INFORMATION-TIME PROXIMAL POLICY OPTIMIZATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Reinforcement Learning with Verifiable Rewards (RLVR) has substantially improved the reasoning capabilities of Large Language Models (LLMs). However, existing methods typically measure temporal progression by token count, despite the highly non-uniform information flow along autoregressive trajectories. We propose Information-Time Proximal Policy Optimization (InfoPPO), which measures temporal distance by accumulated state-wise information rather than raw token count. This provides a common basis for temporal credit propagation and policy updates. InfoPPO enables non-trivial discounting in long-horizon reasoning, retaining effective-horizon contraction in information time while mitigating excessive attenuation of terminal supervision. Moreover, the information-time policy-improvement analysis naturally leads to a state-dependent update constraint, which we implement through adaptive clipping. It scales the update range with local information density, enabling targeted policy updates while preserving proximal control. Theoretically, we extend performance-difference and policy-improvement analyses to the information-time Markov Decision Process (MDP), deriving a policy-improvement lower bound. We further connect this analysis to practical LLM policy optimization by relating state-wise information density to local policy movement, while also providing theoretical grounding for the adaptive update mechanism. Experiments on Qwen3 models demonstrate consistent gains over competitive baselines across five challenging competition-style mathematical reasoning benchmarks. InfoPPO also maintains stable accuracy and response length across non-trivial discount settings under which token-time PPO deteriorates.

1 INTRODUCTION

Although RLVR has substantially improved LLM performance on challenging reasoning tasks (Guo et al., 2025; Yang et al., 2025; Jaech et al., 2024; Shao et al., 2024; Yu et al., 2025), reliably optimizing policies over long reasoning trajectories remains challenging. Existing RLVR methods commonly formulate each generated token as one transition of the MDP, measuring temporal progression by raw token count. This is a natural and widely adopted modeling choice. However, this formulation does not explicitly capture how information content varies across token transitions. This raises the question of whether token count alone provides the most suitable temporal metric for long-horizon LLM reasoning. Under sparse outcome-level supervision, token-wise discounting can excessively attenuate terminal rewards, whereas removing discounting forfeits effective-horizon control. This tension motivates exploring temporal metrics beyond token count for long-horizon LLM reasoning.

One natural basis for such an alternative is the non-uniform information structure of autoregressive trajectories. Conventional RL domains, such as robotic control and games, index time by successive environment interactions (Brockman et al., 2016; Schulman et al., 2015a; 2017). When this convention is applied to language generation, each generated token is likewise treated as one unit of temporal progression. Yet token progression need not match information progression: many locally predictable tokens, including common function words such as articles, contribute mainly to local syntactic or semantic continuity, whereas a smaller number of semantic transition points can substantially redirect the subsequent reasoning trajectory (Wang et al., 2025; Qu et al., 2025; Fu et al., 2025). The resulting information landscape therefore interleaves locally sparse and dense regions. In sparse regions, continuations are relatively constrained, whereas dense regions involve a broader range of plausible continuations and larger increments of information. Consequently, token spans of equal

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

length can accumulate markedly different amounts of reasoning-relevant information, even though token-based discounting applies the same cumulative decay to each. Raw token count thus captures sequence distance, but not the information distance traversed along the sequence. This motivates parameterizing temporal progression by accumulated information rather than token count.

Building on this analysis, we propose Information-Time Proximal Policy Optimization (InfoPPO), which preserves the token-level state and action structure while measuring temporal distance through accumulated state-wise information rather than raw token count. Under this formulation, discounting and trace decay operate over information time, retaining effective-horizon contraction while mitigating the excessive attenuation of terminal supervision over long token sequences. The corresponding policy-improvement analysis further motivates an information-dependent policy-update scale, implemented through adaptive clipping to support more sample-efficient learning. Information density thus provides a common basis for temporal credit propagation and policy-update regulation.

Theoretically, we extend the performance-difference lemma to the information-time MDP and derive a policy-improvement lower bound for policy updates constrained by information density. We further connect the general information-time analysis to practical LLM policy optimization by relating state-wise information density to local policy movement, while also providing theoretical grounding for the adaptive update mechanism in InfoPPO. Empirically, InfoPPO delivers consistent gains over PPO and DAPO-based baselines across model scales and five challenging competition-style mathematical reasoning benchmarks, while maintaining stable response lengths across training and remaining robust to non-trivial discounting that substantially degrades token-time PPO.

2 Related Works

Reinforcement Learning with Verifiable Rewards (RLVR). RLVR improves LLM reasoning through rule-based verification functions to generate binary reward signals (Guo et al., 2025; Yang et al., 2025; Jaech et al., 2024; Lambert et al., 2024). Within this paradigm, PPO (Schulman et al., 2017) remains a central framework and constitutes the most direct comparison for our method. Group-based methods such as GRPO (Shao et al., 2024) replace a separate value network with group-relative advantage estimates from multiple responses per query, as adopted in DeepSeek-R1 (Guo et al., 2025). Concurrently, DAPO (Yu et al., 2025) improves performance and stability through refined engineering implementations. DAPO-FT (Wang et al., 2025) shows that a minority of high-entropy tokens account for a disproportionate share of effective policy learning and accordingly restrict gradient updates to these positions. Collectively, these methods refine how learning signals are estimated, sampled, and allocated, while generally retaining token-based temporal progression. In contrast, InfoPPO uses information density to define temporal progression and regulate policy updates, thereby enabling effective-horizon control and supporting more sample-efficient policy learning.

Temporal Abstraction and Information Dynamics. RL has explored temporal abstraction to move beyond uniformly indexed primitive transitions. Foundational frameworks such as Semi-Markov Decision Processes and the options framework (Sutton et al., 1999; Precup, 2000; Bacon et al., 2017) aggregate primitive transitions into temporally extended actions. Such approaches typically require higher-level actions, units, or boundaries to be specified or learned, which can be challenging in autoregressive generation. A complementary line of work generalizes geometric discounting through weighted or transition-based discount factors, allowing temporal weighting to vary across states or transitions (Feinberg & Shwartz, 1994; White, 2017). InfoPPO is related to this work and grounds non-uniform temporal progression specifically in state-wise information density. Unlike explicit temporal-abstraction methods, InfoPPO preserves token-level states and actions, capturing non-uniform temporal progression without explicit segmentation or macro-action modeling.

3 Preliminaries

Markov Decision Process (MDP). We model language generation as a discounted MDP $\mathcal{M} = (\mathcal{S}, \mathcal{A}, \mathcal{P}, r, \gamma)$, where s_t is the sequence prefix and $a_t \in \mathcal{V}$ is the next token. The transition dynamics $\mathcal{P} : \mathcal{S} \times \mathcal{A} \times \mathcal{S} \to \mathbb{R}$ are deterministic, given by $s_{t+1} = s_t \oplus a_t$, where $\oplus$ denotes concatenation. $r : \mathcal{S} \to \mathbb{R}$ is the reward function and γ is the discount factor. For convenience, we write $r_t := r(s_t)$. The objective is to maximize the expected discounted return $\eta(\pi) = \mathbb{E}_{\tau\sim\pi}[\sum_{t=0}^{\infty} \gamma^t r(s_t)]$. The value and action-value functions are $V_\pi(s_t) = \mathbb{E}_\pi[\sum_{k=0}^{\infty} \gamma^k r(s_{t+k})|s_t]$ and $Q_\pi(s_t, a_t) =$

$\mathbb{E}_\pi[\sum_{k=0}^{\infty}\gamma^k r(s_{t+k})|s_t,a_t]$, with the advantage function $A_\pi(s,a) \coloneqq Q_\pi(s,a) - V_\pi(s)$. Finally, we define the (unnormalized) discounted state visitation distribution as $d_\pi(s) \coloneqq \sum_{t=0}^{\infty}\gamma^t P(s_t = s|\pi)$. In many sparse reward settings, the agent only receives a non-zero signal R_T at the terminal timestep T, i.e., $r_t = 0$ for $t < T$ and $r_T = R_T$ for $t = T$.

Policy Optimization. The *Performance Difference Lemma* (Kakade & Langford, 2002) provides the theoretical foundation for monotonic policy improvement, which expresses the return difference as:

$$\eta(\tilde{\pi}) - \eta(\pi) = \mathbb{E}_{\tau\sim\tilde{\pi}}\left[\sum_{t=0}^{\infty}\gamma^t A_\pi(s_t,a_t)\right] = \sum_s d_{\tilde{\pi}}(s)\sum_a \tilde{\pi}(a|s)A_\pi(s,a). \tag{1}$$

In practice, to render the optimization tractable, we employ a local approximation by replacing $d_{\tilde{\pi}}$ with d_π, yielding the following surrogate objective function:

$$L_\pi(\tilde{\pi}) = \eta(\pi) + \sum_s d_\pi(s)\sum_a \tilde{\pi}(a|s)A_\pi(s,a). \tag{2}$$

TRPO (Schulman et al., 2015a) then controls the discrepancy between the surrogate objective and the true return by constraining policy divergence. PPO (Schulman et al., 2017) approximately limits policy updates through the following clipped surrogate objective:

$$\mathcal{L}^{\text{CLIP}}(\theta) = \mathbb{E}_{\pi_{\theta_{\text{old}}}}\Big[\min\left(\omega_t(\theta)A_{\theta_{\text{old}}}, \text{clip}(\omega_t(\theta), 1-\epsilon_{\text{low}}, 1+\epsilon_{\text{high}})A_{\theta_{\text{old}}}\right)\Big], \tag{3}$$

where $\omega_t(\theta) = \frac{\pi_\theta(a_t|s_t)}{\pi_{\theta_{\text{old}}}(a_t|s_t)}$ is the importance ratio, ϵ_{low} and ϵ_{high} are the clipping hyperparameters.

Advantage Estimation. Generalized Advantage Estimation (GAE) (Schulman et al., 2015b) is used to estimate the advantage from a parameterized value function V_ϕ. Given the TD residual $\delta_t = r_t + \gamma V_\phi(s_{t+1}) - V_\phi(s_t)$, the GAE estimator is given by $A_t^{\text{GAE}(\gamma,\lambda)}(s_t,a_t) = \sum_{l=0}^{\infty}(\gamma\lambda)^l\delta_{t+l}$, where $\lambda \in (0,1]$ controls the bias-variance trade-off.

4 METHODOLOGY

In this section, we develop Information-Time Proximal Policy Optimization (InfoPPO) in three stages. First, we formulate an Information-Time MDP that measures temporal progression through accumulated state-dependent information while preserving token-level states and actions. Second, we extend the performance-difference and policy-improvement analyses to this formulation, deriving a lower bound under information-dependent policy-update constraints. Finally, we specialize the framework to LLM policy optimization by using the old policy's uncertainty over next-token continuations as a tractable proxy for local information density and analyze the resulting policy movement.

4.1 INFORMATION-TIME MARKOV DECISION PROCESS

Temporal Misalignment under Uniform Token Time. Conventional RL formulations assign each generated token a unit temporal increment, $\Delta\mu_{\text{unif}}(t) = 1$, which overlooks how the local continuation structure changes along the trajectory. At each state, however, the autoregressive policy induces a distribution over possible continuations, and the structure of this distribution can vary substantially (Wang et al., 2025; Fu et al., 2025). Some transitions occur where the continuation is relatively constrained, whereas others occur where a broader set of plausible continuations remains. The same unit token step can therefore correspond to different information increments. Consequently, equal token distances need not imply equal information distances. We refer to this mismatch between token distance and information distance as ***Temporal Misalignment***. This mismatch becomes consequential under temporal decay. Over long reasoning trajectories, non-trivial decay can excessively attenuate terminal supervision, whereas removing it forfeits effective-horizon contraction. We therefore introduce a temporal coordinate that advances according to state-wise information density.

The Information-Time MDP Construction. We define the Information-Time MDP as $\mathcal{M}_{\text{info}} = (\mathcal{S},\mathcal{A},\mathcal{P},r,\gamma,\rho)$, where $\rho : \mathcal{S} \to [0,1]$ specifies the information density of each state, reflecting the local intensity of information change along the trajectory. For example, in RL for LLM reasoning, the state corresponds to the self-generated context, while the old policy defines a predictive distribution over the next token at each visited state. Since the old policy remains fixed during the current

update, this distribution provides a stable characterization of the information profile of the state. This motivates using quantities derived from the old policy distribution, such as next-token uncertainty, as proxies for the information density $\rho(s_t)$. Details are provided in Section 4.3.

To formalize temporal progression based on state-wise information density, we define the information-time measure μ_{info} with respect to the uniform time measure μ_{unif}: $\frac{d\mu_{\text{info}}}{d\mu_{\text{unif}}}(t) = \rho(s_t)$. Equipped with this, we define the corresponding objective as $\eta^{\text{info}}(\pi) = \mathbb{E}_{\tau\sim\pi}\left[\sum_{t=0}^{\infty} \Gamma^{\text{info}}(0,t) r(s_t)\right]$, where $\Gamma^{\text{info}}(t_1,t_2) \coloneqq \gamma^{\sum_{t=t_1}^{t_2-1} \Delta\mu_{\text{info}}(t)} = \gamma^{\sum_{t=t_1}^{t_2-1} \rho(s_t)}$ is the information-time discount factor. For $\gamma = 1$, this recovers the undiscounted RLVR objective, whereas for $\gamma < 1$, information time determines how temporal decay is accumulated along the trajectory. The corresponding value and action-value are

$$V_\pi^{\text{info}}(s_t) \coloneqq \mathbb{E}_\pi\left[\sum_{u=t}^{\infty} \Gamma^{\text{info}}(t,u) r(s_u) \,\middle|\, s_t\right], \quad Q_\pi^{\text{info}}(s_t,a_t) \coloneqq \mathbb{E}_\pi\left[\sum_{u=t}^{\infty} \Gamma^{\text{info}}(t,u) r(s_u) \,\middle|\, s_t, a_t\right].$$

Therefore, the information-time advantage function is $A_\pi^{\text{info}}(s,a) \coloneqq Q_\pi^{\text{info}}(s,a) - V_\pi^{\text{info}}(s)$.

4.2 Policy Improvement on Information-Time MDP

To analyze policy updates in the Information-Time MDP, we first extend the classic performance difference lemma (Kakade & Langford, 2002) to the information-time setting.

Lemma 4.1. (Information-Time Performance Difference). *Given two policies π and $\tilde{\pi}$ within the Information-Time MDP framework, the following identity holds:*

$$\eta^{info}(\tilde{\pi}) - \eta^{info}(\pi) = \mathbb{E}_{\tau\sim\tilde{\pi}}\left[\sum_{t=0}^{\infty} \Gamma^{info}(0,t) A_\pi^{info}(s_t,a_t)\right]. \tag{4}$$

The detailed proof is provided in Appendix A.1. To express the information-time performance difference in state space, we define the unnormalized information-discounted state visitation measure $\nu_\pi^{\text{info}}(s) \coloneqq \mathbb{E}_{\tau\sim\pi}\left[\sum_{t=0}^{\infty} \Gamma^{\text{info}}(0,t)\mathbf{1}\{s_t = s\}\right]$. With this definition, Eq. 4 can be written as

$$\eta^{\text{info}}(\tilde{\pi}) - \eta^{\text{info}}(\pi) = \sum_s \nu_{\tilde{\pi}}^{\text{info}}(s) \sum_a \tilde{\pi}(a|s) A_\pi^{\text{info}}(s,a). \tag{5}$$

The derivation is provided in Appendix A.2. However, direct optimization of this expression is difficult because the state visitation measure depends on the candidate policy $\tilde{\pi}$. We therefore replace $\nu_{\tilde{\pi}}^{\text{info}}$ with the visitation measure induced by the current policy π, yielding the local surrogate

$$L_\pi^{\text{info}}(\tilde{\pi}) = \eta^{\text{info}}(\pi) + \sum_s \nu_\pi^{\text{info}}(s) \sum_a \tilde{\pi}(a|s) A_\pi^{\text{info}}(s,a). \tag{6}$$

This replacement introduces an approximation error between the surrogate objective and the true return of $\tilde{\pi}$, which must be controlled as $\tilde{\pi}$ departs from π. Unlike the standard discounted setting, the information-time discount $\Gamma^{\text{info}}(0,t)$ depends on the accumulated information density along the trajectory. As a result, $\Gamma^{\text{info}}(0,t)$ no longer forms a standard geometric sequence, and its cumulative sum cannot in general be reduced to $(1-\gamma)^{-1}$ as in the uniform-time setting. The standard geometric-series argument used to control the approximation error therefore does not directly apply.

The underlying difference is that each transition now advances the temporal coordinate by $\rho(s)$, rather than by a uniform unit step. This state-dependent temporal increment motivates expressing the local policy-update scale in terms of information density. Specifically, we impose the following state-wise total-variation constraint between the candidate policy $\tilde{\pi}$ and the current policy π:

$$D_{\text{TV}}\left(\tilde{\pi}(\cdot|s), \pi(\cdot|s)\right) \le \alpha\rho(s), \qquad \forall s. \tag{7}$$

Under this condition, the divergence is constrained by information density, with α setting its overall scale. We justify this in Section 4.3. We next show that it bounds the surrogate approximation error.

Theorem 4.2. *Let π be the current policy and $\tilde{\pi}$ a candidate policy. Suppose that the divergence condition in Eq. 7 holds for some $\alpha \ge 0$. Then, the following lower bound holds:*

$$\eta^{info}(\tilde{\pi}) \ge L_\pi^{info}(\tilde{\pi}) - \frac{4C\alpha^2}{(1-\gamma)^2}, \qquad \textit{where } C = \max_{s,a} |A_\pi^{info}(s,a)|. \tag{8}$$

The proof in Appendix A.6 uses a coupling argument linking state-wise total variation to action disagreement between policies. If the surrogate gain exceeds the penalty in Theorem 4.2, the candidate policy achieves a non-decreasing information-time return.



(a) Highest average predictive entropy.

(b) Lowest average predictive entropy.

Figure 1: **Predictive entropy patterns in mathematical reasoning.** We show up to 100 tokens with the highest and lowest average predictive entropy in Qwen3-4B, 8B, and 14B responses to AIME24 and AIME25, restricting to tokens occurring more than 100 times in each model to reduce noise.

4.3 Practical Construction of the Information Clock

The analysis in Section 4.2 is stated in the general infinite-horizon form. Episodic language generation is naturally covered by this formulation by treating the state reached after termination as absorbing, with zero continuation reward and value. We now specialize this framework to the practical RLVR setting considered in this work. Each generated response forms a finite trajectory $\tau = (s_0, a_0, \ldots, s_T)$, and supervision is provided through a terminal outcome reward R_T.

Instantiating the Information Density. The Information-Time MDP specifies the role of $\rho(s)$, but leaves its functional form unspecified. Because $\rho(s_t)$ determines the information-time increment assigned to the transition from s_t to s_{t+1}, its practical instantiation should capture the local uncertainty over the next token. In language generation, the next transition is determined by the next-token choice $\pi_{\text{old}}(\cdot|s_t)$. The conditional next-token distribution therefore provides a state-local characterization of the uncertainty over the next transition. Shannon entropy summarizes this uncertainty as the expected next-token surprisal. Unlike the surprisal of a realized token, it depends on the full predictive distribution before sampling, making it a state-wise quantity consistent with ρ. A concentrated distribution yields low entropy and indicates a locally constrained continuation, whereas a dispersed distribution yields high entropy and reflects a broader set of plausible continuations. Moreover, entropy can be computed directly from the policy distribution at the current state, varies continuously with the predictive distribution, and requires neither future trajectory information nor additional supervision. We therefore use the normalized entropy of the frozen old policy as the proxy for $\rho(s_t)$.

Figure 1 provides a qualitative view of token patterns associated with different levels of predictive entropy in mathematical reasoning. Tokens with higher average entropy tend to occur in discourse and reasoning transitions, whereas lower entropy is more common in mathematical symbols, numerals, and predictable subword fragments. This contrast supports predictive entropy as a practical proxy for local information density, since higher entropy reflects a broader set of plausible next-token continuations, while lower entropy indicates more constrained transitions. Figures 10–15 further illustrate this local variation along individual reasoning trajectories.

Formally, for each state s_t visited under the frozen old policy, we define its predictive entropy as

$$\mathcal{H}_{\text{old}}(s_t) := \mathcal{H}(\pi_{\text{old}}(\cdot|s_t)) = -\sum_{a\in\mathcal{V}} \pi_{\text{old}}(a|s_t)\log \pi_{\text{old}}(a|s_t), \tag{9}$$

and instantiate the information density as $\rho(s_t) = \frac{\mathcal{H}_{\text{old}}(s_t)}{\mathcal{H}_{\max}}$, where $\mathcal{H}_{\max} > 0$ is a normalization scale held fixed throughout the current policy update, so that $\rho(s_t) \in [0, 1]$ over the states considered. This construction keeps ρ fixed within each update while allowing it to track changes in predictive uncertainty across policy iterations. Connecting this practical construction to the preceding analysis requires addressing two additional points. Theorem 4.2 assumes that policy divergence scales with $\rho(s)$, so this relation must first be verified for the entropy construction. Moreover, the theorem compares policies under a common information density, whereas the practical algorithm recomputes ρ after each policy update. We therefore quantify the change in information-time return introduced by this recomputation. We first verify the required divergence relation. For a fixed state s, the action-dependent term of the surrogate in Eq. 6 admits the equivalent importance-ratio representation

$$\sum_a \tilde{\pi}(a|s) A^{\text{info}}_{\pi_{\text{old}}}(s,a) = \mathbb{E}_{a\sim\pi_{\text{old}}(\cdot|s)}\left[\frac{\tilde{\pi}(a|s)}{\pi_{\text{old}}(a|s)} A^{\text{info}}_{\pi_{\text{old}}}(s,a)\right]. \tag{10}$$

Because the divergence condition in Eq. 7 is imposed separately at each state, we characterize the policy movement induced by this conditional objective at a fixed s.

Proposition 4.3. *Suppose that π_{old} is parameterized by a softmax distribution and that π_{new} is obtained by one local gradient step on the conditional surrogate in Eq. 10, with the local step size bounded above by $\bar{\beta}$. Then, for every state s,*

$$D_{\mathrm{TV}}\left(\pi_{\mathrm{new}}(\cdot|s), \pi_{\mathrm{old}}(\cdot|s)\right) \leq \frac{\bar{\beta}C}{2}\left(1 - \exp\{-\mathcal{H}_{\mathrm{old}}(s)\}\right) \leq \frac{\bar{\beta}C\mathcal{H}_{\max}}{2}\rho(s), \tag{11}$$

where $C = \max_{s,a} |A^{info}_{\pi_{old}}(s,a)|$. Hence, with $\alpha = \bar{\beta}C\mathcal{H}_{\max}/2$, the local update satisfies the state-dependent divergence scaling in Eq. 7.

The proof is provided in Appendix A.2. Having established the required divergence scaling, we next examine the effect of recomputing the information density across policy iterations. Let ρ_k denote the density induced by π_k, and write $\eta^{\mathrm{info}}_{\rho}(\pi)$ for the return under density ρ. With ρ_k fixed during the update $\pi_k \to \pi_{k+1}$, Theorem 4.2 compares both policies under a common clock. After the update, recomputing ρ introduces a return perturbation bounded by Proposition A.8. Combining this with Theorem 4.2 yields the following comparison between policy iterates under a fixed reference clock.

Proposition 4.4. *Let π_x and π_y, with $x < y$, be two policy iterates from the same sequence of updates, and let ρ_r be any fixed information-density snapshot used as a common reference clock. Suppose that $|R_T| \leq R_{\max} < \infty$. Define the cumulative certified gain between the two iterates as*

$$\mathcal{G}_{x,y} \coloneqq \sum_{t=x}^{y-1}\left[L^{\mathrm{info}}_{\pi_t}(\pi_{t+1}) - \eta^{\mathrm{info}}_{\rho_t}(\pi_t) - \frac{4C_t\alpha_t^2}{(1-\gamma)^2} - \frac{8R_{\max}}{e}\alpha_t\right], \tag{12}$$

where C_t and α_t denote the corresponding quantities for the update $\pi_t \to \pi_{t+1}$, and all information-time quantities in the t-th summand are evaluated under ρ_t. For a trajectory $\tau = (s_0, a_0, \ldots, s_T)$, define $\bar{\rho}_j(\tau) \coloneqq \frac{1}{T}\sum_{u=0}^{T-1}\rho_j(s_u)$, and $\varepsilon^{(r)}_{x,y} \coloneqq \frac{1}{2}\sum_{j\in\{x,y\}}\mathbb{E}_{\tau\sim\pi_j}\left[\left|\log\frac{\bar{\rho}_r(\tau)}{\bar{\rho}_j(\tau)}\right|\right]$. Then

$$\eta^{\mathrm{info}}_{\rho_r}(\pi_y) - \eta^{\mathrm{info}}_{\rho_r}(\pi_x) \geq \mathcal{G}_{x,y} - \frac{2R_{\max}}{e}\varepsilon^{(r)}_{x,y}. \tag{13}$$

The proof is in Appendix A.9. Here, $\varepsilon^{(r)}_{x,y}$ measures the average relative discrepancy between the trajectory-level information densities of the two endpoint policies and the common reference clock. Proposition 4.4 therefore shows that moderate variation in the information clock induces a limited perturbation to comparisons between policy iterates, making their relative comparison less sensitive to clock recomputation across training. In the experiments, we further track the evolution of normalized entropy throughout training to empirically assess how the information clock evolves in practice.

4.4 Information-Time Policy Optimization

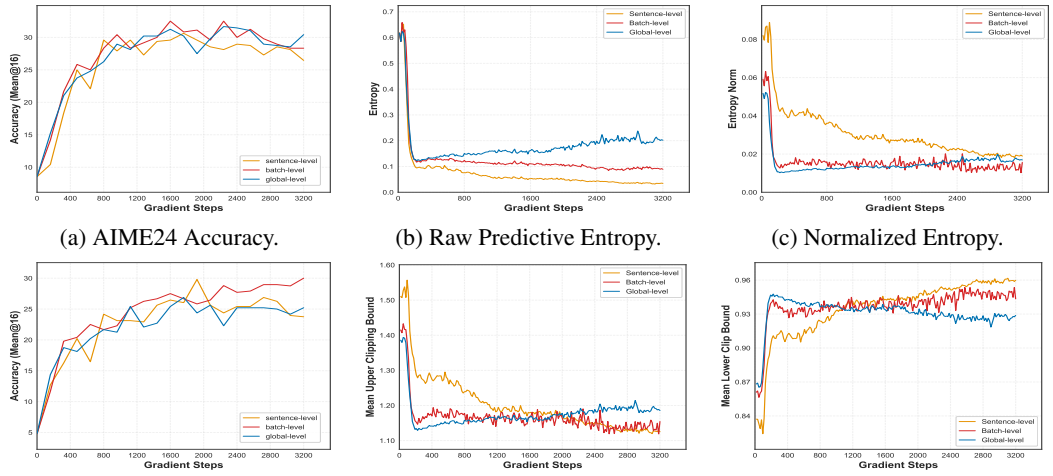
With the information clock specified, we incorporate ρ_t into advantage estimation and policy optimization. Replacing token-time decay in standard GAE with information-time decay gives

$$\delta^{\mathrm{info}}_t = r_t + \gamma^{\rho_t}V_\phi(s_{t+1}) - V_\phi(s_t), \qquad \hat{A}^{\mathrm{info}}_t = \delta^{\mathrm{info}}_t + (\gamma\lambda)^{\rho_t}\hat{A}^{\mathrm{info}}_{t+1}. \tag{14}$$

By defining the information-time trace decay as $\Lambda^{\mathrm{info}}(t_1,t_2) \coloneqq \lambda^{\sum_{j=t_1}^{t_2-1}\rho_j}$, the recursion can be equivalently written as $\hat{A}^{\mathrm{info}}_t = \sum_{l=0}^{T-t-1}\Gamma^{\mathrm{info}}(t,t+l)\Lambda^{\mathrm{info}}(t,t+l)\delta^{\mathrm{info}}_{t+l}$.

Theorem 4.2 motivates a state-dependent update scale. Following the PPO paradigm, we realize this update geometry through an adaptive clipping range. A direct choice would scale the bounds linearly as $[1-\epsilon^{\mathrm{info}}_{\mathrm{low}}\rho_t, 1+\epsilon^{\mathrm{info}}_{\mathrm{high}}\rho_t]$. While this linear construction captures the desired dependence on information density, it can yield relatively permissive updates at high-information states. We therefore adopt a more conservative logarithmic scaling. Specifically, we define the upper and lower clipping bounds as $\xi^{\mathrm{info}}_{\mathrm{high},t} \coloneqq 1 + \log\left(1+\epsilon^{\mathrm{info}}_{\mathrm{high}}\rho_t\right)$ and $\xi^{\mathrm{info}}_{\mathrm{low},t} \coloneqq \frac{1}{1+\log\left(1+\epsilon^{\mathrm{info}}_{\mathrm{low}}\rho_t\right)}$, respectively. Thus, the resulting clipping range expands with information density, while its logarithmic scaling moderates policy updates in high-information regions. Finally, we obtain the optimization objective of InfoPPO:

$$\mathcal{L}^{\text{info-CLIP}}(\theta) = \mathbb{E}_{\pi_{\theta_{\mathrm{old}}}}\left[\min\left(\omega_t(\theta)\hat{A}^{\mathrm{info}}_t, \mathrm{clip}\big(\omega_t(\theta), \xi^{\mathrm{info}}_{\mathrm{low},t}, \xi^{\mathrm{info}}_{\mathrm{high},t}\big)\,\hat{A}^{\mathrm{info}}_t\right)\right]. \tag{15}$$



(a) AIME24 Accuracy. (b) Raw Predictive Entropy. (c) Normalized Entropy.

(d) AIME25 Accuracy. (e) Mean Upper Clipping Bound. (f) Mean Lower Clipping Bound.

Figure 2: Effects of Entropy Normalization. We empirically compare alternative normalization schemes by examining their effects on normalized entropy, adaptive clipping, and performance.

5 Experiments

We organize our experiments to examine three aspects of InfoPPO. We first characterize the information-time quantities induced during training. We then compare information-time and token-time PPO under non-trivial discount and trace-decay settings. Finally, we evaluate InfoPPO against representative RLVR baselines across model scales and mathematical reasoning benchmarks.

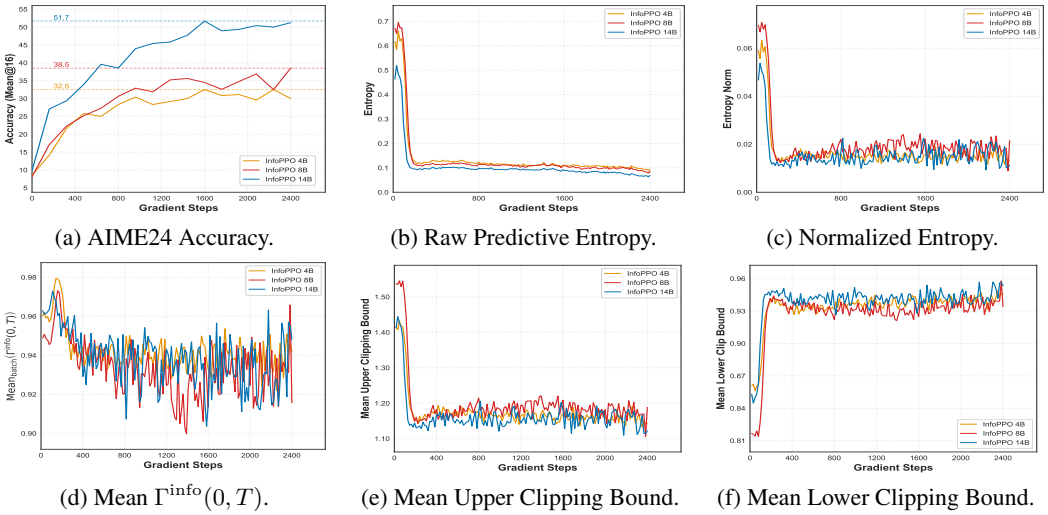
Experimental Setup. We train Qwen3 models (Yang et al., 2025) on DAPO-Math-17K (Yu et al., 2025) and compare InfoPPO against PPO, DAPO, and DAPO-FT. We evaluate on five challenging competition-style mathematical reasoning benchmarks: AMC23 (Li et al., 2024), AIME24 (Zhang & Math-AI, 2024), AIME25 (Zhang & Math-AI, 2025), AIME26 (Zhang & Math-AI, 2026), and BeyondAIME (ByteDance-Seed, 2025). For each prompt, we independently sample 16 responses in a zero-shot setting at temperature $T = 1.0$ and report their average accuracy (Mean@16). Training configurations and implementation details are provided in Appendix C.

5.1 Empirical Characterization of the Information Clock

Section 4.3 instantiates the information density as normalized entropy, $\rho(s_t) = \mathcal{H}_{\text{old}}(s_t)/\mathcal{H}_{\max}$. We examine two complementary aspects of this construction. First, using globally fixed normalization as the reference, we further evaluate batch- and sentence-level alternatives empirically, focusing on the trade-off between local scale adaptation and information-clock stability. Second, we then track information density, temporal discounting, and adaptive clipping during training across model scales.

Entropy Normalization. Figure 2 compares global-, batch-, and sentence-level entropy normalization. Relative to a globally fixed reference scale, more local normalization can make better use of the effective dynamic range of normalized entropy by adapting the scale to the entropy statistics of the current samples. This increased local adaptivity, however, makes the normalization scale itself data-dependent and can introduce additional variation into the resulting information density. The three schemes therefore exhibit different trade-offs. Global normalization preserves a common reference scale but can compress locally relevant entropy variation into a relatively narrow range. Sentence-level normalization provides the strongest local rescaling, but exhibits larger fluctuations and removes the absolute entropy-scale information across responses. Batch-level normalization lies between these two extremes, preserving meaningful local variation while retaining a more consistent reference across samples. Empirically, it maintains relatively stable normalized-entropy and clipping dynamics while achieving competitive performance. We therefore adopt it in remaining experiments.

Information-Time Dynamics Across Model Scales. Figure 3 characterizes the information-time dynamics across Qwen3 model scales. After an initial transient, the normalized entropy defining ρ_t remains relatively stable, suggesting that the overall scale of the information clock does not exhibit large systematic drift as the policy evolves. This is consistent with the moderate clock variation regime characterized by the cross-iteration analysis in Section 4.3. The mean terminal information-



(a) AIME24 Accuracy. (b) Raw Predictive Entropy. (c) Normalized Entropy.

(d) Mean $\Gamma^{\text{info}}(0,T)$. (e) Mean Upper Clipping Bound. (f) Mean Lower Clipping Bound.

Figure 3: Information-time training dynamics of InfoPPO on Qwen3-4B, 8B, and 14B base models. (a) reports AIME24 accuracy; (b)–(c) show the raw predictive entropy and its normalized form used as the information density ρ_t; (d) reports the mean terminal information-time discount $\Gamma^{\text{info}}(0,T)$; and (e)–(f) show the corresponding upper and lower adaptive clipping bounds.

time discount $\Gamma^{\text{info}}(0,T)$ remains substantial throughout training, indicating that information-time discounting avoids severe attenuation of terminal supervision along typical trajectories. Meanwhile, the adaptive clipping bounds remain within stable ranges, and AIME24 accuracy improves over training across model scales. These results show that the information clock induces well-behaved temporal weighting and state-dependent update scales across model scales.

5.2 Information-Time vs. Token-Time PPO

We next compare information-time PPO with token-time PPO under non-trivial discount and trace-decay settings to examine their optimization behavior in long-horizon reasoning.

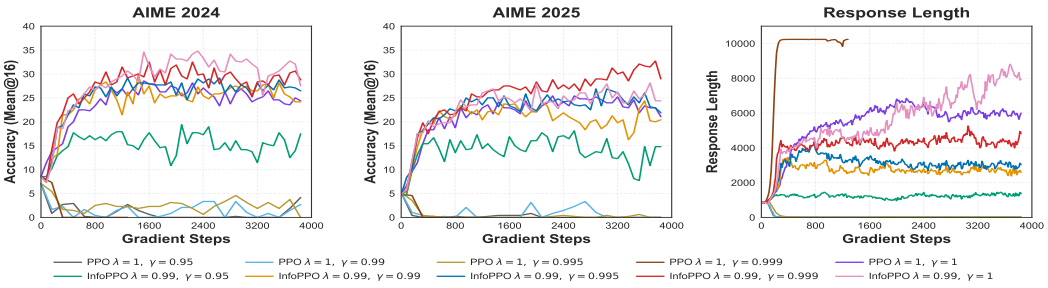
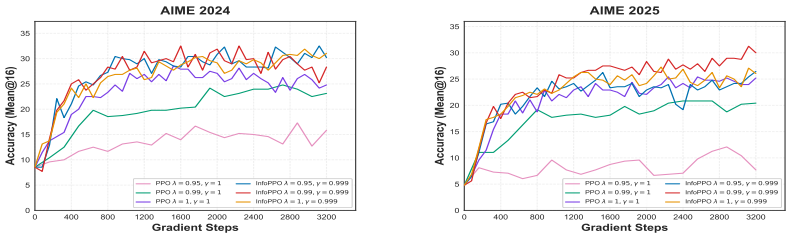


Figure 4: Comparison of information-time and token-time discounting under different γ settings. All experiments are conducted using the Qwen3-4B Base Model. (a) and (b) report the accuracy results on AIME24 and AIME25, respectively. (c) illustrates the average response length.

Discounting over Information Time. Figure 4 compares the two temporal parameterizations across γ settings. Token-time PPO is highly sensitive to non-trivial discounting. For $\gamma < 1$, accuracy deteriorates sharply on both AIME24 and AIME25, while response length exhibits pronounced and often degenerate changes. Notably, at $\gamma = 0.999$, response length rapidly saturates near the maximum generation length. This sharp variation highlights the instability of token-time discounting. In contrast, information-time PPO maintains substantially stronger accuracy and more gradual response-length changes across the tested γ range. This behavior is consistent with the different accumulation of temporal decay. Token-time discounting compounds directly with raw sequence length as γ^T, whereas information-time discounting accumulates according to $\gamma^{\sum_t \rho_t}$. Therefore, the latter retains effective-horizon control while mitigating excessive attenuation of terminal supervision.

Trace Decay over Information Time. Figure 5 compares the two parameterizations across different λ settings. Reducing λ below one substantially degrades token-time PPO performance, whereas InfoPPO remains competitive across non-trivial λ values. Notably, InfoPPO with $\lambda = 0.99$ achieves particularly strong performance, suggesting that non-trivial trace decay can provide a more favorable



(a) AIME24 Accuracy (b) AIME25 Accuracy

Figure 5: Comparison of information-time and token-time trace decay under different λ settings.

bias–variance trade-off in GAE. Under token time, $\lambda^{t_2-t_1}$ decays with token distance, rapidly attenuating credit over long trajectories. In InfoPPO, decay instead accumulates over information distance, supporting long-range credit propagation while retaining control over trace decay.

These results show that information time enables non-trivial temporal decay while preserving effective-horizon control. Based on the results, we use $\gamma = 0.999$ and $\lambda = 0.99$ for InfoPPO in the remaining experiments, while retaining $\gamma = \lambda = 1$ for PPO.

5.3 Overall Performance Comparison

We now evaluate InfoPPO against representative RLVR baselines across model scales and mathematical reasoning benchmarks to assess whether the observed benefits of information time translate into consistent reasoning gains while maintaining stable response lengths.

Table 1: Performance comparison of InfoPPO with PPO, DAPO, and DAPO-FT. Accuracy is averaged over 16 independently sampled responses per prompt (Mean@16). **Bold** indicates the best result.

Benchmark	DAPO		DAPO-FT		PPO		InfoPPO	
	Score	Len.	Score	Len.	Score	Len.	Score	Len.
Qwen3-4B Base Model								
AMC23	73.4	3938.7	75.0	4013.0	72.8	6583.0	**76.6**	3792.5
AIME24	30.4	5650.1	31.1	6286.2	28.1	8552.0	**32.5**	5855.8
AIME25	25.2	5098.3	25.4	5415.1	25.4	8322.0	**28.8**	5545.9
AIME26	21.7	5457.5	20.6	6019.3	20.2	8632.6	**23.8**	5863.3
BeyondAIME	10.3	4820.6	12.3	5580.7	12.3	8428.7	**12.8**	4714.9
Average	32.2	4993.0	32.9	5462.9	31.8	8103.7	**34.9**	5154.5
Qwen3-8B Base Model								
AMC23	80.5	3206.6	81.7	2855.4	76.9	5230.8	**84.4**	3233.2
AIME24	32.3	4789.9	31.3	4919.0	31.3	7263.6	**38.5**	5657.9
AIME25	27.5	4099.5	23.1	3956.1	25.4	6499.4	**30.8**	4795.9
AIME26	28.1	4833.6	26.0	4758.3	26.5	7483.6	**33.5**	5495.4
BeyondAIME	13.0	3924.7	12.4	3957.5	12.6	7160.8	**18.8**	4878.8
Average	36.3	4170.9	34.9	4089.3	34.5	6727.6	**41.2**	4812.2
Qwen3-14B Base Model								
AMC23	87.2	3680.5	86.1	4382.5	86.4	5066.1	**89.1**	3497.4
AIME24	47.7	6519.0	47.5	7622.8	45.4	8254.1	**51.7**	6334.6
AIME25	34.8	6326.4	37.1	7231.8	30.8	8024.7	**37.5**	6751.6
AIME26	37.1	6459.9	41.3	7508.1	37.3	8045.8	**45.8**	6801.0
BeyondAIME	19.8	5746.1	20.9	7245.0	19.4	7954.2	**22.7**	6260.1
Average	45.3	5746.4	46.6	6798.0	43.9	7469.0	**49.4**	5928.9

Table 1 summarizes the results. InfoPPO achieves higher average accuracy than all three baselines. Its average response lengths remain broadly comparable to those of DAPO and DAPO-FT and are substantially shorter than those of PPO. These results indicate that InfoPPO achieves a better accuracy–length trade-off by improving reasoning efficiency rather than increasing trajectory length. The corresponding training dynamics across model scales are provided in Appendix D, which further show that InfoPPO improves accuracy while maintaining relatively stable response lengths throughout training. Together, these results show that InfoPPO translates the information-time formulation into improved reasoning performance across model scales, while maintaining response efficiency.

6 Conclusion

We propose InfoPPO, which measures temporal progression through accumulated information rather than raw token count. We extend the performance-difference analysis to information time and derive a policy-improvement bound linking policy divergence to local information density. Experiments across Qwen3 model scales show consistent reasoning gains with stable response lengths.

AI USE STATEMENT

We used generative AI tools to assist with grammar correction, language editing, and manuscript presentation and organization. These tools were not used to originate the core research question. All final decisions regarding methodology, problem formulation, theoretical claims, experimental design, and interpretation of results were made by the authors. We reviewed all AI-assisted revisions and take responsibility for the final content of this work.

REPRODUCIBILITY STATEMENT

The information-time formulation and optimization objective are described in Section 4, with theoretical proofs provided in Appendix A.1 and Appendix A.2. The datasets and evaluation protocol are described in Section 5, and training configurations and hyperparameters are provided in Appendix C. We plan to publicly release the code for our method.

REFERENCES

Pierre-Luc Bacon, Jean Harb, and Doina Precup. The option-critic architecture. In *Proceedings of the AAAI conference on artificial intelligence*, volume 31, 2017.

Greg Brockman, Vicki Cheung, Ludwig Pettersson, Jonas Schneider, John Schulman, Jie Tang, and Wojciech Zaremba. Openai gym. *arXiv preprint arXiv:1606.01540*, 2016.

ByteDance-Seed. BeyondAIME: Advancing math reasoning evaluation beyond high school olympiads, 2025. URL `https://huggingface.co/datasets/ByteDance-Seed/BeyondAIME`.

Eugene A Feinberg and Adam Shwartz. Markov decision models with weighted discounted criteria. *Mathematics of Operations Research*, 19(1):152–168, 1994.

Yichao Fu, Xuewei Wang, Yuandong Tian, and Jiawei Zhao. Deep think with confidence. *arXiv preprint arXiv:2508.15260*, 2025.

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

Aaron Jaech, Adam Kalai, Adam Lerer, Adam Richardson, Ahmed El-Kishky, Aiden Low, Alec Helyar, Aleksander Madry, Alex Beutel, Alex Carney, et al. Openai o1 system card. *arXiv preprint arXiv:2412.16720*, 2024.

Sham Kakade and John Langford. Approximately optimal approximate reinforcement learning. In *Proceedings of the nineteenth international conference on machine learning*, pp. 267–274, 2002.

Nathan Lambert, Jacob Morrison, Valentina Pyatkin, Shengyi Huang, Hamish Ivison, Faeze Brahman, Lester James V Miranda, Alisa Liu, Nouha Dziri, Shane Lyu, et al. T\" ulu 3: Pushing frontiers in open language model post-training. *arXiv preprint arXiv:2411.15124*, 2024.

Jia Li, Edward Beeching, Lewis Tunstall, Ben Lipkin, Roman Soletskyi, Shengyi Huang, Kashif Rasul, Longhui Yu, Albert Q Jiang, Ziju Shen, et al. Numinamath: The largest public dataset in ai4maths with 860k pairs of competition math problems and solutions. *Hugging Face repository*, 13:9, 2024.

Doina Precup. *Temporal abstraction in reinforcement learning*. University of Massachusetts Amherst, 2000.

Xingwei Qu, Shaowen Wang, Zihao Huang, Kai Hua, Fan Yin, Rui-Jie Zhu, Jundong Zhou, Qiyang Min, Zihao Wang, Yizhi Li, et al. Dynamic large concept models: Latent reasoning in an adaptive semantic space. *arXiv preprint arXiv:2512.24617*, 2025.

John Schulman, Sergey Levine, Pieter Abbeel, Michael Jordan, and Philipp Moritz. Trust region policy optimization. In *International conference on machine learning*, pp. 1889–1897. PMLR, 2015a.

John Schulman, Philipp Moritz, Sergey Levine, Michael Jordan, and Pieter Abbeel. High-dimensional continuous control using generalized advantage estimation. *arXiv preprint arXiv:1506.02438*, 2015b.

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

Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, YK Li, Y Wu, et al. Deepseekmath: Pushing the limits of mathematical reasoning in open language models. *arXiv preprint arXiv:2402.03300*, 2024.

Richard S Sutton, Doina Precup, and Satinder Singh. Between mdps and semi-mdps: A framework for temporal abstraction in reinforcement learning. *Artificial intelligence*, 112(1-2):181–211, 1999.

Shenzhi Wang, Le Yu, Chang Gao, Chujie Zheng, Shixuan Liu, Rui Lu, Kai Dang, Xionghui Chen, Jianxin Yang, Zhenru Zhang, et al. Beyond the 80/20 rule: High-entropy minority tokens drive effective reinforcement learning for llm reasoning. *arXiv preprint arXiv:2506.01939*, 2025.

Martha White. Unifying task specification in reinforcement learning. In *International Conference on Machine Learning*, pp. 3742–3750. PMLR, 2017.

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

Qiying Yu, Zheng Zhang, Ruofei Zhu, Yufeng Yuan, Xiaochen Zuo, Yu Yue, Tiantian Fan, Gaohong Liu, Lingjun Liu, Xin Liu, et al. Dapo: An open-source llm reinforcement learning system at scale. *arXiv preprint arXiv:2503.14476*, 2025.

Yifan Zhang and Team Math-AI. American invitational mathematics examination (aime) 2024, 2024.

Yifan Zhang and Team Math-AI. American invitational mathematics examination (aime) 2025, 2025.

Yifan Zhang and Team Math-AI. American invitational mathematics examination (aime) 2026, 2026.

A Mathematical Derivations

A.1 Proofs for Policy Improvement on the Information-Time MDP

Lemma A.1 (Information-Time Performance Difference)**.** *Given two policies π and $\tilde{\pi}$ within the Information-Time MDP framework, the following identity holds:*

$$\eta^{info}(\tilde{\pi}) - \eta^{info}(\pi) = \mathbb{E}_{\tau\sim\tilde{\pi}}\left[\sum_{t=0}^{\infty}\Gamma^{info}(0,t)A_{\pi}^{info}(s_t,a_t)\right], \tag{16}$$

where $\Gamma^{info}(0,t)$ represents the cumulative information-time discount factor.

Proof. First, we verify that the cumulative information discount factor $\Gamma^{\text{info}}(t_1,t_2)$ satisfies the following multiplicative property:

$$\Gamma^{\text{info}}(t_1,t_2) = \gamma^{\sum_{t=t_1}^{t_2-1}\Delta\mu_{\text{info}}(t)} = \gamma^{\sum_{t_1}^{t_{\text{mid}}-1}\Delta\mu_{\text{info}}} \times \gamma^{\sum_{t_{\text{mid}}}^{t_2-1}\Delta\mu_{\text{info}}} = \Gamma^{\text{info}}(t_1,t_{\text{mid}})\Gamma^{\text{info}}(t_{\text{mid}},t_2). \tag{17}$$

Recalling the definition $Q_{\pi}^{\text{info}}(s_t,a_t) \coloneqq \mathbb{E}_{\pi}\left[\sum_{l=0}^{\infty}\Gamma^{\text{info}}(t,t+l)r(s_{t+l})\middle|s_t,a_t\right]$, we proceed to derive the Bellman equation on the information-time:

$$\begin{aligned}
Q_{\pi}^{\text{info}}(s_t,a_t) &= \mathbb{E}_{\pi}\left[\sum_{l=0}^{\infty}\Gamma^{\text{info}}(t,t+l)r(s_{t+l})\middle|s_t,a_t\right] &(18)\\
&= \mathbb{E}_{\pi}\left[r(s_t) + \Gamma^{\text{info}}(t,t+1)\sum_{l=t+1}^{\infty}\Gamma^{\text{info}}(t+1,l)r(s_l)\middle|s_t,a_t\right] &(19)\\
&= r(s_t) + \Gamma^{\text{info}}(t,t+1)\mathbb{E}_{\pi}\left[\sum_{l=t+1}^{\infty}\Gamma^{\text{info}}(t+1,l)r(s_l)\middle|s_t,a_t\right] &(20)\\
&= r(s_t) + \Gamma^{\text{info}}(t,t+1)V_{\pi}^{\text{info}}(s_{t+1}). &(21)
\end{aligned}$$

Consequently, for the advantage function $A_{\pi}^{\text{info}}(s,a)$, we obtain

$$A_{\pi}^{\text{info}}(s,a) = Q_{\pi}^{\text{info}}(s,a) - V_{\pi}^{\text{info}}(s) = r(s_t) + \Gamma^{\text{info}}(t,t+1)V_{\pi}^{\text{info}}(s_{t+1}) - V_{\pi}^{\text{info}}(s_t). \tag{22}$$

Therefore,

$$\begin{aligned}
&\mathbb{E}_{\tau\sim\tilde{\pi}}\left[\sum_{t=0}^{\infty}\Gamma^{\text{info}}(0,t)A_{\pi}^{\text{info}}(s_t,a_t)\right] &(23)\\
&= \mathbb{E}_{\tau\sim\tilde{\pi}}\left[\sum_{0}^{\infty}\Gamma^{\text{info}}(0,t)\left(r(s_t) + \Gamma^{\text{info}}(t,t+1)V_{\pi}^{\text{info}}(s_{t+1}) - V_{\pi}^{\text{info}}(s_t)\right)\right] &(24)\\
&= \mathbb{E}_{\tau\sim\tilde{\pi}}\left[\sum_{0}^{\infty}\left(\Gamma^{\text{info}}(0,t+1)V_{\pi}^{\text{info}}(s_{t+1}) - \Gamma^{\text{info}}(0,t)V_{\pi}^{\text{info}}(s_t)\right) + \sum_{t=0}^{\infty}\Gamma^{\text{info}}(0,t)r(s_t)\right] &(25)\\
&= \mathbb{E}_{\tau\sim\tilde{\pi}}\left[-\Gamma^{\text{info}}(0,0)V_{\pi}^{\text{info}}(s_0) + \sum_{t=0}^{\infty}\Gamma^{\text{info}}(0,t)r(s_t)\right] &(26)\\
&= \mathbb{E}_{\tau\sim\tilde{\pi}}\left[-V_{\pi}^{\text{info}}(s_0) + \sum_{t=0}^{\infty}\Gamma^{\text{info}}(0,t)r(s_t)\right] &(27)\\
&= -\eta^{\text{info}}(\pi) + \eta^{\text{info}}(\tilde{\pi}). &(28)
\end{aligned}$$

For the infinite-horizon telescoping argument above, the remaining boundary term is required to vanish:

$$\lim_{N\to\infty}\mathbb{E}_{\tau\sim\tilde{\pi}}\left[\Gamma^{\text{info}}(0,N+1)V_{\pi}^{\text{info}}(s_{N+1})\right] = 0.$$

This condition is naturally satisfied in episodic environments with terminal states, since the continuation value is zero after termination. In particular, it holds for the language-generation setting considered in this work, where each response eventually terminates and therefore has zero continuation value thereafter. □

To connect the trajectory form of the Information-Time Performance Difference with the state-space formulation used in Section 4.2, we now derive its equivalent representation in terms of the information-discounted state visitation measure.

Corollary A.2. (State-Space Representation of the Information-Time Performance Difference). *Define the unnormalized information-discounted state visitation measure under policy π as*

$$\nu_\pi^{\mathrm{info}}(s) \coloneqq \mathbb{E}_{\tau\sim\pi}\left[\sum_{t=0}^{\infty}\Gamma^{\mathrm{info}}(0,t)\mathbf{1}\{s_t=s\}\right].$$

Then

$$\eta^{\mathrm{info}}(\tilde\pi)-\eta^{\mathrm{info}}(\pi)=\sum_s \nu_{\tilde\pi}^{\mathrm{info}}(s)\sum_a \tilde\pi(a|s)A_\pi^{\mathrm{info}}(s,a). \tag{29}$$

Proof. Starting from the Information-Time Performance Difference in Lemma A.1,

$$\eta^{\mathrm{info}}(\tilde\pi)-\eta^{\mathrm{info}}(\pi)=\mathbb{E}_{\tau\sim\tilde\pi}\left[\sum_{t=0}^{\infty}\Gamma^{\mathrm{info}}(0,t)A_\pi^{\mathrm{info}}(s_t,a_t)\right]. \tag{30}$$

For each fixed time t, we introduce an indicator over the current state:

$$A_\pi^{\mathrm{info}}(s_t,a_t)=\sum_s \mathbf{1}\{s_t=s\}A_\pi^{\mathrm{info}}(s,a_t). \tag{31}$$

Substituting this identity into Eq. 30 gives

$$\begin{aligned}\eta^{\mathrm{info}}(\tilde\pi)-\eta^{\mathrm{info}}(\pi)&=\sum_{t=0}^{\infty}\mathbb{E}_{\tau\sim\tilde\pi}\left[\Gamma^{\mathrm{info}}(0,t)\sum_s \mathbf{1}\{s_t=s\}A_\pi^{\mathrm{info}}(s,a_t)\right]\\&=\sum_{t=0}^{\infty}\sum_s \mathbb{E}_{\tau\sim\tilde\pi}\left[\Gamma^{\mathrm{info}}(0,t)\mathbf{1}\{s_t=s\}A_\pi^{\mathrm{info}}(s,a_t)\right].\end{aligned} \tag{32}$$

Let h_t denote the trajectory history up to the current state s_t, before a_t is sampled. Conditional on h_t, both $\Gamma^{\mathrm{info}}(0,t)$ and $\mathbf{1}\{s_t=s\}$ are determined, whereas the current action is sampled according to $a_t\sim\tilde\pi(\cdot|s_t)$. Hence, applying iterated expectation first over the current action and then over the history gives

$$\begin{aligned}&\mathbb{E}_{\tau\sim\tilde\pi}\left[\Gamma^{\mathrm{info}}(0,t)\mathbf{1}\{s_t=s\}A_\pi^{\mathrm{info}}(s,a_t)\right]\\&=\mathbb{E}_{h_t\sim\tilde\pi}\left[\Gamma^{\mathrm{info}}(0,t)\mathbf{1}\{s_t=s\}\,\mathbb{E}_{a_t\sim\tilde\pi(\cdot|s_t)}\left[A_\pi^{\mathrm{info}}(s,a_t)\right]\right].\end{aligned} \tag{33}$$

On the event $\{s_t=s\}$,

$$\mathbb{E}_{a_t\sim\tilde\pi(\cdot|s_t)}\left[A_\pi^{\mathrm{info}}(s,a_t)\right]=\sum_a \tilde\pi(a|s)A_\pi^{\mathrm{info}}(s,a). \tag{34}$$

This quantity depends only on the fixed state s and is therefore constant with respect to the outer expectation over h_t. Hence,

$$\begin{aligned}&\mathbb{E}_{\tau\sim\tilde\pi}\left[\Gamma^{\mathrm{info}}(0,t)\mathbf{1}\{s_t=s\}A_\pi^{\mathrm{info}}(s,a_t)\right]\\&=\mathbb{E}_{h_t\sim\tilde\pi}\left[\Gamma^{\mathrm{info}}(0,t)\mathbf{1}\{s_t=s\}\right]\sum_a \tilde\pi(a|s)A_\pi^{\mathrm{info}}(s,a).\end{aligned} \tag{35}$$

Substituting Eq. 35 into Eq. 32 and collecting the contributions associated with each state yields

$$\begin{aligned}\eta^{\text{info}}(\tilde{\pi}) - \eta^{\text{info}}(\pi) &= \sum_s \left[\sum_{t=0}^{\infty} \mathbb{E}_{\tau\sim\tilde{\pi}} \left[\Gamma^{\text{info}}(0,t)\mathbf{1}\{s_t = s\}\right]\right] \sum_a \tilde{\pi}(a|s) A_\pi^{\text{info}}(s,a) \\ &= \sum_s \mathbb{E}_{\tau\sim\tilde{\pi}} \left[\sum_{t=0}^{\infty} \Gamma^{\text{info}}(0,t)\mathbf{1}\{s_t = s\}\right] \sum_a \tilde{\pi}(a|s) A_\pi^{\text{info}}(s,a) \\ &= \sum_s \nu_{\tilde{\pi}}^{\text{info}}(s) \sum_a \tilde{\pi}(a|s) A_\pi^{\text{info}}(s,a), \end{aligned} \tag{36}$$

where the final equality follows from the definition of $\nu_{\tilde{\pi}}^{\text{info}}(s)$. □

Corollary A.2 provides the state-space representation used to define the surrogate in Eq. 6. We now turn to bounding the approximation error introduced by replacing $\nu_{\tilde{\pi}}^{\text{info}}$ with ν_{π}^{info}. For this purpose, it is convenient to return to the equivalent trajectory representation and follow the coupling argument of TRPO (Schulman et al., 2015a).

To do so, we first define $\bar{A}^{\text{info}}(s)$ as the expected information-time advantage of the candidate policy $\tilde{\pi}$ relative to the current policy π at state s:

$$\bar{A}^{\text{info}}(s) = \mathbb{E}_{a\sim\tilde{\pi}(\cdot|s)} \left[A_\pi^{\text{info}}(s,a)\right]. \tag{37}$$

The Information-Time Performance Difference (Lemma A.1) can then be written as:

$$\eta^{\text{info}}(\tilde{\pi}) = \eta^{\text{info}}(\pi) + \mathbb{E}_{\tau\sim\tilde{\pi}} \left[\sum_{t=0}^{\infty} \Gamma^{\text{info}}(0,t)\bar{A}^{\text{info}}(s_t)\right]. \tag{38}$$

The surrogate objective can be written as

$$L_\pi^{\text{info}}(\tilde{\pi}) = \eta^{\text{info}}(\pi) + \mathbb{E}_{\tau\sim\pi} \left[\sum_{t=0}^{\infty} \Gamma^{\text{info}}(0,t)\bar{A}^{\text{info}}(s_t)\right]. \tag{39}$$

The two expressions differ only in the trajectory distribution. The true return uses trajectories generated by the candidate policy $\tilde{\pi}$, whereas the surrogate uses trajectories generated by the current policy π. To compare these distributions under the state-wise total-variation constraint in Eq. 7, we place the two policies on a common probability space through a state-wise coupling.

Definition A.3. $(\pi, \tilde{\pi})$ is an *$\alpha\rho$-coupled policy pair* if, for every state s, there exists a joint distribution $(a, \tilde{a})|s$ satisfying $P(a \neq \tilde{a}|s) \leq \alpha\rho(s)$, where $\pi(\cdot|s)$ and $\tilde{\pi}(\cdot|s)$ are the marginal distributions of a and $\tilde{a}$, respectively.

By the standard coupling characterization of total variation, the divergence condition in Eq. 7 guarantees the existence of such a coupling. Thus, the coupling condition used below is the coupling representation of the state-wise total-variation condition assumed in Theorem 4.2.

Lemma A.4. *Given that $\pi, \tilde{\pi}$ are $\alpha\rho$-coupled policies, for all s,*

$$|\bar{A}^{\textit{info}}(s)| \leq 2\alpha\rho(s) \max_{s,a} |A_\pi^{\textit{info}}(s,a)| \tag{40}$$

Proof.

$$\bar{A}^{\text{info}}(s) = \mathbb{E}_{\tilde{a}\sim\tilde{\pi}} \left[A_\pi^{\text{info}}(s,\tilde{a})\right] = \mathbb{E}_{(a,\tilde{a})\sim(\pi,\tilde{\pi})} \left[A_\pi^{\text{info}}(s,\tilde{a}) - A_\pi^{\text{info}}(s,a)\right] \tag{41}$$

$$= P(a \neq \tilde{a}|s) \mathbb{E}_{(a,\tilde{a})\sim(\pi,\tilde{\pi})|a\neq\tilde{a}} \left[A_\pi^{\text{info}}(s,\tilde{a}) - A_\pi^{\text{info}}(s,a)\right] \tag{42}$$

$$|\bar{A}^{\text{info}}(s)| \leq \alpha\rho(s) \cdot 2 \max_{s,a} |A_\pi^{\text{info}}(s,a)| = 2\alpha\rho(s) \max_{s,a} |A_\pi^{\text{info}}(s,a)| \tag{43}$$

□

Next we establish a bound on the cumulative discounted information weights, which will be used to control the surrogate approximation error.

Lemma A.5. *For any sequence of information densities $\{\rho(s_t)\}_{t=0}^{\infty}$, the following bound holds:*

$$\sum_{t=0}^{\infty} \Gamma^{\textit{info}}(0,t)\rho(s_t) \leq \frac{1}{1-\gamma}. \tag{44}$$

Proof. For brevity, let $\rho_t \coloneqq \rho(s_t)$. Consider the function $f(x) = 1 - \gamma^x$. Since f is concave, the chord between endpoints 0 and 1 gives

$$1 - \gamma^{\rho_t} = f(\rho_t) \geq (1-\rho_t)f(0) + \rho_t f(1) = \rho_t(1-\gamma). \tag{45}$$

Rearranging terms yields

$$\rho_t \leq \frac{1-\gamma^{\rho_t}}{1-\gamma}. \tag{46}$$

Multiplying by $\Gamma^{\text{info}}(0,t)$ and summing over t:

$$\sum_{t=0}^{\infty} \Gamma^{\text{info}}(0,t)\rho_t \leq \frac{1}{1-\gamma}\sum_{t=0}^{\infty} \Gamma^{\text{info}}(0,t)(1-\gamma^{\rho_t}) \tag{47}$$

$$= \frac{1}{1-\gamma}\sum_{t=0}^{\infty}(\Gamma^{\text{info}}(0,t) - \Gamma^{\text{info}}(0,t+1)) \quad (\text{since } \Gamma^{\text{info}}(0,t+1) = \Gamma^{\text{info}}(0,t)\gamma^{\rho_t}) \tag{48}$$

The sum on the right-hand side is a telescoping sum:

$$\sum_{t=0}^{\infty}(\Gamma^{\text{info}}(0,t) - \Gamma^{\text{info}}(0,t+1)) = \Gamma^{\text{info}}(0,0) - \lim_{t\to\infty}\Gamma^{\text{info}}(0,t) \leq 1. \tag{49}$$

Thus, $\sum_{t=0}^{\infty}\Gamma^{\text{info}}(0,t)\rho(s_t) \leq \frac{1}{1-\gamma}$. □

Theorem A.6. *Let $L_{\pi}^{\textit{info}}(\tilde{\pi})$ be the surrogate objective and $\eta^{\textit{info}}(\tilde{\pi})$ be the true expected return. Assume the policies $(\pi, \tilde{\pi})$ are $\alpha\rho$-coupled. Then:*

$$|\eta^{\textit{info}}(\tilde{\pi}) - L_{\pi}^{\textit{info}}(\tilde{\pi})| \leq \frac{4C\alpha^2}{(1-\gamma)^2}, \tag{50}$$

where $C = \max_{s,a}|A_{\pi}^{\textit{info}}(s,a)|$.

Proof. Let $\Delta = |\eta^{\text{info}}(\tilde{\pi}) - L_{\pi}^{\text{info}}(\tilde{\pi})|$. The difference between the true and surrogate objectives can be expressed as the expected sum of advantages over the trajectories generated by $\tilde{\pi}$ vs π:

$$\Delta = \left|\mathbb{E}_{\tau\sim\tilde{\pi}}\left[\sum_{t=0}^{\infty}\Gamma^{\text{info}}(0,t)\bar{A}^{\text{info}}(s_t)\right] - \mathbb{E}_{\tau\sim\pi}\left[\sum_{t=0}^{\infty}\Gamma^{\text{info}}(0,t)\bar{A}^{\text{info}}(s_t)\right]\right|. \tag{51}$$

We analyze this difference under the joint coupling distribution of the two policies. Let T_{div} be the time step of the first action divergence between the coupled policies. For $t < T_{\text{div}}$, the trajectories are identical, so the difference is zero. We can decompose the expectation using the indicator function $\mathbf{1}\{T_{\text{div}} = k\}$:

$$\Delta = \left|\mathbb{E}_{(\tau,\tilde{\tau})}\left[\sum_{k=0}^{\infty}\mathbf{1}\{T_{\text{div}}=k\}\left(\sum_{t=k}^{\infty}\Gamma^{\text{info}}(0,t)\bar{A}^{\text{info}}(s_t)\Big|_{\tilde{\tau}\sim\tilde{\pi}} - \sum_{t=k}^{\infty}\Gamma^{\text{info}}(0,t)\bar{A}^{\text{info}}(s_t)\Big|_{\tau\sim\pi}\right)\right]\right|. \tag{52}$$

Using the triangle inequality and the fact that $\Gamma^{\text{info}}(0,t) = \Gamma^{\text{info}}(0,k)\Gamma^{\text{info}}(k,t)$, we factor out the common discount term up to step k:

$$\Delta \leq \sum_{k=0}^{\infty}\mathbb{E}_{(\tau,\tilde{\tau})}\Bigg[\mathbf{1}\{T_{\text{div}}=k\}\Gamma^{\text{info}}(0,k)\Bigg(\left|\sum_{t=k}^{\infty}\Gamma^{\text{info}}(k,t)\bar{A}^{\text{info}}(s_t)\Big|_{\tilde{\tau}\sim\tilde{\pi}} + \left|\sum_{t=k}^{\infty}\Gamma^{\text{info}}(k,t)\bar{A}^{\text{info}}(s_t)\Big|_{\tau\sim\pi}\right)\Bigg]. \tag{53}$$

Leveraging the result from Lemma A.4, we obtain:

$$\left|\sum_{t=k}^{\infty}\Gamma^{\text{info}}(k,t)\bar{A}^{\text{info}}(s_t)\right| \le \sum_{t=k}^{\infty}\Gamma^{\text{info}}(k,t)\left(2\alpha\rho(s_t)C\right) = 2\alpha C\sum_{t=k}^{\infty}\Gamma^{\text{info}}(k,t)\rho(s_t) \tag{54}$$

Applying Lemma A.5, we bound the summation term inside the expectation:

$$\Delta \le \sum_{k=0}^{\infty}\mathbb{E}_{(\tau,\tilde{\tau})}\left[\mathbf{1}\{T_{\text{div}}=k\}\Gamma^{\text{info}}(0,k)\left(\frac{2\alpha C}{1-\gamma}+\frac{2\alpha C}{1-\gamma}\right)\right] \tag{55}$$

$$= \frac{4\alpha C}{1-\gamma}\sum_{k=0}^{\infty}\mathbb{E}_{(\tau,\tilde{\tau})}\left[\mathbf{1}\{T_{\text{div}}=k\}\Gamma^{\text{info}}(0,k)\right]. \tag{56}$$

Next, we bound the probability of divergence. Let H_k denote the coupled history up to state s_k, before the actions at step k are sampled. Since $\Gamma^{\text{info}}(0,k)=\gamma^{\sum_{t=0}^{k-1}\rho(s_t)}$ is determined by this history, the law of iterated expectations gives

$$\mathbb{E}_{(\tau,\tilde{\tau})}\left[\mathbf{1}\{T_{\text{div}}=k\}\Gamma^{\text{info}}(0,k)\right] \tag{57}$$

$$= \mathbb{E}_{(\tau,\tilde{\tau})}\left[\Gamma^{\text{info}}(0,k)\,\mathbb{E}\left[\mathbf{1}\{T_{\text{div}}=k\}|H_k\right]\right]. \tag{58}$$

The event $\{T_{\text{div}}=k\}$ occurs if the two trajectories have not diverged before step k and their actions differ at step k. Hence,

$$\{T_{\text{div}}=k\} = \{T_{\text{div}}\ge k\}\cap\{a_k\neq\tilde{a}_k\}. \tag{59}$$

Because whether $T_{\text{div}}\ge k$ holds is already determined by the history H_k,

$$\mathbb{E}\left[\mathbf{1}\{T_{\text{div}}=k\}|H_k\right] = \mathbb{E}\left[\mathbf{1}\{T_{\text{div}}\ge k\}\mathbf{1}\{a_k\neq\tilde{a}_k\}|H_k\right] \tag{60}$$

$$= \mathbf{1}\{T_{\text{div}}\ge k\}P(a_k\neq\tilde{a}_k|H_k) \tag{61}$$

$$\le \mathbf{1}\{T_{\text{div}}\ge k\}\alpha\rho(s_k), \tag{62}$$

where the inequality follows from the $\alpha\rho$-coupling assumption, since on $\{T_{\text{div}}\ge k\}$ the two trajectories share the same state s_k.

Substituting this inequality into the previous expression gives

$$\mathbb{E}_{(\tau,\tilde{\tau})}\left[\mathbf{1}\{T_{\text{div}}=k\}\Gamma^{\text{info}}(0,k)\right] \tag{63}$$

$$\le \alpha\mathbb{E}_{(\tau,\tilde{\tau})}\left[\mathbf{1}\{T_{\text{div}}\ge k\}\Gamma^{\text{info}}(0,k)\rho(s_k)\right] \tag{64}$$

$$\le \alpha\mathbb{E}_{(\tau,\tilde{\tau})}\left[\Gamma^{\text{info}}(0,k)\rho(s_k)\right] \tag{65}$$

$$= \alpha\mathbb{E}_{\tau\sim\pi}\left[\Gamma^{\text{info}}(0,k)\rho(s_k)\right], \tag{66}$$

where the second inequality uses $\mathbf{1}\{T_{\text{div}}\ge k\}\le 1$. In the last equality, s_k and $\Gamma^{\text{info}}(0,k)$ are taken along the π-trajectory; the equality follows because the π-marginal of the trajectory coupling is distributed according to π.

Substituting this bound back into the summation:

$$\Delta \le \frac{4\alpha^2 C}{1-\gamma}\mathbb{E}_{\tau\sim\pi}\left[\sum_{k=0}^{\infty}\Gamma^{\text{info}}(0,k)\rho(s_k)\right]. \tag{67}$$

Finally, applying Lemma A.5 again to the summation term in Eq. 67:

$$\Delta \le \frac{4\alpha^2 C}{1-\gamma}\cdot\frac{1}{1-\gamma} = \frac{4C\alpha^2}{(1-\gamma)^2}. \tag{68}$$

□

A.2 Theoretical Analysis of Information-Time Policy Optimization

To connect the entropy-based information density to the policy update analysis, we examine the policy change induced by the information-time surrogate at a fixed state s. With the information-time advantage evaluated under the old policy, its action-dependent component can be written as

$$\sum_a \tilde{\pi}(a|s) A^{\text{info}}_{\pi_{\text{old}}}(s,a) = \mathbb{E}_{a\sim\pi_{\text{old}}(\cdot|s)}\left[\frac{\tilde{\pi}(a|s)}{\pi_{\text{old}}(a|s)} A^{\text{info}}_{\pi_{\text{old}}}(s,a)\right]. \tag{69}$$

The following proposition characterizes the policy divergence induced by a local gradient step on this objective.

Proposition A.7. *Suppose that π_{old} is parameterized by a softmax distribution and that π_{new} is obtained by one local softmax-logit gradient step induced by the conditional surrogate, with step size $0 \le \beta \le \bar{\beta}$. Then, for every state s,*

$$D_{\text{TV}}\left(\pi_{\text{new}}(\cdot|s), \pi_{\text{old}}(\cdot|s)\right) \le \frac{\bar{\beta}C}{2}\left(1 - \exp\{-\mathcal{H}_{\text{old}}(s)\}\right) \le \frac{\bar{\beta}C\mathcal{H}_{\max}}{2}\rho(s), \tag{70}$$

where $C = \max_{s,a} |A^{\text{info}}_{\pi_{\text{old}}}(s,a)|$.

Proof. Fix an arbitrary state s. By the definition of C,

$$\left|A^{\text{info}}_{\pi_{\text{old}}}(s,a)\right| \le C \qquad \text{for every } a \in \mathcal{A}. \tag{71}$$

Moreover, the information-time advantage has zero expectation under the old policy,

$$\sum_{a\in\mathcal{A}} \pi_{\text{old}}(a|s) A^{\text{info}}_{\pi_{\text{old}}}(s,a) = 0. \tag{72}$$

Therefore, for any action a,

$$\begin{aligned} A^{\text{info}}_{\pi_{\text{old}}}(s,a) &= A^{\text{info}}_{\pi_{\text{old}}}(s,a) - \sum_{b\in\mathcal{A}} \pi_{\text{old}}(b|s) A^{\text{info}}_{\pi_{\text{old}}}(s,b) \\ &= \sum_{b\neq a} \pi_{\text{old}}(b|s)\left[A^{\text{info}}_{\pi_{\text{old}}}(s,a) - A^{\text{info}}_{\pi_{\text{old}}}(s,b)\right]. \end{aligned} \tag{73}$$

Using Eq. 71,

$$\begin{aligned} \left|A^{\text{info}}_{\pi_{\text{old}}}(s,a)\right| &\le \sum_{b\neq a} \pi_{\text{old}}(b|s)\left|A^{\text{info}}_{\pi_{\text{old}}}(s,a) - A^{\text{info}}_{\pi_{\text{old}}}(s,b)\right| \\ &\le 2C \sum_{b\neq a} \pi_{\text{old}}(b|s) \\ &= 2C\left(1 - \pi_{\text{old}}(a|s)\right). \end{aligned} \tag{74}$$

Multiplying by $\pi_{\text{old}}(a|s)$ and summing over actions gives

$$\begin{aligned} &\sum_{a\in\mathcal{A}} \pi_{\text{old}}(a|s)\left|A^{\text{info}}_{\pi_{\text{old}}}(s,a)\right| \\ &\qquad \le 2C \sum_{a\in\mathcal{A}} \pi_{\text{old}}(a|s)\left(1 - \pi_{\text{old}}(a|s)\right) \\ &\qquad = 2C\left(1 - \sum_{a\in\mathcal{A}} \pi_{\text{old}}(a|s)^2\right). \end{aligned} \tag{75}$$

By concavity of the logarithm,

$$-\mathcal{H}_{\text{old}}(s) = \sum_{a\in\mathcal{A}} \pi_{\text{old}}(a|s)\log\pi_{\text{old}}(a|s) \le \log\left(\sum_{a\in\mathcal{A}} \pi_{\text{old}}(a|s)^2\right). \tag{76}$$

Exponentiating both sides yields

$$\exp\{-\mathcal{H}_{\text{old}}(s)\} \le \sum_{a\in\mathcal{A}} \pi_{\text{old}}(a|s)^2. \tag{77}$$

Combining Eq. 75 and Eq. 77,

$$\sum_{a\in\mathcal{A}} \pi_{\text{old}}(a|s) \left|A^{\text{info}}_{\pi_{\text{old}}}(s,a)\right| \le 2C\left(1-\exp\{-\mathcal{H}_{\text{old}}(s)\}\right). \tag{78}$$

Let $z_{\text{old}}(s,\cdot)$ denote the logits of $\pi_{\text{old}}(\cdot|s)$, and let $\pi_z(\cdot|s)$ denote the softmax policy induced by logits $z(s,\cdot)$. Holding the information-time advantage under the old policy fixed, the conditional surrogate in Eq. 69 can be written as

$$\sum_{b\in\mathcal{A}} \pi_z(b|s) A^{\text{info}}_{\pi_{\text{old}}}(s,b). \tag{79}$$

For the softmax mapping,

$$\frac{\partial \pi_z(b|s)}{\partial z(s,a)} = \pi_z(b|s)\left(\mathbf{1}\{a=b\} - \pi_z(a|s)\right). \tag{80}$$

Evaluating the gradient of Eq. 79 at $z = z_{\text{old}}$ and using Eq. 72 gives

$$\begin{aligned}
&\frac{\partial}{\partial z(s,a)} \sum_{b\in\mathcal{A}} \pi_z(b|s) A^{\text{info}}_{\pi_{\text{old}}}(s,b)\bigg|_{z=z_{\text{old}}} \\
&= \pi_{\text{old}}(a|s) A^{\text{info}}_{\pi_{\text{old}}}(s,a) - \pi_{\text{old}}(a|s) \sum_{b\in\mathcal{A}} \pi_{\text{old}}(b|s) A^{\text{info}}_{\pi_{\text{old}}}(s,b) \\
&= \pi_{\text{old}}(a|s) A^{\text{info}}_{\pi_{\text{old}}}(s,a).
\end{aligned} \tag{81}$$

Hence, the local softmax-logit gradient step in the proposition satisfies

$$\Delta z(s,a) \coloneqq z_{\text{new}}(s,a) - z_{\text{old}}(s,a) = \beta \pi_{\text{old}}(a|s) A^{\text{info}}_{\pi_{\text{old}}}(s,a). \tag{82}$$

Taking the ℓ_1 norm and using Eq. 78,

$$\begin{aligned}
\|\Delta z(s,\cdot)\|_1 &= \beta \sum_{a\in\mathcal{A}} \pi_{\text{old}}(a|s) \left|A^{\text{info}}_{\pi_{\text{old}}}(s,a)\right| \\
&\le 2\bar{\beta} C\left(1-\exp\{-\mathcal{H}_{\text{old}}(s)\}\right).
\end{aligned} \tag{83}$$

Consider the line segment

$$z_\xi(s,\cdot) = z_{\text{old}}(s,\cdot) + \xi \Delta z(s,\cdot), \qquad \xi \in [0,1]. \tag{84}$$

Let $J(\xi)$ denote the Jacobian of the softmax probability vector with respect to the logits, evaluated at $z_\xi(s,a)$. Its entries are

$$J_{ij}(\xi) = \frac{\partial \pi_z(i|s)}{\partial z(s,j)}\bigg|_{z=z_\xi} = \pi_{z_\xi}(i|s)\left(\mathbf{1}\{i=j\} - \pi_{z_\xi}(j|s)\right). \tag{85}$$

For every column j,

$$\sum_i |J_{ij}(\xi)| = 2\pi_{z_\xi}(j|s)\left(1-\pi_{z_\xi}(j|s)\right) \le \frac{1}{2}. \tag{86}$$

Therefore,

$$\|J(\xi)\|_{1\to 1} \le \frac{1}{2}. \tag{87}$$

By the fundamental theorem of calculus,

$$\begin{aligned}
&\|\pi_{\text{new}}(\cdot|s) - \pi_{\text{old}}(\cdot|s)\|_1 \\
&= \left\| \int_0^1 J(\xi)\Delta z(s,\cdot)\, d\xi \right\|_1 \\
&\le \int_0^1 \|J(\xi)\|_{1\to 1} \|\Delta z(s,\cdot)\|_1 \, d\xi \\
&\le \frac{1}{2}\|\Delta z(s,\cdot)\|_1.
\end{aligned} \tag{88}$$

Thus,

$$\begin{aligned} &D_{\mathrm{TV}}\left(\pi_{\mathrm{new}}(\cdot|s), \pi_{\mathrm{old}}(\cdot|s)\right) \\ &\quad = \frac{1}{2}\left\|\pi_{\mathrm{new}}(\cdot|s) - \pi_{\mathrm{old}}(\cdot|s)\right\|_1 \\ &\quad \leq \frac{1}{4}\|\Delta z(s,\cdot)\|_1 \\ &\quad \leq \frac{\bar{\beta}C}{2}\left(1 - \exp\{-\mathcal{H}_{\mathrm{old}}(s)\}\right). \end{aligned} \tag{89}$$

Finally, since $1 - e^{-x} \leq x$ for $x \geq 0$ and $\mathcal{H}_{\mathrm{old}}(s) = \mathcal{H}_{\max}\rho(s)$,

$$D_{\mathrm{TV}}\left(\pi_{\mathrm{new}}(\cdot|s), \pi_{\mathrm{old}}(\cdot|s)\right) \leq \frac{\bar{\beta}C\mathcal{H}_{\max}}{2}\rho(s). \tag{90}$$

Since s was arbitrary, the result holds for every state. □

By setting $\alpha = \bar{\beta}C\mathcal{H}_{\max}/2$, the proposition yields

$$D_{\mathrm{TV}}\left(\pi_{\mathrm{new}}(\cdot|s), \pi_{\mathrm{old}}(\cdot|s)\right) \leq \alpha\rho(s), \qquad \forall s, \tag{91}$$

which is the state-dependent divergence condition in Eq. 7.

To connect this bound to the coupling formulation used in the proof of Theorem A.6, we first define the common probability mass:

$$m_a \coloneqq \min\left\{\pi_{\mathrm{old}}(a|s), \pi_{\mathrm{new}}(a|s)\right\}, \qquad c \coloneqq \sum_{a\in\mathcal{A}} m_a. \tag{92}$$

Using

$$\min\{x, y\} = \frac{x + y - |x - y|}{2}, \tag{93}$$

we obtain

$$\begin{aligned} 1 - c &= \frac{1}{2}\sum_{a\in\mathcal{A}}\left|\pi_{\mathrm{new}}(a|s) - \pi_{\mathrm{old}}(a|s)\right| \\ &= D_{\mathrm{TV}}\left(\pi_{\mathrm{new}}(\cdot|s), \pi_{\mathrm{old}}(\cdot|s)\right). \end{aligned} \tag{94}$$

The quantity m_a represents the probability mass shared by the two policies at action a. We now construct a coupling of $\pi_{\mathrm{old}}(\cdot|s)$ and $\pi_{\mathrm{new}}(\cdot|s)$ by assigning this common mass to identical action pairs. Specifically, for each action a, we assign

$$P\left(a_{\mathrm{old}} = a, a_{\mathrm{new}} = a \mid s\right) = m_a. \tag{95}$$

Since $\sum_{a\in\mathcal{A}} m_a = c$, the total probability assigned to identical action pairs is c.

After removing this common mass, the remaining probability masses of the two policies are

$$\pi_{\mathrm{old}}(a|s) - m_a \qquad \text{and} \qquad \pi_{\mathrm{new}}(a|s) - m_a,$$

respectively. Moreover,

$$\begin{aligned} \sum_{a\in\mathcal{A}}\left(\pi_{\mathrm{old}}(a|s) - m_a\right) &= 1 - c, \\ \sum_{a\in\mathcal{A}}\left(\pi_{\mathrm{new}}(a|s) - m_a\right) &= 1 - c. \end{aligned} \tag{96}$$

Thus, when $c < 1$, the normalized remaining distributions are

$$\frac{\pi_{\mathrm{old}}(a|s) - m_a}{1 - c}, \qquad \frac{\pi_{\mathrm{new}}(a|s) - m_a}{1 - c}. \tag{97}$$

The remaining probability mass $1 - c$ is assigned according to these two distributions.

By the definition of m_a, for every action a, at least one of

$$\pi_{\text{old}}(a|s) - m_a \qquad \text{and} \qquad \pi_{\text{new}}(a|s) - m_a$$

is zero. Hence, the two remaining distributions have disjoint supports. The sampled actions therefore agree on the common mass and differ almost surely on the remaining mass. Consequently,

$$P\left(a_{\text{old}} \neq a_{\text{new}} \mid s\right) = 1 - c = D_{\text{TV}}\left(\pi_{\text{old}}(\cdot|s), \pi_{\text{new}}(\cdot|s)\right) \leq \alpha\rho(s). \tag{98}$$

Hence, $(\pi_{\text{old}}, \pi_{\text{new}})$ forms an $\alpha\rho$-coupled policy pair in the sense used in the proof of Theorem A.6.

Proposition A.8. *Suppose that the terminal reward is bounded by $|R_T| \leq R_{\max} < \infty$. Under the update from π_k to π_{k+1} in Proposition 4.3, with $\alpha = \bar{\beta}C\mathcal{H}_{\max}/2$, the change in information-time return caused by recomputing the information density satisfies*

$$\left|\eta^{\text{info}}_{\rho_{k+1}}(\pi_{k+1}) - \eta^{\text{info}}_{\rho_k}(\pi_{k+1})\right| \leq \frac{8R_{\max}}{e}\alpha. \tag{99}$$

Proof. We first bound the relative change in the state-wise information density, then propagate this bound to the accumulated information time along a trajectory, and finally control the resulting change in the terminal information-time discount.

Step 1: Control of the local logit movement. Fix an arbitrary state s and define

$$\Delta z_k(s, \cdot) \coloneqq z_{k+1}(s, \cdot) - z_k(s, \cdot).$$

From Proposition 4.3,

$$\begin{aligned}
\|\Delta z_k(s, \cdot)\|_1 &\leq 2\bar{\beta}C\left(1 - e^{-\mathcal{H}_k(s)}\right) \\
&\leq 2\bar{\beta}C\mathcal{H}_k(s) \\
&= 2\bar{\beta}C\mathcal{H}_{\max}\rho_k(s) \\
&= 4\alpha\rho_k(s),
\end{aligned} \tag{100}$$

where the second inequality follows from $1 - e^{-x} \leq x$ for $x \geq 0$. Since $\|v\|_\infty \leq \|v\|_1$, we have

$$\|\Delta z_k(s, \cdot)\|_\infty \leq 4\alpha\rho_k(s). \tag{101}$$

Step 2: Multiplicative stability of the information density. For a generic softmax distribution $\pi_z(\cdot|s)$, we write the corresponding predictive entropy as

$$\mathcal{H}_z(s) = -\sum_a \pi_z(a|s)\log\pi_z(a|s).$$

Its derivative with respect to the logit z_a is

$$\frac{\partial\mathcal{H}_z(s)}{\partial z(s, a)} = -\pi_z(a|s)\left(\log\pi_z(a|s) + \mathcal{H}_z(s)\right). \tag{102}$$

Hence

$$\begin{aligned}
\|\nabla_z\mathcal{H}_z(s)\|_1 &= \sum_a \pi_z(a|s)\left|\log\pi_z(a|s) + \mathcal{H}_z(s)\right| \\
&\leq \sum_a \pi_z(a|s)\left(-\log\pi_z(a|s) + \mathcal{H}_z(s)\right) \\
&= 2\mathcal{H}_z(s).
\end{aligned} \tag{103}$$

Consider the interpolation

$$z_\xi(s, \cdot) = z_k(s, \cdot) + \xi\Delta z_k(s, \cdot), \qquad \xi \in [0, 1],$$

By the chain rule, Hölder's inequality, and Eq. 103,

$$\begin{aligned}
\left|\frac{d}{d\xi}\mathcal{H}_{z_\xi}(s)\right| &= \left|\nabla_z\mathcal{H}_{z_\xi}(s)^\top\Delta z_k(s, \cdot)\right| \\
&\leq \|\nabla_z\mathcal{H}_{z_\xi}(s)\|_1\|\Delta z_k(s, \cdot)\|_\infty \\
&\leq 2\mathcal{H}_{z_\xi}(s)\|\Delta z_k(s, \cdot)\|_\infty.
\end{aligned} \tag{104}$$

For positive entropy,

$$\left|\frac{d}{d\xi}\log \mathcal{H}_{z_\xi}(s)\right| \le 2\|\Delta z_k(s,\cdot)\|_\infty. \tag{105}$$

Integrating over $\xi \in [0,1]$ yields

$$\begin{aligned}|\log \mathcal{H}_{z_1}(s) - \log \mathcal{H}_{z_0}(s)| &= \left|\int_0^1 \frac{d}{d\xi}\log \mathcal{H}_{z_\xi}(s)\, d\xi\right| \\ &\le \int_0^1 \left|\frac{d}{d\xi}\log \mathcal{H}_{z_\xi}(s)\right| d\xi \\ &\le 2\|\Delta z_k(s,\cdot)\|_\infty. \end{aligned} \tag{106}$$

By construction,

$$z_0(s,\cdot) = z_k(s,\cdot), \qquad z_1(s,\cdot) = z_{k+1}(s,\cdot),$$

and hence

$$\mathcal{H}_{z_0}(s) = \mathcal{H}_k(s), \qquad \mathcal{H}_{z_1}(s) = \mathcal{H}_{k+1}(s).$$

Therefore,

$$\left|\log \frac{\mathcal{H}_{k+1}(s)}{\mathcal{H}_k(s)}\right| \le 2\|\Delta z_k(s,\cdot)\|_\infty. \tag{107}$$

Finally, applying Eq. 101 gives

$$\left|\log \frac{\mathcal{H}_{k+1}(s)}{\mathcal{H}_k(s)}\right| \le 8\alpha\rho_k(s). \tag{108}$$

Equivalently,

$$e^{-8\alpha\rho_k(s)}\mathcal{H}_k(s) \le \mathcal{H}_{k+1}(s) \le e^{8\alpha\rho_k(s)}\mathcal{H}_k(s). \tag{109}$$

Dividing both sides by the normalization scale $\mathcal{H}_{\max}$ gives

$$e^{-8\alpha\rho_k(s)}\rho_k(s) \le \rho_{k+1}(s) \le e^{8\alpha\rho_k(s)}\rho_k(s). \tag{110}$$

Step 3: Stability of the accumulated information time. Fix a finite trajectory

$$\tau = (s_0, a_0, \dots, s_T)$$

and define

$$I_k(\tau) \coloneqq \sum_{t=0}^{T-1}\rho_k(s_t), \qquad I_{k+1}(\tau) \coloneqq \sum_{t=0}^{T-1}\rho_{k+1}(s_t). \tag{111}$$

Also define

$$\rho_{k,\max}(\tau) \coloneqq \max_{0\le t<T}\rho_k(s_t). \tag{112}$$

By Eq. 110, for every t,

$$e^{-8\alpha\rho_k(s_t)}\rho_k(s_t) \le \rho_{k+1}(s_t) \le e^{8\alpha\rho_k(s_t)}\rho_k(s_t). \tag{113}$$

Since $\rho_k(s_t) \le \rho_{k,\max}(\tau)$,

$$e^{-8\alpha\rho_{k,\max}(\tau)}\rho_k(s_t) \le \rho_{k+1}(s_t) \le e^{8\alpha\rho_{k,\max}(\tau)}\rho_k(s_t). \tag{114}$$

Summing over $t = 0, \dots, T-1$ gives

$$e^{-8\alpha\rho_{k,\max}(\tau)}I_k(\tau) \le I_{k+1}(\tau) \le e^{8\alpha\rho_{k,\max}(\tau)}I_k(\tau). \tag{115}$$

If $I_k(\tau) = 0$, then nonnegativity of the information density implies $\rho_k(s_t) = 0$ for every t. Eq. 110 then implies $\rho_{k+1}(s_t) = 0$ for every t, and therefore $I_{k+1}(\tau) = 0$. In this case the two trajectory discount factors are identical.

Now suppose that $I_k(\tau) > 0$. Eq. 115 implies that $I_{k+1}(\tau) > 0$, and taking logarithms yields

$$|\log I_{k+1}(\tau) - \log I_k(\tau)| \le 8\alpha\rho_{k,\max}(\tau). \tag{116}$$

Step 4: Log-Lipschitz continuity of the information-time discount. We use the following elementary inequality. For any $\gamma \in (0,1)$ and any $x, y > 0$,

$$|\gamma^x - \gamma^y| \le \frac{1}{e} |\log x - \log y|. \tag{117}$$

To prove this inequality, consider the function

$$f(x) \coloneqq \gamma^{e^x} = \exp\{(\log \gamma) e^x\}, \qquad x \in \mathbb{R}. \tag{118}$$

Its derivative is

$$f'(x) = (\log \gamma) e^x \exp\{(\log \gamma) e^x\}. \tag{119}$$

Since $\log \gamma < 0$, letting $q = (-\log \gamma) e^x > 0$, we can get

$$|f'(x)| = q e^{-q} \le \frac{1}{e}, \tag{120}$$

where the last inequality follows because qe^{-q} attains its maximum at $q = 1$. By applying the mean value theorem, we finally get Eq. 117.

For the trajectory τ,

$$\Gamma^{\text{info}}_{\rho_k}(0,T) = \prod_{t=0}^{T-1} \gamma^{\rho_k(s_t)} = \gamma^{I_k(\tau)}, \tag{121}$$

$$\Gamma^{\text{info}}_{\rho_{k+1}}(0,T) = \prod_{t=0}^{T-1} \gamma^{\rho_{k+1}(s_t)} = \gamma^{I_{k+1}(\tau)}. \tag{122}$$

Therefore, when $I_k(\tau) > 0$, Eqs. 117 and 116 give

$$\left| \Gamma^{\text{info}}_{\rho_{k+1}}(0,T) - \Gamma^{\text{info}}_{\rho_k}(0,T) \right| \le \frac{1}{e} \left| \log I_{k+1}(\tau) - \log I_k(\tau) \right| \le \frac{8\alpha}{e} \rho_{k,\max}(\tau). \tag{123}$$

As shown above, the same inequality also holds when $I_k(\tau) = 0$, because in that case the left-hand side is zero. Hence Eq. 123 holds for every trajectory.

Step 5: From pathwise discount stability to return stability. Under the terminal-reward setting,

$$\eta^{\text{info}}_{\rho}(\pi_{k+1}) = \mathbb{E}_{\tau \sim \pi_{k+1}} \left[\Gamma^{\text{info}}_{\rho}(0,T) R_T \right]. \tag{124}$$

Thus

$$\begin{aligned}
&\left| \eta^{\text{info}}_{\rho_{k+1}}(\pi_{k+1}) - \eta^{\text{info}}_{\rho_k}(\pi_{k+1}) \right| \\
&= \left| \mathbb{E}_{\tau \sim \pi_{k+1}} \left[R_T \left(\Gamma^{\text{info}}_{\rho_{k+1}}(0,T) - \Gamma^{\text{info}}_{\rho_k}(0,T) \right) \right] \right| \\
&\le R_{\max} \mathbb{E}_{\tau \sim \pi_{k+1}} \left[\left| \Gamma^{\text{info}}_{\rho_{k+1}}(0,T) - \Gamma^{\text{info}}_{\rho_k}(0,T) \right| \right] \\
&\le \frac{8 R_{\max} \alpha}{e} \mathbb{E}_{\tau \sim \pi_{k+1}} \left[\rho_{k,\max}(\tau) \right].
\end{aligned} \tag{125}$$

Finally, since $\rho_k(s) \in [0,1]$,

$$\rho_{k,\max}(\tau) \le 1$$

for every trajectory. Therefore,

$$\left| \eta^{\text{info}}_{\rho_{k+1}}(\pi_{k+1}) - \eta^{\text{info}}_{\rho_k}(\pi_{k+1}) \right| \le \frac{8 R_{\max}}{e} \alpha. \tag{126}$$

This proves the proposition. □

The preceding proposition characterizes the variation in the information clock induced by policy updates under a fixed normalization scheme. In Section 5.1, we further empirically examine different normalization schemes. Although changing the normalization scale may introduce additional variation in the resulting information density, more adaptive normalization can also better capture local information dynamics. Among the three normalization schemes considered, batch-level normalization provides a favorable balance between local adaptivity and the stability of the induced information-density scale.

Proposition A.9. *Let π_x and π_y, with $x < y$, be two policy iterates from the same sequence of updates, and let ρ_r be any fixed information-density snapshot used as a common reference clock. Suppose that $|R_T| \le R_{\max} < \infty$. Define the cumulative certified gain between the two iterates as*

$$\mathcal{G}_{x,y} \coloneqq \sum_{t=x}^{y-1} \left[L^{\mathrm{info}}_{\pi_t}(\pi_{t+1}) - \eta^{\mathrm{info}}_{\rho_t}(\pi_t) - \frac{4C_t\alpha_t^2}{(1-\gamma)^2} - \frac{8R_{\max}}{e}\alpha_t \right], \tag{127}$$

where C_t and α_t denote the corresponding quantities for the update $\pi_t \to \pi_{t+1}$, and all information-time quantities in the t-th summand are evaluated under ρ_t. For a trajectory $\tau = (s_0, a_0, \dots, s_T)$, define $\bar{\rho}_j(\tau) \coloneqq \frac{1}{T}\sum_{u=0}^{T-1} \rho_j(s_u)$, and $\varepsilon^{(r)}_{x,y} \coloneqq \frac{1}{2}\sum_{j\in\{x,y\}} \mathbb{E}_{\tau\sim\pi_j}\left[\left|\log \frac{\bar{\rho}_r(\tau)}{\bar{\rho}_j(\tau)}\right|\right]$. Then

$$\eta^{\mathrm{info}}_{\rho_r}(\pi_y) - \eta^{\mathrm{info}}_{\rho_r}(\pi_x) \ge \mathcal{G}_{x,y} - \frac{2R_{\max}}{e}\varepsilon^{(r)}_{x,y}. \tag{128}$$

Proof. For each $t \in \{x, \dots, y-1\}$, the information density ρ_t is fixed during the update from π_t to π_{t+1}. Hence, by Theorem 4.2,

$$\eta^{\mathrm{info}}_{\rho_t}(\pi_{t+1}) - \eta^{\mathrm{info}}_{\rho_t}(\pi_t) \ge L^{\mathrm{info}}_{\pi_t}(\pi_{t+1}) - \eta^{\mathrm{info}}_{\rho_t}(\pi_t) - \frac{4C_t\alpha_t^2}{(1-\gamma)^2}. \tag{129}$$

Moreover, Proposition A.8 gives

$$\eta^{\mathrm{info}}_{\rho_{t+1}}(\pi_{t+1}) - \eta^{\mathrm{info}}_{\rho_t}(\pi_{t+1}) \ge -\frac{8R_{\max}}{e}\alpha_t. \tag{130}$$

Adding the two inequalities yields

$$\eta^{\mathrm{info}}_{\rho_{t+1}}(\pi_{t+1}) - \eta^{\mathrm{info}}_{\rho_t}(\pi_t) \ge L^{\mathrm{info}}_{\pi_t}(\pi_{t+1}) - \eta^{\mathrm{info}}_{\rho_t}(\pi_t) - \frac{4C_t\alpha_t^2}{(1-\gamma)^2} - \frac{8R_{\max}}{e}\alpha_t. \tag{131}$$

Summing over $t = x, \dots, y-1$ and telescoping therefore gives

$$\eta^{\mathrm{info}}_{\rho_y}(\pi_y) - \eta^{\mathrm{info}}_{\rho_x}(\pi_x) \ge \mathcal{G}_{x,y}. \tag{132}$$

It remains to evaluate the two endpoint policies under the common reference clock ρ_r. For $j \in \{x, y\}$, let

$$I_j(\tau) = T\bar{\rho}_j(\tau), \qquad I_r(\tau) = T\bar{\rho}_r(\tau).$$

Under the terminal-reward setting and the log-Lipschitz bound in Eq. 117,

$$\left|\eta^{\mathrm{info}}_{\rho_r}(\pi_j) - \eta^{\mathrm{info}}_{\rho_j}(\pi_j)\right| \le \frac{R_{\max}}{e}\mathbb{E}_{\tau\sim\pi_j}\left[\left|\log\frac{I_r(\tau)}{I_j(\tau)}\right|\right] = \frac{R_{\max}}{e}\mathbb{E}_{\tau\sim\pi_j}\left[\left|\log\frac{\bar{\rho}_r(\tau)}{\bar{\rho}_j(\tau)}\right|\right], \tag{133}$$

where the trajectory length T cancels in the ratio.

Finally, define

$$\Delta^{(r)}_j \coloneqq \eta^{\mathrm{info}}_{\rho_r}(\pi_j) - \eta^{\mathrm{info}}_{\rho_j}(\pi_j).$$

Then

$$\begin{aligned}
\eta^{\mathrm{info}}_{\rho_r}(\pi_y) - \eta^{\mathrm{info}}_{\rho_r}(\pi_x) &= \eta^{\mathrm{info}}_{\rho_y}(\pi_y) - \eta^{\mathrm{info}}_{\rho_x}(\pi_x) + \Delta^{(r)}_y - \Delta^{(r)}_x \\
&\ge \mathcal{G}_{x,y} - |\Delta^{(r)}_y| - |\Delta^{(r)}_x| \\
&\ge \mathcal{G}_{x,y} - \frac{2R_{\max}}{e}\varepsilon^{(r)}_{x,y},
\end{aligned} \tag{134}$$

where the last inequality follows from Eq. 133 and the definition of $\varepsilon^{(r)}_{x,y}$. This proves the proposition. □

Proposition A.9 shows that recomputing the information clock across training does not by itself invalidate comparisons between policy iterates. If the cumulative certified gain exceeds the discrepancy introduced by re-anchoring the endpoint policies to a common reference clock, the later policy still achieves a higher information-time return under that clock. Consequently, smaller clock drift makes the comparison more robust, although the information clock need not remain fixed throughout training. We therefore empirically track the evolution of normalized entropy throughout training to characterize how the information clock changes over the full optimization trajectory.

B METHOD OVERVIEW

Figure 6 compares the temporal weighting and clipping mechanisms of standard PPO and InfoPPO.

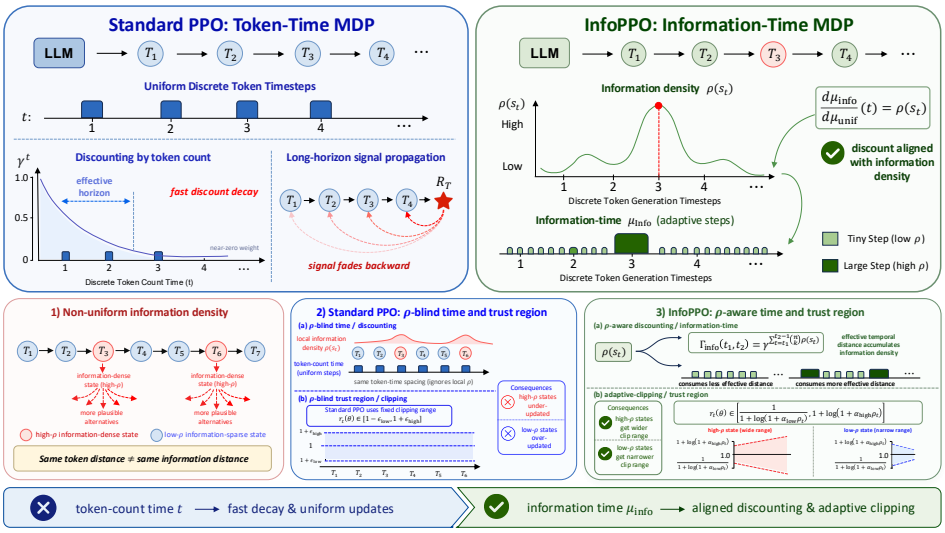


Figure 6: Comparison of standard PPO and InfoPPO. PPO parameterizes temporal progression by uniform token steps and applies fixed clipping bounds across token positions. In long-horizon reasoning, this creates a tension between long-range credit propagation and effective-horizon contraction. In contrast, InfoPPO preserves token-level transitions while measuring temporal distance through accumulated state-dependent information. This reparameterization retains effective-horizon contraction while moderating long-range attenuation. The same information density also adapts the clipping range across states, improving sample efficiency and accelerating policy learning.

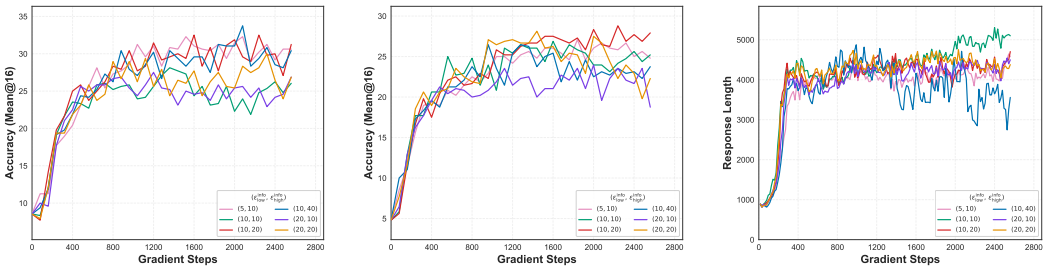
C IMPLEMENTATION DETAILS

Models and Training Settings. We train Qwen3-4B Base Model, Qwen3-8B Base Model, and Qwen3-14B Base Model on DAPO-Math-17K. For both PPO and InfoPPO, the policy learning rate is 1×10^{-6} with a 10-step warmup, while the critic learning rate is 2×10^{-6} without warmup. For the reproduction of DAPO and DAPO-FT, we adhere to the official default configurations. The maximum response length is 10k tokens for the 4B and 8B models and 12k tokens for the 14B model, consistently during training and evaluation.

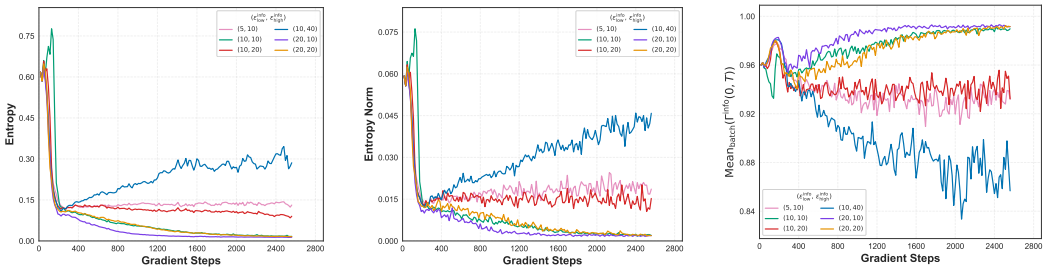
Unless otherwise specified in the sensitivity experiments, PPO uses $\gamma = 1$ and $\lambda = 1$, with clipping parameters $\epsilon_{\text{low}} = 0.2$ and $\epsilon_{\text{high}} = 0.28$. For InfoPPO, we use $\gamma = 0.999$ and $\lambda = 0.99$. The adaptive clipping bounds follow Eq. 15, with $\epsilon_{\text{low}}^{\text{info}} = 10$ and $\epsilon_{\text{high}}^{\text{info}} = 20$, as supported by the ablation study below.

Evaluation. We evaluate the models on five challenging competition-style mathematical reasoning benchmarks: AMC23 (Li et al., 2024), AIME24 (Zhang & Math-AI, 2024), AIME25 (Zhang & Math-AI, 2025), AIME26 (Zhang & Math-AI, 2026), and BeyondAIME (ByteDance-Seed, 2025). All evaluations are performed in a zero-shot setting. For each prompt, we independently sample 16 responses with decoding temperature $T = 1.0$ and report the average accuracy, denoted as Mean@16.

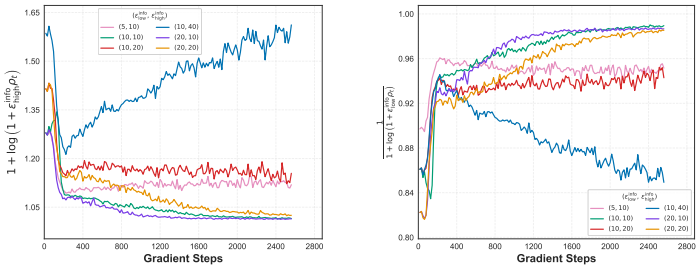
Ablation of Adaptive Clipping Parameters. Figure 7 compares different combinations of $\epsilon_{\text{low}}^{\text{info}}$ and $\epsilon_{\text{high}}^{\text{info}}$ on the Qwen3-4B Base Model. Increasing $\epsilon_{\text{high}}^{\text{info}}$ widens the upper clipping range and is generally associated with higher predictive entropy, whereas increasing $\epsilon_{\text{low}}^{\text{info}}$ lowers the lower clipping bound and tends to reduce predictive entropy over training. Configurations such as $(5, 10)$ and $(10, 20)$, where $\epsilon_{\text{high}}^{\text{info}}$ is approximately twice $\epsilon_{\text{low}}^{\text{info}}$, exhibit comparatively stable entropy dynamics



(a) AIME24 Accuracy. (b) AIME25 Accuracy. (c) Response Length.



(d) Raw Predictive Entropy. (e) Normalized Entropy. (f) Mean $\Gamma^{\text{info}}(0,T)$



(g) Mean Upper Clipping Bound. (h) Mean Lower Clipping Bound.

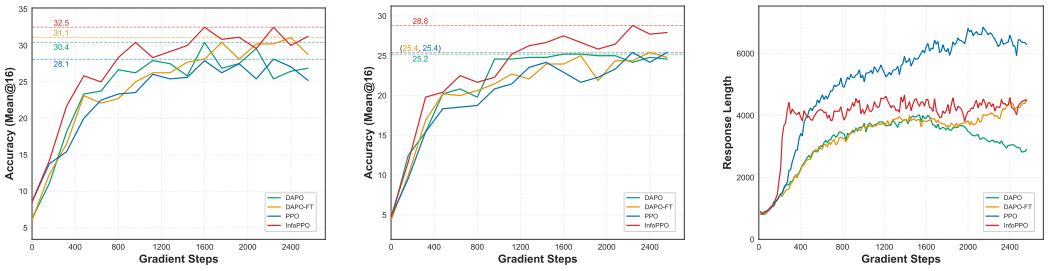
Figure 7: Ablation of $\epsilon_{\text{low}}^{\text{info}}$ and $\epsilon_{\text{high}}^{\text{info}}$ on the Qwen3-4B Base Model.

while maintaining competitive reasoning performance. Such stability is consistent with the moderate variation of the information clock considered in our theoretical analysis, and may also help preserve a favorable balance between exploration and exploitation during RL optimization. Future work could further explore more direct entropy-control mechanisms to provide tighter control over predictive entropy, thereby further stabilizing the information clock and more explicitly regulating the exploration–exploitation trade-off. Response length remains broadly stable across the tested clipping configurations, with substantially smaller variation than that observed in the γ ablation in Figure 4. This suggests that γ provides a more direct and effective control knob for modulating response length. Considering the overall reasoning performance and training dynamics, we use $\epsilon_{\text{low}}^{\text{info}} = 10$ and $\epsilon_{\text{high}}^{\text{info}} = 20$ in the main experiments.

D Training Dynamics Across Qwen3 Model Scales

Figure 8 reports AIME24 and AIME25 accuracy and response length during training on Qwen3-4B, 8B, and 14B base models. Across model scales, InfoPPO steadily improves accuracy while keeping response lengths within a relatively stable range. In contrast, the baselines exhibit larger variations in response length over the course of training. This behavior is consistent with the information-time formulation. Information-time discounting introduces a non-trivial effective-horizon constraint while avoiding excessive accumulation of token-wise decay over long sequences, and information-dependent policy updates adapt the optimization scale to local information density. Together, these

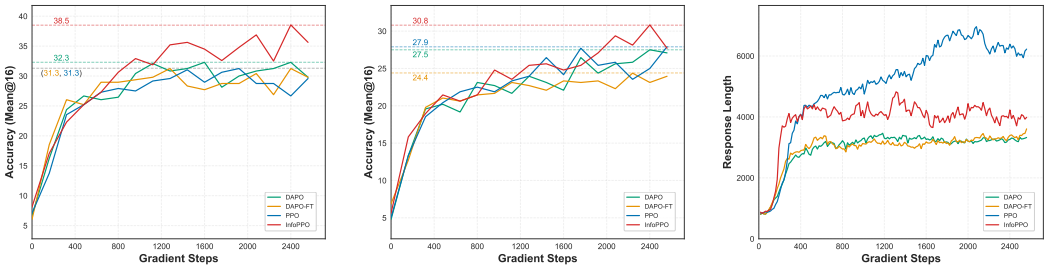
mechanisms enable effective policy improvement within a controlled information horizon without relying on continued trajectory expansion.



(a) AIME24 Accuracy trained from Qwen3-4B Base.

(b) AIME25 Accuracy trained from Qwen3-4B Base.

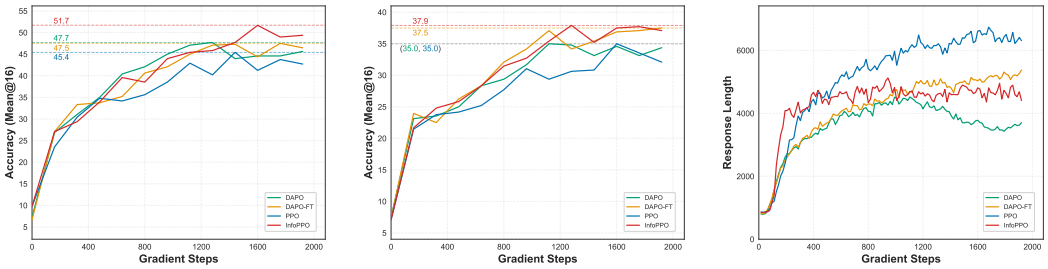
(c) Response lengths trained from Qwen3-4B Base.



(d) AIME24 Accuracy trained from Qwen3-8B Base.

(e) AIME25 Accuracy trained from Qwen3-8B Base.

(f) Response lengths trained from Qwen3-8B Base.



(g) AIME24 Accuracy trained from Qwen3-14B Base.

(h) AIME25 Accuracy trained from Qwen3-14B Base.

(i) Response lengths trained from Qwen3-14B base.

Figure 8: Training Dynamics of Accuracy and Response Length Across Model Scales. InfoPPO exhibits stronger accuracy gains over training while maintaining stable response lengths, indicating more efficient policy improvement without relying on continued length growth.

E Additional Results on Llama Models

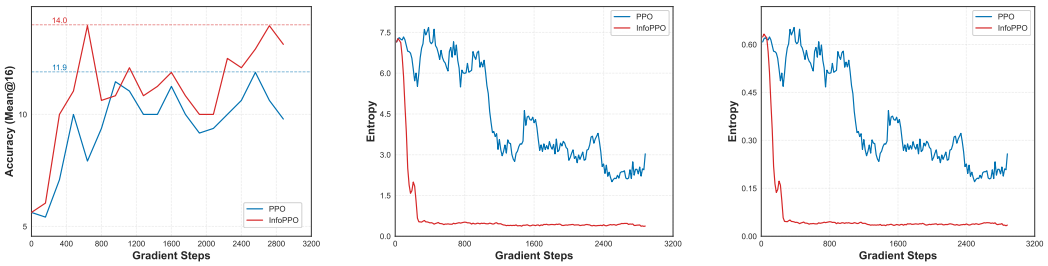
To further examine whether the benefits of InfoPPO extend beyond the Qwen3 model family, we additionally evaluate PPO and InfoPPO on the Llama-3.2-3B Instruct Model under the same overall experimental protocol as the main experiments.

Table 2 summarizes the final evaluation results. InfoPPO achieves clear improvements over PPO on AMC23 and AIME24, while also obtaining higher accuracy on BeyondAIME. On AIME25 and AIME26, both methods achieve very low absolute accuracy, making the comparison less informative on these particularly challenging benchmarks.

Figure 9 further shows the training dynamics on Llama-3.2-3B Instruct. On AIME24, InfoPPO reaches a higher final accuracy than PPO while exhibiting stable predictive-entropy and adaptive clipping dynamics after the initial stage of training. Notably, both methods show a substantial

Table 2: Performance comparison of InfoPPO with PPO on Llama-3.2-3B Instruct model. For each prompt, we independently sample 16 responses and report the average accuracy, denoted as Mean@16.

Benchmark	PPO		InfoPPO	
	Score	Len.	Score	Len.
Llama-3.2-3B Instruct Model				
AMC23	39.2	648.1	42.3	511.9
AIME24	11.9	906.0	14.0	516.4
AIME25	0.0	665.2	0.8	374.7
AIME26	0.2	542.8	0.6	365.7
BeyondAIME	0.6	593.3	1.9	396.1
Average	10.4	671.1	11.9	433.0



(a) AIME24 Accuracy. (b) Raw Predictive Entropy. (c) Normalized Entropy.

(d) Response Length. (e) Mean Upper Clipping Bound. (f) Mean Lower Clipping Bound.

Figure 9: Training dynamics on the Llama-3.2-3B Instruct Model.

decrease in response length during training. This response-length contraction may therefore be partly associated with the training dynamics of the Llama-3.2-3B Instruct Model itself, since a similar and sustained decline is also observed under PPO with $\gamma = 1$. Taken together, these results provide additional evidence that the optimization behavior of InfoPPO remains well behaved on a different model family.

F ENTROPY DYNAMICS

<think>\n Okay , ·so ·I ·need ·to ·solve ·this ·system ·of ·logarithmic ·equations ·and ·then ·find ·the ·absolute ·value ·of ·the ·log ·base · 2 ·of ·x ^ 4 ·y ^ 3 ·z ^ 2 . ·The ·answer ·should ·be ·in ·the ·form ·m /n ·where ·m ·and ·n ·are ·cop r ime , ·and ·then ·find ·m ·+ ·n . ·Let ·me ·start ·by ·writing ·down ·the ·equations ·again ·to ·make ·sure ·I ·have ·them ·right .\n\n First ·equation : ·log ₂ (x /(yz)) ·= · 1 / 2 \n\n Second ·equation : ·log ₂ (y /(xz)) ·= · 1 / 3 \n\n Third ·equation : ·log ₂ (z /(xy)) ·= · 1 / 4 \n\n Hmm . ·So ·each ·equation ·is ·a ·logarithm ·of ·a ·ratio ·of ·variables ·equal ·to ·some ·fraction . ·Maybe ·I ·can ·convert ·these ·logarithmic ·equations ·into ·exponential ·form ·to ·make ·them ·easier ·to ·handle . ·Let ·me ·recall ·that ·log _b (a) ·= ·c ·is ·equivalent ·to ·b ^ c ·= ·a . ·So ·applying ·that ·here .\n\n Starting ·with ·the ·first ·equation : ·log ₂ (x /(yz)) ·= · 1 / 2 . ·That ·means · 2 ^(1 / 2) ·= ·x /(yz). ·So ·sqrt (2) ·= ·x /(yz). ·Similarly ·for ·the ·other ·equations .\n\n Second ·equation : ·log ₂ (y /(xz)) ·= · 1 / 3 , ·so · 2 ^(1 / 3) ·= ·y /(xz). ·Third ·equation : ·log ₂ (z /(xy)) ·= · 1 / 4 , ·so · 2 ^(1 / 4) ·= ·z /(xy).\n\n So ·now ·I ·have ·three ·equations :\n\n 1) ·x ·= ·y z ·* · 2 ^(1 / 2)\n\n 2) ·y ·= ·x z ·* · 2 ^(1 / 3)\n\n 3) ·z ·= ·xy ·* · 2 ^(1 / 4)\n\n Hmm . ·So ·each ·variable ·is ·expressed ·in ·terms ·of ·the ·other ·two . ·Maybe ·I ·can ·substitute ·these ·equations ·into ·each ·other ·to ·solve ·for ·x , ·y , ·z ?\n\n Alternatively, ·maybe ·I ·can ·take ·the ·product ·of ·all ·three ·equations? ·Let ·me ·try ·that .\n\n Multiply ·the ·left ·sides : ·x ·* ·y ·* ·z ·= ·(yz ·* · 2 ^(1 / 2)) ·* ·(xz ·* · 2 ^(1 / 3)) ·* ·(xy ·* · 2 ^(1 / 4))\n\n Wait , ·let ·me ·check ·that . ·If ·I ·take ·the ·left ·sides ·of ·the ·equations :\n\n Left ·side ·of ·equation · 1 : ·x \n\n Left ·side ·of ·equation · 2 : ·y \n\n Left ·side ·of ·equation · 3 : ·z \n\n So ·multiplying ·them ·together : ·x ·* ·y ·* ·z .\n\n On ·the ·right ·side ·of ·each ·equation :\n\n Equ ation · 1 : ·y z ·* · 2 ^(1 / 2)\n\n Equ ation · 2 : ·x z ·* · 2 ^(1 / 3)\n\n Equ ation · 3 : ·xy ·* · 2 ^(1 / 4)\n\n So ·multiplying ·them ·together : ·(yz ·* · 2 ^(1 / 2)) ·* ·(xz ·* · 2 ^(1 / 3)) ·* ·(xy ·* · 2 ^(1 / 4)) .\n\n Let ·me ·compute ·that . ·First , ·multiply ·the ·variables :\n\n yz ·* ·x z ·* ·xy ·= ·x ^ 2 ·y ^ 2 ·z ^ 2 . ·Because ·y z ·* ·x z ·is ·x ·z ^ 2 ·y , ·then ·times ·xy ·is ·x ^ 2 ·y ^ 2 ·z ^ 2 . ·Wait , ·let ·me ·check :\n\n Original ·terms : ·y z ·* ·x z ·* ·xy . ·Let ·me ·write ·all ·variables :\n\n For ·x : ·in ·the ·first ·term ·there 's ·no ·x , ·second ·term ·has ·x , ·third ·term ·has ·x . ·So ·total ·x 's : · 1 ·+ · 1 ·= · 2 .\n\n For ·y : ·first ·term ·has ·y , ·second ·term ·no ·y , ·third ·term ·has ·y . ·So ·total ·y 's : · 1 ·+ · 1 ·= · 2 .\n\n For ·z : ·first ·term ·has ·z , ·second ·term ·has ·z , ·third ·term ·no ·z . ·So ·total ·z 's : · 1 ·+ · 1 ·= · 2 . ·So ·yes , ·x ^ 2 ·y ^ 2 ·z ^ 2 .\n\n Then ·the ·ex ponents ·of · 2 : · 2 ^(1 / 2) ·* · 2 ^(1 / 3) ·* · 2 ^(1 / 4) ·= · 2 ^(1 / 2 ·+ · 1 / 3 ·+ · 1 / 4). ·Let ·me ·compute ·that ·exponent .\n\n First , ·find ·common ·denominator ·for · 2 , · 3 , · 4 . ·Which ·is · 1 2 . ·So · 1 / 2 ·is · 6 / 1 2 , · 1 / 3 ·is · 4 / 1 2 , · 1 / 4 ·is · 3 / 1 2 . ·Adding ·them : · 6 ·+ · 4 ·+ · 3 ·= · 1 3 . ·So ·exponent ·is · 1 3 / 1 2 . ·Therefore, · 2 ^(1 3 / 1 2).\n\n Therefore , ·the ·product ·of ·the ·right ·sides ·is ·x ² ·y ² ·z ² ·* · 2 ^(1 3 / 1 2).\n\n But ·the ·product ·of ·the ·left ·sides ·is ·x ·y ·z . ·Therefore, ·we ·have :\n\n x ·y ·z ·= ·x ² ·y ² ·z ² ·* · 2 ^(1 3 / 1 2)\n\n Div ide

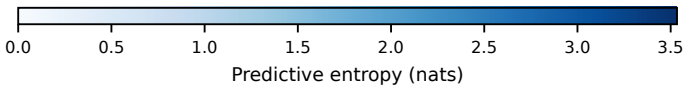


Figure 10: State-wise predictive entropy along a reasoning trajectory (Part 1/6).

·both ·sides ·by ·x ·y ·z ·(since ·x , ·y , ·z ·are ·positive , ·we ·can ·do ·that):\n\n 1 ·= ·x ·y ·z ·* · 2 ^(1 3 / 1 2)\n\n Therefore , ·x ·y ·z ·= · 2 ^ (- 1 3 / 1 2). ·Wait , ·that ·seems ·a ·bit ·complicated . ·Let ·me ·check ·my ·steps ·again .\n\n Wait , ·original ·equation : ·product ·of ·left ·sides ·equals ·product ·of ·right ·sides . ·So :\n\n xyz ·= ·(yz ·* · 2 ^(1 / 2)) ·* ·(xz ·* · 2 ^(1 / 3)) ·* ·(xy ·* · 2 ^(1 / 4)) .\n\n Which ·simpl ifies ·to ·xyz ·= ·(x ^ 2 ·y ^ 2 ·z ^ 2) ·* · 2 ^(1 / 2 ·+ · 1 / 3 ·+ · 1 / 4). ·Therefore , ·dividing ·both ·sides ·by ·xyz :\n\n 1 ·= ·xyz ·* · 2 ^(1 3 / 1 2). ·Therefore , ·xyz ·= · 2 ^ (- 1 3 / 1 2). ·Wait , ·but · 2 ^ (- 1 3 / 1 2) ·is · 1 ·over · 2 ^(1 3 / 1 2). ·So ·that 's ·correct .\n\n But ·maybe ·I ·can ·write ·that ·as · 2 ^ (- 1 3 / 1 2). ·Alternatively , ·maybe ·I ·can ·write ·that ·as · 2 ^ (- 1 3 / 1 2) ·= · 2 ^ (- 1 ·- · 1 / 1 2) ·= · 2 ^ (- 1) ·* · 2 ^ (- 1 / 1 2) ·= ·(1 / 2) ·* · 2 ^ (- 1 / 1 2). ·But ·not ·sure ·if ·that ·helps ·yet .\n\n Alternatively , ·maybe ·I ·can ·use ·this ·value ·of ·xyz ·to ·find ·the ·value ·of ·log ₂ (x ^ 4 ·y ^ 3 ·z ^ 2). ·Let ·me ·recall ·that ·the ·problem ·asks ·for ·| log ₂ (x ^ 4 ·y ^ 3 ·z ^ 2)| . ·Let ·me ·note ·that ·log ₂ (x ^ 4 ·y ^ 3 ·z ^ 2) ·= · 4 ·log ₂ x ·+ · 3 ·log ₂ y ·+ · 2 ·log ₂ z . ·Maybe ·I ·can ·find ·log ₂ x , ·log ₂ y , ·log ₂ z ·individually?\n\n Alternatively , ·since ·I ·have ·xyz , ·maybe ·I ·can ·find ·log ₂ (x yz) ·and ·relate ·it ·to ·the ·expression ·I ·need ?\n\n Wait , ·log ₂ (x ^ 4 ·y ^ 3 ·z ^ 2) ·is · 4 ·log ·x ·+ · 3 ·log ·y ·+ · 2 ·log ·z . ·If ·I ·can ·find ·log ·x , ·log ·y , ·log ·z , ·then ·I ·can ·compute ·that .\n\n Alternatively , ·maybe ·I ·can ·find ·log ₂ (x yz) ·first . ·Since ·xyz ·= · 2 ^ (- 1 3 / 1 2), ·then ·log ₂ (x yz) ·= ·- 1 3 / 1 2 . ·So ·log ₂ x ·+ ·log ₂ y ·+ ·log ₂ z ·= ·- 1 3 / 1 2 .\n\n But ·how ·does ·that ·help ·me ·with · 4 ·log ·x ·+ · 3 ·log ·y ·+ · 2 ·log ·z ?\n\n Hmm . ·Maybe ·I ·need ·more ·equations . ·Let ·me ·think . ·Let ·me ·denote ·a ·= ·log ₂ x , ·b ·= ·log ₂ y , ·c ·= ·log ₂ z . ·Then , ·the ·original ·equations ·can ·be ·rewritten ·in ·terms ·of ·a , ·b , ·c .\n\n Original ·equations :\n\n First ·equation : ·log ₂ (x /(yz)) ·= · 1 / 2 . ·Which ·is ·log ₂ x ·- ·log ₂ y ·- ·log ₂ z ·= · 1 / 2 . ·So ·a ·- ·b ·- ·c ·= · 1 / 2 .\n\n Second ·equation : ·log ₂ (y /(xz)) ·= · 1 / 3 . ·Which ·is ·log ₂ y ·- ·log ₂ x ·- ·log ₂ z ·= · 1 / 3 . ·So ·b ·- ·a ·- ·c ·= · 1 / 3 .\n\n Third ·equation : ·log ₂ (z /(xy)) ·= · 1 / 4 . ·Which ·is ·log ₂ z ·- ·log ₂ x ·- ·log ₂ y ·= · 1 / 4 . ·So ·c ·- ·a ·- ·b ·= · 1 / 4 .\n\n So ·now ·we ·have ·three ·equations :\n\n 1) ·a ·- ·b ·- ·c ·= · 1 / 2 \n\n 2) ·- a ·+ ·b ·- ·c ·= · 1 / 3 \n\n 3) ·- a ·- ·b ·+ ·c ·= · 1 / 4 \n\n And ·also , ·from ·earlier , ·we ·have ·that ·a ·+ ·b ·+ ·c ·= ·log ₂ (x yz) ·= ·- 1 3 / 1 2 .\n\n So ·maybe ·we ·can ·solve ·this ·system ·of ·equations ·for ·a , ·b , ·c ?\n\n Let ·me ·write ·the ·three ·equations ·again :\n\n Equ ation · 1 : ·a ·- ·b ·- ·c ·= · 1 / 2 \n\n Equ ation · 2 : ·- a ·+ ·b ·- ·c ·= · 1 / 3 \n\n Equ ation · 3 : ·- a ·- ·b ·+ ·c ·= · 1 / 4 \n\n And ·equation · 4 : ·a ·+ ·b ·+ ·c ·= ·- 1 3 / 1 2 \n\n So ·we ·have ·four ·equations . ·But ·since ·equations · 1 , · 2 , · 3 ·are ·linear ·equations ·in ·a , ·b , ·c , ·and ·equation · 4 ·is ·another ·equation . ·Let ·me ·check ·if ·they ·are ·consistent .\n\n Alternatively , ·maybe ·I ·can ·solve ·equations · 1 , · 2 , · 3 ·first . ·Let ·me ·try ·that .\n\n Let ·me ·write ·equations · 1 , · 2 , · 3 :\n\n Equ ation · 1 : ·a ·- ·b ·- ·c ·= · 1 / 2 \n\n Equ ation · 2 : ·- a ·+ ·b ·- ·c ·= · 1 / 3 \n\n Equ ation · 3 : ·- a ·- ·b ·+ ·c ·= · 1 / 4

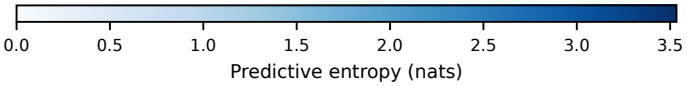


Figure 11: State-wise predictive entropy along a reasoning trajectory (Part 2/6).

\n\n Let ·me ·try ·to ·solve ·these ·equations . ·Maybe ·adding ·equations · 1 ·and · 2 ?\n\n Equ ation · 1 ·+ ·Equation · 2 : ·(a ·- ·b ·- ·c) ·+ ·(- a ·+ ·b ·- ·c) ·= · 1 / 2 ·+ · 1 / 3 .\n\n S implify ·left ·side : ·a ·- ·a ·- ·b ·+ ·b ·- ·c ·- ·c ·= ·- 2 c . ·Right ·side : · 1 / 2 ·+ · 1 / 3 ·= · 5 / 6 . ·Therefore , ·- 2 c ·= · 5 / 6 ·=> ·c ·= ·- 5 / 1 2 .\n\n So ·c ·is ·- 5 / 1 2 . ·Let ·me ·note ·that .\n\n Now , ·let ·me ·substitute ·c ·= ·- 5 / 1 2 ·into ·equations · 1 ·and · 2 .\n\n Equ ation · 1 : ·a ·- ·b ·- ·(- 5 / 1 2) ·= · 1 / 2 ·=> ·a ·- ·b ·+ · 5 / 1 2 ·= · 1 / 2 ·=> ·a ·- ·b ·= · 1 / 2 ·- · 5 / 1 2 ·= · 6 / 1 2 ·- · 5 / 1 2 ·= · 1 / 1 2 . ·So ·equation · 1 ·becomes ·a ·- ·b ·= · 1 / 1 2 .\n\n Equ ation · 2 : ·- a ·+ ·b ·- ·(- 5 / 1 2) ·= · 1 / 3 ·=> ·- a ·+ ·b ·+ · 5 / 1 2 ·= · 1 / 3 ·=> ·- a ·+ ·b ·= · 1 / 3 ·- · 5 / 1 2 ·= · 4 / 1 2 ·- · 5 / 1 2 ·= ·- 1 / 1 2 . ·So ·equation · 2 ·becomes ·- a ·+ ·b ·= ·- 1 / 1 2 .\n\n So ·now , ·from ·equation · 1 : ·a ·- ·b ·= · 1 / 1 2 \n\n From ·equation · 2 : ·- a ·+ ·b ·= ·- 1 / 1 2 \n\n These ·are ·the ·same ·equations . ·Let ·me ·check : ·if ·I ·multiply ·equation · 2 ·by ·- 1 , ·I ·get ·a ·- ·b ·= · 1 / 1 2 , ·which ·is ·same ·as ·equation · 1 . ·So ·equations · 1 ·and · 2 ·are ·dependent ·once ·c ·is ·found . ·Therefore , ·we ·need ·another ·equation ·to ·find ·a ·and ·b . ·That 's ·equation · 3 ·or ·equation · 4 .\n\n Let ·me ·use ·equation · 3 : ·- a ·- ·b ·+ ·c ·= · 1 / 4 . ·Since ·c ·= ·- 5 / 1 2 , ·substitute ·that :\n\n - ·a ·- ·b ·+ ·(- 5 / 1 2) ·= · 1 / 4 ·=> ·- a ·- ·b ·= · 1 / 4 ·+ · 5 / 1 2 ·= · 3 / 1 2 ·+ · 5 / 1 2 ·= · 8 / 1 2 ·= · 2 / 3 . ·Therefore , ·- a ·- ·b ·= · 2 / 3 ·=> ·a ·+ ·b ·= ·- 2 / 3 .\n\n So ·now ·we ·have :\n\n From ·equation · 1 : ·a ·- ·b ·= · 1 / 1 2 \n\n From ·equation · 3 : ·a ·+ ·b ·= ·- 2 / 3 \n\n So ·now ·we ·can ·solve ·for ·a ·and ·b . ·Let ·me ·add ·these ·two ·equations :\n\n (a ·- ·b) ·+ ·(a ·+ ·b) ·= · 1 / 1 2 ·+ ·(- 2 / 3)\n\n Left ·side : · 2 a \n\n Right ·side : · 1 / 1 2 ·- · 8 / 1 2 ·= ·- 7 / 1 2 \n\n Therefore , · 2 a ·= ·- 7 / 1 2 ·=> ·a ·= ·- 7 / 2 4 \n\n Then , ·from ·equation ·a ·+ ·b ·= ·- 2 / 3 , ·so ·b ·= ·- 2 / 3 ·- ·a ·= ·- 2 / 3 ·- ·(- 7 / 2 4) ·= ·- 2 / 3 ·+ · 7 / 2 4 .\n\n Convert ·- 2 / 3 ·to ·- 1 6 / 2 4 , ·so ·- 1 6 / 2 4 ·+ · 7 / 2 4 ·= ·- 9 / 2 4 ·= ·- 3 / 8 .\n\n Therefore , ·b ·= ·- 3 / 8 .\n\n So ·now ·we ·have :\n\n a ·= ·log ₂ x ·= ·- 7 / 2 4 \n\n b ·= ·log ₂ y ·= ·- 3 / 8 \n\n c ·= ·log ₂ z ·= ·- 5 / 1 2 \n\n Let ·me ·check ·if ·these ·values ·satisfy ·equation · 4 : ·a ·+ ·b ·+ ·c ·= ·- 7 / 2 4 ·- · 3 / 8 ·- · 5 / 1 2 .\n\n Convert ·all ·to · 2 4 ·denomin ators :\n\n - 7 / 2 4 ·- · 9 / 2 4 ·- · 1 0 / 2 4 ·= ·(- 7 ·- · 9 ·- 1 0)/ 2 4 ·= ·(- 2 6)/ 2 4 ·= ·- 1 3 / 1 2 . ·Which ·matches ·equation · 4 . ·Good .\n\n So ·that 's ·correct .\n\n Now , ·the ·problem ·asks ·for ·| log ₂ (x ^ 4 ·y ^ 3 ·z ^ 2)| .\n\n Which ·is ·| 4 a ·+ · 3 b ·+ · 2 c | .\n\n Let ·me ·compute · 4 a ·+ · 3 b ·+ · 2 c .\n\n Compute ·each ·term :\n\n 4 a ·= · 4 *(- 7 / 2 4) ·= ·- 2 8 / 2 4 ·= ·- 7 / 6 \n\n 3 b ·= · 3 *(- 3 / 8) ·= ·- 9 / 8 \n\n 2 c ·= · 2 *(- 5 / 1 2) ·= ·- 1 0 / 1 2 ·= ·- 5 / 6 \n\n So ·total ·is ·- 7 / 6 ·- 9 / 8 ·- 5 / 6 .\n\n Let ·me ·combine ·these ·terms . ·First , ·combine ·- 7 / 6 ·- 5 / 6 ·= ·- 1 2 / 6 ·= ·- 2 .\n\n Then , ·- 2 ·- 9 / 8 . ·Convert ·- 2 ·to ·- 1 6 / 8 , ·so ·- 1 6 / 8 ·- 9 / 8 ·= ·- 2 5 / 8 .\n\n Therefore , · 4 a ·+ · 3 b ·+ · 2 c ·= ·- 2 5 / 8 . ·Therefore , ·the ·absolute ·value ·is ·| ·- 2 5 / 8 ·| ·= · 2 5 / 8 .\n\n So ·the ·answer ·is · 2 5 / 8 . ·Then ·m ·is · 2 5 , ·n ·is · 8 . ·They ·are ·cop r ime ·since · 2 5 ·and · 8 ·share ·no ·common ·factors ·besides · 1 . ·Therefore , ·m ·+ ·n ·= · 2 5 ·+ · 8 ·= · 3 3 .\n\n Wait , ·but

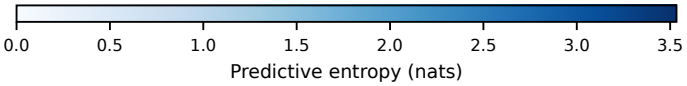


Figure 12: State-wise predictive entropy along a reasoning trajectory (Part 3/6).

·let ·me ·check ·my ·calculation·again . ·Because·I ·might ·have ·made ·a ·mistake·in ·adding .\n\n So :\n\n 4 a ·= · 4 *(- 7 / 2 4) ·= ·- 2 8 / 2 4 ·= ·- 7 / 6 \n\n 3 b ·= · 3 *(- 3 / 8) ·= ·- 9 / 8 \n\n 2 c ·= · 2 *(- 5 / 1 2) ·= ·- 1 0 / 1 2 ·= ·- 5 / 6 \n\n So ·adding ·them ·together : ·- 7 / 6 ·- 9 / 8 ·- 5 / 6 .\n\n First , ·combine ·- 7 / 6 ·- 5 / 6 ·= ·- 1 2 / 6 ·= ·- 2 . ·Then , ·- 2 ·- 9 / 8 . ·Let ·me ·convert ·- 2 ·to ·- 1 6 / 8 . ·So ·- 1 6 / 8 ·- 9 / 8 ·= ·- 2 5 / 8 . ·Yes , ·that 's ·correct . ·So ·the ·value ·is ·- 2 5 / 8 , ·absolute·value ·is · 2 5 / 8 . ·Therefore , ·m ·= · 2 5 , ·n ·= · 8 , ·m ·+ ·n ·= · 3 3 .\n\n But ·wait , ·let ·me ·check ·if ·my ·values ·of ·a , ·b , ·c ·are ·correct . ·Let ·me ·verify ·with ·the ·original ·equations .\n\n First ·equation : ·a ·- ·b ·- ·c ·= ·(- 7 / 2 4) ·- ·(- 3 / 8) ·- ·(- 5 / 1 2) ·= ·- 7 / 2 4 ·+ · 3 / 8 ·+ · 5 / 1 2 .\n\n Convert ·all ·to · 2 4 ·denomin ators :\n\n - 7 / 2 4 ·+ · 9 / 2 4 ·+ · 1 0 / 2 4 ·= ·(- 7 ·+ · 9 ·+ · 1 0)/ 2 4 ·= · 1 2 / 2 4 ·= · 1 / 2 . ·Which ·matches ·the ·first ·equation . ·Good .\n\n Second ·equation : ·- a ·+ ·b ·- ·c ·= ·- (- 7 / 2 4) ·+ ·(- 3 / 8) ·- ·(- 5 / 1 2) ·= · 7 / 2 4 ·- · 3 / 8 ·+ · 5 / 1 2 .\n\n Convert ·to · 2 4 ·denomin ators :\n\n 7 / 2 4 ·- · 9 / 2 4 ·+ · 1 0 / 2 4 ·= ·(7 ·- · 9 ·+ · 1 0)/ 2 4 ·= · 8 / 2 4 ·= · 1 / 3 . ·Which ·matches ·the ·second ·equation . ·Good .\n\n Third ·equation : ·- a ·- ·b ·+ ·c ·= ·- (- 7 / 2 4) ·- ·(- 3 / 8) ·+ ·(- 5 / 1 2) ·= · 7 / 2 4 ·+ · 3 / 8 ·- 5 / 1 2 .\n\n Convert ·to · 2 4 ·denomin ators :\n\n 7 / 2 4 ·+ · 9 / 2 4 ·- 1 0 / 2 4 ·= ·(7 ·+ · 9 ·- 1 0)/ 2 4 ·= · 6 / 2 4 ·= · 1 / 4 . ·Which ·matches ·the ·third ·equation . ·Perfect .\n\n So ·all ·equations ·are ·satisfied . ·Therefore , ·the ·calculation·for · 4 a ·+ · 3 b ·+ · 2 c ·is ·correct . ·Therefore , ·the ·answer ·is · 2 5 / 8 , ·so ·m ·+ ·n ·is · 3 3 .\n\n Wait , ·but ·let ·me ·check ·again ·if ·I ·did ·the ·problem ·correctly . ·The ·problem ·says ·" the ·value ·of ·| log ₂ (x ^ 4 ·y ^ 3 ·z ^ 2)| ". ·So ·log ₂ (x ^ 4 ·y ^ 3 ·z ^ 2) ·is · 4 ·log ·x ·+ · 3 ·log ·y ·+ · 2 ·log ·z , ·which ·is · 4 a ·+ · 3 b ·+ · 2 c . ·Which ·is ·- 2 5 / 8 . ·Absolute ·value ·is · 2 5 / 8 . ·So ·yes , · 2 5 / 8 . ·So ·m ·is · 2 5 , ·n ·is · 8 . · 2 5 ·and · 8 ·are ·cop r ime , ·so · 2 5 ·+ · 8 ·= · 3 3 . ·Therefore , ·the ·answer ·is · 3 3 .\n\n But ·just ·to ·be ·thorough , ·let ·me ·check ·if ·I ·can ·compute ·x , ·y , ·z ·and ·then ·compute ·log ₂ (x ^ 4 ·y ^ 3 ·z ^ 2) ·directly . ·Maybe ·that 's ·a ·good ·idea ·to ·verify .\n\n We ·have ·a ·= ·log ₂ x ·= ·- 7 / 2 4 ·=> ·x ·= · 2 ^ (- 7 / 2 4)\n\n Similarly , ·y ·= · 2 ^ (- 3 / 8) ·= · 2 ^ (- 9 / 2 4)\n\n z ·= · 2 ^ (- 5 / 1 2) ·= · 2 ^ (- 1 0 / 2 4)\n\n Therefore , ·x ^ 4 ·= ·[2 ^ (- 7 / 2 4)] ^ 4 ·= · 2 ^ (- 2 8 / 2 4) ·= · 2 ^ (- 7 / 6)\n\n y ^ 3 ·= ·[2 ^ (- 9 / 2 4)] ^ 3 ·= · 2 ^ (- 2 7 / 2 4) ·= · 2 ^ (- 9 / 8)\n\n z ^ 2 ·= ·[2 ^ (- 1 0 / 2 4)] ^ 2 ·= · 2 ^ (- 2 0 / 2 4) ·= · 2 ^ (- 5 / 6)\n\n Therefore , ·x ^ 4 ·y ^ 3 ·z ^ 2 ·= · 2 ^ (- 7 / 6) ·* · 2 ^ (- 9 / 8) ·* · 2 ^ (- 5 / 6) ·= · 2 ^ (- 7 / 6 ·- 9 / 8 ·- 5 / 6)\n\n Compute ·the ·exponent :\n\n First , ·- 7 / 6 ·- 5 / 6 ·= ·- 1 2 / 6 ·= ·- 2 . ·Then , ·- 2 ·- 9 / 8 ·= ·- 2 ·- · 1 . 1 2 5 ·= ·- 3 . 1 2 5 . ·Which ·is ·- 2 5 / 8 . ·Therefore , ·log ₂ (x ^ 4 ·y ^ 3 ·z ^ 2) ·= ·- 2 5 / 8 . ·Absolute ·value ·is · 2 5 / 8 . ·So ·same ·result . ·Therefore , ·correct .\n\n Therefore , ·the ·answer ·is · 2 5 / 8 , ·so ·m ·+ ·n ·is · 3 3 . ·Therefore , ·Answer : · 3 3 .\n\n ** Final ·Answer **\n \ boxed { 3 3 }\n </think>\n\n To ·solve ·the ·given ·system ·of ·logarithmic ·equations :\n\n $$ \n \ log _ 2 \ left (\ frac {x }{ yz }\ right) ·= ·\ frac { 1 }{ 2

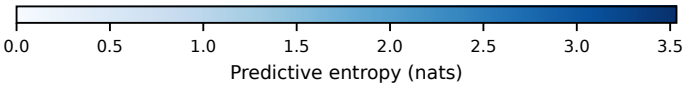


Figure 13: State-wise predictive entropy along a reasoning trajectory (Part 4/6).



Figure 14: State-wise predictive entropy along a reasoning trajectory (Part 5/6).

frac { 1 0 }{ 1 2 } ·= ·- \ frac { 2 0 }{ 2 4 }\n $$ \n\n Add :\n\n $$ \n -\ frac { 2 8 }{ 2 4 } ·- ·\ frac { 2 7 }{ 2 4 } ·- ·\ frac { 2 0 }{ 2 4 } ·= ·- \ frac { 7 5 }{ 2 4 } ·= ·- \ frac { 2 5 }{ 8 }\n $$ \n\n Thus :\n\n $$ \n |\ log _ 2 (x ^ 4 ·y ^ 3 ·z ^ 2)| ·= ·\ left | ·- \ frac { 2 5 }{ 8 } ·\ right | ·= ·\ frac { 2 5 }{ 8 }\n $$ \n\n So ·$ ·m ·= · 2 5 ·$, ·$ ·n ·= · 8 ·$, ·and ·$ ·m ·+ ·n ·= · 3 3 ·$.\n\n ---\n\n ### ·Final ·Answer \n\n $$ \n \ boxed { 3 3 }\n $$

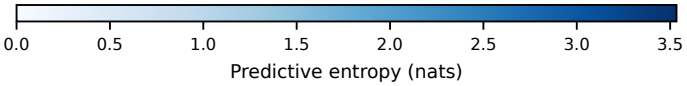


Figure 15: State-wise predictive entropy along a reasoning trajectory (Part 6/6).