

PR #2184 完整报告

THUDM/slime

sync source_names

合并时间: 2026-07-06 18:49

原文链接: <http://prhub.com.cn/THUDM/slime/pull/2184>

执行摘要

- 一句话: 新增 `source_names` 追踪数据来源
- 推荐动作: 建议合并。变更清晰、影响范围小, 文档完善。值得关注的是 `get_source` 的优先级设计, 为未来扩展提供了灵活性。

功能与动机

当训练混合多个数据源时, 需要追踪每个样本的来源, 以支持自定义奖励、过滤、日志统计以及未来的按源路由 (如 OPD teacher 选择)。PR body 虽未提供详细 issue, 但文档中明确说明了该动机。

实现拆解

1. `slime/utils/data.py`: 新增 `get_source(sample: Sample) -> str` 函数, 按优先级提取 `source`: 优先返回 `sample.source` (动态设置), 其次 `metadata["source_name"]`, 否则返回 `"unknown"`。将该函数加入 `__all__` 导出。
2. `slime/ray/rollout.py`: 导入 `get_source`; 在 `_convert_samples_to_train_data` 中, 当 `samples[0].metadata` 不为 `None` 时, 为每个 `sample` 调用 `get_source`, 将结果列表存入 `train_data["source_names"]`; 在 `_split_train_data_by_dp` 的 `skip` 列表中添加 `"source_names"`, 避免被误当作指标处理。
3. `slime/backends/megatron_utils/data.py`: 在 `log_rollout_data` 的 `skip` 列表中添加 `"source_names"`, 防止非张量字段被尝试聚合日志。
4. 文档更新: 中英文 `usage.md` 均增加了关于 `source_name` 的说明和示例 JSON 配置, 并清晰列出了查找优先级和用途。

关键文件:

- `slime/utils/data.py` (模块 数据工具; 类别 `source`; 类型 `core-logic`; 符号 `get_source`): 核心新增 `get_source` 函数, 定义 `source` 提取逻辑和优先级。
- `slime/ray/rollout.py` (模块 `Rollout`; 类别 `source`; 类型 `dependency-wiring`): 导入并使用 `get_source`, 在训练数据中注入 `source_names`, 同时修改 DP 拆分逻辑跳过该字段。
- `slime/backends/megatron_utils/data.py` (模块 训练后端; 类别 `source`; 类型 `core-logic`): 在日志函数中跳过 `source_names`, 避免非张量字段处理出错。

- docs/en/get_started/usage.md (模块 文档; 类别 docs; 类型 documentation) : 英文文档说明 source_name 配置方式和查找优先级。
- docs/zh/get_started/usage.md (模块 文档; 类别 docs; 类型 documentation) : 中文文档同步更新。

关键符号: get_source

关键源码片段

slime/utils/data.py

核心新增 get_source 函数, 定义 source 提取逻辑和优先级。

```
def get_source(sample: Sample) -> str:
    # 优先返回动态设置的 sample.source
    if getattr(sample, "source", None):
        return sample.source
    # 其次从 metadata 中取 source_name
    metadata = getattr(sample, "metadata", None) or {}
    if metadata.get("source_name"):
        return metadata["source_name"]
    # 都没有时标记为 unknown
    return "unknown"
```

评论区精华

本 PR 无 review 评论或讨论。

- 暂无高价值评论线程

风险与影响

- 风险: 风险较低。新增字段仅在 metadata 非空时写入, 不影响现有逻辑。get_source 对未设置 source 的样本返回 "unknown", 兼容旧数据。需注意: 如果其他模块遍历 train_data 时未处理 "source_names", 可能导致意外行为, 但目前 skip 列表已覆盖日志和 DP 拆分路径。
- 影响: 对用户: 需在 metadata 中设置 source_name 或动态设置 sample.source 才能生效; 未配置时自动回退为 "unknown"。对系统: 新增 ~44 行代码, 无性能影响。对团队: 为后续按源路由、分源统计等功能奠定基础。
- 风险标记: 缺少测试覆盖

关联脉络

- 暂无明显关联 PR