

veRL 2026 年第 19 周周报 (05/04-05/10)

verl-project/verl

周期: 2026-05-04 至 2026-05-10

来源 PR: 34 • 重点 PR: 24 • 自动生成

原文链接: <http://prhub.com.cn/verl-project/verl/reports/2026-05-04-to-2026-05-10>

执行摘要

本周 (05/04-05/10) 共合入 34 个 PR, 工具、奖励、rollout 和 Megatron 四大模块均有重要改动。工具模块经历了大规模重构 (移除废弃组件、引入函数式注册、集中化初始化); 奖励模块统一配置并支持多输出异步评分; Rollout 引擎新增 TRTLLM 异步 RL、SGLang Prefill-Decode 分解和在线策略蒸馏; Megatron 修复了 CP position_ids 和 CP+TP 对齐等关键 bug。CI 与 NPU 支持持续增强, 但测试覆盖不足和配置兼容性仍是主要风险点。

本周重点变化

1. 工具系统大规模重构: PR #6302 移除课程采样器、动态数据集及多个废弃工具 (BREAKING), PR #6300 将工具初始化移至 AgentLoopWorker 并废弃 MCP, PR #6189 引入 `@function_tool` 装饰器, 支持 Python 函数自动注册。工具架构从分散走向统一, 降低了用户使用门槛。
2. 奖励模块统一与扩展: PR #6265 统一 RewardModelConfig 字段 (inference→rollout), 修复 world_size 计算; PR #6228 支持多输出轨迹异步评分, 扩展自定义 reward manager 接口; PR #6242 提取分数组装为类方法, 方便覆写。奖励模块的可维护性和灵活性显著提升。
3. Rollout 引擎异步化与多样化: PR #5631 为 TRTLLM 实现完整异步 RL (含 abort/resume); PR #6117 引入 SGLang Prefill-Decode 分解部署; PR #6056 集成多教师在线策略蒸馏。Rollout 层在异步和异构部署上迈出关键步伐。
4. Megatron 核心并行修复: PR #6267 修复 CP 场景 position_ids 计算错误; PR #6206 修正 CP+TP 序列对齐公式并适配新 MTP API。这些修复直接影响大规模训练的数值正确性。
5. CI 与 NPU 持续改进: 新增 GSPO Qwen3-30B NPU nightly CI (#6273); 修复内存可视化 NPU 兼容性 (#6216, #6248); 精简 TRTLLM CI (#6275)。基础设施保持活跃迭代。

模块与主题趋势

- 工具模块: 正经历从混乱到有序的重构。移除 MCP 和旧工具减少了维护负担, function_tool 降低了用户注册新工具的复杂度。但 BREAKING 变更需要下游适配, MCP 移除影响相关用户。
- 奖励模块: 围绕配置统一、多输出支持和接口扩展。配置字段不一致导致 world_size 计算反复 (#6226→#6258→#6265), 提醒模块间配置依赖需细粒度管理。

- Rollout: 异步、分解、蒸馏是三大关键词。TRTLLM 异步 RL 为大模型部署提供更灵活的中断 / 恢复能力, SGLang PD 分解支持更优化的 GPU 布局, OPD 扩展知识蒸馏场景。但 TRTLLM Docker 构建存在版本依赖风险。
- Megatron: 并行机制的精准修复是重点。position_ids 和对齐公式的修复直接影响训练稳定性, 但缺少测试覆盖, 需加强回归。
- CI/Infra: 效率优化 (任务合并、依赖精简) 和硬件兼容 (NPU, aarch64) 双线推进。多处依赖升级和移除, 需警惕对现有流水线的影响。

风险观察

1. 测试覆盖不足: 8 个 PR 明确标记缺少测试覆盖, 尤其 #6267 和 #6206 涉及核心路径, 回归风险高。
2. 配置兼容性断裂: #6302 的 BREAKING 变更和 #6300 的 MCP 移除可能影响用户现有配置和工具使用。
3. reward world_size 计算反复: 世界大小计算从 #6226 修复到 #6258 回滚再到 #6265 解决, 暴露了配置字段缺失的问题, 类似连锁反应可能重现。
4. TRTLLM Docker 构建风险: #6230 中 cupy 和 ray 版本可能存在, 若构建失败将阻塞 CI 和用户部署。
5. 多 output teacher logprobs 缺失: #6228 标记为 TODO, 影响使用 KL 惩罚的多输出蒸馏场景, 需跟进实现。

重点 PR 速览

- #6302 [BREAKING]: 移除废弃工具、课程采样器和动态数据集, 进行大方清理。
- #6300: 工具初始化集中化, 移除 MCP 支持, 统一工具注册表。
- #6189: 新增 @function_tool 装饰器, 支持 Python 函数自动注册为工具。
- #6265: 统一 RewardModelConfig 字段, 修复 reward world_size 计算。
- #6228: 支持多输出轨迹异步评分, 扩展自定义 reward manager。
- #5631: TRTLLM 异步 RL 完整实现 (abort/resume、KV cache 清理、profile)。
- #6117: SGLang Prefill-Decode 分解, 非对称 TP 布局。
- #6056: 在线策略蒸馏集成多教师, 完全异步框架扩展。
- #6267: 修复 Megatron CP position_ids 计算错误。
- #6206: 修复 CP+TP 对齐并适配新 MTP API。
- #6273: 新增 GSPO Qwen3-30B NPU nightly CI。
- #6248: 内存快照采集后自动清除历史, 提升调试效率。
- #6258: 回滚 reward world_size 计算 (临时, 待 #6265 修复)。
- #6230: 修复 TRTLLM Docker 和 CI, 但依赖版本存疑。

后续建议

- 对测试覆盖不足的核心修复 (如 #6267、#6206) 尽快补充集成测试。

- 在 release notes 中说明 #6302 和 #6300 的 BREAKING 变更，提供迁移指南。
- 验证 TRTLLM Docker 构建，确保 cupy 和 ray 版本有效，或调整变通方案。
- 推进 #6228 遗留的 teacher logprobs 实现，完善多输出训练能力。
- 继续保持工具和奖励模块的文档更新（如 #6287、#6272），改善新用户体验。