

PR #36393 完整报告

sgl-project/sglang

[AMD][CI] Restore MiniMax-M2.5 4-GPU MI35x nightly job

合并时间: 2026-08-26 09:34

原文链接: <http://prhub.com.cn/sgl-project/sglang/pull/36393>

执行摘要

- 一句话: 恢复 MiniMax-M2.5 MI35x 夜间 CI 任务
- 推荐动作: 该 PR 值得快速浏览, 但无需深入阅读。它展示了如何安全地恢复 CI 任务, 并强调了在改动 CI 配置时进行逐字节对比和本地验证的重要性。

功能与动机

PR #36142 移除了 `nightly-4-gpu-mi35x-minimax-m25-rocm720` 任务, 将 MI35x 槽位让给 MiniMax-M3。本 PR 恢复 MiniMax-M2.5 任务, 使两个模型都能在各自独立的任务中获得夜间 MI35x 覆盖。

实现拆解

1. 定位变更入口: 修改 `.github/workflows/nightly-test-amd-rocm720.yml`, 在该工作流的 `on` 触发器的 `job_select` 下拉列表、`jobs` 定义和 `check-all-jobs` 的 `needs` 列表中重新添加 `nightly-4-gpu-mi35x-minimax-m25-rocm720` 任务。
2. 恢复任务块: 将任务块恢复到 #36142 移除前的原始位置 (紧接着 MiniMax-M3 任务之前)。任务体与之前版本逐字节一致, 包括相同的精度测试套件 `nightly-amd-4-gpu-mi35x-minimax-m25-tp4`、相同的 `aiter` 和 `aiter_unified_attn` 环境变量, 以及 120 分钟超时。
3. 更新下拉列表和 `needs` 依赖: 在 `job_select` 下拉列表和 `check-all-jobs` 的 `needs` 列表中添加相应条目, 确保任务可选且依赖关系正确。
4. 验证: 本地验证通过, 确认 YAML 语法正确、恢复的任务块与 #36142 前的版本逐字节一致, 且引用的精度测试套件 / 测试文件在 `main` 分支上仍然存在。

关键文件:

- `.github/workflows/nightly-test-amd-rocm720.yml` (模块 部署脚本; 类别 `infra`; 类型 `infrastructure`): 这是唯一修改的文件, 重新添加了 MiniMax-M2.5 4-GPU MI35x 夜间任务, 包括任务块、`job_select` 下拉条目和 `check-all-jobs` 的 `needs` 依赖。

关键符号: 未识别

评论区精华

无评审评论, 评审者 bingxche 直接批准。

- 暂无高价值评论线程

风险与影响

- 风险：风险极低。仅修改 CI workflow 配置，不涉及任何模型、内核或运行时代码。主要潜在风险是 YAML 语法错误或任务依赖配置不正确，但已通过本地验证和逐字节对比确保正确性。由于测试套件文件未被删除，恢复的任务应能成功运行。
- 影响：影响范围限于 ROCm 7.2 夜间 CI 流程。恢复后将增加一个夜间任务，为 MiniMax-M2.5 提供 MI35x 精度覆盖，不影响其他任务或运行时行为。对用户和系统无影响，对团队而言，保证了模型回归检测的完整性。
- 风险标记：CI 配置变更，无测试覆盖

关联脉络

- PR #36142 [AMD][CI] Add MiniMax-M3-MXFP8 MI35x nightly perf benchmark: 本 PR 恢复了 #36142 中移除的 MiniMax-M2.5 任务，两者直接相关。