

THUDM/slime 2026 第 26 周周报 (06-22 至 06-28)

THUDM/slime

周期: 2026-06-22 至 2026-06-28

来源 PR: 4 • 重点 PR: 4 • 自动生成

原文链接: <http://prhub.com.cn/THUDM/slime/reports/2026-06-22-to-2026-06-28>

执行摘要

本周 (06-22 至 06-28) 共合并 4 个 Pull Request, 涉及 Agent 系统、Rollout 层和后端架构。核心主线是提升 Coding-Agent RL 的工程健壮性与可扩展性, 同时通过内部代码同步引入 Teacher Server 丰富训练模式。两个作者 (jingshenghang、zhuzilin) 各贡献 2 个 PR, 分工明确。重点 PR 包括 Agent 多后端选择、稳定性修复、监控修复以及 Teacher Server 集成。

本周重点变化

- Agent 多后端选择: PR #2125 通过 SWE_AGENT 环境变量引入注册表模式, 解耦 Agent Harness 与 Adapter, 支持 claude_code 与 codex 两种实现, 默认保持兼容。
- 稳定性修复: PR #2124 解决 SWE coding-agent RL 训练中会话清理不当、异常处理不完善等问题, 新增 drop_session 方法分离清理与样本生成, 并设置 rollout_log_probs=[0.0] 及 remove_sample=True 优雅处理中止样本。
- Rollout 监控修复: PR #2121 删除 sglang_rollout 中因 reward 类型非浮点 (int/dict) 而崩溃的平均 reward 统计逻辑, 但移除了相关监控展示。
- 架构同步: PR #2118 从内部同步引入 Megatron Teacher Server (独立 HTTP 预填充与 logprob 服务) 和无状态 Adam 优化器, 重构 checkpoint 加载流程, 并改进 trace 日志。

模块与主题趋势

- Agent 模块: 本周变动最大, 涉及配置读取 (SWE_AGENT)、会话管理 (drop_session)、异常处理 (abort_result), 体现工程化打磨趋势。
- Rollout 模块: 仅有一处监控统计删除, 说明该模块整体趋于稳定, 但 reward 类型处理仍需更健壮的方案。
- 后端架构: Teacher Server 的引入标志着项目开始支持 teacher-forcing 模式, 架构从单一 rollout 向多服务协同演进, 但新组件缺少负载测试与监控集成。
- 测试覆盖: 本周所有 PR 均未附带新增测试, 尤其 PR #2124 涉及核心路径变更, 但无测试验证, 风险较高。

风险观察

1. 核心路径变更无测试覆盖: PR #2124 和 #2118 修改了 agent 训练核心逻辑和后端架构, 缺乏自动化测试, 回归风险较大。建议针对关键流程补充集成测试。

2. 监控指标缺失：PR #2121 删除了平均 reward 统计，导致该监控盲点。后续应考虑引入健壮的类型转换或统一 reward 格式。
3. Teacher Server 运维复杂性：新组件作为一个独立的 HTTP 服务，增加了部署和调试难度，需要明确维护职责与监控告警。
4. 无状态 Adam 行为差异：StatelessAdam 不维护动量状态，适用于特殊场景，但可能对收敛行为产生影响，需要实验验证。
5. 环境变量配置管理：SWE_AGENT 等环境变量决定了组件实例化，未设校验或提供清晰错误提示时，错误配置可能导致难以诊断的启动失败。

重点 PR 速览

- PR #2118 (feat: sync from internal)：重要性 9.4，是本周最重要的 PR。它引入了 Megatron Teacher Server、StatelessAdam、优化 logprob 计算（避免全词汇表 gather）等。关键函数包括 `sample_from_vocab_parallel_logits_without_full_gather` 和 `StatelessAdam.step`。建议重点学习其高效采样算法和 teacher-forcing 架构。
- PR #2124 (fix: SWE coding-agent RL stability bugs)：重要性 6.3，修复了多项稳定性问题。关键变更：新增 `drop_session`、设置中止样本处理、修复 `int('')` crash。虽然无测试覆盖，但改动经过充分思考，建议精读异常处理模式。
- PR #2125 (feat: select agent via SWE_AGENT)：重要性 5.3，实现 agent 选择扩展。采用注册表 + 环境变量模式，简单优雅，易于扩展。适合作为多后端实现的参考样例。
- PR #2121 (fix: non-float reward bug)：重要性 5.3，修复了非浮点 reward 导致的 crash。但删除了监控统计，后续应重建健壮的监控方案。

后续建议

1. 追加测试：为 PR #2124 和 #2118 补写集成测试，尤其是 agent 异常路径和 Teacher Server 的基本请求响应测试。
2. 恢复并增强 Reward 监控：设计一个统一 reward 类型转换或使用 try-except 的健壮统计逻辑，避免直接删除监控。
3. 建立 Teacher Server 运维规范：包括健康检查、日志采集和性能基准，确保新组件的可靠性。
4. 环境变量校验：在读取 SWE_AGENT 等环境变量时增加有效值校验和日志输出，帮助快速定位配置错误。
5. 关注无状态 Adam 效果：在特定场景（如 sparse gradient 或 online learning）下验证 StatelessAdam 的表现，并更新优化器选型建议。