DAPO: An Open-Source LLM Reinforcement Learning System at Scale

[1]ByteDance Seed [2]Institute for AI Industry Research (AIR), Tsinghua University
[3]The University of Hong Kong
[4]SIA-Lab of Tsinghua AIR and ByteDance Seed

Full author list in Contributions

Abstract

Inference scaling empowers LLMs with unprecedented reasoning ability, with reinforcement learning as the core technique to elicit complex reasoning. However, key technical details of state-of-the-art reasoning LLMs are concealed (such as in OpenAI o1 blog and DeepSeek R1 technical report), thus the community still struggles to reproduce their RL training results. We propose the **D**ecoupled Clip and **D**ynamic s**A**mpling **P**olicy **O**ptimization (**DAPO**) algorithm, and fully open-source a state-of-the-art large-scale RL system that achieves 50 points on AIME 2024 using Qwen2.5-32B base model. Unlike previous works that withhold training details, we introduce four key techniques of our algorithm that make large-scale LLM RL a success. In addition, we open-source our training code, which is built on the **verl** framework [a], along with a carefully curated and processed dataset. These components of our open-source system enhance reproducibility and support future research in large-scale LLM RL.

Date: March 17, 2025

Correspondence: Qiying Yu at yuqy22@mails.tsinghua.edu.cn

Project Page: https://dapo-sia.github.io/

[a] https://github.com/volcengine/verl

DAPO: An Open-Source LLM Reinforcement Learning System at Scale

[1]ByteDance Seed [2]Institute for AI Industry Research (AIR), Tsinghua University
[3]The University of Hong Kong
[4]SIA-Lab of Tsinghua AIR and ByteDance Seed

Full author list in Contributions

Abstract

推理扩展赋予了大语言模型前所未有的推理能力， 其中强化学习是激发复杂推理的核心技术。 然而， 最先进的推理型大语言模型的关键技术细节被隐藏 （例如在OpenAI的o1博客和DeepSeek的R1技术报告中）， 因此社区仍然难以重现它们的强化学习训练结果。 我们提出了**D**ecoupled Clip and **D**ynamic s**A**mpling **P**olicy **O**ptimization (**DAPO**)算法，并完全开源了一个最先进的大规模强化学习系统，该系统使用Qwen2.5-32B基础模型在2024年的AIME考试中获得了50分。与之前隐瞒训练细节的工作不同，我们介绍了算法成功的四个关键技术。此外，我们开源了基于**verl**框架 [a]的训练代码，以及精心策划和处理的数据集。我们开源系统的这些组成部分增强了可重复性，并支持未来在大规模语言模型强化学习领域的研究。

Date: March 17, 2025

Correspondence: Qiying Yu at yuqy22@mails.tsinghua.edu.cn

Project Page: https://dapo-sia.github.io/

[a] https://github.com/volcengine/verl

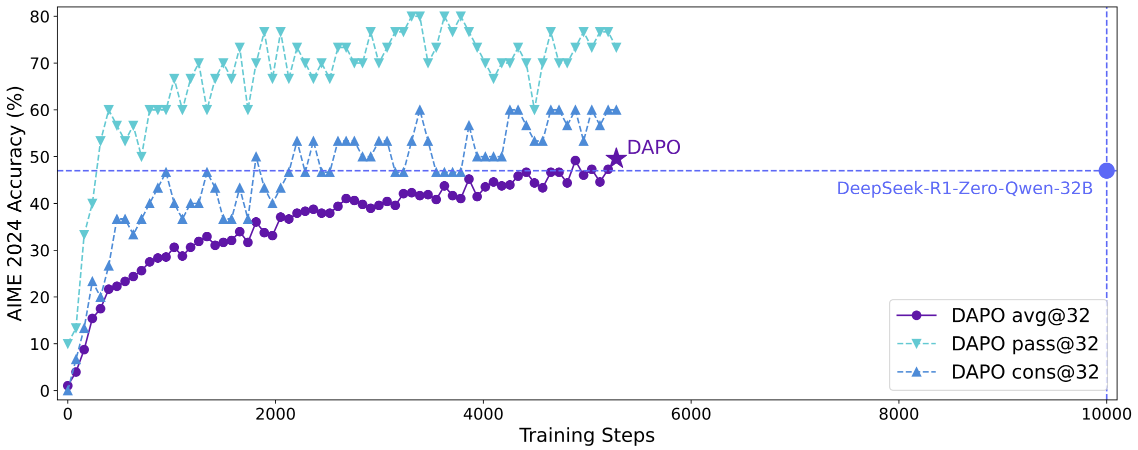


Figure 1 AIME 2024 scores of **DAPO** on the Qwen2.5-32B base model, outperforming the previous SoTA DeepSeek-R1-Zero-Qwen-32B using 50% training steps.

1 Introduction

Test-time scaling such as OpenAI's o1 [1] and DeepSeek's R1 [2] brings a profound paradigm shift to Large Language Models (LLMs) [3–7]. Test-time scaling enables longer Chain-of-Thought thinking and induces sophisticated reasoning behaviors, which makes the models superior in competitive math and coding tasks like AIME and Codeforces.

The central technique driving the revolution is large-scale Reinforcement Learning (RL), which elicits complex reasoning behaviors such as self-verification and iterative refinement. However, the actual algorithm and key recipe for scalable RL training remains a myth, hidden from technical reports of existing reasoning models [1, 2, 8–11]. In this paper, we reveal significant obstacles in large-scale RL training and open-source a scalable RL system with fully open-sourced algorithm, training code and dataset that provides democratized solutions with industry-level RL results.

We experiment over Qwen2.5-32B [12] as the pretrained model for RL. In our initial GRPO run, we achieved only 30 points on AIME — a performance significantly below DeepSeek’ s RL (47 points). A thorough analysis reveals that the naive GRPO baseline suffers from several key issues such as entropy collapse, reward noise, and training instability. The broader community has encountered similar challenges in reproducing DeepSeek's results [13–19] suggesting that critical training details may have been omitted in the R1 paper that are required to develop an industry-level, large-scale, and reproducible RL system.

To close this gap, we release an open-source state-of-the-art system for large-scale LLM RL, which achieves 50 points on AIME 2024 based on Qwen2.5-32B model, outperforming previous state-of-the-art results achieved by DeepSeek-R1-Zero-Qwen-32B [2] (47 points) using 50% training steps (Figure 1). We propose the **D**ecoupled Clip and **D**ynamic s**A**mpling **P**olicy **O**ptimization (**DAPO**) algorithm, and introduce 4 key techniques to make RL shine in the long-CoT RL scenario. Details are presented in Section 3.

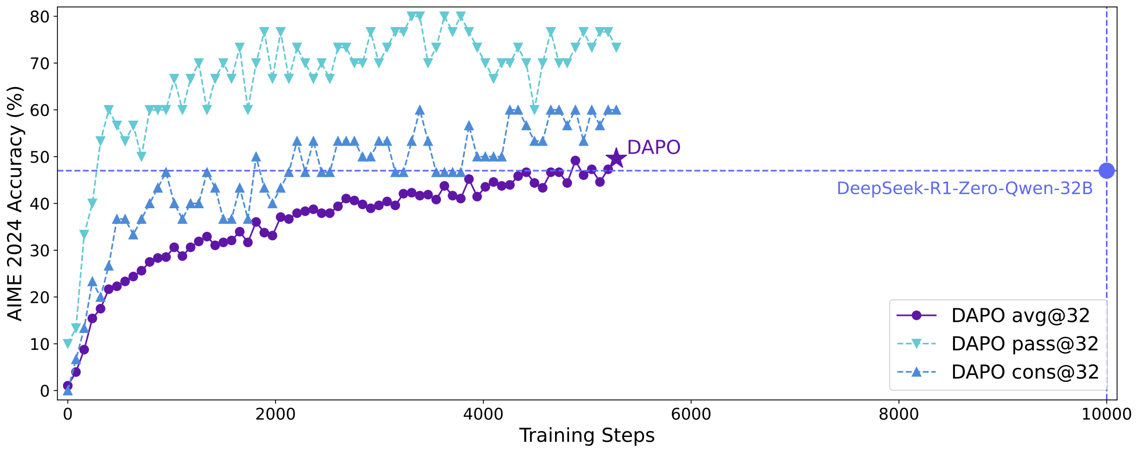


Figure 1 在Qwen2.5-32B基础模型上，**DAPO**的AIME 2024得分超过了之前使用50%训练步数的SoTA DeepSeek-R1-Zero-Qwen-32B。

1 Introduction

测试时扩展，例如OpenAI的o1 [1]和DeepSeek的R1 [2]，为大型语言模型（LLMs） [3–7]带来了深刻的范式转变。测试时扩展能够实现更长的链式思维，并引发复杂的推理行为，从而使这些模型在诸如AIME和Codeforces等数学和编程竞赛任务中表现出色。

推动这一革命的核心技术是大规模强化学习（RL），它能够激发复杂的推理行为，例如自我验证和迭代改进。然而，可扩展RL训练的实际算法和关键方法仍然是一个谜，隐藏在现有推理模型的技术报告中 [1, 2, 8–11]。在本文中，我们揭示了大规模RL训练中的重要障碍，并开源了一个具有完全开源算法、训练代码和数据集的可扩展RL系统，该系统提供了行业级RL结果的民主化解决方案。

我们以Qwen2.5-32B [12]作为预训练模型进行RL实验。在我们的初始GRPO运行中，我们在AIME上仅获得了30分——这一表现显著低于DeepSeek的RL（47分）。深入分析表明，简单的GRPO基线存在几个关键问题，如熵崩溃、奖励噪声和训练不稳定。更广泛的社区在重现DeepSeek的结果时也遇到了类似的挑战 [13–19]，这表明R1论文中可能遗漏了某些关键的训练细节，而这些细节对于开发行业级、大规模且可重现的RL系统至关重要。

为了弥补这一差距，我们发布了一个开源的最先进的大规模LLM RL系统，该系统基于Qwen2.5-32B模型，在AIME 2024上取得了50分的成绩，超过了DeepSeek-R1-Zero-Qwen-32B [2]（47分）之前的最佳结果，并且仅使用了50%的训练步骤（图1）。我们提出了**D**ecoupled Clip和**D**ynamic s**A**mpling **P**olicy **O**ptimization (**DAPO**) 算法，并引入了4项关键技术，使RL在长链式思维（CoT）RL场景中大放异彩。详细内容见Section 3。

1. **Clip-Higher**，它促进系统多样性并避免熵崩溃；
2. **动态采样**，它提高了训练效率和稳定性；
3. **Token-Level 策略梯度损失**，这在长链推理的强化学习场景中至关重要；
4. **过度奖励塑形**，它可以减少奖励噪声并稳定训练。

1. **Clip-Higher**, which promotes the diversity of the system and avoids entropy collapse;
2. **Dynamic Sampling**, which improves training efficiency and stability;
3. **Token-Level Policy Gradient Loss**, which is critical in long-CoT RL scenarios;
4. **Overlong Reward Shaping**, which reduces reward noise and stabilizes training.

Our implementation is based on verl [20]. By fully releasing our state-of-the-art RL system including training code and data, we aim to reveal valuable insights to large-scale LLM RL that benefit the larger community.

2 Preliminary

2.1 Proximal Policy Optimization (PPO)

PPO [21] introduces a clipped surrogate objective for policy optimization. By constraining the policy updates within a proximal region of the previous policy using clip, PPO stabilizes training and improves sample efficiency. Specifically, PPO updates the policy by maximizing the following objective:

$$\mathcal{J}_{\text{PPO}}(\theta) = \mathbb{E}_{(q,a)\sim\mathcal{D}, o_{\le t}\sim\pi_{\theta_{\text{old}}}(\cdot|q)} \left[\min\left(\frac{\pi_\theta(o_t \mid q, o_{<t})}{\pi_{\theta_{\text{old}}}(o_t \mid q, o_{<t})} \hat{A}_t,\ \text{clip}\left(\frac{\pi_\theta(o_t \mid q, o_{<t})}{\pi_{\theta_{\text{old}}}(o_t \mid q, o_{<t})}, 1-\varepsilon, 1+\varepsilon \right) \hat{A}_t \right) \right], \tag{1}$$

where (q,a) is a question-answer pair from the data distribution $\mathcal{D}$, ε is the clipping range of importance sampling ratio, and $\hat{A}_t$ is an estimator of the advantage at time step t. Given the value function V and the reward function R, $\hat{A}_t$ is computed using the Generalized Advantage Estimation (GAE) [22]:

$$\hat{A}_t^{\text{GAE}(\gamma,\lambda)} = \sum_{l=0}^{\infty} (\gamma\lambda)^l \delta_{t+l}, \tag{2}$$

where

$$\delta_l = R_l + \gamma V(s_{l+1}) - V(s_l), \quad 0 \le \gamma, \lambda \le 1. \tag{3}$$

2.2 Group Relative Policy Optimization (GRPO)

Compared to PPO, GRPO eliminates the value function and estimates the advantage in a group-relative manner. For a specific question-answer pair (q,a), the behavior policy $\pi_{\theta_{\text{old}}}$ samples a group of G individual responses $\{o_i\}_{i=1}^G$. Then, the advantage of the i-th response is calculated by normalizing the group-level rewards $\{R_i\}_{i=1}^G$:

$$\hat{A}_{i,t} = \frac{r_i - \text{mean}(\{R_i\}_{i=1}^G)}{\text{std}(\{R_i\}_{i=1}^G)}. \tag{4}$$

Similar to PPO, GRPO adopts a clipped objective, together with a directly imposed KL penalty term:

$$\begin{aligned} \mathcal{J}_{\text{GRPO}}(\theta) &= \mathbb{E}_{(q,a)\sim\mathcal{D}, \{o_i\}_{i=1}^G\sim\pi_{\theta_{\text{old}}}(\cdot|q)} \\ &\left[\frac{1}{G}\sum_{i=1}^{G} \frac{1}{|o_i|}\sum_{t=1}^{|o_i|} \left(\min\left(r_{i,t}(\theta)\hat{A}_{i,t},\ \text{clip}\left(r_{i,t}(\theta), 1-\varepsilon, 1+\varepsilon \right)\hat{A}_{i,t} \right) - \beta D_{\text{KL}}(\pi_\theta \| \pi_{\text{ref}}) \right) \right], \end{aligned} \tag{5}$$

我们的实现基于 verl [20]。通过完全开源我们最先进的强化学习系统，包括训练代码和数据，我们希望能够揭示对大规模语言模型的强化学习有价值的见解，从而惠及更广泛的社区。

2 Preliminary

2.1 Proximal Policy Optimization (PPO)

PPO [21] 引入了一种带有裁剪的替代目标来进行策略优化。通过使用裁剪约束策略更新在先前策略的近端区域内，PPO 稳定了训练并提高了样本效率。具体来说，PPO 通过最大化以下目标来更新策略：

$$\mathcal{J}_{\text{PPO}}(\theta) = \mathbb{E}_{(q,a)\sim\mathcal{D}, o_{\le t}\sim\pi_{\theta_{\text{old}}}(\cdot|q)} \left[\min\left(\frac{\pi_\theta(o_t \mid q, o_{<t})}{\pi_{\theta_{\text{old}}}(o_t \mid q, o_{<t})} \hat{A}_t,\ \text{clip}\left(\frac{\pi_\theta(o_t \mid q, o_{<t})}{\pi_{\theta_{\text{old}}}(o_t \mid q, o_{<t})}, 1-\varepsilon, 1+\varepsilon \right) \hat{A}_t \right) \right], \tag{1}$$

其中 (q,a) 是来自数据分布 $\mathcal{D}$ 的问题-答案对，ε 是重要性采样比率的截断范围，$\hat{A}_t$ 是时间步 t 处的优势函数估计量。给定价值函数 V 和奖励函数 R，$\hat{A}_t$ 使用广义优势估计 (GAE) [22] 进行计算：

$$\hat{A}_t^{\text{GAE}(\gamma,\lambda)} = \sum_{l=0}^{\infty} (\gamma\lambda)^l \delta_{t+l}, \tag{2}$$

where

$$\delta_l = R_l + \gamma V(s_{l+1}) - V(s_l), \quad 0 \le \gamma, \lambda \le 1. \tag{3}$$

2.2 Group Relative Policy Optimization (GRPO)

与PPO相比，GRPO去除了价值函数，并以组相对的方式估计优势。对于特定的问答对(q,a)，行为策略$\pi_{\theta_{\text{old}}}$采样一组包含$G$个个体响应$\{o_i\}_{i=1}^G$。然后，通过归一化组级奖励$\{R_i\}_{i=1}^G$来计算第$i$个响应的优势：

$$\hat{A}_{i,t} = \frac{r_i - \text{mean}(\{R_i\}_{i=1}^G)}{\text{std}(\{R_i\}_{i=1}^G)}. \tag{4}$$

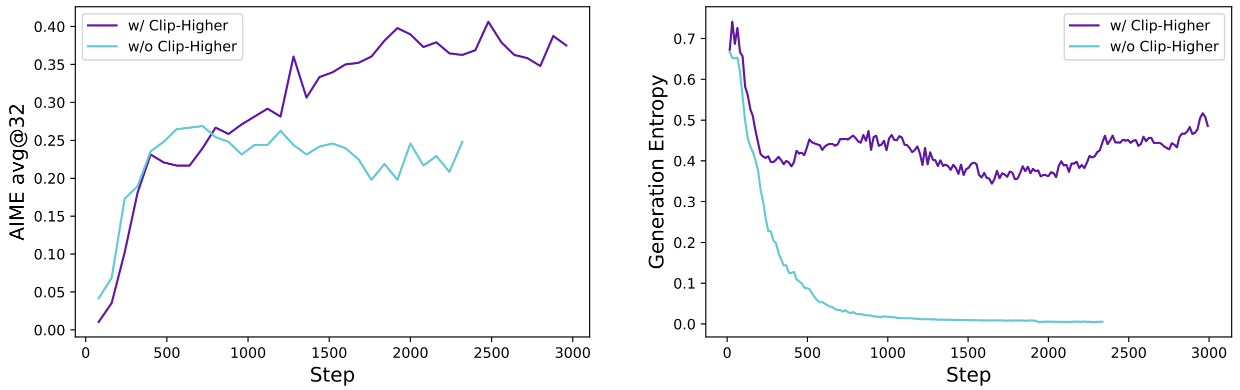
与PPO类似，GRPO采用了一个裁剪目标函数，并直接施加了KL惩罚项：

$$\begin{aligned} \mathcal{J}_{\text{GRPO}}(\theta) &= \mathbb{E}_{(q,a)\sim\mathcal{D}, \{o_i\}_{i=1}^G\sim\pi_{\theta_{\text{old}}}(\cdot|q)} \\ &\left[\frac{1}{G}\sum_{i=1}^{G} \frac{1}{|o_i|}\sum_{t=1}^{|o_i|} \left(\min\left(r_{i,t}(\theta)\hat{A}_{i,t},\ \text{clip}\left(r_{i,t}(\theta), 1-\varepsilon, 1+\varepsilon \right)\hat{A}_{i,t} \right) - \beta D_{\text{KL}}(\pi_\theta \| \pi_{\text{ref}}) \right) \right], \end{aligned} \tag{5}$$

where

$$r_{i,t}(\theta) = \frac{\pi_\theta(o_{i,t} \mid q, o_{i,<t})}{\pi_{\theta_{\text{old}}}(o_{i,t} \mid q, o_{i,<t})}. \tag{6}$$

还值得注意的是，GRPO 在样本级别计算目标函数。确切地说，GRPO 首先计算每个生成序列内的平均损失，然后才对不同样本的损失进行平均。正如我们将在第 3.3节中讨论的那样，这种差异可能会影响算法的性能。



(a) Accuracies on AIME. **(b)** Entropy of actor model.

Figure 2 The accuracy on the AIME test set and the entropy of the actor model's generated probabilities during the RL training process, both before and after applying **Clip-Higher** strategy.

where

$$r_{i,t}(\theta) = \frac{\pi_\theta(o_{i,t} \mid q, o_{i,<t})}{\pi_{\theta_{\text{old}}}(o_{i,t} \mid q, o_{i,<t})}. \tag{6}$$

It is also worth noting that GRPO computes the objective at the sample-level. To be exact, GRPO first calculates the mean loss within each generated sequence, before averaging the loss of different samples. As we will be discussing in Section 3.3, such difference may have an impact on the performance of the algorithm.

2.3 Removing KL Divergence

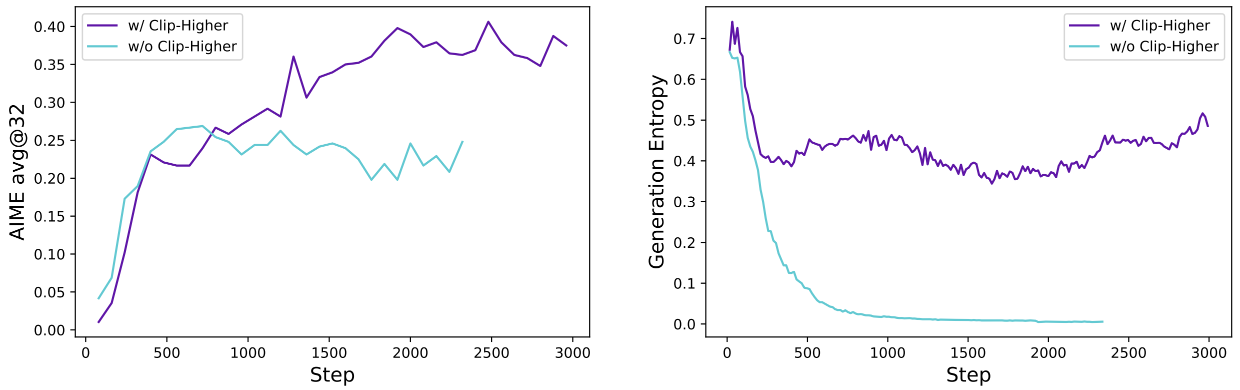
The KL penalty term is used to regulate the divergence between the online policy and the frozen reference policy. In the RLHF scenario [23], the goal of RL is to align the model behavior without diverging too far from the initial model. However, during training the long-CoT reasoning model, the model distribution can diverge significantly from the initial model, thus this restriction is not necessary. Therefore, we will exclude the KL term from our proposed algorithm.

2.4 Rule-based Reward Modeling

The use of reward model usually suffers from the reward hacking problem [24–29]. Instead, we directly use the final accuracy of a verifiable task as the outcome reward, computed using the following rule:

$$R(\hat{y}, y) = \begin{cases} 1, & \texttt{is\_equivalent}(\hat{y}, y) \\ -1, & \text{otherwise} \end{cases} \tag{7}$$

where y is the ground-truth answer and $\hat{y}$ is the predicted answer. This is proved to be an effective approach to activating the base model's reasoning capability, as shown in multiple domains such as automated theorem proving [30–33], computer programming [34–37], and mathematics competition [2].



(a) Accuracies on AIME. **(b)** Entropy of actor model.

Figure 2 在应用**Clip-Higher**策略前后，AIME测试集上的准确率和演员模型在强化学习训练过程中生成的概率的熵。

2.3 Removing KL Divergence

KL惩罚项用于调节在线策略与冻结的参考策略之间的差异。在RLHF场景 [23]中，强化学习的目标是使模型行为对齐，同时不与初始模型相差太远。然而，在训练长链推理模型时，模型分布可能会显著偏离初始模型，因此这种限制并非必要。因此，我们将从我们提出的算法中排除KL项。

2.4 Rule-based Reward Modeling

奖励模型的使用通常会遇到奖励劫持问题 [24–29]。相反，我们直接将可验证任务的最终准确率作为结果奖励，其计算规则如下：

$$R(\hat{y}, y) = \begin{cases} 1, & \texttt{is\_equivalent}(\hat{y}, y) \\ -1, & \text{otherwise} \end{cases} \tag{7}$$

其中 y 是真实答案，$\hat{y}$ 是预测答案。这已被证明是激活基础模型推理能力的有效方法，如在多个领域中所展示的那样，例如自动定理证明 [30–33]、计算机编程 [34–37] 和数学竞赛 [2]。

3 DAPO

我们提出了**D**ecouple Clip and **D**ynamic s**A**mpling **P**olicy **O**ptimization（DAPO）算法。DAPO为每个问题q（配对有答案a）采样一组输出$\{o_i\}_{i=1}^G$，并通过以下目标优化策略：

$$\begin{aligned} \mathcal{J}_{\text{DAPO}}(\theta) = \; & \mathbb{E}_{(q,a)\sim\mathcal{D},\{o_i\}_{i=1}^G\sim\pi_{\theta_{\text{old}}}(\cdot\mid q)} \\ & \left[\frac{1}{\sum_{i=1}^G |o_i|}\sum_{i=1}^G\sum_{t=1}^{|o_i|}\min\Big(r_{i,t}(\theta)\hat{A}_{i,t},\ \text{clip}\Big(r_{i,t}(\theta), 1-\varepsilon_{\text{low}}, 1+\varepsilon_{\text{high}}\Big)\hat{A}_{i,t}\Big)\right] \\ \text{s.t.} \quad & 0 < \Big|\{o_i \mid \texttt{is\_equivalent}(a, o_i)\}\Big| < G, \end{aligned} \tag{8}$$

where

$$r_{i,t}(\theta) = \frac{\pi_\theta(o_{i,t} \mid q, o_{i,<t})}{\pi_{\theta_{\text{old}}}(o_{i,t} \mid q, o_{i,<t})}, \quad \hat{A}_{i,t} = \frac{R_i - \text{mean}(\{R_i\}_{i=1}^G)}{\text{std}(\{R_i\}_{i=1}^G)}. \tag{9}$$

3 DAPO

We propose the **D**ecouple Clip and **D**ynamic s**A**mpling **P**olicy **O**ptimization (DAPO) algorithm. DAPO samples a group of outputs $\{o_i\}_{i=1}^G$ for each question q paired with the answer a, and optimizes the policy via the following objective:

$$\begin{aligned}\mathcal{J}_{\text{DAPO}}(\theta) = \quad & \mathbb{E}_{(q,a)\sim\mathcal{D},\{o_i\}_{i=1}^G\sim\pi_{\theta_{\text{old}}}(\cdot|q)} \\ & \left[\frac{1}{\sum_{i=1}^G|o_i|}\sum_{i=1}^G\sum_{t=1}^{|o_i|}\min\Big(r_{i,t}(\theta)\hat{A}_{i,t},\ \text{clip}\Big(r_{i,t}(\theta),1-\varepsilon_{\text{low}},1+\varepsilon_{\text{high}}\Big)\hat{A}_{i,t}\Big)\right] \\ \text{s.t.} \quad & 0<\Big|\{o_i\mid \texttt{is\_equivalent}(a,o_i)\}\Big|<G,\end{aligned} \tag{8}$$

where

$$r_{i,t}(\theta)=\frac{\pi_\theta(o_{i,t}\mid q,o_{i,<t})}{\pi_{\theta_{\text{old}}}(o_{i,t}\mid q,o_{i,<t})},\quad \hat{A}_{i,t}=\frac{R_i-\text{mean}(\{R_i\}_{i=1}^G)}{\text{std}(\{R_i\}_{i=1}^G)}. \tag{9}$$

The full algorithm can be found in Algorithm 1. In this section, we will introduce the key techniques associated with DAPO.

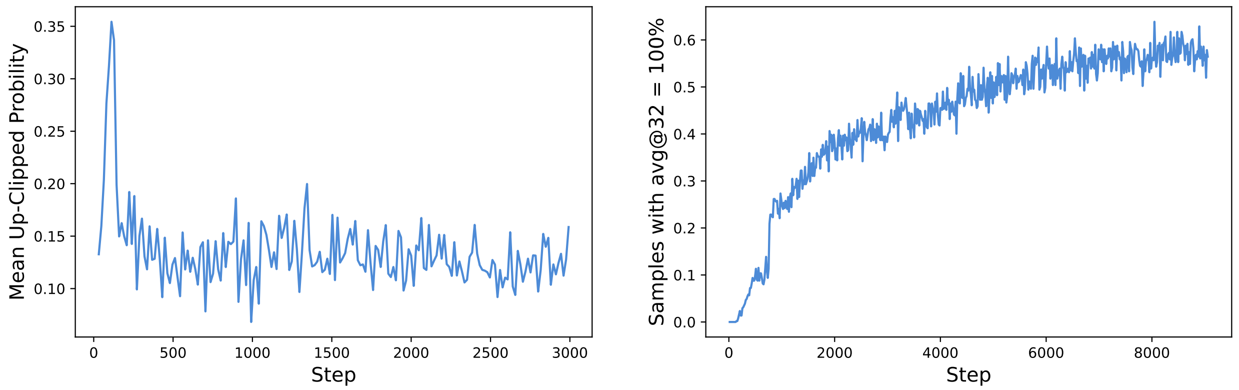
3.1 Raise the Ceiling: Clip-Higher

In our initial experiments using naive PPO [21] or GRPO [38], we observed the entropy collapse phenomenon: the entropy of the policy decreases quickly as training progresses (Figure 2b). The sampled responses of certain groups tend to be nearly identical. This indicates limited exploration and early deterministic policy, which can hinder the scaling process.

We propose the **Clip-Higher** strategy to address this issue. Clipping over the importance sampling ratio is introduced in Clipped Proximal Policy Optimization (PPO-Clip) [21] to restrict the trust region and enhance the stability of RL. We identify that the upper clip can restrict the exploration of the policy. In this case, it is much easier to make an 'exploitation token' more probable, than to uplift the probability of an unlikely 'exploration token'.

Concretely, when $\varepsilon = 0.2$ (the default value of most algorithms), consider two actions with probabilities $\pi_{\theta_{\text{old}}}(o_i \mid q) = 0.01$ and 0.9. The maximum possible updated probabilities $\pi_\theta(o_i \mid q)$ are 0.012 and 1.08, respectively. This implies that for tokens with a higher probability (*e.g.*, 0.9) is less constrained. Conversely, for low-probability tokens, achieving a non-trivial increase in probability is considerably more challenging. Empirically, we also observe that the maximum probability of clipped tokens is approximately $\pi_\theta(o_i \mid q) < 0.2$ (Figure 3a). This finding supports our analysis that the upper clipping threshold indeed restricts the probability increase of low-probability tokens, thereby potentially constraining the diversity of the system.

Adhering to the **Clip-Higher** strategy, we decouple the lower and higher clipping range as ε_{low} and $\varepsilon_{\text{high}}$,



(a) Maximum clipped probabilities.

(b) 准确率为1的样本比例。

Figure 3 演员模型的概率分布的熵，以及响应长度的变化。

完整的算法可以在算法 1中找到。在本节中，我们将介绍与DAPO相关的关键技术。

3.1 Raise the Ceiling: Clip-Higher

在我们最初的实验中，使用了朴素的PPO [21] 或 GRPO [38]，我们观察到了策略熵崩溃现象：随着训练的进行，策略的熵迅速下降（Figure 2b）。某些组别的采样响应趋于几乎完全相同。这表明探索能力有限，并且策略过早地变得确定性，这可能会阻碍模型的扩展过程。

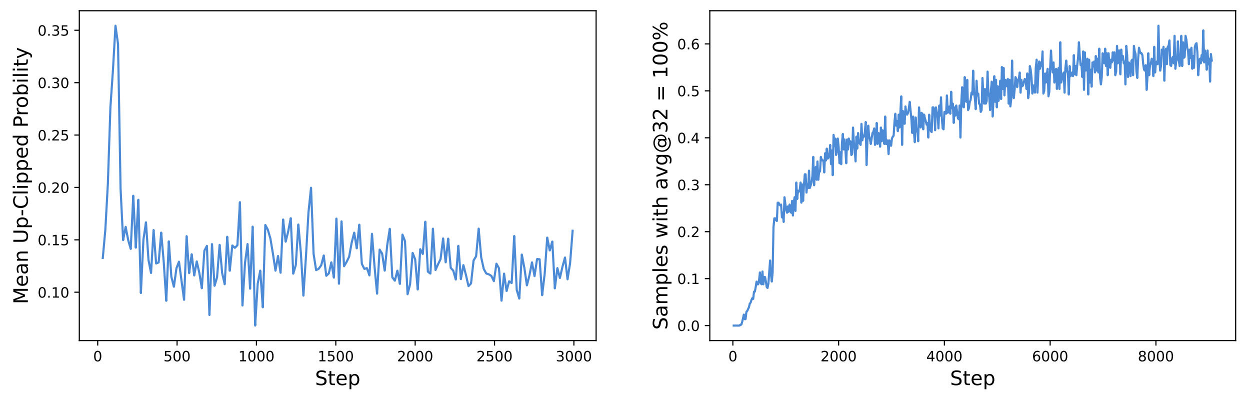
为了解决这一问题，我们提出了 **Clip-Higher** 策略。在截断近端策略优化 (PPO-Clip) [21] 中引入了对重要性采样比率的截断操作，以限制信任区域并增强强化学习的稳定性。我们发现，上界截断会限制策略的探索能力。在这种情况下，使一个"利用型标记"（exploitation token）变得更可能要比提升一个不太可能的"探索型标记"（exploration token）的概率容易得多。

具体来说，当 $\varepsilon = 0.2$（大多数算法的默认值）时，考虑两个动作的概率分别为 $\pi_{\theta_{\text{old}}}(o_i \mid q) = 0.01$ 和 0.9。其最大可能更新后的概率 $\pi_\theta(o_i \mid q)$ 分别为 0.012 和 1.08。这意味着对于高概率标记（例如，0.9），其约束较小。相反，对于低概率标记，实现非平凡的概率提升则要困难得多。从经验上看，我们也观察到被截断的标记的最大概率约为 $\pi_\theta(o_i \mid q) < 0.2$（Figure 3a）。这一结果支持了我们的分析，即上界截断阈值确实限制了低概率标记的概率提升，从而可能限制系统的多样性。

遵循 **Clip-Higher** 策略，我们将下界和上界截断范围分别解耦为 ε_{low} 和 $\varepsilon_{\text{high}}$，如公式 10 所示。

$$\begin{aligned}\mathcal{J}_{\text{DAPO}}(\theta) = \quad & \mathbb{E}_{(q,a)\sim\mathcal{D},\{o_i\}_{i=1}^G\sim\pi_{\theta_{\text{old}}}(\cdot|q)} \\ & \left[\frac{1}{\sum_{i=1}^G|o_i|}\sum_{i=1}^G\sum_{t=1}^{|o_i|}\min\Big(r_{i,t}(\theta)\hat{A}_{i,t},\ \text{clip}\Big(r_{i,t}(\theta),1-\varepsilon_{\text{low}},1+\varepsilon_{\text{high}}\Big)\hat{A}_{i,t}\Big)\right] \\ \text{s.t.} \quad & 0<\Big|\{o_i\mid \texttt{is\_equivalent}(a,o_i)\}\Big|<G.\end{aligned} \tag{10}$$

我们增大 $\varepsilon_{\text{high}}$ 的值，以留出更多空间来提升低概率词元的权重。如Figure 2所示，这一调整有效地提升了策略的熵，促进了更多样本的生成。我们选择保持 ε_{low} 相对较小，因为增大它会将这些词元的概率压制到 0，从而导致采样空间的坍塌。



(a) Maximum clipped probabilities. **(b)** The proportion of samples with an accuracy of 1.

Figure 3 The entropy of the probability distribution of the actor model, as well as the changes in response length.

as highlighted in Equation 10:

$$\begin{aligned}\mathcal{J}_{\text{DAPO}}(\theta) = &\ \mathbb{E}_{(q,a)\sim\mathcal{D},\{o_i\}_{i=1}^G\sim\pi_{\theta_{\text{old}}}(\cdot|q)}\\ &\left[\frac{1}{\sum_{i=1}^G|o_i|}\sum_{i=1}^G\sum_{t=1}^{|o_i|}\min\Big(r_{i,t}(\theta)\hat{A}_{i,t},\ \text{clip}\Big(r_{i,t}(\theta),1-\varepsilon_{\text{low}},1+\varepsilon_{\text{high}}\Big)\hat{A}_{i,t}\Big)\right]\\ \text{s.t.}\quad & 0<\Big|\{o_i\mid \texttt{is\_equivalent}(a,o_i)\}\Big|<G.\end{aligned} \tag{10}$$

We increase the value of $\varepsilon_{\text{high}}$ to leave more room for the increase of low-probability tokens. As shown in Figure 2, this adjustment effectively enhances the policy's entropy and facilitates the generation of more diverse samples. We opt to keep ε_{low} relatively small, because increasing it will suppress the probability of these tokens to 0, resulting in the collapse of the sampling space.

3.2 The More the Merrier: Dynamic Sampling

Existing RL algorithm suffers from the gradient-decreasing problem when some prompts have accuracy equal to 1. For example for GRPO, if all outputs $\{o_i\}_{i=1}^G$ of a particular prompt are correct and receive the same reward 1, the resulting advantage for this group is *zero*. A zero advantage results in no gradients for policy updates, thereby reducing sample efficiency. Empirically, the number of samples with accuracy equal to 1 continues to increase, as shown in Figure 3b. This means that the effective number of prompts in each batch keeps decreasing, which can lead to larger variance in gradient and dampens the gradient signals for model training.

To this end, we propose to **over-sample and filter out prompts with the accuracy equal to 1 and 0** illustrated in Equation 11, leaving all prompts in the batch with effective gradients and keeping a consistent number of prompts. Before training, we keep sampling until the batch is fully filled with samples whose accuracy is neither 0 nor 1.

3.2 The More the Merrier: Dynamic Sampling

现有的强化学习算法在某些提示的准确率等于1时会遇到梯度下降的问题。例如，在GRPO中，如果某个特定提示的所有输出$\{o_i\}_{i=1}^G$都是正确的，并且收到相同的奖励1，则该组的优势值为零。零优势值导致策略更新没有梯度，从而降低了样本效率。从经验来看，准确率为1的样本数量持续增加，如Figure 3b所示。这意味着每批样本中有效的提示数量不断减少，这可能导致梯度的方差更大，并削弱模型训练中的梯度信号。

为此，我们提出**过采样并过滤掉准确率为1和0的提示**，如公式 11所示，从而使批次中的所有提示都具有有效的梯度，并保持提示数量的一致性。在训练之前，我们会继续采样，直到批次完全由准确率既不是0也不是1的样本填充。

$$\begin{aligned}\mathcal{J}_{\text{DAPO}}(\theta) = &\ \mathbb{E}_{(q,a)\sim\mathcal{D},\{o_i\}_{i=1}^G\sim\pi_{\theta_{\text{old}}}(\cdot|q)}\\ &\left[\frac{1}{\sum_{i=1}^G|o_i|}\sum_{i=1}^G\sum_{t=1}^{|o_i|}\min\Big(r_{i,t}(\theta)\hat{A}_{i,t},\ \text{clip}\Big(r_{i,t}(\theta),1-\varepsilon_{\text{low}},1+\varepsilon_{\text{high}}\Big)\hat{A}_{i,t}\Big)\right]\\ \text{s.t.}\quad & 0<\Big|\{o_i\mid \texttt{is\_equivalent}(a,o_i)\}\Big|<G.\end{aligned} \tag{11}$$

请注意，这种策略并不一定会妨碍训练效率，因为如果RL系统是同步的且生成阶段没有进行管道化处理，那么生成时间通常由长尾样本的生成所主导。此外，我们发现使用动态采样时，实验可以更快地达到相同的性能，如Figure 6所示。

3.3 Rebalancing Act: Token-Level Policy Gradient Loss

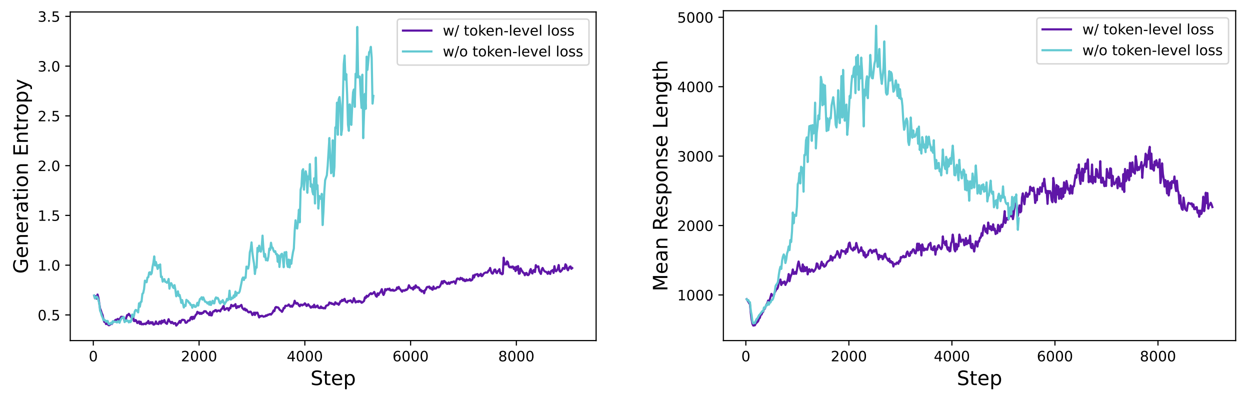
原始的GRPO算法采用样本级别的损失计算方法，该方法首先按样本内的标记（token）平均损失，然后在样本之间聚合损失。在这种方法中，每个样本在最终损失计算中被赋予相同的权重。然而，我们发现，在长链推理（long-CoT）强化学习场景中，这种损失缩减方法引入了几个挑战。

由于所有样本在损失计算中都被赋予相同的权重，较长响应（包含更多标记）中的标记对总体损失的贡献可能会不成比例地降低，这可能导致两种不利影响。首先，对于高质量的长样本，这种效应可能阻碍模型学习其中与推理相关的模式的能力。其次，我们观察到过长的样本往往表现出低质量的模式，例如胡言乱语和重复词语。因此，样本级别的损失计算由于无法有效惩罚长样本中的这些不良模式，导致熵和响应长度不健康地增加，如Figure 4a和Figure 4b所示。

我们在长链推理的强化学习场景中引入了**Token-level Policy Gradient Loss**，以解决上述限制：

$$\begin{aligned}\mathcal{J}_{\text{DAPO}}(\theta) = &\ \mathbb{E}_{(q,a)\sim\mathcal{D},\{o_i\}_{i=1}^G\sim\pi_{\theta_{\text{old}}}(\cdot|q)}\\ &\left[\frac{1}{\sum_{i=1}^G|o_i|}\sum_{i=1}^G\sum_{t=1}^{|o_i|}\min\Big(r_{i,t}(\theta)\hat{A}_{i,t},\ \text{clip}\Big(r_{i,t}(\theta),1-\varepsilon_{\text{low}},1+\varepsilon_{\text{high}}\Big)\hat{A}_{i,t}\Big)\right],\\ \text{s.t.}\quad & 0<\Big|\{o_i\mid \texttt{is\_equivalent}(a,o_i)\}\Big|<G.\end{aligned} \tag{12}$$

在这种设置下，更长的序列相比更短的序列可能对整体梯度更新产生更大的影响。此外，从单个标记的角度来看，如果某个特定的生成模式会导致奖励的增加或减少，那么无论该模式出现在多长的回答



(a) Entropy of actor model's generation probabilities. **(b)** Average length of actor model-generated responses

Figure 4 The entropy of the probability distribution of the actor model, as well as the changes in response length.

$$\mathcal{J}_{\mathrm{DAPO}}(\theta) = \mathbb{E}_{(q,a)\sim\mathcal{D},\{o_i\}_{i=1}^{G}\sim\pi_{\theta_{\mathrm{old}}}(\cdot|q)} \left[\frac{1}{\sum_{i=1}^{G}|o_i|}\sum_{i=1}^{G}\sum_{t=1}^{|o_i|}\min\Big(r_{i,t}(\theta)\hat{A}_{i,t},\ \mathrm{clip}\Big(r_{i,t}(\theta),1-\varepsilon_{\mathrm{low}},1+\varepsilon_{\mathrm{high}}\Big)\hat{A}_{i,t}\Big)\right] \quad (11)$$

$$\text{s.t.}\quad 0 < \Big|\{o_i \mid \texttt{is\_equivalent}(a,o_i)\}\Big| < G.$$

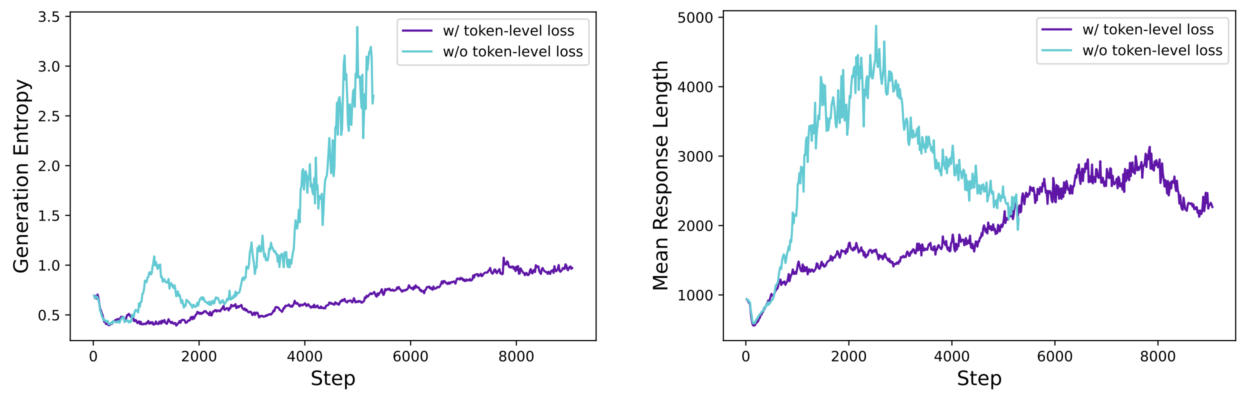
Note that this strategy does not necessarily impede training efficiency, because the generation time is typically dominated by the generation of long-tail samples if the RL system is synchronized and the generation stage is not pipelined. Besides, we find that with dynamic sampling the experiment achieves the same performance faster as shown in Figure 6.

3.3 Rebalancing Act: Token-Level Policy Gradient Loss

The original GRPO algorithm employs a sample-level loss calculation, which involves first averaging the losses by token within each sample and then aggregating the losses across samples. In this approach, each sample is assigned an equal weight in the final loss computation. However, we find that this method of loss reduction introduces several challenges in the context of long-CoT RL scenarios.

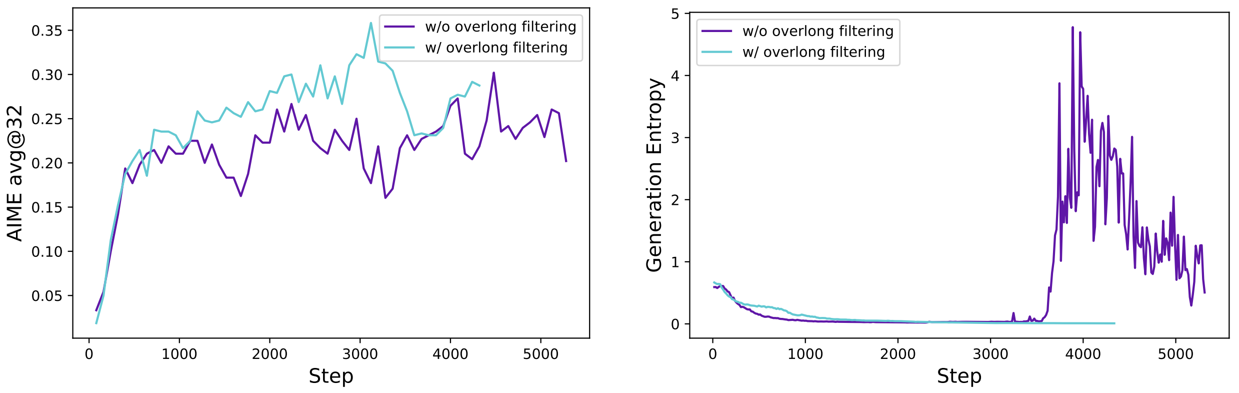
Since all samples are assigned the same weight in the loss calculation, tokens within longer responses (which contain more tokens) may have a disproportionately lower contribution to the overall loss, which can lead to two adverse effects. First, for high-quality long samples, this effect can impede the model's ability to learn reasoning-relevant patterns within them. Second, we observe that excessively long samples often exhibit low-quality patterns such as gibberish and repetitive words. Thus, sample-level loss calculation, due to its inability to effectively penalize those undesirable patterns in long samples, leads to an unhealthy increase in entropy and response length, as shown in Figure 4a and Figure 4b.

We introduce a **Token-level Policy Gradient Loss** in the long-CoT RL scenario to address the above



(a) 演员模型生成概率的熵。 **(b)** 演员模型生成响应的平均长度

Figure 4 演员模型的概率分布的熵，以及响应长度的变化。



(a) Performance on AIME. **(b)** Entropy of actor model.

Figure 5 演员模型在AIME上的准确性及其生成概率的熵，在应用**Overlong Reward Shaping**策略前后的情况。

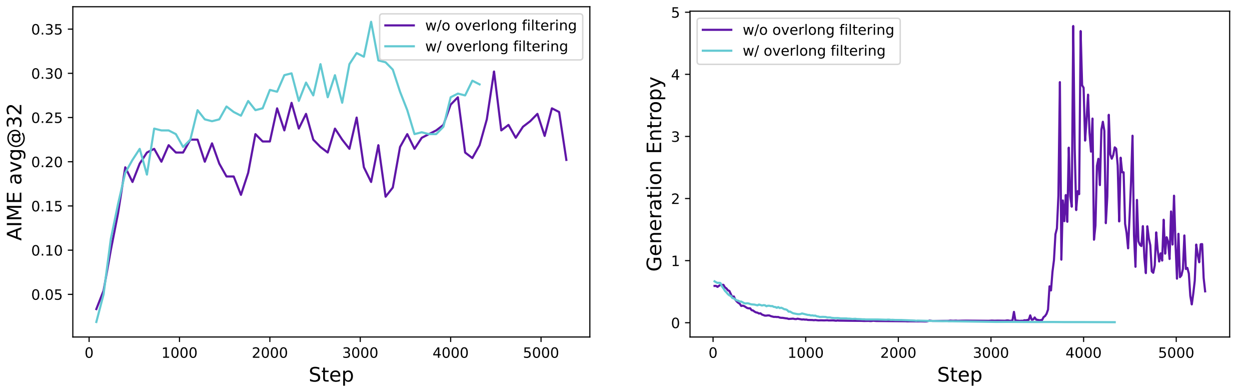
中，它都会被同等程度地鼓励或抑制。

3.4 Hide and Seek: Overlong Reward Shaping

在强化学习（RL）训练中，我们通常为生成设置最大长度，超过该长度的样本将被相应截断。我们发现，对于截断样本的奖励塑形不当可能会引入奖励噪声，并严重干扰训练过程。

默认情况下，我们会给截断样本分配一个惩罚性的奖励。这种方法可能会在训练过程中引入噪声，因为一个合理的推理过程可能仅仅因为其过长而受到惩罚。这种惩罚可能会使模型对其推理过程的有效性产生混淆。

为了研究这种奖励噪声的影响，我们首先应用了一种**过长过滤**策略，该策略屏蔽了截断样本的损失。我们发现这种方法显著稳定了训练并提高了性能，如Figure 5所示。
此外，我们提出了**Soft Overlong Punishment**（公式 13），这是一种针对截断样本设计的长度感知惩罚机制。具体来说，当响应长度超过预定义的最大值时，我们定义一个惩罚区间。在此区间内，响应越长，受到的惩罚越大。该惩罚被加到原始基于规则的正确性奖励上，从而向模型发出避免过长响应的



(a) Performance on AIME.

(b) Entropy of actor model.

Figure 5 The accuracy of the actor model on AIME and the entropy of its generation probabilities, both before and after applying **Overlong Reward Shaping** strategy.

limitations:

$$
\begin{aligned}
\mathcal{J}_{\text{DAPO}}(\theta) = \quad & \mathbb{E}_{(q,a)\sim\mathcal{D},\{o_i\}_{i=1}^G\sim\pi_{\theta_{\text{old}}}(\cdot|q)} \\
& \left[\frac{1}{\sum_{i=1}^G |o_i|}\sum_{i=1}^G\sum_{t=1}^{|o_i|}\min\left(r_{i,t}(\theta)\hat{A}_{i,t},\ \text{clip}\Big(r_{i,t}(\theta), 1-\varepsilon_{\text{low}}, 1+\varepsilon_{\text{high}}\Big)\hat{A}_{i,t}\right)\right], \\
\text{s.t.} \quad & 0 < \Big|\{o_i \mid \texttt{is\_equivalent}(a, o_i)\}\Big| < G.
\end{aligned}
\tag{12}
$$

In this setting, longer sequences can have more influence on the overall gradient update compared to shorter sequences. Moreover, from the perspective of individual tokens, if a particular generation pattern can lead to an increase or decrease in reward, it will be equally prompted or suppressed, regardless of the length of the response in which it appears.

3.4 Hide and Seek: Overlong Reward Shaping

In RL training, we typically set a maximum length for generation, with overlong samples truncated accordingly. We find that improper reward shaping for truncated samples can introduce reward noise and significantly disrupt the training process.

By default, we assign a punitive reward to truncated samples. This approach may introduce noise into the training process, as a sound reasoning process can be penalized solely due to its excessive length. Such penalties can potentially confuse the model regarding the validity of its reasoning process.

To investigate the impact of this reward noise, we first apply an **Overlong Filtering** strategy which masks the loss of truncated samples. We find that this approach significantly stabilizes training and enhances performance, as demonstrated in Figure 5.
Furthermore, we propose **Soft Overlong Punishment** (Equation 13), a length-aware penalty mechanism designed to shape the reward for truncated samples. Specifically, when the response length exceeds the predefined maximum value, we define a punishment interval. Within this interval, the longer the

Algorithm 1 **DAPO**: **D**ecoupled Clip and **D**ynamic s**A**mpling **P**olicy **O**ptimization

```
Input initial policy model $\pi_\theta$; reawrd model $R$; task prompts $\mathcal{D}$; hyperparameters $\varepsilon_{\texttt{low}}, \varepsilon_{\texttt{high}}$
for step = 1,...,M do
    Sample a batch $\mathcal{D}_b$ from $\mathcal{D}$
    Update the old policy model $\pi_{\theta_{old}} \leftarrow \pi_\theta$
    Sample $G$ outputs $\{o_i\}_{i=1}^G \sim \pi_{\theta_{\text{old}}}(\cdot|q)$ for each question $q \in \mathcal{D}_b$
    Compute rewards $\{r_i\}_{i=1}^G$ for each sampled output $o_i$ by running $R$
    Filter out $o_i$ and add the remaining to the dynamic sampling buffer (Dynamic Sampling Equation (11))
    if buffer size $n_b < N$:
        continue
    For each $o_i$ in the buffer, compute $\hat{A}_{i,t}$ for the $t$-th token of $o_i$ (Equation (9))
    for iteration = 1, ..., $\mu$ do
        Update the policy model $\pi_\theta$ by maximizing the DAPO objective (Equation (8))
Output $\pi_\theta$
```

信号。

$$
R_{\text{length}}(y) = \begin{cases} 0, & |y| \le L_{\max} - L_{\text{cache}} \\ \frac{(L_{\max} - L_{\text{cache}}) - |y|}{L_{\text{cache}}}, & L_{\max} - L_{\text{cache}} < |y| \le L_{\max} \\ -1, & L_{\max} < |y| \end{cases}
\tag{13}
$$

3.5 Dataset Transformation

我们的数据集来源于AoPS[1] 网站和官方竞赛主页，通过网络爬取和人工标注相结合的方式获取。数学数据集的答案通常有多种格式，例如表达式、公式和数字，这使得设计全面的规则来解析它们变得具有挑战性。为了使用规则提供准确的奖励信号，并尽量减少公式解析器引入的错误，我们受到AIME的启发，选择并将答案转换为易于解析的整数。例如，如果原始答案以 $\frac{a+\sqrt{b}}{c}$ 的形式表示，我们会指示大语言模型修改问题，使预期答案变为 $a+b+c$。经过选择和转换后，我们获得了 **DAPO-Math-17K** 数据集，该数据集包含17K个提示，每个提示都配有一个整数作为答案。

4 Experiments

4.1 Training Details

在本工作中，我们专注于数学任务来评估我们的算法，该算法可以轻松转移到其他任务。我们采用 verl 框架 [20] 进行训练，并使用 naive GRPO [38] 作为我们的基准算法，同时通过组奖励归一化估计优势。

对于超参数，我们使用 AdamW [39] 优化器，学习率恒定为 1×10^{-6}，并在 20 个 rollout 步骤中加入线性热身。对于 rollout，提示批次大小为 512，我们为每个提示采样 16 个响应。在训练中，我们将 mini-batch 大小设置为 512，即每个 rollout 步骤进行 16 次梯度更新。对于 **Overlong Reward Shaping**，我们将预期最大长度设为 16,384 个 token，并额外分配 4,096 个 token 作为软惩罚缓存。因此，生成

[1] https://artofproblemsolving.com/

Algorithm 1 **D**APO: **D**ecoupled Clip and **D**ynamic s**A**mpling **P**olicy **O**ptimization

```
Input initial policy model $\pi_\theta$; reawrd model $R$; task prompts $\mathcal{D}$; hyperparameters $\varepsilon_{\text{low}}, \varepsilon_{\text{high}}$
for step = 1,...,M do
    Sample a batch $\mathcal{D}_b$ from $\mathcal{D}$
    Update the old policy model $\pi_{\theta_{old}} \leftarrow \pi_\theta$
    Sample $G$ outputs $\{o_i\}_{i=1}^G \sim \pi_{\theta_{\text{old}}}(\cdot|q)$ for each question $q \in \mathcal{D}_b$
    Compute rewards $\{r_i\}_{i=1}^G$ for each sampled output $o_i$ by running $R$
    Filter out $o_i$ and add the remaining to the dynamic sampling buffer (Dynamic Sampling Equation (11))
    if buffer size $n_b < N$:
        continue
    For each $o_i$ in the buffer, compute $\hat{A}_{i,t}$ for the $t$-th token of $o_i$ (Equation (9))
    for iteration = 1, ..., $\mu$ do
        Update the policy model $\pi_\theta$ by maximizing the DAPO objective (Equation (8))
Output $\pi_\theta$
```

response, the greater the punishment it receives. This penalty is added to the original rule-based correctness reward, thereby signaling to the model to avoid excessively long responses.

$$R_{\text{length}}(y) = \begin{cases} 0, & |y| \le L_{\text{max}} - L_{\text{cache}} \\ \frac{(L_{\text{max}} - L_{\text{cache}}) - |y|}{L_{\text{cache}}}, & L_{\text{max}} - L_{\text{cache}} < |y| \le L_{\text{max}} \\ -1, & L_{\text{max}} < |y| \end{cases} \tag{13}$$

3.5 Dataset Transformation

Our dataset is sourced from the AoPS[1] website and official competition homepages through a combination of web scraping and manual annotation. The answers of math dataset typically come in a variety of formats, such as expression, formula and number, which makes it challenging to design comprehensive rules to parse them. To provide accurate reward signals using rules and minimize errors introduced by formula parsers, inspired by AIME, we select and transform the answers into integers, which are easy to parse. For example, if the original answer is expressed in the form of $\frac{a+\sqrt{b}}{c}$, we instruct the LLM to modify the question so that the expected answer becomes $a+b+c$. After selection and transformation, we obtained the **DAPO-Math-17K** dataset, which consists of 17K prompts, each paired with an integer as the answer.

4 Experiments

4.1 Training Details

In this work, we focus specifically on mathematical tasks to evaluate our algorithm, which can be readily transferred to other tasks. We adopt the verl framework [20] for training. We use naive GRPO [38] as

[1] https://artofproblemsolving.com/

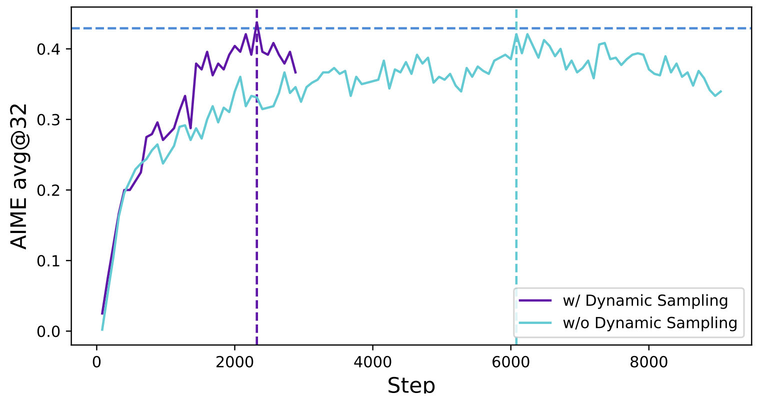


Figure 6 在基线设置中应用动态采样前后的训练进度。

的最大 token 数量被设置为 20,480 个 token。至于 **Clip-Higher** 机制，我们将裁剪参数 ε_{low} 设置为 0.2，$\varepsilon_{\text{high}}$ 设置为 0.28，这有效地平衡了探索与利用之间的权衡。在 AIME 的评估中，我们将评估集重复 32 次，并报告 avg@32 以确保结果的稳定性。评估的推理超参数被设置为温度 1.0 和 topp 0.7。

4.2 Main Results

AIME 2024上的实验表明，**DAPO**成功地将Qwen-32B Base模型训练成了一个强大的推理模型，其性能优于DeepSeek在Qwen2.5-32B上使用R1方法的实验结果。在图 1中，我们观察到AIME 2024上的性能有了显著提升，准确率从接近0%提高到了50%。值得注意的是，这一提升仅需DeepSeek-R1-Zero-Qwen-32B所需训练步数的50%。

我们在Table 1中详细分析了每种训练技术在我们方法中的贡献。所观察到的改进展示了这些技术在强化学习训练中的有效性，每种技术都在AIME 2024上贡献了几个准确率百分点。特别地，在标准的GRPO设置下，通过从Qwen2.5-32B基础模型开始训练，只能达到30%的准确率。

对于token级别的损失，尽管它带来的性能提升较少，但我们发现它增强了训练的稳定性，并使长度增长更加健康。

在应用**动态采样**时，虽然由于过滤掉零梯度数据需要采样更多数据，但整体训练时间并未受到显著影响。如Figure 6所示，尽管采样实例的数量增加，但由于所需的训练步数减少，模型的收敛时间甚至缩短了。

4.3 Training Dynamics

强化学习在大规模语言模型上的应用不仅是一个前沿的研究方向，也是一个本质上复杂的系统工程挑战，其特征在于各个子系统的相互依赖性。对任何一个子系统的修改都可能在整个系统中传播，并由于这些组件之间的复杂相互作用而导致意想不到的后果。即使是初始条件中的看似微小的变化，例如数据和超参数的变化，也可能通过迭代的强化学习过程放大，从而导致结果的巨大偏差。这种复杂性常常使研究人员面临一个困境：即使经过仔细分析并有充分的理由预期某项修改将改善训练过程的某些方面，实际结果却经常偏离预期轨迹。因此，在实验过程中监控关键的中间结果对于快速识别差异来源，并最终优化系统至关重要。

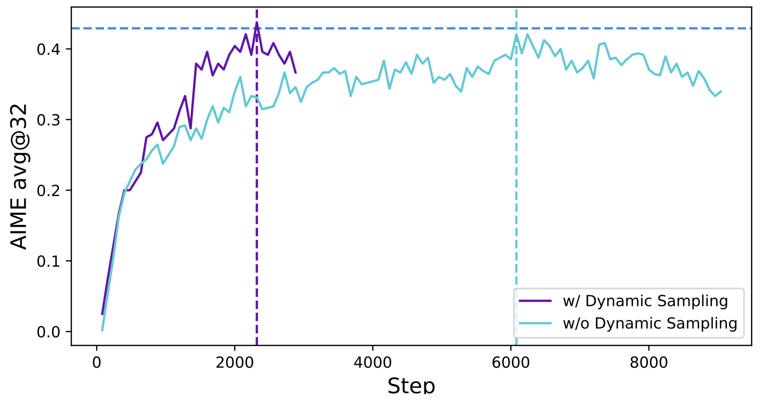


Figure 6 The training progress before and after applying dynamic sampling on a baseline setting.

our baseline algorithm and estimate advantages using group reward normalization.

For hyper-parameters, we utilize the AdamW [39] optimizer with a constant learning rate of 1×10^{-6}, incorporating a linear warm-up over 20 rollout steps. For rollout, the prompt batch size is 512 and we sample 16 responses for each prompt. For training, the mini-batch size is set to 512, i.e., 16 gradient updates for each rollout step. For **Overlong Reward Shaping**, we set the expected maximum length as 16,384 tokens and allocate additional 4,096 tokens as the soft punish cache. Therefore, the maximum number of tokens for generation is set to 20,480 tokens. As for the **Clip-Higher** mechanism, we set the clipping parameter ε_{low} to 0.2 and $\varepsilon_{\text{high}}$ to 0.28, which effectively balance the trade-off between exploration and exploitation. For evaluation on AIME, we repeat the evaluation set for 32 times and report avg@32 for results stability. The inference hyperparameters of evaluation are set to temperature 1.0 and topp 0.7.

4.2 Main Results

Experiments on AIME 2024 demonstrate that **DAPO** has successfully trained the Qwen-32B Base model into a powerful reasoning model, achieving performance superior to DeepSeek's experiments on Qwen2.5-32B using the R1 approach. In Figure 1, we observe a substantial improvement of performance on AIME 2024, with accuracy increasing from near 0% to 50%. Notably, this improvement is achieved with only 50% of the training steps required by DeepSeek-R1-Zero-Qwen-32B.

We analyze the contributions of each training technique in our methodology, as detailed in Table 1. The observed improvements demonstrate the effectiveness of these techniques in RL training, each contributing several accuracy points in AIME 2024. Notably, given the vanilla GRPO setting, only 30% accuracy can be reached by training from a Qwen2.5-32B base model.

For token-level loss, although it brings less performance improvement, we find it enhances training stability and makes the length increase more healthily.

When applying **Dynamic Sampling**, although more data needs to be sampled due to the filtering out of

Table 1 应用于**DAPO**的主要渐进技术结果

Model	**AIME24**$_{\text{avg@32}}$
DeepSeek-R1-Zero-Qwen-32B	47
Naive GRPO	30
+ Overlong Filtering	36
+ Clip-Higher	38
+ Soft Overlong Punishment	41
+ Token-level Loss	42
+ Dynamic Sampling (**DAPO**)	**50**

- **生成响应的长度** 是一个与训练稳定性及性能密切相关的重要指标，如Figure 7a所示。长度的增加为模型提供了更大的探索空间，使得更复杂的推理行为能够被采样，并通过训练逐渐得到强化。然而，需要注意的是，在训练过程中，长度并不总是保持持续上升的趋势。在某些相当长的阶段内，长度可能会出现停滞甚至下降的趋势，这一点在[2]中也得到了证实。我们通常将长度与验证集准确率结合，作为评估实验是否恶化的指标。
- **奖励的动力学** 在训练过程中一直是最关键的监控指标之一，如Figure 7b所示。在大多数实验中，奖励增加的趋势相对稳定，并不会由于实验设置的调整而产生显著波动或下降。这表明，只要有可靠的奖励信号，语言模型就能稳健地拟合训练集的分布。然而，我们发现训练集上的最终奖励往往与验证集上的准确率几乎没有相关性，这表明模型对训练集产生了过拟合。
- **演员模型的熵和生成概率** 与模型的探索能力相关，是我们实验中密切监控的关键指标。直观上，模型的熵需要保持在一个合适的范围内。过低的熵意味着概率分布过于尖锐，导致探索能力的丧失。相反，过高的熵通常伴随着过度探索的问题，例如生成无意义的内容和重复内容。对于生成概率，情况正好相反。如Section 3.1所示，通过应用Clip-Higher策略，我们有效地解决了熵崩溃的问题。在后续实验中，我们发现保持熵缓慢上升的趋势有助于提升模型性能，如Figure 7c和Figure 7d所示。

4.4 Case Study

在强化学习（RL）的训练过程中，我们观察到一个有趣的现象：演员模型（actor model）的推理模式会随着时间动态演化。具体来说，算法不仅强化了有助于正确解决问题的现有推理模式，还逐渐产生了最初并不存在的全新推理方式。这一发现揭示了强化学习算法的适应能力和探索能力，并为模型的学习机制提供了新的见解。

例如，在模型训练的早期阶段，几乎不会出现检查和反思先前推理步骤的行为。然而，随着训练的进行，模型表现出明显的反思和回溯行为，如表Table 2所示。这一观察结果为进一步探讨强化学习过程中推理能力的产生机制提供了启示，我们将此留待未来研究。

5 Conclusion

本文中，我们发布了一个完全开源的大规模LLM强化学习系统，包括算法、代码基础设施和数据集。该系统实现了最先进的大规模LLM强化学习性能（使用Qwen-32B预训练模型的AIME 50）。我们提出

Table 1 Main results of progressive techniques applied to **DAPO**

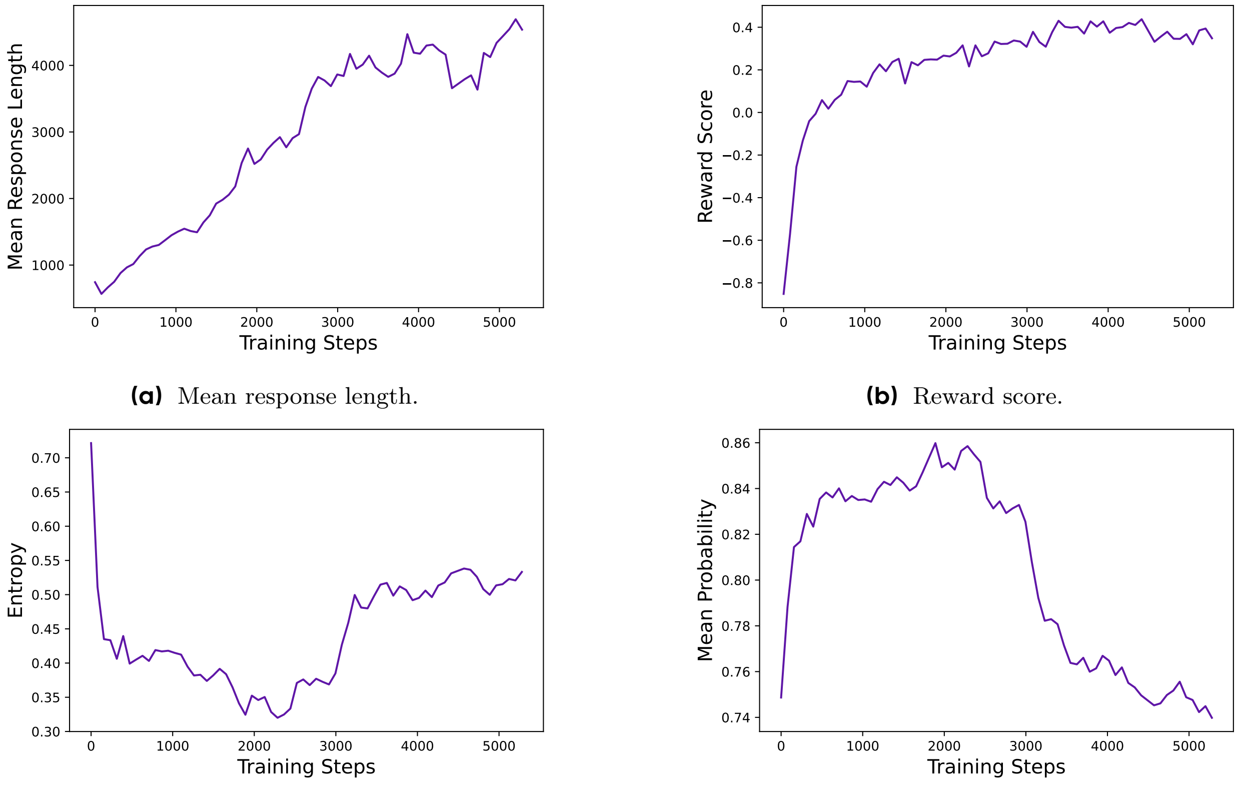
Model	**AIME24**$_{\text{avg@32}}$
DeepSeek-R1-Zero-Qwen-32B	47
Naive GRPO	30
+ Overlong Filtering	36
+ Clip-Higher	38
+ Soft Overlong Punishment	41
+ Token-level Loss	42
+ Dynamic Sampling (**DAPO**)	**50**

zero-gradient data, the overall training time is not significantly affected. As shown in Figure 6, although the number of sampling instances increases, the model's convergence time is even reduced, due to fewer training steps required.

4.3 Training Dynamics

Reinforcement learning on large language models is not only a cutting-edge research direction but also an intrinsically complex systems engineering challenge, characterized by the interdependence of its various subsystems. Modifications to any single subsystem can propagate through the system, leading to unforeseen consequences due to the intricate interplay among these components. Even seemingly minor changes in initial conditions, such as variations in data and hyperparameters, can amplify through iterative reinforcement learning processes, yielding substantial deviations in outcomes. This complexity often confronts researchers with a dilemma: even after meticulous analysis and well-founded expectations that a modification will enhance specific aspects of the training process, the actual results frequently diverge from the anticipated trajectory. Therefore, monitoring of key intermediate results during experimentation is essential for swiftly identifying the sources of discrepancies and, ultimately, for refining the system.

- **The Length of Generated Responses** is a metric closely related to training stability and performance, as shown in Figure 7a. The increase in length provides the model with a larger space for exploration, allowing more complex reasoning behaviors to be sampled and gradually reinforced through training. However, it is important to note that length does not always maintain a continuous upward trend during training. In some considerable periods, it can exhibit a trend of stagnation or even decline, which has also been demonstrated in [2]. We typically use length in conjunction with validation accuracy as indicators to assess whether an experiment is deteriorating.
- **The Dynamics of Reward** during training has always been one of the crucial monitoring indicators in reinforcement learning, as shown in Figure 7b. In the majority of our experiments, the trend of reward increase is relatively stable and does not fluctuate or decline significantly due to adjustments in experimental settings. This indicates that, given a reliable reward signal, language models can robustly fit the distribution of training set. However, we find that the final reward on the training set often exhibits little correlation with the accuracy on the validation set, which indicates



(a) Mean response length.

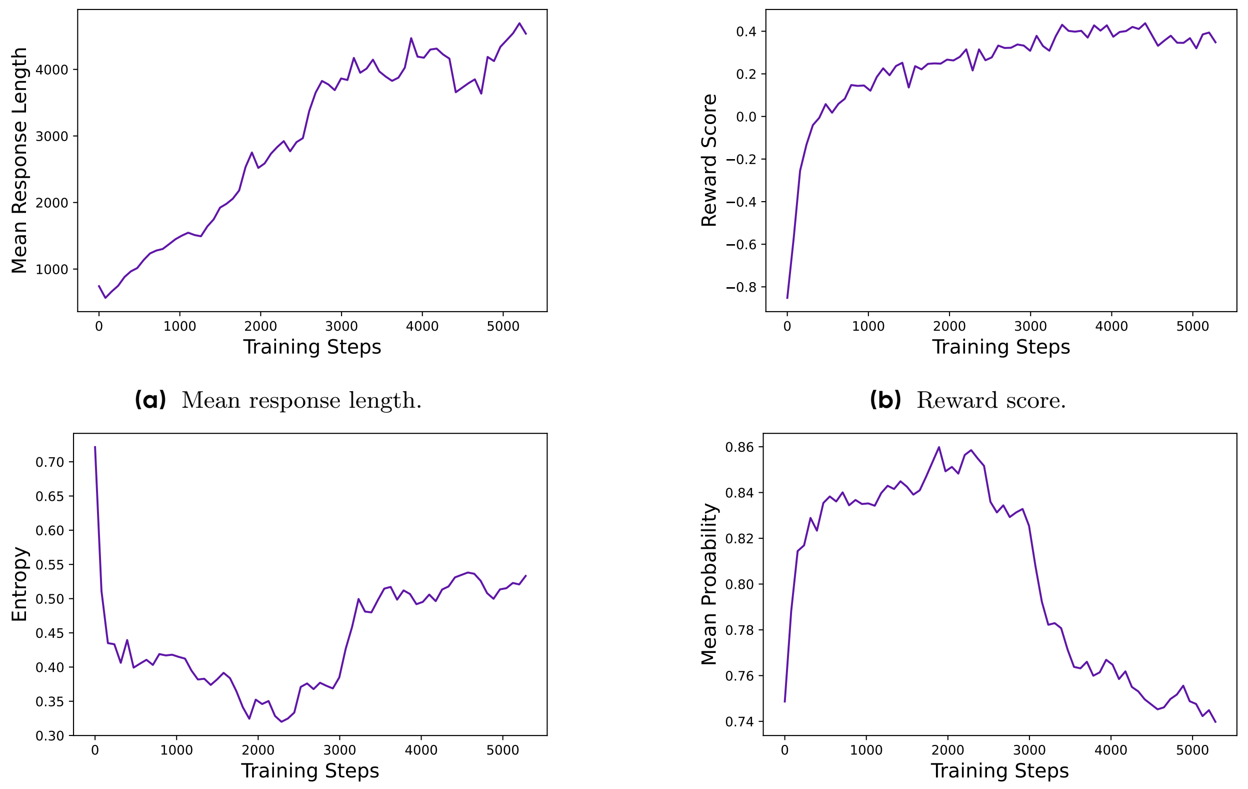
(b) Reward score.

(c) Generation entropy.

(d) Mean probability.

Figure 7 响应长度、奖励分数、生成熵和**DAPO**的平均概率的度量曲线，展示了RL训练的动力学特性，并作为识别潜在问题的重要监控指标。

了**解耦裁剪与动态采样策略优化（DAPO）**算法，并引入了4项关键技术，以在长链推理（long-CoT）的RL场景中大幅提升效率和效果。此外，通过开源训练代码和数据集，我们为更广泛的研究社区和社会提供了可实际操作的、可扩展的强化学习解决方案，使所有人能够从这些进步中受益。



(a) Mean response length.

(b) Reward score.

(c) Generation entropy.

(d) Mean probability.

Figure 7 The metric curves of response length, reward score, generation entropy, and the mean probability of **DAPO**, which show the dynamics of RL training and serve as essential monitoring indicators to identify potential issues.

overfitting to the training set.

- **The Entropy of the Actor Model and Generation Probability** are related to the model's exploration capability and are key metrics that we closely monitor in our experiments. Intuitively, the model's entropy needs to be maintained within an appropriate range. An excessively low entropy indicates that the probability distribution is overly sharp, leading to a loss of exploration capability. Conversely, an excessively high entropy is often associated with issues of over-exploration such as gibberish and repetitive generation. For the generation probability, the situation is exactly the opposite. As demonstrated in Section 3.1, by applying the Clip-Higher strategy, we effectively addressed the issue of entropy collapse. In subsequent experiments, we find that maintaining a slow upward trend in entropy is conducive to the improvement of model performance, shown in Figure 7c and Figure 7d.

4.4 Case Study

During the RL training process, we observe an interesting phenomenon: the reasoning patterns of the actor model evolve dynamically over time. Specifically, the algorithm not only reinforces existing reasoning patterns that facilitate correct problem-solving but also gradually gives rise to entirely new

Question:
Given that the base ABC of the tetrahedron $S-ABC$ is an equilateral triangle, the projection H of point A on the face SBC is the orthocenter of $\triangle SBC$, the dihedral angle $H-AB-C$ is 30°, and $SA=2$, find the volume of the tetrahedron. The answer is in the form $\frac{k}{m}$, so give the value of $k+m$.

Response:
Let us consider the given problem: We need to determine ...
...
Next, let S have coordinates (x_0, y_0, z_0). Since $SA = 2$, the distance formula gives us $\sqrt{(x_0)^2+(y_0)^2+(z_0)^2}=2$, so we can write $x_0^2+y_0^2+z_0^2=4$. Now, remember that H is the orthogonal projection of
...
However, wait a moment, let's rethink about the dihedral angle involving planes in a more thoughtful geometric way. Consider the plane $\alpha_1 = ABC$, which lies entirely on the xy coordinate plane (as $z=0$). Let the plane $\alpha_2 = SBC$. The point A projected perpendicularly to plane α_2 lands on H. The line $l = AB$ ...
...

Table 2 强化学习中反射行为的出现

Contributions

项目负责人

余启颖[1,2,4]

算法

余启颖[1,2,4]，张征[1]，朱若飞[1]，袁玉峰[1]，左晓辰[1]，岳雨

基础设施*

范甜甜[1]，刘高鸿[1]，刘灵君[1]，刘欣[1]，林海斌[1]，林志奇[1]，马博乐[1]，盛光明[1,3]，佟宇轩[1,2,4]，余启颖[1,2,4]，张驰[1]，张墨凡[1]，张旺[1]，朱航[1]，朱金花[1]

*按姓氏字母顺序排列

数据集

陈佳泽[1]，陈江杰[1,4]，王成毅[1]，余红丽[1,2,4]，戴文南[1,2,4]，宋玉轩[1,2,4]，魏翔鹏[1]，余启颖[1,2,4]

监督

周浩[2,4]，刘婧婧[2,4]，马维英[2,4]，张亚勤[2,4]，颜琳[1,4]，乔穆[1,4]，吴永辉[1]，王明轩[1,4]

隶属关系

[1]字节跳动种子

[2]清华大学人工智能产业研究院 (AIR)

[3]香港大学

Question:
Given that the base ABC of the tetrahedron $S-ABC$ is an equilateral triangle, the projection H of point A on the face SBC is the orthocenter of $\triangle SBC$, the dihedral angle $H-AB-C$ is 30°, and $SA = 2$, find the volume of the tetrahedron. The answer is in the form $\frac{k}{m}$, so give the value of $k + m$.

Response:
Let us consider the given problem: We need to determine ...

...

Next, let S have coordinates (x_0, y_0, z_0). Since $SA = 2$, the distance formula gives us $\sqrt{(x_0)^2 + (y_0)^2 + (z_0)^2} = 2$, so we can write $x_0^2 + y_0^2 + z_0^2 = 4$. Now, remember that H is the orthogonal projection of

...

However, wait a moment, let's rethink about the dihedral angle involving planes in a more thoughtful geometric way. Consider the plane $\alpha_1 = ABC$, which lies entirely on the xy coordinate plane (as $z = 0$). Let the plane $\alpha_2 = SBC$. The point A projected perpendicularly to plane α_2 lands on H. The line $l = AB$ ...

...

Table 2 Emergence of Reflective Behavior in Reinforcement Learning

modes of reasoning that were initially absent. This finding reveals the adaptability and exploration capability of RL algorithms and offers new insights into the learning mechanisms of the model.

For example, in the early stages of model training, there was virtually no occurrence of checking and reflecting on previous reasoning steps. However, as training progresses, the model exhibits distinct behaviors of reflection and backtracking, as shown in Table 2. This observation sheds light on further exploration into interpreting the emergence of reasoning abilities during RL, which we leave for future research.

5 Conclusion

In this paper, we release a fully open-sourced system for large-scale LLM RL, including algorithm, code infrastructure, and dataset. The system achieves state-of-the-art large-scale LLM RL performance (AIME 50 using Qwen-32B pretrained model). We propose the **D**ecoupled Clip and **D**ynamic s**A**mpling **P**olicy **O**ptimization (**DAPO**) algorithm, and introduce 4 key techniques to make RL powerfully effective and efficient in the long-CoT RL scenario. Additionally, by open-sourcing the training code and dataset, we provide the broader research community and society with practical access to a scalable reinforcement learning solution, enabling all to benefit from these advancements.

[4] 清华大学AIR与字节跳动种子联合实验室(SIA-Lab)

Acknowledgments

我们要感谢杜正银、沈凯、占天阳、肖振、郑仁杰、韩历、江凯华以及字节跳动的其他同事对**DAPO**项目的支持。

Contributions

Project Lead

Qiying Yu[1,2,4]

Algorithm

Qiying Yu[1,2,4], Zheng Zhang[1], Ruofei Zhu[1], Yufeng Yuan[1], Xiaochen Zuo[1], Yu Yue[1]

Infrastructure*

Tiantian Fan[1], Gaohong Liu[1], Lingjun Liu[1], Xin Liu[1], Haibin Lin[1], Zhiqi Lin[1], Bole Ma[1], Guangming Sheng[1,3], Yuxuan Tong[1,2,4], Qiying Yu[1,2,4], Chi Zhang[1], Mofan Zhang[1], Wang Zhang[1], Hang Zhu[1], Jinhua Zhu[1]

*Last-Name in Alphabetical Order

Dataset

Jiaze Chen[1], Jiangjie Chen[1,4], Chengyi Wang[1], Hongli Yu[1,2,4], Weinan Dai[1,2,4], Yuxuan Song[1,2,4], Xiangpeng Wei[1], Qiying Yu[1,2,4]

Supervision

Hao Zhou[2,4], Jingjing Liu[2,4], Wei-Ying Ma[2,4], Ya-Qin Zhang[2,4], Lin Yan[1,4], Mu Qiao[1,4], Yonghui Wu[1], Mingxuan Wang[1,4]

Affiliation

[1]ByteDance Seed

[2]Institute for AI Industry Research (AIR), Tsinghua University

[3]The University of Hong Kong

[4]SIA-Lab of Tsinghua AIR and ByteDance Seed

Acknowledgments

We thank Zhengyin Du, Kai Shen, Tianyang Zhan, Zhen Xiao, Renjie Zheng, Li Han, Kaihua Jiang as well as other colleagues at ByteDance for their support for the **DAPO** project.

References

[1] OpenAI. Learning to reason with llms, 2024.

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

[3] OpenAI. GPT4 technical report. arXiv preprint arXiv:2303.08774, 2023.

[4] Anthropic. Claude 3.5 sonnet, 2024.

[5] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. Advances in neural information processing systems, 33:1877–1901, 2020.

[6] Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, et al. Palm: Scaling language modeling with pathways. Journal of Machine Learning Research, 24(240):1–113, 2023.

[7] Aixin Liu, Bei Feng, Bing Xue, Bingxuan Wang, Bochao Wu, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chenyu Zhang, Chong Ruan, et al. Deepseek-v3 technical report. arXiv preprint arXiv:2412.19437, 2024.

[8] XAI. Grok 3 beta — the age of reasoning agents, 2024.

[9] Google DeepMind. Gemini 2.0 flash thinking, 2024.

[10] Qwen. Qwq-32b: Embracing the power of reinforcement learning, 2024.

[11] Kimi Team, Angang Du, Bofei Gao, Bowei Xing, Changjiu Jiang, Cheng Chen, Cheng Li, Chenjun Xiao, Chenzhuang Du, Chonghua Liao, et al. Kimi k1. 5: Scaling reinforcement learning with llms. arXiv preprint arXiv:2501.12599, 2025.

[12] An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, et al. Qwen2. 5 technical report. arXiv preprint arXiv:2412.15115, 2024.

[13] Zhipeng Chen, Yingqian Min, Beichen Zhang, Jie Chen, Jinhao Jiang, Daixuan Cheng, Wayne Xin Zhao, Zheng Liu, Xu Miao, Yang Lu, et al. An empirical study on eliciting and improving r1-like reasoning models. arXiv preprint arXiv:2503.04548, 2025.

[14] Jingcheng Hu, Yinmin Zhang, Qi Han, Daxin Jiang, and Heung-Yeung Shum Xiangyu Zhang. Open-reasoner-zero: An open source approach to scaling reinforcement learning on the base model. `https://github.com/Open-Reasoner-Zero/Open-Reasoner-Zero`, 2025.

[15] Jian Hu. Reinforce++: A simple and efficient approach for aligning large language models. arXiv preprint arXiv:2501.03262, 2025.

[16] Ganqu Cui, Lifan Yuan, Zefan Wang, Hanbin Wang, Wendi Li, Bingxiang He, Yuchen Fan, Tianyu Yu, Qixin Xu, Weize Chen, et al. Process reinforcement through implicit rewards. arXiv preprint arXiv:2502.01456, 2025.

[17] Jung Hyun Lee, June Yong Yang, Byeongho Heo, Dongyoon Han, and Kang Min Yoo. Token-supervised value models for enhancing mathematical reasoning capabilities of large language models. arXiv preprint arXiv:2407.12863, 2024.

References

[1] OpenAI. Learning to reason with llms, 2024.

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

[3] OpenAI. GPT4 technical report. arXiv preprint arXiv:2303.08774, 2023.

[4] Anthropic. Claude 3.5 sonnet, 2024.

[5] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. Advances in neural information processing systems, 33:1877–1901, 2020.

[6] Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, et al. Palm: Scaling language modeling with pathways. Journal of Machine Learning Research, 24(240):1–113, 2023.

[7] Aixin Liu, Bei Feng, Bing Xue, Bingxuan Wang, Bochao Wu, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chenyu Zhang, Chong Ruan, et al. Deepseek-v3 technical report. arXiv preprint arXiv:2412.19437, 2024.

[8] XAI. Grok 3 beta — the age of reasoning agents, 2024.

[9] Google DeepMind. Gemini 2.0 flash thinking, 2024.

[10] Qwen. Qwq-32b: Embracing the power of reinforcement learning, 2024.

[11] Kimi Team, Angang Du, Bofei Gao, Bowei Xing, Changjiu Jiang, Cheng Chen, Cheng Li, Chenjun Xiao, Chenzhuang Du, Chonghua Liao, et al. Kimi k1. 5: Scaling reinforcement learning with llms. arXiv preprint arXiv:2501.12599, 2025.

[12] An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, et al. Qwen2. 5 technical report. arXiv preprint arXiv:2412.15115, 2024.

[13] Zhipeng Chen, Yingqian Min, Beichen Zhang, Jie Chen, Jinhao Jiang, Daixuan Cheng, Wayne Xin Zhao, Zheng Liu, Xu Miao, Yang Lu, et al. An empirical study on eliciting and improving r1-like reasoning models. arXiv preprint arXiv:2503.04548, 2025.

[14] Jingcheng Hu, Yinmin Zhang, Qi Han, Daxin Jiang, and Heung-Yeung Shum Xiangyu Zhang. Open-reasoner-zero: An open source approach to scaling reinforcement learning on the base model. https://github.com/Open-Reasoner-Zero/Open-Reasoner-Zero, 2025.

[15] Jian Hu. Reinforce++: A simple and efficient approach for aligning large language models. arXiv preprint arXiv:2501.03262, 2025.

[16] Ganqu Cui, Lifan Yuan, Zefan Wang, Hanbin Wang, Wendi Li, Bingxiang He, Yuchen Fan, Tianyu Yu, Qixin Xu, Weize Chen, et al. Process reinforcement through implicit rewards. arXiv preprint arXiv:2502.01456, 2025.

[17] Jung Hyun Lee, June Yong Yang, Byeongho Heo, Dongyoon Han, and Kang Min Yoo. Token-supervised value models for enhancing mathematical reasoning capabilities of large language models. arXiv preprint arXiv:2407.12863, 2024.

[18] Amirhossein Kazemnejad, Milad Aghajohari, Eva Portelance, Alessandro Sordoni, Siva Reddy, Aaron Courville, and Nicolas Le Roux. Vineppo: Unlocking rl potential for llm reasoning through refined credit assignment. arXiv preprint arXiv:2410.01679, 2024.

[19] Yufeng Yuan, Yu Yue, Ruofei Zhu, Tiantian Fan, and Lin Yan. What's behind ppo's collapse in long-cot? value optimization holds the secret. arXiv preprint arXiv:2503.01491, 2025.

[20] Guangming Sheng, Chi Zhang, Zilingfeng Ye, Xibin Wu, Wang Zhang, Ru Zhang, Yanghua Peng, Haibin Lin, and Chuan Wu. Hybridflow: A flexible and efficient rlhf framework. arXiv preprint arXiv:2409.19256, 2024.

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

[22] John Schulman, Philipp Moritz, Sergey Levine, Michael Jordan, and Pieter Abbeel. High-dimensional continuous control using generalized advantage estimation, 2018.

[23] Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll 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. Training language models to follow instructions with human feedback. In S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh, editors, Advances in Neural Information Processing Systems, volume 35, pages 27730–27744. Curran Associates, Inc., 2022.

[24] Dario Amodei, Chris Olah, Jacob Steinhardt, Paul Christiano, John Schulman, and Dan Mané. Concrete problems in ai safety, 2016.

[25] Tom Everitt, Victoria Krakovna, Laurent Orseau, Marcus Hutter, and Shane Legg. Reinforcement learning with a corrupted reward channel, 2017.

[26] Victoria Krakovna, Jonathan Uesato, Vladimir Mikulik, Matthew Rahtz, Tom Everitt, Ramana Kumar, Zac Kenton, Jan Leike, and Shane Legg. Specification gaming: the flip side of ai ingenuity, 2020.

[27] Tom Everitt, Marcus Hutter, Ramana Kumar, and Victoria Krakovna. Reward tampering problems and solutions in reinforcement learning: A causal influence diagram perspective, 2021.

[28] Leo Gao, John Schulman, and Jacob Hilton. Scaling laws for reward model overoptimization, 2022.

[29] Lilian Weng. Reward hacking in reinforcement learning. lilianweng.github.io, Nov 2024.

[30] Stanislas Polu and Ilya Sutskever. Generative language modeling for automated theorem proving, 2020.

[31] Trieu H Trinh, Yuhuai Wu, Quoc V Le, He He, and Thang Luong. Solving olympiad geometry without human demonstrations. Nature, 625(7995):476–482, 2024.

[32] Trieu Trinh and Thang Luong. Alphageometry: An olympiad-level ai system for geometry, 2024.

[33] AlphaProof and AlphaGeometry Teams. Ai achieves silver-medal standard solving international mathematical olympiad problems, 2024.

[34] Hung Le, Yue Wang, Akhilesh Deepak Gotmare, Silvio Savarese, and Steven Chu Hong Hoi. Coderl: Mastering code generation through pretrained models and deep reinforcement learning. Advances in Neural Information Processing Systems, 35:21314–21328, 2022.

[35] Noah Shinn, Federico Cassano, Edward Berman, Ashwin Gopinath, Karthik Narasimhan, and Shunyu Yao. Reflexion: Language agents with verbal reinforcement learning, 2023.

[18] Amirhossein Kazemnejad, Milad Aghajohari, Eva Portelance, Alessandro Sordoni, Siva Reddy, Aaron Courville, and Nicolas Le Roux. Vineppo: Unlocking rl potential for llm reasoning through refined credit assignment. arXiv preprint arXiv:2410.01679, 2024.

[19] Yufeng Yuan, Yu Yue, Ruofei Zhu, Tiantian Fan, and Lin Yan. What's behind ppo's collapse in long-cot? value optimization holds the secret. arXiv preprint arXiv:2503.01491, 2025.

[20] Guangming Sheng, Chi Zhang, Zilingfeng Ye, Xibin Wu, Wang Zhang, Ru Zhang, Yanghua Peng, Haibin Lin, and Chuan Wu. Hybridflow: A flexible and efficient rlhf framework. arXiv preprint arXiv:2409.19256, 2024.

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

[22] John Schulman, Philipp Moritz, Sergey Levine, Michael Jordan, and Pieter Abbeel. High-dimensional continuous control using generalized advantage estimation, 2018.

[23] Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll 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. Training language models to follow instructions with human feedback. In S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh, editors, Advances in Neural Information Processing Systems, volume 35, pages 27730–27744. Curran Associates, Inc., 2022.

[24] Dario Amodei, Chris Olah, Jacob Steinhardt, Paul Christiano, John Schulman, and Dan Mané. Concrete problems in ai safety, 2016.

[25] Tom Everitt, Victoria Krakovna, Laurent Orseau, Marcus Hutter, and Shane Legg. Reinforcement learning with a corrupted reward channel, 2017.

[26] Victoria Krakovna, Jonathan Uesato, Vladimir Mikulik, Matthew Rahtz, Tom Everitt, Ramana Kumar, Zac Kenton, Jan Leike, and Shane Legg. Specification gaming: the flip side of ai ingenuity, 2020.

[27] Tom Everitt, Marcus Hutter, Ramana Kumar, and Victoria Krakovna. Reward tampering problems and solutions in reinforcement learning: A causal influence diagram perspective, 2021.

[28] Leo Gao, John Schulman, and Jacob Hilton. Scaling laws for reward model overoptimization, 2022.

[29] Lilian Weng. Reward hacking in reinforcement learning. lilianweng.github.io, Nov 2024.

[30] Stanislas Polu and Ilya Sutskever. Generative language modeling for automated theorem proving, 2020.

[31] Trieu H Trinh, Yuhuai Wu, Quoc V Le, He He, and Thang Luong. Solving olympiad geometry without human demonstrations. Nature, 625(7995):476–482, 2024.

[32] Trieu Trinh and Thang Luong. Alphageometry: An olympiad-level ai system for geometry, 2024.

[33] AlphaProof and AlphaGeometry Teams. Ai achieves silver-medal standard solving international mathematical olympiad problems, 2024.

[34] Hung Le, Yue Wang, Akhilesh Deepak Gotmare, Silvio Savarese, and Steven Chu Hong Hoi. Coderl: Mastering code generation through pretrained models and deep reinforcement learning. Advances in Neural Information Processing Systems, 35:21314–21328, 2022.

[35] Noah Shinn, Federico Cassano, Edward Berman, Ashwin Gopinath, Karthik Narasimhan, and Shunyu Yao. Reflexion: Language agents with verbal reinforcement learning, 2023.

[36] Xinyun Chen, Maxwell Lin, Nathanael Schärli, and Denny Zhou. Teaching large language models to self-debug, 2023.

[37] Jonas Gehring, Kunhao Zheng, Jade Copet, Vegard Mella, Quentin Carbonneaux, Taco Cohen, and Gabriel Synnaeve. Rlef: Grounding code llms in execution feedback with reinforcement learning, 2025.

[38] Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Mingchuan Zhang, YK Li, Y Wu, and Daya Guo. Deepseekmath: Pushing the limits of mathematical reasoning in open language models. arXiv preprint arXiv:2402.03300, 2024.

[39] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In International Conference on Learning Representations, 2019.

[36] Xinyun Chen, Maxwell Lin, Nathanael Schärli, and Denny Zhou. Teaching large language models to self-debug, 2023.

[37] Jonas Gehring, Kunhao Zheng, Jade Copet, Vegard Mella, Quentin Carbonneaux, Taco Cohen, and Gabriel Synnaeve. Rlef: Grounding code llms in execution feedback with reinforcement learning, 2025.

[38] Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Mingchuan Zhang, YK Li, Y Wu, and Daya Guo. Deepseekmath: Pushing the limits of mathematical reasoning in open language models. arXiv preprint arXiv:2402.03300, 2024.

[39] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In International Conference on Learning Representations, 2019.

Appendix

A Dataset Transformation

以下是一个数据转换的示例，我们展示了原始问题、转换后的问题以及在转换过程中的模型输出（原始内容为 Markdown 文本，已渲染以提高可读性）。

为了确保模型能够以高精度完成此任务，我们的目标是让模型进行全面推理，同时避免产生幻觉。我们通过提供一个清晰的思维框架来实现这一点，该框架鼓励深入推理。具体来说，我们采用了连贯推理（Chain of Thought, CoT）方法，并结合四个明确定义的步骤：提取答案格式、重述问题陈述、解决修改后的问题并以整数形式提供最终答案。对于每个步骤，我们都提供了少量示例解决方案或详细的指导，以引导模型的推理过程。

通过这种方法，我们观察到在大多数情况下，大语言模型（LLM）可以生成格式和质量均令人满意的重述版本。

Appendix

A Dataset Transformation

Here is an example of data transformation, where we demonstrate the original problem, the transformed problem, and the model outputs during the transformation process (the original is in markdown text, which has been rendered for readability).

To ensure the model can complete this task with high accuracy, our goal is to have it perform comprehensive reasoning while avoiding hallucinations. We achieve this by providing a clear thinking framework that encourages thorough reasoning. Specifically, we employ Chain of Thought (CoT) reasoning along with four well-defined steps: extract the answer format, rewrite the problem statement, solve the modified problem and provide an integer as the final answer. For each step, we provide few-shot solutions or detailed guidelines to guide the model's reasoning.

Through this approach, we have observed that in most cases, the LLM can generate reformulations with both format and quality that are satisfactory.

Original Problem

设 x 和 y 为满足 $x^2+y^2-22x-16y+113=0$ 的实数。确定 x 的最小可能值。

答案:$11-2\sqrt{6}$

Transformed Problem

设 x 和 y 为满足 $x^2+y^2-22x-16y+113=0$ 的实数。确定 x 的最小可能值。原答案的形式为 $k-m\sqrt{n}$，其中 k、m 和 n 为整数。请找出 $k+m+n$ 的值。**答案:**19

模型输出（Markdown文本）

转化思维过程

问题的原始答案是一个代数表达式 $11-2\sqrt{6}$. 我们需要将其转换为整数形式。首先，注意答案由一个整数部分和一个平方根项组成。根据给定规则，我们将整数部分和平方根部分用未定系数代替。原问题使用了变量 x 和 y，所以我们选择其他字母作为未定系数。我们把答案表示为 $k-m\sqrt{n}$ 并要求学生求出 $k+m+n$。

答案格式

$k-m\sqrt{n}$

转换后的问题

设 x 和 y 为满足以下条件的实数： $x^2+y^2-22x-16y+113=0$. 确定 x 的最小可能值。原答案的形式为 $k-m\sqrt{n}$，其中 k、m 和 n 为整数。请计算 $k+m+n$ 的值。

Answer Calculation for the Transformed Problem

- 原答案为：$11-2\sqrt{6}$。
- 所需格式为：$k-m\sqrt{n}$。
- 匹配系数：$k=11$，$m=2$，$n=6$。
- 最终结果为：$k+m+n=11+2+6=19$.

Final Answer

19

Original Problem

Let x and y be real numbers such that $x^2 + y^2 - 22x - 16y + 113 = 0$. Determine the smallest possible value of x. **Answer:** $11 - 2\sqrt{6}$

Transformed Problem

Let x and y be real numbers such that $x^2 + y^2 - 22x - 16y + 113 = 0$. Determine the smallest possible value of x. The original answer is in the form $k - m\sqrt{n}$, where k, m, and n are integers. Please find the value of $k + m + n$. **Answer:** 19

Model's output (Markdown Text)

Transformation Thought Process

The original answer to the problem is an algebraic expression $11 - 2\sqrt{6}$. We need to transform this into an integer form. First, note that the answer consists of an integer part and a square root term. According to the given rule, we replace the integer and square root components with undetermined coefficients. The original problem uses x and y as variables, so we choose other letters as undetermined coefficients. We express the answer in the form $k - m\sqrt{n}$ and require the student to find $k + m + n$.

Answer Format

$k - m\sqrt{n}$

Transformed Problem

Let x and y be real numbers such that $x^2 + y^2 - 22x - 16y + 113 = 0$. Determine the smallest possible value of x. The original answer is in the form $k - m\sqrt{n}$, where k, m, and n are integers. Please find the value of $k + m + n$.

Answer Calculation for the Transformed Problem

- The original answer is: $11 - 2\sqrt{6}$.
- The required format is: $k - m\sqrt{n}$.
- Matching the coefficients: $k = 11$, $m = 2$, $n = 6$.
- The final result is: $k + m + n = 11 + 2 + 6 = 19$.

Final Answer

19

B Supplementary Case

Question:

Among the 900 residents of Aimeville, there are 195 who own a diamond ring, 367 who own a set of golf clubs, and 562 who own a garden spade. In addition, each of the 900 residents owns a bag of candy hearts. There are 437 residents who own exactly two of these things, and 234 residents who own exactly three of these things. Find the number of residents of Aimeville who own all four of these things.

Response:

Step by Step Solution:
1. Let's define the following variables:
Let A be the set of residents who own a diamond ring.
Let B be the set of residents who own a set of golf clubs.
Let C be the set of residents who own a garden spade.
Let D be the set of residents who own a bag of candy hearts. Since all 900 residents own a bag of candy hearts, $|D| = 900$.
Let a_i denote the number of residents who own exactly i out of the 4 items.
Let's use the principle of inclusion-exclusion for the union of 4 sets A, B, C, D.
...
$a_4 = \frac{-219}{-4} = \frac{219}{4} = 54.75$.
Since a_4, the number of people owning all four items, must be a whole number, our current approach needs to be reconsidered in terms of the combinatorial counting of people owning different combinations of the items.
Instead of directly using the inclusion-exclusion principle, we can use a different approach involving the number of people owning combinations of the items A, B, C, and D.
Let x_i be the number of people owning exactly i item among A, B, C (excluding the item D which everyone owns).
Then the total number of people owning exactly k items among the A, B, C, and D can be broken down:
...

Table 3 A case of reflective behavior emergence.

B Supplementary Case

Question:

Among the 900 residents of Aimeville, there are 195 who own a diamond ring, 367 who own a set of golf clubs, and 562 who own a garden spade. In addition, each of the 900 residents owns a bag of candy hearts. There are 437 residents who own exactly two of these things, and 234 residents who own exactly three of these things. Find the number of residents of Aimeville who own all four of these things.

Response:

Step by Step Solution:
1. Let's define the following variables:
Let A be the set of residents who own a diamond ring.
Let B be the set of residents who own a set of golf clubs.
Let C be the set of residents who own a garden spade.
Let D be the set of residents who own a bag of candy hearts. Since all 900 residents own a bag of candy hearts, $|D| = 900$.
Let a_i denote the number of residents who own exactly i out of the 4 items.
Let's use the principle of inclusion-exclusion for the union of 4 sets A, B, C, D.
...
$a_4 = \frac{-219}{-4} = \frac{219}{4} = 54.75$.
Since a_4, the number of people owning all four items, must be a whole number, our current approach needs to be reconsidered in terms of the combinatorial counting of people owning different combinations of the items.
Instead of directly using the inclusion-exclusion principle, we can use a different approach involving the number of people owning combinations of the items A, B, C, and D.
Let x_i be the number of people owning exactly i item among A, B, C (excluding the item D which everyone owns).
Then the total number of people owning exactly k items among the A, B, C, and D can be broken down:
...

Table 3 A case of reflective behavior emergence.