PR #21371 完整报告

sgl-project/sglang

[CI] Fix TestQwen35WithHiCache

合并时间：2026-03-25 15:05

原文链接：http://prhub.com.cn/sgl-project/sglang/pull/21371

执行摘要

- 一句话：修复 HiCache 测试失败，将测试拆分为独立文件以优化 CI 执行。
- 推荐动作：对于技术管理者，此 PR 无需深入评审，可快速合并以稳定 CI。对于工程师，可作为测试文件重构的简单案例，但无关键技术洞察，不建议精读。建议关注类似 CI 修复 PR（如 #21305）以积累测试稳定性优化经验。

功能与动机

PR body 中明确说明 'Fix this failure: https://github.com/sgl-project/sglang/actions/runs/23505618811/job/68438441879'，表明目标是修复特定 CI 工作流失败。Issue 评论中 Fridge003 多次触发 CI 重跑（如 '/rerun-ut' 和 '/rerun-stage'），并报告手动测试结果（score: 0.92），验证修复效果。

实现拆解

实现方案分为两部分：1) 新增文件 test/registered/4-gpu-models/test_qwen35_hicache.py，包含完整的 TestQwen35WithHiCache 测试类，设置 HiCache 相关服务器参数（如 --enable-hierarchical-cache、--hicache-storage-backend file）并执行 gsm8k 评估。2) 修改文件 test/registered/4-gpu-models/test_qwen35_models.py，移除原 TestQwen35WithHiCache 类及相关导入和变量定义（如 QWEN35_27B_MODEL），仅保留其他测试类，代码行数从 103 删除至 1 新增。

关键文件：

- test/registered/4-gpu-models/test_qwen35_hicache.py（模块 test/ci）：新增的 HiCache 测试文件，包含完整的 TestQwen35WithHiCache 测试类，是 PR 核心变更，负责验证 HiCache 功能准确性。
- test/registered/4-gpu-models/test_qwen35_models.py（模块 test/ci）：原文件移除 TestQwen35WithHiCache 类及相关代码，简化测试结构，避免冗余和潜在冲突。

关键符号：TestQwen35WithHiCache.setUpClass, TestQwen35WithHiCache.test_gsm8k, TestQwen35WithHiCache._run_gsm8k

评论区精华

Review 区无评论。Issue 评论中主要是 CI 触发和手动测试结果：Fridge003 使用命令验证修复，并报告 'Manual run qwen3.5 hi cache test on 4*H200 ... Score: 0.920'，确认测试通过。无技术争议或深度讨论，重点是确认 CI 稳定性。

- CI 重跑验证修复效果 (question): 手动测试通过（score: 0.92），表明 CI 失败已修复，测试结构优化成功。

风险与影响

- 风险：风险较低：1) 文件移动可能导致导入错误或 CI 调度依赖问题，但新增文件逻辑与原测试一致，风险可控。2) 测试参数未变，但拆分后可能影响测试分区或执行顺序；历史 PR #21305 已增加缓存刷新超时，表明 HiCache 测试有 flakiness 风险，需关注后续稳定性。无安全或性能风险。
- 影响：影响有限：1) 对用户无直接影响，是内部 CI 优化。2) 对系统，提升测试组织清晰度和可维护性，减少 CI 失败概率。3) 对团队，测试文件结构更模块化，便于后续维护和扩展；影响程度为低，主要集中在测试基础设施层面。
- 风险标记：文件结构变更，潜在导入错误

关联脉络

- PR #21305 Increase flush cache timeout in hicache CI: 同样涉及 HiCache 测试的 CI 稳定性优化，修改了缓存刷新逻辑，与本 PR 的测试拆分共同提升测试可靠性。
- PR #21370 Update skip condition for TestQwen35PPAccuracy: 同为 Qwen3.5 模型相关测试的 CI 修复，涉及测试条件调整，反映团队在持续优化测试套件以避免阻塞。