PR #21564 完整报告

sgl-project/sglang

Fix flaky test_pp_single_node

合并时间：2026-03-28 05:33

原文链接：http://prhub.com.cn/sgl-project/sglang/pull/21564

执行摘要

- 一句话：修复分布式单节点测试中 GSM8K 的 flaky 断言。
- 推荐动作：该 PR 变更简单，无需精读代码；对于工程师，可关注类似 flaky 测试修复模式（如阈值放宽策略），并可参考 PR 21562 等历史 PR 学习 CI 稳定性优化。

功能与动机

PR body 中引用的 CI 运行失败日志（链接 https://github.com/sgl-project/sglang/actions/runs/23627607137/job/68950721355?pr=20972#step:7:8385）显示 test_gsm8k 准确性断言失败，表明该测试存在不稳定性（flaky）。变更旨在通过调整断言条件来减少此类失败，确保 CI 测试更可靠。

实现拆解

仅修改测试文件 test/registered/distributed/test_pp_single_node.py 中的 test_gsm8k 方法：将准确性断言的阈值从大于 0.65 放宽至大于等于 0.65，即 self.assertGreater(metrics["accuracy"], 0.65) 改为 self.assertGreaterEqual(metrics["accuracy"], 0.65)。没有其他代码或模块变更。

关键文件：

- test/registered/distributed/test_pp_single_node.py（模块 测试 / 分布式）：唯一修改的文件，包含 flaky 测试断言修复，直接影响 CI 测试稳定性。

关键符号：test_gsm8k

评论区精华

review 中只有 gemini-code-assist[bot] 的评论，指出变更内容并确认无问题，未引发争议或深入讨论。Issue 评论中作者执行了 /rerun-ut 重新运行测试以验证修复效果。

- 测试断言变更 (testing): 无争议，变更被接受并合并。

风险与影响

- 风险：风险较低：变更仅放宽测试断言阈值，可能掩盖潜在的性能回归问题（例如模型准确性降至 0.65 时不再触发失败）。但由于是纯测试调整，不影响生产代码、系统功能或安全，且无兼容性或性能风险。

- 影响：影响局限于 CI 测试流程：减少 test_pp_single_node.py 测试的 flaky 失败，提高开发体验和 CI 效率。无用户界面、系统性能或代码功能的影响。
- 风险标记：测试阈值放宽

关联脉络

- PR #21562 [CI] Relax several thresholds in flaky CIs: 同样修复 flaky CI 测试，通过放宽测试阈值以减少不稳定性，涉及类似技术手法。