Beijing Forest Studio
北京理工大学信息系统及安全对抗实验中心

强化学习中的信用分配

硕士研究生 贺晨阳

2026年1月11日

问题回溯

- 总结反思
  - 算法原理部分讲解不够细致
  - 实验结果等图片太小，排版存在问题
- 相关内容
  - 2024.4.3 杨宗源《LLM中的强化学习》
  - 2022.3.28 门元昊《强化学习基础与实战》

内容提要

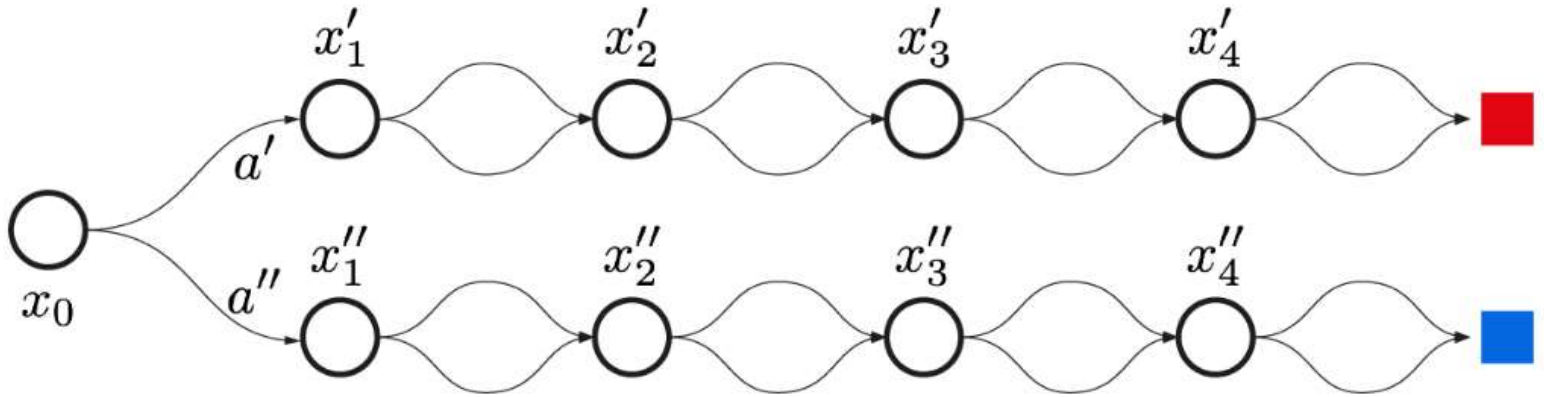
- 预期收获
- 内涵解析与研究目标
- 研究背景
- 知识基础
- 研究历史与现状
- 算法原理
  - LaRe
  - VinePPO
- 特点总结与未来展望
- 参考文献

预期收获

- 预期收获
  - 掌握强化学习信用分配的基本概念
  - 了解信用分配的研究背景
  - 了解信用分配常见方法及其原理

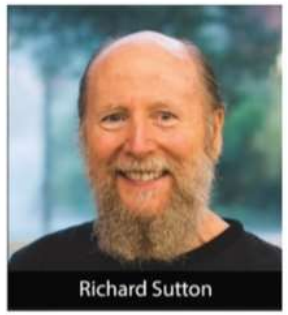
内涵解析与研究目标

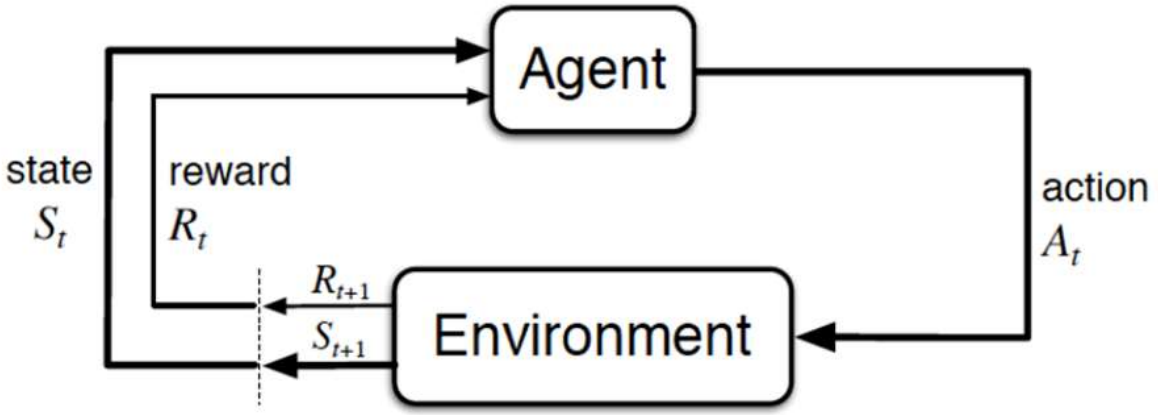
- 题目内涵解析
  - 信用(Credit)：强化学习中衡量一个动作对实现特定结果所产生的潜在影响
  - 信用分配问题(CAP)：强化学习智能体将信用分配给所涉及的众多动作
- 研究目标
  - 如何在延迟奖励、稀疏奖励情况下识别出对结果起决定性作用的关键动作
  - 如何将环境的随机性与智能体的能力解耦，实现更稳健的策略改进
  - 如何有效利用数据，减少训练所需的样本量



研究背景

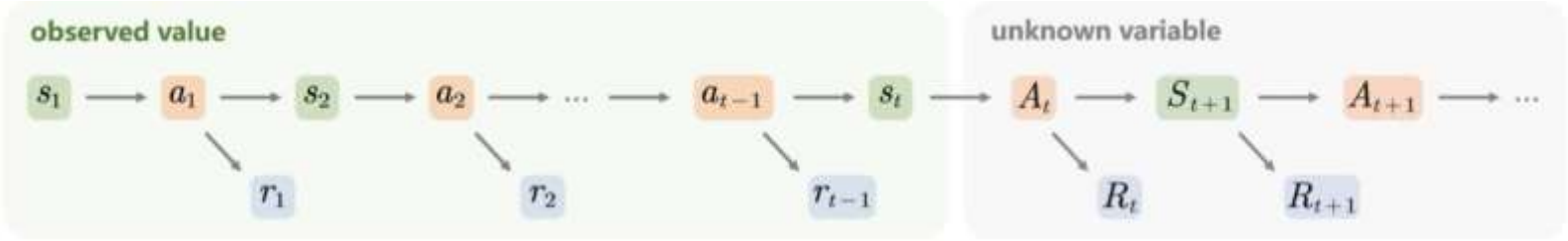
- 强化学习
  - 智能体在复杂、不确定的环境中最大化它能获得的奖励，从而达到自主决策的目的
  - 基本概念
    - 智能体Agent、环境
    - 动作a、状态s、奖励r
  - Agent依据策略决策从而执行动作，然后通过感知环境从而获取环境的状态，进而得到奖励
  - 找到一个策略来最大化奖励





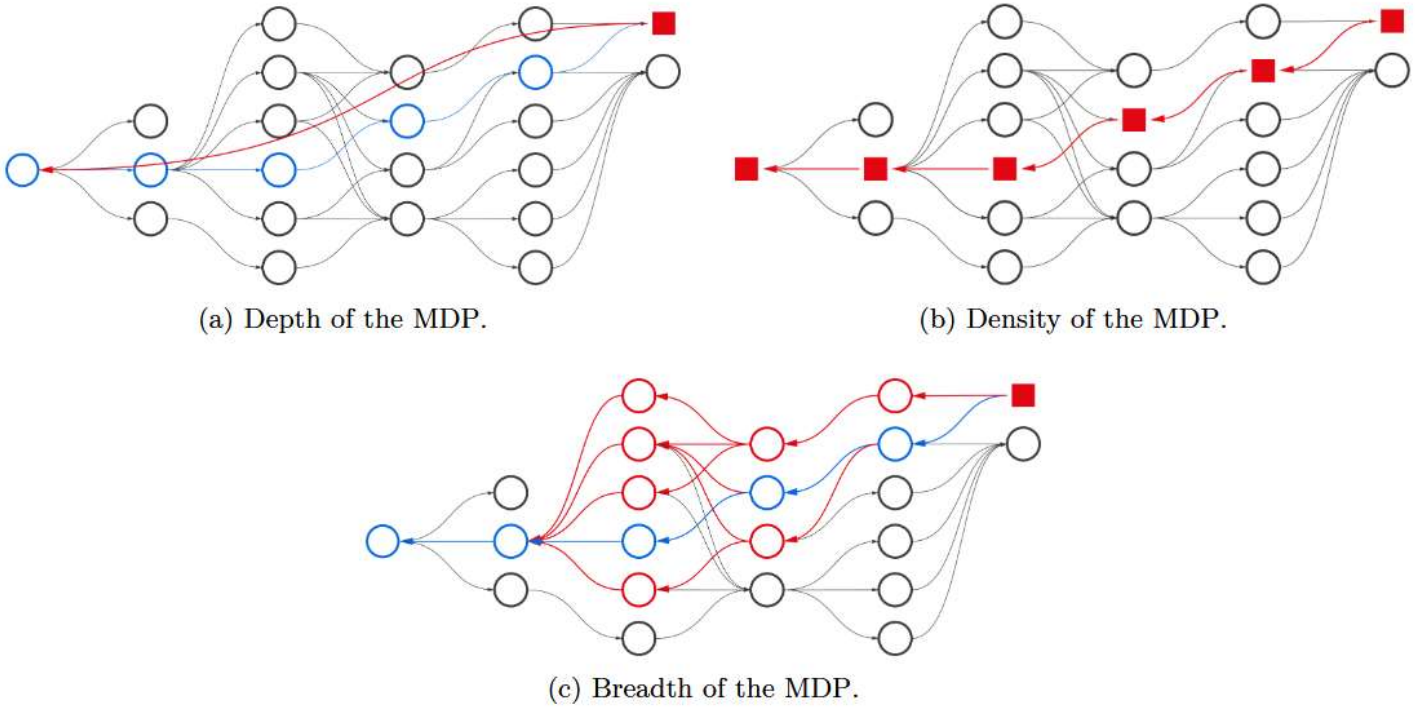
研究背景

- 马尔可夫决策过程
  - 回报：$U_t = \sum_{k=t}^{n} \gamma^{k-t} R_k$
  - 策略：$\pi(a|s)$
    - 在状态s下做出a动作概率分布
  - 状态转移：$P(s'|s,a)$
    - 在状态s下做出动作a到达下一状态的概率分布
  - 状态价值：$V_\pi(s) = \mathbb{E}_\pi(U_t|S_t = s) = \mathbb{E}_\pi(R_t + \gamma U_{t+1}|S_t = s)$
  - 动作价值：$Q_\pi(s,a) = \mathbb{E}_\pi(U_t|S_t = s, A_t = a) = \mathbb{E}_\pi(R_t + \gamma U_{t+1}|S_t = s, A_t = a)$
  - 状态价值与动作价值关系：$V_\pi(s) = \sum_{a \in A} \pi(a|s) Q_\pi(s,a)$
    - 状态的价值取决于做出所有动作的概率分布乘每个动作的价值
- 强化学习两类思路
  - 基于价值：评估动作的价值，选择价值最大的动作
  - 基于策略：训练一个网络，得到所有动作概率分布，取概率最大的动作



研究背景

- 信用分配面临的挑战
  - 延迟奖励：在决定性动作发生很久之后才获得奖励
  - 稀疏奖励：奖励在各处均为零，极少出现峰值，导致缺乏信息
- MDP三个特征带来的影响
  - 深度：动作链很长
  - 密度：误认为动作对结果的影响力微乎其微
  - 广度：多条路径有同样结果，无法判断决定性动作



(a) Depth of the MDP.

(b) Density of the MDP.

(c) Breadth of the MDP.

研究历史与现状

1988 Richard等人提出了时序差分TD learning，奠定了现代强化学习的基础。它提出了利用当前预测与未来预测的差异来更新价值

2015 Schaul等人提出了UVFA提出了将目标作为价值网络的显式输入。这使得智能体能够泛化到不同的目标，为后来的方法提供了架构基础

2019 Jose等人针对长延迟奖励，提出RUDDER算法，用LSTM预测回报，并将预测值的差分作为人造的即时奖励分发给关键动作。理论上证明了这与原问题是回报等价的

2021 Thomas等人提出反事实信用分配方法CCA,引入了因果推理。通过数学方法剥离了环境噪音和运气的影响，试图量化动作变化对结果的影响

2021 Chen等人提出Decision Transformer，利用注意力机制处理长序列。将信用分配隐式地转化为序列预测中的注意力权重分配

2024 Gupta等人提出了一种双向价值函数，该方法对过去的期望回报进行建模，解决传统TD方法在线非线性函数逼近中因梯度信息过时而导致信用分配错误的问题

2025 Qu等人提出LaRe通过引入潜在奖励的概念，利用大模型提供了一种多维度的、语义可解释的性能评估方式，解决信用分配问题

2025 Kazemnejad等人提出VinePPO，在PPO基础上，用蒙特卡洛估计取代价值网络，提升了大模型在数学推理任务中的训练效率、准确率和泛化能力，证明了精细信用分配在LLM强化学习中的关键作用。

研究历史与现状

- 方法分类
  - 基于时间邻近性的方法
    - 按时间链式传导
    - 基于优势的方法、重新加权更新与复合目标
  - 回报分解
    - 将最终的回报直接重新分配到每一个时间步上
    - 一般需要训练辅助模型预测回报
  - 预定义目标
    - 定义好一组不同于主任务的辅助目标，同时学习目标的价值函数
    - 需要预定义目标，限制了应对不同任务灵活性
  - 序列建模
    - 视为序列预测，用 Transformer 隐式解决信用分配

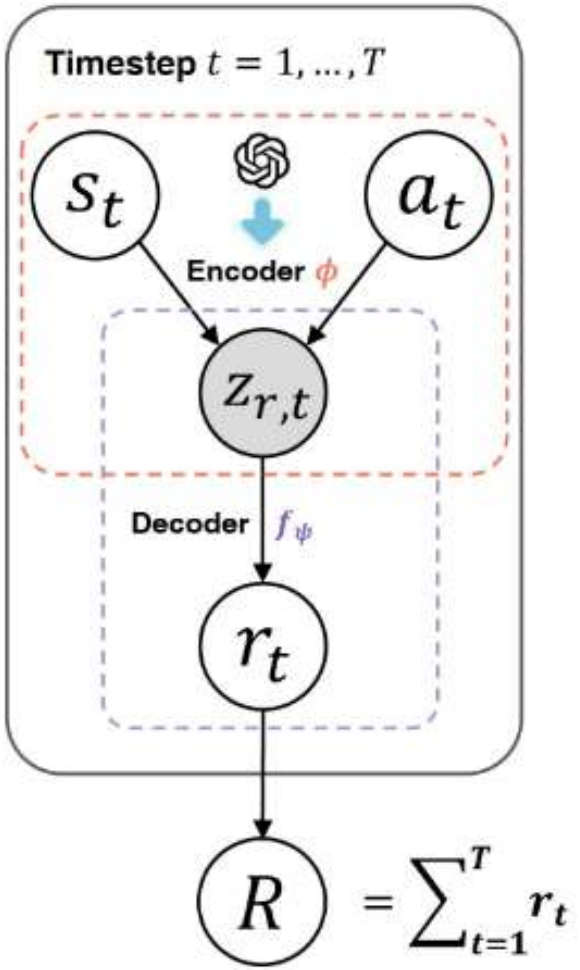
【AAAI-2025】
Latent Reward: LLM-Empowered Credit Assignment in Episodic Reinforcement Learning

LaRe TIPO

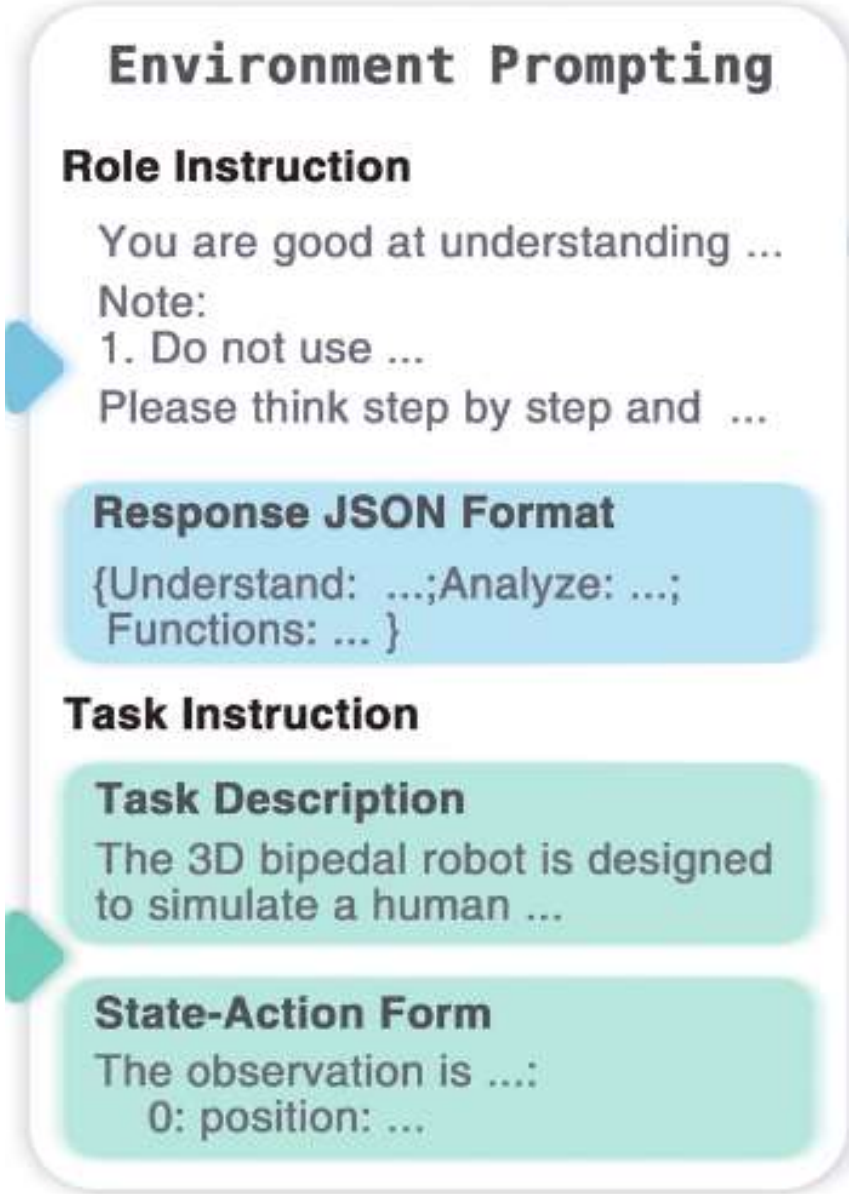
T	目标	利用LLM设计函数提取潜在奖励，过滤无关特征
I	输入	LLM*1，设计好的提示词*1
P	处理	1.构建环境提示，设计通用提示模板 2.潜在奖励自我验证，大模型生成编码规则并自我迭代 3.贡献分配，训练奖励解码器用来预测奖励
O	输出	奖励编码函数*1、训练好的解码器*1

P	问题	现有方法并未去除原始状态中无关的冗余特征
C	条件	需要访问大模型、需要结合已有强化学习算法
D	难点	如何将奖励分配给中间过程 如何避免LLM推理时随机性和幻觉带来的影响
L	水平	2025 CCF A

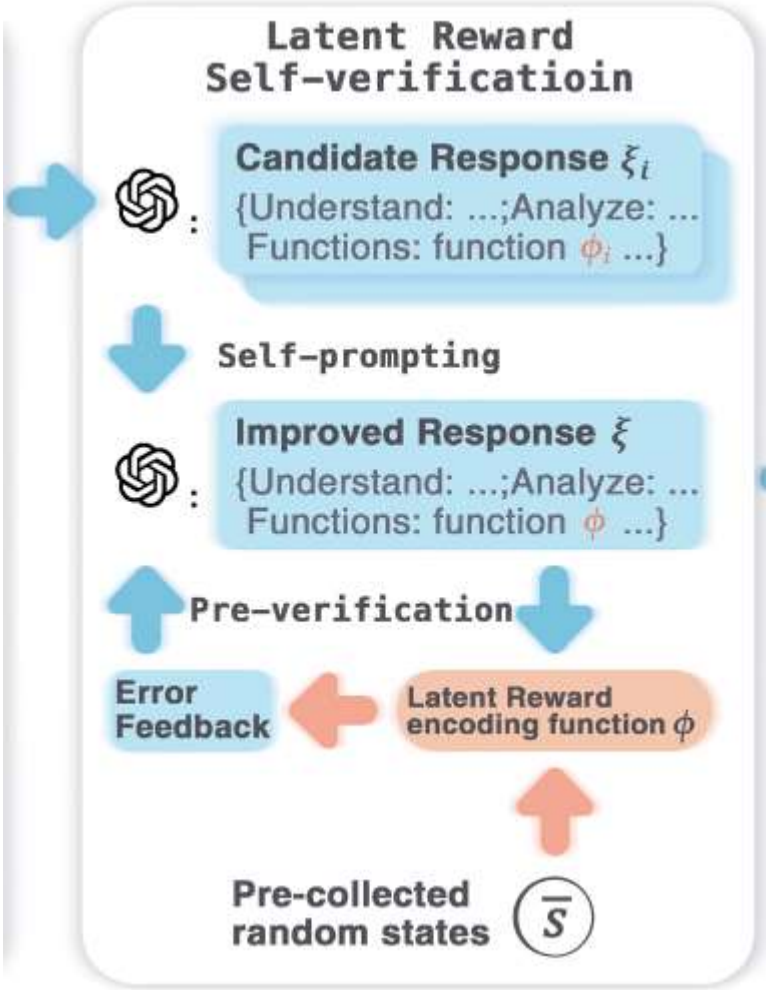
- 潜在奖励
  - 定义：从原始状态经过提炼的、具有明确物理意义的特征向量
  - 原始状态
    - 是强化学习环境在每一个时间步反馈的原始观测数据，
    - 通常为状态动作对(s_t, a_t)得到的长向量
    - 维度过高、存在冗余
  - 需解决问题：减少与奖励无关的冗余来简化网络训练
  - 潜在奖励优势：与直接从原始状态估计逐步奖励相比，可解释性方面具有优势



- 构建环境提示
  - 角色指令
    - 角色定义
    - 强制思维链：理解任务和状态->识别奖励相关因素->生成潜在奖励编码函数
    - 输出格式：以json格式输出
  - 任务指令
    - 任务描述:说明任务背景以及环境信息中各维度的含义

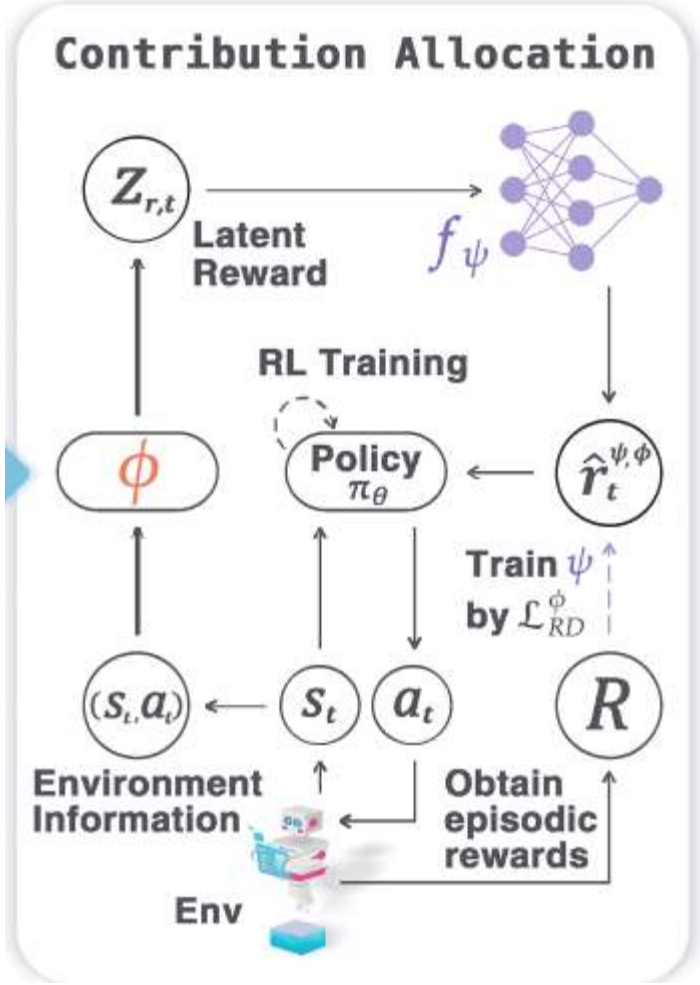


- 潜在奖励自我验证
  - 解决问题：LLM 推理中的随机性和幻觉
  - 自我提示
    - 生成n个候选，每个响应包括潜在奖励编码函数的一个代码实现$\xi_1, \xi_2, \ldots, \xi_n \leftarrow \mathcal{M}(task, role)$
    - 在n个候选中，选出最优
  - 预验证
    - 提取最优函数的python代码实现
    - 利用预收集数据运行测试
    - 根据反馈报错迭代最终得到潜在奖励编码函数ϕ



- 贡献分配
  - 目标：通过潜在奖励向量解码得到奖励
  - 训练解码器f_ψ，对给定的特征向量z_r打分，输出标量$\hat{r}$
  - 将Agent每一轮轨迹τ存入缓存区，随机抽取一批轨迹，将每步的原始状态编码为潜在奖励向量，作为解码器训练数据
  - 所有步骤奖励之和尽可能接近总奖励R

$$\mathcal{L}_{RD}(\psi) = \mathbb{E}[(R(\tau) - \sum_{t=1}^{T} f_\psi(\phi(s_t, a_t)))^2]$$

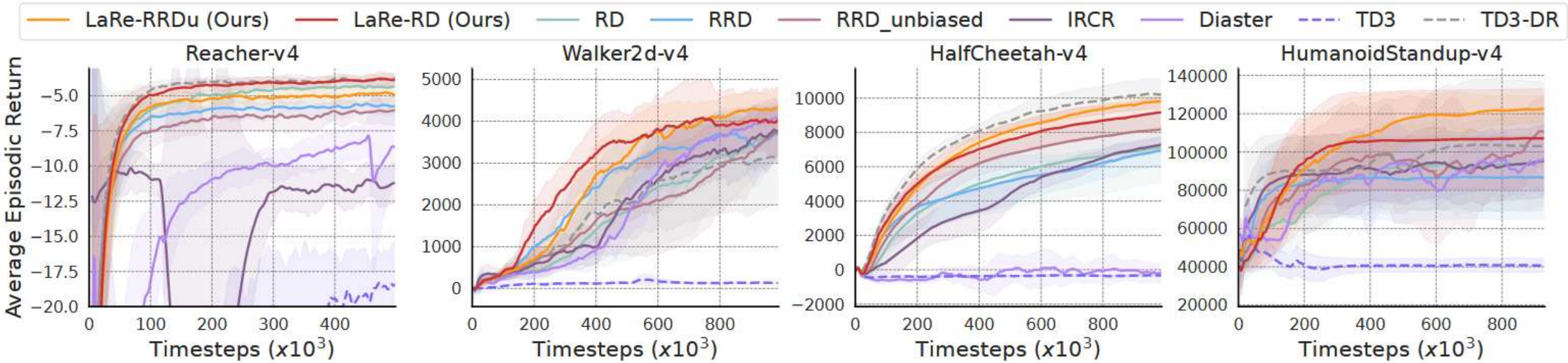


- 测试基准
  - MuJoCo 运动基准：机器人强化学习任务
  - 智能体粒子环境MPE：多智能体协作基准
  - 任务具有中间密集奖励，但训练时只参考最终奖励
- 评价指标
  - 平均情节回报 ： 五次测试获得的平均最终奖励
    - 通过测试训练后智能体的能力来验证方法的有效性
  - 皮尔逊相关系数：计算潜在奖励和真实密集奖励之间的相关性
- 实验设置
  - 大模型选用GPT-4o
  - 强化学习算法选择TD3 、IPPO

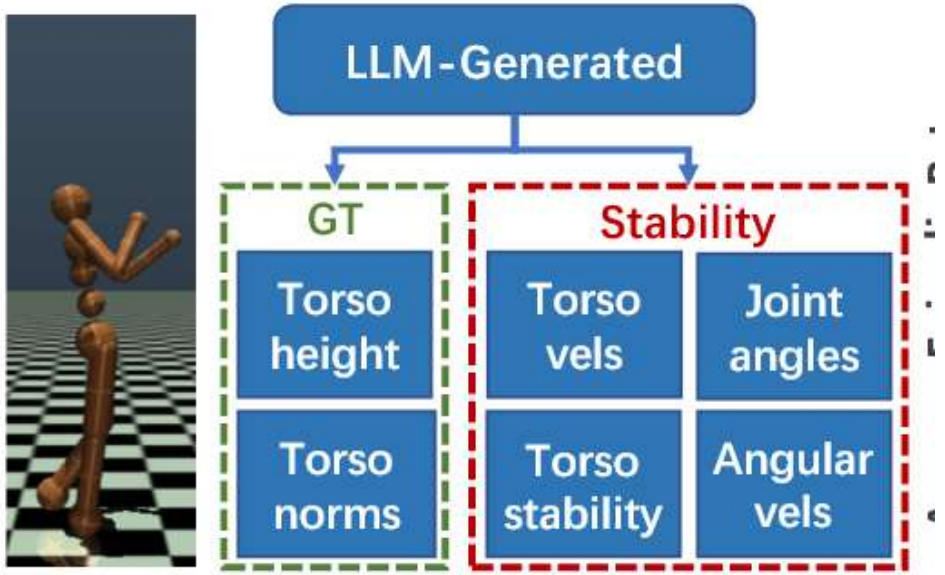
- 对比方法
  - RD(2021):基础回报分解，训练模型预测每一步的奖励
  - RRD(2021):随机回报分解
  - IRCR (2020) :基于轨迹空间平滑的奖励学习方法
  - Diaster (AAAI 2024):随机切割子轨迹，利用差异来推断每一步的贡献
  - AREL (2022):基于注意力的多智能体奖励再分配
  - STAS (2023): 时空回报分解
  - TD3-DR / IPPO-DR：使用环境原本自带的每一步密集奖励

LaRe 实验结果

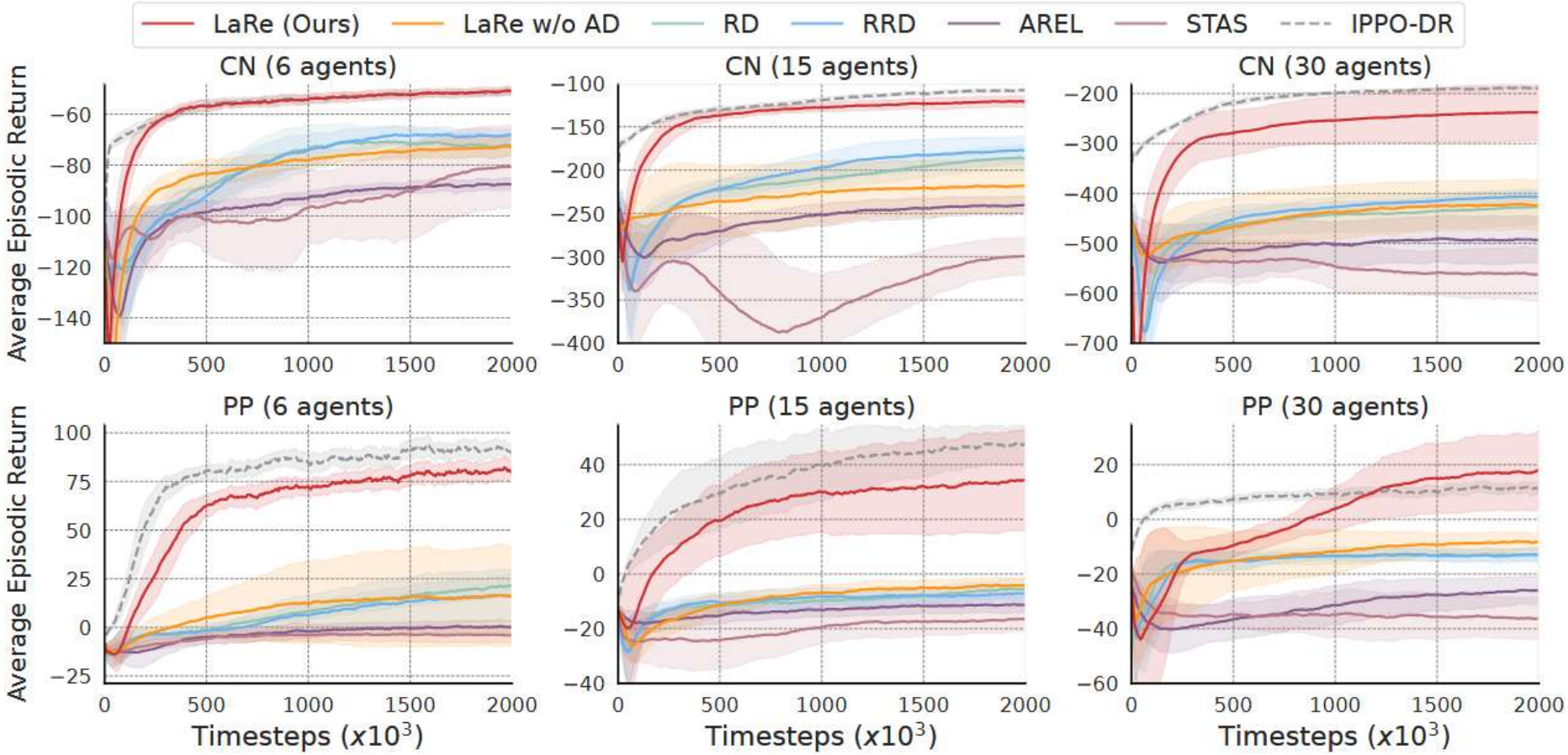
- 评估不同状态空间维度任务的平均回合收益
  - 未做奖励分解的方法垫底
  - LaRe用更短的时间达到更高的奖励
  - LaRe效果优于使用环境原本自带的每一步密集奖励的最佳方法



- 现象解释
  - LaRe效果有时优于最佳方法
  - 自带奖励也是由人为定义的，忽略了造成影响隐含因素
  - 举例说明
    - 机器人站立保持平衡任务
    - 人类设计奖励函数关注躯干高度和躯干范数
    - 大模型额外关注关节角度、躯干速度、角速度等稳定性相关因素



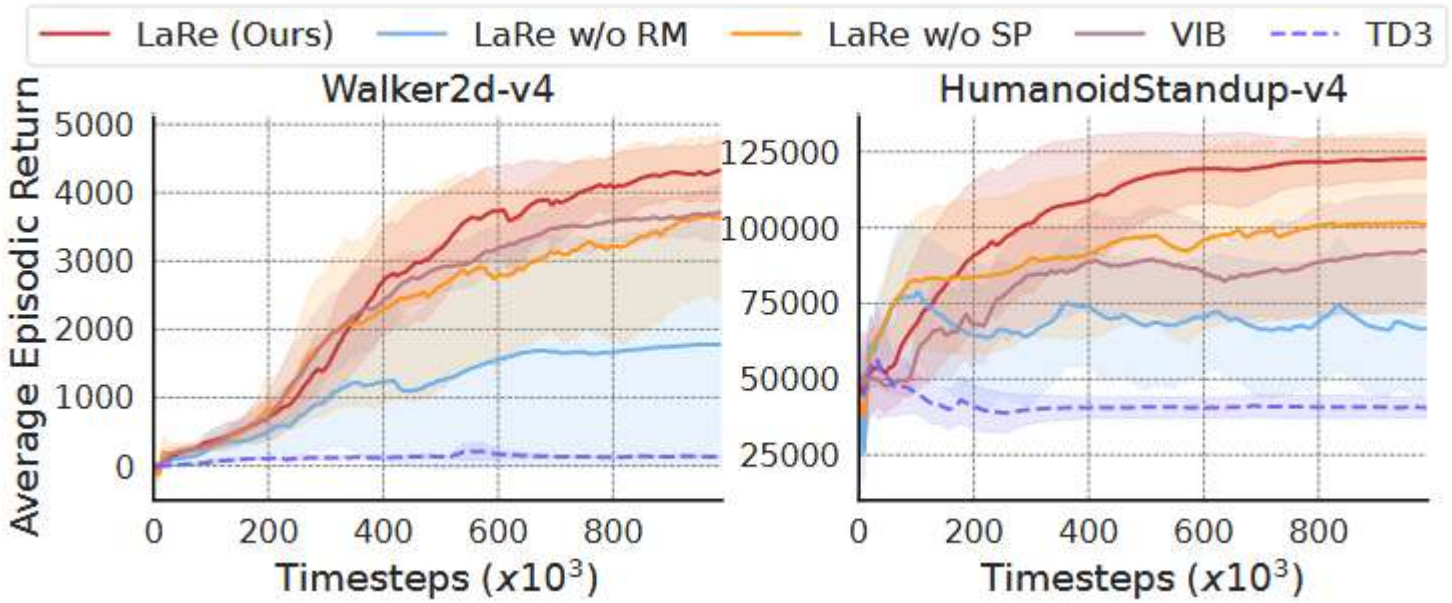
- 评估多智能体环境下效果
  - 多智能体情况下依然保持性能



- 冗余消除效果实验
  - corr 表示平均皮尔逊相关系数,dims表示原始状态或潜在奖励的平均维度数
  - 比较与真实密集奖励的相关性，相比于原始状态，潜在奖励的相关性更高

Tasks	corr (dims)	
	States	Latent Rewards
CN (6 agents)	0.02 (26)	**0.50** (5.6)
PP (6 agents)	0.01 (28)	**0.12** (5.4)
HalfCheetah-v4	0.22 (17)	**0.53** (4.8)
HumanoidStandup-v4	0.20 (376)	**0.49** (5.6)

- 消融实验
  - LaRe w/o RM解码器消融
  - LaRe w/o SP自我提示消融
  - VIB 用纯数学的算法压缩状态，代替大模型



LaRe 算法总结

- 算法贡献
  - 提出了潜在奖励概念，并设计了基于潜在奖励的框架LaRe
  - 旨在解决稀疏奖励带来的信用分配难题
  - 消耗算力较少、具有可解释性
  - 适用于时间信用分配外，还适用于多智能体信用分配
- 算法不足
  - 编码函数是静态的且高度依赖状态描述
  - 难以应对环境状态无法用语言描述的情况
  - 强烈依赖 LLM 的能力

【 ICML 2025 】
VinePPO: Refining Credit Assignment in RL Training of LLMs

VinePPO TIPO

T	目标	利用无偏的估计方法代替PPO中价值网络为模型提供更精准的梯度信息
I	输入	LLM*1、数学问题数据集*1
P	处理	1.主轨迹生成 2.状态重置与蒙特卡洛采样 3.价值计算 4.优势计算与更新
O	输出	优化后的LLM*1

P	问题	现有方法依赖Critic网络预测价值，存在较大偏差
C	条件	可以修改大模型参数
D	难点	如何合理的将奖励分配给中间过程 如何修改PPO算法使其无偏的预测价值
L	水平	2025 CCF A

- **Policy Gradient**
  - 如何训练智能体
  - 轨迹$\tau = \{s_0, a_0, s_1, a_1 \ldots . s_T, a_T\}$
  - 轨迹概率
    - $p(\tau|\theta) = p(s_0) \prod_{t=0}^{T} \pi_\theta(a_t|s_t) p(s_{t+1}|s_t, a_t)$
  - 优势函数
    - $A(s,a) = Q(s,a) - V(s)$
  - 利用目标函数梯度更新参数
    - $J(\pi_\theta) = \mathbb{E}_{\tau\sim\pi_\theta}[R(\tau)]$
    - $\nabla_\theta J(\pi_\theta) = \nabla_\theta \mathbb{E}_{\tau\sim\pi_\theta}[R(\tau)]$

$$= \mathbb{E}_{\tau\sim\pi_\theta}[\sum_{t=0}^{T} \nabla_\theta \log \pi_\theta(a_t|s_t)\, R(\tau)]$$

- **梯度计算过程**

$$\nabla_\theta J(\pi_\theta) = \nabla_\theta \mathbb{E}_{\tau\sim\pi_\theta}[R(\tau)]$$

$$= \int \nabla_\theta P(\tau|\theta) R(\tau) d\tau$$

$$= \int P(\tau|\theta) \nabla_\theta \log P(\tau|\theta)\, R(\tau) d\tau$$

$$= \mathbb{E}_{\tau\sim\pi_\theta}[\nabla_\theta \log P(\tau|\theta) R(\tau)]$$

$$= \mathbb{E}_{\tau\sim\pi_\theta}[\nabla_\theta \log p(s_0) R(\tau) +$$

$$\nabla_\theta \sum \log p(s_{t+1}|s_t, a_t) R(\tau) +$$

$$\nabla_\theta \sum \log \pi_\theta(a_t|s_t)\ R(\tau)]$$

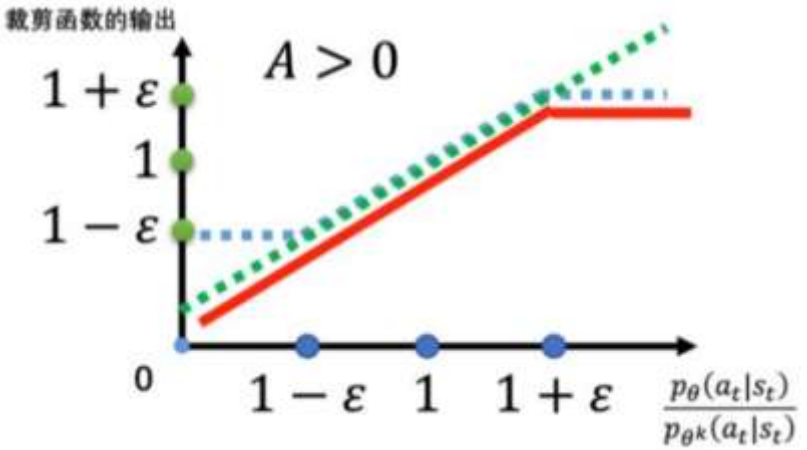
$$= \mathbb{E}_{\tau\sim\pi_\theta}[\sum_{t=0}^{T} \nabla_\theta \log \pi_\theta(a_t|s_t)\, R(\tau)]$$

VinePPO PPO

- Policy Gradient存在的问题
  - 方差大，容易受到运气等情况的干扰
  - 学习的步长难以调整，容易崩溃
- Actor-Critic的改进
  - 引入Critic网络评估价值函数
  - 引入优势函数代替回报
  - 不仅看最终回报的正负，也关注回报是否强于基准
- PPO的改进
  - 以旧策略为参考，限制更新幅度
  - 保证了策略是平滑迭代，不会轻易训练崩溃

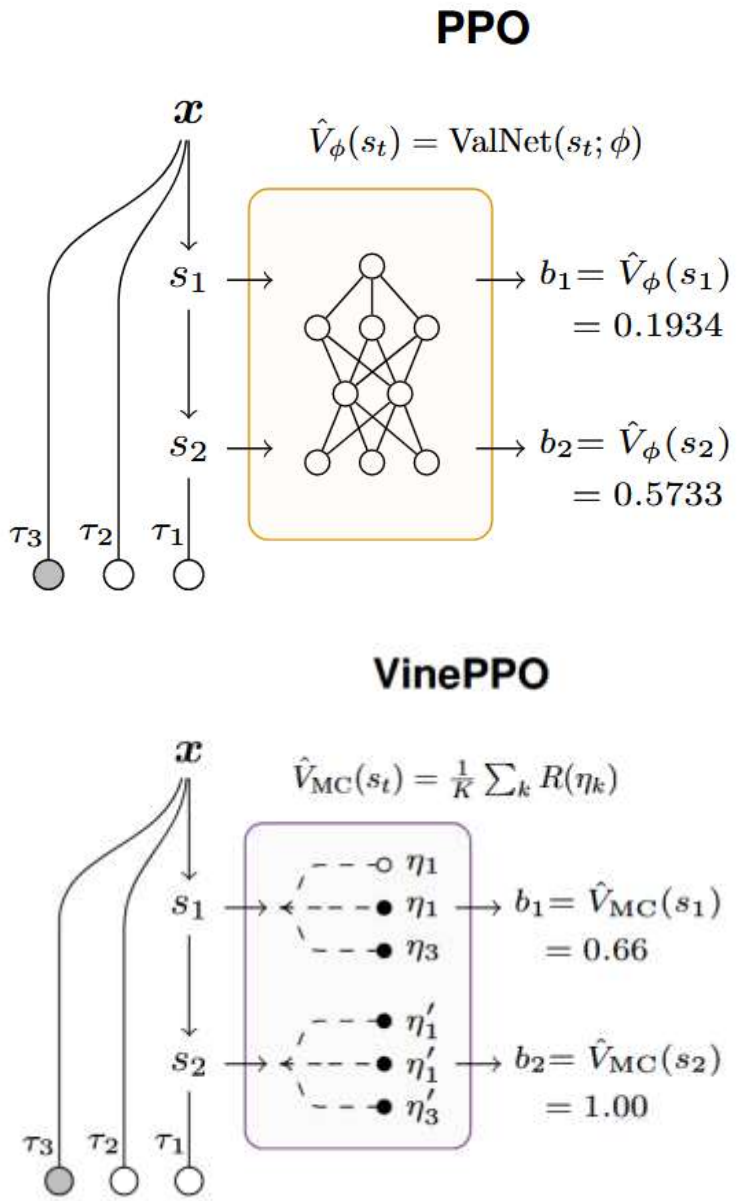
- **PPO算法**
  - 优化目标$J_{PPO}(\vartheta)=\mathbb{E}_t[\min\left(\frac{\pi_\theta(a_t|s_t)}{\pi_{\theta_{old}}(a_t|s_t)}A_t, clip(\frac{\pi_\theta(a_t|s_t)}{\pi_{\theta_{old}}(a_t|s_t)}, 1-\epsilon, 1+\epsilon)A_t\right)]$
  - 优势估计：用A_t代替$R(\tau)$，训练Critic网络预测价值
  - 重要性比率 $r_t(\theta)=\frac{\pi_\theta(a_t|s_t)}{\pi_{\theta_{old}}(a_t|s_t)}$
  - 截断函数：防止策略崩溃
    - $r_t(\theta)>1+\epsilon$时新策略动作概率显著大于旧策略

      裁剪为常数梯度为0，模型不更新
    - $1-\epsilon<r_t(\theta)<1+\epsilon$时，保持不变
    - $r_t(\theta)<1-\epsilon$时min函数生效



VinePPO 算法原理

- 算法原理
  - 语言模型可以轻易的重置状态
  - 蒙特卡罗采样：不停抽样，用样本均值去逼近真实期望
  - 算法过程
    - 对于训练轨迹τ中每个状态s_t，重置状态多次随机采样后续内容，得到辅助轨迹η_k
    - 计算平均回报来获得蒙特卡洛估计，且η_k仅参与价值估计，不用做训练
      - $V_{MC}(s_t) = \frac{1}{K}\sum_{k=1}^{K} R(\eta_k)$
    - 计算优势函数后带入PPO算法继续训练

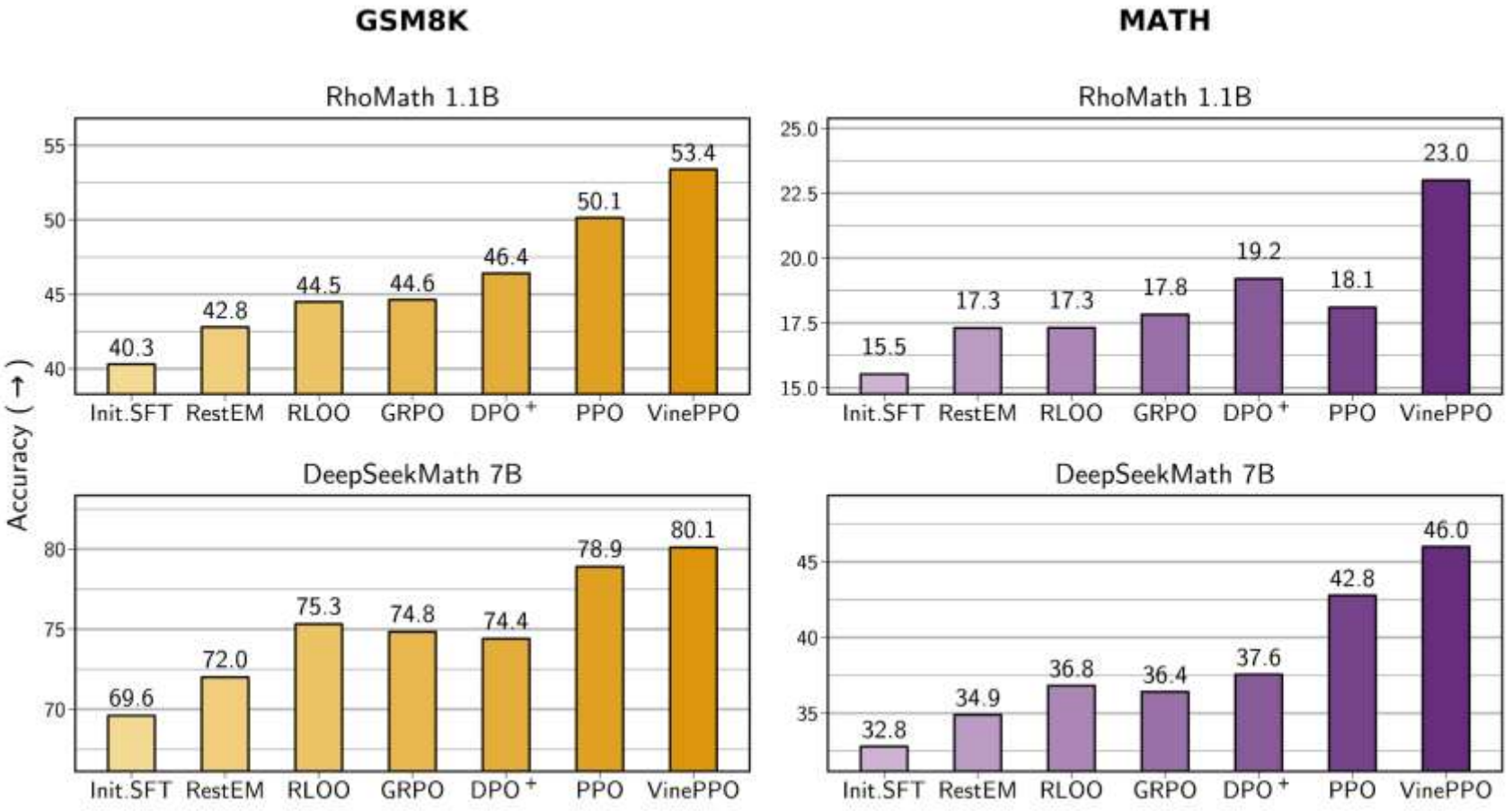


VinePPO 实验设计

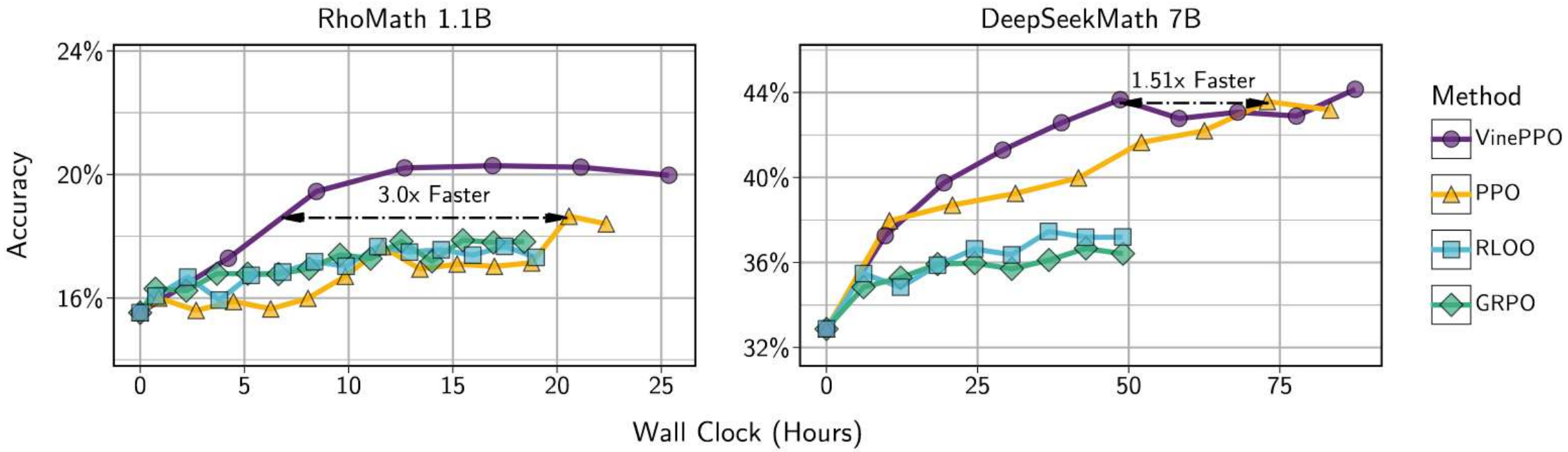
- 基准模型
  - DeepSeekMath 7B　RhoMath 1.1B
- 测试数据
  - MATH、GSM8K
  - 不同难度的数学问题，在最后给出二元奖励
- 评价指标
  - 准确率Pass@1
- 对比方法
  - PPO（2022），RLOO(2024)、GRPO(2024)、RestEM(2017)、DPO+（2024）
  - 除VinePPO和PPO外都省略了显示信用分配，给所有token同样价值

VinePPO 实验结果

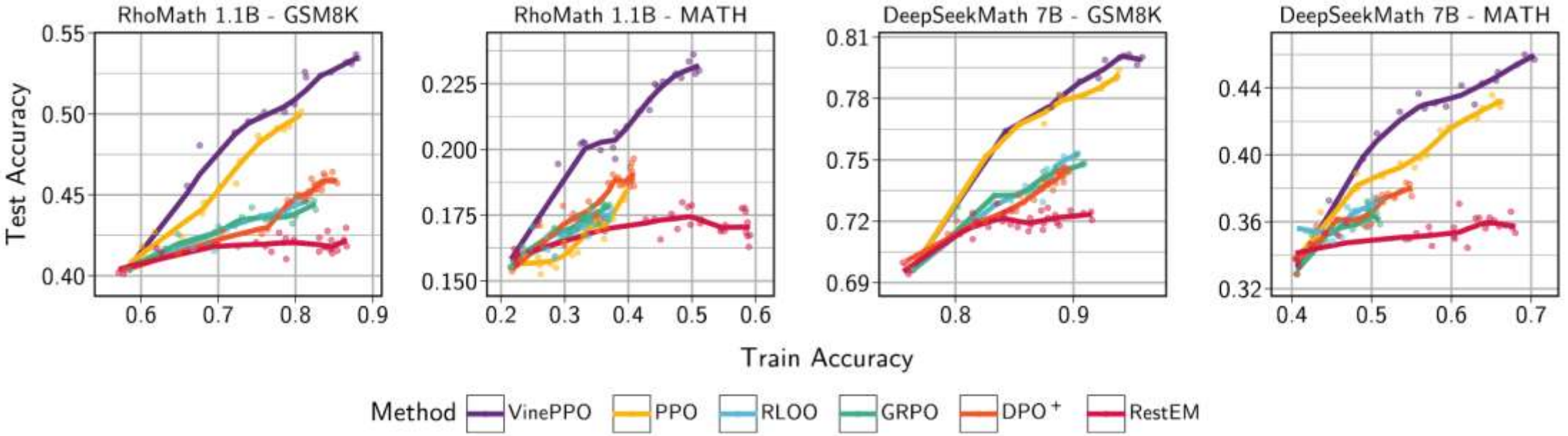
- 对比实验
  - 在不同模型和数据集上效果最佳
  - 在更难的数据集上，与其他方法差距更大



- 效率实验
  - 相同硬件的条件下到达一定准确率用时比其他方法更短
  - 单步慢，但是梯度方向精准，优化效率反而变高

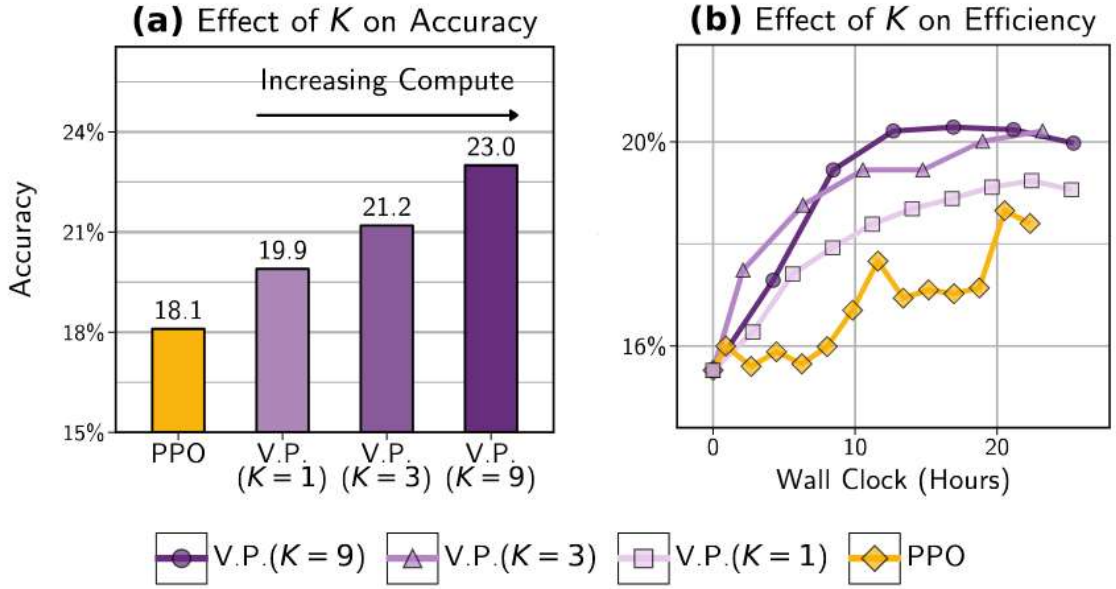


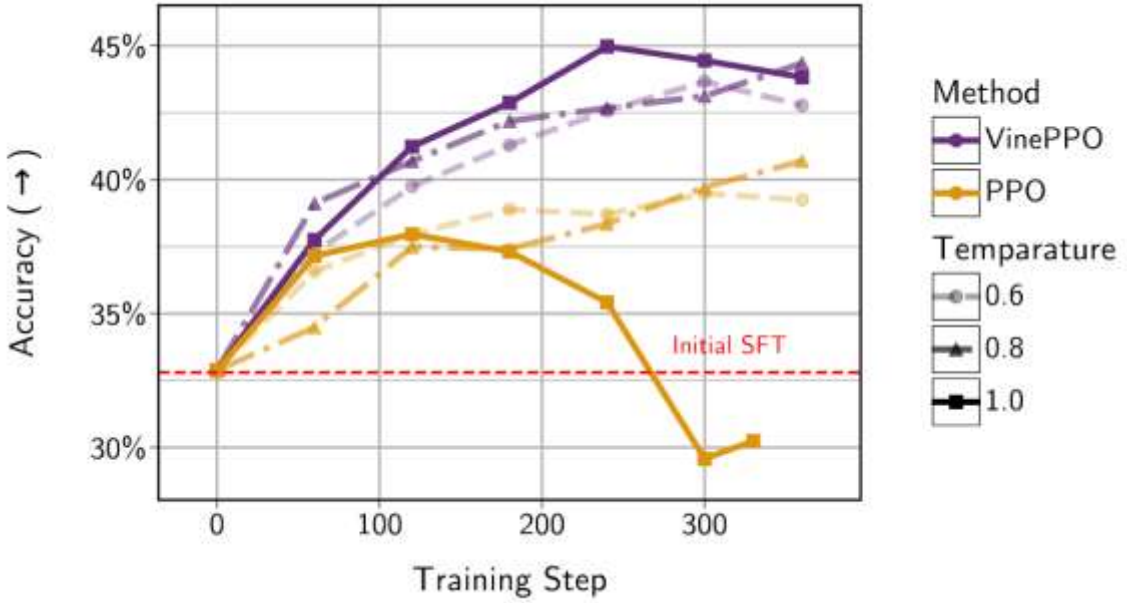
- 泛化实验
  - 泛化增益最高，同一的训练准确率下达到更高的测试准确率
  - 训练中学到的知识能很好地迁移到测试集



VinePPO 实验结果

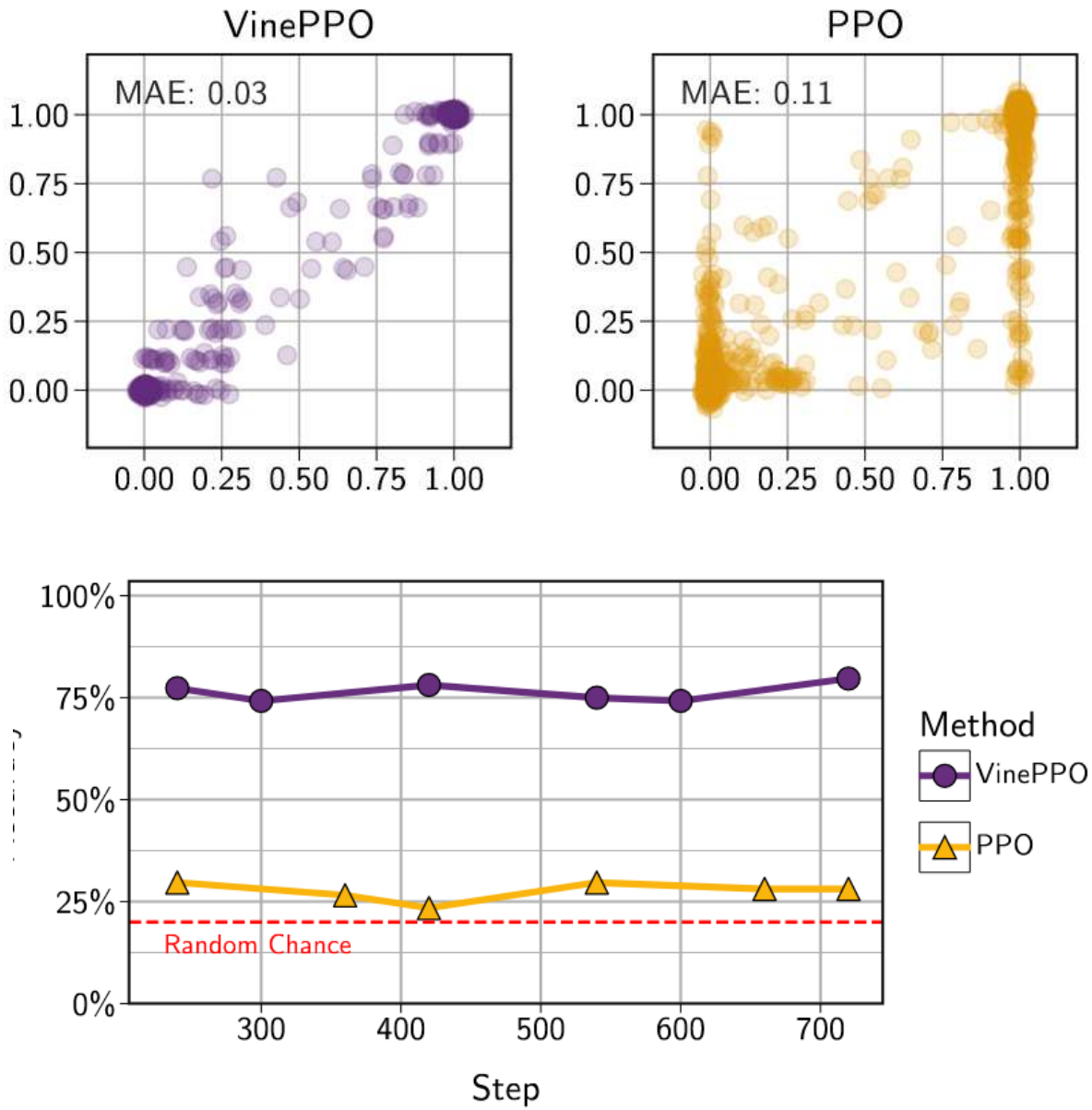
- 参数实验
  - 辅助采样轨迹数量K的影响，无论k取值都优于PPO
  - 蒙特卡洛估计是无偏的，且大模型经过预训练后生成轨迹正确率高于从零探索
  - 大模型温度的影响，高温度下PPO崩溃，但VinePPO效果反而好
  - VinePPO不需要泛化没见过状态，体现采样的优越性





VinePPO 实验结果

- 价值预测实验
  - 用高密度蒙特卡洛256次取平均来代表真实价值
  - PPO的价值网络预测效果偏差很大
  - VinePPO效果接近无偏
  - 给出五个选项选择最PPO效果甚至仅与随机选取差不多



- 算法贡献
  - 指出了传统PPO算法中价值网络效果差的问题
  - 指出了采样方法在大模型上的优势
    - 状态回溯零成本
    - 状态转移的确定性
  - 在PPO基础上加入无偏的价值估计机制
  - 基于蒙特卡洛方法进行价值估计，实际代替预测，将预测问题变成统计问题

- 算法不足
  - 只适用于可以LLM等可以简单回溯的场景，泛用性差
  - 更侧重于推理任务

特点总结与未来展望

特点总结与未来展望

- 特点总结
  - LaRe
    - 利用大模型生成编码函数，训练解码器进行奖励分配
    - 利用大模型先验知识，具有语义可解释性
    - 高度依赖状态描述以及LLM能力
  - VinePPO
    - 通过重置状态多次随机采样获得蒙特卡洛估计作为价值估计
    - 改善了传统PPO方法中价值估计网络在LLM领域效果差的问题
    - 仅适用于可以简单做到回溯的场景
  - 未来发展
    - 从相关性走向因果性
    - 用元学习让模型学会如何分配信用，在新任务上自动调整

参考文献

[1] Kazemnejad A, Aghajohari M, Portelance E, et al. VinePPO: Refining Credit Assignment in RL Training of LLMs[J]. arXiv preprint arXiv:2410.01679, 2024.

[2] Qu Y, Jiang Y, Wang B, et al. Latent reward: Llm-empowered credit assignment in episodic reinforcement learning[C].Proceedings of the AAAI Conference on Artificial Intelligence. 2025, 39(19): 20095-20103.

知人者智，自知者明。胜人者有力，自胜者强。知足者富。强行者有志。不失其所者久。死而不亡者，寿。

谢谢！