

FEPO: Future-Entropy Policy Optimization with Selective Update Modulation

Cunjie He¹, Tianhao Cheng¹, Zhiqi Li¹, Jiaxin Gu²,
Xiaoshuang Huang², Jiayin Cai², Xiaolong Jiang², Yao Hu²,
Jing Xu³, Yuan Qi¹, Zenglin Xu^{1,3}

¹Fudan University, Shanghai, China ²Xiaohongshu, Shanghai, China

³Shanghai Academy of AI for Science, Shanghai, China

{cjhe24@m., zenglinxu}@fudan.edu.cn

Abstract

Existing RL methods for language models typically use token-level entropy, which captures uncertainty at the current position rather than over future generation. We study *future entropy*, i.e., uncertainty over the remaining suffix conditioned on the current prefix. By the entropy chain rule, local entropy is the expected decrease in future entropy, while the realized decrease depends on the chosen token, yielding an action-dependent future-entropy gap. We evaluate this signal at inference time with MIN-F-MC, an oracle-style probe that selects the candidate token with the lowest estimated future-entropy rate at high-entropy positions. On mathematical reasoning tasks, MIN-F-MC yields higher mean accuracy than greedy and Random TopK across the tested checkpoints and datasets. We then propose **FEPO** (Future Entropy Policy Optimization), which uses realized future-entropy rates at a sparse set of high-entropy positions to scale update magnitudes. Across the tested model–benchmark pairs, FEPO improves Mean@32 over GRPO.¹

1 Introduction

Reinforcement-learning post-training is widely used to improve language-model capabilities, including mathematical reasoning, tool use, and planning (Ouyang et al., 2022; Nakano et al., 2021; Shao et al., 2024; Guo et al., 2025a). Group-based methods such as GRPO avoid a learned value head but assign one sequence-level advantage to all tokens, echoing long-standing credit-assignment issues in sequence-level training (Ranzato et al., 2016; Bengio et al., 2015). Finer-grained credit signals can be obtained from process verifiers (Lightman et al., 2024; Li and Li, 2025), token-wise variants of PPO- and DPO-style preference optimization (Zheng et al., 2023; Rafailov et al., 2023; Zhong et al., 2025), and Monte Carlo or

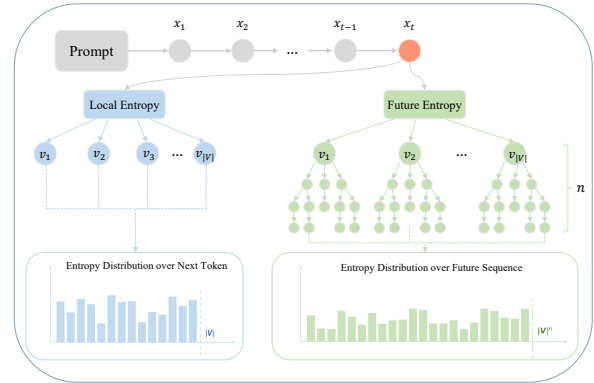


Figure 1: Future entropy vs. local entropy.

segment-level estimators (Kazemnejad et al., 2025; Guo et al., 2025b). These methods derive additional credit information through process supervision, auxiliary models, or additional rollouts.

Recent work shows that learning signals can concentrate at high-entropy or otherwise critical token positions (Wang et al., 2025a; Meng et al., 2026), with related divergent-token patterns appearing in efficient routing systems (Fu et al., 2025). FEPO builds on this sparse-intervention view and derives its signal from existing GRPO rollouts.

Entropy is a natural uncertainty signal (Shannon, 1948; Cover and Thomas, 2006), and recent RL methods use policy entropy for exploration, entropy control, or token-level update shaping (Tan et al., 2025; Xiong et al., 2026; Ahmed and Singh, 2026). Token entropy H_t (1) is local: it measures next-token uncertainty before an action is chosen but does not describe how the realized token changes downstream uncertainty. We define *future entropy* (5) as the conditional entropy of the remaining suffix given the current prefix. The entropy chain rule decomposes the realized future-entropy decrease into token entropy and an action-dependent future-entropy gap $B_t(x_t)$ (7).

At inference time, MIN-F-MC (Figure 2, Section 3.3) probes this signal by selecting low-future-

¹Code: github.com/hecunjie/feipo.

entropy candidates at high-entropy positions. At training time, FEPO reuses realized suffix-entropy rates from existing GRPO rollouts to modulate token-update *magnitude* on the low-entropy tail of sparse high-entropy positions.

Our contributions are:

- We formulate *future entropy* for autoregressive reasoning and derive a decomposition of the realized future-entropy decrease into token entropy and an action-dependent future-entropy gap.
- We introduce MIN-F-MC, an oracle-style inference probe that uses Monte Carlo continuation sampling to test whether lower future entropy at high-entropy decision points is aligned with better mathematical reasoning outcomes.
- We propose FEPO, a training-time update modulator that ranks rollout-level suffix-entropy rates and adjusts the magnitude of GRPO advantages on the low-entropy tail of sparse high-entropy positions.

2 Related Work

2.1 GRPO and the credit-assignment gap

Group Relative Policy Optimization (GRPO) is widely used for RLVR because it estimates trajectory-level advantages by normalizing rewards across multiple completions sampled for the same prompt, thereby avoiding the learned value function used in PPO-style training (Schulman et al., 2017; Shao et al., 2024; Guo et al., 2025a). Recent work further refines GRPO along several axes, including reward normalization, length effects, token aggregation, success amplification, and critic-free credit propagation (Yu et al., 2025; Liu et al., 2025; Mroueh, 2025; Parthasarathi et al., 2025; Wang et al., 2025b). Group-relative policy optimization has also been extended to heterogeneous groups of multi-agent LLM rollouts for critic-free end-to-end optimization of search systems (Chen et al., 2026). Complementary token-level analysis shows that critic-free RL can induce implicit credit assignment through gradient coupling and cancellation across positive and negative rollouts, particularly for shared low-confidence tokens (Cheng et al., 2026b). However, the group-relative advantage remains a sequence-level scalar that is shared by all tokens in the rollout. Thus tokens with potentially different contributions to the final answer receive the same scalar advantage.

2.2 Entropy as a signal: from local to future uncertainty

Two relevant uses of token entropy in RLVR are policy-stability control and token-level update shaping. The first includes entropy regularization to avoid premature collapse (Williams and Peng, 1991; Mnih et al., 2016; Meister et al., 2020) and entropy-aware clipping mechanisms for controlling entropy during PPO-style training (Yu et al., 2025; Cui et al., 2025; Lei and Baehr, 2026). The second is token-level update shaping: GTPO and its variants reweight reasoning steps by local token entropy, and GRPO-S and ETR use token-entropy-based reshaping of the trajectory reward to emphasize high-uncertainty branches (Tan et al., 2025; Cheng et al., 2026a; Xiong et al., 2026; Cao et al., 2025). A common feature of all of these methods is that the entropy quantity used is *local*: it depends only on the next-token distribution at the current prefix.

EPIC (Ahmed and Singh, 2026) incorporates a finite-horizon Rao–Blackwellized lookahead future entropy into *decoding* and aligns the sampling distribution’s stepwise uncertainty with the estimated downstream uncertainty. Related work also studies uncertainty- or entropy-controlled decoding in information-seeking and target-perplexity settings (Hu et al., 2024; Basu et al., 2021). FEPO instead uses future entropy at training time as a within-rollout magnitude modulator of the GRPO advantage (Section 4).

3 Theory

3.1 Setup and notation

Consider an autoregressive language model over vocabulary $\mathcal{V}$ that generates $x = (x_1, \dots, x_n)$. Let $x_{<t} = (x_1, \dots, x_{t-1})$ and $x_{\leq t} = (x_1, \dots, x_t)$. At position t , write $p_t(v) \triangleq p(v \mid x_{<t})$ for the next-token distribution and define the *token entropy*

$$\begin{aligned} H_t &\triangleq H(X_t \mid x_{<t}) \\ &= - \sum_{v \in \mathcal{V}} p_t(v) \log p_t(v). \end{aligned} \quad (1)$$

GRPO objective. For a prompt q , GRPO (Shao et al., 2024) samples K rollouts $\{x^{(k)}\}_{k=1}^K$ from π_{old} and assigns each a scalar reward $R^{(k)}$. The

Model	Method	Math 500				D			
		Acc.	Δ	Branch	Length	Acc.	Δ	Branch	Length
Qwen3-4B	Greedy	42.0	+0.0	0	721.7	19.0	+0.0	0	1036.7
	Random TopK	43.6	+1.6	39.5	732.7	17.7	-1.3	25.9	1076.2
	MIN-F-MC	51.6	+9.6	18.1	860.6	30.0	+11.0	24.5	1200.2
Qwen3-8B	Greedy	40.6	+0.0	0	675.8	29.7	+0.0	0	1460.5
	Random TopK	31.6	-9.0	30.4	709.8	20.7	-9.0	25.3	1385.6
	MIN-F-MC	47.4	+6.8	11.7	820.4	35.3	+5.6	47.9	1609.8

Table 1: Test-time probe comparison on MATH 500 and D (a 300-example subset drawn from the training set). All reported statistics are averaged across multiple independent runs with different random seeds. In MIN-F-MC, future-entropy rate is estimated with 128 Monte Carlo continuations per candidate token. Random TopK samples from the top- k candidate set at high-entropy positions after renormalizing their probabilities. Δ is computed against greedy under the same model and dataset.

path. Then

$$\begin{aligned}\hat{F}_t &\triangleq \sum_{s=t+1}^n H_s, \\ \hat{E}_t &\triangleq \hat{F}_t / (n - t).\end{aligned}\tag{8}$$

Under on-policy sampling, $\mathbb{E}[\hat{F}_t \mid x_{\leq t}] = F(x_{\leq t})$ and $\mathbb{E}[\hat{E}_t \mid x_{\leq t}] = F(x_{\leq t}) / (n - t)$; see Appendix A.5.

The bias term B_t is the only action-dependent component of $\delta_t(x_t)$. Estimating its ordering requires counterfactual continuations for multiple candidate tokens at the same prefix, adding rollout cost at each queried position.

3.3 Test-time probe: MIN-F-MC

To test whether lower future entropy is reward-aligned, MIN-F-MC intervenes only at positions with $H_t > \tau$. It forms $\mathcal{C}_t$ from the top- k tokens of p_t , samples M continuations for each $v \in \mathcal{C}_t$, and estimates

$$\begin{aligned}\hat{F}_{\text{MC}}(x_{<t}, v) &= \frac{1}{M} \sum_{m=1}^M \frac{1}{\min(H_{\max}, n_m - t)} \\ &\quad \cdot \sum_{s=t+1}^{t+\min(H_{\max}, n_m - t)} H_s^{(m)},\end{aligned}\tag{9}$$

This length-normalized rate is used to select $x_t^* = \arg \min_{v \in \mathcal{C}_t} \hat{F}_{\text{MC}}(x_{<t}, v)$; other positions fall back to greedy. For a fixed prefix, this selects the candidate with the largest empirical future-entropy gap among the candidates.

Random TopK uses the same routed positions but samples from the renormalized top- k distribution, isolating the future-entropy criterion from the effect of intervening itself. MIN-F-MC yields higher mean accuracy than greedy and Random

TopK for both checkpoints and datasets (Table 1). This within-checkpoint trend motivates the suffix-entropy ranking used in FEPO. Because absolute accuracy is not monotonic across the 4B and 8B checkpoints, comparisons are made within each checkpoint relative to greedy.

4 FEPO

Section 3.3 evaluates future-entropy-based sparse interventions at inference time through MIN-F-MC. At training time, FEPO uses realized suffix-entropy rates from existing GRPO rollouts to modulate update magnitudes at sparse high-entropy positions, without external process labels or counterfactual rollouts.

4.1 Working assumptions

Assumption 1 (Sparse vital-few). Reward-relevant decisions in RLVR concentrate in a small high-token-entropy set $\mathcal{T} = \{(k, t) : H_t^{(k)} > \tau\}$, where τ is an entropy threshold. FEPO restricts update modulation to this set (Meng et al., 2026).

Assumption 2 (Low-future-entropy / large-gap alignment). Within $\mathcal{T}$, positions in the low- q tail of the batch-level length-bucket-normalized suffix-entropy rank have a higher probability of positive GRPO advantage than positions in the high- q head. FEPO treats low q as a batch-level proxy for larger future-entropy gaps and evaluates this correspondence empirically.

Assumption 1 (A1) is motivated by recent RLVR diagnostics, and Section 5.3 evaluates Assumption 2 (A2).

Off-policy suffix scoring. Proposition 2 is exact only when the suffix is sampled and scored under the same policy. In FEPO, rollouts are drawn from

π_{old} whereas suffix entropies are scored under the current policy π_θ . Thus the future-entropy estimate in Equation (10) is biased under off-policy sampling. Under the old–current policy proximity condition in Assumption 3 (A3), Proposition 3 bounds the distribution-shift component of the expectation-level score drift; Section 5.3 separately evaluates ranks formed from realized suffixes.

4.2 Training-time update modulator: FEPO

Length-bucket normalization. FEPO ranks realized suffix-entropy rates across high-entropy positions in a batch. Because remaining suffix lengths vary across positions and rollouts, we z-score each rate within a suffix-length bucket before ranking.

Definition 2 (Suffix-entropy signal and length-bucket-normalized score). For rollout k and position t , define

$$\begin{aligned}\widehat{E}_t^{(k)} &\triangleq \frac{1}{n^{(k)} - t} \sum_{s=t+1}^{n^{(k)}} H_s^{(k)}, \\ Z_t^{(k)} &\triangleq \frac{\widehat{E}_t^{(k)} - \mu_{b(L_t^{(k)})}}{\sigma_{b(L_t^{(k)})} + \varepsilon},\end{aligned}\quad (10)$$

where $L_t^{(k)} = n^{(k)} - t$, $b(\cdot)$ maps suffix lengths to fixed-width buckets, and (μ_b, σ_b^2) are running bucket statistics (Appendix C). In our implementation, each bucket covers three consecutive suffix lengths. Let $q_t^{(k)} \in [0, 1]$ be the percentile rank of $Z_t^{(k)}$ within the high-entropy set $\mathcal{T}$ after applying the per-rollout cap ρ_{cap} . Lower $q_t^{(k)}$ indicates a lower bucket-normalized realized suffix-entropy rate. FEPO uses this rank as a proxy for a larger future-entropy gap.

Definition 3 (FEPO modulator). Given the GRPO advantage $A_t^{(k)}$, FEPO applies a sign-asymmetric multiplier on the low- q tail:

$$m_t^{(k)} = \begin{cases} 1 + \alpha_{\text{pos}}, & q_t^{(k)} \leq \beta, A_t^{(k)} > 0, \\ 1 - \alpha_{\text{neg}}, & q_t^{(k)} \leq \beta, A_t^{(k)} < 0, \\ 1, & \text{otherwise,} \end{cases} \quad (11)$$

$$\widetilde{A}_t^{(k)} \triangleq m_t^{(k)} A_t^{(k)},$$

where $\alpha_{\text{pos}} \geq 0$, $0 \leq \alpha_{\text{neg}} \leq 1$, and $\beta \in (0, 1/2)$ is a fixed constant. $\widetilde{A}_t^{(k)}$ replaces $A_t^{(k)}$ inside the GRPO surrogate (3); the importance ratio, the clip, and the KL term are unchanged.

Why this asymmetry. FEPO amplifies low- q , positive-advantage updates and attenuates low- q , negative-advantage updates. The attenuation reduces penalties on tokens that may be locally valid despite an unsuccessful trajectory (Section 5.4). The high- q region is left unchanged; Section 5.4 reports the corresponding high-head modulation ablation.

Proposition 3 (Low-tail stability under bounded off-policy shift). *Under Assumption 3 (A3), global proximity between π_{old} and π_θ bounds the mean expectation-level score drift used in the analysis. For these expectation-level scores, changes in low-tail membership are confined to positions with large score drift or scores near the empirical β -quantile boundary.*

Proof in Appendix B. FEPO applies the rank-based modulator only on $\mathcal{T}$. Its additional computation consists of a backward cumulative sum and running bucket statistics; it requires no additional model forward passes, counterfactual rollouts, or auxiliary models. At training time, same-prefix counterfactual branching is replaced by a batch-level rank relaxation. Low- q positions have the smallest bucket-normalized realized suffix-entropy rates among the observed rollouts and are treated as candidates for larger future-entropy gaps. Table 3 evaluates this proxy against an MC=128 same-prefix gap-direction reference.

5 Experiments

5.1 Experimental Setup

Tasks and Datasets. We train on the 17,398-prompt processed DAPO-Math-17k release (Yu et al., 2025) and evaluate on the AIME24 (HuggingFaceH4, 2025) and AIME25 (Zhang and Math-AI Team, 2025) dataset snapshots, MATH 500 (Lightman et al., 2024), AMC23 (Math-AI, 2025), OlympiadBench (He et al., 2024), and Minerva (Lewkowycz et al., 2022). **Evaluation Metrics.** Our primary evaluation metrics are Pass@32 and Mean@32. For each prompt, Mean@32 is the fraction of correct completions among 32 samples, whereas Pass@32 indicates whether at least one of the 32 samples is correct; benchmark scores average these prompt-level quantities. Results on AIME24 and AIME25 are further averaged across multiple random-seed runs, so the reported percentages need not lie on the 1/30 grid of a single evaluation run. **Models and Baselines.** We evaluate Qwen3-4B and Qwen3-8B from the Qwen3

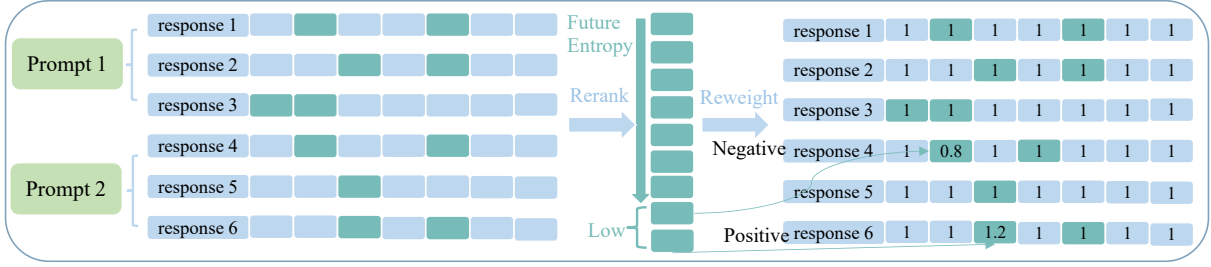


Figure 3: Overview of FEPO. At sparse high-entropy positions, suffix entropy is estimated from rollouts and used to modulate token-level advantage magnitude via a low-tail rank-based selective update signal.

family (Yang et al., 2025), and compare FEPO against GRPO, GTPO, GRPO-S, and ETR. For each baseline, we use the method-specific hyperparameters reported as optimal in its source paper. **Implementation details.** Training hyperparameters, optimizer settings, and generation settings are reported in Appendix E. **Artifact documentation.** Artifact sources, use conditions, intended use, and data statistics are summarized in Appendix F. **FEPO defaults.** Unless otherwise stated, FEPO uses $\alpha_{\text{pos}} = \alpha_{\text{neg}} = 0.2$, tail fraction $\beta = 0.2$, entropy threshold $\tau = 1.0$, and per-response cap $\rho_{\text{cap}} = 5\%$ of the response length. For length normalization, FEPO uses bucket-wise z-scoring with fixed-width suffix-length buckets $b(L) = \lfloor (L - 1)/3 \rfloor$ for $L \geq 1$, minimum bucket count $N_{\text{min}} = 5$, EMA rate $\eta = 0.05$, and score clipping $c = 3$.

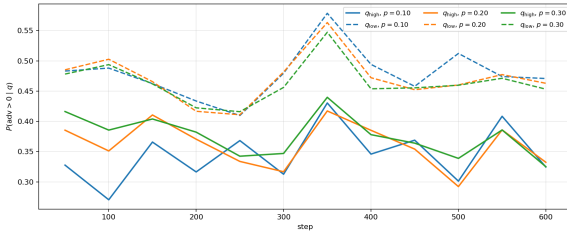


Figure 4: Suffix-entropy quantile diagnostic. We compare $\mathbb{P}(A_t^{(k)} > 0 | q_t^{(k)})$ across low-tail and high-head suffix-entropy quantiles in the high-entropy set $\mathcal{T}$.

5.2 Main results

Table 2 reports Mean@32 (M@32) and Pass@32 (P@32) for GRPO, GTPO, GRPO-S, ETR, and FEPO across six math benchmarks; Figure 5 shows the AIME evaluation curves.

FEPO achieves the highest Mean@32 in 11 of the 12 model–benchmark comparisons and exceeds GRPO in all 12. Relative to GRPO, its largest Pass@32 gains occur on AIME24 and AIME25.

Because the Base results are not monotonic with model size, comparisons are made within each backbone.

5.3 Diagnostics

(A) Realized analogue of MIN-F-MC inside $\mathcal{T}$.

Figure 4 shows that lower realized future entropy is associated with a higher positive-advantage probability inside $\mathcal{T}$, a pattern consistent with Assumption 2 (A2).

Agreement decreases from 0.81 at $q \leq 0.05$ to 0.56 at $q \leq 0.50$. At the default $\beta = 0.2$, agreement is 0.72, compared with 0.73 for MC=1 and 0.64–0.65 for one- and two-step lookahead.

(B) Off-policy: low-tail rank stability.

Suffix entropies are scored under π_θ on π_{old} rollouts. This diagnostic isolates whether rescoreing the same old-policy suffixes under π_θ preserves the ranking used by FEPO; fresh π_θ rollouts would measure the full new-policy future-entropy proxy. Appendix B bounds the global average expected score shift from the average suffix KL $D_{\text{KL}}(\pi_{\text{old}} \| \pi_\theta)$. It then characterizes expectation-level low-tail membership changes at large-drift positions or near the empirical quantile boundary. Figure 6 reports the Spearman correlation between π_θ - and π_{old} -based realized scores on $\mathcal{T}$ during training.



Figure 6: Off-policy diagnostic. Spearman rank correlation between π_θ -based and π_{old} -based suffix-entropy scores, averaged across prompts during training.

Method	AIME24		AIME25		MATH 500		AMC23		Olympiad		Minerva		Average	
	M@32	P@32	M@32	P@32	M@32	P@32	M@32	P@32	M@32	P@32	M@32	P@32	M@32	P@32
<i>Qwen3-4B</i>														
Base	6.1	36.7	5.3	33.3	39.7	90.8	26.4	82.5	21.7	66.5	12.9	44.5	18.7	59.1
GRPO	16.4	<u>43.3</u>	11.5	40.0	79.8	93.0	58.9	<u>92.5</u>	44.1	<u>68.9</u>	40.6	61.4	41.9	66.5
GTPO	16.5	<u>36.7</u>	13.5	38.0	74.8	89.4	58.6	<u>87.5</u>	39.5	<u>61.9</u>	37.4	62.6	40.1	62.7
GRPO-S	<u>18.8</u>	<u>43.3</u>	<u>16.7</u>	<u>45.6</u>	<u>83.5</u>	96.2	64.8	<u>92.5</u>	43.2	64.4	<u>40.9</u>	59.2	<u>44.6</u>	<u>66.9</u>
ETR	16.4	36.7	12.1	40.0	79.7	95.0	57.9	95.0	<u>46.4</u>	71.1	38.6	57.4	41.9	65.9
FEPO	19.8	50.0	17.8	46.7	85.6	96.2	<u>63.5</u>	90.0	49.6	68.0	42.8	64.7	46.5	69.3
<i>Qwen3-8B</i>														
Base	5.8	40	4.9	30	40.8	92	24.6	92.5	18.9	67.4	10.9	49.3	17.7	61.9
GRPO	25.3	60.0	17.9	<u>56.7</u>	<u>86.6</u>	<u>97.2</u>	67.0	<u>95.0</u>	<u>54.1</u>	76.1	<u>43.6</u>	62.1	<u>49.1</u>	74.5
GTPO	26.4	73.3	16.7	50	82.7	93.2	68.6	97.5	50.1	71.2	41.7	63.3	47.7	<u>74.8</u>
GRPO-S	27.6	60.0	18.9	46.7	85.8	<u>97.2</u>	<u>69.1</u>	<u>95.0</u>	49.6	74.6	41.0	66.7	48.7	73.4
ETR	26.7	<u>66.7</u>	15.6	46.7	84.1	<u>97.2</u>	64.0	92.5	51.8	<u>77.1</u>	41.1	63.2	47.2	73.9
FEPO	30.1	73.3	20.3	66.7	87.6	97.4	72.0	<u>95.0</u>	56.0	79.2	44.3	<u>65.6</u>	51.7	79.5

Table 2: Main experiment results. Each benchmark reports Mean@32 (M@32) and Pass@32 (P@32); AIME24 and AIME25 results are averages across multiple random-seed runs. Average reports the mean over the six benchmarks separately for M@32 and P@32. Within each column and each model block, the best result among trained methods (GRPO, GTPO, GRPO-S, ETR, FEPO) is shown in **bold** and the second best is underlined.

Diagnostic synthesis. Figure 4 and Table 3 test complementary properties of the low-tail selection rule: the former relates suffix-entropy ranks to realized advantage signs, while the latter compares the implied gap direction with a higher-budget same-prefix reference. Agreement increases from 0.56 to 0.81 as the selected tail narrows, exposing a precision–coverage trade-off; at $\beta = 0.2$, the 0.72 agreement is close to MC=1 without counterfactual

branching during training. In Figure 6, rank correlation recovers above 0.90 after the early updates and remains near 0.92 later in training. Because FEPO uses quantile ranks, rank preservation is the stability property relevant to the modulator.

5.4 Ablations

We ablate the design choices in Section 4.2. Table 6 (Appendix G.1) reports each setting relative

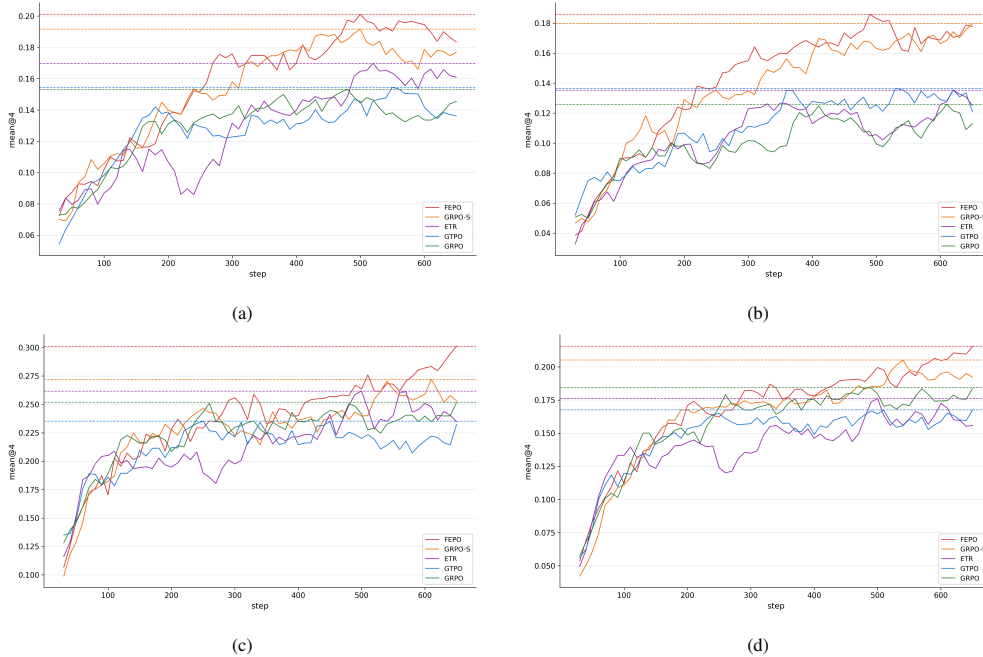


Figure 5: Evaluation curves on the AIME suite. Panels (a) and (b) show Qwen3-4B Mean@4 on AIME24 and AIME25; panels (c) and (d) show the corresponding Qwen3-8B Mean@4 curves.

Estimator	Selection	Large- B_t agree.
FEPO proxy	$q \leq 0.05$	0.81
	$q \leq 0.10$	0.76
	$q \leq 0.20$	0.72
	$q \leq 0.30$	0.67
	$q \leq 0.50$	0.56
Lookahead	1 step	0.64
Lookahead	2 steps	0.65
Monte Carlo	MC=1	0.73

Table 3: Low- q alignment with a large- B_t reference. MC=128 same-prefix continuations estimate the future-entropy gap direction; for increasingly narrow FEPO low- q selections, we report agreement with this reference. Higher agreement at smaller q indicates that the low tail increasingly concentrates samples whose direction matches larger B_t . Lookahead and MC=1 are included as alternative low-cost references.

to the FEPO default. All tested alternatives have negative deltas relative to the FEPO default. **Sign asymmetry** (α_{neg}). With α_{pos} fixed, $\alpha_{\text{neg}} = 0.2$ performs best among the tested in-range values; the deltas for 0.1 and 0 are -0.02 and -0.12 . We additionally test $\alpha_{\text{neg}} = -0.2$ as an out-of-range diagnostic that amplifies negative-advantage updates; its delta is -0.18 . **High-head suppression**. Relative to no high-head modulation, $\alpha_{\text{high}} = -0.2$ and $+0.2$ have deltas of -0.03 and -0.07 , respectively. **Length-bucket correction**. Without bucket correction, the reported delta is -0.15 relative to the FEPO default. **Tail fraction, gating, and cap** ($\beta, \tau, \rho_{\text{cap}}$). Within the tested values, $\beta = 0.2$, $\tau = 1.0$, and $\rho_{\text{cap}} = 5\%$ yield the highest reported scores. Relative to these defaults, the deltas are -0.13 and -0.01 for $\beta \in \{0.1, 0.3\}$, -0.13 and -0.12 for the $\tau = 0.8$ and top-10% gates, and -0.10 , -0.01 , and -0.05 for no cap, 2%, and 10% caps.

Ablation pattern. The largest single degradation is caused by removing length-bucket correction (-0.15), showing that suffix length must be controlled before comparing realized entropy rates across positions. Reducing α_{neg} from 0.2 to 0.1 or 0 weakens performance, and changing it to -0.2 produces the largest drop in that sweep. Both tested modifications to the high-head region are negative, supporting the default choice to leave that region unchanged. The small deltas for $\beta = 0.3$ and a 2% cap indicate local tolerance around the defaults, whereas larger changes to the selected set or cap incur greater losses.

Empirical summary. Relative to GRPO, FEPO raises average Mean@32 from 41.9 to 46.5 for

Qwen3-4B and from 49.1 to 51.7 for Qwen3-8B; average Pass@32 rises from 66.5 to 69.3 and from 74.5 to 79.5. The AIME Pass@32 gains are 6.7 points on both datasets for Qwen3-4B and 13.3 and 10.0 points for Qwen3-8B. Mean@32 measures average sample correctness, whereas Pass@32 measures solution coverage, so their joint improvement distinguishes a broader change from a gain confined to one aggregation rule.

6 Conclusion

We introduce *future entropy* as a longer-horizon complement to token-level entropy. Token-level entropy characterizes uncertainty over the next token, whereas future entropy characterizes uncertainty over the remaining suffix. At sparse high-entropy decision points, this longer-horizon signal further distinguishes among candidate continuations. In our tests, MIN-F-MC yields higher mean accuracy than greedy and Random TopK by selecting low-future-entropy candidates at high-entropy positions. FEPO replaces this same-prefix comparison with batch-level ranks of realized suffix entropy. It treats low- q ranks as a proxy for relatively large future-entropy gaps and modulates the GRPO advantage only on the low-suffix-entropy tail of the sparse high-entropy set, amplifying positive advantages and attenuating negative ones. Our results indicate that future entropy can complement token entropy when weighting GRPO updates at sparse high-entropy positions.

Limitations

Our study is currently validated only on mathematical reasoning tasks and on Qwen3-4B/8B backbones; transfer to code, tool use, planning, multilingual reasoning, and other model families remains to be tested. FEPO also relies on the empirical assumption that, within sparse high-token-entropy positions, lower realized suffix-entropy rates are more likely to align with positive reward advantages. Our diagnostics and ablations support this assumption in the studied setting, but it may not hold uniformly across tasks where exploration, diversity, or multiple valid reasoning paths are more important. During training, FEPO needs to estimate suffix-entropy rates for each selected high-entropy position, which introduces additional runtime overhead. Finally, stronger mathematical-reasoning models may be misused for automated exam or contest assistance.

Acknowledgments

We use AI to assist with language polishing and code-release preparation. We independently review and verify all outputs and take full responsibility for the paper and accompanying code.

References

- Kareem Ahmed and Sameer Singh. 2026. [Entropy-aligned decoding of LMs for better writing and reasoning](#). *arXiv preprint arXiv:2601.01714*.
- Sourya Basu, Govardana Sachitanandam Ramachandran, Nitish Shirish Keskar, and Lav R. Varshney. 2021. [Mirostat: A neural text decoding algorithm that directly controls perplexity](#). In *International Conference on Learning Representations*.
- Samy Bengio, Oriol Vinyals, Navdeep Jaitly, and Noam Shazeer. 2015. [Scheduled sampling for sequence prediction with recurrent neural networks](#). In *Advances in Neural Information Processing Systems*, volume 28, pages 1171–1179.
- Hongye Cao, Zhixin Bai, Ziyue Peng, Boyan Wang, Tianpei Yang, Jing Huo, Yuyao Zhang, and Yang Gao. 2025. [Efficient reinforcement learning with semantic and token entropy for LLM reasoning](#). *arXiv preprint arXiv:2512.04359*.
- Guanzhong Chen, Shaoxiong Yang, Chao Li, Wei Liu, Jian Luan, and Zenglin Xu. 2026. [End-to-end optimization of LLM-driven multi-agent search systems via heterogeneous-group-based reinforcement learning](#). In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 30319–30338, San Diego, California, United States. Association for Computational Linguistics.
- Daixuan Cheng, Shaohan Huang, Xuekai Zhu, Bo Dai, Xin Zhao, Zhenliang Zhang, and Furu Wei. 2026a. [Reasoning with exploration: An entropy perspective](#). *Proceedings of the AAAI Conference on Artificial Intelligence*, 40(36):30377–30385.
- Tianhao Cheng, Zeyu Huang, Zihan Qiu, Yu Cheng, Edoardo Ponti, Yinghui Xu, Ivan Titov, and Zenglin Xu. 2026b. [The cancellation hypothesis in critic-free RL: From outcome rewards to token credits](#). *arXiv preprint arXiv:2605.08666*.
- Thomas M. Cover and Joy A. Thomas. 2006. *Elements of Information Theory*, second edition. Wiley-Interscience.
- 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. 2025. [The entropy mechanism of reinforcement learning for reasoning language models](#). *arXiv preprint arXiv:2505.22617*.
- Tianyu Fu, Yi Ge, Yichen You, Enshu Liu, Zhihang Yuan, Guohao Dai, Shengen Yan, Huazhong Yang, and Yu Wang. 2025. [R2R: Efficiently navigating divergent reasoning paths with small-large model token routing](#). In *Advances in Neural Information Processing Systems*, volume 38, pages 124108–124145.
- Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Peiyi Wang, Qihao Zhu, Runxin Xu, Ruoyu Zhang, Shirong Ma, Xiao Bi, Xiaokang Zhang, Xingkai Yu, Yu Wu, Z. F. Wu, Zhibin Gou, Zhihong Shao, Zhuoshu Li, Ziyi Gao, Aixin Liu, and 175 others. 2025a. [DeepSeek-R1 incentivizes reasoning in LLMs through reinforcement learning](#). *Nature*, 645(8081):633–638.
- Yiran Guo, Lijie Xu, Jie Liu, Dan Ye, and Shuang Qiu. 2025b. [Segment policy optimization: Effective segment-level credit assignment in RL for large language models](#). In *Advances in Neural Information Processing Systems*, volume 38, pages 114399–114431.
- Chaoqun He, Renjie Luo, Yuzhuo Bai, Shengding Hu, Zhen Thai, Junhao Shen, Jinyi Hu, Xu Han, Yujie Huang, Yuxiang Zhang, Jie Liu, Lei Qi, Zhiyuan Liu, and Maosong Sun. 2024. [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)*, pages 3828–3850.
- Zhiyuan Hu, Chumin Liu, Xidong Feng, Yilun Zhao, See-Kiong Ng, Anh Tuan Luu, Junxian He, Pang Wei Koh, and Bryan Hooi. 2024. [Uncertainty of thoughts: Uncertainty-aware planning enhances information seeking in LLMs](#). In *Advances in Neural Information Processing Systems*, volume 37, pages 24181–24215. Curran Associates, Inc.
- HuggingFaceH4. 2025. [AIME 2024](#). Hugging Face dataset. 30-problem default train split; revision 2fe88a2; accessed 2026-08-30.
- Amirhossein Kazemnejad, Milad Aghajohari, Eva Portelance, Alessandro Sordoni, Siva Reddy, Aaron Courville, and Nicolas Le Roux. 2025. [VinePPO: Refining credit assignment in RL training of LLMs](#). In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pages 29557–29590. PMLR.
- Ming Lei and Christophe Baehr. 2026. [A comparative theoretical analysis of entropy control methods in reinforcement learning](#). *arXiv preprint arXiv:2604.09676*.
- Aitor Lewkowycz, Anders Andreassen, David Dohan, Ethan Dyer, Henryk Michalewski, Vinay Ramasesh, Ambrose Slone, Cem Anil, Imanol Schlag, Theo Gutman-Solo, Yuhuai Wu, Behnam Neyshabur, Guy Gur-Ari, and Vedant Misra. 2022. [Solving quantitative reasoning problems with language models](#). In

- Advances in Neural Information Processing Systems*, volume 35, pages 3843–3857.
- Wendi Li and Yixuan Li. 2025. [Process reward model with Q-value rankings](#). In *International Conference on Learning Representations*, volume 2025, pages 14708–14726.
- Hunter Lightman, Vineet Kosaraju, Yuri Burda, Harrison Edwards, Bowen Baker, Teddy Lee, Jan Leike, John Schulman, Ilya Sutskever, and Karl Cobbe. 2024. [Let’s verify step by step](#). In *International Conference on Learning Representations*, volume 2024, pages 39578–39601.
- Zichen Liu, Changyu Chen, Wenjun Li, Penghui Qi, Tianyu Pang, Chao Du, Wee Sun Lee, and Min Lin. 2025. [Understanding R1-Zero-Like Training: A Critical Perspective](#). In *Second Conference on Language Modeling*.
- Math-AI. 2025. [AMC23](#). Hugging Face dataset. 40-problem test split; revision 80815d3; accessed 2026-08-30.
- Clara Meister, Ryan Cotterell, and Tim Vieira. 2020. [If beam search is the answer, what was the question?](#) In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2173–2185. Association for Computational Linguistics.
- Haoming Meng, Kexin Huang, Shaohang Wei, Chiyu Ma, Shuo Yang, Xue Wang, Guoyin Wang, Bolin Ding, and Jingren Zhou. 2026. [Sparse but critical: A token-level analysis of distributional shifts in RLVR fine-tuning of LLMs](#). In *International Conference on Learning Representations*, volume 2026, pages 124171–124210.
- Volodymyr Mnih, Adria Puigdomenech Badia, Mehdi Mirza, Alex Graves, Timothy Lillicrap, Tim Harley, David Silver, and Koray Kavukcuoglu. 2016. [Asynchronous methods for deep reinforcement learning](#). In *Proceedings of the 33rd International Conference on Machine Learning*, volume 48 of *Proceedings of Machine Learning Research*, pages 1928–1937. PMLR.
- Youssef Mroueh. 2025. [Reinforcement learning with verifiable rewards: GRPO’s effective loss, dynamics, and success amplification](#). *arXiv preprint arXiv:2503.06639*.
- Reiichiro Nakano, Jacob Hilton, Suchir Balaji, Jeff Wu, Long Ouyang, Christina Kim, Christopher Hesse, Shantanu Jain, Vineet Kosaraju, William Saunders, Xu Jiang, Karl Cobbe, Tyna Eloundou, Gretchen Krueger, Kevin Button, Matthew Knight, Benjamin Chess, and John Schulman. 2021. [WebGPT: Browser-assisted question-answering with human feedback](#). *arXiv preprint arXiv:2112.09332*.
- Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul Christiano, Jan Leike, and Ryan Lowe. 2022. [Training language models to follow instructions with human feedback](#). In *Advances in Neural Information Processing Systems*, volume 35, pages 27730–27744.
- Prasanna Parthasarathi, Mathieu Reymond, Boxing Chen, Yufei Cui, and Sarath Chandar. 2025. [GRPO- \$\lambda\$: Credit assignment improves LLM reasoning](#). *arXiv preprint arXiv:2510.00194*.
- Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D. Manning, Stefano Ermon, and Chelsea Finn. 2023. [Direct preference optimization: Your language model is secretly a reward model](#). In *Advances in Neural Information Processing Systems*, volume 36, pages 53728–53741.
- Marc’Aurelio Ranzato, Sumit Chopra, Michael Auli, and Wojciech Zaremba. 2016. [Sequence level training with recurrent neural networks](#). In *International Conference on Learning Representations*.
- John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. 2017. [Proximal policy optimization algorithms](#). *arXiv preprint arXiv:1707.06347*.
- Claude E. Shannon. 1948. [A mathematical theory of communication](#). *Bell System Technical Journal*, 27(3):379–423.
- Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, Y. K. Li, Y. Wu, and Daya Guo. 2024. [DeepSeekMath: Pushing the limits of mathematical reasoning in open language models](#). *arXiv preprint arXiv:2402.03300*.
- Hongze Tan, Zihan Wang, Jianfei Pan, Jinghao Lin, Hao Wang, Yifan Wu, Tao Chen, Zhihang Zheng, Zhihao Tang, and Haihua Yang. 2025. [GTPO and GRPO-S: Token and sequence-level reward shaping with policy entropy](#). *arXiv preprint arXiv:2508.04349*.
- Shenzhi Wang, Le Yu, Chang Gao, Chujie Zheng, Shixuan Liu, Rui Lu, Kai Dang, Xiong-Hui Chen, Jianxin Yang, Zhenru Zhang, Yuqiong Liu, An Yang, Andrew Zhao, Yang Yue, Shiji Song, Bowen Yu, Gao Huang, and Junyang Lin. 2025a. [Beyond the 80/20 rule: High-entropy minority tokens drive effective reinforcement learning for LLM reasoning](#). In *Advances in Neural Information Processing Systems*, volume 38, pages 115452–115486.
- Yining Wang, Jinman Zhao, Chuangxin Zhao, Shuhao Guan, Gerald Penn, and Shinan Liu. 2025b. [\$\lambda\$ -GRPO: Unifying the GRPO frameworks with learnable token preferences](#). *arXiv preprint arXiv:2510.06870*.
- Ronald J. Williams and Jing Peng. 1991. [Function optimization using connectionist reinforcement learning algorithms](#). *Connection Science*, 3(3):241–268.

Xuan Xiong, Huan Liu, Li Gu, Zhixiang Chi, Yue Qiu, Yuanhao Yu, and Yang Wang. 2026. [ETR: Entropy trend reward for efficient chain-of-thought reasoning](#). In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 17576–17594, San Diego, California, United States. Association for Computational Linguistics.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, Chujie Zheng, Dayiheng Liu, Fan Zhou, Fei Huang, Feng Hu, Hao Ge, Haoran Wei, Huan Lin, Jialong Tang, and 41 others. 2025. [Qwen3 technical report](#). *arXiv preprint arXiv:2505.09388*.

Qiyang Yu, Zheng Zhang, Ruofei Zhu, Yufeng Yuan, Xiaochen Zuo, Yu Yue, Weinan Dai, Tiantian Fan, Gao-hong Liu, Juncui Liu, Lingjun Liu, Xin Liu, Haibin Lin, Zhiqi Lin, Bole Ma, Guangming Sheng, Yuxuan Tong, Chi Zhang, Mofan Zhang, and 17 others. 2025. [DAPO: An open-source LLM reinforcement learning system at scale](#). In *Advances in Neural Information Processing Systems*, volume 38, pages 113222–113244. Curran Associates, Inc.

Yifan Zhang and Math-AI Team. 2025. [American invitational mathematics examination \(AIME\) 2025](#). Hugging Face dataset. 30-problem default test split; revision 563bb84; accessed 2026-08-30.

Rui Zheng, Shihan Dou, Songyang Gao, Yuan Hua, Wei Shen, Binghai Wang, Yan Liu, Senjie Jin, Qin Liu, Yuhao Zhou, Limao Xiong, Lu Chen, Zhiheng Xi, Nuo Xu, Wenbin Lai, Minghao Zhu, Cheng Chang, Zhangyue Yin, Rongxiang Weng, and 8 others. 2023. [Secrets of RLHF in large language models part I: PPO](#). *arXiv preprint arXiv:2307.04964*.

Han Zhong, Zikang Shan, Guhao Feng, Wei Xiong, Xinle Cheng, Li Zhao, Di He, Jiang Bian, and Liwei Wang. 2025. [DPO meets PPO: Reinforced token optimization for RLHF](#). In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pages 78498–78521. PMLR.

A Proofs

This appendix contains the derivations used in the main text.

A.1 Proof of Proposition 1 (chain-rule identity)

By the entropy chain rule (Cover and Thomas, 2006),

$$\begin{aligned} H(X_t, X_{t+1}, \dots, X_n \mid x_{<t}) \\ &= H(X_t \mid x_{<t}) \\ &\quad + H(X_{t+1}, \dots, X_n \mid X_t, x_{<t}). \end{aligned}$$

The left-hand side is exactly $F(x_{<t})$ by definition (5). The first term on the right-hand side is the token entropy H_t (1). The second term is a conditional entropy over the random variable X_t :

$$\begin{aligned} H(X_{t+1}, \dots, X_n \mid X_t, x_{<t}) \\ &= \sum_{v \in \mathcal{V}} p_t(v) H(X_{t+1}, \dots, X_n \mid X_t=v, x_{<t}) \\ &= \sum_{v \in \mathcal{V}} p_t(v) F(x_{<t}, v) \\ &= \mathbb{E}_{V \sim p_t}[F(x_{<t}, V)] \\ &= \bar{F}_t. \end{aligned}$$

Combining the two sides gives

$$F(x_{<t}) = H_t + \bar{F}_t.$$

A.2 Proof of Proposition 1 (decomposition identity)

Starting from the definition of per-token future-entropy change,

$$\delta_t(x_t) = F(x_{<t}) - F(x_{\leq t}),$$

and substituting the chain-rule identity in Proposition 1:

$$\begin{aligned} \delta_t(x_t) &= (H_t + \bar{F}_t) - F(x_{\leq t}) \\ &= H_t + (\bar{F}_t - F(x_{\leq t})) \\ &= H_t + B_t(x_t). \end{aligned}$$

A.3 Proof of the zero-mean property of B_t and of the identity $\mathbb{E}[\delta_t(X_t)] = H_t$

Taking expectation under $X_t \sim p_t$:

$$\begin{aligned} \mathbb{E}_{X_t \sim p_t}[B_t(X_t)] &= \mathbb{E}_{X_t \sim p_t}[\bar{F}_t - F(x_{<t}, X_t)] \\ &= \bar{F}_t - \mathbb{E}_{X_t \sim p_t}[F(x_{<t}, X_t)] \\ &= 0. \end{aligned}$$

The second line uses linearity of expectation and the fact that $\bar{F}_t$ is a constant with respect to X_t . The last equality follows directly from the definition of $\bar{F}_t$ in (6).

Applying expectation to the decomposition in Proposition 1:

$$\begin{aligned} \mathbb{E}_{X_t \sim p_t}[\delta_t(X_t)] &= H_t + \mathbb{E}_{X_t \sim p_t}[B_t(X_t)] \\ &= H_t. \end{aligned}$$

A.4 Per-token future-entropy change can be negative

A common intuition is that “each additional token can only reduce uncertainty,” which would suggest $\delta_t(x_t) \geq 0$ always. This intuition is incorrect for specific token realizations. The well-known inequality “conditioning reduces entropy” states

$$\mathbb{E}_{X_t \sim p_t}[F(x_{<t}, X_t)] \leq H(X_{t+1}, \dots, X_n \mid x_{<t}).$$

which holds only *in expectation* over X_t , not pointwise. For a low-probability token x_t that opens a wide set of diverse continuations, $F(x_{\leq t})$ can exceed $F(x_{<t})$, yielding $\delta_t(x_t) < 0$.

Minimal example. Consider a binary vocabulary $\mathcal{V} = \{A, B\}$ at position 2 with $p_2(A) = 0.99$ and $p_2(B) = 0.01$. Suppose choosing A makes X_3 deterministic, $F(x_1, A) = 0$, while choosing B makes X_3 uniform, $F(x_1, B) = \ln 2 \approx 0.693$. Then $F(x_{<2}) = H_2 + \bar{F}_2 \approx 0.056 + 0.007 = 0.063$, and for the realized token B , $\delta_2(B) = 0.063 - 0.693 < 0$. The rare token B therefore *increases* future uncertainty rather than reducing it. From the credit-assignment perspective, $\delta_t(x_t) < 0$ identifies tokens that *diverge* the trajectory—a signal that is invisible when only the non-negative quantity H_t is used.

A.5 Proof of Proposition 2 (unbiased suffix-entropy estimator)

Proof. Fix a prefix $x_{\leq t}$ and consider the random suffix $(X_{t+1}, X_{t+2}, \dots, X_n)$ sampled autoregressively from the model distribution $p(\cdot \mid x_{\leq t})$. By Equation (10),

$$\hat{E}_t = \frac{1}{n-t} \sum_{s=t+1}^n h_s(X_{<s}),$$

where

$$h_s(x_{<s}) \triangleq - \sum_{v \in \mathcal{V}} p(v \mid x_{<s}) \log p(v \mid x_{<s}).$$

Note that h_s depends only on the prefix up to position $s-1$, and is itself a deterministic function of $(X_{t+1}, \dots, X_{s-1})$ once $x_{\leq t}$ is fixed.

By linearity of expectation,

$$\mathbb{E}[\hat{E}_t \mid x_{\leq t}] = \frac{1}{n-t} \sum_{s=t+1}^n \mathbb{E}[h_s(X_{<s}) \mid x_{\leq t}]. \quad (12)$$

For each s , the random variable $h_s(X_{<s})$ depends only on $(X_{t+1}, \dots, X_{s-1})$. Writing $\xi = x_{t+1:s-1}$, the expectation is

$$\mathbb{E}[h_s(X_{<s}) \mid x_{\leq t}] = \sum_{\xi} p(\xi \mid x_{\leq t}) h_s(x_{\leq t}, \xi).$$

By definition of conditional entropy,

$$\begin{aligned} H(X_s \mid x_{\leq t}) &= \sum_{\xi} p(\xi \mid x_{\leq t}) H(X_s \mid x_{\leq t}, \xi) \\ &= \sum_{\xi} p(\xi \mid x_{\leq t}) h_s(x_{\leq t}, \xi), \end{aligned}$$

where we used $H(X_s \mid x_{<s}) = h_s(x_{<s})$. Here and below, $H(X_s \mid x_{\leq t})$ is shorthand for $H(X_s \mid X_{t+1:s-1}, x_{\leq t})$: the conditional entropy of X_s averaged over the random intermediate tokens $X_{t+1:s-1}$, given the fixed prefix $x_{\leq t}$.

Therefore $\mathbb{E}[h_s(X_{<s}) \mid x_{\leq t}] = H(X_s \mid X_{t+1:s-1}, x_{\leq t})$ for every $s \in \{t+1, \dots, n\}$. Summing across s and applying the entropy chain rule (Cover and Thomas, 2006),

$$\begin{aligned} \sum_{s=t+1}^n H(X_s \mid X_{t+1:s-1}, x_{\leq t}) \\ &= H(X_{t+1}, \dots, X_n \mid x_{\leq t}) \\ &= F(x_{\leq t}), \end{aligned}$$

where the final equality is the definition (5) of future entropy. Substituting back into (12) yields

$$\mathbb{E}[\hat{E}_t \mid x_{\leq t}] = \frac{F(x_{\leq t})}{n-t}.$$

□

Off-distribution bias. If the suffix is drawn from a distribution $q(\cdot \mid x_{\leq t}) \neq p(\cdot \mid x_{\leq t})$ —for example, from a stale policy π_{old} while the entropies h_s are computed under the current policy π_{θ} , or from a fixed dataset in a teacher-forcing evaluation—the same derivation yields

$$\begin{aligned} \mathbb{E}_{X_{t+1:n} \sim q}[\hat{E}_t \mid x_{\leq t}] \\ &= \frac{1}{n-t} \sum_{s=t+1}^n \sum_{x_{<s}} q(x_{<s} \mid x_{\leq t}) h_s^{(p)}(x_{<s}), \end{aligned}$$

where $h_s^{(p)}$ denotes the local entropy under the model p . The bias relative to $F(x_{\leq t})/(n-t)$ is

$$\begin{aligned} \text{Bias}_t &= \frac{1}{n-t} \sum_{s=t+1}^n \sum_{x_{<s}} (q(x_{<s} \mid x_{\leq t}) \\ &\quad - p(x_{<s} \mid x_{\leq t})) h_s^{(p)}(x_{<s}), \end{aligned} \quad (13)$$

which depends on the correlation between $h_s^{(p)}$ and the likelihood-ratio gap $q - p$ across prefixes. The sign of the bias is not determined a priori. In our GRPO setting, rollouts are drawn from the current policy (or a very recently updated one within the standard importance-weighting range), so $q \approx p$ at the positions used for credit modulation; Appendix B makes this statement explicit by bounding the distribution-shift component under a suffix-level KL condition. Because our method further consumes $\hat{E}_t$ only through *ranks* within the batch-level high-entropy set $\mathcal{T}$, a uniform additive shift in $\hat{E}_t$ across positions in $\mathcal{T}$ does not change the ranking and hence does not affect the modulator $m_t^{(k)}$.

Assumption 3 (Global old-current policy proximity). Let $\mathcal{U}$ be the set of all sampled nonterminal token positions in the batch, and let $\mathcal{T} \subseteq \mathcal{U}$ be the high-entropy routed subset. For each position $i = (k, t) \in \mathcal{U}$, let $q_i = \pi_{\text{old}}(x_{>t} \mid x_{\leq t}^{(k)})$ and $p_i = \pi_{\theta}(x_{>t} \mid x_{\leq t}^{(k)})$ denote the old and current suffix distributions from the same prefix. The average suffix-level KL over the sampled positions is small:

$$\bar{\kappa}_{\mathcal{U}} \triangleq \frac{1}{|\mathcal{U}|} \sum_{i \in \mathcal{U}} D_{\text{KL}}(q_i \| p_i) \ll 1.$$

This is a policy-level proximity condition and does not impose a separate KL condition on the high-entropy subset $\mathcal{T}$.

B Off-policy low-tail stability of FEPO

This appendix makes the off-policy argument explicit. Rollouts are sampled from π_{old} , but suffix entropies are scored under π_{θ} . The theory compares the *expected* scores $\bar{Z}^{q \rightarrow p}$ and Z^p ; FEPO training uses a *single* realized suffix from π_{old} to form (10), so single-sample noise is not part of the deterministic proof and is checked in Section 5.3. We first show that if the old and current suffix distributions have small average KL on the sampled position set $\mathcal{U}$, then the global average induced shift in the expected future-entropy rate is bounded. We then use Markov’s inequality, restrict the resulting bad set to $\mathcal{T}$, and apply a deterministic quantile argument to show that this bounded shift can change low-tail membership only at inherited large-drift positions or near the oracle β -quantile boundary.

Formal statement (full version of Proposition 3). The main text states the conclusion only; this appendix gives notation, bounds, and proof. Let $\mathcal{U}$ be all sampled nonterminal token positions and

let $\mathcal{T} \subseteq \mathcal{U}$ be the routed high-entropy set, with density $\rho_{\mathcal{T}} \triangleq |\mathcal{T}|/|\mathcal{U}|$. For each $i = (k, t) \in \mathcal{U}$, let $p_i = \pi_{\theta}(\cdot \mid x_{\leq t}^{(k)})$ and $q_i = \pi_{\text{old}}(\cdot \mid x_{\leq t}^{(k)})$ be suffix distributions, and let Z_i^p and $\bar{Z}_i^{q \rightarrow p}$ be the fixed-normalization scores obtained from the true π_{θ} future-entropy rate and from the q_i -expected suffix entropy scored under π_{θ} , respectively. FEPO routing uses a single q_i -sampled suffix to form the realized score (10); the bounds below apply to the expectation-level scores $\bar{Z}^{q \rightarrow p}$ and Z^p , while single-suffix noise is assessed in Section 5.3. Let

$$\bar{\kappa}_{\mathcal{U}} \triangleq \frac{1}{|\mathcal{U}|} \sum_{i \in \mathcal{U}} D_{\text{KL}}(q_i \| p_i),$$

$$C_Z \triangleq c_{\text{norm}} \log |\mathcal{V}|.$$

Then the global average expected-score drift obeys

$$\frac{1}{|\mathcal{U}|} \sum_{i \in \mathcal{U}} |\bar{Z}_i^{q \rightarrow p} - Z_i^p| \leq C_Z \sqrt{\bar{\kappa}_{\mathcal{U}}/2}. \quad (14)$$

For any threshold $\eta > 0$, define the global large-drift set and its restriction to $\mathcal{T}$:

$$\mathcal{B}_{\eta}^{\mathcal{U}} \triangleq \{i \in \mathcal{U} : |\bar{Z}_i^{q \rightarrow p} - Z_i^p| > \eta\},$$

$$\mathcal{B}_{\eta}^{\mathcal{T}} \triangleq \mathcal{B}_{\eta}^{\mathcal{U}} \cap \mathcal{T},$$

$$\varepsilon_{\eta}^{\mathcal{T}} \triangleq \frac{|\mathcal{B}_{\eta}^{\mathcal{T}}|}{|\mathcal{T}|}.$$

Markov’s inequality gives a global bad-position fraction, and restriction to $\mathcal{T}$ gives

$$\varepsilon_{\eta}^{\mathcal{T}} \leq \frac{1}{\rho_{\mathcal{T}}} \frac{C_Z}{\eta} \sqrt{\bar{\kappa}_{\mathcal{U}}/2}. \quad (15)$$

Let $Q_{\gamma}(Z^p)$ be the empirical γ -quantile of the oracle current-policy scores on $\mathcal{T}$, with γ clipped to $[0, 1]$. If $\varepsilon_{\eta}^{\mathcal{T}}$ is comparable to β or $1 - \beta$, the clipped bound remains valid but becomes nearly vacuous; the useful regime is $\varepsilon_{\eta}^{\mathcal{T}} \ll \min\{\beta, 1 - \beta\}$. Define the expanded oracle boundary band

$$\begin{aligned} \mathcal{C}_{\eta} &\triangleq \{i \in \mathcal{T} : Q_{\beta - \varepsilon_{\eta}^{\mathcal{T}}}(Z^p) - 2\eta \leq Z_i^p, \\ &\quad Z_i^p \leq Q_{\beta + \varepsilon_{\eta}^{\mathcal{T}}}(Z^p) + 2\eta\}. \end{aligned}$$

Then, up to the standard one-sample ambiguity of empirical quantiles,

$$L_{\beta}(\bar{Z}^{q \rightarrow p}) \triangle L_{\beta}(Z^p) \subseteq \mathcal{B}_{\eta}^{\mathcal{T}} \cup \mathcal{C}_{\eta}. \quad (16)$$

Thus the expected off-policy low tail can differ from the oracle current-policy low tail only on the large-drift positions inherited by $\mathcal{T}$ and on an expanded empirical boundary band. Training uses one realized suffix per position rather than $\bar{Z}^{q \rightarrow p}$, so the bound above is for the distribution-shift (expectation) term only.

Setup. Index the sampled positions in $\mathcal{U}$ by $i = (k, t)$, and write the fixed prefix as $s_i = x_{\leq t}^{(k)}$ and the remaining length as $L_i = n^{(k)} - t$. Let

$$p_i(\cdot) = \pi_\theta(\cdot \mid s_i), \quad q_i(\cdot) = \pi_{\text{old}}(\cdot \mid s_i)$$

denote the current-policy and old-policy suffix distributions from the same prefix, both viewed as distributions over the same L_i -token horizon (or equivalently their L_i -token truncations). For a suffix $y_{>t}$, define the current-policy entropy-rate functional

$$g_i^p(y_{>t}) \triangleq \frac{1}{L_i} \sum_{u=t+1}^{n^{(k)}} H_{\pi_\theta}(Y_u \mid s_i, y_{t+1:u-1}).$$

Since each token entropy is at most $\log |\mathcal{V}|$, we have $0 \leq g_i^p \leq G$ with $G = \log |\mathcal{V}|$. The true current-policy future-entropy rate and its off-policy expectation are

$$e_i^p \triangleq \mathbb{E}_{Y \sim p_i}[g_i^p(Y)], \\ \bar{e}_i^{q \rightarrow p} \triangleq \mathbb{E}_{Y \sim q_i}[g_i^p(Y)].$$

Let Z_i^p and $\bar{Z}_i^{q \rightarrow p}$ be obtained from e_i^p and $\bar{e}_i^{q \rightarrow p}$ using the same length-bucket affine normalization as in Definition 2. We treat the bucket statistics as fixed during routing and write

$$c_{\text{norm}} \triangleq \max_b \frac{1}{\sigma_b + \varepsilon}.$$

For a score vector Z , let $Q_\beta(Z)$ be its empirical β -quantile and

$$L_\beta(Z) \triangleq \{i : Z_i \leq Q_\beta(Z)\}$$

be the low-tail set used by FEPO.

Distribution-shift bound. For each position i , because $g_i^p/G \in [0, 1]$, the variational characterization of total variation gives

$$|\bar{e}_i^{q \rightarrow p} - e_i^p| = |\mathbb{E}_{q_i} g_i^p - \mathbb{E}_{p_i} g_i^p| \\ \leq G \text{TV}(q_i, p_i).$$

Pinsker's inequality then yields

$$|\bar{e}_i^{q \rightarrow p} - e_i^p| \leq G \sqrt{\frac{1}{2} D_{\text{KL}}(q_i \| p_i)}.$$

Applying the fixed length-bucket normalization gives

$$|\bar{Z}_i^{q \rightarrow p} - Z_i^p| \leq C_Z \sqrt{\frac{1}{2} D_{\text{KL}}(q_i \| p_i)}.$$

Averaging over $i \in \mathcal{U}$ and using Jensen's inequality for the concave square-root function,

$$\frac{1}{|\mathcal{U}|} \sum_{i \in \mathcal{U}} |\bar{Z}_i^{q \rightarrow p} - Z_i^p| \leq C_Z \sqrt{\bar{\kappa}_{\mathcal{U}}/2},$$

which proves (14). Markov's inequality then proves (15). Thus a global average suffix-level KL bound between π_{old} and π_θ controls the fraction of positions with large expected score drift, and the factor $1/\rho_{\mathcal{T}}$ accounts for restricting that global bad set to the routed subset.

Low-tail stability. By definition of $\mathcal{B}_\eta^{\mathcal{T}}$, every non-bad position in $\mathcal{T}$ has score drift at most η . This gives the following robust empirical-quantile bounds:

$$Q_\beta(\bar{Z}^{q \rightarrow p}) \geq Q_{\beta - \varepsilon_\eta^{\mathcal{T}}}(Z^p) - \eta, \\ Q_\beta(\bar{Z}^{q \rightarrow p}) \leq Q_{\beta + \varepsilon_\eta^{\mathcal{T}}}(Z^p) + \eta,$$

up to the usual one-sample ambiguity caused by ties. For the upper bound, at least a β fraction of positions in $\mathcal{T}$ are non-bad and have $Z_i^p \leq Q_{\beta + \varepsilon_\eta^{\mathcal{T}}}(Z^p)$, hence their shifted scores are at most $Q_{\beta + \varepsilon_\eta^{\mathcal{T}}}(Z^p) + \eta$. For the lower bound, at most a $(\beta - \varepsilon_\eta^{\mathcal{T}})$ fraction of positions in $\mathcal{T}$ have $Z_i^p < Q_{\beta - \varepsilon_\eta^{\mathcal{T}}}(Z^p)$, and adding all bad positions contributes at most another $\varepsilon_\eta^{\mathcal{T}}$ fraction. Therefore fewer than a β fraction of shifted scores can fall below $Q_{\beta - \varepsilon_\eta^{\mathcal{T}}}(Z^p) - \eta$, which gives the stated lower quantile bound. When the empirical score density around $Q_\beta(Z^p)$ is bounded, choosing η to balance the inherited Markov term and the boundary-band width gives a total membership-change rate that vanishes with $\bar{\kappa}_{\mathcal{U}}$ for fixed $\rho_{\mathcal{T}} > 0$; for example, the natural scaling is $\eta \asymp (\sqrt{\bar{\kappa}_{\mathcal{U}}}/\rho_{\mathcal{T}})^{1/2}$ up to constants.

Now consider a non-bad position whose membership changes between the true current-policy low tail and the off-policy expected low tail. If $i \in L_\beta(Z^p)$ but $i \notin L_\beta(\bar{Z}^{q \rightarrow p})$, then

$$Z_i^p \leq Q_\beta(Z^p), \\ \bar{Z}_i^{q \rightarrow p} > Q_\beta(\bar{Z}^{q \rightarrow p}).$$

Using $\bar{Z}_i^{q \rightarrow p} \leq Z_i^p + \eta$ and $Q_\beta(\bar{Z}^{q \rightarrow p}) \geq Q_{\beta - \varepsilon_\eta^{\mathcal{T}}}(Z^p) - \eta$ gives

$$Z_i^p > Q_{\beta - \varepsilon_\eta^{\mathcal{T}}}(Z^p) - 2\eta.$$

Together with $Z_i^p \leq Q_\beta(Z^p)$, this puts i in the lower part of the expanded boundary band. The

reverse case, $i \notin L_\beta(Z^p)$ but $i \in L_\beta(\bar{Z}^{q \rightarrow p})$, similarly implies $Z_i^p \leq Q_{\beta+\varepsilon\tau_\eta}(Z^p) + 2\eta$. Adding the bad positions $\mathcal{B}_\eta^\tau$ proves (16). Thus global old/current suffix-distribution shift can only affect the off-policy *expected* low-tail selection through a Markov-controlled bad set inherited by $\mathcal{T}$ and the empirical mass in an expanded band around the true current-policy boundary. The remaining risk beyond this deterministic bound is single-suffix sampling noise (realized Z versus $\bar{Z}^{q \rightarrow p}$).

Relation to policy updates. The KL quantity above is a global suffix-level average over sampled positions:

$$\bar{\kappa}_\mathcal{U} = \frac{1}{|\mathcal{U}|} \sum_{i \in \mathcal{U}} D_{\text{KL}}(q_i(x_{>t}) \| p_i(x_{>t})).$$

By the chain rule for KL, define the stepwise divergence

$$d_u(x_{<u}) \triangleq D_{\text{KL}}(\pi_{\text{old}}(\cdot | x_{<u}) \| \pi_\theta(\cdot | x_{<u})).$$

Then the suffix-level KL decomposes as

$$D_{\text{KL}}(q_i \| p_i) = \mathbb{E}_{q_i} \sum_{u=t+1}^{n^{(k)}} d_u(x_{<u}). \quad (17)$$

PPO-style small updates and KL regularization therefore provide a mechanism for keeping the global average $\bar{\kappa}_\mathcal{U}$ small on sampled prefixes, without requiring the high-entropy routed positions to satisfy a separate KL assumption. The empirical diagnostic below checks the resulting rank fidelity directly.

Empirical verification. Because FEPO ultimately depends on ranks and uses a single realized suffix rather than the off-policy expectation $\bar{e}^{q \rightarrow p}$, we assess the gap left by the proof in two ways. First, Section 5.3 reports low- q /large- B_t agreement against an MC=128 same-prefix reference, which probes whether smaller q more often matches the larger-gap direction. Second, Figure 6 reports the Spearman correlation between old-policy and current-policy suffix-entropy scores inside $\mathcal{T}$, computed per prompt and averaged across training. Spearman is a complementary *global* ranking-stability statistic: high correlation indicates that few pairwise orderings change; together with (16), this supports the rank-stability consequence of Assumption 3 in practice.

C Length-bucket statistics for FEPO

This appendix details how FEPO maintains the length-bucket statistics (μ_b, σ_b^2) used in the length-corrected suffix-entropy score (10).

Welford’s online update. For each bucket b we maintain a running count N_b , a running mean μ_b , and a running second-moment accumulator $M_{2,b}$. For each in-bucket sample x (i.e. a value $\hat{E}_t^{(k,b)}$ whose length bucket equals b), the update is

$$\begin{aligned} N_b &\leftarrow N_b + 1, \\ \delta_1 &\leftarrow x - \mu_b, \\ \mu_b &\leftarrow \mu_b + \delta_1 / N_b, \\ \delta_2 &\leftarrow x - \mu_b, \\ M_{2,b} &\leftarrow M_{2,b} + \delta_1 \delta_2, \\ \sigma_b^2 &\leftarrow M_{2,b} / \max(N_b - 1, 1). \end{aligned} \quad (18)$$

Welford’s algorithm is the standard numerically stable estimator of the running variance and avoids the catastrophic cancellation that arises in the naive mean-of-squares formulation.

Lagged use within a training step. At each training step, the routing decision in batch i uses only the statistics accumulated through batch $i - 1$. The new samples in batch i are folded into the bucket statistics via Equation (18) only after the routing decision has been made. This lagged scheme guarantees that the modulator is independent of the samples it scores, eliminating an obvious source of self-referential bias.

EMA smoothing. To reduce step-to-step fluctuations, we additionally apply an exponential moving-average smoothing on top of the Welford estimates,

$$\begin{aligned} \mu_b^{\text{EMA}} &\leftarrow (1 - \eta) \mu_b^{\text{EMA}} + \eta \mu_b, \\ (\sigma_b^2)^{\text{EMA}} &\leftarrow (1 - \eta) (\sigma_b^2)^{\text{EMA}} + \eta \sigma_b^2, \end{aligned} \quad (19)$$

with a small EMA rate $\eta \in (0, 1]$. The score in (10) is then computed using $(\mu_b^{\text{EMA}}, (\sigma_b^2)^{\text{EMA}})$.

Warm-up and bucket fallback. When a bucket has too few samples, the variance estimator is unreliable. We therefore use the following deterministic fallback rule. Let $N_{\min}$ be a minimum sample-count threshold; for each query bucket b , FEPO returns the statistics of the smallest superset that has at least $N_{\min}$ samples, in the following order: (i) bucket b itself; (ii) the union of b and its two adjacent buckets $b - 1, b + 1$; (iii) coarser length buckets

obtained by widening the bucket boundaries; (iv) the global pool over all lengths. During an initial warm-up phase ($N_b < N_{\min}$ for all buckets), FEPO does not modulate any positions, and the modulator in (11) reduces to $m_t^{(k)} = 1$ everywhere; this disables FEPO’s intervention until enough statistics have been collected.

Score clipping. Finally, the length-corrected score is clipped to a bounded range

$$Z_t^{(k)} \leftarrow \text{clip}(Z_t^{(k)}, -c, c), \quad (20)$$

with c a small constant (e.g. $c = 3$). Clipping limits the influence of bucket outliers; since FEPO consumes $Z_t^{(k)}$ only through the batch-level rank $q_t^{(k)}$, clipping has no effect on positions whose ranks are unambiguously inside or outside the low-tail set.

Pseudocode. The complete bucketed scoring procedure is summarized in Algorithm 1.

D Entropy-rate normalization

The future entropy $F(s)$ as defined in (5) is a *sum* over all remaining positions. When continuations have different lengths, the raw sum is dominated by length rather than per-token uncertainty, making cross-candidate comparisons unreliable.

We address this by normalizing to the *entropy rate*: for a continuation of length T following prefix s ,

$$F^{\text{rate}}(s) \triangleq \frac{F(s)}{T}.$$

Under the model’s own distribution, the expected entropy rate equals the expected log-perplexity of the continuation (Cover and Thomas, 2006):

$$\mathbb{E} \left[\frac{1}{T} \sum_{\tau=1}^T (-\log p(y_\tau | y_{<\tau})) \right] = \frac{1}{T} \sum_{\tau=1}^T H(Y_\tau | y_{<\tau}),$$

which is exactly $F^{\text{rate}}(s)$ when the continuation is sampled from the model itself.

After normalization, F^{rate} and the token entropy H_t share the same unit (nats/token), making the bias term $\bar{F}_t^{\text{rate}} - F^{\text{rate}}(x_{\leq t})$ directly comparable across positions and across candidates with different continuation lengths.

E Implementation Details

Training uses 16×H800 with a global batch size of 128, a mini batch size of 64, a group size of

Algorithm 1 Length-bucket scoring for FEPO at training step i .

- 1: **Input:** batch of suffix-entropy rates $\{\hat{E}_t^{(k)}\}$ with lengths $\{L_t^{(k)}\}$; cached statistics $(\mu_b^{\text{EMA}}, (\sigma_b^2)^{\text{EMA}}, N_b)$ from step $i - 1$.
- 2: **Output:** length-corrected scores $\{Z_t^{(k)}\}$, updated statistics for step i .
- 3: **for** each (k, t) in the batch **do**
- 4: $b \leftarrow b(L_t^{(k)})$
- 5: $(\mu, \sigma^2, N) \leftarrow$ statistics of the smallest superset of b with $N \geq N_{\min}$
- 6: **if** $N < N_{\min}$ (warm-up) **then**
- 7: $Z_t^{(k)} \leftarrow 0$ ▷ disables FEPO modulation
- 8: **else**
- 9: $Z_t^{(k)} \leftarrow (\hat{E}_t^{(k)} - \mu) / (\sqrt{\sigma^2} + \varepsilon)$
- 10: $Z_t^{(k)} \leftarrow \text{clip}(Z_t^{(k)}, -c, c)$
- 11: **end if**
- 12: **end for**
- 13: **Routing decision:** compute batch-level ranks $\{q_t^{(k)}\}$ inside $\mathcal{T}$ and modulators $\{m_t^{(k)}\}$ from $\{Z_t^{(k)}\}$.
- 14: **for** each (k, t) in the batch **do**
- 15: Update bucket $b(L_t^{(k)})$ statistics $(N_b, \mu_b, M_{2,b}, \sigma_b^2)$ via the Welford recursion (18).
- 16: **end for**
- 17: Apply EMA smoothing in (19) to update $(\mu_b^{\text{EMA}}, (\sigma_b^2)^{\text{EMA}})$.

16, and a learning rate of 1×10^{-6} . For the optimizer we use the standard GRPO surrogate of (4) with a symmetric clip range $\epsilon_{\text{clip}} = 0.2$, a low-variance KL penalty against π_{ref} with coefficient $\beta_{\text{KL}} = 10^{-3}$, and a cosine learning-rate schedule with 10% warm-up. For generation, we use temperature 1.0, top- p 1.0, a maximum prompt length of 2048 tokens, and a maximum response length of 8192 tokens. The runtime summary is reported in Table 5.

F Artifacts and Data Documentation

Using Qwen3-4B/8B backbones (Yang et al., 2025), we train on the processed DAPO-Math-17k release (Yu et al., 2025) and evaluate on the six benchmarks listed in Table 4; their dataset sources are cited in Section 5.1. Artifacts follow their release terms; we redistribute neither datasets nor modified weights. The code supports reproducibility.

Artifact	Role	Statistics used here	Domain
DAPO-17k (processed)	train	17,398 prompts; 300-example probe subset	math reasoning
MATH 500	eval/probe	500 problems	math reasoning
AIME24/AIME25	eval	30 problems each; 32 samples/prompt	contest math
AMC23	eval	40 problems; 32 samples/prompt	contest math
OlympiadBench	eval	674 English open-ended problems; 32 samples/prompt	olympiad math
Minerva	eval	272 problems; 32 samples/prompt	math reasoning
Qwen3-4B/8B	model	4B and 8B backbones	language model

Table 4: Artifacts and data used in this work.

The prompts and solutions are public mathematical-reasoning data, primarily in English. We collect no new human-subject, user-conversation, or social-media data and add no personally identifying information. We rely on the original providers’ curation for content checks.

Table 4 summarizes artifact coverage. We use the repositories’ default train split for AIME24 and default test splits for AIME25 and AMC23, without repartitioning. The other evaluations use the 500-problem MATH test subset, the 674-problem English open-ended competition-math subset of OlympiadBench, and the 272-problem Minerva Math benchmark. Each main benchmark is evaluated with Mean@32 and Pass@32 (32 completions per prompt).

G Training dynamics and cost

This appendix complements the diagnostics in Section 5.3. For GRPO, GTPO, and FEPO on both models, we log policy entropy H_t , policy KL loss, and mean response length at every step. FEPO’s entropy, KL-loss, and response-length trajectories remain comparable to the baselines. For FEPO, we additionally log suffix length, low-tail and high-head positive/negative ratios, low-entropy means inside $\mathcal{T}$, and suffix-rate percentiles; Figure 7 shows that these traces remain stable across training and are consistent with the sign-asymmetric modulation design.

Model	Metric	GRPO	FEPO
Qwen3-4B	s/step	~122	~146
	h/total	~23	~26
Qwen3-8B	s/step	~272	~302
	h/total	~51	~57

Table 5: Training cost summary for GRPO and FEPO on Qwen3-4B and Qwen3-8B, reported as seconds per step (s/step) and total training time (h/total).

G.1 Ablation summary

Table 6 summarizes the major design sweeps on AIME24/AIME25.

Ablation	Setting	Avg.
α_{pos}	0.2	0.0
	0.3	-0.06
	0.4	-0.07
Entropy threshold	1.0	0.0
	0.8	-0.13
	top10%	-0.12
Length corr.	w/ bucket correction	0.0
	w/o bucket correction	-0.15
α_{neg}	0.2	0.0
	0.1	-0.02
	0.0	-0.12
	-0.2	-0.18
High-head	none	0.0
	$\alpha_{\text{high}} = -0.2$	-0.03
	$\alpha_{\text{high}} = +0.2$	-0.07
Tail fraction β	0.1	-0.13
	0.2	0.0
	0.3	-0.01
Cap ratio	no_cap	-0.1
	0.02cap	-0.01
	0.05cap	0.0
	0.1cap	-0.05

Table 6: Ablations and hyperparameters. Average test accuracy is reported as a delta against the FEPO default on AIME24/AIME25.

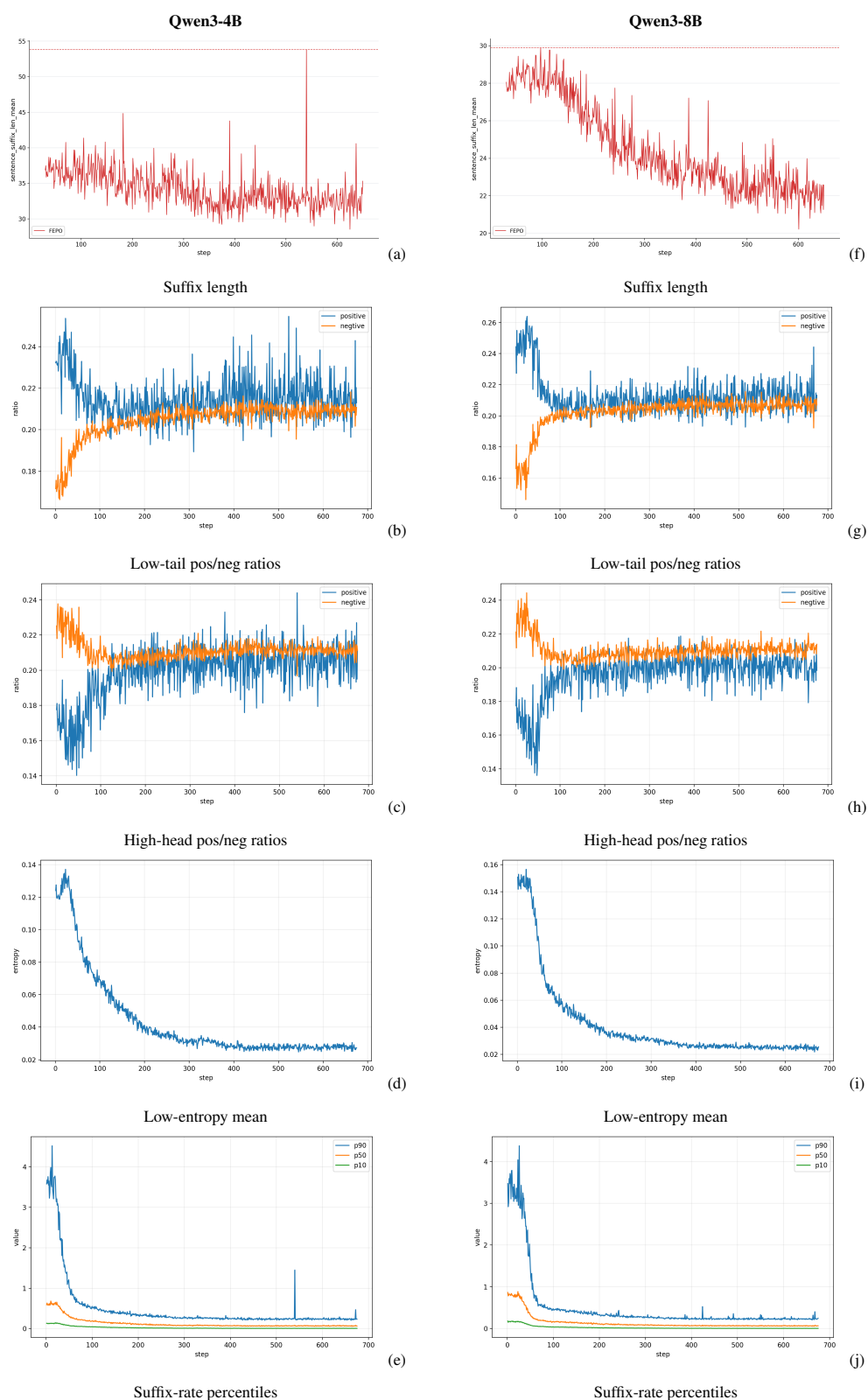


Figure 7: FEPO-specific training diagnostics. Left column: Qwen3-4B (a)–(e); right column: Qwen3-8B (f)–(j). From top to bottom: suffix length, low-tail positive/negative ratios, high-head positive/negative ratios, low-entropy mean, and suffix-rate percentiles.