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-based **A**ugmented Proximal **P**olicy **O**ptimization framework for reasoning models., a novel framework tailored for reasoning models within the value-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-based reinforcement learning framework. We pinpoint three key challenges that plague value-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: 2025 年 4 月 11 日

Correspondence: Yu Yue at yueyu@bytedance.com

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.

ByteDance | Seed

VAPO: Efficient and Reliable Reinforcement Learning for Advanced Reasoning Tasks

ByteDance Seed

Full author list in Contributions

Abstract

*警告： 该PDF由GPT-Academic开源项目调用大语言模型+Latex翻译插件一键生成， 版权归原文作者所有。 翻译内容可靠性无保障， 请仔细鉴别并以原文为准。 项目Github地址 https://github.com/binary-husky/gpt_academic/。当前大语言模型：qwen-plus，当前语言模型温度设定：0.2。为了防止大语言模型的意外谬误产生扩散影响，禁止移除或修改此警告。

我们提出了VAPO（**V**alue-based **A**ugmented Proximal **P**olicy **O**ptimization framework for reasoning models），一个专为基于价值范式的推理模型设计的全新框架。在AIME 2024数据集上的基准测试中，基于Qwen 32B预训练模型构建的VAPO取得了**60.4**的最先进分数。在相同的实验设置下，VAPO的表现比之前报道的DeepSeek-R1-Zero-Qwen-32B和DAPO高出超过10分。VAPO的训练过程以其稳定性和高效性脱颖而出。它仅需5,000步即可达到最先进水平。此外，在多次独立运行中，未发生任何训练崩溃的情况，证明了其可靠性。本研究探讨了使用基于价值的强化学习框架进行长链推理（long-CoT）的能力。我们确定了困扰基于价值方法的三个关键挑战：价值模型偏差、异构序列长度的存在以及奖励信号的稀疏性。通过系统的设计，VAPO提供了一个综合解决方案，有效缓解了这些挑战，从而提升了在长链推理任务中的表现。

Date: 2025 年 4 月 11 日

Correspondence: Yu Yue at yueyu@bytedance.com

1 Introduction

推理模型 [5, 19, 26]，例如 OpenAI O1 [16] 和 DeepSeek R1 [6]，通过在数学推理等复杂任务中表现出显著的性能，这些任务需要逐步分析和通过长时间链式思考（CoT） [27] 来解决问题，极大地推动了人工智能的发展。强化学习（RL）在这些模型的成功中起到了关键作用 [1, 8, 10, 13, 22, 24, 26, 29]。它通过在可验证的问题上不断探索通往正确答案的推理路径，逐步提高模型的性能，实现了前所未有的推理能力。

在大规模语言模型（LLM） [2–4, 11, 15, 25, 28] 的强化学习训练中，无价值方法如 GRPO [22] 和 DAPO [29] 展现了显著的效果。这些方法消除了学习价值模型的计算开销，而是仅根据整个轨迹的最终奖励来

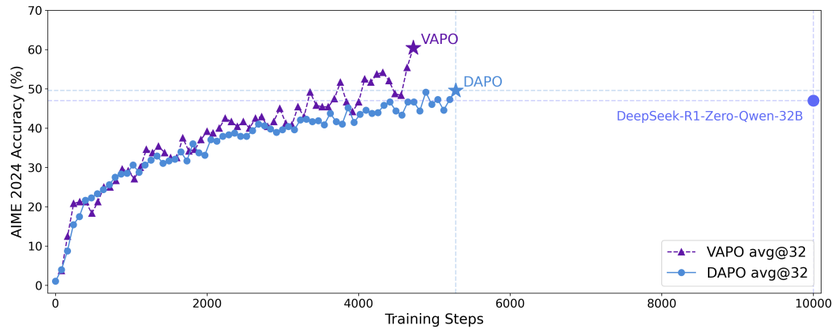


Figure 1 AIME 2024 scores of **VAPO** on the Qwen2.5-32B base model, demonstrates significant superiority over the previous state-of-the-art (SOTA) method DAPO, achieving this with notably fewer training steps. The x-axis denotes the gradient update steps.

In the Large Language Models (LLM) [2–4, 11, 15, 25, 28] RL training, value-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-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-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-free methods, we argue that value-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-free frameworks [30]. Secondly, in contrast to the advantage estimates derived from Monte Carlo methods in value-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,

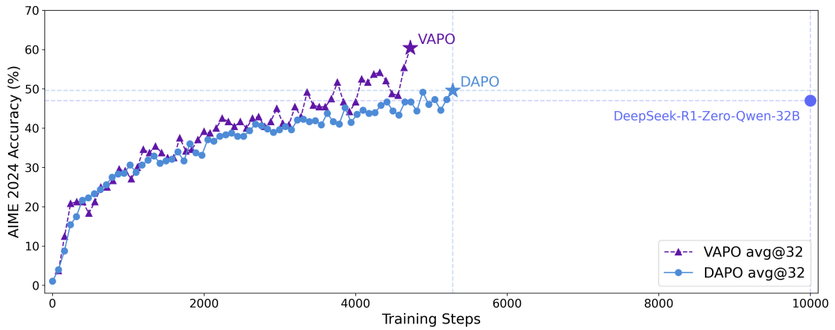


Figure 1 AIME 2024 在 Qwen2.5-32B 基础模型上的 **VAPO** 得分，显示出相对于之前的最先进（SOTA）方法 DAPO 的显著优越性，并且在明显更少的训练步骤中实现了这一点。x轴表示梯度更新步数。

计算优势。然后将轨迹级别的优势直接分配为序列中每个位置的标记级别优势。当训练一个可靠的价值模型特别具有挑战性时，无价值方法通过在组内多个轨迹之间平均奖励，提供了一个准确且稳定的基线来进行优势计算。这种基于组的奖励聚合减轻了对显式价值估计的需求，而显式价值估计在复杂任务中通常会受到不稳定性的影响。因此，无价值方法在解决诸如长 CoT 推理等困难问题方面获得了显著的关注，大量研究工作集中在优化其框架上。

尽管无价值方法取得了显著的成功，我们认为如果能够解决训练价值模型中的挑战，基于价值的方法具备更高的性能上限。首先，价值模型可以通过准确追踪每个动作对后续回报的影响，实现更精确的信用分配，从而促进更精细的优化 [21]。这对复杂的推理任务尤为重要，因为在这些任务中，单个步骤中的细微错误往往会导致灾难性的失败，而这是在无价值框架下进行模型优化所面临的挑战 [30]。其次，与无价值方法中从蒙特卡罗方法得出的优势估计相比，价值模型可以为每个标记提供更低方差的价值估计，从而增强训练的稳定性。此外，经过良好训练的价值模型具有内在的泛化能力，能够更高效地利用在线探索过程中遇到的样本。这显著提升了强化学习算法的优化上限。因此，尽管为复杂问题训练价值模型面临巨大挑战，但克服这些困难所带来的潜在收益是巨大的。

然而，在长 CoT 任务中训练一个完美的价值模型存在重大挑战。首先，鉴于轨迹较长以及以引导方式学习价值的不稳定性，学习一个低偏差的价值模型并非易事。其次，同时处理短响应和长响应也具有挑战性，因为它们在优化过程中可能对偏差-方差权衡表现出非常不同的偏好。最后但同样重要的是，来自验证者的稀疏奖励信号由于长 CoT 模式而进一步加剧，这本质上要求更好的机制来平衡探索和利用。为了解决上述挑战并充分释放基于价值的方法在推理任务中的潜力，我们提出了 **V**alue **A**ugmented proximal **P**olicy **O**ptimization (**VAPO**)，这是一种基于价值的强化学习训练框架。VAPO 受益于先前的研究工作，如 VC-PPO [30] 和 DAPO [29]，并进一步扩展了它们的概念。

我们总结我们的主要贡献如下：

1. 我们引入了VAPO，这是第一个在长链COT任务上显著优于无价值方法的价值基础强化学习训练框架。VAPO不仅在性能方面表现出显著的优越性，还展示了增强的训练效率，简化了学习过程，并突显了其作为该领域新基准的潜力。

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-based methods in reasoning tasks, we present **V**alue **A**ugmented proximal **P**olicy **O**ptimization (**VAPO**), a value-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-based RL training framework to outperform value-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.
3. We systematically integrate techniques from prior work, such as Clip-Higher and Token-level Loss from DAPO [29], Value-Pretraining and Decoupled-GAE from VC-PPO [30], self-imitation learning from SIL [14], and Group-Sampling from GRPO [22]. Additionally, we further validate their necessity through ablation studies.

VAPO is an effective reinforcement learning system that brings together these improvements. These enhancements work together smoothly, leading to a combined result that's better than the sum of the individual parts. We conduct experiments using the Qwen2.5-32B pre-trained model, ensuring no SFT data is introduced in any of the experiments, to maintain comparability with related works (DAPO and DeepSeek-R1-Zero-Qwen-32B). The performance of **VAPO** improves from vanilla PPO a score of 5 to 60, surpassing the previous SOTA value-free methods DAPO [29] by 10 points. More importantly, **VAPO** is highly stable — we don't observe any crashes during training, and the results across multiple runs are consistently similar.

2 Preliminaries

This section presents the fundamental concepts and notations that serve as the basis for our proposed algorithm. We first explore the basic framework of representing language generation as a reinforcement learning task. Subsequently, we introduce Proximal Policy Optimization and Generalized Advantage Estimation.

2. 我们提出了长度自适应GAE（Length-adaptive GAE），它根据响应长度自适应地调整GAE计算中的λ参数。通过这种方式，它有效地满足了与高度可变长度的响应相关的不同偏差-方差权衡需求。因此，它优化了优势估计过程的准确性和稳定性，特别是在数据序列长度变化较大的场景中。
3. 我们系统地整合了来自先前工作的技术，例如 DAPO [29] 中的 Clip-Higher 和 Token-level Loss，VC-PPO [30] 中的 Value-Pretraining 和 Decoupled-GAE，SIL [14] 中的自我模仿学习，以及 GRPO [22] 中的 Group-Sampling。此外，我们还通过消融研究进一步验证了它们的必要性。

VAPO 是一个有效的强化学习系统，它将这些改进结合在一起。这些增强功能协同工作，带来了比各个部分简单相加更好的综合结果。我们使用Qwen2.5-32B预训练模型进行实验，并确保在所有实验中不引入SFT数据，以保持与相关工作的可比性（如DAPO和DeepSeek-R1-Zero-Qwen-32B）。**VAPO** 的性能从原始PPO的得分5提升到60，超过了之前的价值无关方法DAPO [29] 10分。更重要的是，**VAPO** 非常稳定——我们在训练过程中没有观察到任何崩溃现象，多次运行的结果也始终一致。

2 Preliminaries

本节介绍作为我们所提出算法基础的基本概念和符号。我们首先探讨将语言生成表示为强化学习任务的基本框架。随后，我们介绍近端策略优化（Proximal Policy Optimization）和广义优势估计（Generalized Advantage Estimation）。

2.1 Modeling Language Generation as Token-Level MDP

强化学习的核心在于学习一个策略，该策略在智能体与环境交互过程中最大化累积奖励。在这项研究中，我们将语言生成任务置于马尔可夫决策过程（MDP）的框架内 [17]。

设提示为 x，对该提示的响应为 y。x 和 y 均可分解为标记序列。例如，提示 x 可表示为 $x = (x_0, \ldots, x_m)$，其中标记来自固定的离散词汇表 $\mathcal{A}$。

我们定义标记级别的 MDP 为元组 $\mathcal{M} = (\mathcal{S}, \mathcal{A}, \mathbb{P}, R, d_0, \omega)$。以下是每个组成部分的详细说明：

- **状态空间 ($\mathcal{S}$)：** 此空间包含由截至给定时间步生成的标记形成的所有可能状态。在时间步t，状态s_t定义为$s_t = (x_0, \ldots, x_m, y_0, \ldots, y_t)$。
- **动作空间 ($\mathcal{A}$)：** 它对应于固定的离散词汇表，生成过程中从中选择标记。
- **动态 ($\mathbb{P}$)：** 这些表示标记之间的确定性转移模型。给定一个状态 $s_t = (x_0, \ldots, x_m, y_0, \ldots, y_t)$，动作 $a = y_{t+1}$，以及后续状态 $s_{t+1} = (x_0, \ldots, x_m, y_0, \ldots, y_t, y_{t+1})$，概率 $\mathbb{P}(s_{t+1}|s_t, a) = 1$。
- **终止条件：** 当执行终止动作ω时，语言生成过程结束，该终止动作通常为句子结束标志。
- **奖励函数 ($R(s,a)$)：** 此函数提供标量反馈，用于评估代理在状态 s 下采取行动 a 后的表现。在来自人类反馈的强化学习 (RLHF) [18, 23] 的背景下，奖励函数可以基于人类偏好进行学习，或者由与任务相关的规则集定义。
- **初始状态分布 (d_0)：** 它是关于提示 x 的概率分布。初始状态 s_0 由提示 x 中的标记组成。

2.1 Modeling Language Generation as Token-Level MDP

Reinforcement learning centers around the learning of a policy that maximizes the cumulative reward for an agent as it interacts with an environment. In this study, we cast language generation tasks within the framework of a Markov Decision Process (MDP) [17].

Let the prompt be denoted as x, and the response to this prompt as y. Both x and y can be decomposed into sequences of tokens. For example, the prompt x can be expressed as $x = (x_0, \ldots, x_m)$, where the tokens are drawn from a fixed discrete vocabulary $\mathcal{A}$.

We define the token-level MDP as the tuple $\mathcal{M} = (\mathcal{S}, \mathcal{A}, \mathbb{P}, R, d_0, \omega)$. Here is a detailed breakdown of each component:

- **State Space ($\mathcal{S}$)**: This space encompasses all possible states formed by the tokens generated up to a given time step. At time step t, the state s_t is defined as $s_t = (x_0, \ldots, x_m, y_0, \ldots, y_t)$.
- **Action Space ($\mathcal{A}$)**: It corresponds to the fixed discrete vocabulary, from which tokens are selected during the generation process.
- **Dynamics ($\mathbb{P}$)**: These represent a deterministic transition model between tokens. Given a state $s_t = (x_0, \ldots, x_m, y_0, \ldots, y_t)$, an action $a = y_{t+1}$, and the subsequent state $s_{t+1} = (x_0, \ldots, x_m, y_0, \ldots, y_t, y_{t+1})$, the probability $\mathbb{P}(s_{t+1}|s_t, a) = 1$.
- **Termination Condition**: The language generation process concludes when the terminal action ω, typically the end-of-sentence token, is executed.
- **Reward Function ($R(s, a)$)**: This function offers scalar feedback to evaluate the agent's performance after taking action a in state s. In the context of Reinforcement Learning from Human Feedback (RLHF) [18, 23], the reward function can be learned from human preferences or defined by a set of rules specific to the task.
- **Initial State Distribution (d_0)**: It is a probability distribution over prompts x. An initial state s_0 consists of the tokens within the prompt x.

2.2 RLHF Learning Objective

We formulate the optimization problem as a KL-regularized RL task. Our objective is to approximate the optimal KL-regularized policy, which is given by:

$$\pi^* = \arg\max_{\pi} \mathbb{E}_{\pi, s_0 \sim d_0} \left[\sum_{t=0}^{H} \left(R(s_t, a_t) - \beta \mathrm{KL}\big(\pi(\cdot|s_t) \| \pi_{\mathrm{ref}}(\cdot|s_t)\big)\right)\right] \tag{1}$$

In this equation, H represents the total number of decision steps, s_0 is a prompt sampled from the dataset, $R(s_t, a_t)$ is the token-level reward obtained from the reward function, β is a coefficient that controls the strength of the KL-regularization, and π_{ref} is the initialization policy.

In traditional RLHF and most tasks related to LLMs, the reward is sparse and is only assigned at the

2.2 RLHF Learning Objective

我们将优化问题表述为一个KL正则化的强化学习任务。我们的目标是近似最优的KL正则化策略，该策略由以下公式给出：

$$\pi^* = \arg\max_{\pi} \mathbb{E}_{\pi, s_0 \sim d_0} \left[\sum_{t=0}^{H} \left(R(s_t, a_t) - \beta \mathrm{KL}\big(\pi(\cdot|s_t) \| \pi_{\mathrm{ref}}(\cdot|s_t)\big)\right)\right] \tag{1}$$

在该公式中，H 表示决策步数的总数，s_0 是从数据集中采样的提示，$R(s_t, a_t)$ 是从奖励函数获得的逐标记奖励，β 是控制 KL 正则化强度的系数，而 π_{ref} 是初始化策略。

在传统的基于人类反馈的强化学习 (RLHF) 和大多数与大语言模型 (LLM) 相关的任务中，奖励是稀疏的，并且仅在终止动作 ω 时分配，即句子结束标记 `<eos>`。

2.3 Proximal Policy Optimization

PPO [21] 使用一个裁剪的替代目标来更新策略。其核心思想是在每次更新步骤中限制策略的变化，防止可能导致不稳定性的大规模策略更新。

设 $\pi_\theta(a|s)$ 是由参数 θ 定义的策略，$\pi_{\theta_{\mathrm{old}}}(a|s)$ 是前一次迭代中的旧策略。PPO 的替代目标函数定义为：

$$\mathcal{L}^{CLIP}(\theta) = \hat{\mathbb{E}}_t \left[\min \left(r_t(\theta)\hat{A}_t, \mathrm{clip}(r_t(\theta), 1-\epsilon, 1+\epsilon)\hat{A}_t \right) \right] \tag{2}$$

其中 $r_t(\theta) = \frac{\pi_\theta(a_t|s_t)}{\pi_{\theta_{\mathrm{old}}}(a_t|s_t)}$ 是概率比，$\hat{A}_t$ 是时间步 t 的估计优势值，而 ϵ 是控制裁剪范围的超参数。

广义优势估计 [20] 是一种用于在 PPO 中更准确地估计优势函数的技术。它结合了多步引导法以减少优势估计的方差。对于长度为 T 的轨迹，时间步 t 的优势估计值 $\hat{A}_t$ 计算如下：

$$\hat{A}_t = \sum_{l=0}^{T-t-1} (\gamma\lambda)^l \delta_{t+l} \tag{3}$$

其中 γ 是折扣因子，$\lambda \in [0, 1]$ 是广义优势估计（GAE）参数，$\delta_t = R(s_t, a_t) + \gamma V(s_{t+1}) - V(s_t)$ 是时间差分（TD）误差。这里，$R(s_t, a_t)$ 是时间步 t 的奖励，$V(s)$ 是价值函数。由于在基于人类反馈的强化学习（RLHF）中通常使用折扣因子 $\gamma = 1.0$，为简化符号表示，我们在本文后续章节中省略 γ。

3 Challenges in Long-CoT RL for Reasoning Tasks

长链推理（Long-CoT）任务对强化学习（RL）训练提出了独特的挑战，特别是对于那些使用价值模型来减少方差的方法。在本节中，我们系统地分析了由序列长度动态变化、价值函数不稳定性和奖励稀疏性引发的技术问题。

terminal action ω, that is, the end-of-sentence token <eos>.

2.3 Proximal Policy Optimization

PPO [21] uses a clipped surrogate objective to update the policy. The key idea is to limit the change in the policy during each update step, preventing large policy updates that could lead to instability.

Let $\pi_\theta(a|s)$ be the policy parameterized by θ, and $\pi_{\theta_{\text{old}}}(a|s)$ be the old policy from the previous iteration. The surrogate objective function for PPO is defined as:

$$\mathcal{L}^{CLIP}(\theta) = \hat{\mathbb{E}}_t \left[\min \left(r_t(\theta)\hat{A}_t, \text{clip}(r_t(\theta), 1-\epsilon, 1+\epsilon)\hat{A}_t \right) \right] \tag{2}$$

where $r_t(\theta) = \frac{\pi_\theta(a_t|s_t)}{\pi_{\theta_{\text{old}}}(a_t|s_t)}$ is the probability ratio, $\hat{A}_t$ is the estimated advantage at time step t, and ϵ is a hyperparameter that controls the clipping range.

Generalized Advantage Estimation [20] is a technique used to estimate the advantage function more accurately in PPO. It combines multiple-step bootstrapping to reduce the variance of the advantage estimates. For a trajectory of length T, the advantage estimate $\hat{A}_t$ at time step t is computed as:

$$\hat{A}_t = \sum_{l=0}^{T-t-1} (\gamma\lambda)^l \delta_{t+l} \tag{3}$$

where γ is the discount factor, $\lambda \in [0, 1]$ is the GAE parameter, and $\delta_t = R(s_t, a_t) + \gamma V(s_{t+1}) - V(s_t)$ is the temporal-difference (TD) error. Here, $R(s_t, a_t)$ is the reward at time step t, and $V(s)$ is the value function. Since it is a common practice to use discount factor $\gamma = 1.0$ in RLHF, to simplify our notation, we omit γ in later sections of this paper.

3 Challenges in Long-CoT RL for Reasoning Tasks

Long-CoT tasks present unique challenges to RL training, especially for methods that employ a value model to reduce variance. In this section, we systematically analyze the technical issues arising from sequence length dynamics, value function instability, and reward sparsity.

3.1 Value Model Bias over Long Sequences

As identified in VC-PPO [30], initializing the value model with a reward model introduces significant initialization bias. This positive bias arises from an objective mismatch between the two models. The reward model is trained to score on the <EOS> token, incentivizing it to assign lower scores to earlier tokens due to their incomplete context. In contrast, the value model estimates the expected cumulative reward for all tokens preceding <EOS> under a given policy. During early training phases, given the backward computation of GAE, there will be a positive bias at every timestep t that accumulates along the trajectory.

3.1 Value Model Bias over Long Sequences

正如在 VC-PPO [30] 中指出的，用奖励模型初始化价值模型会引入显著的初始化偏差。这种正向偏差源于两个模型之间的目标不匹配。奖励模型被训练为对 <EOS> 令牌进行评分，这促使它由于不完整的上下文而给前面的令牌分配较低的分数。相比之下，价值模型估计的是在给定策略下所有位于 <EOS> 之前的令牌的预期累积奖励。在早期训练阶段，由于 GAE 的反向计算，每个时间步 t 都会出现一个正向偏差，并且该偏差会沿着轨迹累积。

另一种使用 GAE 的标准做法是设置 $\lambda = 0.95$，这可能会加剧这一问题。终止令牌处的奖励信号 $R(s_T, <\text{EOS}>)$ 以 $\lambda^{T-t}R(s_T, <\text{EOS}>)$ 的形式向后传播到第 t 个令牌。对于长序列（其中 $T - t \gg 1$），这种折扣会使有效的奖励信号接近于零。结果，价值更新几乎完全依赖于自举方法，依靠高度偏差的估计值，从而削弱了价值模型作为可靠方差缩减基线的作用。

3.2 Heterogeneous Sequence Lengths during Training

在复杂的推理任务中，生成正确的答案需要较长的思维链（CoT），模型往往会生成长度高度可变的响应。这种可变性要求算法足够强大，能够处理从非常短到极其长的序列。因此，通常使用的具有固定 λ 参数的 GAE 方法遇到了重大挑战。

即使价值模型是完美的，静态的 λ 也可能无法有效适应不同长度的序列。对于短响应，通过 GAE 获得的估计值往往存在高方差。这是因为 GAE 是偏差和方差之间的一种权衡。在短响应的情况下，估计值倾向于偏向方差主导的一侧。另一方面，对于长响应，GAE 通常由于自举（bootstrapping）而导致高偏差。GAE 的递归性质依赖于未来的状态值，在长序列上会累积误差，从而加剧了偏差问题。这些局限性深深植根于 GAE 计算框架的指数衰减性质之中。

3.3 Sparsity of Reward Signal in Verifier-based Tasks

复杂推理任务经常使用验证者作为奖励模型 [6, 16]。与提供密集信号的传统基于语言模型的奖励模型（例如从 -4 到 4 的连续值）不同，基于验证者的奖励模型通常提供二元反馈，例如 0 和 1。由于长链推理 (CoT)，奖励信号的稀疏性进一步加剧。由于 CoT 显著延长了输出长度，这不仅增加了计算时间，还降低了接收非零奖励的频率。在策略优化中，带有正确答案的采样响应可能极为稀少且有价值。

这种情况引发了一个独特的探索-利用困境。一方面，模型必须保持相对较高的不确定性，这使得它可以采样多样化的响应，从而增加为给定提示生成正确答案的可能性。另一方面，算法需要有效地利用通过艰苦探索获得的正确采样响应，以提高学习效率。如果无法在探索和利用之间找到适当的平衡，模型可能会因为过度利用而陷入次优解，或者浪费计算资源在无成效的探索上。

4 VAPO: Addressing the Challenges in Long-CoT RL

4.1 Mitigating Value Model Bias over Long Sequences

基于第 3.1 节中对基于价值模型的分析，我们提出使用值预训练（Value-Pretraining）和解耦广义优势估计（decoupled-GAE）来解决在长序列中价值模型偏差的关键挑战。这两种技术都借鉴了VC-PPO之前引入的方法。

Another standard practice of using GAE with $\lambda = 0.95$ might exacerbates this issue. The reward signal $R(s_T, \text{<EOS>})$ at the termination token propagates backward as $\lambda^{T-t}R(s_T, \text{<EOS>})$ to the t-th token. For long sequences where $T - t \gg 1$, this discounting reduces the effective reward signal to near zero. Consequently, value updates become almost entirely bootstrapped, relying on highly biased estimates that undermine the value model's role as a reliable variance-reduction baseline.

3.2 Heterogeneous Sequence Lengths during Training

In complex reasoning tasks where a long CoT is essential for arriving at the correct answer, models often generate responses with highly variable lengths. This variability requires algorithms to be robust enough to manage sequences that can range from very short to extremely long. As a result, the commonly-applied GAE method with a fixed λ parameter encounters significant challenges.

Even when the value model is perfect, a static λ may not effectively adapt to sequences of varying lengths. For short-length responses, the estimates obtained through GAE tend to suffer from high variance. This is because GAE represents a trade-off between bias and variance. In the case of short responses, the estimates are skewed towards the variance-dominated side. On the other hand, for long-length responses, GAE often leads to high bias due to bootstrapping. The recursive nature of GAE, which relies on future state values, accumulates errors over long sequences, exacerbating the bias issue. These limitations are deeply rooted in the exponentially-decaying nature of GAE's computational framework.

3.3 Sparsity of Reward Signal in Verifier-based Tasks

Complex reasoning tasks frequently deploy a verifier as a reward model [6, 16]. Unlike traditional language-model-based reward models that provide a dense signal, such as a continuous value ranging from -4 to 4, verifier-based reward models typically offer binary feedback, such as 0 and 1. The sparsity of the reward signal is further compounded by long CoT reasoning. As CoT significantly elongates output lengths, it not only increases computational time but also reduces the frequency of receiving non-zero rewards. In policy optimization, the sampled responses with correct answer could be extremely scarce and valuable.

This situation poses a distinct exploration-exploitation dilemma. On one hand, the model must maintain relatively high uncertainty. This enables it to sample a diverse range of responses, increasing the likelihood of generating the correct answer for a given prompt. On the other hand, algorithms need to effectively utilize the correctly sampled responses—obtained through painstaking exploration—to enhance learning efficiency. By failing to strike the right balance between exploration and exploitation, the model may either get stuck in suboptimal solutions due to excessive exploitation or waste computational resources on unproductive exploration.

4 VAPO: Addressing the Challenges in Long-CoT RL

值预训练（Value-Pretraining） 被提出以缓解价值初始化偏差。将PPO直接应用于长链推理任务时，会导致输出长度崩溃和性能下降等问题。其原因是价值模型从奖励模型初始化，而奖励模型与价值模型的目标不一致。这一现象首先在VC-PPO [30] 中被识别并解决。在本文中，我们遵循值预训练技术，具体步骤如下：

1. 持续地从固定的策略中采样生成响应，例如 π_{sft}，并使用蒙特卡洛返回值更新价值模型。
2. 训练价值模型，直到关键的训练指标（包括价值损失和解释方差 [7]）达到足够低的值。
3. 保存值检查点并加载此检查点以进行后续实验。

Decoupled-GAE 在 VC-PPO [30] 中被证明是有效的。该技术将价值和策略的优势计算分离。对于价值更新，建议使用 $\lambda = 1.0$ 来计算价值更新目标。这种选择会导致无偏的梯度下降优化，有效解决长链推理任务中的奖励衰减问题。

然而，对于策略更新，在计算和时间限制下，建议使用较小的 λ 来加速策略收敛。在 VC-PPO 中，这是通过在优势计算中使用不同的系数实现的：$\lambda_{\text{critic}} = 1.0$ 和 $\lambda_{\text{policy}} = 0.95$。在本文中，我们采用了分离 GAE 计算的核心思想。

4.2 Managing Heterogeneous Sequence Lengths during Training

为了解决训练过程中异构序列长度的挑战，我们提出了**Length-Adaptive GAE**。该方法根据序列长度动态调整GAE中的参数，从而实现对不同长度序列的自适应优势估计。此外，为了增强混合长度序列的训练稳定性，我们将传统的样本级策略梯度损失替换为标记级策略梯度损失。以下是关键技术细节的详细说明：

Length-Adaptive GAE 特别提出以解决在不同长度序列中λ_{policy}最优值不一致的问题。在VC-PPO中，λ_{policy}被设定为常数值$\lambda_{\text{policy}} = 0.95$。然而，在考虑GAE计算时，对于长度$l > 100$的较长输出序列，与奖励相对应的TD误差系数为$0.95^{100} \approx 0.006$，实际上接近于零。因此，当$\lambda_{\text{policy}} = 0.95$固定时，GAE计算可能主要由潜在偏差的引导TD误差主导。这种方法可能不适合处理极长的输出序列。

为了解决这一缺陷，我们提出了用于策略更新的**Length-Adaptive GAE**。我们的方法旨在确保TD误差在短序列和长序列之间分布更加均匀。我们设计了λ_{policy}系数的总和与输出长度l成比例：

$$\sum_{t=0}^{\infty} \lambda_{\text{policy}}^{t} \approx \frac{1}{1 - \lambda_{\text{policy}}} = \alpha l, \tag{4}$$

其中 α 是控制整体偏差-方差权衡的超参数。通过求解公式 4 中的 λ_{policy}，我们推导出一个长度自适应的公式：

$$\lambda_{\text{policy}} = 1 - \frac{1}{\alpha l} \tag{5}$$

这种在GAE计算中对λ_{policy}的长度自适应方法使得处理不同长度的序列更加有效。

Token-Level Policy Gradient Loss. 按照DAPO [29]的方法，我们还修改了策略梯度损失的计算方法，以调整在长链推理（COT）场景下的损失权重分配。具体来说，在之前的实现中，策略梯度损失的计算

4.1 Mitigating Value Model Bias over Long Sequences

Building upon the analysis of value-based models presented in section 3.1, we propose to use Value-Pretraining and decoupled-GAE to address the critical challenges in value model bias over long sequences. Both of these two techniques draw upon methodologies previously introduced in VC-PPO.

Value-Pretraining is proposed to mitigate the value initialization bias. Naively applying PPO to long-CoT tasks leads to failures such as collapsed output lengths and degraded performance. The reason is that the value model is initialized from the reward model while the reward model shares a mismatched objective with the value model. This phenomenon is first identified and addressed in VC-PPO [30]. In this paper, we follow the Value-Pretraining technique and the specific steps are outlined as follows:

1. Continuously generate responses by sampling from a fixed policy, for instance, π_{sft}, and update the value model with Monte-Carlo return.
2. Train the value model until key training metrics, including value loss and explained variance [7], attain sufficiently low values.
3. Save the value checkpoint and load this checkpoint for subsequent experiments.

Decoupled-GAE is proven effective in VC-PPO [30]. This technique decouples the advantage computation for the value and the policy. For value updates, it is recommended to compute the value-update target with $\lambda = 1.0$. This choice results in an unbiased gradient-descent optimization, effectively addressing the reward-decay issues in long CoT tasks.

However, for policy updates, using a smaller λ is advisable to accelerate policy convergence under computational and time constraints. In VC-PPO, this is achieved by employing different coefficients in advantage computation: $\lambda_{\text{critic}} = 1.0$ and $\lambda_{\text{policy}} = 0.95$. In this paper, we adopt the core idea of decoupling GAE computation.

4.2 Managing Heterogeneous Sequence Lengths during Training

To address the challenge of heterogeneous sequence lengths during training, we propose the **Length-Adaptive GAE**. This method dynamically adjusts the parameter in GAE according to the sequence length, enabling adaptive advantage estimation for sequences of varying lengths. Additionally, to enhance the training stability of mixed-length sequences, we replace the conventional sample-level policy gradient loss with a token-level policy gradient loss. The key technical details are elaborated as follows:

Length-Adaptive GAE is specifically proposed to to address the inconsistency in optimal λ_{policy} values across sequences of varying lengths. In VC-PPO, λ_{policy} is set to a constant value of $\lambda_{\text{policy}} = 0.95$. However, when considering the GAE computation, for longer output sequences with lengths $l > 100$, the coefficient of the TD-error corresponding to the reward is $0.95^{100} \approx 0.006$, which is effectively zero. As a result, with a fixed $\lambda_{\text{policy}} = 0.95$, the GAE computation becomes dominated by potentially biased bootstrapping TD-errors. This approach may not be optimal for handling extremely long output sequences.

方式如下：

$$\mathcal{L}_{\text{PPO}}(\theta) = -\frac{1}{G}\sum_{i=1}^{G}\frac{1}{|o_i|}\sum_{t=1}^{|o_i|}\min\left(r_{i,t}(\theta)\hat{A}_{i,t}, \text{clip}\left(r_{i,t}(\theta), 1-\varepsilon, 1+\varepsilon\right)\hat{A}_{i,t}\right), \tag{6}$$

其中 G 是训练批次的大小，o_i 是第 i 个样本的轨迹。在这种损失函数的定义中，所有标记（tokens）的损失首先在序列级别上被平均，然后再在批次级别上进一步平均。这种方法导致来自更长序列的标记对最终损失值的贡献较小。因此，如果模型在处理长序列时遇到关键问题，这种情况在强化学习（RL）训练的探索阶段容易发生，由于其权重减少而导致的抑制不足可能会引起训练不稳定甚至崩溃。为了解决标记级别对最终损失贡献的这种不平衡问题，我们将损失函数修改为以下形式：

$$\mathcal{L}_{\text{PPO}}(\theta) = -\frac{1}{\sum_{i=1}^{G}|o_i|}\sum_{i=1}^{G}\sum_{t=1}^{|o_i|}\min\left(r_{i,t}(\theta)\hat{A}_{i,t}, \text{clip}\left(r_{i,t}(\theta), 1-\varepsilon, 1+\varepsilon\right)\hat{A}_{i,t}\right), \tag{7}$$

其中，单个训练批次中的所有标记都被赋予均匀的权重，从而能够以更高的效率解决长序列带来的问题。

4.3 Dealing with Sparsity of Reward Signal in Verifier-based Tasks

如第 3.3 节中分析的，在高度稀疏奖励信号的情况下，提高强化学习训练中探索-利用权衡的效率变得极具挑战性。为了解决这一关键问题，我们采用了三种方法：Clip-Higher、Positive Example LM Loss 和 Group-Sampling。技术细节如下所述：

Clip-Higher 用于缓解在 PPO 和 GRPO 训练过程中遇到的熵崩溃问题，该方法最初在 DAPO [29] 中被提出。我们将较低和较高的裁剪范围解耦为 ε_{low} 和 $\varepsilon_{\text{high}}$。

$$\mathcal{L}_{\text{PPO}}(\theta) = -\frac{1}{\sum_{i=1}^{G}|o_i|}\sum_{i=1}^{G}\sum_{t=1}^{|o_i|}\min\left(r_{i,t}(\theta)\hat{A}_{i,t}, \text{clip}\left(r_{i,t}(\theta), 1-\varepsilon_{\text{low}}, 1+\varepsilon_{\text{high}}\right)\hat{A}_{i,t}\right), \tag{8}$$

我们增大了$\varepsilon_{\text{high}}$的值，以留出更多空间来增加低概率词元的增长。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抑制为0，从而导致采样空间的坍塌。

正例语言模型损失旨在提高强化学习（RL）训练过程中正样本的利用效率。在复杂推理任务的RL场景中，某些任务表现出极低的准确率，大多数训练样本产生错误答案。传统的抑制错误样本生成概率的策略在RL训练中效率低下，因为试错机制带来了巨大的计算成本。鉴于这一挑战，最大化利用正确答案变得至关重要，尤其是在它们被策略模型采样到时。为应对这一挑战，我们采用了一种模仿学习方法，在RL训练期间对采样的正确结果引入额外的负对数似然（NLL）损失。相应的公式如下：

$$\mathcal{L}_{\text{NLL}}(\theta) = -\frac{1}{\sum_{o_i\in\mathcal{T}}|o_i|}\sum_{o_i\in\mathcal{T}}\sum_{t=1}^{|o_i|}\log\pi_\theta\left(a_t|s_t\right), \tag{9}$$

其中 $\mathcal{T}$ 表示正确答案的集合。最终的 NLL 损失通过加权系数 μ 与策略梯度损失结合，共同作为更新

To address this shortcoming, we propose **Length-Adaptive GAE** for policy updates. Our method aims to ensure a more uniform distribution of TD-errors across both short and long sequences. We design the sum of the coefficients λ_{policy} to be proportional to the output length l:

$$\sum_{t=0}^{\infty} \lambda_{\text{policy}}^{t} \approx \frac{1}{1-\lambda_{\text{policy}}} = \alpha l, \tag{4}$$

where α is a hyper-parameter controlling the overall bias-variance trade-off. By solving Equation 4 for λ_{policy}, we derive a length-adaptive formula:

$$\lambda_{\text{policy}} = 1 - \frac{1}{\alpha l} \tag{5}$$

This length-adaptive approach to λ_{policy} in GAE calculation allows for a more effective handling of sequences of varying lengths.

Token-Level Policy Gradient Loss. Following DAPO [29], we have also modified the computation method of the policy gradient loss to adjust the loss weight allocation in long COT scenarios. Specifically, in previous implementations, the policy gradient loss was computed as follows:

$$\mathcal{L}_{\text{PPO}}(\theta) = -\frac{1}{G}\sum_{i=1}^{G}\frac{1}{|o_i|}\sum_{t=1}^{|o_i|}\min\left(r_{i,t}(\theta)\hat{A}_{i,t}, \text{clip}\left(r_{i,t}(\theta), 1-\varepsilon, 1+\varepsilon\right)\hat{A}_{i,t}\right), \tag{6}$$

where G is the size of training batch, o_i is the trajectory of the ith sample. In this loss formulation, the losses of all tokens are first averaged at the sequence level before being further averaged at the batch level. This approach results in tokens from longer sequences contributing less to the final loss value. Consequently, if the model encounters critical issues in processing long sequences, a scenario that is prone to occur during the exploration phase of RL training, the insufficient suppression caused by their diminished weighting may lead to training instability or even collapse. To address this imbalance in token-level contribution to the final loss, we revise the loss function into the following form:

$$\mathcal{L}_{\text{PPO}}(\theta) = -\frac{1}{\sum_{i=1}^{G}|o_i|}\sum_{i=1}^{G}\sum_{t=1}^{|o_i|}\min\left(r_{i,t}(\theta)\hat{A}_{i,t}, \text{clip}\left(r_{i,t}(\theta), 1-\varepsilon, 1+\varepsilon\right)\hat{A}_{i,t}\right), \tag{7}$$

where all tokens within a single training batch are assigned uniform weights, thereby enabling the problems posed by long sequences to be addressed with enhanced efficiency.

4.3 Dealing with Sparsity of Reward Signal in Verifier-based Tasks

As analyzed in Section 3.3, enhancing the efficiency of exploration-exploitation tradeoff in RL training becomes critically challenging under scenarios with highly sparse reward signals. To address this key issue, we adopt three methods: Clip-Higher, Positive Example LM Loss and Group-Sampling. The technical details are elaborated as follows:

Clip-Higher is used to mitigate the entropy collapse issue encountered in PPO and GRPO training

策略模型的目标：

$$\mathcal{L}(\theta) = \mathcal{L}_{\text{PPO}}(\theta) + \mu * \mathcal{L}_{\text{NLL}}(\theta). \tag{10}$$

Group-Sampling 用于在相同提示符内采样判别性的正样本和负样本。在给定固定的计算预算下，存在两种主要的计算资源分配方法。第一种方法尽可能多地使用提示符，并且每个提示符仅采样一次。第二种方法减少每批不同的提示符数量，将计算资源重新导向多次生成。我们观察到后一种方法能够带来略微更好的性能，这归因于它引入了更丰富的对比信号，从而增强了策略模型的学习能力。

5 Experiments

5.1 Training Details

在本工作中，我们基于Qwen-32B模型对PPO算法进行了多种改进，从而提升了模型的数学性能。这些技术同样适用于其他推理任务，例如与代码相关的任务。对于基础的PPO算法，我们使用AdamW作为优化器，将actor的学习率设置为1×10^{-6}，critic的学习率设置为2×10^{-6}，因为critic需要更快地更新以跟上策略的变化。学习率采用了warmup-constant调度器。批量大小为8192个提示（prompt），每个提示采样一次，每个小批量（mini-batch）大小设置为512。价值网络通过奖励模型进行初始化，GAE λ 设置为0.95，γ 设置为1.0。使用样本级损失，并将clip ϵ 设置为0.2。

与原始的PPO相比，VAPO做了以下参数调整：

1. 在启动策略训练之前，根据奖励模型（RM）实现了价值网络的50步热身。
2. 使用了解耦的GAE，其中价值网络从使用λ=1.0估计的回报中学习，而策略网络从使用单独lambda获得的优势中学习。
3. 根据序列长度自适应地设置优势估计的lambda，公式为：$\lambda_{\text{policy}} = 1 - \frac{1}{\alpha l}$，其中$\alpha = 0.05$。
4. 将裁剪范围调整为ϵ_{high}=0.28和ϵ_{low}=0.2。
5. 使用了基于标记级别的策略梯度损失。
6. 在策略梯度损失中添加了正例语言模型（LM）损失，权重为0.1。
7. 在采样过程中使用了512个提示，每个提示采样16次，并将小批量大小设置为512。

我们还将展示从VAPO中单独移除这七项修改中的每一项后的最终效果。对于评估指标，我们使用AIME24在32次运行中的平均通过率，采样参数设置为topp=0.7和temperature=1.0。

5.2 Ablation Results

在Qwen-32b上，DeepSeek R1 使用 GRPO 在 AIME24 上获得了 47 分，而 DAPO 在更新步骤的 50% 处达到了 50 分。如图 1 所示，我们提出的 VAPO 仅使用 DAPO 步骤的 60%，就匹配了这一表现，并且在仅仅 5,000 步内达到了新的 SOTA 得分 60.4，展示了 VAPO 的高效性。此外，VAPO 保持了稳定的熵——既不会坍缩也不会变得过高——并在三次重复实验中始终达到 60-61 的峰值分数，突显了我们算法的可靠性。

process, which is first proposed in DAPO [29]. We decouple the lower and higher clipping range as ε_{low} and $\varepsilon_{\text{high}}$

$$\mathcal{L}_{\text{PPO}}(\theta) = -\frac{1}{\sum_{i=1}^{G}|o_i|}\sum_{i=1}^{G}\sum_{t=1}^{|o_i|}\min\left(r_{i,t}(\theta)\hat{A}_{i,t}, \text{clip}\left(r_{i,t}(\theta), 1-\varepsilon_{\text{low}}, 1+\varepsilon_{\text{high}}\right)\hat{A}_{i,t}\right), \tag{8}$$

We increase the value of $\varepsilon_{\text{high}}$ to leave more room for the increase of low-probability tokens. 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.

Positive Example LM Loss is designed to enhance the utilization efficiency of positive samples during RL training process. In the context of RL for complex reasoning tasks, some tasks demonstrate remarkably low accuracy, with the majority of training samples yielding incorrect answers. Traditional policy optimization strategies that suppress the generation probability of erroneous samples suffer from inefficiency during RL training, as the trial-and-error mechanism incurs substantial computational costs. Given this challenge, it is critical to maximize the utility of correct answers when they are sampled by the policy model. To address this challenge, we adopt an imitation learning approach by incorporating an additional negative log-likelihood (NLL) loss for the correct outcomes sampled during RL training. The corresponding formula is as follows:

$$\mathcal{L}_{\text{NLL}}(\theta) = -\frac{1}{\sum_{o_i\in\mathcal{T}}|o_i|}\sum_{o_i\in\mathcal{T}}\sum_{t=1}^{|o_i|}\log \pi_\theta\left(a_t|s_t\right), \tag{9}$$

where $\mathcal{T}$ denotes the set of correct answers. The final NLL loss is combined with the policy gradient loss through a weighting coefficient μ, which collectively serves as the objective for updating the policy model:

$$\mathcal{L}(\theta) = \mathcal{L}_{\text{PPO}}(\theta) + \mu * \mathcal{L}_{\text{NLL}}(\theta). \tag{10}$$

Group-Sampling is used to sample discriminative positive and negative samples within the same prompt. Given a fixed computational budget, there exist two primary approaches to allocating computational resources. The first approach utilizes as many prompts as possible, with each prompt sampled only once. The second approach reduces the number of distinct prompts per batch and redirects computational resources toward repeated generations. We observed that the latter approach yields marginally better performance, attributed to the richer contrastive signals it introduces, which enhance the policy model's learning capability.

5 Experiments

5.1 Training Details

In this work we enhanced the model's mathematical performance by introducing various modifications to the PPO algorithm based on the Qwen-32B model. These techniques are also effective for other

Table 1 Abalation results of **VAPO**

Model	AIME24$_{\text{avg@32}}$
Vanilla PPO	5
DeepSeek-R1-Zero-Qwen-32B	47
DAPO	50
VAPO w/o Value-Pretraining	11
VAPO w/o Decoupled-GAE	33
VAPO w/o Length-adaptive GAE	45
VAPO w/o Clip-Higher	46
VAPO w/o Token-level Loss	53
VAPO w/o Positive Example LM Loss	54
VAPO w/o Group-Sampling	55
VAPO	**60**

Table 1 系统地展示了我们的实验结果。Vanilla PPO 方法由于价值模型学习的坍缩，在训练后期仅能达到 5 分，其特征是响应长度急剧减少，模型直接回答问题而不进行推理。我们的 VAPO 方法最终达到了 60 分，这是一个显著的提升。我们进一步通过单独消融这七项改进来验证它们的有效性：

1. 如果没有值预训练，模型在训练过程中会经历与 Vanilla PPO 相同的崩溃，收敛到大约 11 分的最大值。
2. 移除解耦的GAE会导致在反向传播过程中奖励信号呈指数级衰减，阻止模型充分优化长篇响应，从而导致下降27分。
3. 自适应GAE平衡了对短响应和长响应的优化，带来了15个点的提升。
4. 更高的剪辑鼓励充分的探索和利用；其移除将模型的最大收敛限制在46分。
5. Token-level loss 隐式地增加了长回答的权重，贡献了 7 个点的提升。
6. 加入正例语言模型损失提升了模型近6个点。
7. 使用组采样来生成更少的提示，但增加重复次数，也带来了5个点的提升。

5.3 Training Dynamics

在强化学习（RL）训练过程中生成的曲线可以提供关于训练稳定性的实时洞察，不同曲线之间的比较可以突出算法之间的差异。人们普遍认为，这些曲线的平滑变化和快速增长是其理想的特征。通过对比VAPO和DAPO的训练过程，我们得出了以下观察结果：

- Figure 2 显示 VAPO 的训练曲线比 DAPO 的更平滑，这表明 VAPO 的算法优化更稳定。
- 如Figure 2a所示，VAPO在长度缩放方面表现出优于DAPO的性能。在现代背景下，更好的长度缩放被广泛认为是模型性能提升的一个标志，因为它增强了模型的泛化能力。
- Figure 2b 表明 VAPO 的分数增长速度比 DAPO 快，因为价值模型为模型提供了更精细的信号以加速优化。

reasoning tasks, such as code-related tasks. For the basic PPO, we used AdamW as the optimizer, setting the actor learning rate to 1×10^{-6} and the critic learning rate to 2×10^{-6}, as the critic needs to update faster to keep pace with policy changes. The learning rate employed a warmup-constant scheduler. The batch size was 8192 prompts, with each prompt sampled once, and each mini-batch size set to 512. The value network was initialized using a reward model, with the GAE λ set to 0.95 and γ set to 1.0. Sample-level loss was used, and the clip ϵ was set to 0.2.

Compared to vanilla PPO, VAPO made the following parameter adjustments:

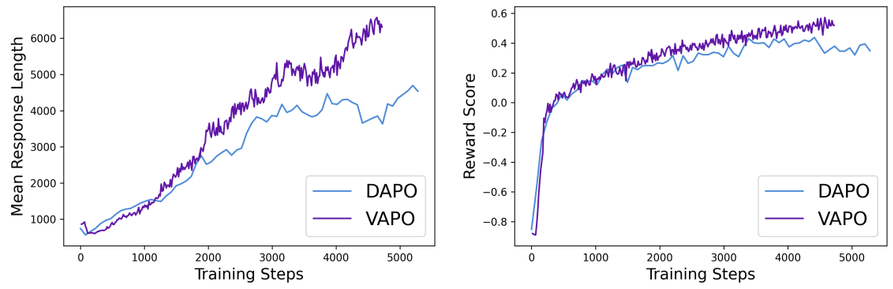
1. Implemented a value network warmup for 50 steps based on the reward model (RM) before initiating policy training.
2. Utilized decoupled GAE, where the value network learns from returns estimated with λ=1.0, while the policy network learns from advantages obtained using a separate lambda.
3. Adaptively set the lambda for advantage estimation based on sequence length, following the formula: $\lambda_{\text{policy}} = 1 - \frac{1}{\alpha l}$, where $\alpha = 0.05$.
4. Adjusted the clip range to ϵ_{high}=0.28 and ϵ_{low}=0.2.
5. Employed token-level policy gradient loss.
6. Added a positive-example language model (LM) loss to the policy gradient loss, with a weight of 0.1.
7. Used 512 prompts per sampling, with each prompt sampled 16 times, and set the mini-batch size to 512.

We will also demonstrate the final effects of removing each of these seven modifications from VAPO individually. For the evaluation metric, we use the average pass rate of AIME24 over 32 times, with sampling parameters set to topp=0.7 and temperature=1.0.

5.2 Ablation Results

On Qwen-32b, DeepSeek R1 using GRPO achieves 47 points on AIME24, while DAPO reaches 50 points with 50% of the update steps. In Figure 1, our proposed VAPO matches this performance using only 60% of DAPO's steps and achieves a new SOTA score of 60.4 within just 5,000 steps, demonstrating VAPO's efficiency. Additionally, VAPO maintains stable entropy—neither collapsing nor becoming excessively high—and consistently achieves peak scores of 60-61 across three repeated experiments, highlighting the reliability of our algorithm.

Table 1 systematically presents our experimental results. The Vanilla PPO method, hindered by value model learning collapse, only achieves 5 points in the later stages of training, characterized by a drastic reduction in response length and the model directly answering questions without reasoning. Our VAPO method finally achieves 60 points, which is a significant improvement. We further validated the effectiveness of the seven proposed modifications by ablating them individually:



(a) Mean response length. **(b)** Reward score.

(c) Generation entropy.

Figure 2 VAPO关于响应长度、奖励分数和生成熵的指标曲线。

- 根据Figure 2c，VAPO的熵在训练后期降到比DAPO更低。这是事情的两面：一方面，它可能会阻碍探索；但另一方面，它提高了模型的稳定性。从VAPO的最终结果来看，较低的熵对性能的负面影响很小，而其可重复性和稳定性则证明了其高度的优势。

6 Related Work

OpenAI O1 [16] 引入了大型语言模型（LLMs）中的一个深刻范式转变，其特点是在给出最终响应之前进行扩展推理 [5, 19, 28]。DeepSeek R1 [6] 开源了其训练算法（无价值的 GRPO [22]）和模型权重，其性能与 O1 相当。DAPO [29] 发现了在无价值 LLM 强化学习扩展过程中遇到的以前未披露的挑战，例如熵崩溃，并提出了四种有效技术来克服这些挑战，实现了行业级别的最先进（SOTA）性能。最近，Dr. GRPO [12] 去除了 GRPO 中的长度和标准差归一化项。另一方面，ORZ [9] 遵循 PPO 方法并使用价值模型进行优势估计，提出了蒙特卡洛估计而不是广义优势估计。然而，他们的方法仅能达到与无价值方法（如 GRPO 和 DAPO）相当的性能。在本文中，我们同样遵循价值模型方法并提出了 VAPO，其性能优于最先进的无价值算法 DAPO。

7 Conclusion

在本文中，我们提出了一种名为VAPO的算法，该算法利用Qwen2.5-32B模型，在AIME24基准测试中达到了最先进（SOTA）的性能。通过在PPO的基础上引入七种新颖的技术，这些技术专注于优化价值

Table 1 Abalation results of **VAPO**

Model	AIME24$_{\text{avg@32}}$
Vanilla PPO	5
DeepSeek-R1-Zero-Qwen-32B	47
DAPO	50
VAPO w/o Value-Pretraining	11
VAPO w/o Decoupled-GAE	33
VAPO w/o Length-adaptive GAE	45
VAPO w/o Clip-Higher	46
VAPO w/o Token-level Loss	53
VAPO w/o Positive Example LM Loss	54
VAPO w/o Group-Sampling	55
VAPO	**60**

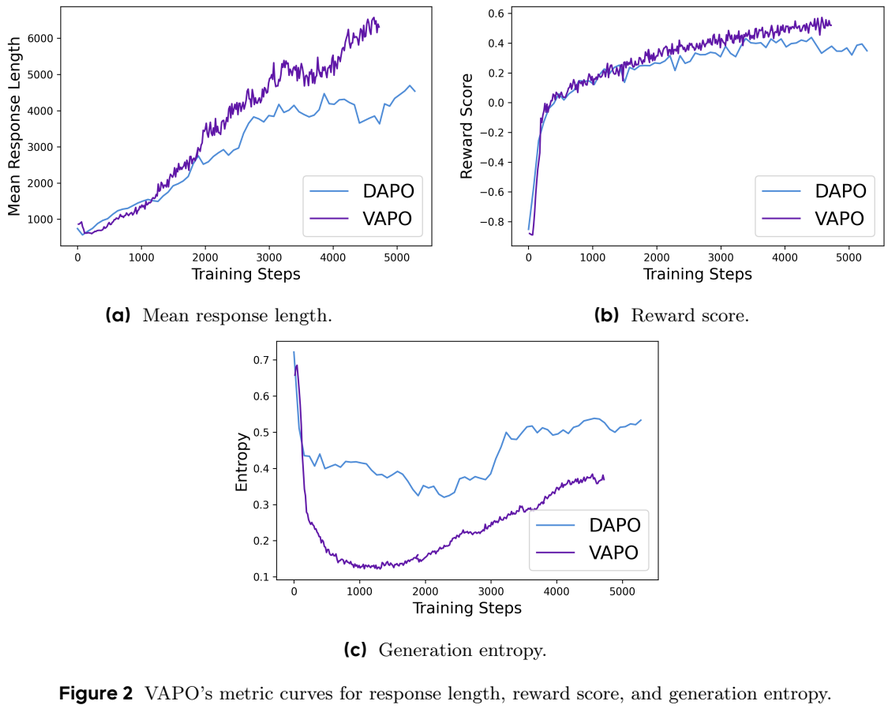
1. Without Value-Pretraining, the model experiences the same collapse as Vanilla PPO during training, converging to a maximum of approximately 11 points.
2. Removing the decoupled GAE causes reward signals to exponentially decay during backpropagation, preventing the model from fully optimizing long-form responses and leading to a 27-point drop.
3. Adaptive GAE balances optimization for both short and long responses, yielding a 15-point improvement.
4. Clip higher encourages thorough exploration and exploitation; its removal limited the model's maximum convergence to 46 points.
5. Token-level loss implicitly increased the weight of long responses, contributing to a 7-point gain.
6. Incorporating positive-example LM loss boosted the model by nearly 6 points.
7. Using Group-Sampling to generate fewer prompts but with more repetitions also resulted in a 5-point improvement.

5.3 Training Dynamics

The curves generated during RL training provide real-time insights into training stability, and comparisons between different curves can highlight algorithmic differences. It is generally believed that smoother changes and faster growth are the desirable characteristics of these curves. Through a comparison of the training processes of VAPO and DAPO, we made the following observations:

- Figure 2 shows that VAPO's training curve is smoother than DAPO's, indicating more stable algorithmic optimization in VAPO.
- As depicted in Figure 2a, VAPO exhibits superior length scaling compared to DAPO. In modern contexts, better length scaling is widely recognized as a marker of improved model performance, as it enhances the model's generalization capabilities.

学习和平衡探索，我们的基于价值的方法超越了如GRPO和DAPO等当代无价值的方法。本研究为推动大型语言模型在推理密集型任务中的发展提供了一个坚实的框架。



(a) Mean response length.

(b) Reward score.

(c) Generation entropy.

Figure 2 VAPO's metric curves for response length, reward score, and generation entropy.

- Figure 2b demonstrates that VAPO's score grows faster than DAPO's, as the value model provides the model with more granular signals to accelerate optimization.
- According to Figure 2c, VAPO's entropy drops lower than DAPO's in the later stages of training. This is two sides of the coin: on one hand, it may hinder exploration, but on the other hand, it improves the model stability. From VAPO's final results, the lower entropy has minimal negative impact on performance, while the reproducibility and stability proves highly advantageous.

6 Related Work

OpenAI O1 [16] introduces a profound paradigm shift in LLMs, characterized by extended reasoning before delivering a final response [5, 19, 28]. DeepSeek R1 [6] open-sources both its training algorithm (the value-free GRPO [22]) and its model weights, which are comparable in performance to O1. DAPO [29] identifies previously undisclosed challenges such as entropy collapse encountered during the scaling of value-free LLM RL, and proposes four effective techniques to overcome these challenges, achieving SOTA industry-level performance. Recently, Dr. GRPO [12] removes both the length and std normalization terms in GRPO. On the other hand, ORZ [9] follows PPO and utilizes a value model for advantage estimation, proposing Monte Carlo estimation instead of Generalized Advantage Estimation. However, they could just achieves a comparable performance to value-free method like GRPO and DAPO. In

Contributions

项目负责人

余越[1]

算法

余越[1], 袁玉峰[1], 余启颖[1,2], 左晓辰[1], 朱若飞[1], 徐文远[1], 陈家泽[1], 王成一[1], 范天天[1], 杜正银[1], 魏向鹏[1], 余翔宇[1]

基础设施*

刘高鸿[1], 刘俊才[1], 刘灵军[1], 林海斌[1], 林智奇[1], 马博磊[1], 张弛[1], 张墨凡[1], 张旺[1], 朱航[1], 张茹[1]

*按姓氏字母顺序排列

指导

刘欣[1], 王明轩[1], 吴永辉[1], 严琳[1]

所属机构

[1] 字节跳动百炼实验室

[2] 清华大学AIR-SIA实验室及字节跳动百炼实验室

this paper, we also follow the value-model approach and propose VAPO, which outperforms the SOTA value-free algorithm DAPO.

7 Conclusion

In this paper, we propose an algorithm named VAPO, which leveraging the Qwen2.5-32B model, achieves the SOTA performance on the AIME24 benchmark. By introducing seven novel techniques atop PPO, which focus on refining value learning and balancing exploration, our value-based approach outperforms contemporary value-free methods like GRPO and DAPO. The work provides a robust framework for advancing large language models in reasoning-intensive tasks.

References

[1] Arash Ahmadian, Chris Cremer, Matthias Gallé, Marzieh Fadaee, Julia Kreutzer, Olivier Pietquin, Ahmet Üstün, and Sara Hooker. Back to basics: Revisiting reinforce style optimization for learning from human feedback in llms, 2024. URL https://arxiv.org/abs/2402.14740.

[2] Anthropic. Claude 3.5 sonnet, 2024. URL https://www.anthropic.com/news/claude-3-5-sonnet.

[3] 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.

[4] 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.

[5] Google DeepMind. Gemini 2.0 flash thinking, 2024. URL https://deepmind.google/technologies/gemini/flash-thinking/.

[6] DeepSeek-AI. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning, 2025. URL https://arxiv.org/abs/2501.12948.

[7] Ron Good and Harold J. Fletcher. Reporting explained variance. Journal of Research in Science Teaching, 18(1): 1–7, 1981. doi: https://doi.org/10.1002/tea.3660180102. URL https://onlinelibrary.wiley.com/doi/abs/10.1002/tea.3660180102.

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

[9] Jingcheng Hu, Yinmin Zhang, Qi Han, Daxin Jiang, Xiangyu Zhang, and Heung-Yeung Shum. Open-reasoner-zero: An open source approach to scaling up reinforcement learning on the base model, 2025. URL https://arxiv.org/abs/2503.24290.

[10] Wouter Kool, Herke van Hoof, and Max Welling. Buy 4 REINFORCE samples, get a baseline for free! In Deep Reinforcement Learning Meets Structured Prediction, ICLR 2019 Workshop, New Orleans, Louisiana, United States, May 6, 2019. OpenReview.net, 2019. URL https://openreview.net/forum?id=r1lgTGL5DE.

[11] 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.

[12] Zichen Liu, Changyu Chen, Wenjun Li, Penghui Qi, Tianyu Pang, Chao Du, Wee Sun Lee, and Min Lin. Understanding r1-zero-like training: A critical perspective, 2025. URL https://arxiv.org/abs/2503.20783.

[13] Zhiyu Mei, Wei Fu, Kaiwei Li, Guangju Wang, Huanchen Zhang, and Yi Wu. Real: Efficient rlhf training of large language models with parameter reallocation. In Proceedings of the Eighth Conference on Machine Learning and Systems, MLSys 2025, Santa Clara, CA, USA, May 12-15, 2025. mlsys.org, 2025.

[14] Junhyuk Oh, Yijie Guo, Satinder Singh, and Honglak Lee. Self-imitation learning. In Jennifer Dy and Andreas Krause, editors, Proceedings of the 35th International Conference on Machine Learning, volume 80 of Proceedings of Machine Learning Research, pages 3878–3887. PMLR, 10–15 Jul 2018. URL https://proceedings.mlr.press/v80/oh18b.html.

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

Contributions

Project Lead

Yu Yue[1]

Algorithm

Yu Yue[1], Yufeng Yuan[1], Qiying Yu[1,2], Xiaochen Zuo[1], Ruofei Zhu[1], Wenyuan Xu[1], Jiaze Chen[1], Chengyi Wang[1], TianTian Fan[1], Zhengyin Du[1], Xiangpeng Wei[1], Xiangyu Yu[1]

Infrastructure*

Gaohong Liu[1], Juncai Liu[1], Lingjun Liu[1], Haibin Lin[1], Zhiqi Lin[1], Bole Ma[1], Chi Zhang[1], Mofan Zhang[1], Wang Zhang[1], Hang Zhu[1], Ru Zhang[1]

*Last-Name in Alphabetical Order

Supervision

Xin Liu[1], Mingxuan Wang[1], Yonghui Wu[1], Lin Yan[1]

Affiliation

[1] ByteDance Seed

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

[16] OpenAI. Learning to reason with llms, 2024. URL https://openai.com/index/learning-to-reason-with-llms/.

[17] Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. Training language models to follow instructions with human feedback. Advances in neural information processing systems, 35:27730–27744, 2022.

[18] Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. Training language models to follow instructions with human feedback. Advances in neural information processing systems, 35:27730–27744, 2022.

[19] Qwen. Qwq-32b: Embracing the power of reinforcement learning, 2024. URL https://qwenlm.github.io/blog/qwq-32b/.

[20] John Schulman, Philipp Moritz, Sergey Levine, Michael Jordan, and Pieter Abbeel. High-dimensional continuous control using generalized advantage estimation. arXiv preprint arXiv:1506.02438, 2015.

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

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

[23] Wei Shen, Guanlin Liu, Zheng Wu, Ruofei Zhu, Qingping Yang, Chao Xin, Yu Yue, and Lin Yan. Exploring data scaling trends and effects in reinforcement learning from human feedback. arXiv preprint arXiv:2503.22230, 2025.

[24] Richard S Sutton, Andrew G Barto, et al. Reinforcement learning: An introduction, volume 1. MIT press Cambridge, 1998.

[25] Gemini Team, Rohan Anil, Sebastian Borgeaud, Yonghui Wu, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M Dai, Anja Hauth, et al. Gemini: a family of highly capable multimodal models. arXiv preprint arXiv:2312.11805, 2023.

[26] 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.

[27] Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed H. Chi, Quoc V. Le, and Denny Zhou. Chain-of-thought prompting elicits reasoning in large language models. In Sanmi Koyejo, S. Mohamed, A. Agarwal, Danielle Belgrave, K. Cho, and A. Oh, editors, Advances in Neural Information Processing Systems 35: Annual Conference on Neural Information Processing Systems 2022, NeurIPS 2022, New Orleans, LA, USA, November 28 - December 9, 2022, 2022.

[28] XAI. Grok 3 beta — the age of reasoning agents, 2024. URL https://x.ai/news/grok-3.

[29] Qiying Yu, Zheng Zhang, Ruofei Zhu, Yufeng Yuan, Xiaochen Zuo, Yu Yue, Tiantian Fan, Gaohong Liu, Lingjun Liu, Xin Liu, Haibin Lin, Zhiqi Lin, Bole Ma, Guangming Sheng, Yuxuan Tong, Chi Zhang, Mofan Zhang, Wang Zhang, Hang Zhu, Jinhua Zhu, Jiaze Chen, Jiangjie Chen, Chengyi Wang, Hongli Yu, Weinan Dai, Yuxuan Song, Xiangpeng Wei, Hao Zhou, Jingjing Liu, Wei-Ying Ma, Ya-Qin Zhang, Lin Yan, Mu Qiao, Yonghui Wu, and Mingxuan Wang. Dapo: An open-source llm reinforcement learning system at scale, 2025. URL https://arxiv.org/abs/2503.14476.

References

[1] Arash Ahmadian, Chris Cremer, Matthias Gallé, Marzieh Fadaee, Julia Kreutzer, Olivier Pietquin, Ahmet Üstün, and Sara Hooker. Back to basics: Revisiting reinforce style optimization for learning from human feedback in llms, 2024. URL https://arxiv.org/abs/2402.14740.

[2] Anthropic. Claude 3.5 sonnet, 2024. URL https://www.anthropic.com/news/claude-3-5-sonnet.

[3] 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.

[4] 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.

[5] Google DeepMind. Gemini 2.0 flash thinking, 2024. URL https://deepmind.google/technologies/gemini/flash-thinking/.

[6] DeepSeek-AI. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning, 2025. URL https://arxiv.org/abs/2501.12948.

[7] Ron Good and Harold J. Fletcher. Reporting explained variance. Journal of Research in Science Teaching, 18(1): 1–7, 1981. doi: https://doi.org/10.1002/tea.3660180102. URL https://onlinelibrary.wiley.com/doi/abs/10.1002/tea.3660180102.

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

[9] Jingcheng Hu, Yinmin Zhang, Qi Han, Daxin Jiang, Xiangyu Zhang, and Heung-Yeung Shum. Open-reasoner-zero: An open source approach to scaling up reinforcement learning on the base model, 2025. URL https://arxiv.org/abs/2503.24290.

[10] Wouter Kool, Herke van Hoof, and Max Welling. Buy 4 REINFORCE samples, get a baseline for free! In Deep Reinforcement Learning Meets Structured Prediction, ICLR 2019 Workshop, New Orleans, Louisiana, United States, May 6, 2019. OpenReview.net, 2019. URL https://openreview.net/forum?id=r1lgTGL5DE.

[11] 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.

[12] Zichen Liu, Changyu Chen, Wenjun Li, Penghui Qi, Tianyu Pang, Chao Du, Wee Sun Lee, and Min Lin. Understanding r1-zero-like training: A critical perspective, 2025. URL https://arxiv.org/abs/2503.20783.

[13] Zhiyu Mei, Wei Fu, Kaiwei Li, Guangju Wang, Huanchen Zhang, and Yi Wu. Real: Efficient rlhf training of large language models with parameter reallocation. In Proceedings of the Eighth Conference on Machine Learning and Systems, MLSys 2025, Santa Clara, CA, USA, May 12-15, 2025. mlsys.org, 2025.

[14] Junhyuk Oh, Yijie Guo, Satinder Singh, and Honglak Lee. Self-imitation learning. In Jennifer Dy and Andreas Krause, editors, Proceedings of the 35th International Conference on Machine Learning, volume 80 of Proceedings of Machine Learning Research, pages 3878–3887. PMLR, 10–15 Jul 2018. URL https://proceedings.mlr.press/v80/oh18b.html.

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

[30] 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, 2025. URL https://arxiv.org/abs/2503.01491.

15

[16] OpenAI. Learning to reason with llms, 2024. URL https://openai.com/index/learning-to-reason-with-llms/.

[17] Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. Training language models to follow instructions with human feedback. Advances in neural information processing systems, 35:27730–27744, 2022.

[18] Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. Training language models to follow instructions with human feedback. Advances in neural information processing systems, 35:27730–27744, 2022.

[19] Qwen. Qwq-32b: Embracing the power of reinforcement learning, 2024. URL https://qwenlm.github.io/blog/qwq-32b/.

[20] John Schulman, Philipp Moritz, Sergey Levine, Michael Jordan, and Pieter Abbeel. High-dimensional continuous control using generalized advantage estimation. arXiv preprint arXiv:1506.02438, 2015.

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

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

[23] Wei Shen, Guanlin Liu, Zheng Wu, Ruofei Zhu, Qingping Yang, Chao Xin, Yu Yue, and Lin Yan. Exploring data scaling trends and effects in reinforcement learning from human feedback. arXiv preprint arXiv:2503.22230, 2025.

[24] Richard S Sutton, Andrew G Barto, et al. Reinforcement learning: An introduction, volume 1. MIT press Cambridge, 1998.

[25] Gemini Team, Rohan Anil, Sebastian Borgeaud, Yonghui Wu, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M Dai, Anja Hauth, et al. Gemini: a family of highly capable multimodal models. arXiv preprint arXiv:2312.11805, 2023.

[26] 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.

[27] Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed H. Chi, Quoc V. Le, and Denny Zhou. Chain-of-thought prompting elicits reasoning in large language models. In Sanmi Koyejo, S. Mohamed, A. Agarwal, Danielle Belgrave, K. Cho, and A. Oh, editors, Advances in Neural Information Processing Systems 35: Annual Conference on Neural Information Processing Systems 2022, NeurIPS 2022, New Orleans, LA, USA, November 28 - December 9, 2022, 2022.

[28] XAI. Grok 3 beta — the age of reasoning agents, 2024. URL https://x.ai/news/grok-3.

[29] Qiying Yu, Zheng Zhang, Ruofei Zhu, Yufeng Yuan, Xiaochen Zuo, Yu Yue, Tiantian Fan, Gaohong Liu, Lingjun Liu, Xin Liu, Haibin Lin, Zhiqi Lin, Bole Ma, Guangming Sheng, Yuxuan Tong, Chi Zhang, Mofan Zhang, Wang Zhang, Hang Zhu, Jinhua Zhu, Jiaze Chen, Jiangjie Chen, Chengyi Wang, Hongli Yu, Weinan Dai, Yuxuan Song, Xiangpeng Wei, Hao Zhou, Jingjing Liu, Wei-Ying Ma, Ya-Qin Zhang, Lin Yan, Mu Qiao, Yonghui Wu, and Mingxuan Wang. Dapo: An open-source llm reinforcement learning system at scale, 2025. URL https://arxiv.org/abs/2503.14476.

[30] 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, 2025. URL https://arxiv.org/abs/2503.01491.