

PR #1961 完整报告

THUDM/slime

[docs] Add docs for agent rl

合并时间: 2026-05-27 15:26

原文链接: <http://prhub.com.cn/THUDM/slime/pull/1961>

执行摘要

- 一句话: 新增 Agent RL 文档与路线图
- 推荐动作: 值得精读。尤其是 customization.md 中关于 custom_generate 返回多个样本的契约 (共享 rollout_id) 和奖励分配模式, 是设计 agent 训练的重要决策。

功能与动机

Agentic RL 是 slime 的重要方向, 需要系统化文档指导用户接入工具调用、多智能体、沙盒交互等工作流。PR 标题和 body 明确为 Agent RL 添加文档。

实现拆解

1. 新增 docs/en/get_started/agent.md 和 docs/zh/get_started/agent.md: 提供 Agentic RL 训练路线图, 以表格形式列出不同场景的推荐入口 (custom-generate-func、custom-rm、fan-out、fully-async、coding-agent 示例等), 并给出推荐集成方式和服务性能配置建议。
2. 修改 docs/en/get_started/customization.md 和 docs/zh/get_started/customization.md: 在 custom_generate 签名中增加 | list[Sample] 返回值, 新增“一个 prompt 产生多个训练样本”小节, 包含代码示例 (使用 copy.copy 和 rollout_id 共享) 和奖励分配建议。
3. 修改 docs/en/index.rst 和 docs/zh/index.rst: 在 toctree 中添加 get_started/agent.md 和 _examples_synced/coding_agent_rl/README.md, 确保新文档出现在导航中。

关键文件:

- docs/en/get_started/agent.md (模块 文档; 类别 docs; 类型 documentation): 新增的英文 Agent RL 路线图文档, 是本次 PR 的核心内容, 定义了 agent 功能入口矩阵和集成模式。
- docs/zh/get_started/agent.md (模块 文档; 类别 docs; 类型 documentation): 中文版 Agent RL 路线图, 与英文版同步翻译, 便于中文用户阅读。
- docs/en/get_started/customization.md (模块 文档; 类别 docs; 类型 documentation; 符号 custom_generate): 修改 custom_generate 签名并新增 fan-out 返回多个样本的详细说明与代码示例, 是 agent 集成的重要契约。
- docs/zh/get_started/customization.md (模块 文档; 类别 docs; 类型 documentation; 符号 custom_generate): 中文版 customization 更新, 与英文同步。

- docs/en/index.rst (模块 文档; 类别 docs; 类型 documentation) : 将 agent.md 和 coding_agent_rl README 加入导航索引, 确保文档可见。
- docs/zh/index.rst (模块 文档; 类别 docs; 类型 documentation) : 中文索引同步更新。

关键符号: custom_generate

关键源码片段

docs/en/get_started/customization.md

修改 custom_generate 签名并新增 fan-out 返回多个样本的详细说明与代码示例, 是 agent 集成的重要契约。

```
import copy

from slime.utils.types import Sample

async def custom_generate(args, sample: Sample, sampling_params: dict) -> list[Sample]:
    # 假设 run_agent_and_split_segments 是你的 agent 执行逻辑,
    # 返回一个 segment 列表, 每个 segment 包含 tokens、response 等。
    segments = await run_agent_and_split_segments(args, sample, sampling_params)

    # 使用原始 sample 的 rollout_id (如果存在), 否则用 index 作为 ID。
    rollout_id = sample.rollout_id if sample.rollout_id is not None else sample.index

    samples: list[Sample] = []
    for segment in segments:
        # 浅拷贝 sample 可以保留原始 meta 信息 (如 prompt、rollout_id)
        s = copy.copy(sample)
        s.tokens = segment.tokens
        s.response = segment.response
        s.response_length = segment.response_length
        s.loss_mask = segment.loss_mask
        s.reward = segment.reward
        s.status = Sample.Status.COMPLETED
        s.rollout_id = rollout_id # 所有 sibling samples 使用相同的 rollout_id
        samples.append(s)
    return samples
```

- 关键: 共享 rollout_id 确保 slime 在训练 step 切分和 loss 聚合时, 把这些样本当作同一次 rollout, 不会重复计数。
- 如果完整 trajectory 只有一个总奖励但被分成了 K 个片段, 常见做法是分配 reward / K 到每个片段, 避免奖励被放大。

评论区精华

该 PR 无 review 讨论和评论。

- 暂无高价值评论线程

风险与影响

- 风险：纯文档变更，未修改任何源码或配置，无运行时风险。但中文版 agent.md 中“router 参数通过 --router-* 传入”的表述可能存在歧义，需确保与现有 CLI 文档一致。
- 影响：对用户：为尝试 Agentic RL 的用户提供清晰的切入路径和最佳实践，降低上手成本。对系统：无影响。对团队：统一了 Agent 相关文档口径，减少重复解释。影响范围：所有阅读文档的用户。
- 风险标记：无运行时风险

关联脉络

- PR #1956 Add slime/agent/ and move sandbox impl inside: 将 sandbox 移至核心模块，为 agent 文档提供架构基础。
- PR #1923 [examples] add coding_agent_rl: agent-in-sandbox RL minimal demo: 提供了 agent RL 的完整示例，文档中多处引用该示例。