

PR #2013 完整报告

THUDM/slime

Revert "rename rollout_ids to group_ids"

合并时间: 2026-06-04 09:43

原文链接: <http://prhub.com.cn/THUDM/slime/pull/2013>

执行摘要

- 一句话: 撤销 rollout_id 到 group_id 的重命名
- 推荐动作: 建议阅读者注意该 PR 与 #1984 的关联, 确保后续开发基于正确的命名。同时关注合并冲突的解决。如果之前使用了 group_id 的新 API, 需要及时迁移回 rollout_id。

功能与动机

PR #1984 将 rollout_id 重命名为 group_id, 但该命名与内部工具存在冲突, 因此需要撤销该变更以恢复正常集成。

实现拆解

该 PR 通过 git revert 撤销了 #1984 的提交, 恢复所有文件中的旧命名。主要包括:

1. slime/utils/types.py: Sample 数据类的字段从 group_id 改回 rollout_id, 并移除了 __getattribute__ 和 __setattr__ 中的向后兼容逻辑 (允许 rollout_id 写访问并发出弃用警告)。
2. slime/ray/rollout.py: 验证函数从 _validate_group_id_annotated 改回 _validate_rollout_id_annotated, 训练数据字典的键从 group_ids 改回 rollout_ids, 相关注释同步更新。
3. slime/utils/dp_schedule.py: 函数参数 group_indices 改回 rollout_indices, 文档和注释恢复。
4. slime/backends/megatron_utils/data.py 等后端文件: 字典键和注释回退。
5. 测试文件: tests/test_sample.py 和 tests/test_dp_schedule.py 恢复对 rollout_id 的测试, 移除针对 group_id 兼容性的测试。

关键文件:

- slime/ray/rollout.py (模块 推理执行; 类别 source; 类型 core-logic; 符号 _validate_group_id_annotated, _validate_rollout_id_annotated): 核心逻辑文件, rollout 数据生成和转换的主路径, 验证函数和训练数据键的变更直接影响训练流程。
- slime/utils/types.py (模块 数据模型; 类别 source; 类型 dependency-wiring; 符号 getattribute, setattr): 定义了核心数据类 Sample, 字段名变更直接影响整个数据流; 移除了向后兼容的 __getattribute__ 和 __setattr__ 方法。

- tests/test_sample.py (模块 样本测试; 类别 test; 类型 test-coverage; 符号 test_group_id_accepts_legacy_rollout_id_assignment_only, test_from_dict_accepts_legacy_rollout_id_without_group_id, test_from_dict_prefers_group_id_over_legacy_rollout_id, test_group_id_does_not_serialize_legacy_rollout_id_alias) : 测试文件, 验证 Sample 类的序列化和兼容性, 随着字段变更而回退。
- tests/test_dp_schedule.py (模块 调度测试; 类别 test; 类型 test-coverage; 符号 test_grouping_keeps_samples_together, test_rollout_grouping_keeps_samples_together, test_trims_trailing_groups_that_dont_fill_a_step, test_trims_trailing_rollouts_that_dont_fill_a_step) : 测试 dp 调度逻辑, 参数名从 group_indices 改回 rollout_indices。
- slime/utils/dp_schedule.py (模块 调度器; 类别 source; 类型 core-logic) : 调度核心逻辑, 函数参数和文档从 group_indices 改回 rollout_indices。

关键符号: _validate_group_id_annotated, _validate_rollout_id_annotated, getattr, setattr, test_group_id_accepts_legacy_rollout_id_assignment_only, test_from_dict_accepts_legacy_rollout_id_without_group_id, test_from_dict_prefers_group_id_over_legacy_rollout_id, test_group_id_does_not_serialize_legacy_rollout_id_alias

关键源码片段

slime/ray/rollout.py

核心逻辑文件, rollout 数据生成和转换的主路径, 验证函数和训练数据键的变更直接影响训练流程。

```
# slime/ray/rollout.py 中 _get_rollout_data 方法片段

# Enforce the rollout_id contract before flattening: any list[Sample]
# encountered in the nested output must have rollout_id set on every
# element. Default rollouts inherit it from the data source; compact /
# subagent paths that split one rollout into N training samples must
# set the same rollout_id on every sibling so the loss reducer counts
# the rollout once instead of N times.
_validate_rollout_id_annotated(data) # 恢复为 rollout_id 验证
```

slime/utils/types.py

定义了核心数据类 Sample, 字段名变更直接影响整个数据流; 移除了向后兼容的 __getattr__ 和 __setattr__ 方法。

```
# slime/utils/types.py 中 Sample 数据类的关键部分

@dataclass
class Sample:
    """The sample generated"""

    group_index: int | None = None
```

```
index: int | None = None
# Id of the rollout this sample came from. Defaults to ``None`` and the
# downstream pipeline falls back to ``index`` (so the default rollout
# path, where one execution = one training sample, sees rollout_id ==
# index). Compact / subagent paths that split one rollout execution into
# multiple training samples should set the same ``rollout_id`` on every
# sibling, so loss aggregation averages within the rollout instead of
# over-counting it.
rollout_id: int | None = None # 恢复为 rollout_id, 不再使用 group_id
# 其他字段省略 ...

# 移除了 __getattr__ 和 __setattr__ 中的 rollout_id 兼容处理,
# 因为 rollout_id 现在作为普通字段存在, 不再需要特殊处理。
```

评论区精华

该 PR 没有 review 评论。根据 PR body, 撤销原因是 [as there is conflict to internal tools](#), 即与内部工具冲突。

- 暂无高价值评论线程

风险与影响

- 风险：恢复旧命名可能带来以下风险：
 - 如果已有数据（如调试数据、checkpoint）使用了新的 group_id 字段，加载时可能出现兼容性问题。
 - 如果其他待合并的 PR 依赖了 group_id，则会产生合并冲突。
 - 回退操作本身可能引入疏漏，例如某些注释或日志未完全还原，需要仔细审计。
- 影响：影响范围：
 - 所有基于 slime 开发的使用者需要将代码中的 group_id 引用改回 rollout_id。
 - 内部工具可以正常集成。
 - 测试覆盖恢复到重命名前的状态，相关测试全部回退。
 - 风险标记：恢复命名导致兼容性问题，与其他待合并 PR 冲突风险，回退疏漏可能残留 group_id 引用

关联脉络

- PR #1984 rename rollout_ids to group_ids: 该 PR 是 #1984 的 revert, 直接撤销了 #1984 的变更。