $W = 1$ and the current entropy level satisfies $W^*(\bar{H}_k) > 1$, then Proposition F.1 gives $\Delta\bar{H}_k < 0$. Since $W^*(h)$ is strictly decreasing in h, the next step satisfies*

$$\bar{H}_{k+1} < \bar{H}_k \quad \Longrightarrow \quad W^*(\bar{H}_{k+1}) \geq W^*(\bar{H}_k).$$

Thus, the unit-weight GRPO baseline moves farther below the critical-weight curve, which strengthens the entropy-decreasing pressure and reinforces entropy collapse.

Conversely, if at some step $W > W^*(\bar{H}_k)$, then Proposition F.1 gives $\Delta\bar{H}_k > 0$. By the same monotonicity condition,

$$\bar{H}_{k+1} > \bar{H}_k \quad \Longrightarrow \quad W^*(\bar{H}_{k+1}) \leq W^*(\bar{H}_k).$$

Therefore, the margin by which the fixed weight exceeds the critical-weight curve increases after the entropy rises. The entropy-increasing pressure is consequently strengthened, which reinforces entropy recovery.

Proof. Both claims follow from Proposition F.1 and the monotonicity of $W^*(h)$. For the collapse regime, if $W = 1$ and $W^*(\bar{H}_k) > 1$, then $\Delta\bar{H}_k < 0$, so $\bar{H}_{k+1} < \bar{H}_k$. Since the critical-weight curve decreases with entropy, evaluating it at the smaller entropy value $\bar{H}_{k+1}$ yields a critical weight that is no smaller than before. The unit-weight baseline is therefore even farther below the critical threshold at the next step. The recovery regime follows by the same argument with the inequalities reversed. This proves the corollary. □

F.4 Local Stability of Batch-Level Target-Entropy Gating

Proposition F.3 (Local stability of batch-level target-entropy gating)**.** *Consider the batch-level target-entropy gating rule introduced in Section 4.3,*

$$g_k = \mathbf{1}[\bar{H}_k < H_{\mathrm{tgt}}].$$

For the one-sided STARE variant, the one-step change in batch-averaged policy entropy is

$$\Delta\bar{H}_k = -\frac{1}{N}\Big[\Lambda(\bar{H}_k) - g_k(W-1)\Gamma(\bar{H}_k)\Big].$$

Define the gated and ungated update maps as

$$F_{\mathrm{on}}(h) \triangleq -\frac{1}{N}\big[\Lambda(h) - (W-1)\Gamma(h)\big],$$

$$F_{\mathrm{off}}(h) \triangleq -\frac{1}{N}\Lambda(h).$$

Assume that there exists a target-entropy neighborhood

$$I = [H_{\mathrm{tgt}} - \delta,\ H_{\mathrm{tgt}} + \delta]$$

such that the gated update points toward higher entropy below the target, while the ungated update points toward lower entropy at and above the target:

$$F_{\mathrm{on}}(h) > 0, \qquad h \in [H_{\mathrm{tgt}} - \delta, H_{\mathrm{tgt}}),$$
$$F_{\mathrm{off}}(h) < 0, \qquad h \in [H_{\mathrm{tgt}}, H_{\mathrm{tgt}} + \delta].$$

Assume in addition that the one-step displacement inside this neighborhood is bounded by the neighborhood half-width:

$$\Delta_{\max} \triangleq \max\left\{ \sup_{h\in[H_{\mathrm{tgt}}-\delta,H_{\mathrm{tgt}})} F_{\mathrm{on}}(h),\ \sup_{h\in[H_{\mathrm{tgt}},H_{\mathrm{tgt}}+\delta]} |F_{\mathrm{off}}(h)| \right\} \leq \delta.$$

Then I is forward invariant for the induced discrete-time entropy dynamics. Equivalently, once $\bar{H}_k$ enters I, all subsequent batch-averaged entropy values remain inside this neighborhood. The closed-loop system therefore exhibits bounded oscillation around the target entropy, rather than continuing to drift monotonically away from it.

Proof. Within I, the assumed sign conditions imply

$$F_{\mathrm{on}}(h) > 0 \quad (h < H_{\mathrm{tgt}}), \qquad F_{\mathrm{off}}(h) < 0 \quad (h \geq H_{\mathrm{tgt}}).$$

The bounded-step condition gives $\Delta_{\max} \leq \delta$.

If $\bar{H}_k \in [H_{\mathrm{tgt}} - \delta, H_{\mathrm{tgt}})$, then $g_k = 1$, and hence

$$\bar{H}_{k+1} = \bar{H}_k + F_{\mathrm{on}}(\bar{H}_k).$$

Since $F_{\mathrm{on}}(\bar{H}_k) > 0$,

$$\bar{H}_{k+1} > \bar{H}_k \geq H_{\mathrm{tgt}} - \delta.$$

Moreover, $F_{\mathrm{on}}(\bar{H}_k) \leq \Delta_{\max} \leq \delta$, so

$$\bar{H}_{k+1} \leq \bar{H}_k + \delta < H_{\mathrm{tgt}} + \delta.$$

Therefore,

$$\bar{H}_{k+1} \in I.$$

If $\bar{H}_k \in [H_{\mathrm{tgt}}, H_{\mathrm{tgt}} + \delta]$, then $g_k = 0$, and hence

$$\bar{H}_{k+1} = \bar{H}_k + F_{\mathrm{off}}(\bar{H}_k).$$

Since $F_{\mathrm{off}}(\bar{H}_k) < 0$,

$$\bar{H}_{k+1} < \bar{H}_k \leq H_{\mathrm{tgt}} + \delta.$$

At the same time, $|F_{\mathrm{off}}(\bar{H}_k)| \leq \Delta_{\max} \leq \delta$, which gives

$$\bar{H}_{k+1} \geq \bar{H}_k - \delta \geq H_{\mathrm{tgt}} - \delta.$$

Thus,

$$\bar{H}_{k+1} \in I.$$

In both cases, one step of the dynamics maps I into itself. Hence I is forward invariant, and any trajectory that enters the target-entropy neighborhood remains there thereafter. This proves the proposition. □

G Single-Polarity Operations and Finer-Grained Closed-Loop Extensions

G.1 Definition of Surprisal-Quantile Proxy Sets

In Section 4.1 of the main text, we define

$$\mathcal{T}^+ = \{(i,t) : \hat{A}_i > 0\}, \qquad \mathcal{T}^- = \{(i,t) : \hat{A}_i < 0\}.$$

Within $\mathcal{T}^+$ and $\mathcal{T}^-$, token positions are sorted separately in descending order of surprisal,

$$\mathfrak{s}_{i,t} = -\ln \pi_\theta(o_{i,t} \mid x_i, o_{i,<t}),$$

and the top $P\%$ positions are selected to form

$$\mathcal{L}_q^+ \subset \mathcal{T}^+, \qquad \mathcal{L}_q^- \subset \mathcal{T}^-.$$

To cover all four single-polarity operations, we also define the corresponding low-surprisal proxy sets. Let $\mathcal{U}_q^+$ denote the bottom $P\%$ positions with the lowest surprisal in $\mathcal{T}^+$, and let $\mathcal{U}_q^-$ denote the bottom $P\%$ positions with the lowest surprisal in $\mathcal{T}^-$.

These four proxy sets approximate the four theoretical quadrants in the advantage-surprisal decomposition. The set $\mathcal{L}_q^+$ corresponds to positive-advantage high-surprisal positions and targets the entropy-increasing quadrant. The set $\mathcal{U}_q^+$ corresponds to positive-advantage low-surprisal positions and targets the entropy-decreasing quadrant. The set $\mathcal{U}_q^-$ corresponds to negative-advantage low-surprisal positions and targets the entropy-increasing quadrant. The set $\mathcal{L}_q^-$ corresponds to negative-advantage high-surprisal positions and targets the entropy-decreasing quadrant.

G.2 Position-Level Logit Update under STARE Reweighting

Proposition G.1 (Position-level logit update under STARE reweighting)**.** *Under the weighted clipped surrogate objective*

$$\mathcal{J}_{\mathrm{STARE}}(\theta) = \frac{1}{N}\sum_{i,t}\omega_{i,t}\min\Big(\rho_{i,t}(\theta)\hat{A}_i,\ \mathrm{clip}(\rho_{i,t}(\theta),1-\epsilon,1+\epsilon)\hat{A}_i\Big),$$

if position (i,t) lies in the unclipped regime, its effective logit update is

$$\Delta z_v = \eta\,\omega_{i,t}\hat{A}_i(\delta_{va}-\pi_v), \qquad v\in\mathcal{V}.$$

Since $\omega_{i,t}>0$ always holds, STARE never reverses any token-level policy-gradient direction. It only selectively rescales the magnitude of the original GRPO learning signal.

Proof. The proof follows the same argument as Proposition D.2. The only change is that the local surrogate term is multiplied by the positive weight $\omega_{i,t}$. Therefore, the corresponding logit gradient is scaled by $\omega_{i,t}$:

$$\frac{\partial}{\partial z_v}\big[\omega_{i,t}\hat{A}_i\log\pi_a\big] = \omega_{i,t}\hat{A}_i(\delta_{va}-\pi_v).$$

Taking one infinitesimal gradient ascent step gives the stated update. Because $\omega_{i,t}>0$, the policy-gradient direction is preserved. This proves the proposition. □

G.3 Exact Batch-Level Entropy Variation under Reweighting an Arbitrary Token Subset

Proposition G.2 (Exact batch-level entropy variation under reweighting an arbitrary token subset)**.** *Let $\mathcal{S}$ be an arbitrary subset of token positions. Apply a uniform weight $r>0$ to positions in $\mathcal{S}$ and unit weight to all remaining positions:*

$$\omega_{i,t} = \begin{cases} r, & (i,t)\in\mathcal{S},\\ 1, & \textit{otherwise}. \end{cases}$$

Then, in the unclipped regime,

$$\left.\frac{d\bar{H}}{d\eta}\right|_{\mathcal{S},r} = -\frac{1}{N}\Big[\Lambda+(r-1)\Xi(\mathcal{S})\Big],$$

where

$$\Lambda \triangleq \sum_{i,t}\hat{A}_i\Phi_{i,t}, \qquad \Xi(\mathcal{S}) \triangleq \sum_{(i,t)\in\mathcal{S}}\hat{A}_i\Phi_{i,t}.$$

Proof. By Proposition G.1 and Theorem 3.1, the first-order entropy contribution at each position is linearly scaled by its token-level weight:

$$-\omega_{i,t}\hat{A}_i\Phi_{i,t}.$$

Thus,

$$\left.\frac{d\bar{H}}{d\eta}\right|_{\omega} = -\frac{1}{N}\sum_{i,t}\omega_{i,t}\hat{A}_i\Phi_{i,t}.$$

Substituting the single-subset reweighting form gives

$$\sum_{i,t}\omega_{i,t}\hat{A}_i\Phi_{i,t} = \sum_{(i,t)\notin\mathcal{S}}\hat{A}_i\Phi_{i,t} + r\sum_{(i,t)\in\mathcal{S}}\hat{A}_i\Phi_{i,t}.$$

Rearranging the expression into the unweighted batch sum plus the reweighting correction yields

$$\sum_{i,t}\omega_{i,t}\hat{A}_i\Phi_{i,t} = \sum_{i,t}\hat{A}_i\Phi_{i,t} + (r-1)\sum_{(i,t)\in\mathcal{S}}\hat{A}_i\Phi_{i,t} = \Lambda+(r-1)\Xi(\mathcal{S}).$$

Substitution completes the proof:

$$\left.\frac{d\bar{H}}{d\eta}\right|_{\mathcal{S},r} = -\frac{1}{N}\Big[\Lambda+(r-1)\Xi(\mathcal{S})\Big].$$

□

G.4 Four Single-Polarity Operations

Operation O1 amplifies $\mathcal{L}_q^+$. Define

$$\omega_{i,t}^{(\mathrm{O1})} = \begin{cases} W_+, & (i,t) \in \mathcal{L}_q^+, \\ 1, & \text{otherwise}, \end{cases} \qquad W_+ > 1.$$

By Proposition G.2,

$$\left.\frac{d\bar{H}}{d\eta}\right|_{\mathrm{O1}} = -\frac{1}{N}\Big[\Lambda + (W_+ - 1)\Xi(\mathcal{L}_q^+)\Big].$$

When $\mathcal{L}_q^+$ is dominated by positive-advantage high-surprisal entropy-increasing positions, we have $\Xi(\mathcal{L}_q^+) < 0$. Define

$$\Gamma_{L^+} \triangleq -\Xi(\mathcal{L}_q^+) = \sum_{(i,t)\in\mathcal{L}_q^+} |\hat{A}_i \Phi_{i,t}|.$$

Then,

$$\left.\frac{d\bar{H}}{d\eta}\right|_{\mathrm{O1}} = -\frac{1}{N}\Big[\Lambda - (W_+ - 1)\Gamma_{L^+}\Big].$$

Operation O2 attenuates $\mathcal{U}_q^+$. Define

$$\omega_{i,t}^{(\mathrm{O2})} = \begin{cases} M_+, & (i,t) \in \mathcal{U}_q^+, \\ 1, & \text{otherwise}, \end{cases} \qquad 0 < M_+ < 1.$$

By Proposition G.2,

$$\left.\frac{d\bar{H}}{d\eta}\right|_{\mathrm{O2}} = -\frac{1}{N}\Big[\Lambda + (M_+ - 1)\Xi(\mathcal{U}_q^+)\Big].$$

When $\mathcal{U}_q^+$ is dominated by positive-advantage low-surprisal entropy-decreasing positions, we have $\Xi(\mathcal{U}_q^+) > 0$. Define

$$\Gamma_{U^+} \triangleq \Xi(\mathcal{U}_q^+) = \sum_{(i,t)\in\mathcal{U}_q^+} |\hat{A}_i \Phi_{i,t}|.$$

Then,

$$\left.\frac{d\bar{H}}{d\eta}\right|_{\mathrm{O2}} = -\frac{1}{N}\Big[\Lambda - (1 - M_+)\Gamma_{U^+}\Big].$$

Operation O3 amplifies $\mathcal{U}_q^-$. Define

$$\omega_{i,t}^{(\mathrm{O3})} = \begin{cases} W_-, & (i,t) \in \mathcal{U}_q^-, \\ 1, & \text{otherwise}, \end{cases} \qquad W_- > 1.$$

By Proposition G.2,

$$\left.\frac{d\bar{H}}{d\eta}\right|_{\mathrm{O3}} = -\frac{1}{N}\Big[\Lambda + (W_- - 1)\Xi(\mathcal{U}_q^-)\Big].$$

When $\mathcal{U}_q^-$ is dominated by negative-advantage low-surprisal entropy-increasing positions, we have $\Xi(\mathcal{U}_q^-) < 0$. Define

$$\Gamma_{U^-} \triangleq -\Xi(\mathcal{U}_q^-) = \sum_{(i,t)\in\mathcal{U}_q^-} |\hat{A}_i \Phi_{i,t}|.$$

Then,

$$\left.\frac{d\bar{H}}{d\eta}\right|_{\mathrm{O3}} = -\frac{1}{N}\Big[\Lambda - (W_- - 1)\Gamma_{U^-}\Big].$$

Operation O4 attenuates $\mathcal{L}_q^-$. Define

$$\omega_{i,t}^{(\mathrm{O4})} = \begin{cases} M_-, & (i,t) \in \mathcal{L}_q^-, \\ 1, & \text{otherwise}, \end{cases} \qquad 0 < M_- < 1.$$

By Proposition G.2,

$$\left.\frac{d\bar{H}}{d\eta}\right|_{\mathrm{O4}} = -\frac{1}{N}\Big[\Lambda + (M_- - 1)\Xi(\mathcal{L}_q^-)\Big].$$

When $\mathcal{L}_q^-$ is dominated by negative-advantage high-surprisal entropy-decreasing positions, we have $\Xi(\mathcal{L}_q^-) > 0$. Define

$$\Gamma_{L^-} \triangleq \Xi(\mathcal{L}_q^-) = \sum_{(i,t)\in\mathcal{L}_q^-} |\hat{A}_i \Phi_{i,t}|.$$

Then,

$$\left.\frac{d\bar{H}}{d\eta}\right|_{\mathrm{O4}} = -\frac{1}{N}\Big[\Lambda - (1 - M_-)\Gamma_{L^-}\Big].$$

G.5 Exact Entropy Variation under Unified Closed-Loop Gating

Proposition G.3 (Exact entropy variation under unified closed-loop gating)**.** *Let $g_{i,t} \in \{0, 1\}$ be an arbitrary binary gate. For the one-sided variant, define*

$$\omega_{i,t}^{(\mathrm{V1},g)} = 1 + g_{i,t}(W - 1)\mathbf{1}[(i,t) \in \mathcal{L}_q^+], \qquad W > 1.$$

Then,

$$\left.\frac{d\bar{H}}{d\eta}\right|_{\mathrm{V1},g} = -\frac{1}{N}\left[\Lambda + (W - 1)\sum_{i,t} g_{i,t}\mathbf{1}[(i,t) \in \mathcal{L}_q^+]\hat{A}_i\Phi_{i,t}\right].$$

If the activated proxy set is dominated by the corresponding entropy-increasing quadrant, the expression reduces to

$$\left.\frac{d\bar{H}}{d\eta}\right|_{\mathrm{V1},g} = -\frac{1}{N}\big[\Lambda - (W - 1)\Gamma_g^+\big],$$

where

$$\Gamma_g^+ \triangleq \sum_{i,t} g_{i,t}\mathbf{1}[(i,t) \in \mathcal{L}_q^+]|\hat{A}_i\Phi_{i,t}|.$$

For the two-sided variant, define

$$\omega_{i,t}^{(\mathrm{V2},g)} = 1 + g_{i,t}(W - 1)\mathbf{1}[(i,t) \in \mathcal{L}_q^+] - g_{i,t}(1 - M)\mathbf{1}[(i,t) \in \mathcal{L}_q^-], \qquad W > 1, \quad 0 < M < 1.$$

Then,

$$\left.\frac{d\bar{H}}{d\eta}\right|_{\mathrm{V2},g} = -\frac{1}{N}\left[\Lambda + (W - 1)\sum_{i,t} g_{i,t}\mathbf{1}[(i,t) \in \mathcal{L}_q^+]\hat{A}_i\Phi_{i,t} + (M - 1)\sum_{i,t} g_{i,t}\mathbf{1}[(i,t) \in \mathcal{L}_q^-]\hat{A}_i\Phi_{i,t}\right].$$

If the two activated proxy sets are respectively dominated by the corresponding entropy-increasing and entropy-decreasing quadrants, the expression reduces to

$$\left.\frac{d\bar{H}}{d\eta}\right|_{\mathrm{V2},g} = -\frac{1}{N}\big[\Lambda - (W - 1)\Gamma_g^+ - (1 - M)\Gamma_g^-\big],$$

where

$$\Gamma_g^- \triangleq \sum_{i,t} g_{i,t}\mathbf{1}[(i,t) \in \mathcal{L}_q^-]|\hat{A}_i\Phi_{i,t}|.$$

Proof. The result follows by substituting the corresponding gated weights into the general weighted entropy identity derived in the proof of Proposition G.2. In the one-sided case,

$$\omega_{i,t}^{(\mathrm{V1},g)} - 1 = g_{i,t}(W - 1)\mathbf{1}[(i,t) \in \mathcal{L}_q^+],$$

which gives

$$\left.\frac{d\bar{H}}{d\eta}\right|_{\mathrm{V1},g} = -\frac{1}{N}\left[\Lambda + (W - 1)\sum_{i,t} g_{i,t}\mathbf{1}[(i,t) \in \mathcal{L}_q^+]\hat{A}_i\Phi_{i,t}\right].$$

The two-sided case follows by the same linearity argument. When the gated proxy sets preserve the dominant signs of their target quadrants, the corresponding signed sums can be rewritten as sums of absolute contributions. This proves the proposition. □

G.6 Unified Instantiations of Three Closed-Loop Granularities

The gate $g_{i,t}$ in Proposition G.3 admits three closed-loop instantiations. For batch-level gating, define

$$g_{i,t}^{(\mathrm{batch})} = \mathbf{1}[\bar{H}_k < H_{\mathrm{tgt}}].$$

For sample-level gating, define

$$g_{i,t}^{(\mathrm{sample})} = \mathbf{1}[\bar{H}_i < H_{\mathrm{tgt}}], \qquad \bar{H}_i \triangleq \frac{1}{T_i}\sum_{t=1}^{T_i} H_{i,t}.$$

For token-level gating, define

$$g_{i,t}^{(\mathrm{token})} = \mathbf{1}[H_{i,t} < H_{\mathrm{tgt}}].$$

Because $\omega_{i,t} > 0$ holds in all three cases, every closed-loop variant preserves the direction of the original GRPO learning signal. The variants differ only in the granularity at which they modulate the local strength of the token-level update.

H Combined Reweighting Operations and Adaptive Weights

H.1 Exact Entropy Variation for Arbitrary Two-Subset Reweighting

Proposition H.1 (Exact entropy variation for arbitrary two-subset reweighting)**.** *Let $\mathcal{S}_1$ and $\mathcal{S}_2$ be two disjoint token subsets. We assign multiplicative token weights $r_1 > 0$ and $r_2 > 0$ to them, respectively, while all remaining token positions receive unit weight:*

$$\omega_{i,t} = \begin{cases} r_1, & (i,t) \in \mathcal{S}_1, \\ r_2, & (i,t) \in \mathcal{S}_2, \\ 1, & \textit{otherwise}. \end{cases}$$

Then, in the unclipped regime of the weighted GRPO surrogate,

$$\left.\frac{d\bar{H}}{d\eta}\right|_{\mathcal{S}_1,r_1;\mathcal{S}_2,r_2} = -\frac{1}{N}\Big[\Lambda + (r_1-1)\Xi(\mathcal{S}_1) + (r_2-1)\Xi(\mathcal{S}_2)\Big].$$

Proof. The proof follows directly from the same argument as Proposition G.2. By the linearity of multiplicative token weighting,

$$\sum_{i,t}\omega_{i,t}\hat{A}_i\Phi_{i,t} = \sum_{i,t}\hat{A}_i\Phi_{i,t} + (r_1-1)\sum_{(i,t)\in\mathcal{S}_1}\hat{A}_i\Phi_{i,t} + (r_2-1)\sum_{(i,t)\in\mathcal{S}_2}\hat{A}_i\Phi_{i,t}.$$

Using the definitions of Λ and $\Xi(\cdot)$, this becomes

$$\sum_{i,t}\omega_{i,t}\hat{A}_i\Phi_{i,t} = \Lambda + (r_1-1)\Xi(\mathcal{S}_1) + (r_2-1)\Xi(\mathcal{S}_2).$$

Substituting this identity into the first-order batch-averaged entropy derivative yields the result. □

H.2 Four Representative Combined Reweighting Operations

We next instantiate Proposition H.1 for four representative two-subset reweighting schemes. Each scheme combines two single-polarity interventions from the four-quadrant entropy decomposition.

(1) Combination C1: amplifying $\mathcal{L}_q^+$ and amplifying $\mathcal{U}_q^-$. C1 amplifies the two dominant entropy-increasing token categories on both advantage sides.

Define

$$\omega_{i,t}^{(\mathrm{C1})} = \begin{cases} W_+, & (i,t) \in \mathcal{L}_q^+, \\ W_-, & (i,t) \in \mathcal{U}_q^-, \\ 1, & \text{otherwise}, \end{cases} \qquad W_+ > 1, \quad W_- > 1.$$

By Proposition H.1,

$$\left.\frac{d\bar{H}}{d\eta}\right|_{\mathrm{C1}} = -\frac{1}{N}\Big[\Lambda + (W_+ - 1)\Xi(\mathcal{L}_q^+) + (W_- - 1)\Xi(\mathcal{U}_q^-)\Big].$$

When the two proxy sets capture the dominant mass of their corresponding entropy-increasing quadrants, we have

$$\left.\frac{d\bar{H}}{d\eta}\right|_{\mathrm{C1}} = -\frac{1}{N}\Big[\Lambda - (W_+ - 1)\Gamma_{L^+} - (W_- - 1)\Gamma_{U^-}\Big].$$

(2) Combination C2: amplifying $\mathcal{L}_q^+$ and attenuating $\mathcal{L}_q^-$. C2 strengthens high-surprisal entropy-increasing tokens under positive advantages and weakens high-surprisal entropy-decreasing tokens under negative advantages. This is the combined operation used by Variant II in the main text.

Define

$$\omega_{i,t}^{(\mathrm{C2})} = \begin{cases} W, & (i,t) \in \mathcal{L}_q^+, \\ M, & (i,t) \in \mathcal{L}_q^-, \\ 1, & \text{otherwise}, \end{cases} \qquad W > 1, \quad 0 < M < 1.$$

By Proposition H.1,

$$\left.\frac{d\bar{H}}{d\eta}\right|_{\mathrm{C2}} = -\frac{1}{N}\Big[\Lambda + (W - 1)\Xi(\mathcal{L}_q^+) + (M - 1)\Xi(\mathcal{L}_q^-)\Big].$$

When the two proxy sets capture the dominant mass of the corresponding entropy-increasing and entropy-decreasing quadrants, respectively, we obtain

$$\left.\frac{d\bar{H}}{d\eta}\right|_{\mathrm{C2}} = -\frac{1}{N}\Big[\Lambda - (W - 1)\Gamma_{L^+} - (1 - M)\Gamma_{L^-}\Big].$$

This expression matches the Variant II batch entropy shift in the main text:

$$\Lambda_{\mathrm{V2}} = \Lambda - (W - 1)\Gamma^+ - (1 - M)\Gamma^-,$$

where

$$\Gamma^+ = \Gamma_{L^+}, \qquad \Gamma^- = \Gamma_{L^-}.$$

(3) Combination C3: attenuating $\mathcal{U}_q^+$ and amplifying $\mathcal{U}_q^-$. C3 regulates the low-surprisal region from both advantage sides. It weakens the entropy-decreasing signal from positive-advantage low-surprisal tokens and strengthens the entropy-increasing signal from negative-advantage low-surprisal tokens.

Define

$$\omega_{i,t}^{(\mathrm{C3})} = \begin{cases} M_+, & (i,t) \in \mathcal{U}_q^+, \\ W_-, & (i,t) \in \mathcal{U}_q^-, \\ 1, & \text{otherwise}, \end{cases} \qquad 0 < M_+ < 1, \quad W_- > 1.$$

By Proposition H.1,

$$\left.\frac{d\bar{H}}{d\eta}\right|_{\mathrm{C3}} = -\frac{1}{N}\Big[\Lambda + (M_+ - 1)\Xi(\mathcal{U}_q^+) + (W_- - 1)\Xi(\mathcal{U}_q^-)\Big].$$

When the proxy sets align with their target quadrants and capture the dominant contributions, we have

$$\left.\frac{d\bar{H}}{d\eta}\right|_{\mathrm{C3}} = -\frac{1}{N}\Big[\Lambda - (1 - M_+)\Gamma_{U^+} - (W_- - 1)\Gamma_{U^-}\Big].$$

(4) Combination C4: attenuating $\mathcal{U}_q^+$ and attenuating $\mathcal{L}_q^-$. C4 attenuates the two dominant entropy-decreasing token categories.

Define

$$\omega_{i,t}^{(\mathrm{C4})} = \begin{cases} M_+, & (i,t) \in \mathcal{U}_q^+, \\ M_-, & (i,t) \in \mathcal{L}_q^-, \\ 1, & \text{otherwise}, \end{cases} \qquad 0 < M_+ < 1, \quad 0 < M_- < 1.$$

By Proposition H.1,

$$\left.\frac{d\bar{H}}{d\eta}\right|_{\mathrm{C4}} = -\frac{1}{N}\Big[\Lambda + (M_+ - 1)\Xi(\mathcal{U}_q^+) + (M_- - 1)\Xi(\mathcal{L}_q^-)\Big].$$

When the two proxy sets capture the dominant mass of their corresponding entropy-decreasing quadrants, we obtain

$$\left.\frac{d\bar{H}}{d\eta}\right|_{\mathrm{C4}} = -\frac{1}{N}\Big[\Lambda - (1 - M_+)\Gamma_{U^+} - (1 - M_-)\Gamma_{L^-}\Big].$$

H.3 Summary of Single-Polarity and Combined Operations

For clarity, we summarize the single-polarity operations O1 to O4 and the combined operations C1 to C4 below.

Single-polarity operations act on one entropy quadrant:

- O1 amplifies $\mathcal{L}_q^+$, the positive-advantage high-surprisal set. It strengthens the entropy-increasing minority under positive advantages.
- O2 attenuates $\mathcal{U}_q^+$, the positive-advantage low-surprisal set. It weakens the dominant entropy-decreasing signal under positive advantages.
- O3 amplifies $\mathcal{U}_q^-$, the negative-advantage low-surprisal set. It strengthens the entropy-increasing effect induced by suppressing low-surprisal tokens.
- O4 attenuates $\mathcal{L}_q^-$, the negative-advantage high-surprisal set. It weakens the entropy-decreasing effect induced by suppressing high-surprisal tokens.

Under the four-quadrant analysis in Section 3.2, O1 and O3 amplify entropy-increasing signals, whereas O2 and O4 attenuate entropy-decreasing signals. Variant I in the main text is a direct instance of O1, since it only amplifies token weights in $\mathcal{L}_q^+$.

Combined operations act on two entropy quadrants:

- C1 jointly amplifies $\mathcal{L}_q^+$ and $\mathcal{U}_q^-$. It strengthens entropy-increasing signals on both the positive- and negative-advantage sides.
- C2 amplifies $\mathcal{L}_q^+$ and attenuates $\mathcal{L}_q^-$. It simultaneously strengthens positive-advantage high-surprisal entropy-increasing tokens and weakens negative-advantage high-surprisal entropy-decreasing tokens.
- C3 attenuates $\mathcal{U}_q^+$ and amplifies $\mathcal{U}_q^-$. It regulates the low-surprisal region by weakening positive-advantage entropy-decreasing tokens and strengthening negative-advantage entropy-increasing tokens.
- C4 jointly attenuates $\mathcal{U}_q^+$ and $\mathcal{L}_q^-$. It suppresses the two dominant entropy-decreasing token categories.

Variant II in the main text is a representative instance of C2. It combines amplification of $\mathcal{L}_q^+$ with attenuation of $\mathcal{L}_q^-$. Since all weights remain positive, the sign of each original GRPO policy-gradient update is preserved. The method only rescales token-level contribution magnitudes, thereby regulating batch-level entropy dynamics through two complementary mechanisms: amplifying entropy-increasing contributions and attenuating entropy-decreasing contributions.

Single-polarity operations implement targeted interventions on individual entropy quadrants, while combined operations perform joint regulation across two quadrants.

Here, $\mathcal{L}_q^+$ and $\mathcal{L}_q^-$ denote high-surprisal proxy subsets within positive- and negative-advantage tokens, respectively, while $\mathcal{U}_q^+$ and $\mathcal{U}_q^-$ denote low-surprisal proxy subsets within positive- and negative-advantage tokens. Amplification refers to assigning weights larger than one, whereas attenuation refers to assigning weights in $(0, 1)$.

Table 15: Summary of O1 to O4 and C1 to C4. The Target subset column specifies the entropy quadrant or pair of quadrants selected by each operation.

Notation	Type	Selected Subset	Weight Adjustment	Mechanistic Interpretation
O1	Single-polarity	$\mathcal{L}_q^+$	Amplification, $W_+ > 1$	Strengthens high-surprisal tokens in positive-advantage samples
O2	Single-polarity	$\mathcal{U}_q^+$	Attenuation, $0 < M_+ < 1$	Weakens low-surprisal tokens in positive-advantage samples
O3	Single-polarity	$\mathcal{U}_q^-$	Amplification, $W_- > 1$	Strengthens the suppression of low-surprisal tokens under negative advantages
O4	Single-polarity	$\mathcal{L}_q^-$	Attenuation, $0 < M_- < 1$	Weakens the suppression of high-surprisal tokens under negative advantages
C1	Combined	$\mathcal{L}_q^+ + \mathcal{U}_q^-$	Joint amplification	Strengthens entropy-increasing token categories on both advantage sides
C2	Combined	$\mathcal{L}_q^+ + \mathcal{L}_q^-$	Amplifies the former and attenuates the latter	Strengthens positive-advantage high-surprisal tokens while weakening negative-advantage high-surprisal entropy-decreasing tokens
C3	Combined	$\mathcal{U}_q^+ + \mathcal{U}_q^-$	Attenuates the former and amplifies the latter	Jointly regulates the low-surprisal region on both advantage sides
C4	Combined	$\mathcal{U}_q^+ + \mathcal{L}_q^-$	Joint attenuation	Suppresses entropy-decreasing token categories on both advantage sides

I Limitations and Broader Impacts

Our analysis characterizes local first-order entropy dynamics in logit space and does not establish near-criticality or global stability under arbitrary finite steps, active clipping, or shared-parameter optimization. STARE depends on trajectory-level advantages and cannot correct erroneous verifiers or prevent reward hacking; robustness to biased, adversarial, non-stationary, or partial-credit rewards remains unverified. Although target-entropy gating limits sustained intervention, suitable hyperparameters may vary across models and tasks, and strong reweighting or noisy feedback can still cause over-exploration or oscillations. Evaluation primarily covers mathematical reasoning and tool use, with limited code-generation evidence; broader task generalization and variability across training seeds require further study. Finally, stable exploration does not ensure factuality, fairness, or safety, and STARE-trained models may generate harmful or misleading content, requiring application-specific verification and evaluation before deployment.

NeurIPS Paper Checklist

The checklist is designed to encourage best practices for responsible machine learning research, addressing issues of reproducibility, transparency, research ethics, and societal impact. Do not remove the checklist: **The papers not including the checklist will be desk rejected.** The checklist should follow the references and follow the (optional) supplemental material. The checklist does NOT count towards the page limit.

Please read the checklist guidelines carefully for information on how to answer these questions. For each question in the checklist:

- You should answer [Yes], [No], or [N/A].
- [N/A] means either that the question is Not Applicable for that particular paper or the relevant information is Not Available.
- Please provide a short (1–2 sentence) justification right after your answer (even for [N/A]).

The checklist answers are an integral part of your paper submission. They are visible to the reviewers, area chairs, senior area chairs, and ethics reviewers. You will also be asked to include it (after eventual revisions) with the final version of your paper, and its final version will be published with the paper.

The reviewers of your paper will be asked to use the checklist as one of the factors in their evaluation. While [Yes] is generally preferable to [No], it is perfectly acceptable to answer [No] provided a proper justification is given (e.g., error bars are not reported because it would be too computationally expensive" or "we were unable to find the license for the dataset we used"). In general, answering [No] or [N/A] is not grounds for rejection. While the questions are phrased in a binary way, we acknowledge that the true answer is often more nuanced, so please just use your best judgment and write a justification to elaborate. All supporting evidence can appear either in the main paper or the supplemental material, provided in appendix. If you answer [Yes] to a question, in the justification please point to the section(s) where related material for the question can be found.

IMPORTANT, please:

- **Delete this instruction block, but keep the section heading "NeurIPS Paper Checklist"**,
- **Keep the checklist subsection headings, questions/answers and guidelines below.**
- **Do not modify the questions and only use the provided macros for your answers**.

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: Yes, please refer to the abstract and introduction section of this article for details.

   Guidelines:

   - The answer [N/A] means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A [No] or [N/A] answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

Justification: Yes, please refer to the Limitations of this article for details.

Guidelines:

- The answer [N/A] means that the paper has no limitation while the answer [No] means that the paper has limitations, but those are not discussed in the paper.
- The authors are encouraged to create a separate "Limitations" section in their paper.
- The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
- The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
- The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
- The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
- If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
- While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [Yes]

   Justification: Yes, please refer to the theoretical analysis section of this article for details.

   Guidelines:

   - The answer [N/A] means that the paper does not include theoretical results.
   - All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
   - All assumptions should be clearly stated or referenced in the statement of any theorems.
   - The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
   - Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
   - Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: Yes, please refer to the experiments of this article for details.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.

- If the paper includes experiments, a [No] answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
- If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
- Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
  (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
  (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
  (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
  (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [No]

   Justification: No. The current anonymous submission does not include an open code package; we describe the experimental details and detailed analysis in the paper and appendix, and plan to release code after the appropriate internal review.

   Guidelines:

   - The answer [N/A] means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (https://neurips.cc/public/guides/CodeSubmissionPolicy) for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so [No] is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
   - The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (https://neurips.cc/public/guides/CodeSubmissionPolicy) for more details.
   - The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
   - The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
   - At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).

- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

   Answer: [Yes]

   Justification: Yes, please refer to the experiments of this article for details.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [No]

   Justification: No, we report average accuracy over repeated evaluation samples (32 evaluations for AIME24/25 and AMC23, and 4 evaluations for the other benchmarks), but do not report error bars or statistical significance tests due to computational cost.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
   - The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
   - The assumptions made should be given (e.g., Normally distributed errors).
   - It should be clear whether the error bar is the standard deviation or the standard error of the mean.
   - It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
   - For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
   - If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: Yes, please refer to the experiments of this article for details. For the 7B Short CoT scenario, we use 2 nodes, each with 8 GPUs, and training 1000 steps only takes one day.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.

- The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
- The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
- The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: Yes. The work is an empirical study on existing datasets and public backbone families, does not involve human subjects, and the submission is prepared in anonymized form.

   Guidelines:

   - The answer [N/A] means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer [No], they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

    Justification: Yes, please refer to the Limitations and Broader Impacts of this article for details.

    Guidelines:

    - The answer [N/A] means that there is no societal impact of the work performed.
    - If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.
    - Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
    - The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
    - The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
    - If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

Answer: [N/A]

Justification: N/A. The paper does not release high-risk generative models, scraped datasets,or similar assets with elevated misuse risk; it studies reinforcement learning for stable training.

Guidelines:

- The answer [N/A] means that the paper poses no such risks.
- Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
- Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
- We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

Answer: [Yes]

Justification: Yes, please refer to the experiments of this article for details.

Guidelines:

- The answer [N/A] means that the paper does not use existing assets.
- The authors should cite the original paper that produced the code package or dataset.
- The authors should state which version of the asset is used and, if possible, include a URL.
- The name of the license (e.g., CC-BY 4.0) should be included for each asset.
- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

Answer: [N/A]

Justification: N/A. The paper does not release new assets.

Guidelines:

- The answer [N/A] means that the paper does not release new assets.
- Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
- The paper should discuss whether and how consent was obtained from people whose asset is used.
- At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [N/A]

    Justification: N/A. The paper does not involve crowdsourcing nor research with human subjects.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [N/A]

    Justification: N/A. The paper does not involve crowdsourcing nor research with human subjects.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
    - We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
    - For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

    Answer: [N/A]

    Justification: N/A. The core method development in this research does not involve LLMs as any important, original, or non-standard components.

    Guidelines:

    - The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
    - Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.