PR #21370 完整报告

sgl-project/sglang

Update skip condition for TestQwen35PPAccuracy

合并时间：2026-03-25 14:28

原文链接：http://prhub.com.cn/sgl-project/sglang/pull/21370

执行摘要

本 PR 更新了 TestQwen35PPAccuracy 测试的跳过条件，从仅针对 AMD CI 扩展到所有 CI 环境，以临时解决 H100 和 AMD GPU 上的精度回归导致的 CI 阻塞，为调查根本原因提供时间。

功能与动机

Qwen35 PP 支持引入后，在 H100 和 AMD 硬件上出现精度问题，但无法在 H20 上复现。根据 PR body 描述，为允许 CI 测试通过，避免开发阻塞，决定暂时跳过该测试。

实现拆解

修改文件 test/registered/distributed/test_pp_single_node.py，将 @unittest.skipIf(is_in_amd_ci(), "PP consistency too flaky on AMD 4-GPU runners") 替换为 @unittest.skipIf(is_in_ci(), "Qwen35 PP consistency too flaky on H100 and AMD 4-GPU runners")，使测试在 CI 环境下全部跳过。

评论区精华

无 review 评论。

风险与影响

风险在于临时跳过可能掩盖精度问题，延迟修复；影响是CI稳定性提升，但测试覆盖暂时降低。

关联脉络

- 关联 PR #19670 和 #21070，涉及 Qwen35 PP 的引入和修复。
- 同文件修改的 PR #20294 显示对分布式测试的持续维护。