

中图法分类号: TP183 文献标识码: 文章编号: 1006-8961(XXXX)XX-0001-22

论文引用格式: Miao Jiyuan, Wang Shiyu, Qi Honggang, Zhao Di. A Comprehensive Review of LLM's Policy Optimization Algorithms[J/OL]. Journal of Image and Graphics, XXXX: 1-22. DOI: 10.11834/jig.260027. (缪济远, 王适雨, 齐洪钢, 赵地. 大语言模型策略优化方法: 分类、进展与展望[J/OL]. 中国图象图形学报, XXXX: 1-22. DOI: 10.11834/jig.260027.) [DOI: 10.11834/jig.260027]

大语言模型策略优化方法: 分类、进展与展望

缪济远^{1,4}, 王适雨^{2,3}, 齐洪钢^{1,2}, 赵地^{1,2*}

1. 中国科学院计算技术研究所, 北京 100190; 2. 中国科学院大学, 北京 100049; 3. 河南理工大学软件学院, 河南 454000; 4. 北京化工大学

摘要: 策略优化(Policy Optimization, PolOpt)是增强大语言模型推理能力的核心技术, 广泛应用于从 RLHF 到 DeepSeek-R1 等主流模型中。然而, 随着 PolOpt 算法的快速演进, 该领域缺乏系统性的梳理与分类, 研究人员难以从零散文献中把握技术全貌与发展脉络。本文对 PolOpt 领域的最新进展进行了系统性综述, 覆盖从基础 PolOpt (PPO、GRPO) 到 GRPO 变体, 再到扩展 PolOpt 的完整发展谱系。我们提出了一个分层分类框架 (B1 - B3、C1 - C7、D1 - D19), 从优化范式演进的角度对现有算法进行了有机组织。基于该框架, 本文总结了不同算法类别的核心改进维度、适用场景与性能特点, 并揭示了当前 PolOpt 研究面临的三大瓶颈: 奖励稀疏性与噪声放大、多目标权重冲突、计算资源与扩展性限制。最后, 本文展望了未来在奖励密集化、多目标自适应优化及分布式训练架构等方向的研究机遇。

关键词: 大语言模型; 推理模型; 策略优化; 综述; 分类学

A Comprehensive Review of LLM's Policy Optimization Algorithms

Miao Jiyuan^{1,4}, Wang Shiyu^{2,3}, Qi Honggang^{1,2}, Zhao Di^{1,2*}

1. Institute of Computing Technology, Chinese Academy of Sciences, Beijing 100190, China; 2. School of Computer Science and Technology, University of Chinese Academy of Sciences, Beijing 100049, China; 3. College of Software, Henan Polytechnic University, Henan 454000, China; 4. Beijing University of Chemical Technology, Beijing 100029, China

Abstract: Policy Optimization (PolOpt) has become the backbone alignment technology for modern reasoning-oriented large language models (LLMs), powering state-of-the-art systems from OpenAI's GPT series, Anthropic Claude to open-source reasoning models represented by DeepSeek-R1, Qwen2.5-MAX and LLaMA-4. Unlike static supervised fine-tuning (SFT) that merely fits fixed labeled corpora, PolOpt leverages reinforcement learning (RL) to iteratively adjust model output distributions via numerical reward signals, bridging the gap between model generation and complex human preferences, logical reasoning rules, multi-step mathematical derivation and coding verification standards. Since the proposal of Proximal Policy Optimization (PPO) in 2017, and the birth of Group Relative Policy Optimization (GRPO) in 2024 optimized for LLM group sampling scenarios, a huge family of derivative PolOpt algorithms has emerged within only two years (2024-2026). These variants have been widely deployed in mathematical reasoning, code generation, multimodal generation, agent task automation and simultaneous machine translation, yet existing academic reviews only focus on partial branches such as RLHF pipeline or single GRPO improvement points. No systematic hierarchical taxonomy covering the

收稿日期: 2026-01-13; 修回日期: 2026-09-15

* 通信作者: zhaodi@ict.ac.cn

基金项目: 本研究获得中国科学院计算技术研究所 2025 年度春晖计划项目资助。

Supported by: This paper was supported by Springboard Program from Institute of Computing Technology, Chinese Academy of Sciences in 2025.

© 中国图象图形学报版权所有

full evolution spectrum of PoOpt has been proposed, which creates heavy information fragmentation and raises high learning barriers for researchers who intend to select, modify or customize PoOpt algorithms for LLM reasoning tasks. To fill this research gap, this paper systematically sorts out all mainstream PoOpt algorithms published from 2017 to mid-2026, and constructs a four-layer hierarchical classification framework (Level A: General PoOpt; Level B: three core paradigms B1 Basic PoOpt / B2 GRPO variants / B3 Extended PoOpt; Level C 7 subdimensions; Level D 19 fine-grained algorithm clusters). The timeline evolution diagram clearly distinguishes the three core technical routes of PoOpt: the early constraint-based baseline paradigm (B1, including PPO and GRPO), the intra-framework optimized GRPO variants (B2 with four orthogonal improvement dimensions: sampling modularization, reward shaping, advantage variance control, training efficiency tuning), and cross-paradigm breakthrough extended PoOpt (B3 that abandons GRPO's core objective function and reconstructs new learning frameworks for cross-domain, multi-agent, diffusion generation scenarios). This taxonomy strictly defines the classification boundary between B2 and B3: algorithms retaining GRPO's group relative reward mechanism and KL constraint architecture are categorized as GRPO variants, while those that restructure policy update logic, learning signals or overall training pipelines are classified as extended PoOpt. On the basis of the proposed taxonomy, this paper conducts a comprehensive comparative analysis of over 50 representative PoOpt algorithms including DAPO, RePO, COPO, MAGPO, DPPPO, MiMo-7B. For each subclass under C1-C7, we elaborate their core innovation logic, mathematical derivation highlights, hyperparameter sensitivity, computational overhead, convergence characteristics and suitable downstream tasks. We also summarize horizontal comparison metrics covering reasoning accuracy, GPU memory consumption, training stability, generalization ability across math/code/multimodal tasks to provide explicit algorithm selection guidelines for industrial and academic practitioners. Through systematic induction of existing experimental results, this paper identifies three universal bottlenecks restricting current PoOpt performance: (1) reward sparsity and noise amplification, where terminal-only binary rewards in reasoning tasks lead to weak gradient signals and group normalization magnifies random noise; (2) multi-objective weight conflict, static linear weighting fails to balance accuracy, conciseness, safety and creativity simultaneously; (3) scalability barriers under trillion-parameter LLMs, including high value network overhead, inefficient distributed communication and reward model coupling costs. Furthermore, this paper forecasts three high-potential future research directions corresponding to the three core bottlenecks: first, reward densification and robust estimation, combining intermediate process reward models, multi-modal joint feedback and Bayesian noise filtering to alleviate sparse gradient issues; second, adaptive multi-objective Pareto optimization, introducing dynamic weight adjustment and hierarchical reward architecture to resolve target competition; third, lightweight distributed PoOpt training frameworks, including reward model distillation, parameter-efficient fine-tuning decoupling and asynchronous group sampling architectures to cut hardware costs. Different from previous review papers that only summarize technical details, this work builds a traceable evolution roadmap for PoOpt and unifies scattered mathematical symbols, loss functions and group sampling formulas into a standardized symbol system, which lowers the reproduction cost of existing algorithms for follow-up researchers. In terms of practical value, the proposed classification system provides clear algorithm selection suggestions for different resource constraints and task scenarios: lightweight tasks with limited computing resources prioritize GFPO and RePO; mathematical/code reasoning tasks adopt GRPO, TreeRPO or COPO; multimodal image/video generation selects DanceGRPO, MixGRPO, DPPPO; safety alignment tasks apply MO-GRPO and GAP; real-time low-latency reasoning uses LAPO and HBPO. For open-source LLM communities such as Qwen and DeepSeek, this paper also sorts out the compatible matching relationship between mainstream PoOpt variants and open reasoning models, which can guide industrial practitioners to build low-cost, high-performance RL post-training pipelines. Finally, this paper briefly distinguishes PoOpt and Preference Optimization (PO, including DPO, CPL) to clarify the two branches of LLM alignment research, and points out the limitations of current GRPO derivative improvements—most incremental optimizations lack rigorous gradient theoretical proof, while extended PoOpt that breaks the original framework will become the mainstream innovative direction in the next two years. This systematic taxonomy and bottleneck analysis can support researchers to carry out targeted PoOpt algorithm innovation for complex LLM reasoning and multi-agent intelligent systems.

Key words: Large language models (LLM); reasoning models; policy optimization (PoOpt); comprehensive review; taxonomy

缪济远, 王适雨, 齐洪钢, 等. 大语言模型策略优化方法: 分类、进展与展望 [J]. 中国图象图形学报, 2026, XX (XX): XXXX-XXXX. DOI: 10. 11834/jig. 260***

Miao J Y, Wang S Y, Qi H G, et al. A Comprehensive Review of LLM's Policy Optimization Algorithms [J]. Journal of Image and Graphics, 2026, XX (XX): XXXX-XXXX. DOI: 10. 11834/jig. 260***

0 引言

近年来, 大规模模型技术的迅猛发展掀起了人工智能研究和应用的热潮, 受到了学术界和工业界的广泛关注。根据 Chatbot Arena 在 2025 年 8 月发布的排名 (2025), 领先的大型语言模型包括: OpenAI 的 GPT-5、谷歌的 Gemini 2.0 Ultra、DeepSeek 的 DeepSeek-R1、Qwen 的 Qwen2.5-MAX、Anthropic 的 Claude 3.7、抖音的 Doubao-1.5、Meta 的 LLaMA-4、百度 (ernie) 的 ernie 4.0 以及华为的 Pangu Model。在这些顶尖模型中, DeepSeek-R1、Qwen2.5-MAX、Doubao-1.5 和 LLaMA-4 采用了开源策略, 而其余模型则

仍保持专有状态。

语言模型专注于自然语言的理解与生成, 广泛应用于对话系统、机器翻译等领域。其典型代表包括 GPT-5、Gemini 2.0、Claude 3.7、Ernie 4.0、Spark X1 和盘古等。作为大型语言模型的一个重要分支, 推理模型在逻辑推理、数学计算以及解决复杂多步骤问题方面表现尤为突出。这类模型在符号推理、决策规划和高精度计算任务中展现出卓越的性能, 例如 GPT-5 在数学领域、Gemini 在多模态应用、DeepSeek-R1 在编程任务以及 Qwen 2.5 在逻辑推理方面均表现优异。

作为推理模型领域的核心方法论, PolOpt 在多个领域展现出巨大的应用潜力。然而, 目前尚缺乏对 PolOpt 最新进展进行全面梳理的系统性综述。这一知识空白导致研究人员在选择、改进或应用扩展 PolOpt 技术时, 不得不从零散的论文和技术报告中筛选信息, 由于缺乏完整的技术路线图, 研究过程变得极为艰难。本文旨在系统性地梳理 PolOpt 领域的最新进展, 并通过分类法预测未来算法发展趋势, 填补现有综述的空白。

Evolution of Policy Optimization (PolOpt) Algorithms for LLMs

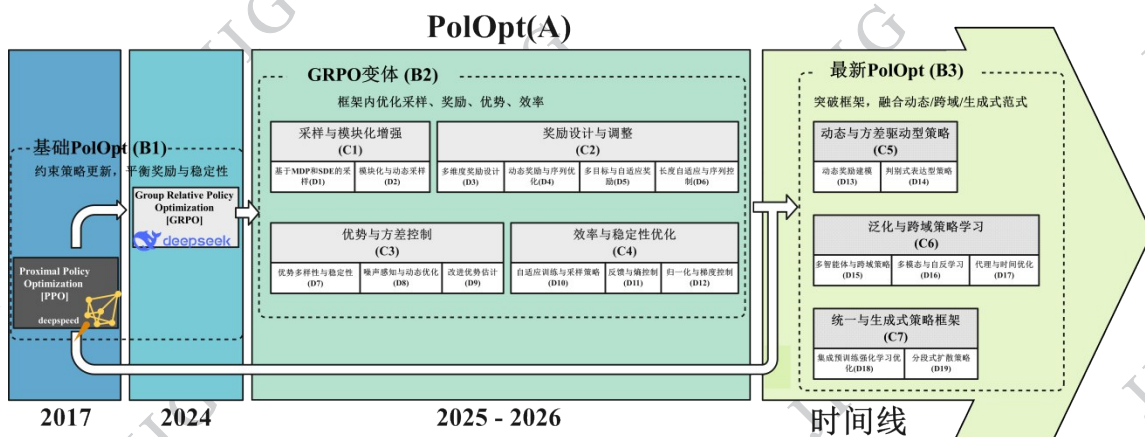


图 1 PolOpt 算法的详细分类体系包含四个层级

Fig. 1 The detailed taxonomy of PolOpt algorithms comprises four levels (A, B, C, D) with specific categories under each. Parentheses following the names indicate their respective category designations. Each category includes several specific PolOpt algorithms.

如图 1 所示, PolOpt 算法可细分为三大类别, 分别是基础 PolOpt (B1)、GRPO 变体 (B2) 以及扩展 PolOpt (B3)。具体的分类详情请参阅第 2 节, 相应的 mermaid 代码收录于附录 1。

本文将深入探讨以下主题: 第一部分介绍

PolOpt 的基本原理及其核心概念; 第二部分介绍 PolOpt 的具体分类; 第三部分阐述基础 PolOpt (B1) 的相关内容; 第四部分分析 GRPO 变体 (B2) 的特点; 第五部分探讨扩展 PolOpt (B3) 的进展; 第六部分总结研究成果, 并讨论潜在的未来研究方向。

1 PolOpt基本原理

本节将深入探讨PolOpt算法的显著优势及其核心概念与定义。为确保表述的清晰性和一致性,特此统一采用以下符号约定:常用数学符号(例如输入提示符 x 和响应序列 y)将在首次出现时进行正式定义,后续则直接使用,不再重复解释。关于所有常见数学符号及其精确含义的完整汇总,请参阅附录2。

1.1 PolOpt在大型语言模型中的优势

语言模型(Language models),特别是基于Transformer架构的模型,通常需要在庞大的数据集上进行预训练,以预测序列中的下一个标记。这一预训练过程能够生成流畅且符合语境的文本,但往往难以精确满足特定目标。例如,确保安全性,以规避有

害、偏见或不当的输出;提升实用性,以提供准确且契合用户需求的回应;以及维护伦理一致性,以遵循社会规范与价值观。

监督微调(Supervised Fine-Tuning, SFT)在经过精心整理的数据集上虽能提升模型的对齐效果,却难以捕捉人类的细微偏好或处理边缘案例。PolOpt通过引入奖励信号,迭代优化模型输出分布,突破了传统方法的局限,使模型能够更精准地适配那些难以用静态数据集编码的复杂目标。

1.2 强化学习

强化学习(Reinforcement Learning, RL)是一种机器学习范式,侧重于智能体通过与环境进行试错交互,学习一系列策略,以实现累积奖励的最大化(Zhao等,2018)。强化学习的概念如图2所示。

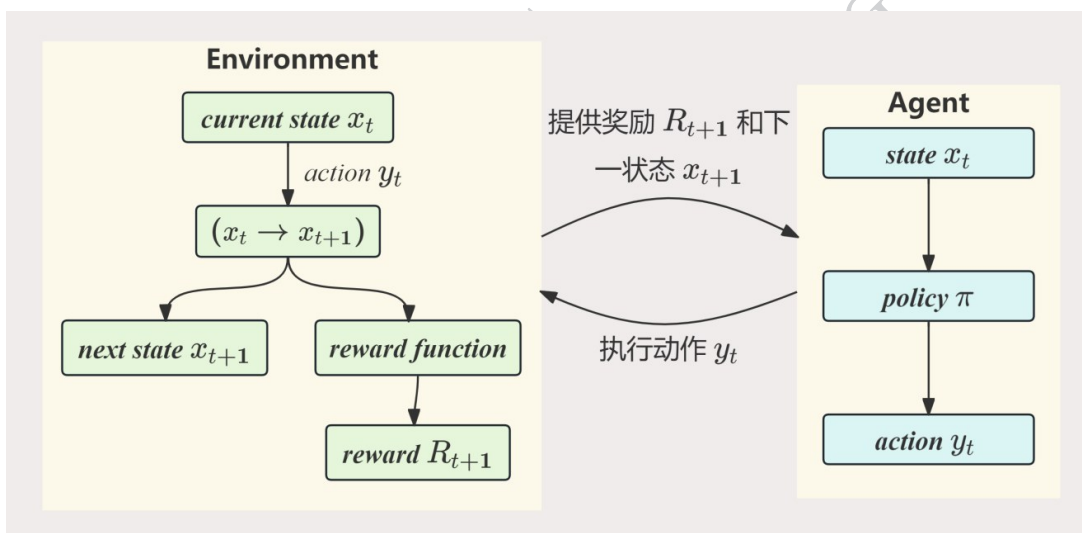


图2 强化学习基础交互框架

Fig. 2 Reinforcement Learning Basic Interaction Framework: At the end of a time step, the environment returns the reward y_{t+1} for the next state, and then transitions to the next time step. The Agent generates the next action based on the reward y_{t+1} and the next state x_{t+1} .

如图2所示,核心概念具体阐述如下:

1) *Environment* 是与 *Agent* 交互的外部系统,提供 *State* 和 *Reward* 反馈。

2) *Agent* 是学习和决策的主体,通过 *Policy* 选择 *Action*。

3) *State* 是给定时刻 *Environment* 的完整描述。

4) *Action* 是 *Agent* 根据 *Reward* 执行的操作。

5) 策略是 *State* 到 *Action* 的映射,通常表示为 $\pi_{\theta}(y|x)$,表示在状态 x 下选择动作 y 的概率。

6) *Reward Model (RM)* 依据输入数据为输出分配标量值,该值体现了输出质量是否符合预期目标。

1.3 PolOpt核心概念

PolOpt的核心概念涵盖以下几个关键要素:第1.3.1节所述的参考策略与旧策略、第1.3.2节定义的KL散度惩罚项、第1.3.3节阐述的价值函数、第1.3.4节描述的优势函数、第1.3.5节说明的目标函数,以及第1.3.6节详述的梯度更新机制(Pan等,2023)。

1.3.1 参考策略与旧策略

参考策略 π_{ref} 是指初始的 SFT 模型或基线策略, 该策略在整个训练过程中保持不变。 π_{ref} 通常用于计算 KL 散度惩罚, 具体如公式(1)所示:

$$KL(\pi_{\theta} \parallel \pi_{ref}) = E_{x \sim Q, y \sim \pi_{\theta}} \left[\log \frac{\pi_{\theta}(y|x)}{\pi_{ref}(y|x)} \right]. \quad (1)$$

旧策略 π_{old} 是指在每次更新之后, 当前策略将取代上次更新前的旧策略。 π_{old} 通常被用于计算 PolOpt 目标函数中的重要性采样比率, 具体公式如(2)所示:

$$r_t(\theta) = \frac{\pi_{\theta}(y_t|x_t, y_{<t})}{\pi_{old}(y_t|x_t, y_{<t})}. \quad (2)$$

1.3.2 KL 散度惩罚项

Kullback - Leibler (KL) 散度惩罚是 PolOpt 算法中的核心约束项, 旨在防止强化学习策略 π_{θ} 在训练过程中相对于参考策略 π_{ref} 出现过度偏离。这一约束在群组相对策略优化 (Group Relative Policy Optimization, GRPO) 和近端策略优化 (Proximal Policy Optimization, PPO) 等算法中尤为显著 (Schulman 等, 2017; Shao 等, 2024), 能够有效稳定训练过程, 避免因灾难性策略更新而导致的性能衰退。

KL 散度, 其定义详见公式(1), 旨在量化当前策略 π_{θ} 与参考策略 π_{ref} 之间的相对熵, 即在使用 π_{ref} 近似 π_{θ} 时所损失的信息量。KL 值越高, 说明这两种策略之间的差异越显著。

在优化目标中, KL 项被引入作为惩罚项, 正如公式(3)所示。通过抑制策略的剧烈波动, KL 散度使得学习过程更加平滑和稳定。这一特性在在线策略和策略梯度方法中尤为关键, 因为过于激进的更新可能会导致系统不稳定或收敛至次优策略。

$$J(\theta) = L(\theta) - \beta \cdot KL(\pi_{\theta} \parallel \pi_{ref}). \quad (3)$$

1.3.3 价值函数

价值函数用于评估状态的质量, 可分为状态价值函数和动作价值函数:

1) 状态价值函数描述了从当前状态 x 出发, 按照既定策略所能累积的所有未来奖励的总和。其参数化模型可以通过公式(4)来表示:

$$V^{\pi}(x) = E_{\pi} \left[\sum_{t=0}^{\infty} \gamma^t r_t | x_0 = x \right]. \quad (4)$$

2) 动作价值函数是指在当前状态 x 下, 根据既定策略采取行动 y 所能获得的所有未来奖励的总和。该函数的参数化模型可以通过公式(5)来表示:

$$Q^{\pi}(x, y) = E_{\pi} \left[\sum_{t=0}^{\infty} \gamma^t r_t | x_0 = x, y_0 = y \right]. \quad (5)$$

3) 状态值是所有可能动作值的加权平均值, 其中权重与策略概率相对应, 具体公式如(6)所示:

$$V^{\pi}(x) = \sum_y \pi(y|x) Q^{\pi}(x, y). \quad (6)$$

1.3.4 优势函数

优势函数通过公式(7)衡量在特定状态下采取某行动相对于策略的平均性能:

$$A^{\pi}(x, y) = Q^{\pi}(x, y) - V^{\pi}(x). \quad (7)$$

1.3.5 目标函数

PolOpt 的目标在于最大化预期累积奖励。在文本生成任务中, 当依据提示 x 生成完整序列 y 时, 此目标可具体表述为最大化整个输出序列的预期奖励, 正如公式(8)所示:

$$\max_{\theta} \{ E_{y \sim \pi_{\theta}(y|x)} \left[\sum_{t=0}^T \gamma^t R(y_t, x_t) \right] \}. \quad (8)$$

1.3.6 梯度更新机制

梯度更新是机器学习和深度学习优化模型参数的核心过程。通过调整模型参数, 该方法旨在使模型的输出更加接近预期的目标。

1) **梯度定义**: 梯度是一个向量, 表示函数在某一点的最大上升方向及其速率。对于模型参数 θ 和目标函数 $J(\theta)$, 梯度可表示为公式(9)所示:

$$\nabla_{\theta} J(\theta) = \left(\frac{\partial J}{\partial \theta_1}, \frac{\partial J}{\partial \theta_2}, \dots \right). \quad (9)$$

正梯度表明, 增加参数 θ 能够提升目标函数的值; 而负梯度则意味着, 减少参数 θ 有助于提高目标函数的值。

2) **梯度下降**: 如公式(10)所示, 通过沿梯度反方向更新参数, 逐步逼近最优解:

$$\theta_{new} = \theta_{old} - \alpha \nabla_{\theta} J(\theta), \quad (10)$$

其中 α 是学习率, 用于控制更新幅度。

3) **梯度更新在强化学习中的作用**: 其核心目标在于通过调整策略 π_{θ} , 以使系统能够获得更高的奖励。策略梯度定理直接优化策略参数 θ , 其梯度公式如公式(11)所示:

$$\nabla_{\theta} J(\theta) = E_{y \sim \pi_{\theta}} \left[\nabla_{\theta} \log \pi_{\theta}(y|x) \cdot A(y) \right], \quad (11)$$

其中, $\log \pi_{\theta}(y|x)$ 表示在给定 x 条件下 y 的响应概率的对数, $A(y)$ 为优势函数, 用于衡量 y 相较于平均性能的优劣。

2 PolOpt 分类

近年来,PolOpt算法迅猛发展,涌现出多种变体以应对推理模型中的各类挑战。为避免简单罗列,本节将对PolOpt的发展历程进行重新梳理,呈现为优化范式的演进过程:从稳定性约束的基础出发,逐步推进结构优化的改进,最终实现范式革新的重大突破。

2.1 分类法

本文提出了如图1所示的四层概念分类体系。该体系依据优化范式对PolOpt进行系统组织,并将其细分为四个相互独立的层级:

(1) **A级**:定义了总体PolOpt。

(2) **B级**:定义PolOpt算法体系涵盖三大核心范式:基础PolOpt(B1)稳定约束型,作为基础方法;GRPO变体(B2)结构优化型,实现结构层面的深度优化;扩展PolOpt(B3)范式突破型,包含该领域最具革命性的前沿成果。

注:为避免GRPO变体(B2)与扩展PolOpt(B3)之间的分类边界模糊,本文采用“是否突破GRPO核心框架”作为划分标准:若算法依然依赖于GRPO的核心目标函数及奖励建模逻辑,则将其归入GRPO变体(B2)类别;若在优化范式、结构约束或学习信号方面对GRPO进行了重构或替代,则将其划分为扩展PolOpt(B3)类别。

(3) **C级**:明确PolOpt算法的核心改进维度,如采样与模块化增强(C1)、奖励设计与调整(C2)、优

势与方差控制(C3)、效率与稳定性优化(C4),这些维度精准刻画了B2的内部演进特征;动态与方差驱动型策略(C5)、泛化与跨域策略学习(C6)以及统一与生成式策略框架(C7)等,突破GRPO框架,在B3框架下实现。

(4) **D级**:详尽阐述具体的算法实例。

这种分层设计能够有效避免概念的混淆,并清晰地呈现复杂算法的演进路径。图1清晰展示了这一结构化框架。

本节将详细阐述第二层中基于PolOpt算法的三大核心范式。关于基础PolOpt(B1)的详尽信息,将在第2.2节中进行介绍;GRPO变体(B2)的详细内容,将在第2.3节中展开探讨;而扩展PolOpt(B3)的详细信息,则将在第2.4节中进行深入阐述。此外,第2.5节将详细对比GRPO变体(B2)与扩展PolOpt(B3)之间的差异。

2.2 基础PolOpt(B1)

基础PolOpt(B1)奠定了策略优化的稳定性约束基础。其核心原理在于通过利用KL散度惩罚或裁剪机制,在最大化奖励的同时,确保分布的稳定性,从而实现两者之间的平衡。PPO和GRPO等算法通过显式控制更新幅度,构建了这一关键层,有效防止奖励模型崩溃,为后续所有变体提供了坚实的理论与实践基准。

基础PolOpt(B1)主要由PPO和GRPO这两个核心组件构成(Schulman等,2017;Shao Zhihong等,2024)。其整体架构详见表1,具体算法的详细总结请参阅第3节。

表1 基础PolOpt(B1)的详细信息

Table 1 The Details of the Basic PolOpt (B1): The table includes the short description, representative, author, publication year, and number of references for the PPO and GRPO.

算法	介绍	代表模型	作者
PPO	通过约束策略更新,平衡样本效率、稳定性和性能。	GPT-5	Schulman, J., et al.
GRPO	无需额外引入价值函数,将KL散度惩罚项作为独立项。	DeepSeek-R1	Shao, Zhihong, et al.

2.3 GRPO变体(B2)的分类

GRPO变体(B2)在基础PolOpt(B1)的稳定性约束框架内进行了创新性拓展,通过四个正交结构维度:采样与模块化增强(C1)、奖励设计与调整(C2)、优势与方差控制(C3)以及效率与稳定性优化(C4),来增强其功能。这一改进旨在提升样本效率、增强

奖励表达力并优化过程平滑性,同时保留GRPO的理论核心。因此,GRPO变体(B2)在该框架内实现了创新性突破,通过优化内部机制,在维持原有概念框架的基础上完成了技术升级。

GRPO变体(B2)可分为四大核心类别:采样与模块化增强(C1),包含基于MDP和SDE的采样

(D1)以及模块化与动态采样(D2);奖励设计与调整(C2),涵盖多维度奖励设计(D3)、动态奖励与序列优化(D4)、多目标与自适应奖励(D5)以及长度自适应与序列控制(D6);优势与方差控制(C3),包含优势多样性与稳定性(D7)、噪声感知与动态优化(D8)以及改进优势估计(D9);效率与稳定性优化(C4),涵盖自适应训练与采样策略(D10)、反馈与熵控制(D11)以及归一化与梯度控制(D12)。

该类别涵盖多种受GRPO启发的算法,这些算法在大规模推理模型中展现出显著的改进效果。具体算法参阅第4节。

2.4 扩展PolOpt(B3)的分类

扩展PolOpt(B3)算法标志着一次突破性的范式转变,超越了传统的GRPO框架。该算法不再局限于奖励微优化或采样技术,而是融合了动态与方差驱动型策略(C5)、泛化与跨域策略学习(C6)以及统一与生成式策略框架(C7)等新型优化范式。这些创新方法旨在突破单一领域稳定性的限制,提升策略优化的泛化能力,有效应对长期规划、多模态处理及不确定性等复杂挑战。

扩展PolOpt(B3)被划分为六大核心类别:动态与方差驱动型策略(C5)包含动态奖励建模(D13)和判别式表达型策略(D14);泛化与跨域策略学习(C6)涵盖多智能体与跨域策略(D15)、多模态与自反学习(D16)以及代理与时间优化(D17);统一与生成式策略框架(C7)则包含集成预训练强化学习优化(D18)和分段式扩散策略(D19)。

相较于B2,B3引入了结构创新,拓展了PolOpt的应用范围,为下一代推理模型奠定了理论基础。代表性算法包括DPPO、MAGPO、HBPO和COPO,具体算法详见第5节。

2.5 B2和B3的差异

尽管GRPO变体(B2)和扩展PolOpt(B3)均标志着政策优化的显著进展,但它们在创新焦点、与GRPO的结构性关联,以及所应对问题范畴方面,存在根本性的差异。

两者的核心差异在于对基础GRPO框架的改进方式。B2变体以框架内优化为显著特征,即在现有GRPO架构中进行针对性的改进。这类变体在保留GRPO核心运作流程的基础上,通过优化采样策略、奖励塑形和优势评估等特定模块,显著提升性能、稳定性和效率。其改进过程呈现出迭代优化的特性,

重点聚焦于采样与模块化增强(C1)、奖励设计与调整(C2)、优势与方差控制(C3)和效率与稳定性优化(C4)等已明确的维度。

相比之下,B3方法最为显著的特征在于其在范式层面上的创新突破,这些突破往往超越了传统GRPO框架,甚至对其进行了彻底的重构。它们并非仅仅是简单的改良方案,而是构建了应对更广泛挑战的全新框架或整合体系。该类别重点关注动态与方差驱动型策略(C5)、泛化与跨域策略学习(C6)、统一与生成式策略框架(C7)(C7)等新兴前沿领域。B3方法的重心已从优化现有组件转向拓展政策优化的范式本身,致力于攻克泛化能力、多模态处理、长期规划等传统GRPO难以应对的复杂难题。

3 基础PolOpt(B1)

基础PolOpt(B1)作为PolOpt的核心算法,涵盖了第4.1节详细阐述的PPO(Proximal Policy Optimization)和第4.2节详细阐述的GRPO(Group Relative Policy Optimization)。这些算法为PolOpt的理论基础奠定了坚实的基石,并在语言模型的训练及推理系统中得到了广泛应用。通过精细的策略优化,在确保系统稳定性的前提下,最大化奖励信号,从而显著提升了模型的计算效率和泛化能力。

3.1 PPO

PPO(近端策略优化)是一种广泛应用且特别适用于大型语言模型的强化学习算法。该算法通过限制策略更新的幅度,有效避免了突变现象,从而在样本效率、稳定性和性能之间实现了均衡(Schulman等,2017)。本节将详细探讨PPO的核心概念,具体步骤可参考附录3中的算法1。

PPO的核心理念在于其独特的双重约束机制:

1) KL惩罚:通过限制对新策略的偏离,可有效避免因过度优化奖励模型而引发的语义坍塌现象,具体如公式(12)所示:

$$KL(\pi_{\theta} \parallel \pi_{ref}) = E_{x \sim Q, y \sim \pi_{\theta}} [\log \frac{\pi_{\theta}(y|x)}{\pi_{ref}(y|x)}]. \quad (12)$$

2) 剪切机制:该机制旨在防止可能对训练稳定性造成破坏的大规模策略更新。通过强制概率

$$r_t(\theta) = \frac{\pi_{\theta}(y_t|x, y_{<t})}{\pi_{old}(y_t|x, y_{<t})}$$

限制在 $[1 - \epsilon, 1 + \epsilon]$ 区间内,从而有效抑制单步更

新的幅度:

$$\text{clip}(r_i(\theta), 1 - \epsilon, 1 + \epsilon), \quad (13)$$

其中 ϵ 是控制裁剪范围的超参数。

PPO算法通过交替进行当前策略的采样输出、奖励估计及策略更新,在确保与先前策略相似性的基础上,最大化奖励收益。详细步骤请参阅附录3中的算法1。

3.2 GRPO

GRPO由DeepSeek开发,作为PPO的改进和扩展,已成为主流的PoLOpt方法,并在推理模型的训练中得到广泛应用(Shao Zhihong等,2024)。本节将探讨GRPO的关键概念,并详细阐述附录3中算法2的完整步骤。

GRPO的核心理念在于摒弃了对附加价值函数的依赖,而PPO则必须依赖这一功能。此外,GRPO还将KL散度惩罚项分解为独立的计算单元。

1) **群体相对奖励**:GRPO采用相对奖励作为同一问题多个采样输出的比较指标,具体如公式(14)所示,而不是以价值函数作为评价基准:

$$\hat{A}_{i,t} = \frac{R(y_i, x) - \bar{R}}{\text{std}(R)}. \quad (14)$$

2) **KL散度作为独立项**:KL散度定义为当前策略与参考策略之间的差异,具体如公式(15)所示,其目的在于防止过度偏离:

$$J(\theta) = L(\theta) - \beta \cdot \text{KL}(\pi_\theta \| \pi_{\text{ref}}). \quad (15)$$

GRPO通过其基于群体的相对奖励机制,有效解决了大型语言模型微调中的两大关键挑战。用简化的组内奖励归一化替代了PPO中的核心价值函

数,从而无需训练独立且计算密集的价值网络,降低了模型复杂度和内存占用。针对复杂推理任务中奖励信号稀疏且存在噪声的问题,GRPO通过对比分析同一提示下的多个响应,自动生成序列中每个标记的有意义相对优势信号,从而大幅提升训练稳定性和学习效率。GRPO的具体实现步骤详见附录3中算法2。

4 GRPO变体(B2)

GRPO变体(B2)是对基础GRPO框架的优化与扩展。尽管这些变体仅在核心结构上进行了细微调整,但它们仍然保留了大部分的操作流程,展现了多维度的创新方法,包括在第4.1节中详细阐述的采样与模块化增强(C1)、第4.2节中详细阐述的奖励设计与调整(C2)、第4.3节中详细阐述的优势与方差控制(C3),以及在第4.4节中详细阐述的效率和稳定性优化(C4)。

4.1 采样与模块化增强(C1)

采样与模块化增强(C1)涵盖两个核心类别:一是基于MDP和SDE的采样(D1),具体内容详见第4.1.1节;二是模块化与动态采样(D2),详细阐述请参阅第4.1.2节。

4.1.1 基于MDP和SDE的采样(D1)

本节将介绍DanceGRPO算法和MixGRPO算法(Xue Zeyue等,2025;Li Junzhe等,2025),其大致框架见表2。

表2 D1算法的详细信息

Table 2 The Details of D1 Algorithms: The table includes the short description, author and publication year for the DanceGRPO and MixGRPO.

算法	介绍	作者	年份
DanceGRPO	采用马尔可夫决策过程和随机微分方程采样建模去噪流程。	Xue, Zeyue, et al.	2025
MixGRPO	重点聚焦于采样策略和优化范围。	Li, Junzhe, et al.	2025

如表2所示D1类算法,以下将对DanceGRPO算法和MixGRPO算法的核心思想与创新价值进行详细阐述:

1) **DanceGRPO**: Dance Group Relative Policy Optimization (DanceGRPO)专为视觉生成任务而设计,涵盖扩散模型和校正流等多个领域。该方法通

过将生成模型的去噪过程建模为马尔可夫决策过程(MDP),并结合随机微分方程(SDE)采样技术,有效解决了传统强化学习方法在视觉生成任务中面临的稳定性难题。具体实现细节请参阅附录3中的算法3。

2) **MixGRPO**: MixGRPO(Flow-based GRPO with

Mixed ODE-SDE)是一种高效的训练框架,通过采用混合 ODE-SDE 采样技术,显著提升了流匹配模型在人类偏好对齐任务中的性能。该框架通过整合确定性常微分方程(ODE)与随机微分方程(SDE)的采样策略,在滑动窗口的关键去噪步骤中引入随机性,并结合强化学习进行优化,从而大幅降低了计算成本,

同时显著加快了模型的收敛速度。具体实现细节请参阅附录 3 中的算法 4。

4.1.2 模块化与动态采样(D2)

本节将介绍 MMGRPO 算法、Scaling Up RL 算法和 GFPO 算法(Ziems 等, 2025; Liu Mingjie 等, 2025; Shrivastava 等, 2025),其大致框架见表 3。

表 3 D2 算法的详细信息

Table 3 The Details of D2 Algorithms: The table includes the short description, author and publication year for the MMGRPO, Scaling Up RL and GFPO.

算法	介绍	作者	年份
MMGRPO	主要在于模块级分组和各模块的独立优化。	Ziems, Noah, et al.	2025
Scaling Up RL	采用解耦裁剪、动态采样、KL 正则化和参考策略重置。	Liu, Mingjie, et al.	2025
GFPO	包含组过滤策略、令牌效率优化及适应性难度提升。	Shrivastava, et al.	2025

如表 3 所示 D2 类算法,以下将对 MMGRPO 算法、Scaling Up RL 算法和 GFPO 算法的核心思想与创新价值进行详细阐述:

1) **MMGRPO**: MMGRPO (Multi-Module Group Relative Policy Optimization) 作为 GRPO 的扩展方案,专为多模块语言模型编程而设计。该方案通过在模块层面引入 GRPO 的分组机制,有效解决了多模块语言模型程序中轨迹长度不一致和结构不协调的难题。MMGRPO 的核心理念在于实现相同模块调用的跨轨迹对齐,进而构建模块级 GRPO 组。这一创新方法不仅继承了 GRPO 的核心优势,更在模块化优化方面取得了显著突破。具体细节参见附录 3 中的算法 5。

2) **Scaling Up RL**: Scaling Up RL 是一项旨在提升小型语言模型在数学、编程和逻辑谜题等复杂推理任务中表现的研究项目。该计划通过延长强化学习训练时长,并整合多种任务优化技术,以达到预期目标。其核心方法包括采用可验证的奖励信号、增强型 GRPO 算法、KL 正则化以及周期性参考策略重置等技术,有效解决了训练过程中出现的熵坍塌和稳定性问题。具体实现细节请参阅附录 3 中的算法 6。

3) **GFPO**: GFPO (Group Filtered Policy Optimization) 是一种针对大型语言模型在强化学习训练过程中响应长度膨胀问题的有效解决方案。该方法通过在训练阶段增加采样响应的数量,并筛选出符合预期学习特征的响应予以保留,从而在推理阶段生成

更为简洁且高效的输出。具体实现细节请参阅附录 3 中的算法 7。

4.2 奖励设计与调整(C2)

奖励设计与调整(C2)包括四个核心类别:多维度奖励设计(D3),详见第 4.2.1 节;动态奖励与序列优化(D4),详见第 4.2.2 节;多目标与自适应奖励(D5),详见第 4.2.3 节;长度自适应与序列控制(D6),详见第 4.2.4 节。

4.2.1 多维度奖励设计(D3)

本节将介绍 AR-GRPO 算法、DAR-GRPO 算法和 GRPO-CARE 算法(Yuan Shihao 等, 2025; Chen Xiwen 等, 2025; Chen Yi 等, 2025),其大致框架见表 4。

如表 4 所示 D3 类算法,以下将对 AR-GRPO 算法、DAR-GRPO 算法和 GRPO-CARE 算法的核心思想与创新价值进行详细阐述:

(1) **AR-GRPO**: AR-GRPO (Autoregressive Group Relative Policy Optimization) 是一种将强化学习应用于自回归图像生成模型的前沿技术。该方法通过改进 GRPO 算法,并引入多维奖励机制,能够有效优化 AR 模型的生成策略,显著提升图像质量,同时更契合人类的审美偏好。具体实现细节请参阅附录 3 中的算法 8。

(2) **DAR-GRPO**: DAR-GRPO (多样性感知奖励调整 GRPO) 是一种专为大型语言模型在强化学习领域量身定制的增强型后训练方法,有效克服了传统 GRPO 算法中由于标量奖励信号引发的“多样性与质量不匹配”难题。该方法通过将语义多样性信

表 4 D3算法的详细信息

Table 4 The Details of D3 Algorithms: The table includes the short description, author and publication year for the AR-GRPO, DAR-GRPO and GRPO-CARE.

算法	介绍	作者	年份
AR-GRPO	The main improvement of AR-GRPO is the multi-dimensional reward design.	Yuan, Shihao, et al.	2025
DRA-GRPO	The main improvement involves explicitly incorporating semantic diversity information into the reward design.	Chen, Xiwen, et al.	2025
GRPO-CARE	The main improvement of GRPO-CARE is its dual-level reward mechanism.	Chen, Yi, et al.	2025

息显式融入奖励计算,并利用语义多样性指数量化样本冗余度,能够动态调整奖励权重。在资源受限的条件下,该方法既能够激励模型探索多样化的推理路径,又能够保障高质量样本的稳定利用。具体实现细节请参阅附录3中的算法9。

(3) **GRPO-CARE**: GRPO-CARE (Consistency-Aware Reinforcement Learning)旨在同步优化推理过程中的答案准确性与逻辑一致性。其采用双层奖励机制:基础奖励确保答案的准确性,而自适应一致性奖励则用于提升推理过程的逻辑连贯性。该方法无需显式过程监督,即可在复杂任务中有效提升模型

的性能和可解释性。具体实现细节请参阅附录3中的算法10。

4.2.2 动态奖励与序列优化(D4)

本节将介绍 GRTPO 算法、GTPO 算法、GRPO-S 算法、P-GRPO 算法和 S-GRPO 算法 (Simoni 等, 2025; Tan Hongze 等, 2025; Fan Lishui 等, 2025; Dai Muzhi 等, 2025), 其大致框架见表5。

注:鉴于存在两个 GTPOs,为清晰起见,本文将 Group-Relative Trajectory-Based Policy Optimization 称为 GRTPO。

表 5 D4算法的详细信息

Table 5 The Details of D4 Algorithms: The table includes the short description, author and publication year for the GRTPO, GTPO, GRPO-S, P-GRPO and S-GRPO.

算法	介绍	作者	年份
GRTPO	主要改进在梯度更新机制和策略稳定性控制。	Simoni, Marco, et al.	2025
GTPO	采用 token 级奖励机制。	Tan, Hongze, et al.	2025
GRPO-S	采用 sequence(序列)级奖励设计。	Tan, Hongze, et al.	2025
P-GRPO	采用 post-hoc 奖励分配策略并解决 GRPO 固有的梯度消失问题。	Fan, Lishui, et al.	2025
S-GRPO	采用序列组生成和递减奖励策略。	Dai, Muzhi, et al.	2025

如表5所示 D4 类算法,以下将对 GRTPO 算法、GTPO 算法、GRPO-S 算法、P-GRPO 算法和 S-GRPO 算法的核心思想与创新价值进行详细阐述:

1) **GRTPO**: GRTPO (Group-Relative Trajectory-Based Policy Optimization)有效解决了 GRPO 所固有的梯度冲突和策略崩溃问题。该算法能够识别并保护那些共享但存在奖励冲突的“冲突令牌”,并在此基础上融合了基于熵的过滤与正则化技术。这些创

新性改进不仅显著增强了训练过程的稳定性,还大幅提升了模型的综合性能。具体实现细节请参阅附录3中的算法11。

2) **GTPO and GRPO-S**: GTPO (Group Token Policy Optimization) 和 GRPO-S (Sequence-Level Group Relative Policy Optimization) 是两种基于动态熵加权的强化学习算法,旨在解决 GRPO 中信用分配不精确的问题,尤其适用于长链推理任务。GTPO

通过为每个令牌生成动态熵加权奖励,实现了精确的信用分配。相比之下,GRPO-S通过调整序列层面的平均熵值,在性能与计算效率之间取得了平衡。这些方法利用成功推理路径中的模型不确定性,将学习信号聚焦于关键推理步骤,从而显著提升了模型的推理能力。具体实现细节请参阅附录3中的算法12和算法13。

3) **P-GRPO**: P-GRPO (Posterior Group Relative Policy Optimization)是一种创新的强化学习方法,该方法通过融合过程奖励与结果奖励,显著提升了大型语言模型的推理能力。其核心机制在于,仅在任务成功时(即生成的代码通过所有测试用例)才激活过程奖励,这一设计既有效防止了奖励作弊,又确保了模型内部推理与最终输出结果的高度一致性。具

体实现细节请参阅附录3中的算法14。

4) **S-GRPO**: S-GRPO (Serial-Group Decaying-Reward Policy Optimization)旨在解决大型语言模型及其他推理模型在生成思维链(CoT)过程中出现的“过度思考”问题。该方法通过引入序列化生成和衰减奖励机制,有效激励模型在早期阶段产出高质量的推理步骤,并在完成充分推理后及时终止。此策略在确保推理准确性的基础上,显著减少了计算成本。具体实现细节请参阅附录3中的算法15。

4.2.3 多目标与自适应奖励(D5)

本节将介绍GHPO算法和MO-GRPO算法(Liu Ziru 等, 2025; Li Xuying 等, 2025),其大致框架见表6。

表6 D1算法的详细信息

Table 6 The Details of D5 Algorithms: The table includes the short description, author and publication year for the GHPO and MO-GRPO.

算法	介绍	作者	年份
GHPO	改进在于采用动态难度检测和自适应引导。	Liu, Ziru, et al.	2025
MO-GRPO	改进在于其多标签奖励模型。	Li, Xuying, et al.	2025

如表6所示D5类算法,以下将对GHPO算法和MO-GRPO算法的核心思想与创新价值进行详细阐述:

1) **GHPO**: 根据样本奖励的稀疏性, GHPO (Guided Hybrid Policy Optimization)会动态将其划分为“简单”或“困难”类别。对于被判定为困难的样本,系统将引入真实答案 $h_{f,x}$ 作为提示,以模拟学习过程;而对于简单样本,则继续采用标准强化学习的探索机制。具体实现细节请参阅附录3中的算法16。

2) **MO-GRPO**: Mo-GRPO (Multi-objective Opti-

mization Framework based on GRPO)旨在实现语言模型的安全性协调性生成。该方法通过训练多标签奖励模型,预测涵盖安全性、礼貌性、可操作性和语义连贯性等多个维度的评分。这些评分经过加权聚合,形成统一的奖励信号,最终利用GRPO算法进行优化。具体实现细节请参阅附录3中的算法17。

4.2.4 长度自适应与序列控制(D6)

本节将介绍LAPO算法和GSPO算法(Wu Xingyu 等, 2025; Zheng Chujie 等, 2025),其大致框架见表7。

表7 D6算法的详细信息

Table 7 The Details of D6 Algorithms: The table includes the short description, classification, author, publication year, and number of references for the LAPO and GSPO.

算法	介绍	作者	年份
LAPO	推理过程的长度从受外部因素制约转变为模型的内在能力。	Wu, Xingyu, et al.	2025
GSPO	采用基于完整序列似然的重要采样和裁剪机制。	Zheng, Chujie, et al.	2025

如表7所示D6类算法,以下将对LAPO算法和GSPO算法的核心思想与创新价值进行详细阐述:

1) **LAPO**: LAPO (Length-Adaptive Policy Optimization)的核心理念在于强调有效的推理必须建立在

对具体问题计算需求的深刻理解之上,而非机械地遵循固定的规则。LAPO算法采用了一种两阶段模型,这种设计促进了自然的演进过程:首先,模型通过经验积累,学习和掌握适当的推理深度标准;随后,逐步培养出主动预判这些需求的能力。相关核心概念详见附录4中的算法18。

2) **GSPO**: GSPO (Group Sequence Policy Optimization)通过序列似然重要性比率进行序列级剪枝,并将归一化奖励作为查询响应的优势值,确保奖励机制与序列级优化过程的高度一致。核心概念详见附录3的算法19。

4.3 优势与方差控制(C3)

优势与方差控制(C3)包括三个核心类别:优势多样性与稳定性(D7),详见第4.3.1节;噪声感知与动态优化(D8),详见第4.3.2节;以及改进优势估计(D9),详见第4.3.3节。

4.3.1 优势多样性与稳定性(D7)

本节将介绍EDGE-GRPO算法、Hybrid GRPO算法、TreeRPO算法和MEML-GRPO算法(Zhang Xingjian等,2025; Soham, 2025; Yang Zhicheng等, 2025; Jia Weibo等, 2025),其大致框架见表8。

表8 D7算法的详细信息

Table 8 The Details of D7 Algorithms: The table includes the short description, author and publication year for the EDGE-GRPO, Hybrid GRPO, TreeRPO and MEML-GRPO.

算法	介绍	作者	年份
EDGE-GRPO	改进在于引导式纠错和熵驱动优势。	Zhang, Xingjian, et al.	2025
Hybrid GRPO	改进在于优势函数保留了价值函数。	Sane, Scham.	2025
TreeRPO	改进包括树采样机制、数据剪枝策略以及连续奖励优势计算方法。	Yang, Zhicheng, et al.	2025
MEML-GRPO	改进包括估算组内相对优势,并基于该优势进行政策优化。	Jia, Weibo, et al.	2025

如表8所示D7类算法,以下将对EDGE-GRPO算法、Hybrid GRPO算法、TreeRPO算法和MEML-GRPO算法的核心思想与创新价值进行详细阐述:

1) **EDGE-GRPO**: EDGE-GRPO (Entropy-Driven GRPO with Guided Error Correction)是一种旨在解决强化学习中优势崩溃问题的算法。该算法通过引入引导式纠错机制,有效提升了响应的多样性;同时,借助熵驱动优势策略优化优势信号,并通过策略熵的提升,显著增强了模型在数学推理等复杂任务中的表现能力。具体实现细节请参阅附录3中的算法20。

2) **Hybrid GRPO**: Hybrid GRPO (Hybrid Group Relative Policy Optimization)结合了PPO和GRPO的双重优势。该方法通过利用价值函数 $V(s)$ 的稳定性,采用多动作采样策略,在每个状态转移时生成多个动作样本,从而获取更多训练数据。这一创新设计不仅提高了样本利用率,还有效稳定了策略更新过程。Hybrid GRPO的核心在于平衡经验采样与价值函数的引导机制,既降低了方差,又避免了纯经验方法所面临的高样本复杂度问题。具体实现细节请参阅附录3中的算法21。

3) **TreeRPO**: TreeRPO (Tree Relative Policy Optimization)是一种基于树采样的强化学习方法,专为大型语言模型中的推理任务设计,能够提供细粒度的奖励信号。其核心机制通过树结构采样,在不同推理步骤中估算奖励期望值,从而摆脱对显式过程奖励模型的依赖。该方法在GRPO的基础上,利用树采样生成的步骤级分组来计算相对奖励,显著提升了模型在数学推理等复杂任务中的性能和效率。具体实现细节请参阅附录3中的算法22。

4) **MEML-GRPO**: MEML-GRPO (Multi-Expert Mutual Learning GRPO)是专为大型语言模型推理任务设计的强化学习算法的升级版,成功解决了奖励稀疏环境下学习信号不足的难题。MEML-GRPO的核心机制在于组内比较分析:它不依赖于绝对奖励值,而是通过比较同一组提示(即“专家”)在相同问题条件下的多个生成结果,计算相对优势。通过引导模型提高高奖励答案的生成概率并抑制低奖励答案,促使模型朝着正确解的方向发展,有效缓解了奖励稀疏环境所带来的挑战。具体实现细节请参阅附录3中的算法23。

4.3.2 噪声感知与动态优化(D8)

本节将介绍 Flow-GRPO 算法和 TempFlow-

GRPO 算法(Liu Jie 等, 2025; He Xiaoxuan 等, 2025), 其大致框架见表9。

表9 D8算法的详细信息

Table 9 The Details of D8 Algorithms: The table includes the short description, author and publication year for the Flow-GRPO and TempFlow-GRPO.

算法	介绍	作者	年份
Flow-GRPO	将普通微分方程转换为随机微分方程以及实现噪声抑制策略。	Liu, Jie, et al.	2025
TempFlow-GRPO	改进包括轨迹分支和噪声感知加权。	He, Xiaoxuan, et al.	2025

如表9所示D8类算法,以下将对Flow-GRPO算法和TempFlow-GRPO算法的核心思想与创新价值进行详细阐述:

1) **Flow-GRPO**: Flow Group Relative Policy Optimization (Flow-GRPO)是一种创新方法,通过将在线强化学习与流匹配模型相结合,借助普通微分方程到随机微分方程的转换和降噪技术,实现高效训练。具体实现细节请参阅附录3中的算法24。

2) **TempFlow-GRPO**: TempFlow-GRPO (Temporal Flow Group Relative Policy Optimization)是一种针对流匹配模型的时序感知强化学习框架。该框架针

对传统GRPO方法在文生图任务中面临的两大瓶颈(终端奖励稀疏性和均匀优化权重),创新性地引入了轨迹分支与噪声感知加权机制。通过在不同生成阶段采用差异化优化策略,该框架显著提升了生成图像在人类偏好对齐和视觉质量方面的综合表现。具体实现细节请参阅附录3中的算法25。

4.3.3 改进优势估计(D9)

本节将介绍 GAE-GRPO 算法 T-PPO 算法、VAPO 算法和 GPG 算法 (Schulman 等, 2015; Fan Tiantian 等, 2025; Yue Yu 等, 2025; Chu Xiangxiang 等, 2025), 其大致框架见表10。

表10 D9算法的详细信息

Table 10 The Details of D9 Algorithms: The table includes the short description, author and publication year for the GAE-GRPO, T-PPO, VAPO and GPG.

算法	介绍	作者	年份
GRPO-GAE	采用GAE集成以计算其优势。	Schulman, John, et al.	2015
T-PPO	采用截断轨迹生成和扩展GAE的应用。	Fan, Tiantian, et al.	2025
VAPO	采用增强型价值建模和自适应优化机制。	Yue, Yu, et al.	2025
GPG	通过组奖励归一化估计优势函数,并提出梯度估计技术来纠正无效样本。	Chu, Xiangxiang, et al.	2025

如表10所示D1类算法,以下将对GAE-GRPO算法、T-PPO算法、VAPO算法和GPG算法的核心思想与创新价值进行详细阐述:

1) **GAE-GRPO**: 优势函数是策略更新的核心驱动力。通过采用广义优势估计(GAE),GRPO能够动态平衡估计偏差与方差,从而更精确地评估动作的长期价值。核心概念详见附录3的算法26。

2) **T-PPO**: T-PPO (Truncated Proximal Policy Optimization)通过截断生成、EGAE优势估计以及动态批次管理,支持部分轨迹优化。该方法在维持优势计算的同时,有效处理不完整响应,实现了策略更新与完整生成过程的解耦。此外,通过整合蒙特卡

洛价值校正,确保了收敛性,从而在计算效率与解精度之间达到了完美平衡。核心概念详见附录3中的算法27。

3) **VAPO**: VAPO (Value-model-based Augmented Proximal Policy Optimization)是专为长链推理(Long-CoT)量身定制的强化学习框架。该框架通过引入价值模型增强和自适应优化机制,高效攻克了传统PPO在长序列任务中面临的三大难题:价值模型偏差、序列长度不均一性以及奖励稀疏性。其核心概念详见附录3中的算法28。

4) **GPG**: GPG (Group Policy Gradient)是一种精简且高效的强化学习方法,其核心在于直接优化原

始的强化学习目标,无需依赖价值模型、参考模型、KL约束或替代损失函数。GPG通过群组奖励归一化技术来估计优势函数,有效减少了梯度估计的方差。此外,该方法引入了精确梯度估计技术,用于修正由无效样本(即完全正向或负向响应)引起的梯度偏差。同时,采用阈值机制来控制有效样本的比例,确保训练过程的稳定性。关于其核心概念的详细说明,请参阅附录3中的算法29。

4.4 效率与稳定性优化(C4)

效率与稳定性优化(C4),我们可以采用三个核心类别:自适应训练与采样策略(D10),详见第4.4.1节;反馈与熵控制(D11),详见第4.4.2节;以及归一化与梯度控制(D12),详见第4.4.3节。

4.4.1 自适应训练与采样策略(D10)

本节将介绍AGPO算法、CPPO算法、DAPO算法和RePO算法(Li Chen等,2025;Lin Zhihang等,2025;Yu Qiyang等,2025;Li Siheng等,2025),其大致框架见表3。

如表11所示D1类算法,以下将对AGPO算法、CPPO算法、DAPO算法和RePO算法的核心思想与创新价值进行详细阐述:

1) **AGPO**: AGPO(Adaptive Group Policy Optimization)是一种改进型强化学习算法,旨在克服GRPO在训练稳定性和推理效率上的不足。通过引入自适应损失函数,AGPO不仅有效解决了GRPO中训练信号消失以及优势估计值为零或负值所导致的波动问题,而且在推理阶段显著提高了令牌的使用效率。具体细节参见附录3中的算法30。

2) **CPPO**:在复杂推理任务中,GRPO模型因奖励稀疏面临重大挑战:模型需要生成大量补全样本,但仅有少数能提供有效奖励信号。通过采用动态补全剪枝技术,CPPO(Completion Pruning Policy Optimization)对已完成样本的筛选机制,仅保留绝对优势值满足特定阈值 $|A_i| \geq \gamma$ 的样本,有效抑制无效更新,从而加速智能体发现有效策略。具体细节参见附录3中的算法31。

3) **DAPO**: DAPO(Decoupled Clip and Dynamic Sampling Policy Optimization)整合了四大核心技术:动态采样、剪切机制、令牌级策略梯度损失和超长奖励塑形。这些技术有效解决了强化学习中处理长文本时的熵坍塌和效率低下的难题,显著提升了模型的探索能力与训练稳定性。具体细节参见附录3中的算法32。

4) **RePO**: RePO(Replay-Enhanced Policy Optimization)是一种先进的强化学习方法,引入了回缓冲机制,该机制通过将当前策略生成的样本(在策略内)与回放缓冲区获取的历史样本(在策略外)相结合,实现策略优化。该方法显著提升了数据效率和训练多样性,有效避免了GRPO等传统方法因依赖单一当前策略样本而产生的梯度信号消失问题。具体细节参见附录3中的算法33。

4.4.2 反馈与熵控制(D11)

本节将介绍Critique-GRPO算法和SEED-GRPO算法(Zhang Xiaoying等,2025;Chen Minghan等,2025),其大致框架见表12。

表 11 D10算法的详细信息

Table 11 The Details of D10 Algorithms: The table includes the short description, author and publication year for the AGPO, CPPO, DAPO and RePO.

算法	介绍	作者	年份
AGPO	改进在于其自适应损失函数。	Li, Chen, et al.	2025
CPPO	改进在于对完成的样本进行剪枝处理。	Lin, Zhihang, et al.	2025
DAPO	包含四大核心技术:动态采样、剪切机制、令牌级策略梯度损失以及超长奖励塑形。	Yu, Qiyang, et al.	2025
RePO	将当前策略生成的样本与回放缓冲区中获取的历史样本相结合。	Li, Siheng, et al.	2025

如表12所示D11类算法,以下将对Critique-GRPO算法和SEED-GRPO算法的核心思想与创新价值进行详细阐述:

1) **Critique-GRPO**: Critique-GRPO的创新之处

在于其将自然语言反馈与强化学习进行了深度整合,突破了传统GRPO纯数值优化框架的局限性。该方法通过引入自然语言反馈作为辅助奖励信号,并结合优势函数(包含初始响应与修正响应)的混合

表 12 D11 算法的详细信息

Table 12 The Details of D11 Algorithms: The table includes the short description, author and publication year for the Critique-GRPO and SEED-GRPO.

算法	介绍	作者	年份
Critique-GRPO	采用双重反馈机制和在线自我改进。	Zhang, Xiaoying, et al.	2025
SEED-GRPO	引入语义熵和不确定性感知的优势调制。	Chen, Minghan, et al.	2025

计算方式,有效优化了策略梯度估计。具体细节参见附录 3 中的算法 34。

2) **SEED-GRPO**: SEED-GRPO (Semantic Entropy Enhanced GRPO)通过引入语义熵和不确定性感知的优化调节机制得以实现。当模型的语义熵值较高(即表明不确定性较大)时,系统会自适应地降低策略更新过程中的优势权重,采取更为保守的

学习步骤;相反,当模型的语义熵值较低时,则维持原有的优势权重不变。具体细节参见附录 3 中的算法 35。

4.4.3 归一化与梯度控制(D12)

本节将介绍 GMPO 算法、GAPO 算法和 ReMix 算法 (Zhao Yuzhong 等, 2025; Li Chengao 等, 2025; Liang Jing 等, 2025),其大致框架见表 13。

表 13 D12 算法的详细信息

Table 13 The Details of D12 Algorithms: The table includes the short description, author and publication year for the GMPO, GAPO and ReMix.

算法	介绍	作者	年份
GMPO	用几何平均优化奖励替代相对平均优化奖励。	Zhao, Yuzhong, et al.	2025
GAPO	采用动态调整各目标函数的梯度权重,确定最优更新方向。	Li, Chengao, et al.	2025
ReMix	采用基于历史策略数据学习的非策略算法。	Liang, Jing, et al.	2025

如表 13 所示 D12 类算法,以下将对 GMPO 算法、GAPO 算法和 ReMix 算法的核心思想与创新价值进行详细阐述:

1) **GMPO**: GMPO(几何平均策略优化)是 GRPO 的稳定改进版本,采用几何平均优化奖励机制,而非相对平均优化奖励机制。通过优化 token 级奖励的几何平均值,有效减少了异常值对策略更新的干扰,进而提升了训练的稳定性。具体细节参见附录 3 中的算法 36。

2) **GAPO**: GAPO(Gradient Adaptive Policy Optimization)通过将梯度自适应归一化机制整合到策略网络中,有效稳定了训练过程并加速了收敛速度。该方法通过动态调整目标函数的梯度权重,实现了多目标优化,确定了最优的更新方向,从而在不同目标之间达到了最佳平衡,有效解决了 GRPO 算法中梯度冲突和收敛不稳定的问题。具体核心概念详见附录 3 中的算法 37。

3) **ReMix**: Reincarnating Mix-policy Proximal Policy Gradient(ReMix)采用了一种离策略算法,通

过利用历史策略数据(即经验)进行学习,使得近策略梯度方法能够高效地利用离策略数据,并将历史策略数据与混合端策略优化相结合。通过运用历史策略生成的非当前策略数据,显著提升了 GRPO 的采样效率,同时引入动态 KL 约束,确保了策略的稳定性。具体实现细节请参阅附录 3 中的算法 38。

5 Extended PolOpt (B3)

扩展 PolOpt(B3)代表了当前研究的前沿成果。它并非对 GRPO 的简单改进,而是专注于 5.1 节详述的动态与方差驱动型策略(C5)、5.2 节详述的泛化与跨域策略学习(C6)、5.3 节详述的统一与生成式策略框架(C7)、5.4 节详述的训练稳定性(C8)、5.5 节详述的跨域应用(C9)以及 5.6 节详述的不确定性处理(C10)等新兴领域。

5.1 动态与方差驱动型策略(C5)

动态与方差驱动型策略(C5)通过第 5.1.1 节所述的动态奖励建模(D13)和第 5.1.2 节所述的判别

式表达型策略(D14)优化学习信号。

5. 1. 1 动态奖励建模(D13)

本节将介绍 SeqPO-SiMT 算法、GVPO 算法、

GRQO 算法、HBPO 算法和 COPO 算法(Xu Ting 等, 2025; ZhangKalchen 等, 2025; Pan Chenbin 等, 2025; Lyu Shangke 等, 2025), 其大致框架见表 14。

表 14 D13 算法的详细信息

Table 14 The Details of D13 Algorithms: The table includes the short description, author and publication year for the SeqPO-SiMT, GVPO, GRQO, HBPO and COPO.

算法	介绍	作者	年份
SeqPO-SiMT	将奖励标准化与延迟截断技术整合到混合奖励机制中。	Xu, Ting, et al.	2025
GVPO	通过比较隐式奖励与实际奖励之间的中心距离差异来设计梯度权重。	Zhang, Kalchen, et al.	2025
GRQO	通过计算查询级别的相对奖励,激励查询结果超越动态分组基线。	Pan, Chenbin, et al.	2025
HBPO	采用层次化预算探索,动态调整推理时长。	Lyu, Shangke, et al.	2025
COPO	采用基于一致性与熵驱动软混合策略的全局奖励机制。	Han, Jinghang, et al.	2025

如表 14 所示 D1 类算法, 以下将对 SeqPO-SiMT 算法、GVPO 算法、GRQO 算法、HBPO 算法和 COPO 算法的核心思想与创新价值进行详细阐述:

1) **SeqPO-SiMT**: SeqPO-SiMT (Sequential Policy Optimization for Simultaneous Machine Translation)将奖励标准化与延迟截断技术整合至混合奖励机制中,通过动态调节权重,同步评估翻译质量和延迟性能。核心概念详见附录 3 中的算法 39。

2) **GVPO**: GVPO(Group Variance Policy Optimization)通过直接将最大化 KL 约束奖励的解析解融入梯度权重,确保策略更新与最优策略的一致性。通过比较隐式奖励与实际奖励的中心距离差异,设计出梯度权重,既有效避免了传统方法中的训练不稳定性,又支持灵活的采样分布。具体实现细节请参阅附录 3 中的算法 40。

3) **GRQO**: GRQO(Group Relative Query Optimization)是一种专为视觉基础模型设计的强化学习范式,通过引入基于群体的相对奖励机制和 KL 散度正则化,在视觉引导场景中显著提升了查询质量与训练稳定性。其核心思想在于计算查询级别的相对奖励,以激励查询超越动态群体基线。此外, KL 散度正则化对目标概率分布进行有效约束,有效防止训练过程中的漂移现象。具体实现细节请参阅附录 3 中的算法 41。

4) **HBPO**: HBPO (Hierarchical Budget Policy Optimization)是一种基于分层预算探索的强化学习框架,该框架能够使推理模型根据问题的复杂度自适应地调整推理长度,从而在保持高性能的同时显

著提升推理效率。其核心概念详见附录 3 中的算法 42。

5) **COPO**: COPO (Consistency-Aware Policy Optimization)是一种强化学习框架,通过引入基于一致性和熵驱动软混合策略的全局奖励机制,有效解决 GRPO 方法中因样本过度一致所导致的梯度消失和样本浪费问题。同时,它充分利用高度一致样本的信息,从而在数学推理等复杂任务中显著提升模型的性能和训练稳定性。核心概念详见附录 3 中的算法 43。

5. 1. 2 判别式表达型策略(D14)

本节将介绍 EXPO 算法和 DisCO 算法(Dong 等, 2025; Li Gang 等, 2025), 其大致框架见表 2。

如表 15 所示 D14 类算法, 以下将对 EXPO 算法和 DisCO 算法的核心思想与创新价值进行详细阐述:

1) **EXPO**: EXPO (Expressive Policy Optimization)是一种专为扩散策略或流匹配策略等表达式策略设计的在线强化学习算法。该算法通过将策略优化与价值最大化分离,有效解决了传统梯度传播方法中存在的稳定性问题。结合了基于模仿学习的表达式基础策略与轻量级高斯编辑策略,并通过实时策略选择来最大化 Q 值,在确保稳定性的基础上显著提高了样本效率。相关核心概念详见附录 3 中的算法 44。

2) **DisCO**: DisCO (Discriminative Constrained Optimization)通过判别学习直接提升正负样本的区分能力,并在动态约束条件下均衡策略更新的稳定

表 15 D14算法的详细信息

Table 15 The Details of D14 Algorithms: The table includes the short description, author and publication year for the EXPO and DisCO.

算法	介绍	作者	年份
EXPO	将策略优化与价值最大化解耦,并将基于模仿学习的表达性基础策略与轻量级高斯编辑策略相结合。	Dong, Perry, et al.	2025
DisCO	通过判别性学习直接优化正负样本间的区分度,并利用动态约束在稳定性和灵活性之间取得平衡。	Li, Gang, et al.	2025

性和灵活性。该方法通过限制策略更新幅度以防止训练过程中的不稳定性,同时最大化正负样本之间的区分度。具体实现细节请参阅附录 3 中的算法 45。

5.2 泛化与跨域策略学习(C6)

泛化与跨域策略学习(C6)通过采用第 5.2.1 节所述的多智能体与跨域策略(D15)、第 5.2.2 节所述

的多模态与自反学习(D16)和第 5.2.3 节所述的代理与时间优化(D17)来细化策略更新并降低方差。

5.2.1 多智能体与跨域策略(D15)

本节将介绍 SRPO 算法和 MAGPO 算法(Zhang Xiaojiang 等, 2025; Li Yueheng 等, 2025),其大致框架见表 16。

表 16 D15算法的详细信息

Table 16 The Details of D15 Algorithms: The table includes the short description, author and publication year for the SRPO and MAGPO.

算法	介绍	作者	年份
SRPO	采用两阶段训练范式和历史重采样机制。	Zhang, Xiaojiang, et al.	2025
MAGPO	将集中式指导与去中心化执行相结合。	Li, Yueheng, et al.	2025

如表 16 所示 D15 类算法,以下将对 SRPO 算法和 MAGPO 算法的核心思想与创新价值进行详细阐述:

1) **SRPO**: SRPO (Two-Staged History-Resampling Policy Optimization) 采用双阶段训练模式,通过历史重采样机制(一种专门用于处理无效样本的技术)协调数学推理能力与编程技能的同步提升。其核心原理详见附录 3 中的算法 46。

2) **MAGPO**: MAGPO (Multi-Agent Guided Policy

Optimization)融合集中式指导与分布式执行,充分发挥集中式训练的独特优势。借助自我强化的联合策略实现高效的协作探索,确保在整个训练过程中与分布式学习器保持持续同步。其核心概念详见附录 3 中的算法 47。

5.2.2 多模态与自反学习(D16)

本节将介绍 PAPO 算法和 SRPO 算法(Wang Zhenhailong, 2025; Wan Zhongwei 2025),其大致框架见表 17。

表 17 D16算法的详细信息

Table 17 The Details of D16 Algorithms: The table includes the short description, author and publication year for the PAPO and SRPO.

算法	介绍	作者	年份
PAPO	通过最大化原始图像与遮罩图像输出分布的差异,被迫依赖视觉输入进行推理。	Wang, Zhenhailong, et al.	2025
SRPO	学习自我反思,并利用反思结果确保整个过程完全自主。	Wan, Zhongwei, et al.	2025

如表 17 所示 D16 类算法,以下将对 PAPO 算法和 SRPO 算法的核心思想与创新价值进行详细阐述:

PAPO: PAPO (Perception-Aware Policy Optimization)通过最大化原始图像 I 与掩膜图像 I_{mask} 之间的输出分布差异,迫使模型在推理过程中更多地依赖

视觉输入。其核心概念详见于附录3中的算法48。

SRPO: Multimodal Self-Reflection Enhanced Reasoning with Group Relative Policy Optimization (SRPO)是一个专为促进模型学习机器实现显式自我反思与修正行为而设计的两阶段反思训练框架。第一阶段采用反思引导的监督微调,旨在减少冗余

推理并修正错误推断;第二阶段则通过具备反思意识的强化学习,主要依托显式奖励模型实现优化。其核心概念详见附录3中的算法49。

5.2.3 代理与时间优化(D17)

本节将介绍 ARPO 算法和 TSPO 算法(Dong 等, 2025;Tang Canhui 等, 2025),其大致框架见表18。

表 18 D17 算法的详细信息

Table 18 The Details of D17 Algorithms: The table includes the short description, author and publication year for the ARPO and TSPO.

算法	介绍	作者	年份
ARPO	引入基于熵的自适应展开机制。	Dong, G, et al.	2025
TSPO	采用事件感知的时间代理模型和双奖励机制。	Tang, Canhui, et al.	2025

如表18所示D17类算法,以下将对ARPO算法和TSPO算法的核心思想与创新价值进行详细阐述:

1) **ARPO**: ARPO (Agentic Reinforced Policy Optimization)是一种专门为训练多轮次大型智能体设计的强化学习算法。该算法通过熵自适应展开机制,动态平衡全局轨迹采样与步骤级采样,进而在工具使用后的高不确定性步骤中有效促进探索。其核心概念详见附录3中的算法50。

2) **TSPO**: TSPO (Temporal Sampling Policy Optimization)是一种基于强化学习框架的先进方法,专门用于优化视频关键帧的采样过程。该框架针对长视频多模态大语言模型中无监督且非微观级的稀疏帧采样难题,通过将关键帧选择与语言生成建模为

一个联合决策过程,采用事件感知的时间代理机制,并结合包含响应准确度与时间定位奖励的双奖励机制,实现了端到端的策略优化。其核心概念详见附录3中的算法51。

5.3 统一与生成式策略框架(C7)

统一与生成式策略框架(C7)通过采用第5.3.1节所述的集成预训练强化学习优化(D18)和第5.3.2节所述的分段式扩散策略(D19),提高数据效率和培训质量。

5.3.1 集成预训练强化学习优化(D18)

本节将介绍 MiMo-7B 算法(Xia Bingquan 等, 2025),其大致框架见表19。

表 19 D18 算法的详细信息

Table 19 The Details of D18 Algorithms: The table includes the short description, author and publication year for MiMo-7B.

算法	介绍	作者	年份
MiMo-7B	预训练阶段,采用三阶段数据融合策略、多标记预测目标以及高质量推理数据增强技术;后训练阶段,强化学习与测试难度驱动的奖励机制及动态采样策略相结合。	Xia, Bingquan, et al.	2025

如表19所示D18类算法,以下将对MiMo-7B算法的核心思想与创新价值进行详细阐述:

1) **MiMo-7B**: MiMo-7B是一款专为应对复杂推理任务而设计的大型语言模型,其核心创新在于对从预训练到强化学习后训练的全流程进行了深度优化。在预训练阶段,该模型采用了三阶段数据融合策略、多标记预测目标以及高质量推理数据增强技术,显著提升了底层模型的推理能力。在后训练阶

段,通过结合测试难度驱动的奖励机制、动态采样策略以及无缝滚动引擎等高效基础设施,强化学习进一步挖掘了模型在数学推导、代码生成和逻辑推理方面的潜力。其核心设计理念详见附录3中的算法52。

5.3.2 分段式扩散策略(D19)

本节将介绍 SPO 算法和 DPPO 算法(Guo Yiran 等,2025;Ren 等,2024),其大致框架见表20。

表 20 D19 算法的详细信息

Table 20 The Details of D19 Algorithms: The table includes the short description, author and publication year for the SPO and DPPO.

算法	介绍	作者	年份
SPO	通过评估中粒度分割的收益,以平衡细粒度与粗粒度方法的优劣。	Guo, Yiran, et al.	2025
DPPO	强化学习策略被建模为一种“扩散-噪声”过程,通过专家数据与环境反馈实现持续优化。	Ren, Allen Z., et al.	2024

如表 20 所示 D19 类算法,以下将对 SPO 算法和 DPPO 算法的核心思想与创新价值进行详细阐述:

1) **SPO**: SPO (Segment Policy Optimization) 是一种基于段落级优势估计的强化学习框架,旨在解决大型语言模型推理任务中的信用分配难题。通过中间粒度的段落优势评估,均衡细粒度与粗粒度方法的优劣。具体概念详见附录 3 中的算法 53。

2) **DPPO**: 通过模仿学习并利用专家数据,GRPO 能够迅速收敛至高质量策略。Diffusion Policy Policy Optimization (DPPO) 采用策略梯度对预训练的扩散策略进行精细微调,将去噪扩散过程分解为具有明确动作概率的马尔可夫决策过程。为此,构建了双层神经网络架构:外层对应环境神经网络,内层则对应去噪神经网络。核心概念详见附录 3 中的算法 54。

6 性能与效率对比

为了更好地指导研究人员依据实际应用场景选择适宜的 PolOpt 算法,本章内容主要基于各原始论文的实验结果与理论分析,并结合典型应用场景及资源约束条件,提供经验性的选择建议。将从性能表现、计算开销、训练稳定性及应用场景适应性四个维度,对本文涵盖的主要 PolOpt 算法展开横向对比分析,具体见表 21 所示:

1) 性能表现:指算法在特定任务(如数学推理、代码生成、多模态理解)上的最终效果,通常通过准确率、奖励得分、人类偏好评分等指标衡量。

2) 计算开销:包括显存占用、训练时间、推理延迟等,反映算法对硬件资源的需求。

3) 训练稳定性:指训练过程中是否容易发生崩溃、模式坍塌、奖励过拟合等问题,以及超参数敏感性。

4) 适用场景:根据算法特点,归纳其最适合解决的问题类型或资源条件。

由表 21 可得 PolOpt 算法的场景化选型建议:

1) 训练资源受限场景下,首选算法包括 GFPO 和 RePO。这是因为 GFPO 能够通过响应过滤减少冗余计算,而 RePO 则通过经验回放机制提升数据利用效率,二者均在有限资源条件下保持了较好的性能表现。

2) 针对强推理任务(如数学、代码等领域),推荐采用 GRPO、COPO 或 TreeRPO。这类算法在设计上侧重于推理能力增强:GRPO 专为推理任务优化,COPO 提升逻辑一致性,TreeRPO 则提供细粒度的奖励机制,因而在复杂推理场景中表现突出。

3) 在多模态与生成任务中,DPPO、AR-GRPO 和 EXPO 是首选算法。DPPO 适用于扩散生成模型,AR-GRP 针对自回归图像生成进行优化,而 EXPO 支持表达性策略学习,因此三者均适合视觉与生成类任务的需求。

4) 在安全与对齐任务方面,建议使用 GAPO 和 MO-GRPO。这两种算法均支持多目标优化,能够平衡性能、安全性和伦理等多个维度,适用于对系统行为有严格对齐要求的场景。

5) 对于实时或低延迟场景,HBPO 和 LAPO 是理想选择。它们均具备自适应推理长度控制能力,在保持准确性的同时有效减少计算时间,因此适合部署在实时性要求较高的系统中。

7 结语与展望

本节将重点概括本文的核心内容,并深入阐述偏好优化(Preference Optimization, PO)的概念,同时探讨 PolOpt 的未来发展趋势。

[结语]本文对多种应用于语言模型和推理模型的 PolOpt 算法进行了全面综述,重点探讨了其技术演进与核心原理。研究将这些算法系统化地组织到一个包含四个层级的分类框架中:三大基础类别(B1-B3)、七个中间分类(C1-C7)和十九个细分类别

表 21 代表性 PolOpt 算法对比分析

Table 21 Comparative Analysis of Representative PolOpt Algorithms: The table summarizes the performance, computational cost, training stability, and applicable scenarios for each algorithm category and variant.

算法类别	代表算法	性能表现	计算开销	训练稳定性	适用场景
基础 PolOpt (B1)	PPO	稳定可靠,泛化性强	高(需价值函数)	高(有KL约束与裁剪)	通用对齐任务、资源充足场景
	GRPO	推理任务表现优异	中(无需价值函数)	中高(组归一化稳定训练)	数学推理、代码生成、中等资源
	GFPO	生成结果简洁,推理效率高	中低(过滤机制减少计算)	中(过滤可能减少探索)	需要简洁输出的推理任务、资源受限
	GAPO	多目标平衡性好	中(梯度归一化增加开销)	高(动态权重缓解冲突)	安全对齐、多目标优化
	DAPO	长文本任务表现好	中高(动态采样与裁剪)	中(超参数敏感)	长链推理、对话生成
扩展 PolOpt (B3)	RePO	数据利用效率高	中(经验回放增加内存)	中(可能策略漂移)	样本稀缺、需重复利用历史数据的场景
	COPO	高一致性任务表现优	中高(全局奖励计算)	高(防梯度消失)	数学证明、逻辑一致性要求高的任务
	MAGPO	多智能体协同效果好	高(集中式训练开销大)	中(需智能体间同步)	多模型协作、分布式推理系统
	DPPO	生成质量高,适合扩散模型	高(扩散过程计算密集)	中(依赖预训练扩散模型)	图像生成、多模态创作
	HBPO	自适应推理长度,效率高	中(分层控制增加复杂度)	中(预算机制需调参)	实时推理、资源动态分配场景

(D1-D19)。该分类体系完整呈现了 PolOpt 技术的发展脉络,涵盖从基础 PolOpt(B1)如 PPO 和 GRPO,到 GRPO 变体(B2),再到扩展 PolOpt(B3)的完整谱系。研究亮点包括对采样优化、奖励机制设计、优势评估及训练稳定性等算法改进的深度剖析,以及多模态与跨领域应用的整合实践。该框架为科

研人员理解并推进语言模型与智能系统中的 PolOpt 技术提供了清晰的路线图。

[偏好优化(PO)]近年来,一个与策略优化(PolOpt)紧密相关但又独具特色的领域——偏好优化(PO)正迅速崛起。PO 可以视为 PolOpt 的专门化演变,主要应用于需要将模型与人类价值观对齐的场景。其核心区别在于依赖人类对行为或结果的相对偏好判断(例如认为选项 A 优于选项 B),而非依赖显式的奖励函数,从而解决复杂、抽象或难以定义的奖励任务。鉴于 PO 在方法论、假设条件和应用场景上与 PolOpt 存在根本性差异,加之该领域已涵盖直接偏好优化(DPO)和比较偏好学习(CPL)等颠覆

性算法,本文因篇幅和深度限制,将不展开详细讨论。我们计划在后续论文中专门撰写,对 PO 进行全面综述与理论分析。

[展望]PolOpt 的未来发展预计将遵循不同的路径。作为基础方法,基础 PolOpt(B1)已达到相对成熟的阶段,其核心框架趋于稳定。其后续发展将更加注重计算效率和可扩展性。相比之下,GRPO 变体(B2)目前已经历多次优化,其创新潜力似乎已接近饱和。实现重大突破可能需要通过细微调整或与互补技术的整合。值得注意的是,那些缺乏深入理论分析、明确动机或基础算法洞见的修改,仅依赖孤立技术改进的做法,往往被视为随意之举,难以通过严格的学术评审。从政策梯度方法的角度来看,GRPO 变体(B2)要求任何改动都必须在梯度层面进行充分论证。事实上,某些所谓的改进从梯度视角来看根本无法解释。相比之下,扩展 PolOpt(B3)有望成为学术界的研究焦点,尤其是在多模态感知、跨域推理、智能体机制以及不确定性处理等关键领域。

例如,将扩散模型或人类反馈强化学习技术进行整合应用,极有可能在这些领域取得突破性进展。

本文提出的分类体系为大型语言模型及推理模型中的PolOpt算法构建了一个全面的分类框架。其层级结构(B1-B3、C1-C7、D1-D19)不仅有效整合了现有技术,还为未来的发展预留了充足的扩展空间。当前PolOpt研究仍存在以下关键瓶颈:

1) 奖励稀疏性与噪声放大问题:当前强化学习训练主要依赖于离散或稀疏的奖励信号,尤其在复杂推理和代码生成等任务中,奖励信号通常仅在最终输出阶段才出现,这会导致梯度传播不稳定,进而降低训练效率;另外,基于相对奖励的机制在高方差环境下容易引发噪声放大效应,从而干扰模型的学习方向。

2) 多目标权重冲突:在多维优化目标共存的情境下,不同奖励维度之间呈现出显著的冲突现象。现有算法普遍采用静态加权或线性加权的方法,这种方式难以灵活应对动态任务的需求,进而导致策略在某些目标上出现过度拟合,而在其他目标上的性能则显著下降。

3) 计算资源与扩展性:当前,PolOpt在千亿参数模型的适配过程中,主要受限于显存占用、通信延迟以及多GPU并行的效率问题,特别是在奖励模型与主模型耦合训练阶段,资源消耗尤为显著。

以下列出了未来可能的研究方向:

1) 针对奖励稀疏与噪声放大问题,未来研究可从奖励密集化与鲁棒化方向入手,通过引入中间过程奖励或自监督信号提高训练阶段的梯度可传播性,同时结合贝叶斯噪声建模与对比学习等方法降低奖励方差的影响;此外,可探索多模态奖励融合机制,将文本、图像及结构化反馈共同纳入强化学习信号体系,从而在复杂推理场景中提升训练稳定性与收敛速度。

2) 针对多目标权重冲突问题,后续研究可引入自适应权重标定与Pareto前沿优化机制,使策略优化能够动态平衡不同目标间的竞争关系;同时,建立层次化的多目标结构,在全局收敛性约束下实现安全性、逻辑性与创造性等子目标的协调优化,以提升模型在多任务场景下的综合性能与稳健性。

3) 针对计算资源与扩展性问题,未来可发展轻量级参数高效微调与分布式PolOpt训练架构;同时研究奖励模型蒸馏与共享机制,实现主模型与奖励

模型的高效协同训练,从而显著提升超大规模语言模型的训练效率与工程可扩展性。

致谢:Acknowledgements

参考文献(References)

- LMSys. Chatbot Arena Leaderboard[EB/OL].
<https://huggingface.co/spaces/lmsys/chatbot-arena-leaderboard>, 2025.
- Zhao D, et al. Reinforcement Learning [M]. Beijing: China Machine Press, 2018.
- Pan W, Ouyang J, Liu Y, Zhao D, Su T. Fundamentals of Pattern Recognition and Machine Learning [M]. Beijing: China Machine Press, 2023.
- Schulman J, Wolski F, Dhariwal P, et al. Proximal Policy Optimization Algorithms[EB/OL]. arXiv:1707.06347, 2017.
- Shao Z H, Li Y D, Zhang J, et al. DeepSeekMath: Pushing the Limits of Mathematical Reasoning in Open Language Models [EB/OL]. arXiv:2402.03300, 2024.
- Xue Z Y, Wang L, Chen H, et al. DanceGRPO: Unleashing GRPO on Visual Generation[EB/OL]. arXiv:2505.07818, 2025.
- Li J Z, Zhou M, Wang H, et al. MixGRPO: Unlocking Flow-based GRPO Efficiency with Mixed ODE-SDE [EB/OL]. arXiv: 2507.21802, 2025.
- Ziems N, Zhang T, Liu S, et al. Multi-module GRPO: Composing Policy Gradients and Prompt Optimization for Language Model Programs[EB/OL]. arXiv:2508.04660, 2025.
- Liu M J, Chen X, Wang Z, et al. Scaling Up RL: Unlocking Diverse Reasoning via Prolonged Training [EB/OL]. arXiv: 2507.12507, 2025.
- Shrivastava V, Das S, Raj A, et al. Sample More to Think Less: Group Filtered Policy Optimization[EB/OL]. arXiv:2508.09726, 2025.
- Yuan S H, Wu P, Lin H, et al. AR-GRPO: Training Autoregressive Image Generation via RL[EB/OL]. arXiv:2508.06924, 2025.
- Chen X W, Li Q, Zhang Y, et al. DAR-GRPO: Diversity-Aware Reward Adjustment for LLM RL Training [EB/OL]. arXiv: 2505.09655, 2025.
- Chen Y, He L, Zhou W, et al. GRPO-CARE: Consistency-Aware RL for Multimodal Reasoning[EB/OL]. arXiv:2506.16141, 2025.
- Simoni M, Rossi F, Bianchi L, et al. GRTPO: Trajectory-Based Group Policy Optimization[EB/OL]. arXiv:2508.03772, 2025.
- Tan H Z, Liang Y, Sun Q, et al. GTPO and GRPO: Token/Sequence Level Entropy Weighted RL[EB/OL]. arXiv:2508.04349, 2025.
- Fan L S, Zeng M, Huang T, et al. P-GRPO: Posterior Reward RL for Code Generation[EB/OL]. arXiv:2306.05170, 2025.
- Dai M Z, Chen L, Yang K, et al. S-GRPO: Serial Group Decaying Reward Optimization[EB/OL]. arXiv:2505.07686, 2025.
- Liu Z R, Wang H, Li P, et al. GHPO: Guided Hybrid Policy Optimiza-

- tion[EB/OL]. arXiv:2507.10626, 2025.
- Li X Y, Zhang H, Wu J, et al. MO-GRPO: Multi-Objective Alignment Framework for LLMs[EB/OL]. arXiv:2503.21819, 2025.
- Wu X Y, Zheng C J, Liu B, et al. LAPO: Length-Adaptive Policy Optimization[EB/OL]. arXiv:2507.15758, 2025.
- Zheng C J, Wu X Y, Wang L, et al. Group Sequence Policy Optimization (GSPO)[EB/OL]. arXiv:2507.16071, 2025.
- Zhang X J, Liu Y, Chen M, et al. EDGE-GRPO: Entropy Driven GRPO with Error Correction[EB/OL]. arXiv:2507.21848, 2025.
- S S. Hybrid Group Relative Policy Optimization[EB/OL]. arXiv:2502.01652, 2025.
- Yang Z C, Tan Y, Li S, et al. TreeRPO: Tree Sampling RL for Mathematical Reasoning[EB/OL]. arXiv:2506.05183, 2025.
- Jia W B, Wang X, Liu H, et al. MEML-GRPO: Multi-Expert Mutual Learning RL[EB/OL]. arXiv:2505.09570, 2025.
- Liu J, He X, Zhang L, et al. Flow-GRPO: RL for Flow Matching Models[EB/OL]. arXiv:2505.05470, 2025.
- He X X, Liu J, Wang Y, et al. TempFlow-GRPO: Temporal Flow RL Framework[EB/OL]. arXiv:2508.04324, 2025.
- Schulman J, Moritz P, Levine S, et al. High-Dimensional Continuous Control Using GAE[EB/OL]. arXiv:1506.02438, 2015.
- Fan T T, Li Y, Zhang Z, et al. T-PPO: Truncated Proximal Policy Optimization[EB/OL]. arXiv:2506.15050, 2025.
- Yue Y, Wang H, Chen S, et al. VAPO: Value-Augmented PPO for Long-CoT Tasks[EB/OL]. arXiv:2504.05118, 2025.
- Chu X X, Peng D, Liu Z, et al. GPG: Simple Group Policy Gradient Baseline[EB/OL]. arXiv:2504.02546, 2025.
- Li C, Wang Y, Zhang L, et al. AGPO: Adaptive Group Policy Optimization[EB/OL]. arXiv:2503.15952, 2025.
- Lin Z H, Huang B, Chen X, et al. CPPPO: Completion Pruning Policy Optimization[EB/OL]. arXiv:2503.22342, 2025.
- Yu Q Y, Li M, Wang S, et al. DAPO: Decoupled Clip Dynamic Sampling RL[EB/OL]. arXiv:2503.14476, 2025.
- Li S H, Zeng T, Zhou H, et al. RePO: Replay-Enhanced Policy Optimization[EB/OL]. arXiv:2506.09340, 2025.
- Zhang X Y, Liu H, Chen L, et al. Critique-GRPO: Natural Language Feedback RL[EB/OL]. arXiv:2508.03106, 2025.
- Chen M H, Wang Z, Xu L, et al. SEED-GRPO: Semantic Entropy Enhanced GRPO[EB/OL]. arXiv:2505.12345, 2025.
- Zhao Y Z, Li C, Wu H, et al. GMPO: Geometric Mean Policy Optimization[EB/OL]. arXiv:2507.20673, 2025.
- Li C G, Zhang Y, Liu B, et al. GAP: Gradient Adaptive Policy Optimization[EB/OL]. arXiv:2507.01915, 2025.
- Liang J, Sun M, Wang Q, et al. ReMix: Reincarnating Mix-Policy PPG[EB/OL]. arXiv:2507.06892, 2025.
- Xu T, Wang H, Zhang Y, et al. SeqPO-SiMT: Sequential Policy Optimization for Simultaneous MT[EB/OL]. arXiv:2505.20622, 2025.
- Zhang K C, Liu P, Chen L, et al. GVPO: Group Variance Policy Optimization[EB/OL]. arXiv:2504.19599, 2025.
- Pan C B, Zhou Y, Wang S, et al. GRQO: Group Relative Query Optimization for Vision Models[EB/OL]. arXiv:2505.24025, 2025.
- Lyu S K, Zhang H, Li M, et al. HBPO: Hierarchical Budget Policy Optimization[EB/OL]. arXiv:2507.15844, 2025.
- Han J H, Chen X, Wu T, et al. COPO: Consistency-Aware Policy Optimization[EB/OL]. arXiv:2508.04138, 2025.
- Dong P, Wang L, Zhang H, et al. EXPO: Expressive Policy Optimization[EB/OL]. arXiv:2507.07986, 2025.
- Li G, Zhao H, Lin S, et al. DisCO: Discriminative Constrained Optimization[EB/OL]. arXiv:2505.12366, 2025.
- Zhang X J, Wang Y, Liu Z, et al. SRPO: Two-Stage History Resampling RL[EB/OL]. arXiv:2304.14298, 2025.
- Li Y H, Chen T, Wu J, et al. MAGPO: Multi-Agent Guided Policy Optimization[EB/OL]. arXiv:2507.15059, 2025.
- Wang Z H, He B, Zhang L, et al. PAPO: Perception-Aware Policy Optimization[EB/OL]. arXiv:2507.06448, 2025.
- Wan Z W, Li Q, Chen H, et al. SRPO: Multimodal Self-Reflection RL[EB/OL]. arXiv:2506.01713, 2025.
- Dong G, Liu Y, Wang H, et al. ARPO: Agentic Reinforced Policy Optimization[EB/OL]. arXiv:2507.19849, 2025.
- Tang C H, Zhou L, Yang M, et al. TSPO: Temporal Sampling Policy Optimization for Video LLMs[EB/OL]. arXiv:2508.04369, 2025.
- Xia B Q, Zhang H, Li M, et al. MiMo-7B: End-to-End Reasoning Model from Pretraining to RL[EB/OL]. arXiv:2505.07608, 2025.
- Guo Y R, Wang T, Chen X, et al. SPO: Segment Policy Optimization for LLMs[EB/OL]. arXiv:2505.23564, 2025.
- Ren A Z, Liu S, Zhang L, et al. DPPO: Diffusion Policy Policy Optimization[EB/OL]. arXiv:2409.00588, 2024.