verl-project/verl 2026 第 26 周周报 (06-22 至 06-28)

verl-project/verl

周期：2026-06-22 至 2026-06-28

来源 PR：22・重点 PR：22・自动生成

原文链接：http://prhub.com.cn/verl-project/verl/reports/2026-06-22-to-2026-06-28

1. 执行摘要

本周（2026-06-22 至 2026-06-28）verl 项目共合并 22 个 PR，重点集中在算法训练能力增强、Ascend NPU 平台支持深化以及基础设施鲁棒性提升。最值得关注的变化包括：连续 Token 机制的引入为多轮 Agentic Rollout 提供了统一 tokenization 方案；完全确定性支持的实现使得训练过程可复现；V1 PPO trainer 正式默认启用（breaking change）。此外，NPU 补丁按模型隔离重构、多个关键 bugfix 以及 Ascend CI 与文档的持续改善，共同推动了项目在功能完整性和工程质量上的稳步前进。

2. 本周重点变化

- 连续 Token 机制（#6779）：该 PR 设计了多模型族的 ContinuousTokenBuilder 架构，通过中央化构建器替代散落在 AgentLoop 中的手动拼接，确保多轮对话中 token ID 与元数据的连续性和对齐。当前默认关闭，待多模态支持后转为默认。
- 完全确定性支持（#6572）：为 vLLM rollout 和 reward model 新增 full_determinism 配置，通过设置确定性环境变量、priority 调度和 hash 路由，实现端到端可重现训练。这是对调试和科学可重复性需求的重要回应。
- V1 Trainer 默认启用（#6823）：将 V1 PPO trainer 设为默认，涉及配置键名统一（model_config→model）和参考策略变量引用修正。用户需在配置中显式设置 use_v1=False 以沿用旧行为。讨论中暴露了 temperature 参数传递的职责边界问题，建议后续统一清理。
- NPU 补丁按模型隔离重构（#6777）：将模块级急切导入改为按模型懒加载，每个模型补丁独立捕获异常，避免单模型缺失导致全部补丁失效。这是 Ascend 平台可靠性的一次重要提升。
- Megatron 优化器状态对齐（#6526）：新增 use_precision_aware_optimizer 配置项，允许将优化器状态和 DDP 梯度桶降为 bf16 以降低显存，默认仍为 fp32。经过收敛性实验验证后，作为可选功能发布。

3. 模块与主题趋势

- Ascend NPU 生态：本周是 NPU 相关 PR 数量最多的一周（9 个），涵盖示例脚本、Docker 镜像、CI 修复、文档更新和功能新增。这表明 Ascend 平台正在快速迭代并逐步成为一等公民。

- 训练器与算法：V1 trainer 默认化标志着 v0.8.0 即将到来。连续 Token、确定性支持和共置奖励模型等特性均围绕训练体验和可复现性展开，体现了对研究严肃性的重视。
- 基础设施与 CI：CI 工作流和 Dockerfile 的频繁调整（如共享内存增大、版本修复）反映出项目在多节点、多硬件环境下的持续磨合。nightly CI 基线检查的引入加强了回归检测。
- Bugfix 密集：本周 7 个 bugfix 覆盖了指标死指标、布局崩溃、日志偏移、配置 typefos 等，表明项目正处于积极扫除遗留问题的高频率开发阶段。

4. 风险观察

- 核心路径变更风险：连续 Token、默认 V1 trainer、NPU 补丁重构和完全确定性均属于核心路径变更，可能对现有工作流产生破坏性影响。尤其是 V1 trainer 的配置键名变更需要用户主动适配。
- 缺少测试覆盖：完全确定性支持和 Megatron 优化器对齐等 PR 明确标注缺少自动化测试覆盖，后续需补充单元测试以防止回归。
- 版本错配风险：Docker 镜像和文档中曾出现版本号错误（如将 v0.8.0 误标为 v0.7.1），说明发布流程中仍存在人为疏忽，建议引入自动化版本检查机制。
- 共置奖励模型 GPU 内存竞争：separate_async 模式下显式禁止共置奖励模型，但仍可能在其他模式下引发内存分配问题，需持续监控。
- 配置变更频繁：多个 PR 涉及配置键名、默认值或类型注解的调整，用户升级时需仔细核对配置迁移指南。

5. 重点 PR 速览

PR 编号	标题要点	重要性	关键贡献
#6779	Continuous Token for Agentic Rollout	9.4	多轮 token 连续性构建器架构，默认关闭
#6572	Full determinism for vLLM rollout/RM	8.7	端到端可重现训练，环境变量与路由
#6526	Align optimizer states with model precision	8.5	Megatron 显存优化选项（可选）
#6777	Per-model NPU patches with fault isolation	8.4	按模型懒加载，错误隔离
#6818	Colocated reward model for V1 trainer	8.2	共置奖励模型评分实现

PR 编号	标题要点	重要性	关键贡献
#6823	Enable V1 trainer by default	6.8	Breaking change，配置键名统一
#6796	Fix aggregated metrics logging in fully_async	6.6	修复指标日志偏移
#6846	Fix fused log_probs/entropy in NO_PADDING	6.5	修复布局不匹配崩溃
#6842	Write HF config before bridge save	6.0	修复 Megatron 断点分片推断
#6807	Qwen3.5-122B long seq launch script for Ascend	5.2	4 节点 64 卡 GRPO 启动脚本

（注：按重要性降序排列，仅列出重要性≥5 的 PR）

6. 后续建议

1. 关注 V1 trainer 迁移：从本周开始，新用户默认使用 V1 trainer，现有 V0 用户应评估迁移成本，重点关注配置键名 change 和 refer 策略调整。
2. 验证确定性训练：对于需要严格复现的实验场景，建议尽早启用 full_determinism 并反馈潜在问题。
3. 评估连续 Token 机制：虽然当前默认关闭，但该机制将对多轮 Agent Loop 产生根本性影响，建议相关团队提前熟悉其设计和配置方式。
4. 补充测试覆盖：针对完全确定性、Megatron 优化器对齐等关键特性，应尽快补充自动化测试，确保长期可维护性。
5. 优化 Ascend CI：鉴于 NPU 平台活跃度持续上升，建议进一步优化 CI 基础架构（如调整共享内存、版本检查），并完善基线测试用例。
6. 版本管理与沟通：在即将发布的 v0.8.0 中，应确保变更日志清晰，特别是 breaking changes 的说明，并通过讨论或文档引导用户平滑升级。

以上为本周周报，建议团队周会上对重点 PR 和技术决策进行深入讨论。