Heterogeneous Adaptive Policy Optimization: Tailoring Optimization to Every Token's Nature

Zheng Liu[1,2,*], **Mengjie Liu**[1,2,*], **Siwei Wen**[3], **Mengzhang Cai**[2], **Bin Cui**[1,5], **Conghui He**[2], **Wentao Zhang**[1,4,5,†]

[1]Peking University, [2]Shanghai AI Laboratory, [3]Beihang University, [4]Zhongguancun Academy, [5]Beijing Key Laboratory of Software and Hardware Cooperative Artificial Intelligence Systems

Abstract

Using entropy as a measure of heterogeneity to guide optimization has emerged as a crucial research direction in Reinforcement Learning for LLMs. However, existing methods typically treat it as a discrete filter or post-hoc regulator rather than a core optimization driver. To fully leverage the potential of entropy and achieve fine-grained regulation, we introduce **H**eterogeneous **A**daptive **P**olicy **O**ptimization (HAPO), a token-aware algorithm that continuously adapts optimization dynamics based on token-level entropy throughout the entire training process. Our algorithm includes four key components: (1) **Adaptive Temperature Sampling** that adjusts sampling temperature in real time, promoting exploration at high-entropy tokens. (2) **Token-Level Group Average Advantage Estimation** that estimates advantages at token level, accounting for sequence-length effects while preserving non-biased treatment.(3) **Differential Advantage Redistribution** that leverages entropy and importance ratios to adjust advantages for tokens with clear signals. (4) **Asymmetric Adaptive Clipping** that adynamically adjusts clipping boundaries based on token-level entropy. Through systematic investigation of entropy, we embed token-level treatment into every stage. Extensive experiments on mathematical reasoning, code, and logic tasks across multiple models demonstrate HAPO's consistent superiority over DAPO. Our code can be found in https://github.com/starriver030515/HAPO.

1 Introduction

Reinforcement Learning from Human Feedback (Ouyang et al., 2022) has emerged as a fundamental technique for enhancing the reasoning capabilities of LLMs. State-of-the-art models including OpenAI o1 (OpenAI, 2024), DeepSeek-R1 (DeepSeek-AI et al., 2025), and Qwen3 series (Yang et al., 2025) have achieved remarkable performance gains, underscoring how carefully designed reinforcement learning frameworks can empower LLMs to tackle tasks requiring reasoning.

However, existing algorithms (Schulman et al., 2017b; Shao et al., 2024; Yu et al., 2025) share a fundamental limitation: they employ a uniform optimization across all tokens, failing to distinguish between tokens that represent critical reasoning paths vs routine patterns. This conflicts with the heterogeneous nature of language generation, where tokens serve different roles in the reasoning.

Recent works have pioneered token heterogeneity by using entropy to distinguish tokens with different roles. DAPO with forking tokens (Wang et al., 2025b) proves that only a minority of high-entropy tokens guide the optimization process. Archer (Wang et al., 2025a) divides tokens into high/low entropy groups, relaxing clipping bounds for high-entropy tokens to encourage exploration. However, these methods function entropy as discrete filters, failing to leverage it for fundamental optimization and adjustment. Other methods leverage entropy to adjust advantages. Entropy Adv (Cheng et al., 2025) adds an entropy term in advantage to encourage extended reasoning, while EDGE-GRPO (Zhang et al., 2025) utilizes sequence-level entropy to reweight advantages in favor of confident outputs. While realizing a form of heterogeneity, they restrict entropy to an auxiliary regulator or post-hoc bonus, leaving the core optimization mechanics fundamentally unchanged.

These limitations motivate us to develop an algorithm that embeds entropy into the core of optimization, thereby fully leveraging its potential and achieve fine-grained regulation. We first conduct a systematic empirical analysis, delving into the mechanisms and requirements for heterogeneous treatment across different training stages:

Rollout Generation: Exploration-Exploitation Imbalance. High-entropy tokens encoding critical

* Equal contribution.
† Corresponding authors.

Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 6023–6045

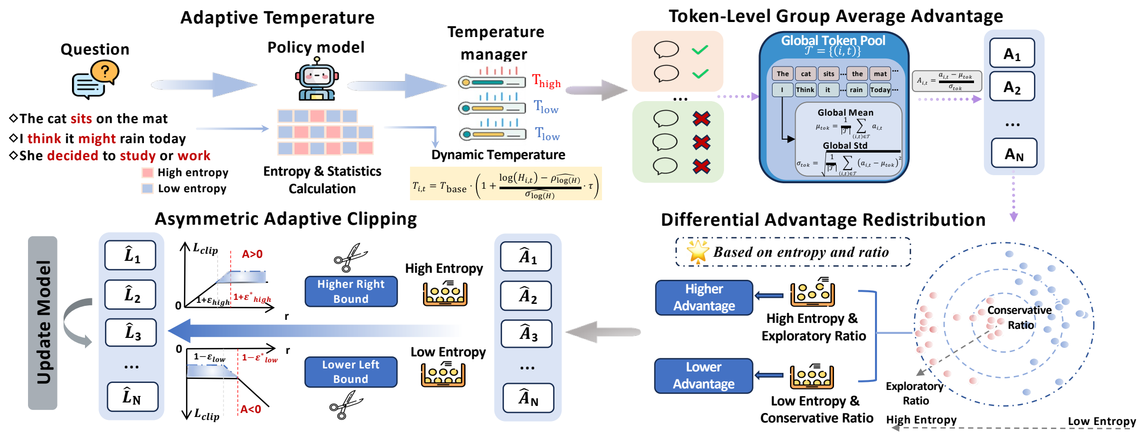


Figure 1: HAPO's overall framework. HAPO integrates Adaptive Temperature Sampling, Token-Level Group Average Advantage Estimation, Differential Advantage Redistribution and Asymmetric Adaptive Clipping, leveraging entropy as a core optimization driver to achieve fine-grained, token-level optimization.

reasoning decisions are rare. We find the temperature dilemma: low temperatures preserve accuracy but suppress critical tokens, while high temperatures increase their occurrence but introduce noise.

Advantage Calculation: Reward Attribution Granularity. Standard sequence-level advantage estimation ignores intra-sequence token heterogeneity and becomes problematic under DAPO's token-mean averaging loss where longer negative samples dominate gradients. Moreover, our importance ratio analysis shows that tokens require different update magnitudes regardless of their entropy values, yet uniform and entropy-based reweighting fail to identify critical reasoning decisions.

Clipping Loss Computation: Clipping Constraint Mismatch. Our clipping analysis reveals paradoxical patterns: low-entropy tokens (mostly formatting symbols) hit left bounds preventing probability reduction, while high-entropy tokens (crucial reasoning elements) hit right bounds limiting exploration. Uniform clipping thus protects noise while constraining exploration.

Based on the discoveries above, we introduce **Heterogeneous Adaptive Policy Optimization (HAPO)**. Moving beyond simple regulation, HAPO formulates optimization parameters as continuous functions of entropy to embed fine-grained treatment into every stage:

- **Adaptive Temperature Sampling**: We dynamically adjust sampling temperature based on token entropy—increasing it for high-entropy tokens to promote exploration while reducing it for low-entropy tokens to maintain coherence. This resolves the accuracy-exploration trade-off in training data.
- **Token-Level Group Average Advantage Estimation**: We estimate advantages at the token level across the group, respecting intra-sequence heterogeneity. This also ensures balanced optimization between positive and negative samples while preserving DAPO's token mean loss benefits for long-sequence learning.
- **Differential Advantage Redistribution**: We leverage entropy and importance ratios to modulate advantages—amplifying advantages for high-entropy tokens with extreme ratios while suppressing low-entropy ones near unity, enabling fine-grained attribution aligned with each token's optimization needs.
- **Asymmetric Adaptive Clipping**: We implement reversed asymmetric clipping boundaries. Low-entropy tokens receive expanded left boundaries to enable aggressive noise suppression, while high-entropy tokens receive expanded right boundaries to facilitate exploration at critical decision points.

HAPO realizes fine-grained token-level heterogeneous optimization based on each token's continuous entropy signal. Our experiments demonstrate that this systematic, continuous treatment yields substantial improvements on mathematical reasoning, code, and logic tasks across multiple models.

2 Background

Given a language model π_θ. States consist of the prompt q and previously generated tokens $o_{<t} = (o_1, o_2, ..., o_{t-1})$. Actions are tokens o_t from vocabulary $\mathcal{V}$. Rewards $R(o)$ are typically provided

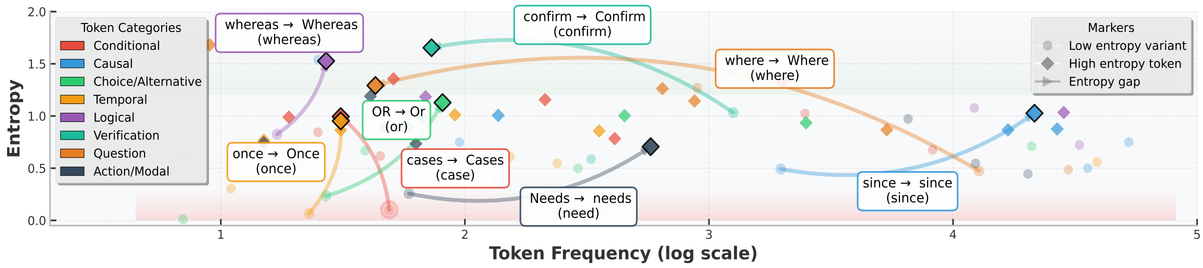
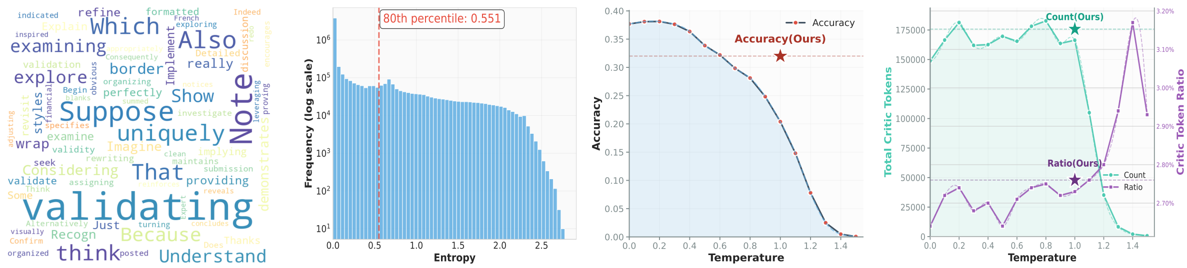


Figure 2: Dual-Entropy Token Frequency-Entropy Landscape



(a) Top100 High Entropy Tokens (b) Entropy Distribution (c) Accuracy (d) Critic Token Counts

Figure 3: Token entropy characteristics and temperature effects on critical tokens.

at the sequence level for the complete response o.

Group Relative Policy Optimization. GRPO (Shao et al., 2024) removes value function by employing group normalization. For each prompt q, GRPO samples G responses $\{o_i\}_{i=1}^{G}$:

$$\mathcal{L}^{\text{GRPO}} = \mathbb{E}_{(q,o)}\Bigg[\frac{1}{G}\sum_{i=1}^{G}\frac{1}{|o_i|}\sum_{t=1}^{|o_i|}\min\Big(r_{i,t}A_{i,t},\ \text{clip}(r_{i,t}, 1-\epsilon, 1+\epsilon)A_{i,t}\Big) - \beta D_{\text{KL}}\left[\pi_\theta \| \pi_{\text{ref}}\right]\Bigg] \quad (1)$$

The advantage is estimated at sequence-level within the group as $A_{i,t} = \frac{R_i - \text{mean}(\{R_j\}_{j=1}^{G})}{\text{std}(\{R_j\}_{j=1}^{G})}$.

GRPO with Entropy Advantage. Entropy Adv (Cheng et al., 2025) augments the advantage function with an entropy-based term, encouraging longer and deeper reasoning chains:

$$\mathcal{L}^{\text{Ent}} = \mathbb{E}_{(q,o)}\Bigg[\frac{1}{G}\sum_{i=1}^{G}\frac{1}{|o_i|}\sum_{t=1}^{|o_i|}\min\Big(r_{i,t}A^{\text{s}}_{i,t},\ \text{clip}(r_{i,t}, 1-\epsilon, 1+\epsilon)A^{\text{s}}_{i,t}\Big) - \beta D_{\text{KL}}\left[\pi_\theta \| \pi_{\text{ref}}\right]\Bigg] \quad (2)$$

where $A^{\text{s}}_{i,t} = A_{i,t} + \psi(\mathcal{H}_t)$, combining the advantage $A_{i,t}$ with an entropy-based bonus $\psi(\mathcal{H}_t)$.

Decoupled Clip and Dynamic Ampling Policy Optimization. DAPO (Yu et al., 2025) modifies GRPO with token-mean loss—averaging over tokens rather than sequences—to preserve gradient contributions from longer sequences. It also introduces clip-higher with asymmetric bounds:

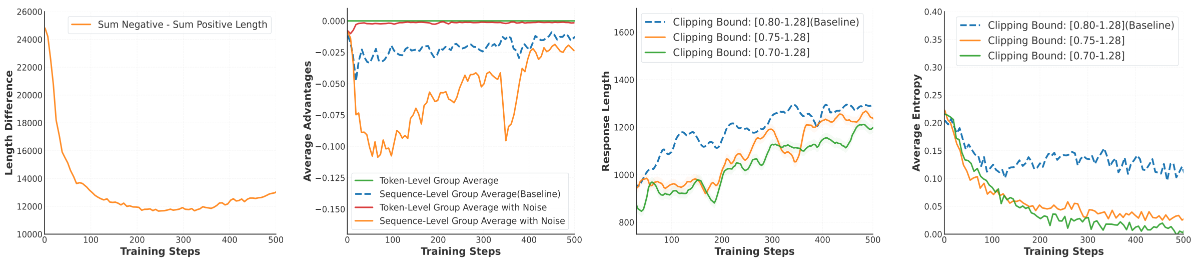
$$\mathcal{L}^{\text{DAPO}} = \mathbb{E}_{(q,o)}\Bigg[\frac{1}{\sum_{i=1}^{G}|o_i|}\sum_{i=1}^{G}\sum_{t=1}^{|o_i|}\min\Big(r_{i,t}A_{i,t}, \text{clip}(r_{i,t}, 1-\epsilon_{\text{low}}, 1+\epsilon_{\text{high}})\hat{A}_{i,t} \quad (3)$$

The asymmetric bounds $\epsilon_{\text{high}} > \epsilon_{\text{low}}$ (typically $\epsilon_{\text{high}} = 0.28$, $\epsilon_{\text{low}} = 0.2$) allow larger probability increases for high-entropy token exploration.

DAPO with Forking Tokens. (Wang et al., 2025b) demonstrates that a minority of high entropy tokens function as critical decision points. This motivates selective optimization focusing on high-entropy tokens:

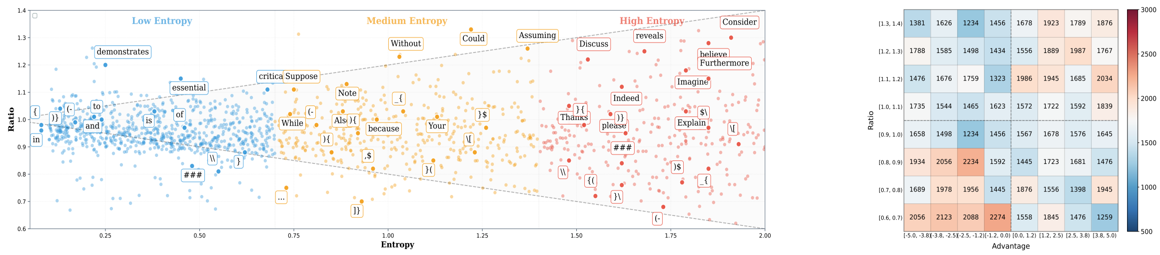
$$\mathcal{L}^{\text{Ent}} = \mathbb{E}_{(q,o)}\Bigg[\frac{1}{\sum_{i=1}^{G}|o_i|}\sum_{i=1}^{G}\sum_{t=1}^{|o_i|}\mathbb{I}[H_t \geq \tau_\rho]\cdot \min\Big(r_{i,t}A_{i,t}, \text{clip}(r_{i,t}, 1-\epsilon_{\text{low}}, 1+\epsilon_{\text{high}})A_{i,t}\Big) \quad (4)$$

where H_t represents the entropy at position t, τ_ρ denotes the ρ-th percentile entropy threshold.



(a) Length Differences (b) Average Advantage (c) Response Length (d) Average Entropy

Figure 4: Bias of token-mean loss toward negative samples affects advantage and clipping.



(a) Token Distribution: Entropy-Dependent Ratio Spread (b) Ratio-Advantage Heatmap

Figure 5: Token update patterns analysis via importance ratios

3 Token Heterogeneity in RLHF

3.1 Why Heterogeneity Is Necessary

The Dual-Entropy Phenomenon. We first analyzed entropy patterns during training. As shown in Figures 2, we discovered "Dual-Entropy Phenomenon": high-entropy tokens frequently have "twin siblings" in low-entropy region—tokens with same word stems but drastically different entropy. We show more visualizations in Appendix G.

This pattern spans multiple semantic categories and frequency ranges, explaining why low-entropy tokens are crucial: they contain important reasoning tokens that differ from high-entropy ones only in syntactic context. This reveals limitations of entropy-based binary discretization and uniform treatment, necessitating a nuanced approach that preserves low-entropy tokens' stabilizing influence without letting them dominate the learning signal.

3.2 Rollout Generation: Exploration-Exploitation Imbalance

Characterizing Tokens in Rollout Generation. We then analyze token distributions and occurrence frequencies in Figure 3a, 3b. We find that high-entropy tokens are rare. These tokens have low sampling probabilities but predominantly represent critical decision points. This creates a paradox: the tokens most valuable for learning are systematically under-sampled during standard rollout.

Temperature Effects on Critic Tokens. To address this, we investigate temperature's role in generating critical tokens. We evaluate rollout across various temperatures. Figure 3c, 3d reveal that while accuracy decreases with temperature, critical token occurrence increases before dropping at extremes. This demonstrates a fundamental trade-off: low temperatures preserve accuracy but suppress critical tokens, while high temperatures generate more meaningful tokens but reduce accuracy.

3.3 Advantage Calculation: Reward Attribution Granularity

Limitations of Sequence-Level Advantage Estimation. Current methods adopt GRPO's sequence-level normalization to estimate advantages—using a single reward to represent an entire sequence for estimation, disregarding length differences across sequences. Figure 4a reveals that negative samples are significantly longer than positive ones. GRPO divides loss by sequence length (Equation 1) to balance contributions but sacrifices individual token gradients. DAPO improves upon this by dividing by total tokens (Equation 3) to preserve token-level gradients—a natural approach for handling token heterogeneity. However, this creates a conflict: sequence-level advantages estimation with token-mean loss causes longer negative samples to dominate gradient, as shown in Figure 4b.

While DAPO's token-mean loss achieves better

(a) Entropy-Clip Patterns (b) Right-Clipped (c) Left-Clipped (d) Critical Token Clipped

Figure 6: Token-Level clipping patterns and word cloud visualizations

heterogeneity, sequence-level advantage estimation introduce negative length bias that impacts token-level optimization. First, when minor perturbations are applied to advantages, the ignorance of length in sequence-level advantage estimation amplifies these fluctuations into substantial deviations, causing training instability as shown in Figure 4b. Second, this bias conflicts with strategies like Archer, which relaxes left clipping bounds for high-entropy tokens. As shown in Figure 4c and Figure 4d, negative advantage dominance combined with relaxed left clipping bounds causes rapid probability degradation for high-entropy critical tokens, leading to entropy collapse and reduced response lengths.

Analyzing Token Update Patterns. Beyond sequence-level advantage ignoring token heterogeneity, another limitation is that all tokens within a sequence receive same advantages. We conduct a fine-grained analysis of advantage dynamics during training. Specifically, we introduce importance ratios $r_{i,t} = \pi_\theta(a_{i,t}|s_{i,t})/\pi_{\text{ref}}(a_{i,t}|s_{i,t})$, which reflect model's existing optimization state and actual update needs for each token.

Figure 5a shows high-entropy tokens have broader ratio distributions, with most tokens deviating from 1.0 as the model preferentially updates them. By analyzing the distribution of top 20% high-entropy tokens across different advantage and ratio values in Figure 5b, we observe a strong correlation between advantages and ratios. Meanwhile, we find that low-entropy tokens' ratios cluster around 1.0, suggesting no clear update direction. The token examples in Figure 5a demonstrate that tokens with large ratio deviations are those requiring the most optimization—either critical tokens or redundant tokens.

These findings demonstrate that tokens with similar entropy often have vastly different optimization needs. This insight exposes critical flaws in prior works that leverage entropy to reshape advantages. Methods like Entropy Adv incorporate entropy as an intrinsic reward added to the advantage (Eq. 2), which boosts all high-entropy tokens. For high-entropy tokens with negative advantages, this entropy reward offsets the negative advantage of erroneous tokens, dampening penalties for errors and thereby weakening the training signal.

3.4 Clipping Loss Computation: Clipping Constraint

Analyzing Clipped Token Properties. Figures 6 reveal distinct clipping patterns: low-entropy tokens, mostly format and useless code tokens, are frequently left-clipped, while high-entropy tokens including critical reasoning tokens are frequently right-clipped. Figure 6d shows that even with DAPO's clip higher, high-entropy tokens remain frequently clipped, with frequency increasing with entropy. Existing clipping strategies prematurely exclude these tokens from gradient updates. Uniform clipping constrains both useless low-entropy tokens that need probability reduction and critical high-entropy tokens that require exploration. DAPO's static high clipping bound is inadequate for adapting to token-specific entropy and update requirements. Furthermore, its restrictive left bound hinders the model from aggressively suppressing useless low-entropy tokens.

4 Heterogeneous Adaptive Policy Optimization

Based on the challenges in Section 3, we propose HAPO that optimizes the RL pipeline according to token heterogeneity. We introduce our core principles and formalize how entropy serves as a continuous regulator, enabling smooth transitions between different tokens while maintaining simplicity.

4.1 Entropy-Based Continuous Regulation

As we use entropy as a continuous indicator, we first prepare the entropy to obtain a unified regulation signal. Following (Wang et al., 2025b), we normalize entropy within each batch and apply asymmetric scaling to map values to $[-1, 1]$:

$$h_{i,t} = \frac{\log(H_{i,t}) - Q_\rho(\log(H))}{\sigma(\log(H))} \tag{5}$$

$$\tilde{h}_{i,t} = \begin{cases} h_{i,t}/h_{\max} & \text{if } h_{i,t} > 0 \\ -h_{i,t}/|h_{\min}| & \text{if } h_{i,t} \leq 0 \end{cases} \tag{6}$$

where $H_{i,t}$ is the entropy at i, t, Q_ρ denotes the ρ-th quantile serving as the boundary between high and low entropy tokens, and σ is the standard deviation.

We apply logarithmic smoothing to handle the significant variance in entropy distribution. The asymmetric scaling ensures balanced regulation across both positive and negative ranges.

With this normalized entropy $\tilde{h}_{i,t}$, we implement adaptive modulation across multiple components:

$$\text{Parameter}_{i,t} = \text{Parameter}_{\text{base}} \cdot f(\tilde{h}_{i,t}) \tag{7}$$

where $f(\cdot)$ is designed for each specific parameters.

Equation 7 serves as a meta-equation for our adaptive strategy. We apply this logic by substituting the parameters with sampling temperature, advantage estimates, and clipping bounds.

4.2 Adaptive Temperature Sampling

Now that we understand temperature's impact on token generation and accuracy, a balanced strategy is to introduce dynamic temperature. We apply high temperature to high-entropy tokens to encourage exploration while using low temperature for low-entropy ones to maintain coherence. Since rollout proceeds token-by-token without sequence-level context, we maintain the formulation of Equation 7 but dynamically adjust temperature based on current state, computing normalization on-the-fly:

$$T_{i,t} = T_{\text{base}} \cdot \left(1 + \frac{\log(H_{i,t}) - \hat{\rho}_{\log(H)}}{\hat{\sigma}_{\log(H)}} \cdot \tau\right) \tag{8}$$

$\hat{\rho}_{\log(H)}$ is the ρ-th quantile of entropy and serves as the direction of temperature adjustments. $\hat{\sigma}_{\log(H)}$ is the variance, and τ determines the maximum bounds of temperature adjustment. We compute $\hat{\rho}_{\log(H)}$ and $\hat{\sigma}_{\log(H)}$ by leveraging token statistics from previous step to capture more comprehensive information, rather than relying on sequence-level normalization.

4.3 Token-Level Group Average Advantage Estimation

As DAPO computes loss through token-mean, a more compatible approach for advantage estimation is token-level averaging. We estimate advantages across all tokens in the group, distributing rewards to individual tokens before normalization:

$$A_{i,t} = \frac{a_{i,t} - \mu_{\text{tok}}}{\sigma_{\text{tok}}} \tag{9}$$

where $a_{i,t} = r_i \in \{0, 1\}$ is the token-level reward inherited from sequence i, $\mu_{\text{tok}} = \frac{1}{|\mathcal{T}|}\sum_{(i,t)\in\mathcal{T}} a_{i,t}$ is the mean across all tokens, $\sigma_{\text{tok}} = \sqrt{\frac{1}{|\mathcal{T}|}\sum_{(i,t)\in\mathcal{T}}(a_{i,t} - \mu_{\text{tok}})^2}$ is the standard deviation across all tokens, $\mathcal{T} = \{(i,t)\}$ represents all tokens across sequences in the group.

The key distinction from sequence-level advantage estimation lies in granularity. Instead of using a single reward per sequence, we assign rewards to individual tokens then involve all tokens in advantage estimation. This ensures $\sum_{(i,t)\in\mathcal{T}} \hat{A}_{i,t} = 0$, resolving the gradient bias in token-mean loss. More importantly, our method preserves length-dependent gradient scaling within reward categories—longer sequences contribute larger gradients for complex reasoning. This simultaneously maintains GRPO's unbiased gradient contributions and DAPO's token-level granularity advantages.

4.4 Differential Advantage Redistribution

Our core idea remains amplifying advantages for high-entropy tokens while reducing them for low-entropy tokens. Different from existing methods, we additionally use importance ratios for regulation, ensuring updates based on each token's actual conditions to achieve token-level control.

To implement this, we define a neutral zone $[\gamma_L, \gamma_U]$, typically centered around ratio = 1. High-entropy tokens receive amplification only when their importance ratio falls outside the neutral zone, reflecting clear update trends. We enhance them accordingly. For low-entropy tokens, we apply suppression within the neutral zone to reduce the influence of tokens lacking update tendency. This aligns with our analysis, where ratios far from 1 typically represent relatively important tokens, and we assign them relatively larger advantages:

$$\hat{A}_{i,t} = \begin{cases} A_{i,t} \cdot (1 + \tilde{h}_{i,t}) & \text{if } C(\tilde{h}_{i,t}, r_{i,t}) \\ A_{i,t} & \text{otherwise} \end{cases} \tag{10}$$

Table 1: Comparison between *HAPO* and existing methods on Math benchmarks. The models below are from 1000 steps of training. The best results are indicated by **boldface**.

Model	Method	**Avg (1590)**	AIME24(30)	AIME25(30)	AMC(83)	Math(500)	OlympiadBench(675)	Minerva (272)
Qwen2.5-Math-1.5B	Vanilla GRPO	36.83	20.12	15.67	71.23	65.91	29.84	18.23
	Vanilla DAPO	38.34	21.73	17.11	72.85	**67.33**	31.47	19.52
	DAPO w/ Forking Tokens	38.83	22.51	18.56	74.80	65.54	32.81	18.74
	Archer	37.01	19.62	17.24	72.17	64.21	29.98	19.14
	Entropy Env	37.45	19.45	18.02	72.88	65.13	29.56	19.67
	EDGE-GRPO	38.79	22.88	17.86	74.48	66.94	30.67	19.93
	HAPO(ours)	**40.62**	**25.33**	**20.12**	**75.69**	66.83	**33.86**	**21.91**
Qwen2.5-Math-7B	Vanilla GRPO	45.26	35.67	18.91	80.14	74.28	32.76	29.85
	Vanilla DAPO	46.97	37.24	20.24	81.99	76.72	34.35	31.28
	DAPO w/ Forking Tokens	47.43	38.45	21.90	80.55	75.38	35.52	32.79
	Archer	45.63	35.25	19.19	82.53	72.16	33.04	31.61
	Entropy Env	46.15	36.82	18.95	82.94	73.05	33.68	31.43
	EDGE-GRPO	45.89	38.85	21.83	83.23	69.78	32.43	29.24
	HAPO(ours)	**50.04**	**41.31**	**24.34**	**85.47**	**78.45**	**36.94**	**33.73**
Qwen3-8B	Vanilla GRPO	48.15	33.91	23.56	76.84	81.17	42.93	30.47
	Vanilla DAPO	50.00	35.84	25.24	78.37	83.54	44.87	32.13
	DAPO w/ Forking Tokens	50.21	36.22	25.85	80.26	82.75	**46.27**	29.91
	Archer	49.17	34.76	23.33	78.47	82.57	42.69	33.22
	Entropy Env	49.60	34.45	24.12	80.05	83.12	42.88	32.95
	EDGE-GRPO	50.06	35.15	24.12	80.47	83.33	43.35	33.91
	HAPO(ours)	**51.97**	**39.01**	**26.83**	**81.77**	**84.23**	45.26	**34.74**
Qwen3-14B	Vanilla GRPO	55.89	42.36	35.78	83.91	85.63	50.14	37.52
	Vanilla DAPO	58.06	44.82	37.53	85.52	88.25	52.28	39.97
	DAPO w/ Forking Tokens	60.00	46.71	39.12	87.31	88.90	**56.85**	40.84
	Archer	58.88	46.28	38.68	86.73	87.76	53.19	40.62
	Entropy Env	59.27	46.35	39.22	87.04	88.35	54.12	40.56
	EDGE-GRPO	59.16	46.94	39.01	87.15	89.12	52.45	40.31
	HAPO(ours)	**62.09**	**49.74**	**41.71**	**89.66**	**92.47**	56.43	**42.53**

Table 2: Comparison between *HAPO* and existing methods on LLaMA series models.

Model	Method	**Avg (1590)**	AIME24(30)	AIME25(30)	AMC(83)	Math(500)	OlympiadBench(675)	Minerva (272)
LLaMA3.2-3B Instruct	Vanilla GRPO	20.97	11.67	0.00	49.28	38.62	14.35	11.94
	Vanilla DAPO	22.99	13.93	0.00	52.11	41.95	16.63	13.32
	DAPO w/ Forking Tokens	23.86	12.45	0.00	55.96	43.34	16.89	14.54
	HAPO(ours)	**27.42**	**17.50**	**8.00**	**58.74**	**44.86**	**19.37**	**16.05**
LLaMA3.1-8B-Instruct	Vanilla GRPO	22.07	10.83	0.00	54.16	40.78	13.81	12.89
	Vanilla DAPO	24.21	13.33	0.00	57.11	44.05	16.03	14.76
	DAPO w/ Forking Tokens	25.30	14.50	0.00	58.46	46.55	15.89	16.39
	HAPO(ours)	**27.06**	**19.17**	**1.25**	**59.32**	**47.83**	**17.66**	**17.11**

$$C(\tilde{h}_{i,t}, r_{i,t}) = \begin{cases} r_{i,t} \notin [\gamma_L, \gamma_U] & \text{if } \tilde{h}_t > 0 \\ r_{i,t} \in [\gamma_L, \gamma_U] & \text{if } \tilde{h}_t \le 0 \end{cases} \tag{11}$$

4.5 Asymmetric Adaptive Clipping

Unlike Archer (Wang et al., 2025a) that uniformly adjusts clipping bounds, we apply asymmetric adjustments respecting natural update tendencies of each token. By lowering the left boundary for low-entropy tokens, we allow these tokens to continue participating in the update process and are more likely to decrease substantially. By raising the right boundary for high-entropy tokens, we encourage exploration at critical decision points.

$$\epsilon_L(i,t) = \begin{cases} \epsilon_L^{\text{base}}(1 - \tilde{h}_{i,t}) & \text{if } \tilde{h}_{i,t} \le 0 \\ \epsilon_L^{\text{base}} & \text{if } \tilde{h}_{i,t} > 0 \end{cases} \tag{12}$$

$$\epsilon_R(i,t) = \begin{cases} \epsilon_R^{\text{base}} & \text{if } \tilde{h}_{i,t} \le 0 \\ \epsilon_R^{\text{base}}(1 + \tilde{h}_{i,t}) & \text{if } \tilde{h}_{i,t} > 0 \end{cases} \tag{13}$$

where $\epsilon_L^{\text{base}}, \epsilon_R^{\text{base}}$ are the base clipping bounds.

These continuous formulations unify our heterogeneous strategy, with all modulations smoothly varying as functions of normalized entropy $\tilde{h}$, ensuring a concise implementation. As entropy is already computed during sampling, HAPO introduces virtually no additional computational overhead. We present the pseudocode of HAPO in Appendix C.

5 Experiments

Experimental Setup. We incorporate our token-level strategy into DAPO (Yu et al., 2025) within verl (Sheng et al., 2025) and vLLM (Kwon et al., 2023). We maintain DAPO's settings: clip-higher of $\epsilon_{\text{high}} = 0.28$ and $\epsilon_{\text{low}} = 0.2$; overlong reward shaping with 10240-token maximum length and 4096-token cache. We use a training batch size of 512 and a mini-batch size of 32. Training proceeds

Table 3: Comparison between *HAPO* and existing methods on Code and Logic benchmarks.

Model	Method	Logic RL					LiveCodeBench		
		Avg	4ppl	5ppl	6ppl	7ppl	**Avg**	V5(24/8/1-25/2/1)	V6(25/2/1-25/5/1)
Qwen2.5-7B-Instruct-1M	Vanilla DAPO	87.13	92.23	90.16	85.72	80.42	29.08	28.22	29.94
	DAPO w/ Forking Tokens	90.15	93.44	92.13	90.26	84.75	31.46	29.72	33.19
	Archer	87.68	94.12	89.75	85.28	81.55	29.75	28.63	30.86
	EDGE-GRPO	88.92	92.35	91.13	89.58	82.63	30.40	29.27	31.53
	HAPO(ours)	**92.69**	**95.41**	**93.34**	**93.26**	**88.74**	**34.02**	**32.55**	**35.48**

Table 4: Component ablation of HAPO. We denote Adaptive Temperature Sampling, Token-Level Group Average Advantage Estimation, Differential Advantage Redistribution, and Asymmetric Adaptive Clipping as A, B, C, and D respectively for convenience.

A	B	C	D	Avg (1560)	AIME24 (30)	AIME25 (30)	AMC (83)	Math (500)	OlympiadBench(675)	Minerva (272)
				46.97	37.24	20.24	81.99	76.72	34.35	31.28
✓				48.85	39.77	23.01	84.20	76.88	**37.34**	31.92
	✓			48.56	39.04	23.77	83.59	77.19	35.02	32.76
		✓		48.28	38.63	22.07	84.19	76.58	36.00	32.19
			✓	48.02	38.38	20.73	82.45	78.56	36.49	31.51
		✓	✓	48.74	39.19	22.55	84.43	77.72	36.13	32.44
	✓	✓	✓	49.42	40.17	23.62	84.41	**79.10**	35.96	33.28
✓	✓	✓	✓	**50.04**	**41.31**	**24.34**	**85.47**	78.45	36.94	**33.73**

with a learning rate of 10^{-6}.

For HAPO's token-level strategies, we follow (Wang et al., 2025b) and set the quantile ρ to 80%, encouraging exploration to the top 20% highest-entropy tokens. For Adaptive Temperature Sampling, we compute $\hat{\rho}_{\log(H)}$ and $\hat{\sigma}_{\log(H)}$ based on all tokens' entropy in the previous step. We set $T_{\text{base}} = 1.0$ and $\tau = 0.1$. For Differential Advantage Redistribution, we determine the neutral zone using the clipping ratios that correspond to each token's actual dynamics, setting it to $[1 - \frac{\epsilon_L}{2}, 1 + \frac{\epsilon_R}{2}]$. For Asymmetric Adaptive Clipping, we set $\epsilon_L^{\text{base}} = 0.2$ and $\epsilon_R^{\text{base}} = 0.28$. We adopt a fixed hyperparameter setting across all models to explicitly showcase the generalizability of our method. We also provide hyperparameter ablations in Appendix D.

We perform experiments across Qwen2.5-Math-1.5B,Qwen2.5-Math-7B (Yang et al., 2024) base models and Qwen3-8B,Qwen3-14B (Yang et al., 2025) base models, using DAPO-Math-17K (Yu et al., 2025) as the training dataset. We trained all models on 4 nodes with 32 A100 GPUs.

Evaluation. We evaluate all the models on mathematical benchmarks: AIME24, AIME25, AMC23, Minerva (Lewkowycz et al., 2022), MATH500 (Hendrycks et al., 2021), and OlympiadBench (He et al., 2024). For each question, we generate 8 responses under a decoding temperature $T = 0.5$, and report the average accuracy.

Main Results. We compare HAPO against vanilla GRPO, vanilla DAPO, DAPO with Forking Tokens, Archer, Entropy Env and EDGE-GRPO in Table 1. On Qwen2.5-Math-7B, HAPO surpasses DAPO w/ Forking Tokens by 2.86 and 2.44 points on AIME'24 and AIME'25. When compared to entropy-based approaches, HAPO outperforms Archer by 2.80-4.41 points and Entropy Adv by 2.37-4.25 points, highlighting the effectiveness of our fine-grained heterogeneous treatment. Overall, HAPO outperforms these methods across models of different scales, demonstrating strong scalability. We visualize training dynamics in Figures 7, 8, 9, comparing HAPO and DAPO throughout the training process. Our method maintains longer response lengths and higher accuracy, indicating that HAPO preserves model exploration capabilities while achieving better task performance.

6 Ablations

Results on Models Other Than Qwen. We employ same experiments on LLaMA models. Results are presented in Table 2. HAPO demonstrates more pronounced improvements on weaker models, suggesting that our fine-grained heterogeneous treatment effectively helps models capture critical information, further validating HAPO's scalability. In Figure 10, we show training dynamics of LLaMA3.2-3B. HAPO exhibits higher entropy and longer responses, enabling the model to maintain continuous exploration.

Generalization to Other Domains. We validate HAPO's generalizability in code and logic. We use code data from Archer (Wang et al., 2025a) and logic data from Logic-RL (Xie et al., 2025). We employ Qwen2.5-7B-Instruct-1M due to its general capabilities and the fact that it has not been pre-trained on these datasets. Evaluation on LiveCodeBench (Jain et al., 2024) and Logic in Table 3 shows that HAPO consistently outperforms DAPO w/ Forking Tokens. This demonstrates that token heterogeneity exists across different domains, and HAPO's approach naturally generalizes to diverse areas, confirming HAPO's broad applicability.

Component-wise Contributions. We examine each component's contribution in Table 4. All components contribute to HAPO's performance, with Adaptive Temperature Sampling proving particularly crucial as it governs token distribution and entropy. Token-Level Group Average Advantage Estimation significantly impacts advantage redistribution and clipping, corroborating our analysis. In fact, all components form an organic unity: Adaptive Temperature Sampling ensures sufficient high-entropy critical tokens, Differential Advantage Redistribution optimizes them with larger updates, and Asymmetric Adaptive Clipping encourages exploration. This synergy enables effective leverage of token heterogeneity throughout optimization.

7 Conclusion

We introduce Heterogeneous Adaptive Policy Optimization (HAPO), which leverages entropy to drive the entire optimization process and implements fine-grained token-level optimization. Experiments demonstrate consistent improvements over DAPO. HAPO establishes that effective RLHF requires adapting to token heterogeneity, opening new ways for developing more capable reasoning models.

Limitations

Despite HAPO's strong empirical performance, we acknowledge several limitations that provide opportunities for future refinement:

Entropy as Primary Heterogeneity Signal. While our entropy-based approach effectively identifies token heterogeneity, it may not capture all aspects of token importance. Tokens with moderate entropy could still be semantically crucial, particularly in structured or deterministic outputs. Future work could explore complementary signals like gradient magnitudes or attention weights.

Static Entropy Thresholds. HAPO uses a fixed 80th percentile to distinguish high/low entropy tokens throughout training, though entropy distributions evolve as training progresses. Dynamic threshold adaptation could potentially improve performance but would add complexity.

These limitations suggest directions for future refinement.

Acknowledgments

This work is supported by Fundamental and Interdisciplinary Disciplines Breakthrough Plan of the Ministry of Education of China (JYB2025XDXM113), National Natural Science Foundation of China (92470121, 62402016), National Key R&D Program of China (2024YFA1014003), Zhongguancun Academy (C20250204, C20250602), Beijing Major Science and Technology Project (Z251100008125043, Z251100008425023), and High-performance Computing Platform of Peking University.

References

Shivam Agarwal, Zimin Zhang, Lifan Yuan, Jiawei Han, and Hao Peng. 2025. The unreasonable effectiveness of entropy minimization in llm reasoning. *Preprint*, arXiv:2505.15134.

Anthropic. 2024. Claude 3.5 sonnet.

Daixuan Cheng, Shaohan Huang, Xuekai Zhu, Bo Dai, Wayne Xin Zhao, Zhenliang Zhang, and Furu Wei. 2025. Reasoning with exploration: An entropy perspective. *Preprint*, arXiv:2506.14758.

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. *Preprint*, arXiv:2505.22617.

DeepSeek-AI, Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, Xiaokang Zhang, Xingkai Yu, Yu Wu, Z. F. Wu, Zhibin Gou, Zhihong Shao, Zhuoshu Li, Ziyi Gao, and 181 others. 2025. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *Preprint*, arXiv:2501.12948.

Kawin Ethayarajh, Winnie Xu, Niklas Muennighoff, Dan Jurafsky, and Douwe Kiela. 2024. Kto: Model alignment as prospect theoretic optimization. *Preprint*, arXiv:2402.01306.

Zitian Gao, Lynx Chen, Joey Zhou, and Bryan Dai. 2025. One-shot entropy minimization. *Preprint*, arXiv:2505.20282.

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. 2024. Olympiadbench: A challenging benchmark for promoting agi with olympiad-level bilingual multimodal scientific problems. *Preprint*, arXiv:2402.14008.

Dan Hendrycks, Collin Burns, Saurav Kadavath, Akul Arora, Steven Basart, Eric Tang, Dawn Song, and Jacob Steinhardt. 2021. Measuring mathematical problem solving with the math dataset. *Preprint*, arXiv:2103.03874.

Jian Hu, Jason Klein Liu, Haotian Xu, and Wei Shen. 2025. Reinforce++: An efficient rlhf algorithm with robustness to both prompt and reward models. *Preprint*, arXiv:2501.03262.

Naman Jain, King Han, Alex Gu, Wen-Ding Li, Fanjia Yan, Tianjun Zhang, Sida Wang, Armando Solar-Lezama, Koushik Sen, and Ion Stoica. 2024. Livecodebench: Holistic and contamination free evaluation of large language models for code. *Preprint*, arXiv:2403.07974.

Woosuk Kwon, Zhuohan Li, Siyuan Zhuang, Ying Sheng, Lianmin Zheng, Cody Hao Yu, Joseph E. Gonzalez, Hao Zhang, and Ion Stoica. 2023. Efficient memory management for large language model serving with pagedattention. *Preprint*, arXiv:2309.06180.

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. *Preprint*, arXiv:2206.14858.

Honglin Lin, Zheng Liu, Yun Zhu, Chonghan Qin, Juekai Lin, Xiaoran Shang, Conghui He, Wentao Zhang, and Lijun Wu. 2026. Mmfinereason: Closing the multimodal reasoning gap via open data-centric methods. *arXiv preprint arXiv:2601.21821*.

Zheng Liu, Honglin Lin, Chonghan Qin, Xiaoyang Wang, Xin Gao, Yu Li, Mengzhang Cai, Yun Zhu, Zhanping Zhong, Qizhi Pei, and 1 others. 2026. Chartverse: Scaling chart reasoning via reliable programmatic synthesis from scratch. *arXiv preprint arXiv:2601.13606*.

Zheng Liu, Mengjie Liu, Jingzhou Chen, Jingwei Xu, Bin Cui, Conghui He, and Wentao Zhang. Flare: Fully integration of vision-language representations for deep cross-modal understanding. In *The Fourteenth International Conference on Learning Representations*.

Zheng Liu, Mengjie Liu, Jingzhou Chen, Jingwei Xu, Bin Cui, Conghui He, and Wentao Zhang. 2025a. Fusion: Fully integration of vision-language representations for deep cross-modal understanding. *arXiv preprint arXiv:2504.09925*.

Zichen Liu, Changyu Chen, Wenjun Li, Penghui Qi, Tianyu Pang, Chao Du, Wee Sun Lee, and Min Lin. 2025b. Understanding r1-zero-like training: A critical perspective. *Preprint*, arXiv:2503.20783.

Yu Meng, Mengzhou Xia, and Danqi Chen. 2024. Simpo: Simple preference optimization with a reference-free reward. *Preprint*, arXiv:2405.14734.

OpenAI. 2024. Learning to reason with llms.

Long Ouyang, Jeff 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. *Preprint*, arXiv:2203.02155.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Stefano Ermon, Christopher D. Manning, and Chelsea Finn. 2024. Direct preference optimization: Your language model is secretly a reward model. *Preprint*, arXiv:2305.18290.

John Schulman, Sergey Levine, Philipp Moritz, Michael I. Jordan, and Pieter Abbeel. 2017a. Trust region policy optimization. *Preprint*, arXiv:1502.05477.

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

ByteDance Seed, :, Jiaze Chen, Tiantian Fan, Xin Liu, Lingjun Liu, Zhiqi Lin, Mingxuan Wang, Chengyi Wang, Xiangpeng Wei, Wenyuan Xu, Yufeng Yuan, Yu Yue, Lin Yan, Qiying Yu, Xiaochen Zuo, Chi Zhang, Ruofei Zhu, Zhecheng An, and 255 others. 2025. Seed1.5-thinking: Advancing superb reasoning models with reinforcement learning. *Preprint*, arXiv:2504.13914.

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. *Preprint*, arXiv:2402.03300.

Guangming Sheng, Chi Zhang, Zilingfeng Ye, Xibin Wu, Wang Zhang, Ru Zhang, Yanghua Peng, Haibin Lin, and Chuan Wu. 2025. Hybridflow: A flexible and efficient rlhf framework. In *Proceedings of the Twentieth European Conference on Computer Systems*, page 1279–1297. ACM.

Jiakang Wang, Runze Liu, Fuzheng Zhang, Xiu Li, and Guorui Zhou. 2025a. Stabilizing knowledge, promoting reasoning: Dual-token constraints for rlvr. *Preprint*, arXiv:2507.15778.

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. 2025b. Beyond the 80/20 rule: High-entropy minority tokens drive effective reinforcement learning for llm reasoning. *Preprint*, arXiv:2506.01939.

Tian Xie, Zitian Gao, Qingnan Ren, Haoming Luo, Yuqian Hong, Bryan Dai, Joey Zhou, Kai Qiu, Zhirong Wu, and Chong Luo. 2025. Logic-rl: Unleashing llm reasoning with rule-based reinforcement learning. *Preprint*, arXiv:2502.14768.

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. *Preprint*, arXiv:2505.09388.

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. 2024. Qwen2.5-math technical report: Toward mathematical expert model via self-improvement. *Preprint*, arXiv:2409.12122.

Qiying Yu, Zheng Zhang, Ruofei Zhu, Yufeng Yuan, Xiaochen Zuo, Yu Yue, Weinan Dai, Tiantian Fan, Gaohong Liu, Lingjun Liu, Xin Liu, Haibin Lin, Zhiqi Lin, Bole Ma, Guangming Sheng, Yuxuan Tong, Chi Zhang, Mofan Zhang, Wang Zhang, and 16 others. 2025. Dapo: An open-source llm reinforcement learning system at scale. *Preprint*, arXiv:2503.14476.

Yu Yue, Yufeng Yuan, Qiying Yu, Xiaochen Zuo, Ruofei Zhu, Wenyuan Xu, Jiaze Chen, Chengyi Wang, TianTian Fan, Zhengyin Du, Xiangpeng Wei, Xiangyu Yu, Gaohong Liu, Juncai Liu, Lingjun Liu, Haibin Lin, Zhiqi Lin, Bole Ma, Chi Zhang, and 8 others. 2025. Vapo: Efficient and reliable reinforcement learning for advanced reasoning tasks. *Preprint*, arXiv:2504.05118.

Weihao Zeng, Yuzhen Huang, Qian Liu, Wei Liu, Keqing He, Zejun Ma, and Junxian He. 2025. Simplerl-zoo: Investigating and taming zero reinforcement learning for open base models in the wild. *Preprint*, arXiv:2503.18892.

Xingjian Zhang, Siwei Wen, Wenjun Wu, and Lei Huang. 2025. Edge-grpo: Entropy-driven grpo with guided error correction for advantage diversity. *Preprint*, arXiv:2507.21848.

Appendix

A Related Work

A.1 Reinforcement Learning from Human Feedback

The evolution of reinforcement learning in language models represents a trajectory from basic preference alignment toward sophisticated reasoning capabilities. Foundational work in constrained optimization emerged through TRPO (Schulman et al., 2017a) and PPO (Schulman et al., 2017b), establishing principles for stable policy updates. Subsequent algorithmic innovations eliminated computational bottlenecks—GRPO (Shao et al., 2024) and REINFORCE++ (Hu et al., 2025) removed value network dependencies via group-based advantage computation. The offline optimization paradigm gained prominence through DPO (Rafailov et al., 2024), KTO (Ethayarajh et al., 2024), and SimPO (Meng et al., 2024), which circumvent reward model training entirely.

A fundamental transformation occurred with the advent of reasoning-centric models. OpenAI's o1 (OpenAI, 2024) pioneered effective multi-step reasoning through reinforcement learning at scale, catalyzing widespread development. DeepSeek-R1 (DeepSeek-AI et al., 2025) demonstrated reasoning emergence without supervised fine-tuning, while Claude3.5 (Anthropic, 2024), Qwen3 (Yang et al., 2025), and Seed-1.5-Thinking (Seed et al., 2025) expanded the reasoning frontier (Liu et al., 2025a; Liu et al.; Lin et al., 2026; Liu et al., 2026). Beyond these state-of-the-art models, various works explored complementary optimization strategies, including SimpleRLZoo (Zeng et al., 2025) and Open-Reasoner-Zero, which investigate alternative training paradigms. Meanwhile, algorithmic refinements continue to enhance existing methods—DAPO (Yu et al., 2025) introduced token-mean averaging and asymmetric clipping for extended sequences, while VAPO (Yue et al., 2025) developed variance-aware optimization strategies to improve training stability and convergence.

A.2 Entropy-Based Optimization in LLMs

Entropy has emerged as a fundamental signal for quantifying model uncertainty and identifying token heterogeneity in reasoning tasks. Pioneering works have established the theoretical and empirical basis for entropy-aware learning (Zeng et al., 2025; Liu et al., 2025b; Cui et al., 2025). analyzed how entropy dynamics correlate with reasoning capabilities, while some works (Wang et al., 2025b) proposed the "80/20 rule," demonstrating that a minority of high-entropy "forking tokens" effectively control reasoning diversity. Other studies (Agarwal et al., 2025; Gao et al., 2025) have shown that entropy minimization alone can serve as a powerful unsupervised signal.

Building upon these identification strategies, recent research has actively explored incorporating entropy into optimization mechanisms, typically following two paradigms:

Differentiated Clipping and Categorization. One line of work focuses on relaxing constraints for critical tokens. DAPO with forking tokens (Wang et al., 2025b) and Archer (Wang et al., 2025a) employ entropy to categorize tokens into high and low entropy groups, adjusting clipping boundaries to encourage exploration. However, these methods generally function as discrete filters, relying on rigid thresholds that create artificial discontinuities in the optimization landscape.

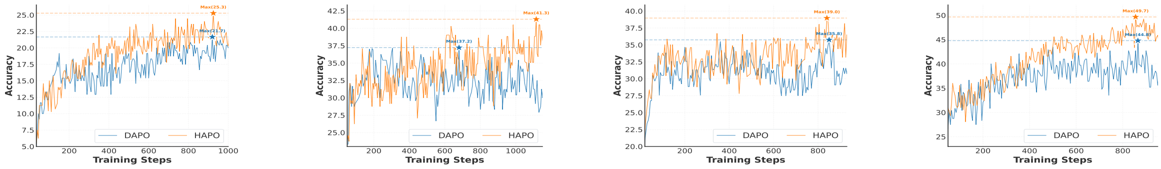
Advantage Modulation and Regularization. Another direction involves reshaping learning signals directly. Entropy Adv (Cheng et al., 2025) and EDGE-GRPO (Zhang et al., 2025) propose modulating advantages based on entropy, assigning higher weights or bonuses to confident or high-entropy tokens. While effective, these approaches typically treat entropy as an additive bonus or an auxiliary regulator. By superimposing entropy onto existing rewards rather than integrating it into the core functional dependencies of optimization, they often function as post-hoc regularizers.

B Training Dynamics of HAPO

Figures 7, 8, and 9 illustrate the training dynamics of HAPO versus DAPO across Qwen2.5 and Qwen 3 models, focusing on response length, entropy, and accuracy. Our observations reveal that HAPO sustains longer reasoning chains and higher accuracy levels, suggesting that it balances extensive exploration with precise execution. Additionally, comparative analysis on the LLaMA series in Figure 10 shows that HAPO exhibits robust training dynamics, further confirming its generalizability.

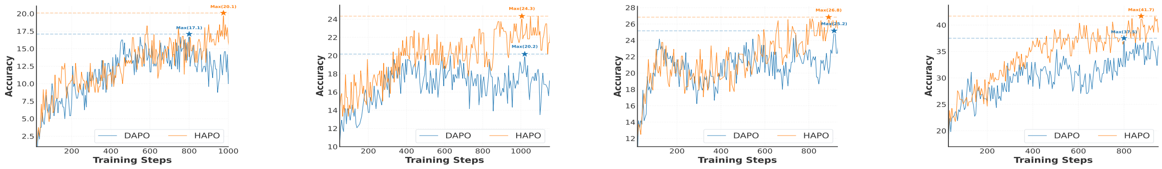
C HAPO Pseudocode

Figure 12 provides the complete pseudocode of HAPO based on the verl and vLLM implementation. The algorithm details the step-by-step process



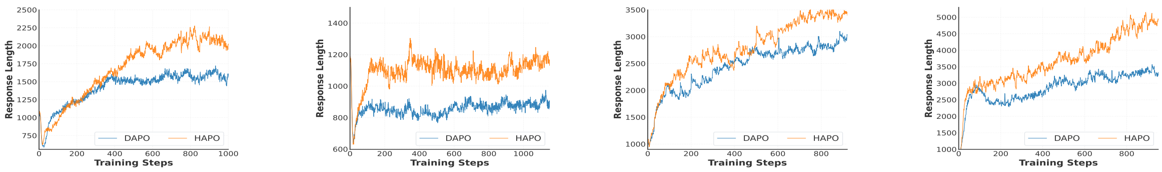
(a) Qwen2.5-Math-1.5B (b) Qwen2.5-Math-7B (c) Qwen3-8B (d) Qwen3-14B

Figure 7: Training dynamics of AIME24 results for HAPO and DAPO.



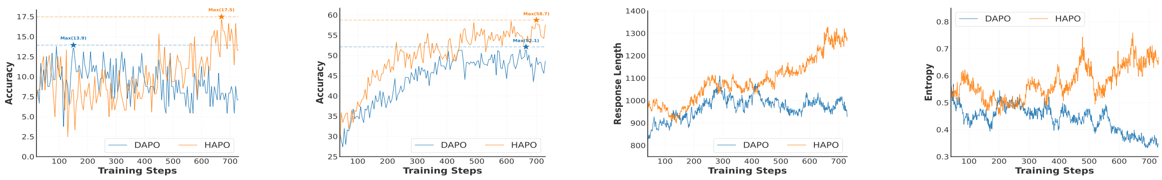
(a) Qwen2.5-Math-1.5B (b) Qwen2.5-Math-7B (c) Qwen3-8B (d) Qwen3-14B

Figure 8: Training dynamics of AIME25 results for HAPO and DAPO.



(a) Qwen2.5-Math-1.5B (b) Qwen2.5-Math-7B (c) Qwen3-8B (d) Qwen3-14B

Figure 9: Training dynamics of response length for HAPO and DAPO.



(a) AIME24 (b) AMC (c) Response Length (d) Entropy

Figure 10: Training dynamics of HAPO and DAPO on LLaMA3.2-3B-Instruct.

of: (1) computing entropy statistics for continuous regulation, (2) applying adaptive temperature during sampling, (3) performing token-level group average advantage estimation, (4) implementing differential advantage redistribution based on importance ratios and entropy values, and (5) executing asymmetric adaptive clipping for policy updates.

The pseudocode highlights HAPO's key innovation: using normalized entropy $\tilde{h}_{i,t} \in [-1, 1]$ as a continuous regulator that smoothly modulates all components without introducing significant computational overhead. The entropy computation required for sampling is reused throughout the pipeline, ensuring efficient implementation. The algorithm maintains simplicity while achieving fine-grained heterogeneous treatment, making it applicable to existing RL frameworks.

D Additional Hyperparameter Ablations

We conduct extensive ablation studies to validate our hyperparameter choices and design decisions. All experiments in this section are performed on Qwen2.5-Math-7B following the same training protocol described in Section 5.

Entropy Quantile Threshold. We investigate the impact of the entropy quantile threshold ρ, which determines the boundary between high and low entropy tokens. Table 5 presents results across different quantile values on Qwen2.5-Math-7B.

The results confirm that the 80th percentile provides optimal performance. The 70th percentile achieves comparable results, suggesting robustness around this range. Setting ρ to 90% degrades performance as this restrictive threshold identifies too few high-entropy tokens (only top 10%), limiting

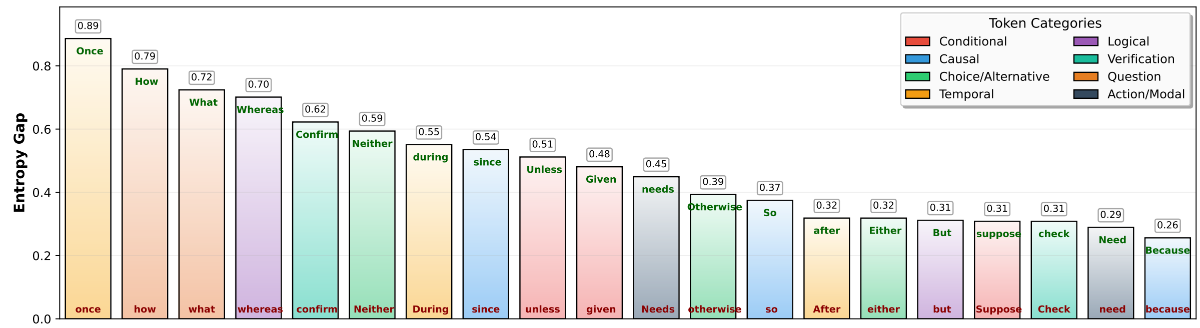


Figure 11: Dual-Entropy Token Frequency-Entropy Landscape

Table 5: Ablation study on entropy quantile threshold ρ using Qwen2.5-Math-7B. The 80th percentile achieves optimal balance.

Quantile	**Avg (1560)**	AIME24 (30)	AIME25 (30)	AMC (83)	Math (500)	OlympiadBench(675)	Minerva (272)
90%	48.96	39.44	22.73	84.11	77.74	36.61	33.14
80%	**50.04**	**41.31**	**24.34**	**85.47**	78.45	36.94	**33.73**
70%	49.74	40.78	24.11	84.96	**78.83**	**37.05**	32.71
50%	48.45	38.92	23.67	82.41	76.98	36.42	32.35
30%	47.10	37.05	23.12	81.58	75.21	34.39	31.26

exploration at critical points. Conversely, lower thresholds (50% and 30%) progressively worsen performance by classifying too many tokens as high-entropy, diluting the focused amplification effect. The 30% threshold particularly suffers, treating most tokens as high-entropy and applying unfocused modifications. This validates the 80th percentile as balancing critical token identification with sufficient coverage.

Temperature Adjustment Factor. The temperature adjustment factor τ controls the magnitude of dynamic temperature modulation. We evaluate its impact across a range of values in Table 6.

Our chosen value of $\tau = 0.1$ achieves the best overall performance. While $\tau = 0.05$ produces competitive results, this smaller adjustment factor limits the temperature modulation range too narrowly, preventing the mechanism from fully leveraging adaptive temperature benefits. The complete absence of temperature adjustment ($\tau = 0$) eliminates this exploration mechanism entirely. Conversely, larger values ($\tau \geq 0.3$ and $\tau \geq 0.5$) lead to excessive temperature modifications. Since high-entropy tokens inherently exhibit strong randomness, applying overly aggressive temperature increases to these already uncertain positions introduces extreme instability, disrupting coherent reasoning chains. This demonstrates the importance of balanced temperature modulation—sufficient to encourage exploration at critical points without transforming productive exploration into chaotic sampling.

Neutral Zone Configuration. The neutral zone $[\gamma_L, \gamma_U]$ determines when differential advantage redistribution is applied based on importance ratios. We examine various configurations based on the clipping bounds in Table 7.

Setting the neutral zone to half of the clipping bounds achieves optimal performance. The No Zone configuration, which applies no differential advantage redistribution, performs worst—without entropy-based modulation, the model treats all tokens equally regardless of their importance. Full Clip Bounds, which only suppresses low-entropy token advantages without amplifying high-entropy ones, shows marginal improvement but fails to encourage exploration at critical decision points. Zero Bounds performs better by only amplifying high-entropy token advantages. While this promotes exploration, it neglects to suppress trivial low-entropy tokens, leading to inefficient updates. **1/2 Clip Bounds** configuration achieves the best results by combining both mechanisms: it suppresses low-entropy tokens within the neutral zone while amplifying high-entropy tokens whose importance ratios indicate meaningful updates. This balanced approach ensures that computational resources focus on genuinely important tokens while reducing noise from trivial updates, demonstrating the necessity of bidirectional entropy-aware modulation.

Table 6: Ablation study on temperature adjustment factor τ using Qwen2.5-Math-7B. Moderate adjustment ($\tau = 0.1$) yields best results.

τ	**Avg (1560)**	AIME24 (30)	AIME25 (30)	AMC (83)	Math (500)	OlympiadBench(675)	Minerva (272)
0.0	49.42	40.17	23.62	84.41	**79.10**	35.96	33.28
0.05	49.58	40.86	24.05	83.98	78.74	36.52	33.31
0.1	**50.04**	**41.31**	**24.34**	**85.47**	78.45	**36.94**	**33.73**
0.3	49.11	40.67	23.61	83.13	77.58	36.89	32.36
0.5	48.12	38.92	22.45	83.27	76.83	35.14	32.08

Table 7: Ablation study on neutral zone configuration using Qwen2.5-Math-7B. Using half of the clipping bounds provides optimal control.

Neutral Zone	**Avg (1560)**	AIME24 (30)	AIME25 (30)	AMC (83)	Math (500)	OlympiadBench(675)	Minerva (272)
Full Clip Bounds	49.57	40.77	24.13	84.86	78.23	36.48	32.93
1/2 Clip Bounds	**50.04**	41.31	**24.34**	**85.47**	78.45	**36.94**	**33.73**
Zero Bounds	49.71	**41.42**	23.51	85.23	**78.62**	36.17	33.32
No Zone	49.37	40.83	23.67	84.58	77.94	36.56	32.61

These comprehensive ablation studies on Qwen2.5-Math-7B confirm that our hyperparameter choices are well-calibrated and that each design decision contributes meaningfully to HAPO's strong performance. The continuous, entropy-based regulation framework proves essential for effectively leveraging token heterogeneity in reinforcement learning optimization.

E Theoretical Analysis of Entropy-Driven Modulation

Below, we provide a formal theoretical analysis showing how entropy values drive each component and why HAPO systematically improves over DAPO by optimizing the gradient estimation process.

E.1 Gradient Estimation under Entropy-Driven Modulation

The ideal policy gradient is defined as:

$$\nabla_\theta J(\theta) = \mathbb{E}_{\tau \sim \pi_\theta}\left[\sum_t Q^\pi(s_t, a_t)\nabla_\theta \log \pi_\theta(a_t|s_t)\right]$$

In practice, the actual gradient estimator $\hat{g}$ differs from the ideal gradient. The mean squared error (MSE) of any gradient estimator admits the classic bias-variance decomposition:

$$\text{MSE}(\hat{g}) = \|\text{Bias}(\hat{g})\|^2 + \text{Var}(\hat{g})$$

where $\text{Bias}(\hat{g}) = \mathbb{E}[\hat{g}] - \nabla_\theta J(\theta)$. HAPO's four components target these two terms from complementary angles: *Adaptive Temperature Sampling* and *Token-Level Group Average Advantage Estimation* primarily reduce $\|\text{Bias}\|^2$, while *Differential Advantage Redistribution* and *Asymmetric Adaptive Clipping* primarily reduce $\text{Var}(\hat{g})$.

E.2 Adaptive Temperature Sampling: Reducing Bias

Standard sampling uses $\pi_T(o_t|s_t) \propto \pi_\theta(o_t|s_t)^{1/T}$. HAPO sets $T_{i,t} = T_{\text{base}}(1 + \tilde{h}_{i,t} \cdot \tau)$. In RLHF, the estimation quality depends on whether rollout samples cover critical reasoning paths.

At a high-entropy position t, a critical action a^* typically has low probability under π_θ. In G independent rollouts, the probability that a^* is sampled at least once is $P(\text{sampled}) = 1 - (1 - \pi_T(a^*|s_t))^G$. When T is fixed at a low value ($1/T > 1$), the exponentiation $\pi_\theta(a^*)^{1/T}$ suppresses already-small probabilities, making $P(\text{sampled})$ very low. This causes the advantage estimate to be biased toward dominant, potentially sub-optimal paths.

HAPO raises the temperature at high-entropy positions ($T_{i,t} > T_{\text{base}}$ when $\tilde{h}_{i,t} > 0$). Since $1/T_{i,t} < 1/T_{\text{base}}$, the suppression effect is reduced:

$$\begin{aligned} \pi_{T_{i,t}}(a^*|s_t) &> \pi_{T_{\text{base}}}(a^*|s_t) \\ &\implies P^{\text{HAPO}} > P^{\text{DAPO}} \end{aligned} \tag{14}$$

More critical actions are covered, yielding a more representative advantage estimate and reducing coverage bias. Conversely, at low-entropy positions, lowering T concentrates the distribution on the single viable action, reducing noise without introducing bias. HAPO breaks the global temper-

Require: Policy π_θ, dataset $\mathcal{D}$
Require: Entropy quantile: ρ
Require: Temperature params: $T_{\text{base}}, \hat{\rho}_{\log(H)}, \hat{\sigma}_{\log(H)}, \tau$ Clipping params: base bounds $\epsilon_L^{\text{base}}, \epsilon_R^{\text{base}}$
Ensure: Updated policy $\pi_{\theta'}$

▷ Sampling Rollout Sequences
for each token position i, t **do** ▷ Continuous Adaptive Temperature Sampling
  Compute entropy: $H_{i,t} \leftarrow -\sum_v \pi_\theta(v|s_{i,t}) \log \pi_\theta(v|s_{i,t})$
  Adaptive temperature: $T_{i,t} \leftarrow T_{\text{base}} \cdot \left(1 + \frac{\log(H_{i,t}) - \hat{\rho}_{\log(H)}}{\hat{\sigma}_{\log(H)}} \cdot \tau\right)$
end for
Assign rewards to tokens: $a_{i,t} \leftarrow r_i, \mathcal{T} \leftarrow \{(i,t)\}$ ▷ Token-Level Group Average Advantage Estimation
Compute token-level statistics: $\mu_{\text{tok}} \leftarrow \frac{1}{|\mathcal{T}|}\sum_{(i,t)} a_{i,t}, \sigma_{\text{tok}} \leftarrow \sqrt{\frac{1}{|\mathcal{T}|}\sum_{(i,t)}(a_{i,t} - \mu_{\text{tok}})^2}$
Normalize advantages: $A_{i,t} \leftarrow (a_{i,t} - \mu_{\text{tok}})/\sigma_{\text{tok}}$
$Q_\rho(\log(H)) \leftarrow \text{Quantile}_\rho(\{\log(H_{i,t}) : (i,t) \in \mathcal{T}\})$ ▷ Compute global entropy statistics
$\sigma(\log(H)) \leftarrow \sqrt{\frac{1}{|\mathcal{T}|}\sum_{(i,t)\in\mathcal{T}}(\log(H_{i,t}) - Q_\rho(\log(H)))^2}$
for each mini batch $\mathcal{B}$ **do** ▷ Training with Heterogeneous Treatment
  for each token i, t in $\mathcal{B}$ **do**
    Compute log-entropies: $h_{i,t} = \frac{\log(H_{i,t}) - Q_\rho(\log(H))}{\sigma(\log(H))}$
    Asymmetric scaling: $\tilde{h}_{i,t} \leftarrow \begin{cases} h_{i,t}/h_{\max} & \text{if } h_{i,t} > 0 \\ -h_{i,t}/|h_{\min}| & \text{if } h_{i,t} \le 0 \end{cases}$
    ▷ Continuous Asymmetric Adaptive Clipping - Compute First
    if $\tilde{h}_{i,t} \le 0$ **then**
      $\epsilon_L(i,t) \leftarrow \epsilon_L^{\text{base}}(1 - \tilde{h}_{i,t}), \epsilon_R(i,t) \leftarrow \epsilon_R^{\text{base}}$
    else
      $\epsilon_L(i,t) \leftarrow \epsilon_L^{\text{base}}, \epsilon_R(i,t) \leftarrow \epsilon_R^{\text{base}}(1 + \tilde{h}_{i,t})$
    end if
    $\gamma_L \leftarrow 1 - \frac{\epsilon_L(i,t)}{2}, \gamma_U \leftarrow 1 + \frac{\epsilon_R(i,t)}{2}$ ▷ Neutral Zone based on Dynamic Clipping
    ▷ Continuous Differential Advantage Redistribution
    Compute importance ratio: $r_{i,t} \leftarrow \pi_\theta(a_{i,t}|s_{i,t})/\pi_{\theta_{\text{old}}}(a_{i,t}|s_{i,t})$
    if ($\tilde{h}_{i,t} > 0$ and $r_{i,t} \notin [\gamma_L, \gamma_U]$) or ($\tilde{h}_{i,t} \le 0$ and $r_{i,t} \in [\gamma_L, \gamma_U]$) **then**
      $\hat{A}_{i,t} \leftarrow A_{i,t} \cdot (1 + \tilde{h}_{i,t})$
    else
      $\hat{A}_{i,t} \leftarrow A_{i,t}$
    end if
  end for
  ▷ Compute HAPO loss
  $\mathcal{L}^{\text{HAPO}}(\theta) = \left[\frac{1}{\sum_{i=1}^G |o_i|}\sum_{i=1}^G \sum_{t=1}^{|o_i|} \min\left(r_{i,t}(\theta)\hat{A}_{i,t}, \text{clip}(r_{i,t}(\theta), 1 - \epsilon_L(i,t), 1 + \epsilon_R(i,t))\hat{A}_{i,t}\right)\right]$
  Update parameters: $\theta' \leftarrow \theta + \eta \cdot \nabla_\theta \mathcal{L}^{\text{HAPO}}(\theta)$
end for
$\hat{\rho}_{\log(H)} \leftarrow Q_\rho(\log(H_t)), \hat{\sigma}_{\log(H)} \leftarrow \sigma(\log(H_t))$ ▷ Update entropy statistics for next step

Figure 12: HAPO algorithm overview. The method dynamically adjusts temperature, advantages, and clipping bounds based on normalized entropy $\tilde{h}_{i,t} \in [-1, 1]$, implementing heterogeneous treatment for tokens throughout the RL pipeline.

ature trade-off:

$$T_{i,t}^{\text{high-ent}} > T_{\text{base}} > T_{i,t}^{\text{low-ent}}$$

thereby ensuring $\|\text{Bias}(\hat{g}^{\text{HAPO}})\|^2 < \|\text{Bias}(\hat{g}^{\text{DAPO}})\|^2$.

E.3 Token-Level Group Average Advantage: Reducing Bias

GRPO divides the loss by sequence length $|o_i|$, while DAPO divides by total tokens $N_{\text{tok}} = \sum_i |o_i|$. Both use sequence-level advantage $A_i = (R_i - \mu_R)/\sigma_R$. The gradient of DAPO's loss (simplified) is:

$$\hat{g}^{\text{DAPO}} = \frac{1}{N_{\text{tok}}}\sum_i A_i \cdot |o_i| \cdot \bar{\nabla}_i$$

where $\bar{\nabla}_i$ is the average per-token gradient. The weight of sequence i is $A_i \cdot |o_i|$. Since negative samples are significantly longer ($|o_i^-| \gg |o_i^+|$), the balance is broken: $\sum_{i:A_i<0}|A_i| \cdot |o_i^-| \gg \sum_{i:A_i>0}|A_i| \cdot |o_i^+|$.

This length-advantage coupling introduces a systematic negative bias. HAPO resolves this by normalizing across the global token pool:

$$A_{i,t}^{\text{tok}} = \frac{a_{i,t} - \mu_{\text{tok}}}{\sigma_{\text{tok}}}$$

By construction, $\sum_{(i,t)} A_{i,t}^{\text{tok}} = 0$ holds at the **token level**. Longer sequences contribute proportionally more tokens to both positive and negative sides, eliminating the bias while preserving the ability to learn from complex reasoning chains.

E.4 Differential Advantage Redistribution: Reducing Variance

Consider the per-token gradient $g_{i,t} = \hat{A}_{i,t} \cdot \nabla_\theta \log \pi_\theta(o_t|s_t)$. Its variance is $\text{Var}(g_{i,t}) = \hat{A}_{i,t}^2 \cdot F_t(\theta)$, where $F_t(\theta)$ is the Fisher Information. High-entropy positions have large F_t, yielding a low Signal-to-Noise Ratio (SNR).

HAPO redistributes advantage such that for high-entropy tokens ($\tilde{h} > 0$) with $r_{i,t} \notin [\gamma_L, \gamma_U]$, the advantage is amplified by $(1 + \tilde{h})$. This boosts the SNR:

$$\text{SNR}_{i,t}^{\text{HAPO}} = \frac{(1 + \tilde{h}_{i,t})^2 \cdot |\mathbb{E}[\nabla_\theta \log \pi_\theta]|^2}{F_t(\theta)}$$

The $(1 + \tilde{h})^2$ factor compensates for the large F_t. Simultaneously, HAPO suppresses low-entropy tokens that have high SNR but low information. This reallocates the gradient budget to high-information tokens, reducing overall variance: $\text{Var}(\hat{g}^{\text{HAPO}}) < \text{Var}(\hat{g}^{\text{DAPO}})$.

E.5 Asymmetric Adaptive Clipping: Reducing Variance

The local KL divergence contribution from a token o_t under a ratio change $\delta = |r_{i,t} - 1|$ is approximately $\Delta D_{\text{KL}}^{(t)}(\delta) \approx \delta^2 \cdot \pi_{\text{old}}(o_t)$.

In DAPO, fixed clipping $[1 - \epsilon_L, 1 + \epsilon_R]$ causes high-entropy tokens (π_{old} is small) to be clipped prematurely despite low KL cost, while low-entropy tokens (π_{old} is large) undergo disproportionately large KL shifts. HAPO makes the clipping threshold a function of entropy:

$$\delta_R^{\max}(\tilde{h}) = \epsilon_R^{\text{base}}(1+\tilde{h}_{i,t}), \quad \delta_L^{\max}(\tilde{h}) = \epsilon_L^{\text{base}}(1-\tilde{h}_{i,t})$$

This ensures the effective trust region in KL space is approximately uniform across entropy levels:

$$\Delta D_{\text{KL}}^{\text{HAPO, high-ent}} \approx \Delta D_{\text{KL}}^{\text{HAPO, low-ent}}$$

This retains informative high-entropy gradients while properly constraining noisy low-entropy updates, further reducing variance.

F Practical Adoption

F.1 Hyperparameter Robustness

HAPO uses a single configuration across all tested models (Qwen, LLaMA) and scales (1.5B–14B). Ablations in Appendix D confirm that performance is stable across a wide range of values, requiring no per-model tuning.

F.2 Implementation and Integration

HAPO is easily integrated into existing RLHF pipelines (e.g., `verl`, `vLLM`) using simple element-wise operations. The core logic follows a unified principle: normalize entropy to $\tilde{h} \in [-1, 1]$ and scale base parameters accordingly. Implementation requires only minimal additions to the sampling, clipping, and advantage estimation modules.We provide full source code, configuration files, and a comprehensive setup guide at: https://github.com/starriver030515/HAPO.

G Example of Dual-Entropy Phenomenon

We show the 20 most frequent dual-entropy token pairs in Figure 11. These dual-entropy tokens are all critical tokens that determine reasoning paths, yet exhibit vastly different entropy values. We present concrete examples of the dual entropy phenomenon in Figure 13, 14, 15, 16, 17, 18. We use blue and red to represent the low-entropy and high-entropy parts of dual-entropy, respectively. These visualizations demonstrate how tokens with identical semantic meanings but different surface forms exhibit drastically different entropy values during model generation.

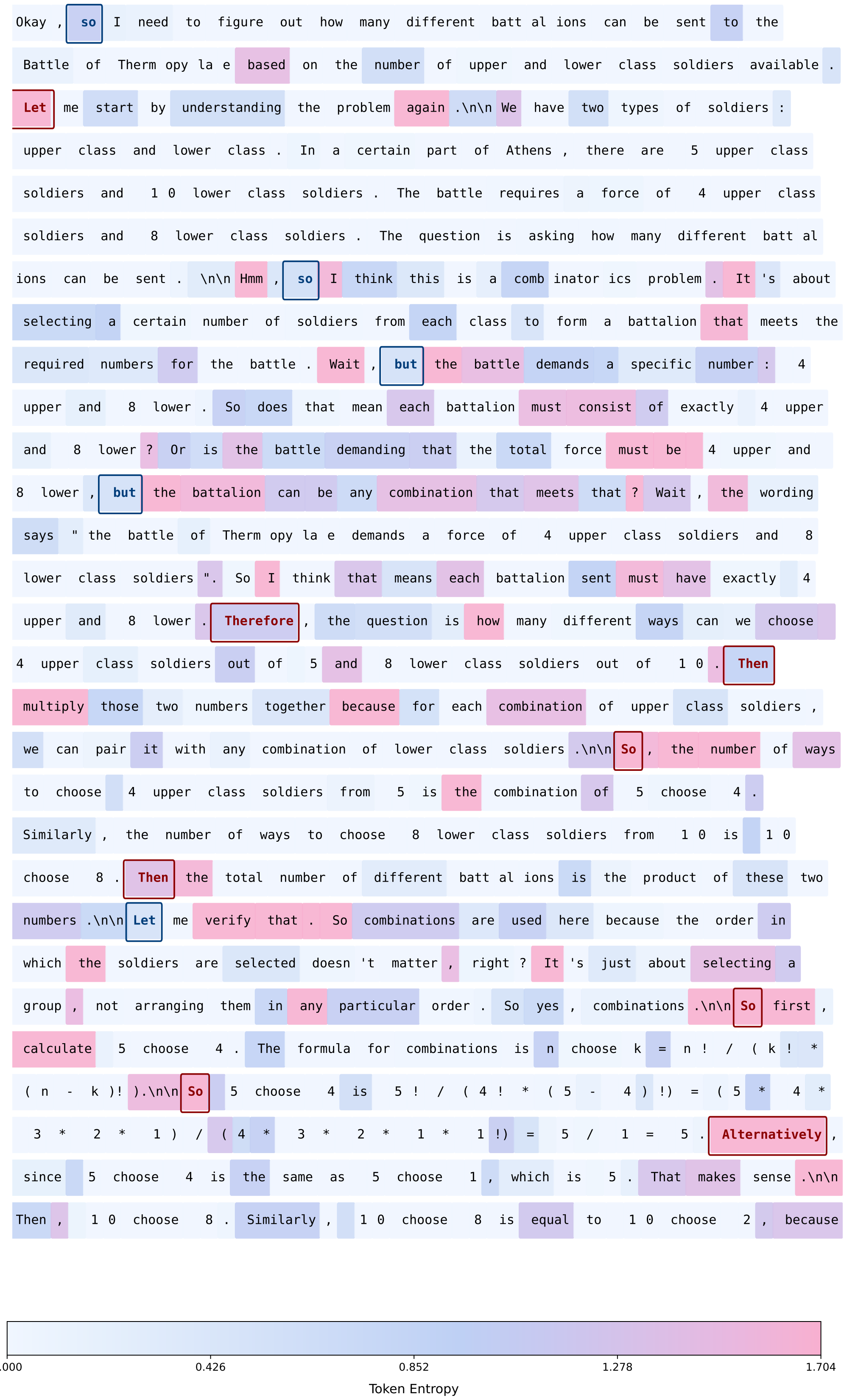


Figure 13: Dual-Entropy Example 1-part1

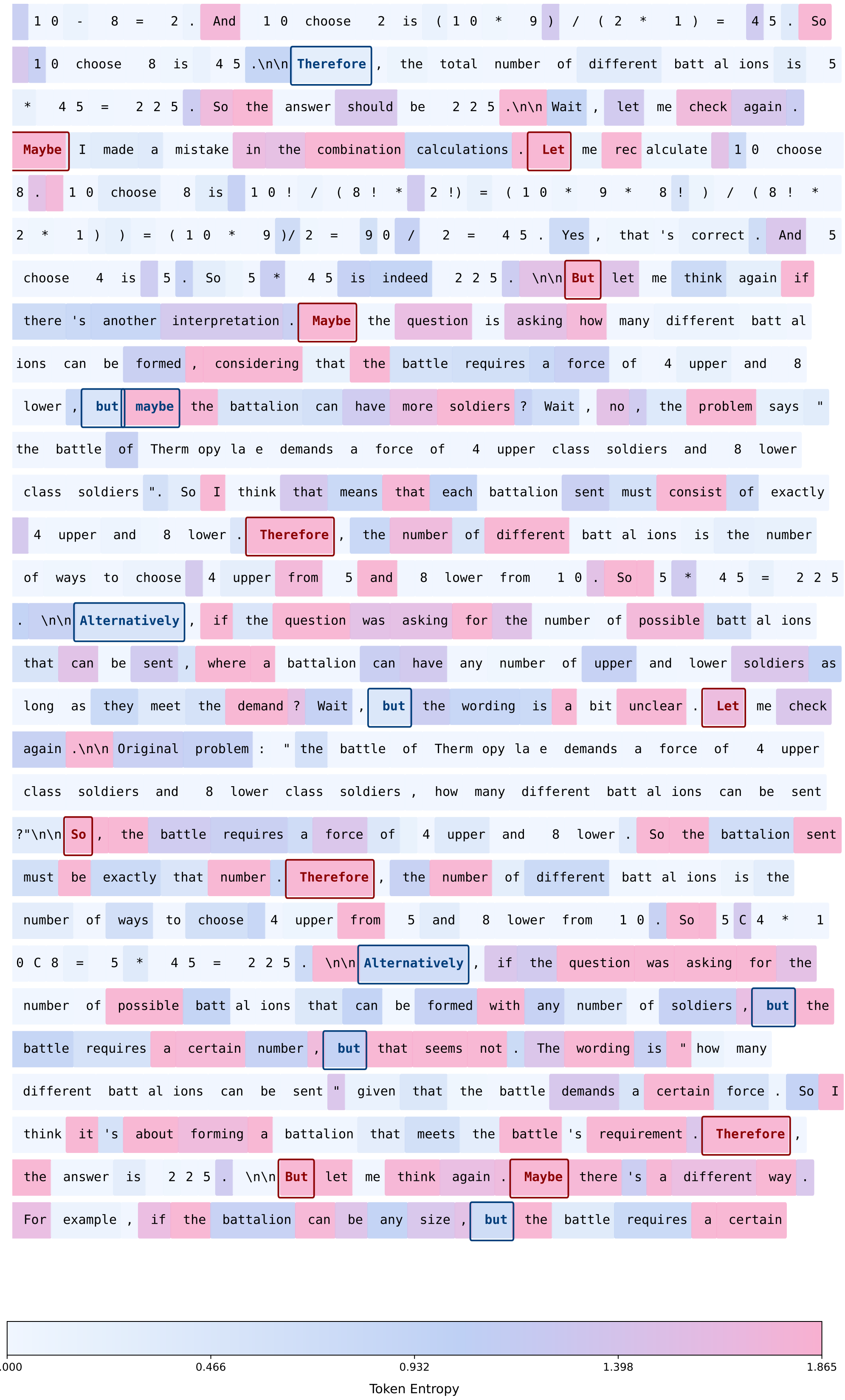


Figure 14: Dual-Entropy Example 1-part2

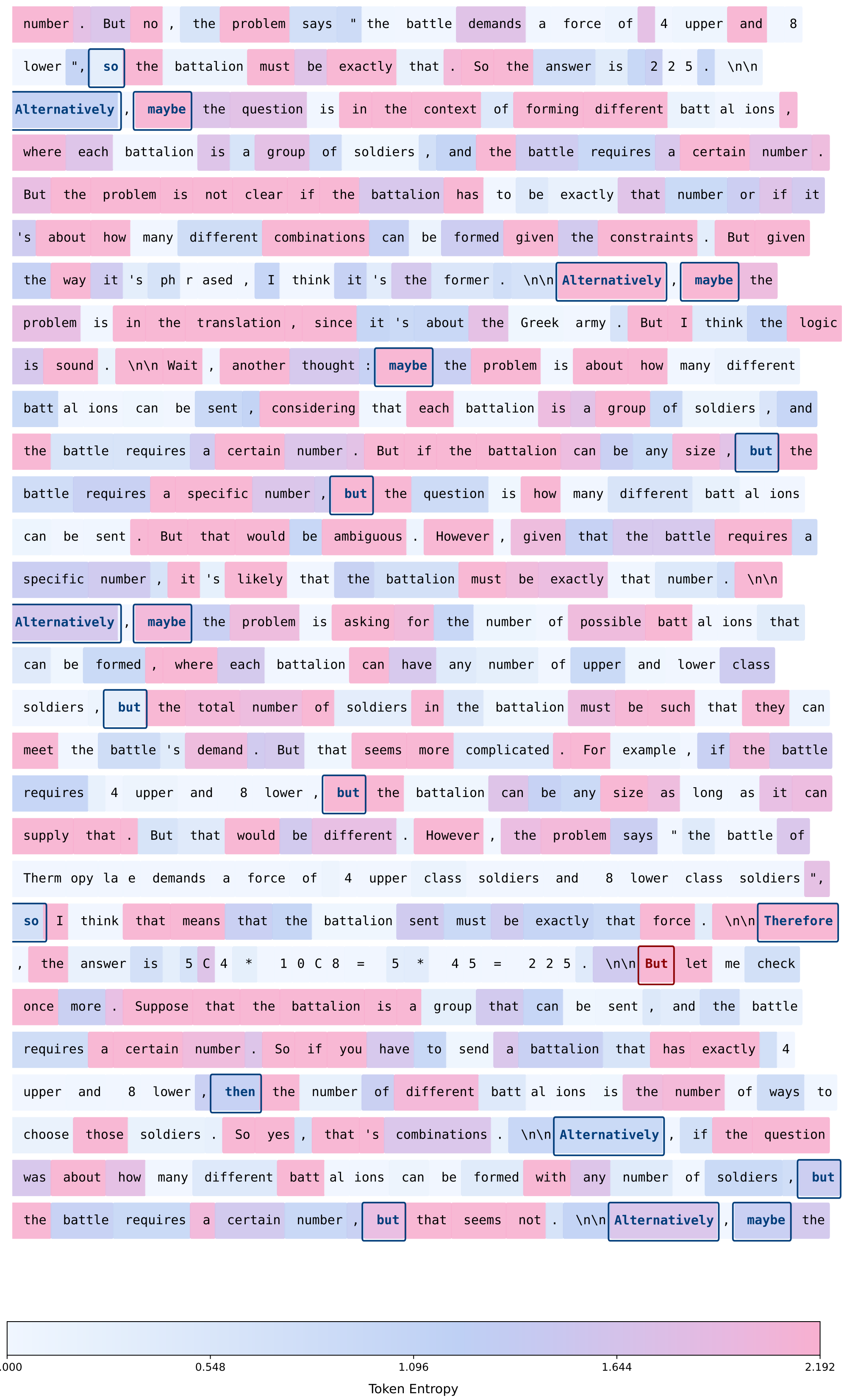


Figure 15: Dual-Entropy Example 1-part3

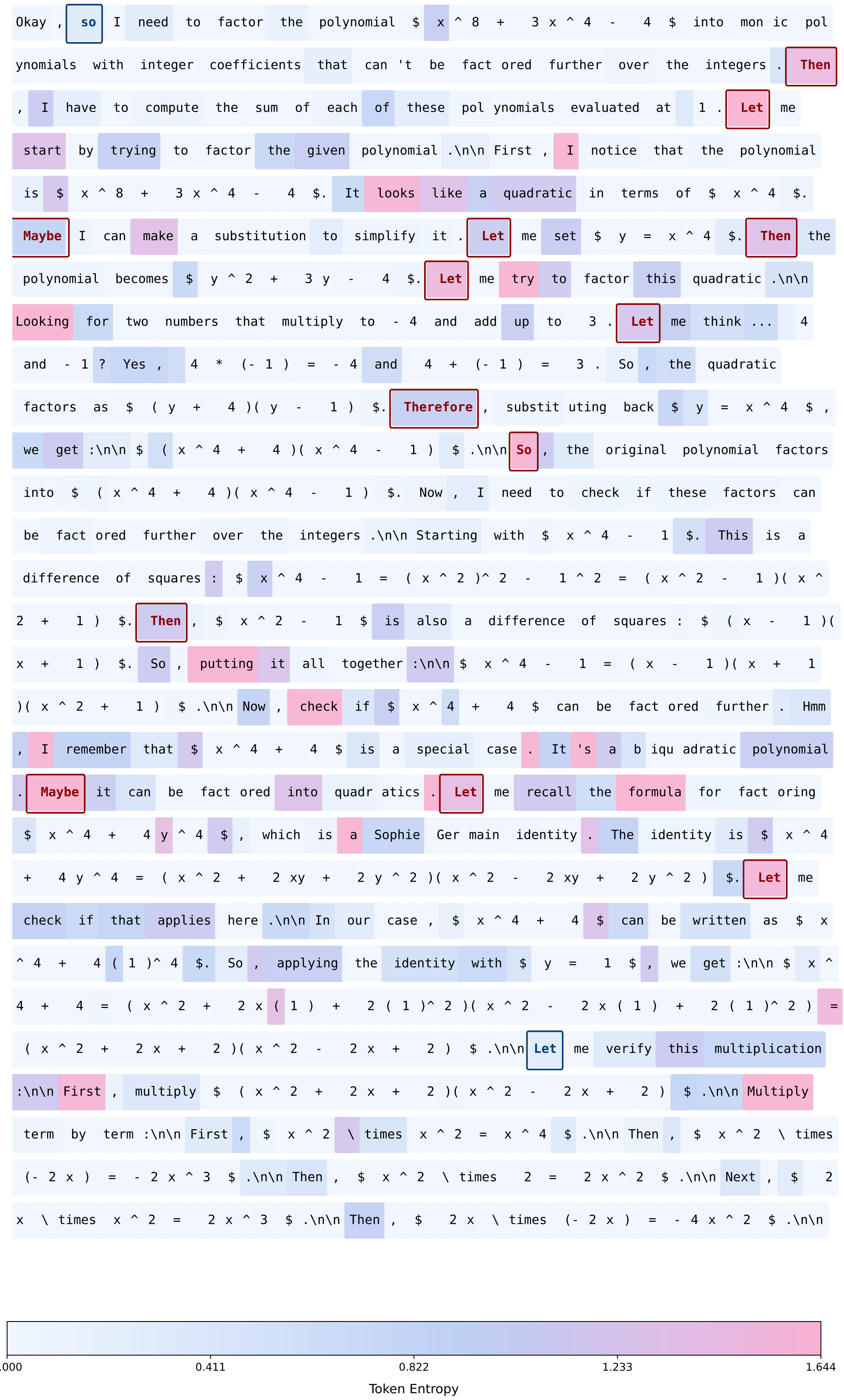


Figure 16: Dual-Entropy Example 2-part1

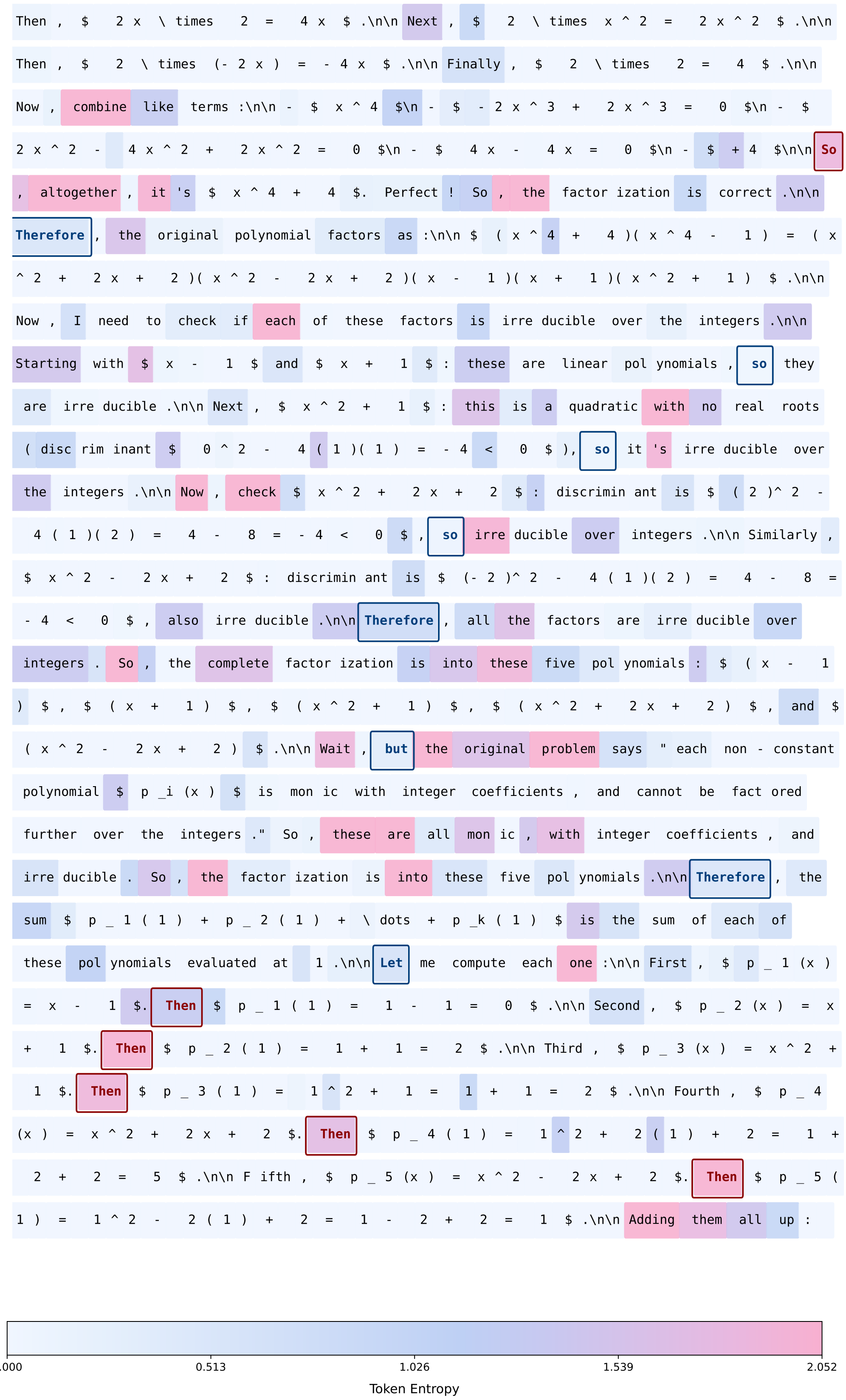


Figure 17: Dual-Entropy Example 2-part2

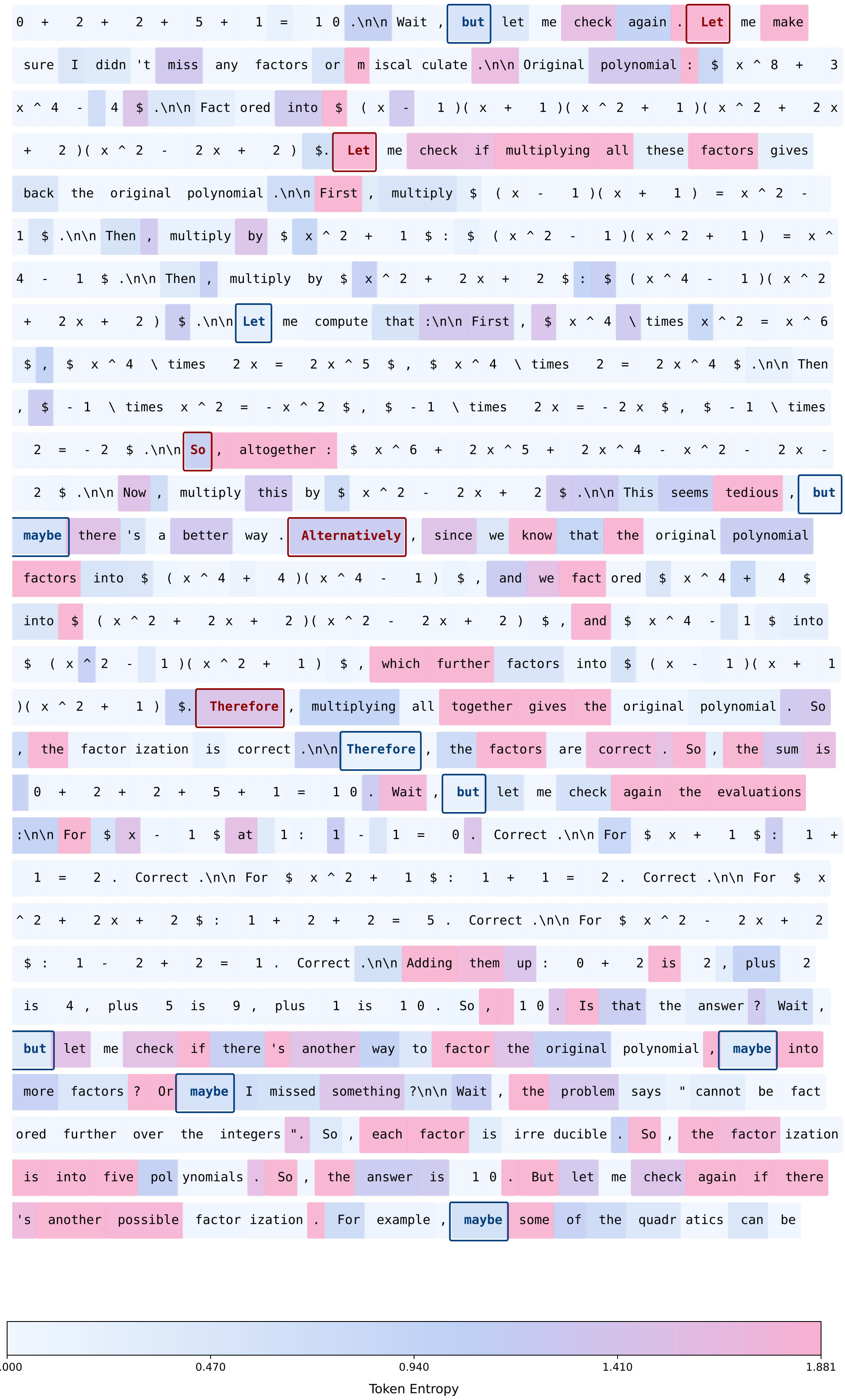


Figure 18: Dual-Entropy Example 2-part3