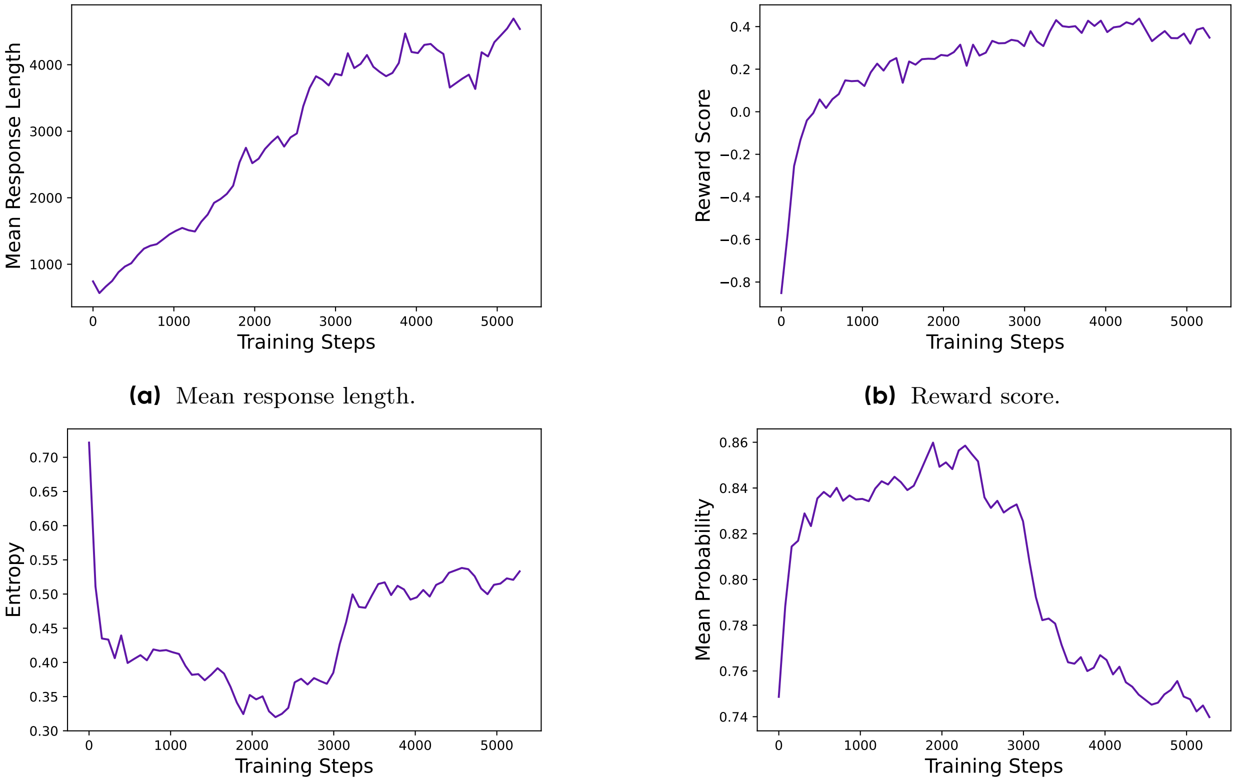

**(a)** Mean response length.

**(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.

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 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

**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 $30^\circ$, 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 $\alpha_2$ lands on $H$. The line $l = AB$ ...
...

**Table 2** Emergence of Reflective Behavior in Reinforcement Learning

**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.

## 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*

Weinan Dai[1,2,4], Tiantian Fan[1], Gaohong Liu[1], Juncai 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], Ru 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], 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, Shengding Hu, 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.