

verl-project/verl 2026 年第 29 周周报 (07-13 至 07-19)

verl-project/verl

周期: 2026-07-13 至 2026-07-19

来源 PR: 28 • 重点 PR: 24 • 自动生成

原文链接: <http://prhub.com.cn/verl-project/verl/reports/2026-07-13-to-2026-07-19>

执行摘要

本周共处理 28 个 PR (重点 24 个), 聚焦于 Ascend NPU 兼容性修复、训练核心功能增强及可观测性建设。NPU 相关 PR 占比 43%, 体现了对国产硬件生态的持续投入。重要新功能包括分片增量权重同步 (#6974)、RL-Insight 可观测性 (#6680)、LoRA-only checkpoint (#7061) 以及流式数据加载与异步恢复 (#7037)。多项严重 bug 得到修复, 如 MoE 指标偏差 (#7075)、Megatron CP>1 梯度爆炸 (#6836) 和视觉占位符 token 采样 (#7038)。整体上, 本周为即将到来的 v0.9 版本铺平了道路, 但部分 review 遗留问题需在后续迭代中解决。

本周重点变化

- NPU 兼容性提升: 恢复因重构丢失的 vLLM 补丁 (#7076), 修复注意力掩码构建 OOM (#7055), 更新 Docker 镜像和 CI 脚本 (#7035、#7081), 同步 Qwen3.5 Dockerfile (#7028)。同时清理了过时的 NPU 补丁代码 (#7042), 降低维护成本。
- 训练性能与功能增强: 引入分片 delta 权重同步引擎 (#6974), 通过字节级 diff 和稀疏收集将权重同步速度提升 2-3 倍; 支持流式数据加载 (#7037), 使 `gen_batch_size` 独立于 `train_batch_size`, 并基于 `TransferQueue` 实现训练器检查点恢复。
- 监控与存储优化: 集成 RL-Insight 可观测性框架 (#6680), 支持 Grafana 展示训练指标和 RL 状态; LoRA-only checkpoint (#7061) 将 27B 模型检查点从 54GiB 压缩至 150MiB。
- 严重 bug 修复: MoE 负载均衡指标排除最后一个未路由 token (#7075); Megatron CP>1 下 MoE 辅助损失梯度爆炸 (#6836); 屏蔽视觉占位符 token 防止训练崩溃 (#7038); 修复 Qwen3 XML 工具解析器对畸形结构的崩溃 (#7044)。

模块与主题趋势

本周标签统计显示 bugfix 居首 (12 次), NPU、CI、rollout 模块改动频繁。rollout 模块涉及工具调用、视觉 token 屏蔽和权重同步, 多处修改集中于 `vllm_async_server.py`。NPU 模块有多达 12 个相关 PR, 但仍有配置兼容性风险未解决。CI 相关 PR (8 个) 持续优化, 表明团队重视测试可靠性。checkpoint 领域同时新增 LoRA-only 和流式恢复能力, 成为基础设施亮点。总体来看, 项目正从核心功能开发转向稳定性增强和生产化准备, 异步训练和可观测性是明显的发展方向。

风险观察

- PR#7035 在更新 Docker 镜像的 shell 脚本中设置 `enforce_eager=False` 且启用 `cuda_graph_mode`，与 Ascend NPU 不兼容，可能导致 vLLM 初始化失败，该问题未修复即合并。
- PR#7032 的 SkipManager 在 `parameter_sync_step=1` 时破坏向后兼容，且中断运行可能引起数据污染，review 指出的问题未解决。
- PR#7024 的中英文 profiling 文档存在命令示例引号缺失和章节编号错误，用户直接复用可能失败。
- PR#6473 对 DeepSeek V4 的 MTP drafter 权重准备覆盖未明确验证，存在遗漏风险。
- PR#7042 移除旧 NPU 补丁后，依赖旧补丁的极端配置可能未完全清理，可能导致运行时错误。

重点 PR 速览

- #6974 分片 delta 权重同步：通过字节级 diff 和 NCCL 稀疏收集，仅传输变化的参数分片，在 Qwen2.5-7B/32B/72B 上取得 1.9-3.1 倍加速，是分离式异步训练的关键组件。
- #7061 LoRA-only checkpoint：新增 `save_lora_only` 选项，自动过滤非 LoRA 权重，节省约 98% 存储空间，支持自动检测与向后兼容，设计清晰。
- #6680 RL-Insight 集成：低入侵的可观测性框架，通过环境变量控制，零成本禁用，支持训练指标和 RL 状态 trace，适合生产监控。
- #7075 MoE 负载均衡指标修复：排除序列末尾未路由的 EOS/token，修正专家 0 负载虚高，确保负载均衡监控准确性。
- #6836 Megatron CP>1 梯度爆炸修复：计算实际路由 token 数并正确归一化 MoE 辅助损失，配合防护逻辑阻止不合理配置，对长上下文训练至关重要。
- #7037 流式数据加载与异步恢复：支持 `gen_batch_size` 独立配置、drop 与 refill 策略、基于 TransferQueue 的检查点恢复，显著提升训练灵活性和弹性。

后续建议

1. 优先修复 NPU 配置兼容性问题：PR#7035 中 `enforce_eager` 设置需立即修正，避免 CI 阻塞；同时对齐 shell 脚本中的并行策略（如 #6973 中的 EP/DP 配置）。
2. 完善 review 遗留问题：对 SkipManager（#7032）增加向后兼容和清理逻辑；确认 MoE 指标修复（#7075）后监控阈值是否需要调整。
3. 补充新模型测试：为 DeepSeek V4 增加 MTP drafter 专项测试，确保权重传输全覆盖。
4. 建立 CI 模板规范：避免硬编码路径（如 #6973 的 `/root`），使用环境变量或通用占位符，提升脚本可移植性。
5. 强化文档验证：在 CI 中加入文档示例的语法检查，防止命令行引号缺失等问题流入主分支。