

PR #6085 完整报告

verl-project/verl

[ci] chore: add sgl_ascend cu for NPU

合并时间: 2026-04-28 10:14

原文链接: <http://prhub.com.cn/verl-project/verl/pull/6085>

执行摘要

- 一句话: 为 Ascend NPU 新增 sglang CI workflow
- 推荐动作: 该 PR 是基础设施常规变更, 不值得精读, 但值得关注其触发条件设计和镜像选择策略, 可作为后续新增 NPU workflow 的模板。

功能与动机

PR body 说明: "Add sglang ci yaml for Ascend NPU". 目的是为 NPU 上的 sglang 功能提供自动化测试, 填补此前仅有 vllm (vllm_ascend) 而无 sglang CI 的空白。

实现拆解

1. 创建 workflow 文件: 在 .github/workflows/ 下新增 sgl_ascend.yml, 定义名称为 sglang_ascend 的 CI workflow。
2. 配置触发条件: 初始为 push/pull_request 到 main/v0.* 分支, 并限定路径过滤; 根据 review 建议修改为类似 sgl.yml 的触发设置, 仅当 sglang 相关文件变化时触发。
3. 指定运行环境: 选用 linux-aarch64-a3-8 节点, 并使用 swr.cn-southwest-2 的 sglang 专用镜像 (verl-sglang-8.5.0-910b-ubuntu22.04-py3.11-latest), 确保 NPU 驱动和库版本匹配。
4. 定义 job 步骤: 包括环境准备、依赖安装、运行 sglang 单元测试和端到端测试; 超时设为 90 分钟。

关键文件:

- .github/workflows/sgl_ascend.yml (模块 CI 配置; 类别 infra; 类型 infrastructure): 唯一变更文件, 定义 Ascend NPU 上的 sglang CI workflow, 包括触发条件、运行环境、测试步骤。

关键符号: 未识别

评论区精华

reviewer @yyyy2000 提出四点意见:

- 删除不再需要的行 (原注释块中的冗余内容)。
- 容器镜像应使用 sglang 专用镜像, 而非 vllm 镜像 (引用具体镜像地址)。

- workflows 名称应为 `sglang_ascend`（原可能误用其他名称）。
- 触发条件需参考 `sgl.yml` 重写，并替换为 `sgl_ascend.yml` 自身路径。所有意见均在合并前被采纳，最终 workflow 配置符合现有 CI 规范。
- 删除不再需要的行 (other): 已采纳，删除冗余注释。
- 容器镜像应使用 `sglang` 专用镜像 (design): 已修改为指定镜像。
- workflows 名称应为 `sglang_ascend` (other): 已修改。
- 触发条件需类似 `sgl.yml` 并引用自身路径 (design): 已采纳，触发条件重写。

风险与影响

- 风险：低风险。新增的 CI workflow 仅影响 NPU CI 场景，不修改任何源代码。主要风险：触发条件配置不当可能导致误触发或漏触发；镜像版本与 NPU 环境不兼容导致测试失败；长时间运行占用 CI 资源。但这些问题通过 review 已缓解。
- 影响：用户：NPU 上的 `sglang` 用户可获得持续验证，减少回归风险。系统：增加 A3 节点上的 CI 负载（约 90 分钟 / 次），但限于 `sglang` 变更时触发。团队：需维护该 workflow，未来修改 `sglang` 代码时必须确保 NPU 兼容性。
- 风险标记：CI 资源消耗，配置错误风险

关联脉络

- PR #6120 [sglang] feat: Patch `sglang` to support on-policy distillation teacher: 涉及 `sglang` 功能扩展，该 PR 的 CI 覆盖将确保 NPU 兼容性。
- PR #6109 [sglang] feat: restrict abort and resume requests to primary server only: 另一个 `sglang` 相关 PR，新增的 CI workflow 可检测其在 NPU 上的回归。