PR #21178 完整报告

sgl-project/sglang

Temporarily disable flaky qwen3 cp test in CI

合并时间：2026-03-23 12:13

原文链接：http://prhub.com.cn/sgl-project/sglang/pull/21178

执行摘要

本 PR 临时禁用了 CUDA CI 中不稳定的 Qwen3-30B 上下文并行测试，以解决 CI 失败问题，确保 CI 流程稳定，但牺牲了测试覆盖。变更简单，被快速合并。

功能与动机

根据 CI 运行失败（https://github.com/sgl-project/sglang/actions/runs/23402689674/job/68108381141），Qwen3-30B 测试在 CUDA CI 中表现不稳定，可能导致 CI 阻塞。因此，通过临时禁用该测试来维护 CI 的可靠性，避免频繁失败影响开发流程。

实现拆解

唯一改动位于 test/registered/4-gpu-models/test_qwen3_30b.py 文件，具体变更如下：

- 原代码：python register_cuda_ci(est_time=300, suite="stage-c-test-4-gpu-h100")
- 新代码：python register_cuda_ci(est_time=300, suite="stage-c-test-4-gpu-h100", disabled="Temporarily disable the flaky test.",) 通过在 register_cuda_ci 函数调用中添加 disabled 参数，该测试在 CI 中将被跳过。

评论区精华

无实质性讨论，只有 Fridge003 的批准评论（内容为空），表明变更被快速接受。

风险与影响

风险：禁用测试可能掩盖 Qwen3-30B 模型在上下文并行下的潜在问题，但鉴于测试不稳定可能源于环境因素，风险较低。具体在文件 test/registered/4-gpu-models/test_qwen3_30b.py 中，禁用后该测试不会执行，导致相关代码路径缺乏验证。 影响：对用户无直接影响；CI 将更稳定，减少失败运行；团队开发效率提升，但测试套件完整性受损。影响范围仅限于 CUDA CI 的特定测试执行。

关联脉络

本 PR 属于 CI 维护系列，与近期 PR 如 #21187（统一测试套件命名）、#21118（移除 Blackwell 环境变量）和 #21162（修复 NPU CI git 问题）相关，共同优化 CI 配置和稳定性，反映团队对持续集成流程的持续改进。