

PR #2369 完整报告

radixark/miles

fix(rollout): normalize rewards per rollout

合并时间: 2026-08-13 00:35

原文链接: <http://prhub.com.cn/radixark/miles/pull/2369>

执行摘要

- 一句话: 按 rollout 归一化奖励, 修复 session v2 变扇出统计偏差
- 推荐动作: 值得精读。重点看 `_reward_group_segments` 的三级 fallback 和 `_normalize_rewards_by_rollout` 的主流样本假设, 这是本 PR 最核心的设计决策; 建议配合 `test_train_data_conversion.py` 中新用例阅读, 理解从旧行级归一化到新 rollout 级归一化的契约变化。若要改动主流选择策略, 需同步审视 `_trainable_token_count` 对 `remove_sample` 的处理。

功能与动机

PR body 描述的症状: `Variable session-v2 leaf fanout gives one rollout multiple segment rewards, so row-level normalization weights prompt statistics by segment count and lacks one rollout target`. 即同一 rollout 因叶子数不同, 在按行归一化时一个 rollout 会被重复计入统计, 且缺少单一代表奖励。同时两个 CI 任务失败: `v2 agentic samples require input Sample.rollout_id or Sample.index` 与 `sample.rollout_id is not bitwise equal`, 说明 session v2 已开始传播 rollout 身份而测试夹具没有跟上。修复目标正如标题所述: `Normalize one mainstream reward per rollout without weighting statistics by leaf fanout`.

实现拆解

1. 恢复 prompt 边界: 新增 `_reward_group_segments` (`miles/ray/rollout/train_data_conversion.py`), 按 `prompt_group_sizes` (Multi-LoRA 显式边界) → 完整 `group_index` → 旧版固定 fanout 连续布局 → 单组兜底的优先级还原每个 prompt 的行范围, 替换旧逻辑里依赖 `n_samples_per_prompt * rollout_batch_size` 的 `reshape` 推断。
2. 每个 rollout 选主流样本: 新增 `_trainable_token_count` 与 `_normalize_rewards_by_rollout`. 在 prompt 组内用 `rollout_id` (或 `index`) 聚合兄弟样本, 选择 `_trainable_token_count` 最大的样本作为主流, 取其原始奖励参与组内零均值 (及可选 std 归一化), 然后把归一化后的 `advantage` 广播给同 rollout 的所有叶子; 无任何身份时每行独立成组。
3. 替换旧归一化逻辑: `_post_process_rewards` 删除原来的 `reshape/view/split` 分组分支, 统一委托给 `_normalize_rewards_by_rollout`; `estimator` 白名单 (`grpo/gspo/reinforce_plus_plus_baseline`) 与 `grpo_std_normalization` 开关逻辑保持兼容。

4. 测试与夹具对齐: tests/fast/ray/rollout/test_train_data_conversion.py 新增主流选择、平局取首个、兄弟广播、prompt 边界恢复、空输入等 8+ 用例;
tests/fast/rollout/generate_hub/test_multi_turn.py 为 agentic variant 补 index=0 并在结构比较时剥离 index/rollout_id 字段; tests/fast/router/test_session_v1_v2_parity.py 的 fixture 显式设置 rollout_id=index。无配置、schema 或部署配套改动。

关键文件:

- miles/ray/rollout/train_data_conversion.py (模块 奖励归一化; 类别 source; 类型 core-logic; 符号 _reward_group_segments, _trainable_token_count, _normalize_rewards_by_rollout, _post_process_rewards) : 核心修复文件: 新增 _reward_group_segments、_trainable_token_count、_normalize_rewards_by_rollout, 并重写 _post_process_rewards, 把奖励归一化从行级改为按 rollout 主流样本归一化后广播。
- tests/fast/ray/rollout/test_train_data_conversion.py (模块 数据转换; 类别 test; 类型 test-coverage; 符号 test_irregular_group_size_uses_explicit_group_index, test_grpo_normalizes_unique_rollouts_with_unequal_fanout, test_grpo_broadcasts_std_normalized_rollout_advantage, test_grpo_uses_most_trainable_sibling_reward_for_rollout) : 新增 8+ 个针对新归一化语义的用例, 覆盖主流选择、平局、广播、prompt 边界恢复、缺失身份兜底、空输入等, 是理解新契约的入口。
- tests/fast/rollout/generate_hub/test_multi_turn.py (模块 多轮生成; 类别 test; 类型 test-coverage) : session v2 agentic 生成测试需要显式设置 index 并在结构比较中剥离身份字段, 否则会与生产契约冲突。
- tests/fast/router/test_session_v1_v2_parity.py (模块 会话对齐; 类别 test; 类型 test-coverage) : v1/v2 对齐测试的 fixture 需要显式设置 rollout_id=index, 并把 rollout_id 纳入 bitwise 对齐断言, 匹配 session v2 生产行为。

关键符号: _reward_group_segments, _trainable_token_count, _normalize_rewards_by_rollout

评论区精华

本 PR 没有 line-level review 评论, guapisolo 直接批准 (APPROVED)。核心设计权衡体现在 6 个提交的演进里: 第一个提交尝试“保留每个 leaf reward、按等权聚合再变换”; 随后 guapisolo 改为“每个 rollout 选主流样本”, 并明确假设“最长可训练兄弟最能代表 rollout”; 再逐步补齐命名 (segment 术语)、文档 (fallback 说明) 和测试夹具对齐。这说明该改动经历过一轮内部设计迭代后收敛, 而非一次性拍板。

- 暂无高价值评论线程

风险与影响

- 风险: 数值语义变化: 所有启用 rewards_normalization 的 GRPO/GSPO/Reinforce++ 任务, advantage 计算从“行级”改为“rollout 级”, 训练曲线会与旧版本不同, 属于预期内但影响面广。主流样本假设: 若最长可训练段的奖励恰是异常值 (如被截断或含噪声), 该

rollout 的整体 advantage 会被带偏，目前没有对主流选择的置信度或回退机制。零方差行为变化：旧代码在 $\text{std}=0$ 时仍除以 $1e-6$ 产生无意义放大，新代码跳过缩放，结果更合理但可能与旧快照不一致。兜底路径：无 rollout_id/index/group_index 的旧数据会走“每行独立”或“整体一组”的兜底，行为与旧版不同，迁移老数据时需验证。回归保护：新的 8+ 用例覆盖主流选择、平局、广播、边界恢复与空输入，CI 全绿，风险可控。

- 影响：影响所有启用 rewards_normalization 的 GRPO/GSPO/Reinforce++ 训练任务，尤其 session v2 agentic 多叶子和 Multi-LoRA 显式分组的场景；行级训练数据结果与历史版本数值不同（语义上更合理）。对用户而言，同一 rollout 的多个叶子现在共享同一个 advantage，prompt 统计不再被扇出倍数放大；对团队而言，reward 语义从“按行”变为“按 rollout 单元”，需要同步更新相关文档和基于旧数值的基准对比。
- 风险标记：核心训练路径数值变更，主流样本启发式假设，旧数据兜底行为差异，测试覆盖充分

关联脉络

- PR #2368 fix(rollout): group session v2 leaf samples: 直接相关：该 PR 让 session v2 多叶子样本共享 rollout_id 并补失败保护，本 PR 正是基于这一 identity 做按 rollout 归一化；PR body 也提及 agentic_tool_call.generate 现在会传播 v2 rollout 身份。
- PR #2278 feat(session): request and assemble additional R3 rows under in-place weight updates: 关联触发场景：该 PR 引入 session v2 增量请求 R3 行的数据形态，使同一 rollout 出现多行（多叶子），为本次修复提供了触发条件。