AIPO: Adaptive Information Guided Token-Level Reinforcement Learning for Large Language Model Reasoning

Bin Chen[*1], **Hongfei Ye**[*1], **Huiyang Wang**[1], **Wenxi Liu**[1], **Yu Zhang**[1], **Furui Liu**[1,2†]

[1]University of Chinese Academy of Sciences, [2]Zhejiang Lab

{chenbin232,yehongfei23,wanghuiyang23,liuwenxi23,zhangyu2312}@mails.ucas.ac.cn, liufurui@zhejianglab.com

Abstract

Reinforcement Learning with Verifiable Rewards (RLVR) improves the reasoning capability of Large Language Models (LLMs). Current RLVR trains LLMs on all generated tokens, rather than exploring which tokens actually contribute to reasoning. We propose **AIPO** (Adaptive–Information Policy Optimization), which focuses updates on those decisive tokens discovered on the fly. AIPO estimates each hidden state's mutual information to score tokens. Policy gradients are then computed only on these critical tokens, using an advantage that verifiable correctness. To improve the efficiency of mutual-information estimation, AIPO adopts a Random–Fourier approximation of the Hilbert–Schmidt Independence Criterion. Across five math and science benchmarks, AIPO yields up to **+20%** accuracy over strong RLVR baselines while updating merely **10%** of tokens, demonstrating superior efficiency and effectiveness. Our findings highlight the importance of information–driven token selection for efficient and effective reinforcement learning of LLM reasoning.

1 Introduction

Large Language Models (LLMs) have rapidly advanced the state of multi-step reasoning in mathematics, programming, and science (Wei et al., 2022; Wang et al., 2025b; Wenjuan et al., 2025). A prevailing technique behind these improvements is Reinforcement Learning with Verifiable Rewards (RLVR), which optimises LLMs against deterministic checkers that validate solution correctness (Shao et al., 2024; Lambert et al., 2024).

However, existing RLVR methods still suffer from limited training effectiveness. Most approaches optimize all generated tokens in a blanket manner, overlooking that different tokens contribute unequally to final performance. Because tokens serve heterogeneous functions during reasoning, such uniform updates devote substantial effort to low-impact tokens, diminishing overall training effectiveness and diluting the gains that could be achieved by focusing on critical tokens (Babu et al., 2025; Pursell and Maiti, 2024). Despite strong evidence that only a small fraction of steps materially affect the final answer, standard RLVR spreads gradient updates evenly across the entire sequence. This uniform token optimization wastes computation and weakens learning signals, ultimately capping the achievable performance of large-scale reasoning models.

Prior work((Vassoyan et al., 2025; Wang et al., 2025a)) has sought to improve training efficiency by selectively updating only the most informative tokens. Prior strategies typically assess token importance either by evaluating whether the token steers generation toward a correct answer, or by computing information–theoretic scores such as token-level entropy. Mutual information (MI) offers another principled information-theoretic measure, with higher MI generally indicating lower prediction bias. However, MI has not yet been exploited for token selection during training. To address this gap, we introduce AIPO, an adaptive MI-guided pruning and optimization framework that targets the most influential tokens.

We introaduce **AIPO** (Adaptive–Information Policy Optimization), an information-theoretic RLVR framework that learns where to update while it learns what to predict. AIPO leverages a Random-Fourier implementation of the Hilbert–Schmidt Independence Criterion (HSIC) to estimate the mutual information (MI) between each intermediate (hidden) representation and the gold-standard answer in real time. Tokens whose MI exceeds an adaptive threshold are labelled critical. Policy gradients are computed only on these tokens and are weighted by a composite advantage that couples

[*]Equal contribution. The first two authors contributed equally to this work. [†]Corresponding author: Furui Liu (liufurui@zhejianglab.com).

Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 44441–44450

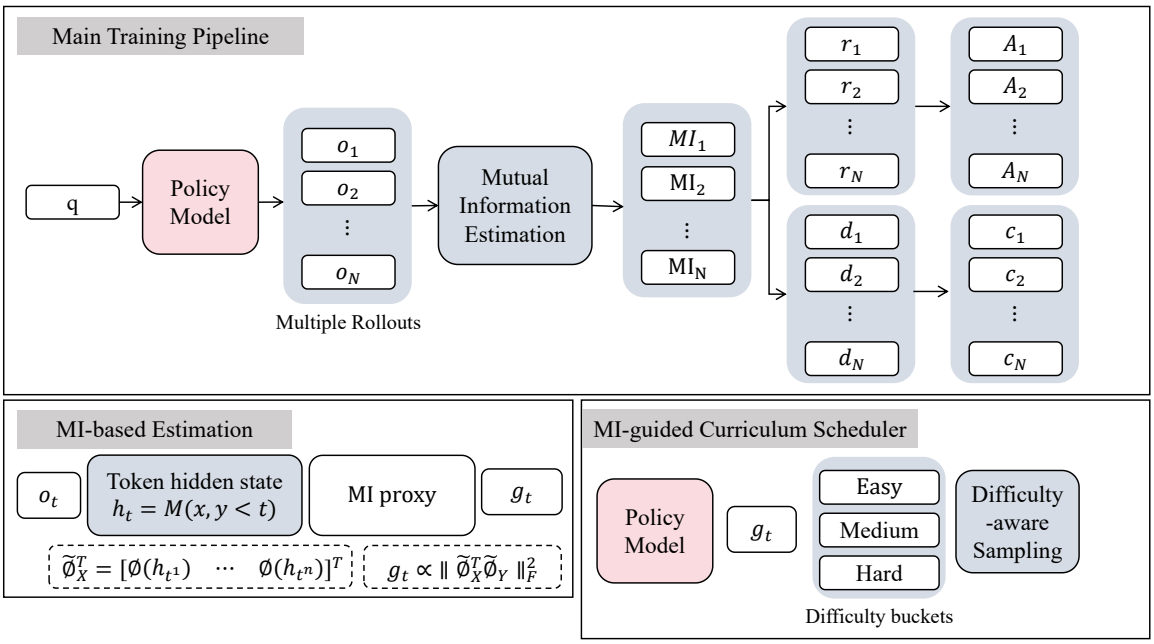

Figure 1: Overview of AIPO. **(a) Main training pipeline** For every question x the policy model π_θ produces N roll-outs $o_1, \ldots, o_N$. A mutual-information module ranks the roll-outs token-wise, after which a reward head supplies task rewards $r_1, \ldots, r_N$. Combining MI scores and rewards yields information-aware advantages $A_1, \ldots, A_N$ that drive the policy update. **(b) MI-based estimation** At each step t the hidden state h_t is mapped through a Random Fourier Feature block to obtain an MI proxy $\widehat{\mathrm{MI}}_t$ with respect to the gold answer representation $h^\star$. The computation scales linearly with batch size n and projection dimension D. **(c) MI-guided curriculum scheduler** The sequence of $\widehat{\mathrm{MI}}_t$ values is scanned for local peaks whose count serves as a difficulty signal. Training batches are routed to easy, medium, or hard buckets, and κ is relaxed whenever validation accuracy saturates so that progressively longer subsequences become trainable.

MI gain with verifiable rewards. A progressive relaxation schedule gradually expands the set of optimised tokens as the model stabilises, ensuring robust early training and thorough later exploration.

- We propose **AIPO**, the novel RLVR algorithm to adaptively detect high-MI tokens during generation and focus optimisation on them.
- To enhance the efficiency of the AIPO training process, we introduce a curriculum learning strategy that organizes samples by difficulty through mutual-information–based stratification.
- Experiments on five challenging reasoning benchmarks show that AIPO surpasses strong baselines by up to 20 % accuracy while updating only 10 % of the tokens, demonstrating superior efficiency and effectiveness.

2 Related Work

RL-based optimisation for LLM reasoning. A growing body of research applies reinforcement learning, most notably Proximal Policy Optimization (PPO)(Schulman et al., 2017), to strengthen LLM reasoning. RLHF pipelines fine-tune models with outcome-level reward models (Ouyang et al., 2022), while more recent work adopts process-level reward models (PRMs) to provide step-wise supervision and improve reasoning transparency (Sun et al., 2024; Wang et al., 2024; Yuan et al., 2024). Several studies narrow the optimisation scope to reduce gradient waste: DAPO introduces selective token refinement via forking tokens (Yu et al., 2025), and GRPO exploits gradient truncation to curb over-correction (Guo et al., 2025). Nonetheless, these methods either rely on static heuristics (e.g., prefix length(Sun et al., 2025)), leaving token-level information signals untapped. Our work differs by using an online mutual-information probe to adaptively identify high-impact tokens and incorporating their information gain directly into the RL objective.

Information-theoretic analysis and critical tokens. Information theory has recently been employed to understand LLM behaviour, from study-

ing information flow during generation (Jeon, 2025; Ton et al.) to analysing inductive biases (Ren and Liu, 2025). In parallel, research on critical tokens shows that a small subset of tokens can drastically influence model outputs, motivating efforts to detect and control them (Lin et al., 2024; Goldshmidt and Horovicz, 2024). Qian et al. (2024) report a Mutual-Information Peaks phenomenon, where "thinking tokens" trigger hidden states highly informative of the correct answer leverages these peaks only at inference time. By computing mutual information in real time during training, AIPO directing policy updates toward MI-critical tokens and thus bridging the gap between analytical insights and training algorithms.

3 Methodology

This section introduces **AIPO** (Adaptive Information Policy Optimization), a reinforcement-learning framework that dynamically focuses training on the informative tokens. Unlike fixed-position or pattern-based heuristics, AIPO employs an efficient information-theoretic probe to identify critical tokens in real time during training and combines their information gain with task correctness when estimating policy advantages. Consequently, the policy is able to highlight any influential tokens and converge with markedly fewer optimisation tokens.

3.1 Overview

Given a question x and its gold answer y, the policy model π_θ produces a sequence $o = (o_1, \ldots, o_T)$ in an auto–regressive manner. At each step t, we extract the hidden representation $\boldsymbol{h}_t$, and a fixed representation $\boldsymbol{h}^\star$ for the ground truth answer. The AIPO comprises three stages. First, during *fast information probing*, an information coefficient $\widehat{\mathrm{MI}}_t$ is obtained between the current hidden state $\mathbf{h}_t$ and the gold representation $\boldsymbol{h}^\star$ via a Random Fourier Feature approximation of the Hilbert–Schmidt Independence Criterion. The computation grows linearly with both the batch size n and the projection dimension D (with D typically much smaller than n). Second, *adaptive token masking* detects and suppresses low–mutual-information tokens by assigning a mask $m_t = 1[\,\widehat{\mathrm{MI}}_t > \mu + \kappa\sigma\,]$, μ denotes the running mean of the set $\{\widehat{\mathrm{MI}}_1, \ldots, \widehat{\mathrm{MI}}_T\}$, σ its running standard deviation, and κ is a hyperparameter. Masked tokens are excluded from gradient updates. Third, information-aware advantage gives each unmasked token an advantage A_t. The reward R comprises correctness, formatting, and length reward and $\bar{R}$ is the mini-batch baseline. Only those unmasked positions participate in the policy–gradient update. To avoid prematurely freezing potentially useful tokens, the threshold κ gradually adjusted when validation accuracy plateaus, enlarging the optimised subsequence in a curriculum.

3.2 FastHSIC-RFF online MI estimation and adaptive key–token selection.

For every training example and model M, we obtain the hidden state of step t as $h_t = M(x,\, o_{<t}) \in \mathbb{R}^d$, while a forward pass over the gold answer y yields $\boldsymbol{h}^\star = M(y)$. A Gaussian kernel is approximated with Random Fourier Features $\phi(h) = \sqrt{2/D}\,\cos(Wh + b)$, where $W \sim \mathcal{N}(0, \sigma^{-2}I)$ and $b \sim \mathcal{U}(0, 2\pi)$, so the projection cost is $\mathcal{O}(nD)$ for $D \ll n$. Setting $\Phi_x = \phi(h_t)$ and $\Phi_y = \phi(\boldsymbol{h}^\star)$, we compute the centred FastHSIC estimate

$$\widehat{\mathrm{MI}}_t \;=\; \frac{\left\|\Phi_x^\top \Phi_y\right\|_F^2}{(n-1)^2}, \tag{1}$$

which remains fully differentiable with respect to h_t. While streaming through $\{\widehat{\mathrm{MI}}_t\}$ we track the running mean μ and standard deviation σ (or the inter-quartile range, IQR) and mark step t as *information-critical* whenever $\widehat{\mathrm{MI}}_t > \mu + \kappa\sigma$. The binary mask m_t produced in this way gates optimisation: only unmasked tokens are allowed to receive policy-gradient updates, ensuring that learning focuses on the positions carrying the most information about the gold answer while keeping computational overhead minimal.

The AIPO optimization objective is then formulated as a masked, token-level clipped policy gradient:

$$\begin{aligned} J_{\mathrm{AIPO}}(\theta) = \mathbb{E}\Bigg[\frac{1}{N}\sum_{i=1}^{N}\frac{1}{\sum_{t=1}^{T} m_{i,t}}\sum_{t=1}^{T} m_{i,t}\,\Big(\min\Big(\\ \rho_{i,t}(\theta)\,A_{i,t},\, \mathrm{clip}\big(\rho_{i,t}(\theta),\,1-\epsilon,\,1+\epsilon\big)\,A_{i,t}\Big) \\ -\beta\, D_{\mathrm{KL}}\big(\pi_\theta \,\|\, \pi_{\mathrm{ref}}\big)\Big)\Bigg], \end{aligned} \tag{2}$$

where

$$\rho_{i,t}(\theta) = \frac{\pi_\theta\big(o_{i,t} \mid x, o_{i,<t}\big)}{\pi_{\theta_{\mathrm{old}}}\big(o_{i,t} \mid x, o_{i,<t}\big)} \tag{3}$$

3.3 The training algorithm via MI-based curriculum learning strategy

Algorithm 1 gives an overview of the full training routine. Unlike standard token-level policy optimisation, AIPO sends gradient only to tokens whose

mutual information is high, so in practice it updates roughly ten percent of the tokens. To keep learning stable while gradually presenting the agent with harder reasoning cases, we add a difficulty-aware curriculum that relies on the mutual-information (MI) peaks already computed by AIPO. For every generated trajectory we locate all local MI peaks whose values exceed the running mean μ plus κ times the running standard deviation σ. The number of such peaks, denoted N_{peak} and bounded between zero and forty, serves as an on-line indicator of instance difficulty. Based on this count we place each example into one of three buckets: Easy when N_{peak} is at most fifteen, Medium when it lies between sixteen and twenty five, and Hard when it is twenty six or above. During training we maintain a time-dependent sampling distribution $P^{(t)}$ over these buckets. The process begins with roughly seventy percent Easy, twenty–five percent Medium and five percent Hard examples. From step t_1, set to thirty percent of the total training steps S, up to step t_2 at seventy percent of S, this mixture is linearly shifted toward twenty-five, thirty-five and forty percent, respectively, and remains fixed thereafter, thus realising an easy-to-hard curriculum. Concurrently, the peak threshold κ is annealed from an initial value of 0.3 down to 0.1, which gradually relaxes the criterion and allows more tokens to be regarded as critical as the model becomes more capable.

3.4 Theoretical Analysis

Below we revisit the theory under the strictly reward-based advantage used by AIPO, i.e. the per–token advantage is

$$A_t \;=\; R - \bar{R}, \tag{4}$$

where R contains correctness, formatting and length bonuses and $\bar{R}$ is a baseline computed inside the mini-batch. Mutual information is *only* employed to decide which tokens are updated; it does not enter the advantage itself.

Let x denote the input, y the gold answer and π_θ an autoregressive policy that produces the sequence $o_{1:T}$. The hidden state at position t is $h_t = f_\theta(x, o_{<t})$. The verifiable reward $R(o_{1:T}, y) \in [0, 1]$ measures task success.

For every step we compute the conditional mutual information

$$\widehat{\mathrm{MI}}_t \;=\; I\big(y; h_t \;\big|\; x, o_{<t}\big), \tag{5}$$

Algorithm 1: Training Procedure of MI–Curriculum AIPO

Input: Dataset $\mathcal{D}$ of (x, y) pairs
Policy network π_θ
Verifier $\mathcal{V}$ (returns $R \in \{0, 1\}$)
Thresholds $(\kappa_{\text{start}}, \kappa_{\text{end}})$, decay span T_{dec}
Curriculum pivots (t_1, t_2), sampling probs $P^{(0)}, P^{(*)}$
Batch size B, RFF dimension D, GRPO KL coef. β
Validation stall window P_{val}, max steps S
Output: Optimised policy parameters θ

For each $(x, y) \in \mathcal{D}$ **do**
  Generate $\hat{y}$ via π_θ; compute $\{\widehat{\mathrm{MI}}_t\}$; detect peaks with κ_{start}
  $N_{\text{peak}} \leftarrow \sum_t \mathbf{1}[\widehat{\mathrm{MI}}_t \text{ is peak}]$
  if $N_{\text{peak}} \le 15$ **then**
    label $\leftarrow$ Easy
  end
  else if $N_{\text{peak}} \le 25$ **then**
    label $\leftarrow$ Medium
  end
  else
    label $\leftarrow$ Hard
  end
  Store (x, y) in $\mathcal{D}_{\textit{label}}$
end
$\kappa \leftarrow \kappa_{\text{start}}$
for $t \leftarrow 1$ **to** S **do**
  if $t < t_1$ **then**
    $P^{(t)} \leftarrow P^{(0)}$
  end
  else if $t < t_2$ **then**
    $\tau \leftarrow \frac{t - t_1}{t_2 - t_1}$; $P^{(t)} \leftarrow (1 - \tau)P^{(0)} + \tau P^{(*)}$
  end
  else
    $P^{(t)} \leftarrow P^{(*)}$
  end
  $\mathcal{B} \leftarrow$ SAMPLEBYDIFFICULTY$(\mathcal{D}_{\text{E}}, \mathcal{D}_{\text{M}}, \mathcal{D}_{\text{H}}, P^{(t)}, B)$
  $\mathcal{S} \leftarrow \emptyset$
  For each $(x, y) \in \mathcal{B}$ **do**
    generate o with π_θ; compute $\{\widehat{\mathrm{MI}}_t\}$; update (μ, σ)
    $m_t \leftarrow \mathbf{1}[\widehat{\mathrm{MI}}_t > \mu + \kappa\sigma]$; $R \leftarrow \mathcal{V}(x, o)$
    normalise $\widehat{\mathrm{MI}}_t$; compute A_t
    For each t *with* $m_t = 1$ **do**
      $\mathcal{S} \leftarrow \mathcal{S} \cup (x, o_{<t}, o_t, A_t)$
    end
  end
  $\theta \leftarrow \theta +$ GRPO_UPDATE$(\mathcal{S}, \beta)$
  if NOVALIMPROVE$(P_{\textit{val}})$ **then**
    $\kappa \leftarrow \max(\kappa_{\text{end}}, \kappa - 0.05)$; $t \leftarrow \min(t_2, t + 0.05S)$
  end
  $\kappa \leftarrow \max(\kappa_{\text{end}}, \kappa_{\text{start}} - 0.2t/T_{\text{dec}})$
end
return θ

and declare a step to be *critical* if $\widehat{\mathrm{MI}}_t \ge \tau$, where τ is an adaptive threshold. The resulting set

$$\mathcal{C}(o_{1:T}) \;=\; \{\, t \in [T] : \widehat{\mathrm{MI}}_t \ge \tau \,\} \tag{6}$$

contains the positions whose internal representations carry the bulk of information about y.

Prediction-error bounds motivate focusing on $\mathcal{C}$. Let p_e be the probability that the model's prediction o is wrong and define the cumulative conditional mutual information

$$\mathcal{I}_{1:T} = \sum_{t=1}^{T} I\big(y; h_t \mid x, h_{<t}\big). \tag{7}$$

The following two inequalities quantify how $\mathcal{I}_{1:T}$ constrains p_e.

Theorem 1 (Lower and upper bounds on the prediction error)**.** *For any autoregressive policy that produces hidden states $h_{1:T}$ as above, the error probability satisfies*

$$p_e \geq \frac{1}{\log(|y|-1)}\Big[H(y) - \mathcal{I}_{1:T} - 1\Big], \tag{8}$$

$$p_e \leq H(y) - \mathcal{I}_{1:T}, \tag{9}$$

where $|y|$ is the label-space size.

Proof sketch. For the lower bound (8), the chain rule for entropy decomposes the conditional entropy as $H(y \mid o) = H(y \neq o \mid o) + p_e\, H(y \mid o, y{=}o)$. So Fano's inequality implies $H(y\,|\,o) \leq 1 + p_e \log(|y|-1)$. Since mutual information is defined as $I(y;o) = H(y) - H(y \mid o)$, and the data-processing inequality guarantees $I(y;o) \leq I(y;h_{1:T}) = \mathcal{I}_{1:T}$, re-arranging terms yields the desired lower bound.

Turning to the upper bound (9), for any discrete distribution $(p_1, \ldots, p_m)$, establishing that $1 - \max_i p_i \leq \frac{1}{2} H(p_1, \ldots, p_m)$. Taking expectations over the hidden trajectory then gives $p_e \leq H(y \mid h_{1:T})$. Finally, substituting the chain-rule identity $H(y\,|\,h_{1:T}) = H(y) - \mathcal{I}_{1:T}$ leads directly to the upper bound.

These bounds show that *increasing* the cumulative CMI $\mathcal{I}_{1:T}$ simultaneously *lowers* the attainable error floor (by tightening the lower bound) and *forces* the actual error to decrease (via the upper bound). Because the dominant contribution to $\mathcal{I}_{1:T}$ stems from the few steps with the largest $\widehat{\mathrm{MI}}_t$, concentrating learning on those *critical* positions is the most effective way to reduce p_e.

Standard RL for language-model verification and repair (RLVR) updates every position:

$$\nabla_\theta J(\theta) = \mathbb{E}_{o_{1:T}\sim\pi_\theta}\Big[\textstyle\sum_{t=1}^{T} \nabla_\theta \log \pi_\theta(o_t|x,o_{<t})\; A_t\Big]. \tag{10}$$

When only a small fraction of steps is decisive, the same scalar advantage is back-propagated through many non-critical tokens whose gradients behave almost like zero-mean noise. These spurious updates inflate gradient variance and waste optimisation budget.

AIPO instead restricts updates to the critical set:

$$\nabla_\theta J_{\text{AIPO}}(\theta) = \mathbb{E}_{o_{1:T}\sim\pi_\theta}\Big[\sum_{t\in\mathcal{C}(o_{1:T})} \nabla_\theta \log \pi_\theta(o_t \mid x, o_{<t})\; A_t\Big], \tag{11}$$

with $m_t = 1[t \in \mathcal{C}]$. Let $K = \mathbb{E}\big[|\mathcal{C}(o_{1:T})|\big]$ empirically $K \ll T$. Near the optimum, gradients from non-critical positions are almost independent of the global reward and have expected value close to zero. Omitting them introduces negligible bias yet removes a large source of variance. Consequently, for a fixed token-update budget, AIPO increases the signal-to-noise ratio of every update and accelerates convergence.

4 Experiments

In this section, we introduce the experimental setup and analyses of our experimental results.

4.1 Experimental Setup

Training details. We employ the Qwen3-1.7B, Qwen3-4B and Qwen3-8B (Yang et al., 2025) running in thinking mode over the DAPO-Math-17K corpus. And we use A100s for training and evaluation. For every question x eight complete roll-outs $\{\mathbf{o}_i\}_{i\in[1,8]}$ are sampled. The optimisation relies on AdamW with a constant learning rate of 10^{-6}. Following Yu et al. (2025), the clipping coefficients are fixed at ε_{low} of 0.2 and $\varepsilon_{\text{high}}$ of 0.28.

Baseline methods. We compare our proposal (AIPO) with the unaligned backbone, GRPO (Shao et al., 2024), DAPO (Yu et al., 2025), INTUITOR (Zhao et al., 2025), DAPO with Forking Tokens (DAPO-FT)(Wang et al., 2025a) and PPPO. All methods are allowed a maximum generation length.

Evaluation protocol. Assessment is carried out on five widely adopted reasoning benchmarks: AIME'24 (of Problem Solving), AIME'25 (of Problem Solving, 2025), MATH 500 (Hendrycks et al., 2021), AMC'23 (of Problem Solving, 2026) and GPQA–Diamond (Rein et al., 2024). Following the protocol of Wang et al. (2025a), we run each model thirty-two independent zero-shot trials per problem and report the mean accuracy (*avg@32*).

4.2 Main Results

As shown in Table 1, AIPO attains the best results across all five evaluation datasets, outperforming the other methods by up to 20%. These findings demonstrate the effectiveness of AIPO.

Training efficiency Beyond absolute accuracy, we quantify how economically each method utilises its optimisation budget. Concretely, we report **MAG** (Mean Accuracy Gain), the increase in validation accuracy over the frozen backbone. **OTF** (Optimised Token Fraction), the share of generated tokens whose advantages are back-propagated during RLVR. **TES** (Training Efficiency Score) is defined as MAG dived OTF. Higher TES indicates a more cost-effective algorithm. As shown in Table 2, our AIPO attains the highest TES across all backbones, achieving up to 20% MAG while updating only 20% of the generated tokens, which translates into a state-of-the-art training efficiency of 92.25 TES points. In the best-case scenario, updating just 10% of the generated tokens can significantly improve the MAG.

4.3 Ablation and Analysis

Ablating the MI-based curriculum learning strategy Table 3 contrasts the full AIPO with a variant trained without the MI-based curriculum, denoted as AIPO w/o cl. On every backbone size and for every benchmark, discarding the curriculum leads to noticeable accuracy drops of roughly four points on average. The gap is especially clear on the challenging AIME'25 and GPQA, showing that the MI-based curriculum learning is a decisive factor for the gains delivered by AIPO.

High MI Tokens are Critical to LRM's Reasoning. We computed mutual information scores and found that the highest values belong chiefly to conjunctions such as *So*, *Let*, *Hmm*, *Okay*. From a semantic perspective these tokens usually carry reasoning-related functions: they can initiate a thought (*So* or *Hmm*), signal logical transitions (*Since* or *Therefore*), or organise discourse (*Let*, *Then*, *To*). Such roles may help the model sustain coherent reasoning. We conduct three experimental settings. The first is an unaltered baseline. In the second setting MI DROP we set the generation probability of a fixed set of high-MI tokens to zero. In the third setting MI FORCE whenever a high-MI token falls within the top twenty percent of the next token distribution the decoder is forced to emit it.

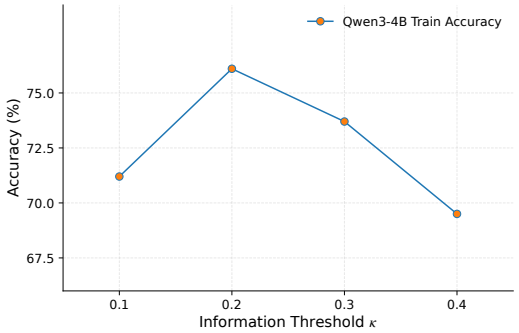

Figure 2: Effect of different information thresholds κ on the upper bound of training accuracy during training on the DAPO-Math-17K dataset.

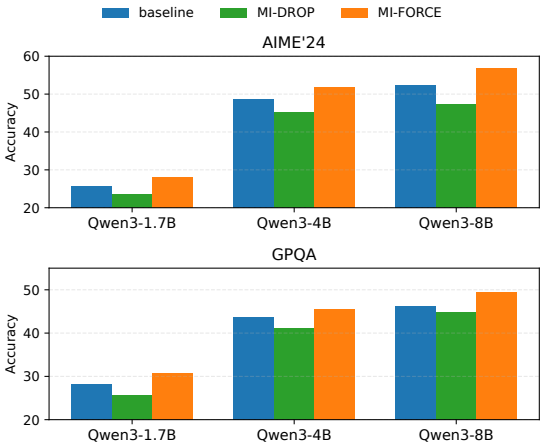

Figure 3: Impact of MI-DROP and MI-FORCE on LRMs' reasoning performance.

Figures 3 shows that suppressing high-MI tokens degrades reasoning accuracy, whereas encouraging their production consistently boosts performance.

4.4 Discussion

Impact of information threshold κ We sweep the initial threshold κ from 0.1 to 0.4 on the QWEN3–4B backbone. As illustrated in Figure 2, choosing κ that is either too small or too large hampers the selection of high-MI tokens. The default κ set to 0.2 achieves the best balance.

Impact of the N_{peak} Bucketing Scheme. As shown in Table 4, we divide the training data into three difficulty groups based on the number of mutual information peaks and report the validation accuracy for each group. Accuracy increases from the easy group to the medium group and then drops in the hard group. The trend shows that focusing only on easy questions limits performance, while starting directly with very difficult cases can severely harm reasoning ability. The results also confirm that the peak count is a reliable indicator of reason-

Method	Accuracy (%)					Avg.
	AIME'24	AIME'25	MATH 500	AMC'23	GPQA	
			Qwen3–1.7B			
Backbone	25.8	18.3	76.1	56.9	28.2	41.1
+ GRPO	27.5	20.0	82.1	60.2	29.8	43.9
+ DAPO	30.8	23.3	83.6	63.0	31.8	46.5
+ INTUITOR	27.3	17.7	78.1	61.0	30.0	42.8
+ DAPO–FT	31.2	23.9	83.9	62.8	31.4	46.7
+ PPPO	38.7	29.0	87.8	62.9	39.5	51.6
+ AIPO **(ours)**	**42.0**	**30.6**	**88.7**	**63.1**	**40.9**	**53.6**
			Qwen3–4B			
Backbone	48.8	35.4	84.5	72.7	43.6	57.0
+ GRPO	52.1	37.7	88.4	76.8	46.8	60.3
+ DAPO	56.5	42.1	92.3	81.6	49.4	64.4
+ INTUITOR	51.0	35.4	88.3	75.8	46.4	59.4
+ DAPO–FT	56.3	42.1	92.4	82.0	49.2	64.4
+ PPPO	63.5	53.4	94.6	83.1	52.1	69.3
+ AIPO **(ours)**	**66.8**	**56.7**	**94.9**	**86.4**	**53.4**	**71.6**
			Qwen3–8B			
Backbone	52.3	38.8	86.1	75.1	46.1	59.7
+ GRPO	58.8	42.3	91.0	79.4	50.5	64.4
+ DAPO	63.1	48.8	93.2	84.0	55.2	68.9
+ INTUITOR	55.4	40.8	91.2	78.2	49.3	63.0
+ DAPO–FT	63.8	49.4	93.6	84.1	54.8	69.1
+ PPPO	72.2	59.7	94.7	86.8	58.1	74.3
+ AIPO **(ours)**	**75.5**	**63.1**	**95.0**	**87.6**	**59.9**	**75.3**

Table 1: Performance on five math reasoning benchmarks. Bold numbers denote the best results and underlined numbers denote the second-best results.

Backbone	Method	MAG (%) ↑	OTF (%) ↓	TES ↑
Qwen3-1.7B	GRPO	2.87	100.00	2.87
	DAPO	5.46	100.00	5.46
	INTUITOR	1.76	100.00	1.76
	DAPO-FT	5.62	**20.00**	28.08
	PPPO	10.52	29.19	36.05
	AIPO (ours)	**12.10**	22.15	**54.61**
Qwen3-4B	GRPO	3.37	100.00	3.37
	DAPO	7.39	100.00	7.39
	INTUITOR	2.42	100.00	2.42
	DAPO-FT	7.39	**20.00**	37.02
	PPPO	12.36	26.17	47.24
	AIPO (ours)	**14.91**	21.07	**70.77**
Qwen3-8B	GRPO	4.74	100.00	4.74
	DAPO	9.19	100.00	9.19
	INTUITOR	3.34	100.00	3.34
	DAPO-FT	9.47	20.00	47.36
	PPPO	14.64	24.83	58.95
	AIPO (ours)	**18.02**	**19.54**	**92.25**

Table 2: Training effectiveness under the proposed metrics. ↑/↓ denote that higher/lower values are better. Best and second-best numbers are shown in **bold** and underlined, respectively.

Method	AIME'25	MATH 500	GPQA
	Qwen3–1.7B		
AIPO (w/o cl)	26.6	84.7	36.9
AIPO (ours)	**30.6**	**88.7**	**40.9**
	Qwen3–4B		
AIPO (w/o cl)	51.2	91.1	49.4
AIPO (ours)	**56.7**	**94.9**	**53.4**
	Qwen3–8B		
AIPO (w/o cl)	56.7	91.0	55.9
AIPO (ours)	**63.1**	**95.0**	**59.9**

Table 3: Single-column results on three math reasoning benchmarks. Removing the MI-based curriculum learning component ("AIPO (w/o cl)") consistently lowers accuracy, whereas "AIPO (ours)" denotes the full method.

ing difficulty. By beginning with easy examples and gradually introducing harder questions, the MI Curriculum yields balanced improvements across all difficulty levels.

Impact of the N_{peak} intervals To identify the two cut-points for the MI-based curriculum, we ran a grid search instead of relying on visual inspection alone. Let t_1 and t_2 be the lower and upper boundaries, constrained by $t_1 < t_2$. We swept $t_1 \in \{5, 10, 15\}$ and $t_2 \in \{15, 20, 25, 30\}$, trained a separate curriculum for every (t_1, t_2) pair, and recorded the corresponding validation accuracy. Figure 4 displays these results: the hori-

Difficulty bucket	Acc. (%) ↑
Easy ($N_{\text{peak}} \leq 15$)	68.8
Medium ($16 \leq N_{\text{peak}} \leq 25$)	75.4
Hard ($N_{\text{peak}} \geq 26$)	58.9

Table 4: Impact of N_{peak}-based Difficulty Bucketing on Validation Accuracy.

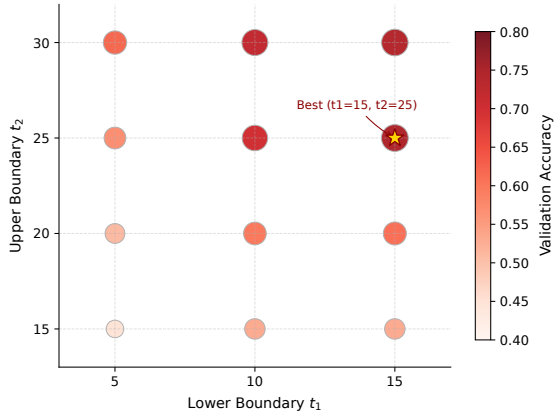

Figure 4: Impact of the N_{peak} intervals on LLMs' reasoning performance.

zontal axis shows t_1, the vertical axis shows t_2. Accuracy increases with larger t_1 until it levels off at fifteen, and it rises steadily with larger t_2 up to roughly twenty-five, beyond which gains taper off. The setting t_1 at fifteen and t_2 at twenty-five delivers the highest accuracy, balancing bucket sizes while maintaining clear difficulty separation. Consequently, all subsequent experiments adopt the three curriculum intervals $[0, 15]$, $[16, 25]$ and $[26, \infty)$.

Impact of the scheduled κ decay. We next remove Step 6 in Alg. 1 and freeze the peak-selection threshold at its initial value κ_{start}. Doing so prevents the model from gradually enlarging the set of MI-critical tokens as learning proceeds. Empirically, this modification yields two negative effects. First, the policy converges to a noticeably worse optimum: the best validation accuracy drops by $2.1\,\%$. Second, because fewer tokens satisfy the strict, fixed threshold, model receives a weaker learning signal and thus requires 23% more training steps to reach the same loss plateau. These results underline the importance of the scheduled linear decay of κ, whose progressive mask relaxation supplies the optimiser with ever-richer, yet still focused, gradient information as the policy matures.

5 Conclusion

We presented **Adaptive–Information Policy Optimization** (AIPO), an RL approach that trains language models by updating only tokens with high mutual information to the reference answer. AIPO blends this information signal with verifiable rewards to form an information-aware advantage, removing static heuristics and reducing unnecessary updates. Experiments on five reasoning benchmarks show that AIPO improves accuracy by up to 20% while using about 10% of the optimisation tokens compared with strong RLVR baselines. Beyond methodological advancements, our findings highlight the importance of trajectory-specific optimization and indicate promising avenues for integrating cognitive insights into LLM research.

Limitations

We compute the MI of the LLM at the token level. Alternative granularities such as computing MI at the level of semantic units or concepts—may reveal additional insights.

Acknowledgment

The work was supported by National Science and Technology Major Project (2023ZD0121401). We use an AI assistant for language translation.

References

Anantha Narayanan Suresh Babu, Akhil Sadam, and Pierre FJ Lermusiaux. 2025. Evaluation of analytical turbulence closures for quasi-geostrophic ocean flows with coastal boundaries. In *OCEANS 2025-Great Lakes*, pages 01–10. IEEE.

Roni Goldshmidt and Miriam Horovicz. 2024. Tokenshap: Interpreting large language models with monte carlo shapley value estimation. *arXiv preprint arXiv:2407.10114*.

Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, and 1 others. 2025. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*.

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

Hong Jun Jeon. 2025. *Information-theoretic foundations for machine learning*. Ph.D. thesis, Stanford University.

Nathan Lambert, Jacob Morrison, Valentina Pyatkin, Shengyi Huang, Hamish Ivison, Faeze Brahman, Lester James V Miranda, Alisa Liu, Nouha Dziri, Shane Lyu, and 1 others. 2024. Tulu 3: Pushing frontiers in open language model post-training. *arXiv preprint arXiv:2411.15124*.

Zicheng Lin, Tian Liang, Jiahao Xu, Qiuzhi Lin, Xing Wang, Ruilin Luo, Chufan Shi, Siheng Li, Yujiu Yang, and Zhaopeng Tu. 2024. Critical tokens matter: Token-level contrastive estimation enhances llm's reasoning capability. *arXiv preprint arXiv:2411.19943*.

Art of Problem Solving. 2024 aime i. https://artofproblemsolving.com/wiki/index.php/2024_AIME_I. Accessed: 2026-01-05.

Art of Problem Solving. 2025. 2025 aime ii problems/problem 1. https://artofproblemsolving.com/wiki/index.php/2025_AIME_II_Problems/Problem_1. Accessed: 2026-01-05.

Art of Problem Solving. 2026. Amc problems and solutions. https://artofproblemsolving.com/wiki/index.php?title=AMC_Problems_and_Solutions. Accessed: 2026-01-05; AMC 8/10/12AIME.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul F. Christiano, Jan Leike, and Ryan Lowe. 2022. Training language models to follow instructions with human feedback. In *Advances in Neural Information Processing Systems 35: Annual Conference on Neural Information Processing Systems 2022, NeurIPS 2022, New Orleans, LA, USA, November 28 - December 9, 2022*.

Noah Pursell and Anindya Maiti. 2024. Generating print-ready personalized ai art products from minimal user inputs. *arXiv preprint arXiv:2405.18247*.

Chen Qian, Dongrui Liu, Jie Zhang, Yong Liu, and Jing Shao. 2024. Dean: Deactivating the coupled neurons to mitigate fairness-privacy conflicts in large language models.

David Rein, Betty Li Hou, Asa Cooper Stickland, Jackson Petty, Richard Yuanzhe Pang, Julien Dirani, Julian Michael, and Samuel R Bowman. 2024. Gpqa: A graduate-level google-proof q&a benchmark. In *First Conference on Language Modeling*.

Ruifeng Ren and Yong Liu. 2025. Revisiting transformers through the lens of low entropy and dynamic sparsity. *arXiv preprint arXiv:2504.18929*.

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

Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, YK Li, Yang Wu, and 1 others. 2024. Deepseekmath: Pushing the limits of mathematical reasoning in open language models. *arXiv preprint arXiv:2402.03300*.

Yiliu Sun, Zicheng Zhao, Yang Wei, Yanfang Zhang, and Chen Gong. 2025. Well begun, half done: Reinforcement learning with prefix optimization for llm reasoning. *arXiv preprint arXiv:2512.15274*.

Zhiqing Sun, Longhui Yu, Yikang Shen, Weiyang Liu, Yiming Yang, Sean Welleck, and Chuang Gan. 2024. Easy-to-hard generalization: Scalable alignment beyond human supervision. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*.

Jean-Francois Ton, Muhammad Faaiz Taufiq, and Yang Liu. Understanding chain-of-thought in llms through information theory, 2024. *URL https://arxiv.org/abs/2411.11984*.

Jean Vassoyan, Nathanaël Beau, and Roman Plaud. 2025. Ignore the kl penalty! boosting exploration on critical tokens to enhance rl fine-tuning. *arXiv preprint arXiv:2502.06533*.

Peiyi Wang, Lei Li, Zhihong Shao, Runxin Xu, Damai Dai, Yifei Li, Deli Chen, Yu Wu, and Zhifang Sui. 2024. Math-Shepherd: Verify and reinforce LLMs step-by-step without human annotations. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 9426–9439, Bangkok, Thailand. Association for Computational Linguistics.

Shenzhi Wang, Le Yu, Chang Gao, Chujie Zheng, Shixuan Liu, Rui Lu, Kai Dang, Xionghui Chen, Jianxin Yang, Zhenru Zhang, and 1 others. 2025a. Beyond the 80/20 rule: High-entropy minority tokens drive effective reinforcement learning for llm reasoning. *arXiv preprint arXiv:2506.01939*.

Yaoting Wang, Shengqiong Wu, Yuecheng Zhang, Shuicheng Yan, Ziwei Liu, Jiebo Luo, and Hao Fei. 2025b. Multimodal chain-of-thought reasoning: A comprehensive survey. *arXiv preprint arXiv:2503.12605*.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, and 1 others. 2022. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35:24824–24837.

Lian Wenjuan, Zhang Chengxin, Zhang Hongbao, Jia Bin, and Liu Baihang. 2025. Rulemaster+: Llm-based automated rule generation framework for intrusion detection systems. *Chinese Journal of Electronics*, 34(5):1402.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao,

Chengen Huang, and Chenxu Lv. 2025. Qwen3 technical report.

Qiying Yu, Zheng Zhang, Ruofei Zhu, Yufeng Yuan, Xiaochen Zuo, Yu Yue, Weinan Dai, Tiantian Fan, Gaohong Liu, Lingjun Liu, and 1 others. 2025. Dapo: An open-source llm reinforcement learning system at scale. *arXiv preprint arXiv:2503.14476*.

Lifan Yuan, Wendi Li, Huayu Chen, Ganqu Cui, Ning Ding, Kaiyan Zhang, Bowen Zhou, Zhiyuan Liu, and Hao Peng. 2024. Free process rewards without process labels. *arXiv preprint arXiv:2412.01981*.

Xuandong Zhao, Zhewei Kang, Aosong Feng, Sergey Levine, and Dawn Song. 2025. Learning to reason without external rewards. *arXiv preprint arXiv:2505.19590*.