ByteDance | Seed

VAPO: Efficient and Reliable Reinforcement Learning for Advanced Reasoning Tasks

ByteDance Seed

Full author list in Contributions

Abstract

We present VAPO, **V**alue-model-based **A**ugmented Proximal **P**olicy **O**ptimization framework for reasoning models., a novel framework tailored for reasoning models within the value-model-based paradigm. Benchmarked the AIME 2024 dataset, VAPO, built on the Qwen 32B pre-trained model, attains a state-of-the-art score of **60.4**. In direct comparison under identical experimental settings, VAPO outperforms the previously reported results of DeepSeek-R1-Zero-Qwen-32B and DAPO by more than 10 points. The training process of VAPO stands out for its stability and efficiency. It reaches state-of-the-art performance within a mere 5,000 steps. Moreover, across multiple independent runs, no training crashes occur, underscoring its reliability. This research delves into long chain-of-thought (long-CoT) reasoning using a value-model-based reinforcement learning framework. We pinpoint three key challenges that plague value-model-based methods: value model bias, the presence of heterogeneous sequence lengths, and the sparsity of reward signals. Through systematic design, VAPO offers an integrated solution that effectively alleviates these challenges, enabling enhanced performance in long-CoT reasoning tasks.

Date: April 14, 2025
Correspondence: Yu Yue at yueyu@bytedance.com

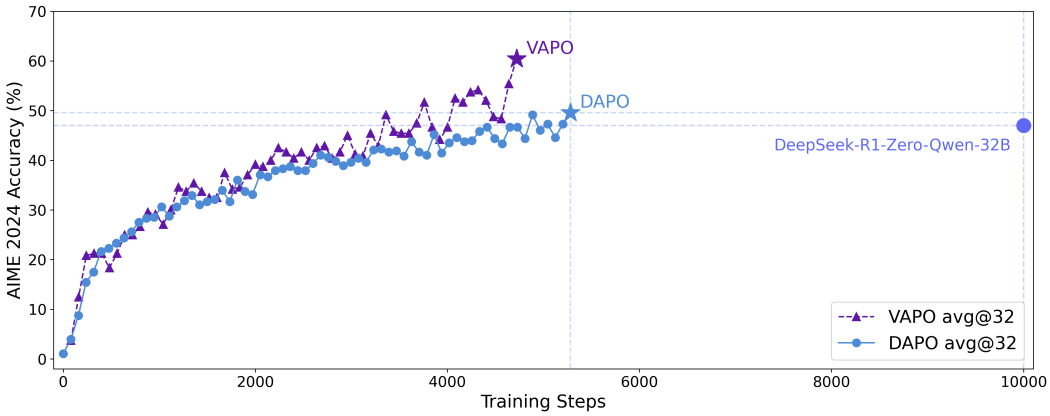

Figure 1 AIME 2024 scores of **VAPO** on the Qwen2.5-32B base model, demonstrates significant superiority over the previous state-of-the-art (SOTA) method DAPO, achieving this with notably fewer training steps. The x-axis denotes the gradient update steps.

1 Introduction

Reasoning models [5, 19, 26] such as OpenAI O1 [16] and DeepSeek R1 [6] have significantly advanced artificial intelligence by exhibiting remarkable performance in complex tasks such as mathematical reasoning, which demand step-by-step analysis and problem-solving through long chain-of-thought (CoT) [27] at test time. Reinforcement learning (RL) plays a pivotal role in the success of these models [1, 8, 10, 13, 22, 24, 26, 29]. It gradually enhances the model's performance by continuously exploring reasoning paths toward correct answers on verifiable problems, achieving unprecedented reasoning capabilities.

In the Large Language Models (LLM) [2–4, 11, 15, 25, 28] RL training, value-model-free methods like GRPO [22] and DAPO [29] have demonstrated remarkable effectiveness. These approaches eliminate the computational overhead of learning a value model, instead computing advantage solely based on the final reward of the entire trajectory. The trajectory-level advantage is then directly assigned as the token-level advantage for each position in the sequence. When training a reliable value model is particularly challenging, value-model-free methods deliver an accurate and stable baseline for advantage calculation by averaging the rewards across multiple trajectories within a group. This group-based reward aggregation mitigates the need for explicit value estimation, which often suffers from instability in complex tasks. Consequently, value-model-free methods have gained significant traction in addressing difficult problems such as long-CoT reasoning, with substantial research efforts focused on optimizing their frameworks.

Despite the notable success achieved by the value-model-free methods, we argue that value-model-based approaches possess a higher performance ceiling if the challenges in training value models can be addressed. First, value models enable more precise credit assignment by accurately tracing the impact of each action on subsequent returns, facilitating finer-grained optimization [21]. This is particularly critical for complex reasoning tasks, where subtle errors in individual steps often lead to catastrophic failures, and it remains challenging for model optimizing under value-model-free frameworks [30]. Secondly, in contrast to the advantage estimates derived from Monte Carlo methods in value-model-free approaches, value models can provide lower-variance value estimates for each token, thereby enhancing training stability. Furthermore, a well-trained value model exhibits inherent generalization capabilities, enabling more efficient utilization of samples encountered during online exploration. This significantly elevates the optimization ceiling of reinforcement learning algorithms. Consequently, despite the formidable challenges in training value models for complex problems, the potential benefits of overcoming these difficulties are substantial.

However, training a perfect value model in Long COT tasks presents significant challenges. First, learning a low-bias value model is non-trivial given the long trajectory and the instability of learning value in a bootstrapped way. Second, handling both short and long responses simultaneously is also challenging, as they might exhibit very distinct preferences towards the bias-variance trade-off during optimization. Last but not least, the sparsity of the reward signal from verifiers is further exacerbated by the long CoT pattern, which intrinsically requires better mechanisms to balance exploration and exploitation. To address the aforementioned challenges and fully unleash the potential of value-model-based methods in reasoning tasks, we present **V**alue **A**ugmented proximal **P**olicy **O**ptimization (**VAPO**), a value-model-based RL training framework. VAPO draws inspiration from prior research works such as VC-PPO [30] and DAPO [29], and further extends their concepts.

We summarize our key contributions as follows:

1. We introduce VAPO, the first value-model-based RL training framework to outperform value-model-free methods on long COT tasks significantly. VAPO not only demonstrates remarkable superiority in terms of performance but also showcases enhanced training efficiency, streamlining the learning process and underscoring its potential as a new benchmark in the field.
2. We propose Length-adaptive GAE, which adaptively adjusts the λ parameter in GAE computation based on response lengths. By doing so, it effectively caters to the distinct bias-variance trade-off requirements associated with responses of highly variable lengths. As a result, it optimizes the accuracy and stability of the advantage estimation process, particularly in scenarios where the length of the data sequences varies widely.