

VAPO: 用于高级推理任务的高效且可靠的强化学习

ByteDance Seed

Full author list in Contributions

Abstract

我们提出了 VAPO, **V**alue-model-based **A**ugmented Proximal **P**olicy **O**ptimization 框架, 用于推理模型的一个新颖框架, 专为基于价值模型范式中的推理模型量身定制。在 AIME 2024 数据集进行基准测试中, VAPO 建立在 Qwen 32B 预训练模型之上, 达到了 **60.4** 的最新得分。在相同实验设置下直接比较, VAPO 的表现超过了此前报告的 DeepSeek-R1-Zero-Qwen-32B 和 DAPO 的结果超过 10 分。VAPO 的训练过程以其稳定性和高效性而脱颖而出, 仅需 5,000 步即可达到最新性能。此外, 在多次独立运行中, 训练过程没有发生崩溃, 突显了其可靠性。本研究深入探讨了使用基于价值模型的强化学习框架进行长链思维 (long-CoT) 推理。我们指出了困扰基于价值模型方法的三个关键挑战: 价值模型偏差、异质序列长度的存在以及奖励信号的稀疏性。通过系统设计, VAPO 提供了一个集成解决方案, 有效缓解了这些挑战, 从而在长链思维推理任务中实现了性能提升。

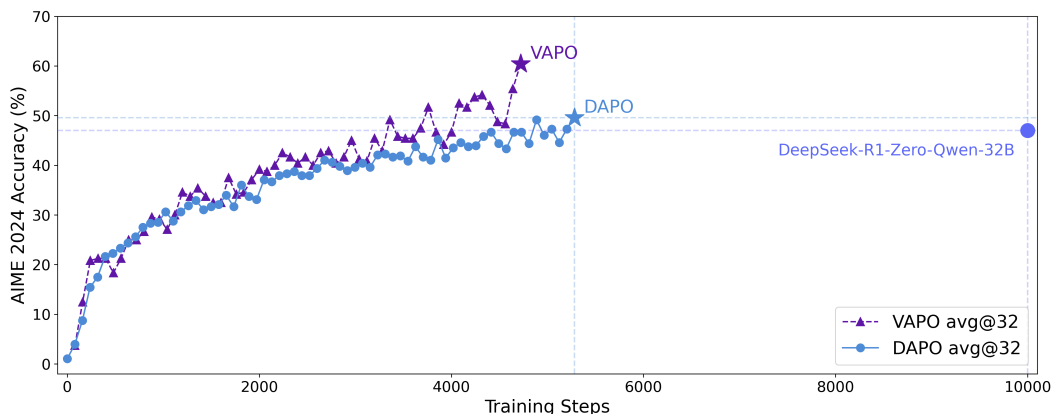
Date: 2025 年 6 月 15 日**Correspondence:** Yu Yue at yueyu@bytedance.com

Figure 1 AIME 2024 中 VAPO 在 Qwen2.5-32B 基础模型上的得分, 显示出相对于之前的最先进 (SOTA) 方法 DAPO 的显著优势, 同时使用显著更少的训练步骤实现。横轴表示梯度更新步骤。

1 介绍

诸如 OpenAI O1 [16] 和 DeepSeek R1 [6] 等推理模型 [5, 19, 26] 显著推动了人工智能的发展，其在复杂任务（如数学推理）中表现出色，这些任务需要在测试时通过长链思维（CoT）[27]进行逐步分析和问题解决。强化学习（RL）在这些模型的成功中发挥了关键作用 [1, 8, 10, 13, 22, 24, 26, 29]。通过在可验证问题上不断探索推理路径以获得正确答案，它逐渐提高了模型的性能，实现了前所未有的推理能力。

在大语言模型（LLM）[2–4, 11, 15, 25, 28]的RL训练中，诸如GRPO [22]和DAPO [29]等无价值模型方法展示了显著的效果。这些方法消除了学习价值模型的计算负担，而是仅基于整个轨迹的最终奖励计算优势。轨迹级别的优势被直接分配为序列中每个位置的令牌级别优势。当训练可靠的价值模型特别具有挑战性时，无价值模型方法通过在组内平均多个轨迹的奖励，提供准确且稳定的优势计算基线。这种基于组的奖励聚合减轻了对显式价值估算的需求，这在复杂任务中通常不稳定。因此，无价值模型方法在解决诸如长链推理等困难问题上获得了显著的关注，研究重点集中在优化其框架。

尽管无价值模型方法取得了显著成功，我们认为如果可以解决价值模型训练中的挑战，基于价值模型的方法可能拥有更高的性能上限。首先，价值模型通过准确追踪每个动作对后续回报的影响，实现更精细的优化 [21]。这对于复杂推理任务尤为重要，因为个别步骤中的细微错误经常导致灾难性失败，而在无价值模型框架下进行模型优化仍具挑战性 [30]。其次，与无价值模型方法中基于蒙特卡罗方法的优势估算相比，价值模型可以为每个令牌提供低方差的价值估算，从而增强训练的稳定性。此外，训练良好的价值模型具有内在的泛化能力，能够更有效地利用在线探索过程中遇到的样本。这显著提高了强化学习算法的优化上限。因此，尽管在复杂问题中训练价值模型面临巨大挑战，克服这些困难的潜在收益是巨大的。

然而，在长链任务中训练完美的价值模型面临重大挑战。首先，鉴于长轨迹和以引导方式学习价值的不稳定性，学习低偏差价值模型并非易事。其次，同时处理短响应和长响应也具有挑战性，因为它们优化过程中可能表现出非常不同的偏差-方差权衡偏好。最后但同样重要的是，来自验证器的奖励信号稀疏性由于长链模式而进一步加剧，这本质上需要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模型的方法在推理任务中的潜力，我们提出了**Value Augmented proximal Policy Optimization (VAPO)**，一个基于价值模型的RL训练框架。VAPO 从先前的研究工作如 VC-PPO [30] 和 DAPO [29] 中汲取灵感，并进一步扩展了它们的概念。

我们总结了我们的主要贡献如下：

1. 我们介绍了 VAPO，这是第一个基于价值模型的强化学习训练框架，在长 COT 任务中显著超越了无价值模型方法。VAPO 不仅在性能方面表现出卓越的优势，还展示了增强的训练效率，简化了学习过程，并强调了其作为该领域新基准的潜力。
2. 我们提出了长度自适应 GAE，它根据响应长度自适应地调整 GAE 计算中的 λ 参数。通过这样做，它有效地满足了与长度高度可变的响应相关的不同偏差-方差权衡要求。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差异较大的情况下。
3. 我们系统地整合了先前工作的技术，如来自 DAPO [29] 的 Clip-Higher 和 Token-level Loss，来自 VC-PPO [30] 的 Value-Pretraining 和 Decoupled-GAE，来自 SIL [14] 的自我模仿学习，以及来自

GRPO [22] 的组采样。此外，我们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，将这些改进结合在一起。这些增强措施协同工作，导致合并结果优于各个部分的简单相加。我们使用Qwen2.5-32B预训练模型进行实验，确保在任何实验中都没有引入SFT数据，以保持与相关工作的可比性（DAPO 和 DeepSeek-R1-Zero-Qwen-32B）。**VAPO** 的性能从原始PPO的得分5提高到60，超过了之前的SOTA无价值模型方法DAPO [29] 10分。更重要的是，**VAPO** 高度稳定——我们在训练期间没有观察到任何崩溃，并且多次运行的结果始终相似。

2 预备知识

本节介绍作为我们所提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和广义优势估计。

2.1 将语言生成建模为令牌级MDP

强化学习的核心是学习一种策略，使代理在与环境交互时最大化累积奖励。在本研究中，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]。

令提示表示为 x ，对该提示的响应表示为 y 。 x 和 y 都可以分解为令牌序列。例如，提示 x 可以表示为 $x = (x_0, \dots, x_m)$ ，其中令牌来自固定的离散词汇 \mathcal{A} 。

我们将令牌级MDP定义为元组 $\mathcal{M} = (\mathcal{S}, \mathcal{A}, \mathbb{P}, R, d_0, \omega)$ 。以下是每个组件的详细分解：

- **状态空间 (\mathcal{S}):** 此空间包含了在给定时间步之前生成的所有可能状态。在时间步 t ，状态 s_t 定义为 $s_t = (x_0, \dots, x_m, y_0, \dots, y_t)$ 。
- **动作空间 (\mathcal{A}):** 它对应于固定的离散词汇表，从中选择生成过程中的标记。
- **动态模型 (\mathbb{P}):** 这些表示标记之间的确定性转换模型。给定状态 $s_t = (x_0, \dots, x_m, y_0, \dots, y_t)$ ，动作 $a = y_{t+1}$ ，以及后续状态 $s_{t+1} = (x_0, \dots, x_m, y_0, \dots, y_t, y_{t+1})$ ，则概率 $\mathbb{P}(s_{t+1}|s_t, a) = 1$ 。
- **终止条件:** 语言生成过程在终止动作 ω 执行时结束，通常是句子结束标记。
- **奖励函数 ($R(s, a)$):** 此函数提供标量反馈，以评估智能体在状态 s 下执行动作 a 后的表现。在从人类反馈中进行强化学习 (RLHF) [18, 23] 的背景下，奖励函数可以从人类偏好中学习，或通过特定任务的规则集定义。
- **初始状态分布 (d_0):** 这是一个关于提示 x 的概率分布。初始状态 s_0 包含提示 x 内的标记。

2.2 RLHF 学习目标

我们将优化问题表述为一个 KL 正则化的 RL 任务。我们的目标是逼近最优的 KL 正则化策略，其表示为：

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

在此方程中， H 表示决策步骤的总数， s_0 是从数据集中采样的提示， $R(s_t, a_t)$ 是从奖励函数中获得的基于 token 的奖励， 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 RLHF 和大多数与 LLM 相关的任务中，奖励是稀疏的，仅在终端动作 ω ，即句子结束 token `<eos>` 时分配。

2.3 近端策略优化

PPO [21] 使用截断的替代目标来更新策略。其关键思想是在每次更新步骤中限制策略的变化，防止过大的策略更新导致不稳定。

设 $\pi_\theta(a|s)$ 为参数化为 θ 的策略， $\pi_{\theta_{\text{old}}}(a|s)$ 为上一迭代中的旧策略。PPO 的替代目标函数定义为：

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

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

广义优势估计 [20] 是一种在 PPO 中用来更准确估计优势函数的技术。它结合了多步引导来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 \hat{A}_t 计算为：

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

其中 γ 是折扣因子， $\lambda \in [0, 1]$ 是 GAE 参数， $\delta_t = R(s_t, a_t) + \gamma V(s_{t+1}) - V(s_t)$ 是时序差分（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 RLHF 中常用的做法是使用折扣因子 $\gamma = 1.0$ ，为简化记号，本文后续部分将省略 γ 。

3 长链式思维路径强化学习在推理任务中的挑战

长链式思维路径任务对强化学习训练带来了独特的挑战，特别是对于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、价值函数不稳定性和奖励稀疏性引发的技术问题。

3.1 长序列上的价值模型偏差

如 VC-PPO 中所指出的 [30]，用奖励模型初始化价值模型会引入显著的初始化偏差。这种正偏差来源于两个模型之间的目标不匹配。奖励模型被训练在 `<EOS>` 标记上打分，激励其对较早的标记给予较低分数，因为这些标记的上下文不完整。相比之下，价值模型估计在给定策略下 `<EOS>` 之前所有标记的预期累计奖励。在训练初期阶段，由于 GAE 的反向计算，每个时间步 t 都会存在一个正偏差，并沿着轨迹累积。

使用 $\lambda = 0.95$ 的 GAE 的另一种常见做法可能会加剧这一问题。在终止标记处的奖励信号 $R(s_T, \text{<EOS>})$ 向后传播为 $\lambda^{T-t} R(s_T, \text{<EOS>})$ 到第 t 个标记。对于长序列而言，当 $T - t \gg 1$ 时，这种折扣会将有效的奖励信号降低到接近于零。因此，价值更新几乎完全依赖于高度有偏差的估计，削弱了价值模型作为可靠的方差降低基线的作用。

3.2 训练期间的异构序列长度

在复杂的推理任务中，长的CoT对得出正确答案至关重要，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具备足够的鲁棒性，以管理从非常短到极长的序列。因此，常用的具有固定 λ 参数的GAE方法面临显著的挑战。

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。对于短长度的响应，通过GAE获得的估计往往受到高方差的影响。这是因为GAE在偏差和方差之间进行权衡。在短响应的情况下，估计值偏向于方差主导的一侧。另一方面，对于长长度的响应，GAE由于引导而常常导致高偏差。GAE的递归性质依赖于未来状态值，在长序列中累积错误，进一步加剧偏差问题。这些限制深深植根于GAE计算框架的指数衰减性质。

3.3 基于验证器任务的奖励信号稀疏性

复杂的推理任务经常使用验证器作为奖励模型 [6, 16]。与传统的基于语言模型的奖励模型提供密集信号（例如从-4到4的连续值）不同，基于验证器的奖励模型通常提供二元反馈，例如0和1。奖励信号的稀疏性随着长链推理（CoT）而进一步加剧。由于CoT显著延长了输出长度，这不仅增加了计算时间，还减少了接收非零奖励的频率。在策略优化中，采样到的正确答案的响应可能极其稀少且珍贵。

这种情况提出了一个明显的探索-利用困境。一方面，模型必须保持相对高的不确定性。这使其能够采样多样化的响应，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。如果未能在探索和利用之间取得适当的平衡，模型可能由于过度利用而陷入次优解，或者在无效的探索上浪费计算资源。

4 VAPO: 解决长链推理强化学习中的挑战

4.1 缓解长序列上的价值模型偏差

基于第3.1节中提出的价值模型分析，我们建议使用价值预训练和解耦GAE来解决长序列上价值模型偏差的关键挑战。这两种技术都借鉴了VC-PPO中先前介绍的方法。

价值预训练被提议用于缓解价值初始化偏差。直接将PPO应用于长CoT任务会导致诸如输出长度崩溃和性能下降等失败。原因在于价值模型是从奖励模型初始化的，而奖励模型与价值模型的目标不匹配。这种现象首先在VC-PPO中被识别和解决[30]。在本文中，我们遵循价值预训练技术，具体步骤如下所示：

1. 通过从固定策略（例如， π_{sft} ）进行采样来连续生成响应，并使用蒙特卡罗回报来更新价值模型。
2. 训练价值模型，直到包括价值损失和解释方差 [7] 在内的关键训练指标达到足够低的值。
3. 保存价值检查点，并加载此检查点以进行后续实验。

解耦GAE在VC-PPO中被证明是有效的[30]。该技术将价值和策略的优势计算解耦。对于价值更新，建议使用 $\lambda = 1.0$ 来计算价值更新目标。这个选择可以实现无偏的梯度下降优化，有效解决长CoT任务中的奖励衰减问题。

然而，对于策略更新，建议使用较小的 λ 以在计算和时间限制下加速策略收敛。在VC-PPO中，这是

通过在优势计算中使用不同的系数来实现的： $\lambda_{\text{critic}} = 1.0$ 和 $\lambda_{\text{policy}} = 0.95$ 。在本文中，我们采用了解耦GAE计算的核心思想。

4.2 训练中异构序列长度的管理

为了解决训练过程中异构序列长度的挑战，我们提出了**长度自适应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，我们用基于token级别的策略梯度损失替代了传统的样本级别策略梯度损失。关键的技术细节如下所述：

长度自适应GAE专门针对不同长度序列中最佳 λ_{policy} 值的不一致性进行了提出。在VC-PPO中， λ_{policy} 被设定为常数值 $\lambda_{\text{policy}} = 0.95$ 。然而，当考虑GAE计算时，对于长度 $l > 100$ 的较长输出序列，与奖励对应的TD误差的系数为 $0.95^{100} \approx 0.006$ ，这实际上为零。结果是，当固定 $\lambda_{\text{policy}} = 0.95$ 时，GAE计算可能会被潜在的偏置的引导TD误差所主导。这种方法可能不适用于处理极长的输出序列。

为了解决这一缺点，我们提出了用于策略更新的**长度自适应GAE**。我们的方法旨在确保TD误差在短序列和长序列之间的分布更加均匀。我们设计了系数和 λ_{policy} 的总和与输出长度 l 成正比：

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

其中 α 是一个控制整体偏差-方差权衡的超参数。通过解方程4得到 λ_{policy} 的长度自适应公式：

$$\lambda_{\text{policy}} = 1 - \frac{1}{\alpha l} \quad (5)$$

这种长度自适应的 λ_{policy} 在GAE计算中的方法能够更有效地处理不同长度的序列。

Token级别的策略梯度损失。参考DAPO [29]，我们还修改了策略梯度损失的计算方法，以调整长COT场景中的损失权重分配。具体而言，在之前的实现中，策略梯度损失的计算如下：

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

其中 G 是训练批次的大小， o_i 是第 i 个样本的轨迹。在这种损失公式中，所有token的损失首先在序列级别上进行平均，然后在批次级别上进一步平均。这种方法导致来自较长序列的token在最终损失值中贡献较少。因此，如果模型在处理长序列时遇到关键问题，这种在RL训练探索阶段容易发生的情况，由于它们缩减的权重导致的抑制不足可能导致训练不稳定甚至崩溃。为了应对token级对最终损失贡献的不平衡，我们将损失函数修改为以下形式：

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

在单个训练批次中，所有token被赋予统一的权重，从而能够更有效地解决长序列带来的问题。

4.3 处理基于验证器任务中奖励信号稀疏的问题

如在第 3.3 节中分析的那样，在奖励信号极为稀疏的场景下，提高RL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、正例语言模型损失和组合采样。技术细节如下所述：

Clip-Higher 用于缓解PPO和GRPO训练过程中遇到的熵崩溃问题，该方法首次在DAPO中提出 [29]。我们将较低和较高的剪辑范围解耦为 ϵ_{low} 和 ϵ_{high}

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

我们增加 ϵ_{high} 的值，以为低概率标记的增加留出更多空间。我们选择保持 ϵ_{low} 相对较小，因为增加它会将这些标记的概率压制到0，从而导致采样空间的崩溃。

正例语言模型损失 旨在提高RL训练过程中正样本的利用效率。在复杂推理任务的RL背景下，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。传统的策略优化策略在RL训练中效率低下，因为试错机制会带来大量计算成本。鉴于这一挑战，当策略模型采样出正确答案时，最大化其效用至关重要。为了解决这一挑战，我们采用模仿学习的方法，通过为RL训练中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：

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

其中 \mathcal{T} 表示正确答案的集合。最终的NLL损失与策略梯度损失通过加权系数 μ 结合在一起，共同作为更新策略模型的目标：

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

组合采样 用于在同一提示中采样有辨别力的正负样本。在给定的计算预算下，存在两种主要的计算资源分配方法。第一种方法是尽可能多地利用提示，每个提示仅采样一次。第二种方法是减少每批次的不同提示数量，并将计算资源引导至重复生成。我们观察到后者表现稍好，归因于其引入了更丰富的对比信号，增强了策略模型的学习能力。

5 实验

5.1 训练细节

在这项工作中，我们通过在Qwen-32B模型的基础上对PPO算法进行多种修改来增强模型的数学性能。这些技术对于其他推理任务（如代码相关任务）也同样有效。对于基础的PPO，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，每个提示采样一次，每个小批量大小设置为512。价值网络使用奖励模型初始化，GAE λ 设置为0.95， γ 设置为1.0。使用样本级损失，剪辑 ϵ 设置为0.2。

Table 1 VAPO 的消融结果

Model	AIME24 _{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

与原始的PPO相比，VAPO做了以下参数调整：

1. 在策略训练开始之前，基于奖励模型（RM）实现了一个50步的价值网络预热。
2. 使用了解耦的GAE，其中价值网络从用 $\lambda=1.0$ 估计的回报中学习，而策略网络则从使用不同 λ 获得的劣势中学习。
3. 根据序列长度自适应地设置优势估计的 λ ，遵循公式： $\lambda_{\text{policy}} = 1 - \frac{1}{\alpha l}$ ，其中 $\alpha = 0.05$ 。
4. 将clip范围调整为 $\epsilon_{\text{high}}=0.28$ 和 $\epsilon_{\text{low}}=0.2$ 。
5. 采用了基于token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. 每次采样使用512个提示词，每个提示词采样16次，并将小批量大小设置为512。

我们还将单独展示从VAPO中移除这七项修改之一的最终效果。对于评估指标，我们使用AIME24在32次中的平均通过率，采样参数设置为 $\text{topp}=0.7$ 和 $\text{temperature}=1.0$ 。

5.2 消融结果

在 Qwen-32b 上，使用 GRPO 的 DeepSeek R1 在 AIME24 上获得 47 分，而 DAPO 在更新步骤的 50% 时达到 50 分。在图 1 中，我们提出的 VAPO 仅使用 DAPO 步骤的 60% 就匹配了这一性能，并在仅 5,000 步内达到了新的 SOTA 分数 60.4，展示了 VAPO 的高效。此外，VAPO 保持了稳定的熵——既没有崩溃，也没有过高——并且在三次重复实验中始终达到 60-61 的最高分，突出了我们算法的可靠性。

Table 1 系统地呈现了我们的实验结果。受限于值模型学习崩溃，Vanilla PPO 方法在训练后期仅获得 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，这是一项显著的改进。我们通过分别消融七个提出的修改，进一步验证了其有效性：

1. 没有值预训练时，模型在训练过程中与基础PPO一样经历崩溃，收敛到大约11分的最大值。

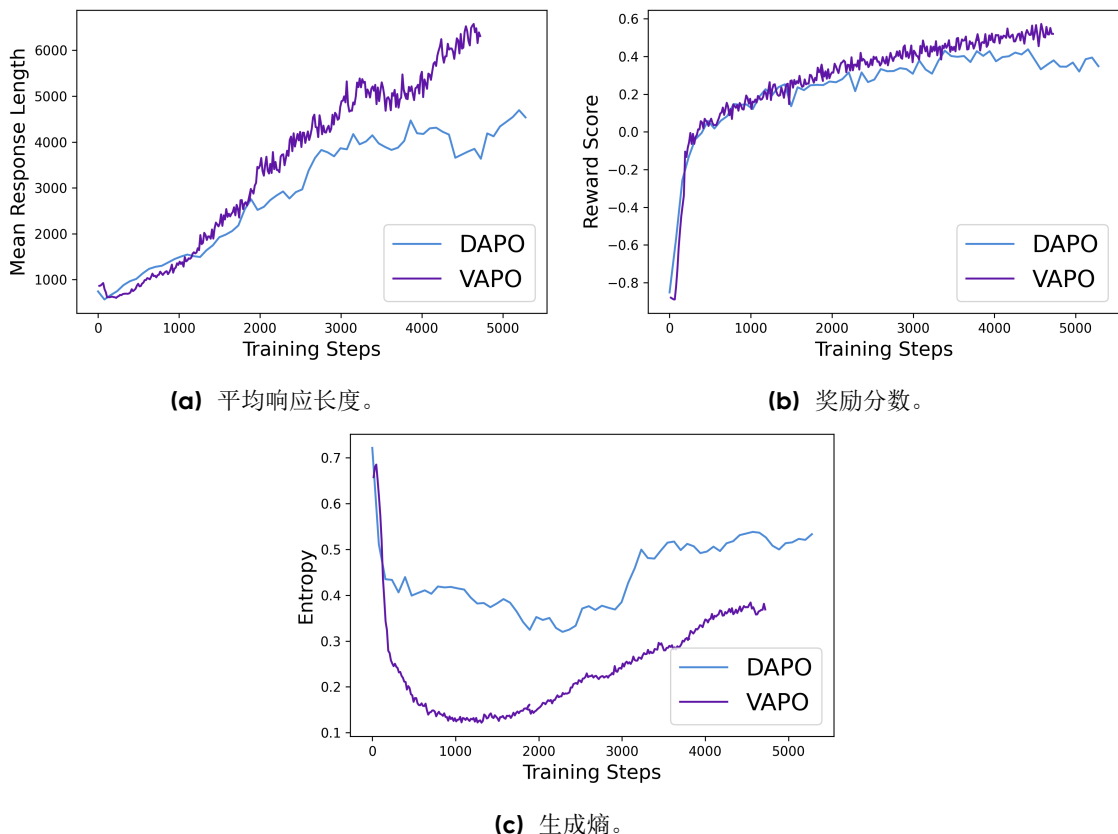


Figure 2 VAPO 的响应长度、奖励分数和生成熵的指标曲线。

2. 移除解耦的GAE会导致奖励信号在反向传播过程中呈指数级衰减，阻止模型完全优化长篇回复，导致27分的下降。
3. 自适应GAE平衡了对短和长回复的优化，带来15分的提升。
4. Clip higher鼓励彻底的探索和开发；其移除限制了模型的最大收敛至46分。
5. 令牌级损失隐含地增加了长回复的权重，贡献了7分的增益。
6. 合并正例LM损失使模型提升了近6分。
7. 使用组采样生成较少提示但重复次数更多，也带来了5分的改进。

5.3 训练动态

在强化学习训练过程中生成的曲线提供了对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的差异。普遍认为，变化越平滑和增长越快是这些曲线的理想特征。通过比较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 更平滑，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优于 DAPO。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标志，因为它增强了模型的泛化能力。

- [Figure 2b](#) 显示 VAPO 的得分增长速度比 DAPO 快，因为价值模型为模型提供了更多细致的信号来加速优化。
- 根据 [Figure 2c](#)，在训练的后期阶段，VAPO 的熵比 DAPO 更低。这是一枚硬币的两面：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微乎其微，而可重复性和稳定性证明是非常有利的。

6 相关工作

OpenAI O1 [16] 在大语言模型中引入了一种深刻的范式转变，其特征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（无价值模型的 GRPO [22]）和模型权重，其性能可与 O1 相媲美。DAPO [29] 识别出之前未公开的挑战，如在不价值模型的大语言模型强化学习扩展中遇到的熵崩溃，并提出了四种有效技术来克服这些挑战，实现了业界最先进的表现。最近，Dr. GRPO [12] 移除了 GRPO 中的长度和标准差归一化项。另一方面，ORZ [9] 遵循 PPO 并利用价值模型进行优势估计，提出使用蒙特卡洛估计代替广义优势估计。然而，他们只能达到与无价值模型方法如 GRPO 和 DAPO 相当的性能。在本文中，我们也遵循基于价值模型的方法，并提出 VAPO，其性能优于最先进的无价值模型算法 DAPO。

7 结论

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 基准测试中实现了 SOTA 性能。通过在 PPO 基础上引入七种新的技术，这些技术侧重于改进价值学习和平衡探索，我们的基于价值模型的方法优于像 GRPO 和 DAPO 等当前的无价值模型方法。该工作为推进大型语言模型在推理密集型任务中的应用提供了一个稳健的框架。

贡献

项目负责人

Yu Yue¹

算法

Yu Yue¹, Yufeng Yuan¹, Qiyang Yu^{1,2}, Xiaochen Zuo¹, Ruofei Zhu¹, Wenyuan Xu¹, Jiaze Chen¹, Chengyi Wang¹, TianTian Fan¹, Zhengyin Du¹, Xiangpeng Wei¹, Xiangyu Yu¹

基础设施*

Gaohong Liu¹, Juncai Liu¹, Lingjun Liu¹, Haibin Lin¹, Zhiqi Lin¹, Bole Ma¹, Chi Zhang¹, Mofan Zhang¹, Wang Zhang¹, Hang Zhu¹, Ru Zhang¹

*姓氏按字母顺序排列

监督

Xin Liu¹, Mingxuan Wang¹, Yonghui Wu¹, Lin Yan¹

所属机构

¹ 字节跳动种子

² 清华大学AIR实验室与字节跳动种子

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.

- [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] Qiyang 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>.