

What Makes Reasoning Invalid: Echo Reflection Mitigation for Large Language Models

Anonymous ACL submission

Abstract

Large Language Models (LLMs) excel in various reasoning tasks, including complex mathematical reasoning. However, when applied to domain-specific tasks, they consistently fail to generate novel insights during the reflection stage. Instead of conducting genuine cognitive refinement, the model tends to *mechanically reiterate earlier reasoning steps without introducing new information or perspectives, a phenomenon referred to as “Echo Reflection”*. We attribute this behavior to two key defects: (1) Uncontrollable information flow during response generation, which allows premature intermediate thoughts to propagate unchecked and distort final decisions; (2) An imbalance between exploration and exploitation of domain-relevant internal knowledge, leading to repeating earlier findings rather than generating new cognitive insights. Building on these findings, we proposed a novel reinforcement learning method termed *Adaptive Entropy Policy Optimization (AEPO)*. Specifically, the AEPO framework consists of two major components: (1) Reflection-aware Information Filtration, which quantifies the cognitive information flow and prevents the final answer from being affected by earlier bad cognitive information; (2) Adaptive-Entropy Optimization, which dynamically balances exploration and exploitation across different reasoning stages, promoting both reflective diversity and answer correctness. Extensive experiments demonstrate that AEPO consistently achieves state-of-the-art performance over mainstream reinforcement learning baselines across diverse benchmarks. Our code is available at <https://anonymous.4open.science/r/AEPO-7F3A>.

1 Introduction

Large Language Models (LLMs) have significantly enhanced reasoning capabilities by enabling longer Chain-of-Thought processes and adopting sophisticated reasoning methodologies. A key technique driving these improvements is Reinforcement

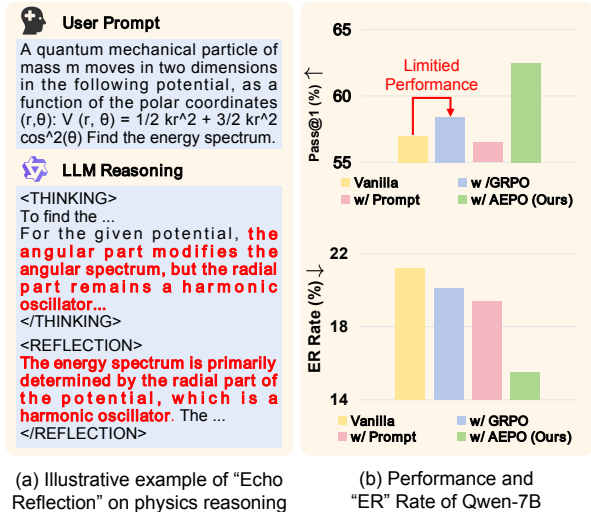


Figure 1: Illustrative examples of: (a) A typical instance of “Echo Reflection” in reasoning on a physics problem (with the echoed incorrect content highlighted in red). (b) Performance (upper figure) and corresponding Echo Reflection rate (lower figure) of Qwen-7B on knowledge-intensive benchmark MedQA.

Learning with Verifiable Rewards (RLVR) (Shao et al., 2024; Yu et al., 2025; Dai et al., 2025; Chen et al., 2025), where models optimize outputs based on RL objectives tied to automated thinking and reflective behavior.

Although these methods show promising performance, we observe consistent failures when applied to tasks involving complex domain-specific knowledge (e.g., biology, physics, chemistry reasoning). As shown in Figure 1(a), we present an example of a failed reasoning attempt by Qwen2.5-7B-Instruct (Team, 2024) on a physics problem. In the “Thinking” stage, the model erroneously states that “the angular part modifies the angular spectrum, but the radial part remains a harmonic oscillator”. In the subsequent “Reflection” stage, the model fails to revise this mistake and instead mechanically reiterates the same incorrect reasoning. Moreover, due to the narrow scope of the initial “Thinking” stage, alternative options are not critically evaluated in the

“Reflection” stage, leading to shallow analysis and limited knowledge utilization. This case illustrates that the model fails to fully take advantage of inherent knowledge during reasoning, thereby performing few genuine cognitive updates during reflection. We refer to the phenomenon of a model merely repeating earlier reasoning during the reflection stage without generating new insights as *Echo Reflection (ER)*. As illustrated in Figure 1(b), the upper image shows that compared with the vanilla Qwen-7B model, existing reinforcement learning methods (w/ GRPO) and explicitly prompting the model not to repeat erroneous content (w/ Prompt) yield only marginal performance gains on knowledge-intensive tasks. More importantly, the lower image reveals that these approaches fail to substantially reduce the occurrence of Echo Reflection. Both the GRPO-based method and the explicit prompting method lead to only minimal decreases in ER rate, indicating that erroneous reasoning patterns continue to persist during reflection. These findings reveal a fundamental limitation of current reinforcement learning frameworks in mitigating the ER phenomenon.

We attribute this failure to two main factors in the “Reflection” stage: (1) Uncontrolled propagation of prior errors, where incorrect intermediate steps are preserved; (2) An imbalance between exploration and exploitation of domain-relevant internal knowledge, which limits the model’s ability to critically assess and revise its initial output. Building on this, we propose a novel reinforcement learning algorithm, termed *Adaptive Entropy Policy Optimization (AEPO)*, designed to mitigate the Echo Reflection problem, thereby enhancing its reasoning capabilities in complex tasks. Specifically, as depicted in Figure 2, the AEPO framework consists of two major components: (1) Reflection-aware Information Filtration (RIF), inspired by Information Bottleneck (IB) theory (Tishby et al., 2000), aims to preserve task-relevant cognitive signals while suppressing misleading intermediate information during reasoning. It helps regulate the internal flow of information to prevent early-stage errors from contaminating final predictions. (2) Adaptive-Entropy Optimization (AEO), which dynamically regulates the policy entropy conditioned on exploration and exploitation balance.

Our contribution can be summarized as follows:

- We identify and characterize the “Echo Reflection” phenomenon, shedding light on a pre-

viously underexplored failure mode in LLM reasoning over knowledge-intensive tasks.

- We take a closer look at the ER from an information-theoretic perspective and propose Reflection-aware Information Filtration, a module grounded in Information Bottleneck theory that constrains the flow of cognitive information during reasoning.
- We introduce the AEPO algorithm, which enables the model to balance exploration and exploitation during reinforcement learning, encouraging meaningful revisions and deeper knowledge utilization during reflection.

2 Related Works

Reinforcement Learning in LLMs. Reinforcement learning from human feedback (RLHF) (Ouyang et al., 2022) has become a key paradigm for aligning large language models with human preferences. Recent advances (Schulman et al., 2017b; Shao et al., 2024; Ahmadian et al., 2024) have extended RLHF beyond instruction-following, introducing verifiable reward functions to better incentivize complex reasoning. OpenAI’s o1 (Jaech et al., 2024) was the first to demonstrate that RL can effectively elicit high-level reasoning abilities in large-scale LLMs. Building on this, models such as DeepSeek R1 (Guo et al., 2025), Qwen QwQ (Team, 2024), and Kimi k1.5 (Team et al., 2025) have sought to match or surpass o1’s performance. Concurrently, a series of studies, including OpenReasoner-Zero (Hu et al., 2025), SimpleRL-Zoo (Zeng et al., 2025), and Logic-RL, (Xie et al., 2025) have explored direct RL-based fine-tuning of base models, omitting the need for an intermediate supervised fine-tuning phase. Other approaches, such as Light-R1 (Wen et al., 2025) and DeepScaler (Meng et al., 2023), propose specially crafted cold-start datasets to encourage fine-grained, step-wise reasoning during early training. The SRPO (Zhang et al., 2025) framework further combines cold-start strategies with GRPO (Shao et al., 2024) to enhance deep reasoning development. In parallel, complementary methods such as VAPO (Yue et al., 2025), DAPO (Yu et al., 2025), and S-GRPO (Dai et al., 2025) aim to refine the GRPO framework by improving reward formulation and advantage estimation, thereby more effectively promoting complex reasoning behaviors in LLMs.

Policy Entropy in Reinforcement Learning. Stemmed in information theory, entropy has long

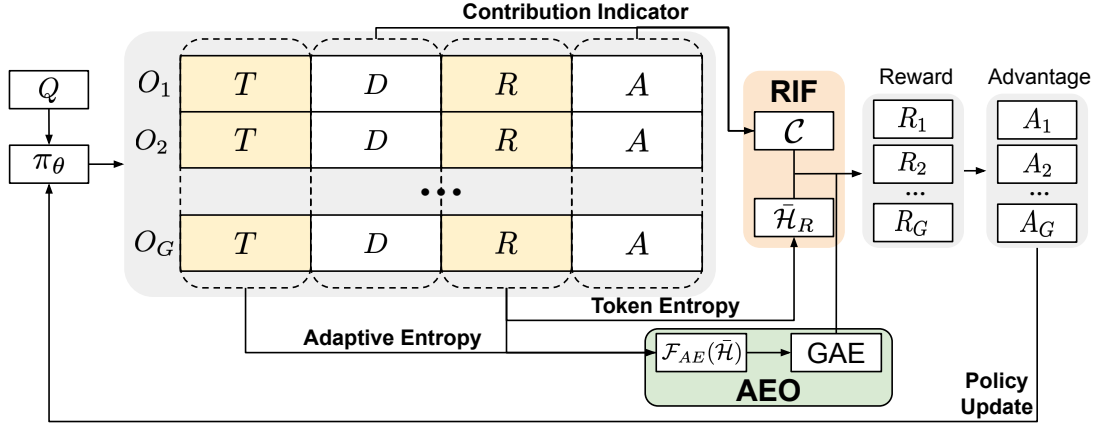


Figure 2: The overall framework of the proposed AEPO algorithm. It consists of two key components: 1) Reflection-aware Information Filtration (RIF), which leverages IB theory to constrain the flow of cognitive information; (2) Adaptive Entropy Optimization (AEO), which balances the behaviors of LLM exploration and exploitation.

166 been a core principle in reinforcement learning
167 (RL). Classical entropy-regularized reinforcement
168 learning (ERL) (Ziebart et al., 2008; Toussaint,
169 2009) frameworks are often coupled with the soft
170 Bellman update (Schulman et al., 2017a), discour-
171 aging overly deterministic policies. A canonical
172 example of this approach is the Soft Actor-Critic
173 (SAC) algorithm (Haarnoja et al., 2018), which ef-
174 fectively constrains policy updates by maintaining
175 proximity to a reference distribution. Recent work
176 has extended entropy-based methods to LLMs train-
177 ing. ETPO (Wen et al., 2024) applies token-level
178 ERL to enhance LLM performance in interactive
179 tasks, while EP-PRM (Zhang et al., 2024) incor-
180 porates entropy into process-level reward model-
181 ing. Beyond regularization, Cui et al. reveal that
182 entropy dynamics reflect the covariance between
183 action probabilities and logit shifts, offering a tool
184 to prevent policy collapse. Cheng et al. shows that
185 higher-entropy reasoning chains can lead to more
186 effective inference by generating longer responses.
187 More abstractly, entropy has been used to model
188 uncertainty at the semantic level. SEED-GRPO
189 (Chen et al., 2025) introduces prompt-level seman-
190 tic entropy to capture ambiguity in instructions,
191 and Wang et al. identify high-entropy tokens as key
192 “forks” during generation.

193 3 Proposed Method

194 3.1 Preliminary

195 Let \mathcal{D} denote the training dataset, where each data
196 point consists of a question–answer pair $\{Q, L\} \in$
197 \mathcal{D} . To optimize the large language model (LLM)
198 policy π_θ , given Q , we sample a group of outputs
199 $\mathcal{O} = \{O_i\}_{i=1}^G$ to estimate the advantage. Note

200 that $O_i = \{o_1, \dots, o_{|O_i|}\}$ is a complete response
201 sequence consisting of tokens.

202 We adopt a four-step response format to elicit
203 structured reasoning from the model. Specifically,
204 the model is prompted to sequentially: (1) think
205 about the question (thinking-stage), (2) propose
206 an initial response (draft-stage), (3) perform self-
207 reflection (reflection-stage), and (4) produce a fi-
208 nal answer (answer-stage). To simplify notation,
209 we also refer to the sampled output O_i as being
210 composed of four semantically coherent segments:
211 $O_i = [T, D, R, A]$, where each segment corre-
212 sponds to a distinct stage of reasoning. Note that
213 while O_i denotes the full token sequence gener-
214 ated by the model, $T, D, R,$ and A are contiguous
215 token subsequences that represent the outputs of each
216 respective step. The prompt used for the four-step
217 response format is detailed in the *Appendix*.

218 3.2 Reflection-aware Information Filtration

219 To prevent LLMs from failing to accurately identify
220 errors in previous reasoning during the reflection
221 process, the Reflection-aware Information Filtra-
222 tion module aims to optimize the model’s reflection
223 by constraining the information flow, suppressing
224 the retention of erroneous or redundant informa-
225 tion, and promoting the transmission and retention
226 of accurate and useful information.

227 **Information Bottleneck.** The Information Bot-
228 tleneck theory (Tishby et al., 2000) introduces an
229 intermediate variable Z , whose objective is to com-
230 press the input as much as possible while retaining
231 sufficient information to recover the target output.

232 As shown in Figure 3, in the context of LLM rea-
233 soning, Q along with T together serve as input. We
234 use $X = [Q, T]$ to represent them. The reflection

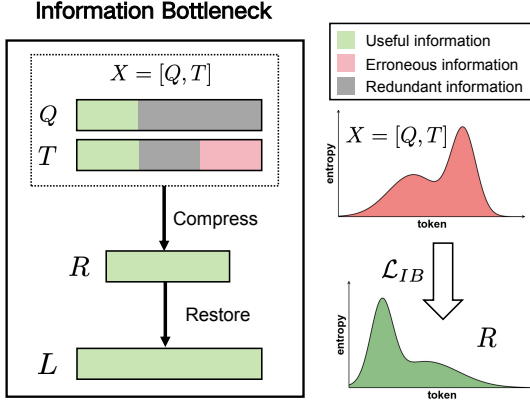


Figure 3: Information Bottleneck adopted in Reflection-aware Information Filtration module. The Information Bottleneck aims to suppress erroneous information and redundant information flow (left). By minimizing \mathcal{L}_{IB} , the token entropy is reduced (right).

R serves as the intermediate representation, and the groundtruth answer L as the output.

IB theory formulates this trade-off using mutual information (MI) to quantify statistical dependence. The MI between X and R is denoted as $I(X; R)$, which measures how much information about X is preserved in R . A higher mutual Information indicates that more of the X can be reconstructed for R , whereas a lower value suggests that R is a compressed, less redundant encoding of X . However, minimizing $I(X; R)$ alone can lead to loss of task-relevant information. To address this, IB theory simultaneously maximizes $I(R; L)$, which is the MI between L and R . In this manner, the IB theory ensures the retained information in R remains sufficient for accurately predicting L .

To this end, the optimization objective is:

$$\mathcal{L}_{IB} = I(X; R) - \beta \cdot I(R; L), \quad (1)$$

where β is a hyper-parameter.

Under this formulation, minimizing $I(X; R)$ encourages the reflection to filter out redundant or erroneous content, reducing mechanical repetition of earlier reasoning. Concurrently, maximizing $I(R; L)$ promotes the retention of task-relevant information within the reflection, thereby alleviating the ‘‘Echo Reflection’’ phenomenon.

Proxy Metrics. While the IB objective provides a principled framework for guiding information flow during reflection, directly computing mutual information terms $I(X; R)$ and $I(R; L)$ is intractable in high-dimensional generative models such as LLMs (Alemi et al., 2017; Shwartz-Ziv and Tishby, 2022). To this end, we propose the Proxy Metrics for an efficient estimate of the mutual information.

By definition, the mutual information $I(X; R)$ is given by $I(X; R) = \mathcal{H}(X) - \mathcal{H}(R|X)$, where $\mathcal{H}(\cdot)$ denotes entropy. Under a fixed X at the reflection stage, $\mathcal{H}(X)$ can be regarded as a constant and therefore minimizing $I(X; R)$ leads to maximize $\mathcal{H}(R|X)$. Specifically, the token-level entropy of the reflection step is denoted as:

$$\mathcal{H}_{o_i} = - \sum_{v \in \mathcal{V}} \pi_{\theta}(v|q, o_{<i}) \log \pi_{\theta}(v|q, o_{<i}). \quad (2)$$

Hence, the average policy entropy $\bar{\mathcal{H}}_R$ over the reflection stage is:

$$\bar{\mathcal{H}}_R = \frac{1}{|R|} \sum_{y_i \in R} \mathcal{H}_{o_i}. \quad (3)$$

The $\bar{\mathcal{H}}_R$, measures the information capacity of the reflection received from the thinking. Lower entropy implies higher predictability of the reflection given the initial reasoning, indicating stronger dependence (i.e., higher mutual information). Thus, we interpret higher entropy as a signal of reduced dependence on T .

To estimate $I(R; L)$, we define a heuristic contribution indicator \mathcal{C} to measure how much the reflection improves the likelihood of producing a correct answer. Higher \mathcal{C} reflects greater task-relevant information retained in R , corresponding to higher $I(R; L)$. Hence \mathcal{C} is a proxy of $I(R; L)$:

$$\mathcal{C} = \begin{cases} \alpha_{pre}, & \text{if } D = L \text{ and } A = L \\ \alpha_{fix}, & \text{if } D \neq L \text{ and } A = L \\ 0, & \text{if } D \neq L \text{ and } A \neq L \\ -\alpha_{reg}, & \text{if } D = L \text{ and } A \neq L \end{cases} \quad (4)$$

where α_{pre} rewards preserving a correct draft, α_{fix} rewards fixing an incorrect draft, and α_{reg} penalizes regressions from correct to incorrect; we enforce $\alpha_{fix} > \alpha_{pre} > 0$ and $\alpha_{reg} > 0$.

3.3 Adaptive Entropy Optimization

Another key factor that leads to ER phenomenon is the insufficient exploration of internal knowledge during reasoning. To tackle this problem, we adopt the Adaptive Entropy Optimization mechanism to balance the model’s exploration and exploitation. **Adaptive Entropy.** To quantify the degree of exploration at each stage of reasoning, we compute the average token-level entropy for the thinking-stage and reflection stage with Equation 2, yielding $\bar{\mathcal{H}}_T$ and $\bar{\mathcal{H}}_R$, respectively.

A common practice in reinforcement learning is to inject entropy into the Bellman backup to encourage exploration. Following this, one might directly maximize $\bar{\mathcal{H}}_T$ and $\bar{\mathcal{H}}_R$. However, excessive entropy can lead to incoherent reasoning by promoting erroneous decision-making, while minimizing entropy may improve precision but at the cost of stifling exploration and creativity.

To balance these extremes, we aim to maintain entropy near a task-suitable target \mathcal{H}^* . Based on this, we adopt an adaptive entropy mechanism as:

$$\mathcal{F}_{AE}(\bar{\mathcal{H}}) = -\|\bar{\mathcal{H}} - \mathcal{H}^*\|_2, \quad (5)$$

where $\bar{\mathcal{H}} \in \{\bar{\mathcal{H}}_T, \bar{\mathcal{H}}_R\}$. This adaptive mechanism guides both thinking and reflection stages to balance exploration and exploitation, respectively.

Gated Adaptive Entropy. While exploration fosters diversity, it also introduces the risk of incorrect reasoning. To address this, we use a Gated Adaptive Entropy (GAE) mechanism, ensuring that entropy rewards for reflection are only applied when the output is correct, thereby gating them by the final correctness.

$$\mathcal{F}_{GAE} = \sum_{\bar{\mathcal{H}} \in \{\bar{\mathcal{H}}_T, \bar{\mathcal{H}}_R\}} \mathcal{F}_{AE}(\bar{\mathcal{H}}) * \mathbb{1}_{\text{correct}}. \quad (6)$$

Minimizing $\mathcal{L}(O_i)$ jointly minimizes the Information Bottleneck term \mathcal{L}_{IB} , which suppresses redundant or misleading information during reflection, and maximizes \mathcal{F}_{GAE} , introducing a preference for controlled entropy dynamics. Hence the total objective for output O_i is:

$$\mathcal{L}(O_i) = \mathcal{L}_{IB} - \mathcal{F}_{GAE}. \quad (7)$$

Formulation of AEPO. Following previous works (Yu et al., 2025; Guo et al., 2025), we use token-level policy gradient loss, which is denoted as:

$$\mathcal{L}_{\text{AEPO}}(\theta) = \mathbb{E}_{(Q,L) \sim \mathcal{D}, \{O_i\}_{i=1}^G \sim \pi_{\theta_{\text{old}}}(\cdot|Q)} \left[\frac{1}{\sum_{i=1}^G |O_i|} \sum_{i=1}^G \left[\mathcal{L}(O_i) + \sum_{t=1}^{|O_i|} \min(r_{i,t}(\theta) \hat{A}_{i,t}, \text{clip}(i,t) \hat{A}_{i,t}) \right] \right], \quad (8)$$

where

$$r_{i,t}(\theta) = \frac{\pi_{\theta}(o_{i,t} | Q, O_{i,<t})}{\pi_{\theta_{\text{old}}}(o_{i,t} | Q, o_{i,<t})},$$

$$\hat{A}_{i,t} = \frac{R_i - \text{mean}(\{R_i\}_{i=1}^G)}{\text{std}(\{R_i\}_{i=1}^G)}, \quad (9)$$

$$\text{clip}(i,t) = \text{clip}(r_{i,t}(\theta), 1 - \epsilon_{\text{low}}, 1 + \epsilon_{\text{high}}).$$

4 Experiments

4.1 Experimental Setup

Datasets. We evaluate the proposed AEPO framework on both in-distribution (I.D.) and out-of-distribution (O.O.D.) datasets. The I.D. datasets include: (1) MedQA (Jin et al., 2021), which contains 10,178 training and 1,273 testing QA pairs from professional medical board exams; (2) MedMcQA (Pal et al., 2022), which contains 182,822 training and 4,183 testing QA pairs from real-world medical entrance exams. The O.O.D datasets include: (1) MMLU-Pro (Wang et al., 2024), which is a multi-task language understanding benchmark comprising 12,000 questions spanning diverse academic domains; (2) GPQA (Rein et al., 2024), which is a subject-specific QA dataset covering biology, physics, and chemistry domains. (3) MATH-500, which contains a subset of 500 problems from the MATH benchmark that OpenAI (Lightman et al., 2023). (4) AIME 24, which contains problems from the American Invitational Mathematics Examination 2024. (5) AMC 23, which is a mathematical dataset with 40 problems.

Training Details. All models are trained on 4 Nvidia A6000 GPU, using the EasyR1 framework (Zheng et al., 2025). We adopt Qwen2.5-7B-Instruct (Team, 2024) (denoted as Qwen-7B) and LLaMA3-8B-Instruct (Dubey et al., 2024) (denoted as LLaMA-8B) as base models for most experiments. Training is conducted with bfloat16 precision using the AdamW optimizer, a constant learning rate of 1×10^{-6} , and a linear warm-up over the first 10 steps. Each batch contains 64 prompts, with 5 responses sampled per prompt. For language model decoding, we set the temperature to 1.0 and top-p to 0.99. Following DAPO (Yu et al., 2025), we use clipped policy ratios with $\epsilon_{\text{low}} = 0.2$ and $\epsilon_{\text{high}} = 0.28$. We set $\alpha_{\text{pre}}=0.4$, $\alpha_{\text{fix}}=0.6$, $\alpha_{\text{reg}}=0.3$, and use $\beta=1$, $\mathcal{H}^*=0.67$ in all experiments.

Inference Details. We report pass@1 on both I.D. and O.O.D. datasets, where pass@1 is defined as the proportion of questions for which the

Model	#Params	MedQA	MedMcQA
Mistral	7B	48.2	44.6
Yi-1.5	9B	50.8	48.7
LLaMA-3.1	8B	58.7	56.0
LLaMA-3.1 + CoT	8B	57.7	56.6
LLaMA-3.1 + GRPO	8B	63.8	57.4
LLaMA-3.1 + DAPO	8B	62.9	59.8
Qwen2.5	7B	57.0	55.6
Qwen2.5 + CoT	7B	57.7	56.4
Qwen2.5 + GRPO	7B	58.4	56.2
Qwen2.5 + DAPO	7B	57.1	57.4
AEPO-L (Ours)	8B	68.5	64.3
AEPO-Q (Ours)	7B	62.5	61.6

Table 1: Comparison results (Pass@1) on MedQA and MedMcQA. AEPO-L and AEPO-Q denote LLaMA-3.1 w/ AEPO and Qwen2.5 w/ AEPO, respectively.

model’s single generated answer exactly matches the ground-truth answer. In the I.D. setting, models are trained on the training split and evaluated on the corresponding test split. In the O.O.D. setting, we directly evaluate the I.D.-trained models on unseen O.O.D. datasets without further adaptation. For inference stability, the decoding temperature is fixed at 0 during inference.

4.2 Overall Comparison Results

Comparisons On In-Distribution Dataset. To evaluate the effectiveness of our proposed AEPO method in enhancing LLM reasoning in knowledge-intensive question answering, we compare AEPO with mainstream reinforcement learning with verifiable reward (RLVR) approaches across multiple LLM backbones. Answer accuracy is reported on both the MedQA and MedMcQA benchmarks, as shown in Table 1, respectively. From the experimental results, we can observe that: (1) AEPO consistently outperforms existing RLVR baselines, including GRPO and DAPO, demonstrating its effectiveness in mitigating the Echoed Reflection phenomenon during multi-stage inference; (2) Specifically, our AEPO outperforms the DAPO by 5.4% and 5.6% on Qwen-7B and LLaMA-8B, respectively. These remarkable results demonstrate that our proposed AEPO method effectively enhances reasoning by suppressing incorrect reasoning in the thinking stage and promotes more accurate final answers. (3) Moreover, AEPO achieves superior performance across both MedQA and MedMcQA, indicating its stronger capacity to explore and leverage the internal knowledge embedded within large language models.

Comparisons On Out-of-Distribution Dataset. To further evaluate the generalization capability of our model in handling O.O.D. problems, we con-

duct additional experiments on unseen datasets and compare our method with state-of-the-art (SOTA) baselines. Specifically, as shown in Figure 4, subfigure (a) is the experimental results of LLaMA-8B trained on the MedMcQA dataset, and subfigure (b) is the experimental results of Qwen-7B trained on MedQA. We directly validate the O.O.D. **without** additional training. The results reveal the following insights: (1) Our proposed AEPO method significantly outperforms existing SOTA methods on both domains, demonstrating stronger generalization to unfamiliar problem types. (2) We attribute this improved generalization to the Adaptive Entropy Optimization module, which enables deeper and more efficient exploration of the solution space while avoiding policy collapse. (3) Moreover, AEPO effectively filters out irrelevant or misleading information during open-ended reasoning, enabling the model to maintain coherent and accurate inference paths even under high uncertainty.

4.3 Further Analysis

Ablation Study. As shown in Table 2, we conducted ablation studies on both MedQA and MedMcQA. The experimental results reveal the following key observations: (1) The comparison between setting No.0 and No.1 demonstrates the effectiveness of the proposed RIF module. By constraining the cognitive information flow and suppressing the influence of erroneous or redundant intermediate reasoning, RIF significantly improves the model’s ability to generate correct answers. (2) The performance gap between settings No.1 and No.6 highlights the impact of Adaptive Entropy Optimization. By dynamically regulating the policy entropy in the thinking stage, AEO enables deeper yet controlled exploration. When combined with RIF, this leads to notably improved reasoning performance. (3) The comparison between settings No.4 and No.7 reveals the importance of the Correctness-Gated Entropy Reward mechanism. By conditioning entropy-based exploration rewards on the correctness of the final answer, this component helps ensure that exploratory behaviors contribute meaningfully to reasoning quality.

Comparisons On Finetuned Model. To comprehensively evaluate the effectiveness of the proposed AEPO in leveraging domain-specific knowledge, we conducted additional experiments on HuatuoGPT-O1-8B (Chen et al., 2024), which has been fine-tuned specifically for the medical domain. As shown in Table 3, the following observations

No.	RIF	AEO		MedQA		MedMcQA	
		AE	GAE	Llama	Qwen	LLaMA	Qwen
0	-	-	-	58.7	57.0	56.0	55.6
1	✓	-	-	62.3	58.4	56.6	56.8
2	-	✓	-	63.7	58.6	57.9	57.8
3	-	-	✓	59.4	57.8	57.4	57.4
4	✓	✓	-	65.8	60.7	59.3	58.8
5	-	-	✓	66.0	61.2	61.5	59.2
6	✓	-	✓	65.4	60.2	59.4	58.6
7	✓	✓	✓	68.5	62.5	64.3	61.6

Table 2: Ablation Study on MedQA and MedMcQA.

Dataset	MedQA	MedMcQA
Baseline	72.6	60.4
w/ GRPO	72.4	57.3
w/ DAPO	71.2	61.7
w/ AEPO (Ours)	75.4	63.9

Table 3: Results on Domain Specific Finetune.

Dataset	MedQA	MedMcQA
Baseline	58.4	56.5
w/ GRPO	59.2	58.2
w/ DAPO	62.4	60.1
w/ AEPO (Ours)	68.5	63.7

Table 4: Results on Additional Prompts.

can be made: (1) Incorporating domain-specific knowledge leads to significant performance gains for HuatuoGPT-O1-8B over the base models. However, existing RLVR methods show limited additional improvements, indicating their inability to effectively utilize domain-specific information. (2) In contrast, our AEPO method consistently outperforms both HuatuoGPT variants by 3.0% and 4.2% on MedQA, demonstrating its superior capacity to exploit internal knowledge. We attribute this to the AEO module, which facilitates deeper and more controlled exploration of the solution space while mitigating the risk of policy collapse.

Analysis on Additional Prompts. To further demonstrate the robustness of our AEPO method across different Chain-of-Thoughts (CoTs), we conduct additional experiments using alternative prompts. In addition to the four-stage reasoning prompt introduced in our main framework, we also adopt an R1-style CoT prompt following Guo et al. (2025). The prompt is provided in the Appendix. Experimental results show that under the R1-style CoT setting, AEPO significantly outperforms the GRPO baseline. It highlights the prompt-robustness of AEPO and its strong transferability across different reasoning paradigms.

Analysis on Creativity Index. To highlight the influence of our proposed RIF module, we visualize the Creativity Index (Lu et al., 2024) of LLaMA models on the MedMcQA dataset under different RL strategies using violin plots. The CI measures the creativity of model outputs, where higher values indicate more novel generations, while lower val-

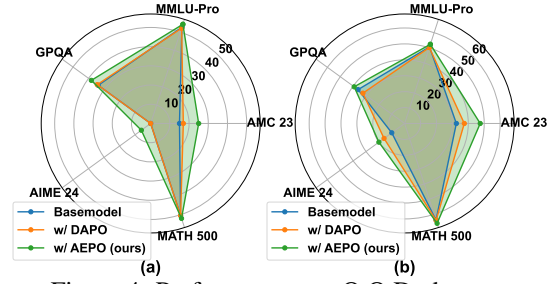


Figure 4: Performances on O.O.D. datasets.

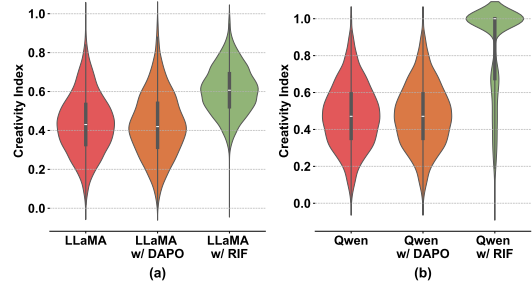


Figure 5: Violin plot of creativity index on MedMcQA.

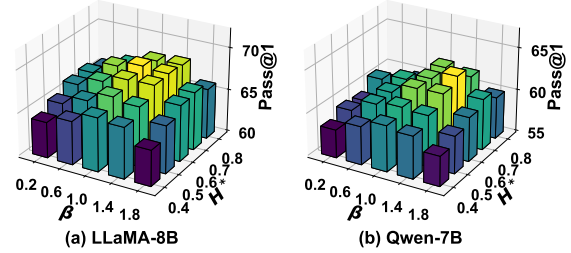


Figure 6: Visualization of hyperparameter sensitivity analysis of β and \mathcal{H}^* on MedQA dataset.

ues suggest a tendency to copy from the reference corpus. In our setting, we construct the reference corpus by concatenating all the input questions with the model’s response of the thinking-stage, and then evaluate the CI of the corresponding reflection stage. The results reveal that both LLaMA-8B w/ RIF and Qwen-7B w/ RIF achieve substantially higher Creativity Index scores compared to their counterparts with DAPO. This suggests that the RIF module effectively suppresses the influence of prior thinking content during reflection, thereby mitigating the Echo Reflection phenomenon and encouraging more original reflective reasoning.

Analysis on Hyperparameter Sensitivity. As shown in Figure 6, we study the impact of the two key hyperparameters, β and \mathcal{H}^* on the MedQA dataset. Note that β and $[\alpha_{pre}, \alpha_{fix}, \alpha_{reg}]$ are coupled; thus, we fix $[\alpha_{pre}, \alpha_{fix}, \alpha_{reg}]$ and only analyze sensitivity to β to avoid redundant hyperparameter sweeps. Several observations can be made: First, small β leads to overly aggressive compression in the reflection stage, suppressing task-relevant information, whereas large β allows erroneous or redun-

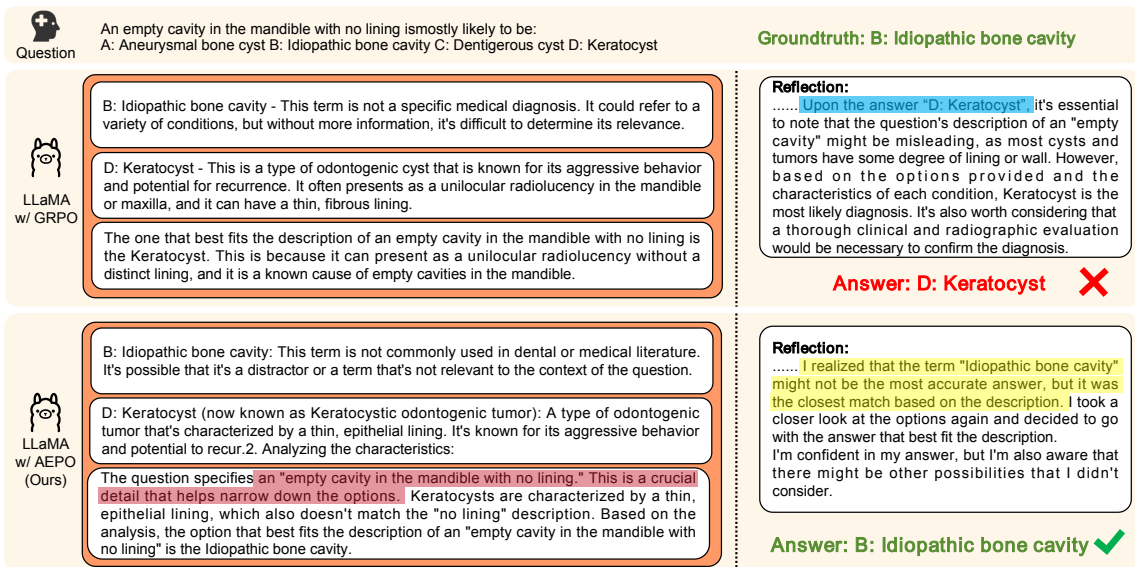


Figure 7: Quality visualization analysis. It presents responses of LLaMA w/GRPO and our proposed LLaMA w/AEPO. The left side of the dashed line illustrates the content of the model’s thinking-stage, while the right side displays the content of the reflection-stage and the final answer. Key contents are highlighted in different colors.

532 dant reasoning signals to propagate. This trend is
533 consistent across both LLaMA-8B and Qwen-7B,
534 highlighting the importance of balancing informa-
535 tion compression and preservation. Second, \mathcal{H}^*
536 plays a crucial role in regulating exploration during
537 reasoning. Low \mathcal{H}^* induces overly deterministic
538 policies, while excessively high \mathcal{H}^* encourages
539 uncontrolled exploration, resulting in noisy or inco-
540 herent reasoning trajectories. Finally, we observe
541 a broad region of high performance in the (β, \mathcal{H}^*)
542 space, indicating that AEPO is robust to reasonable
543 hyperparameter variations.

544 **Qualitative Analysis.** As shown in Figure 7,
545 we present a qualitative analysis that compares
546 the question answer results from the MedMcQA
547 dataset. For clarity, we only visualize the criti-
548 cal reasoning contents leading to divergent deci-
549 sions between LLaMA w/ GRPO and our LLaMA
550 w/ AEPO. The visualization reveals that, during
551 the thinking stage, LLaMA w/ AEPO conducts a
552 more refined analysis of the options. Notably, when
553 evaluating the correct option B, LLaMA w/ AEPO
554 successfully connects “bone cavity” and “no lin-
555 ing” (highlighted with a red rectangle), enabling
556 a correct judgment. This capability is largely at-
557 tributed to our proposed AEO module, which dy-
558 namically adjusts entropy to balance exploration
559 and exploitation. Moreover, during the reflection
560 phase, LLaMA w/ AEPO demonstrates enhanced
561 metacognitive awareness. For instance, it explicitly
562 acknowledges: “Idiopathic bone cavity might not
563 be the most accurate answer” (highlighted with a

564 blue rectangle). This indicates that the model main-
565 tains critical awareness of terminology, avoiding
566 cognitive fixation. In contrast, LLaMA w/ GRPO
567 makes an incorrect judgment early in reasoning
568 and, during reflection, fails to correct the error, in-
569 stead reiterating its flawed reasoning (highlighted
570 with a yellow rectangle). This highlights the vulner-
571 ability of conventional frameworks to error prop-
572 agation and entrenchment. Our Reflection-aware
573 Information Filtering module mitigates such Echo
574 Reflection by selectively filtering and reconstruct-
575 ing intermediate reasoning information. More qual-
576 itative analysis on the medical domain and other
577 domains can be found in the *Appendix*.

578 5 Conclusion

579 In this paper, we identified the Echo Reflection
580 phenomenon, where language models fail to per-
581 form meaningful cognitive updates during reflec-
582 tion. Through step-wise policy entropy analysis,
583 we found that low entropy suppresses the retrieval
584 and utilization of inherent knowledge, leading to
585 ER. To address this, we proposed Adaptive Entropy
586 Policy Optimization (AEPO), which enhances rea-
587 soning with two key components: Reflection-aware
588 Information Filtration, guided by information bot-
589 tleneck theory, and Adaptive Entropy Optimization,
590 formulated via dynamic entropy adjustment. To-
591 gether, they effectively mitigate ER by promoting
592 more informative and adaptive reasoning processes.
593 In the future, we will explore additional strategies
594 to further improve reasoning effectiveness.

595 Limitation

596 While the proposed AEPO framework demon-
597 strates strong empirical performance in mitigating
598 the Echo Reflection phenomenon, several limita-
599 tions merit discussion. (1) Although AEPO is con-
600 ceptually grounded in the IB principle, it relies on
601 proxy metrics rather than direct mutual informa-
602 tion estimation, which is currently infeasible for
603 LLMs. Identifying more expressive yet computa-
604 tionally viable proxies remains an important direc-
605 tion for future work. (2) AEPO improves reflection
606 quality indirectly through entropy regulation and
607 information filtering, rather than explicitly model-
608 ing erroneous beliefs or structured cognitive states;
609 exploring tighter connections between entropy and
610 explicit cognitive modeling may further enhance
611 interpretability and theoretical grounding.

612 References

613 Arash Ahmadian, Chris Cremer, Matthias Gallé,
614 Marzieh Fadaee, Julia Kreutzer, Olivier Pietquin,
615 Ahmet Üstün, and Sara Hooker. 2024. Back to ba-
616 sics: Revisiting reinforce style optimization for learn-
617 ing from human feedback in llms. *arXiv preprint*
618 *arXiv:2402.14740*.

619 Alexander A Alemi, Ian Fischer, Joshua V Dillon, and
620 Kevin Murphy. 2017. Deep variational information
621 bottleneck. In *International Conference on Learning*
622 *Representations*.

623 Junying Chen, Zhenyang Cai, Ke Ji, Xidong Wang,
624 Wanlong Liu, Rongsheng Wang, Jianye Hou, and
625 Benyou Wang. 2024. Huatuogpt-o1, towards med-
626 ical complex reasoning with llms. *arXiv preprint*
627 *arXiv:2412.18925*.

628 Minghan Chen, Guikun Chen, Wenguan Wang, and
629 Yi Yang. 2025. Seed-grpo: Semantic entropy en-
630 hanced grpo for uncertainty-aware policy optimiza-
631 tion. *arXiv preprint arXiv:2505.12346*.

632 Daixuan Cheng, Shaohan Huang, Xuekai Zhu, Bo Dai,
633 Wayne Xin Zhao, Zhenliang Zhang, and Furu Wei.
634 2025. Reasoning with exploration: An entropy per-
635 spective. *arXiv preprint arXiv:2506.14758*.

636 Ganqu Cui, Yuchen Zhang, Jiacheng Chen, Lifan
637 Yuan, Zhi Wang, Yuxin Zuo, Haozhan Li, Yuchen
638 Fan, Huayu Chen, Weize Chen, and 1 others. 2025.
639 The entropy mechanism of reinforcement learning
640 for reasoning language models. *arXiv preprint*
641 *arXiv:2505.22617*.

642 Muzhi Dai, Chenxu Yang, and Qingyi Si. 2025. S-grpo:
643 Early exit via reinforcement learning in reasoning
644 models. *arXiv preprint arXiv:2505.07686*.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, 645
Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, 646
Akhil Mathur, Alan Schelten, Amy Yang, Angela 647
Fan, and 1 others. 2024. The llama 3 herd of models. 648
arXiv e-prints, pages arXiv-2407. 649

Daya Guo, Dejian Yang, Haowei Zhang, Junxiao 650
Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shi- 651
rong Ma, Peiyi Wang, Xiao Bi, and 1 others. 2025. 652
Deepseek-r1: Incentivizing reasoning capability in 653
llms via reinforcement learning. *arXiv preprint* 654
arXiv:2501.12948. 655

Tuomas Haarnoja, Aurick Zhou, Pieter Abbeel, and 656
Sergey Levine. 2018. Soft actor-critic: Off-policy 657
maximum entropy deep reinforcement learning with 658
a stochastic actor. In *International conference on* 659
machine learning, pages 1861–1870. Pmlr. 660

Jingcheng Hu, Yinmin Zhang, Qi Han, Daxin Jiang, Xi- 661
angyu Zhang, and Heung-Yeung Shum. 2025. Open- 662
reasoner-zero: An open source approach to scaling 663
up reinforcement learning on the base model. *arXiv* 664
preprint arXiv:2503.24290. 665

Aaron Jaech, Adam Kalai, Adam Lerer, Adam Richard- 666
son, Ahmed El-Kishky, Aiden Low, Alec Helyar, 667
Aleksander Madry, Alex Beutel, Alex Carney, and 1 668
others. 2024. Openai o1 system card. *arXiv preprint* 669
arXiv:2412.16720. 670

Di Jin, Eileen Pan, Nassim Oufattole, Wei-Hung Weng, 671
Hanyi Fang, and Peter Szolovits. 2021. What disease 672
does this patient have? a large-scale open domain 673
question answering dataset from medical exams. *Ap- 674*
plied Sciences, 11(14):6421. 675

Qiao Jin, Bhuwan Dhingra, Zhengping Liu, William W 676
Cohen, and Xinghua Lu. 2019. Pubmedqa: A dataset 677
for biomedical research question answering. *arXiv* 678
preprint arXiv:1909.06146. 679

Hunter Lightman, Vineet Kosaraju, Yuri Burda, Harri- 680
son Edwards, Bowen Baker, Teddy Lee, Jan Leike, 681
John Schulman, Ilya Sutskever, and Karl Cobbe. 682
2023. Let’s verify step by step. In *The Twelfth Inter- 683*
national Conference on Learning Representations. 684

Ximing Lu, Melanie Sclar, Skyler Hallinan, Niloofar 685
Miresghallah, Jiacheng Liu, Seungju Han, Allyson 686
Ettinger, Liwei Jiang, Khyathi Chandu, Nouha Dziri, 687
and 1 others. 2024. Ai as humanity’s salieri: Quan- 688
tifying linguistic creativity of language models via 689
systematic attribution of machine text against web 690
text. 691

Chunyang Meng, Shijie Song, Haogang Tong, Maolin 692
Pan, and Yang Yu. 2023. Deepscaler: Holistic au- 693
toscaling for microservices based on spatiotemporal 694
gnn with adaptive graph learning. In *2023 38th* 695
IEEE/ACM International Conference on Automated 696
Software Engineering (ASE), pages 53–65. IEEE. 697

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, 698
Carroll Wainwright, Pamela Mishkin, Chong Zhang, 699
Sandhini Agarwal, Katarina Slama, Alex Ray, and 1 700

701	others. 2022. Training language models to follow instructions with human feedback. <i>Advances in neural information processing systems</i> , 35:27730–27744.	755
702		756
703		757
704	Ankit Pal, Logesh Kumar Umapathi, and Malaikanan Sankarasubbu. 2022. Medmcqa: A large-scale multi-subject multi-choice dataset for medical domain question answering. In <i>Conference on health, inference, and learning</i> , pages 248–260. PMLR.	758
705		759
706		760
707		761
708		762
709	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 <i>First Conference on Language Modeling</i> .	763
710		764
711		765
712		766
713		
714	John Schulman, Xi Chen, and Pieter Abbeel. 2017a. Equivalence between policy gradients and soft q-learning. <i>arXiv preprint arXiv:1704.06440</i> .	767
715		768
716		769
717	John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. 2017b. Proximal policy optimization algorithms. <i>arXiv preprint arXiv:1707.06347</i> .	770
718		771
719		
720		
721	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. <i>arXiv preprint arXiv:2402.03300</i> .	772
722		773
723		774
724		775
725		776
726		
727	Ravid Shwartz-Ziv and Naftali Tishby. 2022. Opening the black box of deep neural networks via information. <i>Information Flow in Deep Neural Networks</i> , page 24.	777
728		778
729		779
730		780
731	Kimi Team, Angang Du, Bofei Gao, BOWEI XING, Changjiu Jiang, Cheng Chen, Cheng Li, Chenjun Xiao, Chenzhuang Du, Chonghua Liao, and 1 others. 2025. Kimi k1. 5: Scaling reinforcement learning with llms. <i>arXiv preprint arXiv:2501.12599</i> .	781
732		782
733		783
734		784
735		785
736	Qwen Team. 2024. Qwen2 technical report. <i>arXiv preprint arXiv:2407.10671</i> .	786
737		787
738	Naftali Tishby, Fernando C Pereira, and William Bialek. 2000. The information bottleneck method. <i>arXiv preprint physics/0004057</i> .	788
739		789
740		790
741	Marc Toussaint. 2009. Robot trajectory optimization using approximate inference. In <i>Proceedings of the 26th annual international conference on machine learning</i> , pages 1049–1056.	791
742		792
743		
744		
745	Shenzhi Wang, Le Yu, Chang Gao, Chujie Zheng, Shixuan Liu, Rui Lu, Kai Dang, Xionghui Chen, Jianxin Yang, Zhenru Zhang, and 1 others. 2025. Beyond the 80/20 rule: High-entropy minority tokens drive effective reinforcement learning for llm reasoning. <i>arXiv preprint arXiv:2506.01939</i> .	793
746		794
747		795
748		796
749		797
750		798
751	Yubo Wang, Xueguang Ma, Ge Zhang, Yuansheng Ni, Abhranil Chandra, Shiguang Guo, Weiming Ren, Aaran Arulraj, Xuan He, Ziyang Jiang, and 1 others. 2024. Mmlu-pro: A more robust and challenging	799
752		800
753		801
754		802
	multi-task language understanding benchmark. <i>Advances in Neural Information Processing Systems</i> , 37:95266–95290.	803
		804
		805
		806
	Liang Wen, Yunke Cai, Fenrui Xiao, Xin He, Qi An, Zhenyu Duan, Yimin Du, Junchen Liu, Lifu Tang, Xiaowei Lv, and 1 others. 2025. Light-rl: Curriculum sft, dpo and rl for long cot from scratch and beyond. <i>arXiv preprint arXiv:2503.10460</i> .	
	Muning Wen, Cheng Deng, Jun Wang, Weinan Zhang, and Ying Wen. 2024. Entropy-regularized token-level policy optimization for large language models. <i>CoRR</i> .	
	Tian Xie, Zitian Gao, Qingnan Ren, Haoming Luo, Yuqian Hong, Bryan Dai, Joey Zhou, Kai Qiu, Zhirong Wu, and Chong Luo. 2025. Logic-rl: Unleashing llm reasoning with rule-based reinforcement learning. <i>arXiv preprint arXiv:2502.14768</i> .	
	Qiyang 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. <i>arXiv preprint arXiv:2503.14476</i> .	
	Yu Yue, Yufeng Yuan, Qiyang Yu, Xiaochen Zuo, Ruofei Zhu, Wenyuan Xu, Jiaye Chen, Chengyi Wang, TianTian Fan, Zhengyin Du, and 1 others. 2025. Vapo: Efficient and reliable reinforcement learning for advanced reasoning tasks. <i>arXiv preprint arXiv:2504.05118</i> .	
	Weihao Zeng, Yuzhen Huang, Qian Liu, Wei Liu, Keqing He, Zejun Ma, and Junxian He. 2025. Simplerl-zoo: Investigating and taming zero reinforcement learning for open base models in the wild. <i>arXiv preprint arXiv:2503.18892</i> .	
	Hanning Zhang, Pengcheng Wang, Shizhe Diao, Yong Lin, Rui Pan, Hanze Dong, Dylan Zhang, Pavlo Molchanov, and Tong Zhang. 2024. Entropy-regularized process reward model. <i>arXiv preprint arXiv:2412.11006</i> .	
	Xiaojiang Zhang, Jinghui Wang, Zifei Cheng, Wenhao Zhuang, Zheng Lin, Minglei Zhang, Shaojie Wang, Yinghan Cui, Chao Wang, Junyi Peng, and 1 others. 2025. Srpo: A cross-domain implementation of large-scale reinforcement learning on llm. <i>arXiv preprint arXiv:2504.14286</i> .	
	Yaowei Zheng, Juntao Lu, Shenzhi Wang, Zhangchi Feng, Dongdong Kuang, and Yuwen Xiong. 2025. Easyrl: An efficient, scalable, multi-modality rl training framework.	
	Brian D Ziebart, Andrew L Maas, J Andrew Bagnell, Anind K Dey, and 1 others. 2008. Maximum entropy inverse reinforcement learning. In <i>Aaai</i> , volume 8, pages 1433–1438. Chicago, IL, USA.	

A Overview

B More Detailed Implementation

We provide a more detailed implementation, including hardware platform and prompts, which is difficult to mention due to space limits.

B.1 Implementation Detail

All experiments were conducted on a server equipped with 4 NVIDIA A6000 GPUs on Ubuntu 22.04.5 LTS platform with EaryR1 framework (Zheng et al., 2025). We used NVIDIA driver version 535.247.01 with CUDA 12.2. We set all random seeds to 1 to reduce the impact of stochasticity across runs.

The overall training pipeline of the AEPO method is shown in Algorithm 1.

B.2 Prompt

The prompt used in 4-step response and R1-like (Shao et al., 2024) response is shown in Figure 8 and Figure 9, respectively.

4-Step Response

You are a senior expert in the medical field. Your task is to accurately answer the given medical question through careful analysis, thorough reasoning, and by integrating relevant medical knowledge with a clear understanding of the question.

You must follow a 4-step reasoning process: First, THINK through the reasoning and enclose it within <THINKING></THINKING> tags. Second, TRY to answer and wrap it in \boxed{}. Third, REFLECT on your answer and enclose it within <REFLECTION></REFLECTION> tags. Finally, give your FINAL answer, also wrapped in \boxed{}.

Figure 8: 4-step Response Prompt.

R1-like Response

You are a senior expert in the medical field. Your task is to accurately answer the given medical question through careful analysis, thorough reasoning, and by integrating relevant medical knowledge with a clear understanding of the question.

You must follow a 2-step reasoning process: First, THINK through the reasoning and enclose it within <THINKING></THINKING> tags. Then, give your answer, and wrapped in \boxed{}.

Figure 9: R1-like Response Prompt.

Model	#Params	Pass@1
Mistral	7B	59.5
Yi-1.5	9B	69.8
LLaMA-3.1	8B	75.2
LLaMA-3.1 + CoT	8B	74.6
LLaMA-3.1 + GRPO	8B	75.7
LLaMA-3.1 + DAPO	8B	75.4
Qwen2.5	7B	72.7
Qwen2.5 + CoT	7B	73.1
Qwen2.5 + GRPO	7B	74.9
Qwen2.5 + DAPO	7B	73.6
AEPO-L (Ours)	8B	77.2
AEPO-Q (Ours)	7B	76.3

Table 5: Comparison Results on PubMedQA Dataset. The AEPO-L and AEPO-Q indicate LLaMA-3.1 w/ AEPO and Qwen2.5 w/ AEPO, respectively.

Method	Act.	Adv.	Gen.
LLaMA w/ DAPO	2.870	7.97e-5	0.882
LLaMA w/ AEPO	2.325	8.25e-5	0.897
Qwen w/ DAPO	1.632	1.52e-4	0.703
Qwen w/ AEPO	1.727	1.68e-4	0.756

Table 6: Comparison of Computational Cost per-Token During Training. We report average time cost (ms/-token) for actor update (Act.), advantage computation (Adv.), and token generation (Gen.).

C Further Quantitative Analysis

C.1 Additional Analysis on I.D. Dataset

We conducted a comprehensive evaluation of our proposed AEPO method on the PubMedQA dataset (Jin et al., 2019), utilizing both Qwen2.5-7B-Instruct (Team, 2024) and LLaMA3.1-8B-Instruct (Dubey et al., 2024) as backbone models. The experimental results are summarized in Table 5. (1) AEPO consistently outperforms existing reinforcement learning with value regularization (RLVR) baselines, including GRPO and DAPO. This consistent improvement highlights AEPO’s effectiveness in mitigating the Echo Reflection (ER) phenomenon during multi-stage inference. (2) In particular, AEPO achieves performance gains of 1.4% and 1.5% over GRPO (Shao et al., 2024) when applied to Qwen2.5 and LLaMA3.1, respectively. These significant improvements underscore the capacity of AEPO to enhance reasoning quality by suppressing erroneous inferences introduced during the initial thinking phase, thereby facilitating more accurate final predictions.

Additional Analysis on Computational Cost. As shown in Table 6 We further compare the average computational time (in milliseconds per token) for three major stages in the reinforcement learning training loop: actor update (Act.), advantage cal-

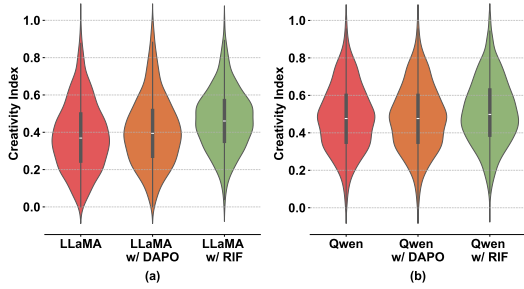


Figure 10: Violin plot of creativity index on MedQA.

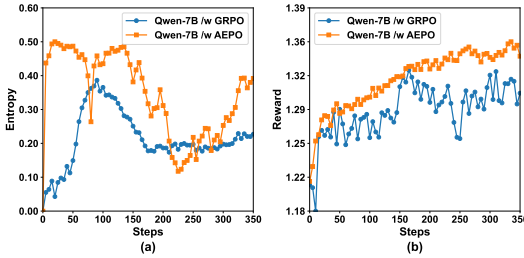


Figure 11: Visualization of training dynamics.

853 culation (Adv.), and token generation (Gen.). The
 854 reported values reflect the average latency required
 855 to process each token, serving as a fine-grained pro-
 856 filing of computational overhead across different
 857 components of the training pipeline. From the table,
 858 we observe that our proposed AEPO method incurs
 859 a comparable computational cost to DAPO (Yu
 860 et al., 2025), while achieving significantly super-
 861 ior performance. This demonstrates the efficiency
 862 and effectiveness of AEPO, highlighting its practi-
 863 cal advantages in balancing reasoning quality and
 864 training overhead.

865 **Analysis on RL Training Dynamics.** To further
 866 illustrate the advantages of the proposed AEPO
 867 algorithm, we visualize the training dynamics of
 868 policy entropy and reward during training for two
 869 methods: GRPO and AEPO. As shown in Fig-
 870 ure 11 (a), the GRPO exhibits a rapid decay in
 871 policy entropy, leading to premature convergence
 872 to a suboptimal policy. However, our proposed
 873 AEPO algorithm effectively mitigates the entropy
 874 collapse problem commonly observed in reinforc-
 875 ement learning while also avoiding the instability
 876 caused by excessive entropy. Figure 11 (b) fur-
 877 ther supports this claim: while DAPO quickly con-
 878 verges after entropy drops, AEPO is able to main-
 879 tain effective exploration throughout training. We
 880 attribute this behavior to the AEO module, which
 881 strikes a dynamic balance between exploration and
 882 exploitation. This balance enables AEPO to sustain

883 high-quality exploration of domain-specific knowl-
 884 edge, ultimately improving reasoning performance.
 885 **Additional Analysis on Creativity Index.** We
 886 further perform supplementary experiments on the
 887 MedQA dataset (Jin et al., 2021) to visualize the
 888 Creativity Index (Lu et al., 2024) of model re-
 889 sponses using violin plots, as shown in Figure 10.
 890 The Creativity Index quantitatively assesses the
 891 originality of model outputs, with higher values in-
 892 dicated more novel generations and lower values
 893 reflecting greater reliance on the reference corpus.
 894 We follow the same experimental setup as in the
 895 main paper’s Creativity Index evaluation, except
 896 that we use a different dataset. The results demon-
 897 strate that both LLaMA-8B with RIF and Qwen-
 898 7B with RIF achieve significantly higher Creativ-
 899 ity Index scores than their respective DAPO-based
 900 counterparts. This indicates that the RIF module
 901 effectively reduces the influence of prior reasoning
 902 during the reflection phase, thereby alleviating the
 903 Echo Reflection phenomenon and promoting more
 904 original and independent reflective reasoning.

905 D Additional Qualitative Analysis

906 D.1 Analysis on I.D. Dataset

907 We present addition qualitative analysis on I.D.
 908 dataset MedMcQA (Pal et al., 2022). As illus-
 909 trated in Figure 12, both our model (LLaMA w/
 910 AEPO) and comparison model (LLaMA w/ GRPO)
 911 correctly identified “Candida is not typically con-
 912 sidered a significant teratogen” (highlighted with
 913 a green rectangle); however, neither produced the
 914 correct final judgment. Notably, the two models di-
 915 verged significantly in the reflection-stage. Specif-
 916 ically, LLaMA w/ GRPO was heavily influenced
 917 by the erroneous reasoning established during the
 918 initial thinking phase and failed to break free from
 919 these misconceptions (highlighted with blue and
 920 red rectangles). This resulted in the propagation
 921 of incorrect reasoning into the reflection stage, ul-
 922 timately leading to an incorrect conclusion. In
 923 contrast, our model successfully recognized during
 924 the reflection the necessity of carefully examining
 925 “the specific effects of each infection on the fetus”
 926 (highlighted with a yellow rectangle), thereby re-
 927 vising the earlier judgment and arriving at a correct
 928 reasoning trajectory. These results demonstrate
 929 that, compared to GRPO, our method AEPO more
 930 effectively regulates the flow of information by
 931 leveraging the IB theory, thus preventing the prop-
 932 agation of faulty reasoning. Moreover, through a

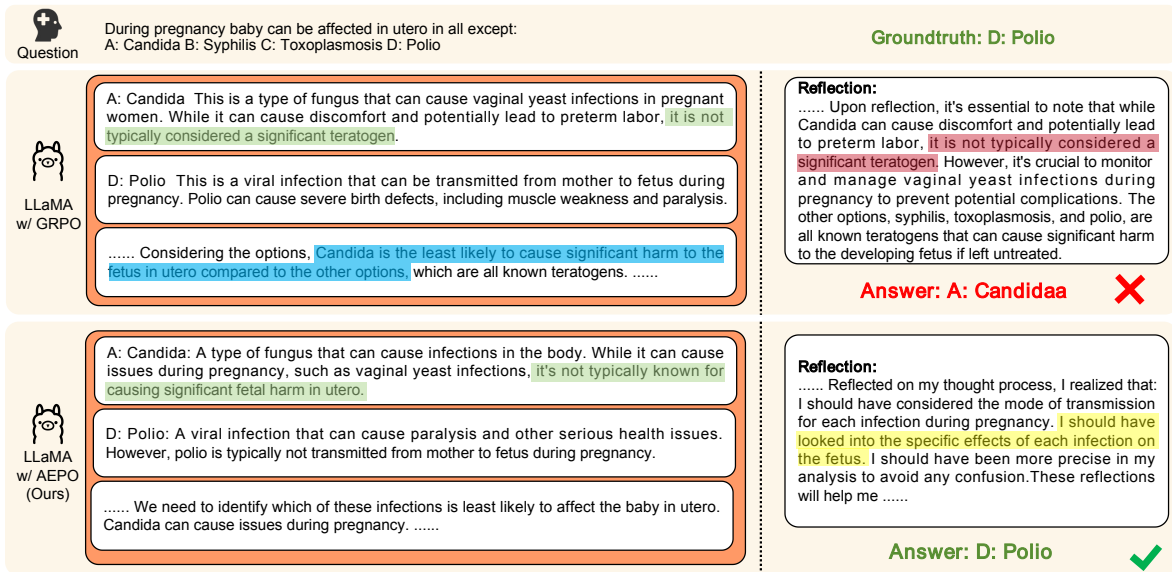


Figure 12: Quality visualization analysis on I.D. Dataset. It presents responses of LLaMA with GRPO (LLaMA w/GRPO) and our proposed AEPO method (LLaMA w/AEPO). The left side of the dashed line illustrates the content of the model’s thinking-stage, while the right side displays the content of the reflection-stage and the final answer. Key contents are highlighted in different colors.

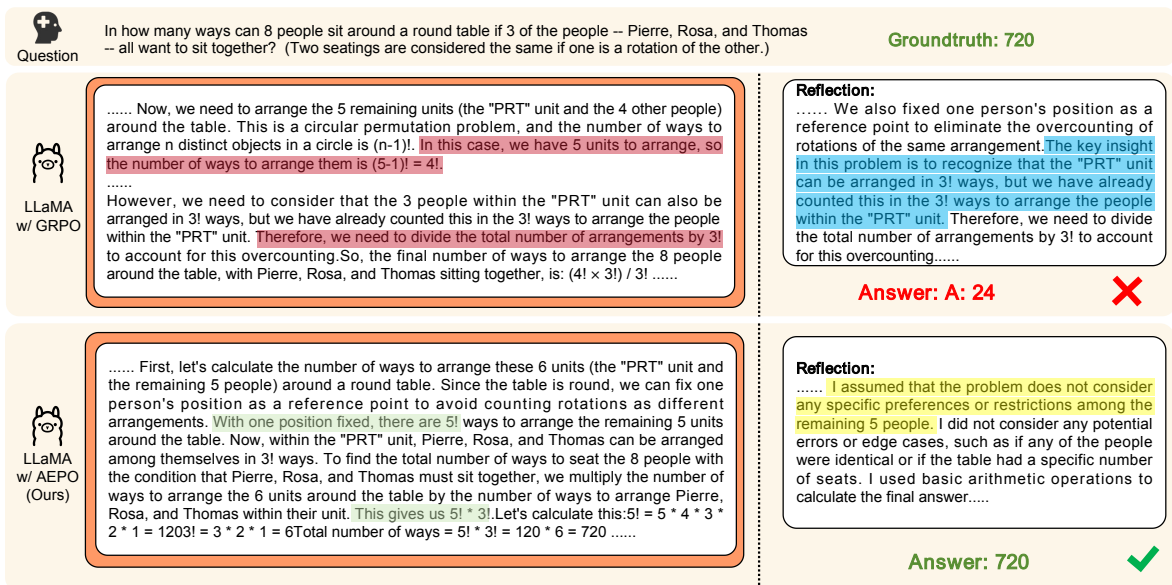


Figure 13: Quality visualization analysis on O.O.D. Dataset. It presents responses of LLaMA with GRPO (LLaMA w/GRPO) and our proposed AEPO method (LLaMA w/AEPO). The left side of the dashed line illustrates the content of the model’s thinking-stage, while the right side displays the content of the reflection-stage and the final answer. Key contents are highlighted in different colors.

933 more efficient exploration of the knowledge space,
934 our model achieves higher reasoning accuracy dur-
935 ing reflection.

936 D.2 Analysis on O.O.D. Dataset

937 We further present addition qualitative analysis on
938 O.O.D. dataset MATH-500 (Lightman et al., 2023).
939 The visualization (Figure 13) demonstrates that our

model (LLaMA w/ AEPO) consistently generates
correct reasoning paths from the outset (high-
lighted with a green rectangle), highlighting the superiority
of the AEO module in exploring the knowledge
space. Further analysis shows that the comparison
model LLaMA w/GRPO model continues to ex-
hibit a strong influence from its initial think-stage
during reflection-stage (highlighted with red and

940
941
942
943
944
945
946
947

948 blue rectangles), manifesting the typical Echo Re-
 949 flection error. In contrast, even when the initial rea-
 950 soning is correct, our model successfully engages
 951 in reflective evaluation of potential weaknesses,
 952 indicating the robustness of the RIF under both cor-
 953 rect and incorrect reasoning scenarios (highlighted
 954 with a yellow rectangle).

955 Additionally, visualizations on the mathematical
 956 dataset further reveal that the ER phenomenon is
 957 not confined to the medical domain but is preva-
 958 lent across diverse reasoning tasks. Our proposed
 959 method demonstrates strong cross-domain general-
 960 ization capabilities, adapting effectively to complex
 961 reasoning environments across various fields with
 962 robustness and reliability.

Algorithm 1 Algorithm for AEPO.

Input: policy model π_θ ; dataset \mathcal{D} ; hyperparame-
 ters ϵ_{low} , ϵ_{low} ; training steps S .

Output: π_θ .

- 1: **for** step = 1, \dots , S **do**
 - 2: Sample a minibatch \mathcal{D}_m form \mathcal{D} .
 - 3: Update the old policy model $\pi_{\theta_{old}} \leftarrow \pi_\theta$.
 - 4: Sample G outputs $\{O_i\}_{i=1}^G \sim \pi_{\theta_{old}}$ for $Q \in \mathcal{D}_m$.
 - 5: **for** $i \in 1, \dots, G$ **do**
 - 6: Compute reward r_i for O_i .
 - 7: Compute $\mathcal{L}(O_i) = \min(\mathcal{L}_{IB} - \mathcal{F}_{GAE})$
 for O_i .
 - 8: Compute advantage $\hat{A}_{i,t}$ for each token
 for O_i .
 - 9: **end for**
 - 10: Update π_θ by maximizing the AEPO objec-
 tive.
 - 11: **end for**
-