

PR #29409 完整报告

sgl-project/sglang

[AMD] Split qwen3.5 triton DCP test into its own nightly job

合并时间: 2026-07-01 15:39

原文链接: <http://prhub.com.cn/sgl-project/sglang/pull/29409>

执行摘要

将 AMD 夜测中 Qwen3.5 Triton DCP 测试从原超时 job 中拆分到独立 job，每个 job 单独拥有 120 分钟超时，彻底解决因两个测试总预估时间 8400 秒超过 7200 秒 timeout 而导致的持续失败。

功能与动机

原 AMD 夜测 job `nightly-8-gpu-mi35x-qwen35` 的单个 step 中顺序执行两个测试文件：`test_qwen35_eval_mi35x.py`（预估 3600 秒）和 `test_qwen3p5_triton_dcp.py`（预估 4800 秒），合计约 8400 秒，超过 `timeout-minutes: 120`（7200 秒），导致 job 频繁超时失败。需要拆分以稳定 CI。

实现拆解

1. 修改测试注册套件名：test/registered/amd/test_qwen3p5_triton_dcp.py 中 register_amd_ci 的 suite 参数从 nightly-amd-accuracy-8-gpu-mi35x-qwen35 改为独立套件名 nightly-amd-accuracy-8-gpu-mi35x-qwen35-triton-dcp。
2. 新增独立 job（默认版）：在 .github/workflows/nightly-test-amd.yml 中添加 job nightly-8-gpu-mi35x-qwen35-triton-dcp，包含完整 CI 步骤，设置 timeout-minutes: 120，运行新套件。
3. 新增独立 job（ROCm 7.2 版）：在 .github/workflows/nightly-test-amd-rocm720.yml 中添加对应 job nightly-8-gpu-mi35x-qwen35-triton-dcp-rocm720。
4. 注册 job：在两个 workflow 的 on 触发列表和 check-all-jobs 的 needs 列表中添加新 job 名称。
5. 验证：PR 作者已分别在 ROCm 7.0 和 ROCm 7.2 上运行通过。

无需源码片段，变更为纯 YAML 配置和一行字符串修改。

评论区精华

无实质讨论，只有 gemini-code-assist[bot] 的自动评论和 HaiShaw 的批准。

风险与影响

- 风险：极低。仅 CI 配置变更，不涉及生产代码。新 job 可能缺少某些环境变量或 runner 标签，但已通过实际运行验证。

- 影响：AMD 夜测流水线更稳定，不再因超时导致整组失败；Qwen3.5 Triton DCP 测试获得独立执行时间和失败报告；额外占用 8-GPU MI35x 机器约 80 分钟。

关联脉络

本 PR 是独立的 CI 稳定性修复，与近期其他 PR（如 HiCache 优化、Kernel 测试增强等）无直接关联。但可参考此模式拆分其他超时 job。