

PR #7517 完整报告

verl-project/verl

[ci] fix: enable fully sharded LoRA in Ascend E2E tests

合并时间: 2026-08-25 09:20

原文链接: <http://prhub.com.cn/verl-project/verl/pull/7517>

执行摘要

- 一句话: Ascend E2E 测试开启 fully sharded LoRA
- 推荐动作: 该 PR 是 CI 配置调整, 不值得精读, 但可关注其规避的 vLLM bug 及后续修复, 有助于理解 LoRA 分片在异构硬件上的兼容性问题。

功能与动机

PR body 明确指出启用 `fully_sharded_loras` 是为了绕过 vLLM 的 tensor-parallel shape bug, 详见 [vllm-project/vllm#47650](#)。该 issue 描述了在 AutoRound int4 模型上启用 `--enable-lora` 后, 在 TP>1 时因无条件 all_gather 导致非分片 LoRA 崩溃的问题。

实现拆解

1. 修改CI工作流文件 ([.github/workflows/e2e_ppo_trainer_megatron_vllm_2_ascend.yml](#))
: 在三个不同的 E2E 测试步骤中追加 `actor_rollout_ref.model.lora.fully_sharded_loras=True` 参数, 分别对应 GRPO LoRA、FSDP2 和 Qwen2.5-VL 场景。
2. 增加 NPU 健康检查: 在 'Check npu and CANN info' 步骤中, 将 `npu-smi info` 输出重定向到临时文件, 并通过 `grep` 检查是否存在 'Warning'、'Alarm'、'Critical'、'UNKNOWN' 等关键字段, 若命中则直接退出, 防止在异常硬件上运行测试。
3. 无额外测试或文档改动: 由于是 CI 配置调整, 未引入新的单元测试或文档更新。

关键文件:

- [.github/workflows/e2e_ppo_trainer_megatron_vllm_2_ascend.yml](#) (模块 CI 配置; 类别 infra; 类型 infrastructure): 唯一变更文件, 定义了多个 Ascend E2E 测试 job, 本 PR 为其追加 LoRA 分片参数并增加 NPU 健康检查。

关键符号: 未识别

评论区精华

Review 过程中没有评论, 仅 wucong25 审批通过。

- 暂无高价值评论线程

风险与影响

- 风险：回归风险：追加参数可能影响测试的 LoRA 行为，但该参数是为规避已知 bug，风险较低。健康检查误报风险：grep 匹配关键字可能因 NPU 输出格式变化导致误判。兼容性风险：依赖 vLLM 对 fully sharded LoRA 的支持，若底层版本变化可能失效。
- 影响：用户影响：仅影响运行 Ascend CI 的开发和测试流程。系统影响：确保 NPU 健康后运行测试，减少失败干扰。团队影响：需要关注相关 CI 测试结果，确保问题被正确规避。
- 风险标记：CI 配置变更，依赖 vLLM 行为

关联脉络

- PR #7538 [vllm, hardware] fix: support modular FusedMoE on NPU: 同仓库 NPU 与 vLLM 相关修复，涉及 NPU 启动问题。
- PR #7508 [vllm] fix: honor explicit False on Optional[bool] engine args in CLI serialization: vLLM 参数序列化修复，可能与 LoRA 参数传递相关。
- PR #7456 [ci] chore: Update ascend ci image: Ascend CI 相关调整，涉及 NPU 配置。