DAPO

💡 字节与清华强强联合，推出了开源的DAPO算法，基于 Qwen2.5-32B 模型在 AIME 2024 上超过了之前最先进的 DeepSeek-RL-Zero-Qwen-32B，在long-CoT场景大放异彩。这个让模型更聪明的秘诀，藏在四大黑科技里：

- **解耦裁剪**：提高clip上界，避免熵崩溃，既保持了思维的严谨性，又让模型的回答充满惊喜创意。
- **动态采样**：过滤掉准确率为 1 和 0 的数据，自动过滤掉太简单或超纲的题目，提升训练效率和稳定性。
- **token级梯度损失**：提升长序列样本中的token对整体损失的影响，使得模型能够更好地学习长序列中的推理模式（long-CoT）。
- **过长样本奖励调整**：对过长样本的惩罚进行平滑处理，用渐进式调整让模型明白：不是文章越长越好，而是要把复杂问题说清楚。

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

作者先是吐槽了下gpt1 o1 和 deepseek R1，虽然他们通过测试时扩展（Test-time scaling）使得模型能够进行更长的链式思考（Chain-of-Thought, CoT），在推理任务上表现出色，但都隐藏了大规模强化学习的细节，导致难以复现其成果。

例如在AIME 2024测试上，DeepSeek-R1-Zero-Qwen-32B取得了47分，但作者复现的成果只有30分。通过分析发现GRPO面临着许多关键问题：

- **熵崩溃**：策略的熵（不确定性）逐渐趋近于零，导致智能体行为过于确定化，失去探索能力。
- **奖励噪音**：奖励信号存在随机波动或误差，导致模型接收到的奖励信息不准确、不稳定，从而影响策略学习的效果。
- **训练不稳定**：在强化学习模型的训练过程中，性能指标（如奖励、损失等）出现大幅波动、震荡或不收敛的现象，导致模型无法稳定地学习到有效的策略。

为了解决以上问题，作者开源了DAPO（Decoupled Clip and Dynamic sAmpling Policy Optimization）算法来提升long-CoT场景的效果，提出了4个关键创新：

🐵
- **Clip-Higher**：通过解耦上下剪裁范围，避免熵崩溃，提升系统的多样性。
- **Dynamic Sampling**：动态采样策略，提升训练效率和稳定性。
- **Token-Level Policy Gradient Loss**：在长链推理（long-CoT）场景中，使用基于token的策略梯度损失，避免长序列样本对梯度的负面影响。

- **Overlong Reward Shaping**：通过软惩罚机制，减少过长样本的奖励噪声，稳定训练过程。

首先回顾下GRPO的公式：

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

where

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

- 可以看到GRPO的训练目标是**sample-level**的，也就是GRPO 首先会计算每个生成序列中的平均损失，然后再对不同样本的损失进行平均。因此DAPO中采用**Token-Level Policy Gradient Loss**来避免这种差异对性能的影响。
- 在RLHF场景中，KL散度用于限制在线策略与冻结参考策略之间的偏离。然而，在long-CoT推理模型的训练中，模型分布可能会显著偏离初始模型，这样的KL-散度约束没有意义。DAPO算法移除了KL散度项，从而允许模型在训练过程中自由探索。
- 传统的奖励模型往往面临**奖励劫持**问题（即模型通过操纵奖励信号来获得高分，而非真正提升推理能力）。DAPO直接使用可验证任务的最终准确率作为奖励（这也是基于规则的奖励模型的一种表现形式），这种方法在自动化定理证明、编程、数学等推理场景效果良好。

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

DAPO的核心在于**解耦裁剪**和**动态采样策略优化**，基于以上改进，再结合后续的优化，先看下DAPO的更新公式：

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

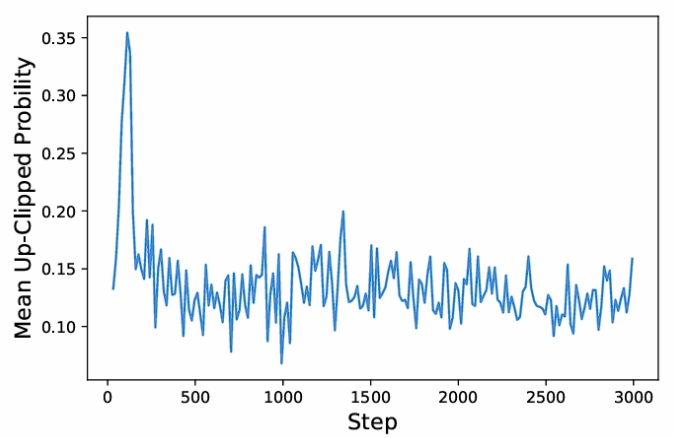
where

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

Clip-Higher

在PPO和GRPO训练中，经常会出现**熵崩溃问题**，即训练过程中，策略的熵值迅速降低，部分采样响应近乎相同，这严重限制了模型的探索能力，导致采样空间崩塌。这是由于采用的固定的clip范围来限制策略更新，但其中的上界裁剪限制了策略的探索性，也就是虽然增强了exploitation能力，而降低了exploration能力（clip操作的下界保障更新速度不太慢，clip操作上界保障更新速度不太快）。

举例说明，假设当前的优势值 $A_{i,t} > 0$，$\epsilon = 0.2$，那就要增大当前回复中token的概率，如果最开始两个token的生成概率分别是 $\pi_{\theta_{old}}(o_i|q) = 0.01$和$0.9$，经过clip更新后的策略的最大概率分别是 $\pi_\theta(o_i|q) = 0.012$和$1.08$，对于低概率token概率几乎没有增长，这表明clip上界限制了低概率token概率的增长。

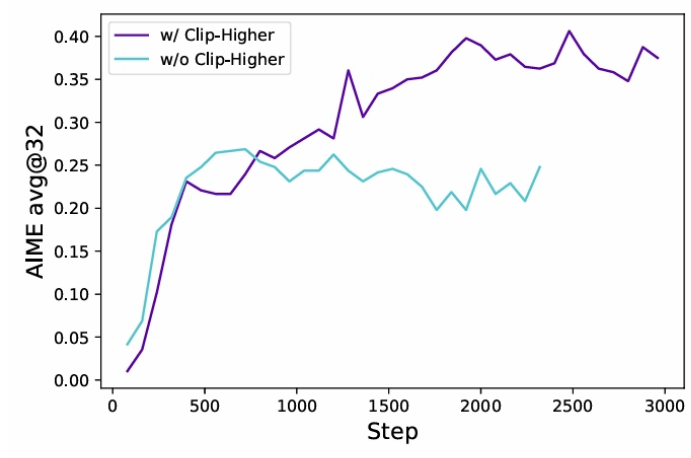


如右图所示，低概率token经过clip后的最大概率约为 $\pi_\theta(o_i|q) < 0.2$。这印证了clip上界限制了对低概率token的探索，会限制long CoT推理性能以及推理范式的多样性。

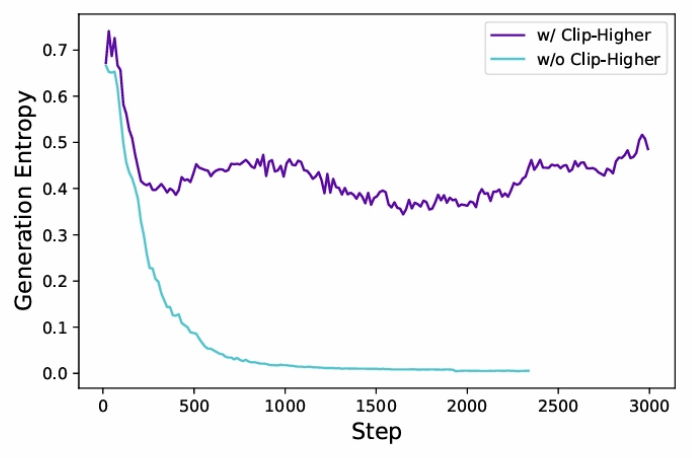
而Clip-Higher，顾名思义，就是通过解耦上下剪裁范围，将上下剪裁范围分别设置为 ϵ_{low} 和 ϵ_{high}，其中 ϵ_{high} 较大，允许低概率token有更大的更新空间:

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

具体来说，将下裁剪范围 ϵ_{low} 设置为0.2，上裁剪范围 ϵ_{high} 设置为0.28，从而在保持稳定性的同时提升策略的多样性。如下图所示，策略的熵显著增加，能生成更丰富的样本。同时 ϵ_{low} 保持较低的值，避免将低概率token的概率压缩至0，从而避免采样空间崩塌。



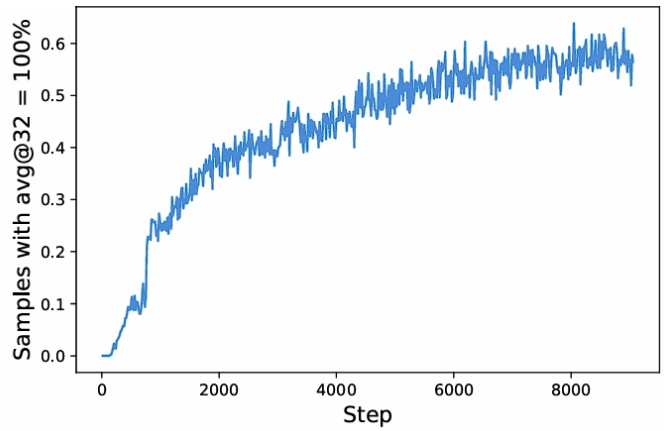
(a) Accuracies on AIME.



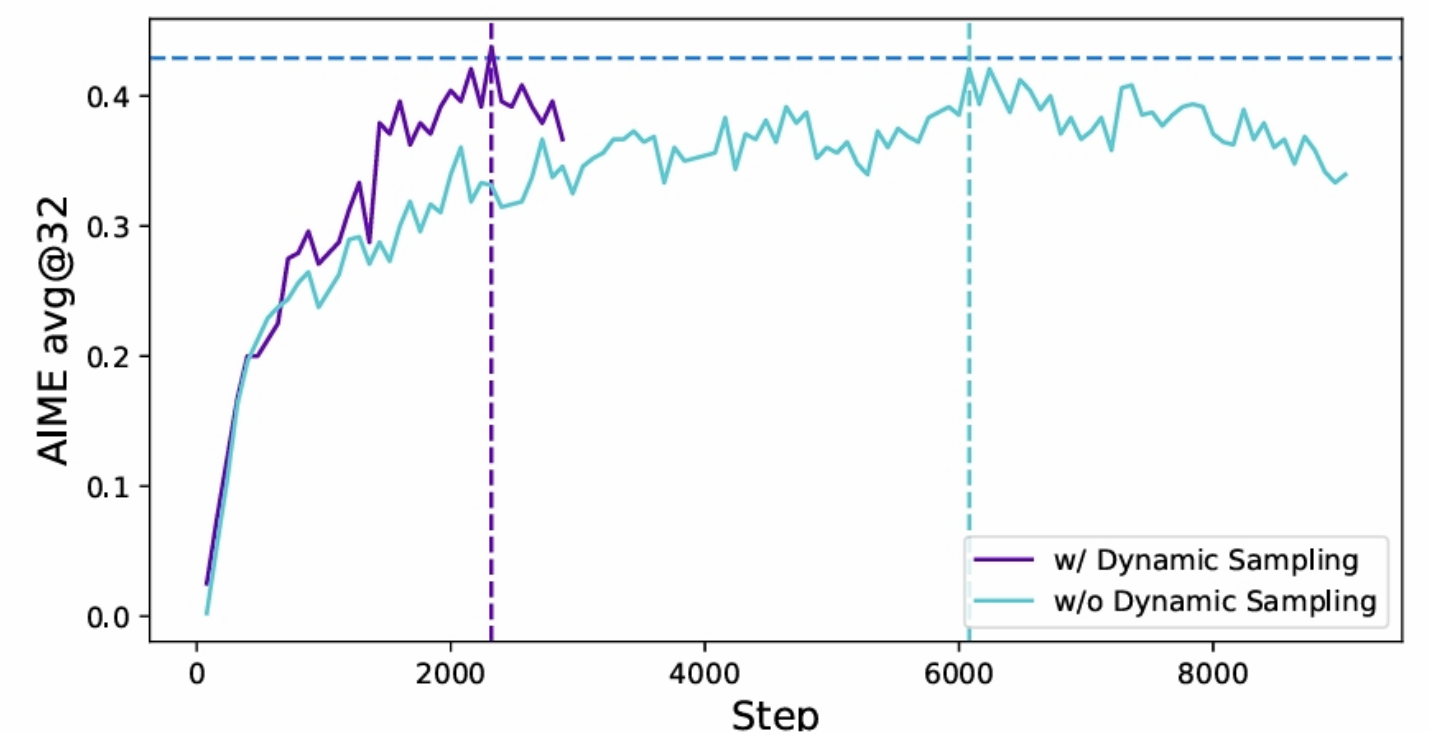
(b) Entropy of actor model.

Dynamic Sampling

在RL训练中，某些prompt准确率是1时(在GRPO中，就是组内所有输出都是正确的，奖励都是1)，会导致这些样本的优势和梯度为零，从而降低训练效率。如右图所示，准确率为1的prompt随着训练而逐渐增加，这会导致每个批次中有效数据量减低，从而导致梯度方差增大。



为此，DAPO提出动态采样策略，进行过采样（over-sample）并过滤掉准确率为 1 和 0 的数据，直到批次中充满准确率既不是 0 也不是 1 的样本，这样能让各个批次中的样本数量一致，并且都保持有效梯度。实验结果如下图证明，动态采样能更快地实现相同的性能。



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

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

Token-Level Policy Gradient Loss

在long-CoT场景中，GRPO算法使用**sample-level**损失计算，即先对每个样本内的token损失求平均，再对所有样本的损失求平均，在最终损失计算中每个样本权重相等。这种计算方式会导致长序列样本中的token对整体损失的贡献较小，造成以下负面影响：

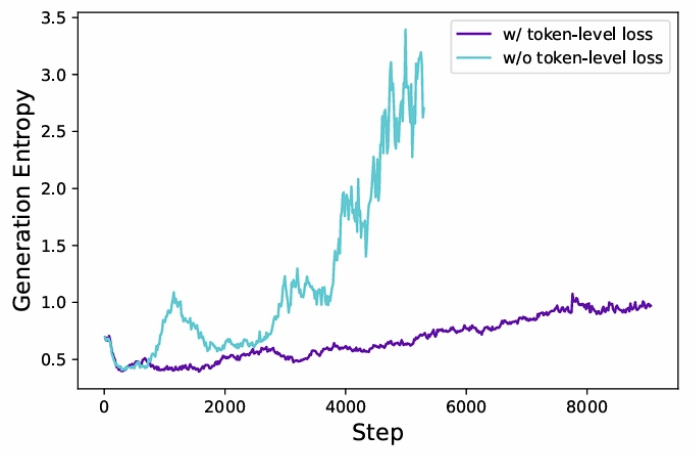
- 对于高质量长样本，阻碍模型学习高质量长样本中的推理相关模式
- 对于低质量样本，无法有效惩罚过长样本中的低质量内容（如乱码和重复词），导致熵和响应长度不合理增加。

$$\frac{1}{G}\sum_{i=1}^{G}\frac{1}{|o_i|}\sum_{t=1}^{|o_i|}$$

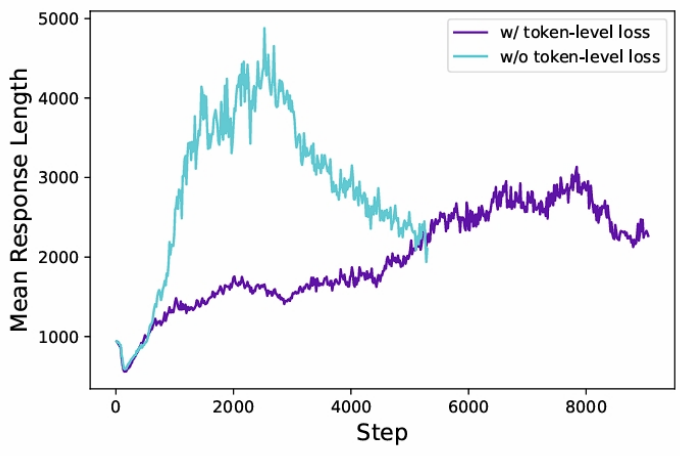
为了解决long-CoT场景中面临的以上问题，DAPO采用**Token-Level Policy Gradient Loss**：

$$\frac{1}{\sum_{i=1}^{G}|o_i|}\sum_{i=1}^{G}\sum_{t=1}^{|o_i|}$$

这样就保证了长序列相比短序列对整体梯度更新影响更大，模型能够更好地学习长序列中的推理模式，减少了低质量长序列样本（如重复或无意义内容）的影响。并且从单个 token 角度考虑，无论其所在响应长度如何，都会被施以同等的激励或抑制信号。



(a) Entropy of actor model's generation probabilities.



(b) Average length of actor model-generated responses

Overlong Reward Shaping

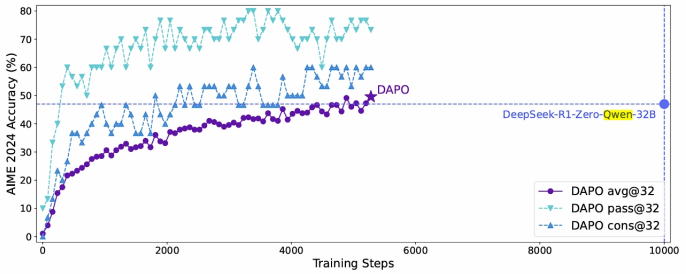
在RL训练中，通常会设置一个最大生成长度，超过该长度的样本会被截断。传统的做法是对截断样本施加惩罚性奖励（如-1），但这种做法会引入奖励噪声和干扰训练过程，尤其是当模型生成了合理的推理过程但仅仅因为长度过长而被截断时。

作者首先采用了**过滤超长样本**策略，屏蔽截断样本的损失，就已经能提升性能。进一步的还提出Soft Overlong Punishment，对过长样本的惩罚进行平滑处理。具体来说，DAPO定义了一个惩罚区间，当样本长度超过最大长度时，惩罚会随着长度的增加而逐渐加重，而不是直接施加一个固定的惩罚。这种惩罚会添加到基于规则的原始正确性奖励中，从而向模型发出信号，避免过长的响应。

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

实验结果

使用Qwen2.5-32B作为预训练模型进行RL训练，模型在AIME 2024竞赛中取得了50分的成绩，超过了DeepSeek的47分，且仅使用了50%的训练步数。



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**