PR #21791 完整报告

sgl-project/sglang

Increase hicache eval to 200 examples

合并时间：2026-04-01 07:58

原文链接：http://prhub.com.cn/sgl-project/sglang/pull/21791

执行摘要

本 PR 通过将 hicache 一致性测试的 GSM8K 样本数从 50 增加到 200，并调整并行数从 10 到 64，解决了测试中 flaky 分数差异超过阈值的问题。这是一个小范围测试改进，旨在提升 CI 稳定性和团队效率，风险较低。

功能与动机

此变更的主要动机是修复测试的 flakiness，避免因样本数不足导致的一致性得分波动。在 PR body 中，作者指出：

> With 50 examples each sample is worth 0.02, causing flaky score diffs > 0.03 threshold. With 200 examples, locally verified diff = 0.01 < 0.03. 通过增加样本数，每个样本的权重降低，从而减少随机差异，确保测试结果更稳定可靠。

实现拆解

实现方案集中在文件 test/registered/hicache/test_hicache_storage_file_backend.py 的 run_eval_accuracy_test 函数中。关键改动如下：

- num_questions从 50 改为 200：增加评估样本数，降低每个样本的分数权重。
- parallel从 10 改为 64：提高测试并行度，可能加快执行速度。 代码示例： 所有变更仅调整测试参数，未触及核心业务逻辑，属于配置优化。

评论区精华

Review 中无实质性讨论，Issue 评论中作者使用 /rerun-test 命令进行测试重运行，但未触发技术交流。这表明变更简单直接，团队共识度高，无需深入辩论。

风险与影响

- 风险分析：增加样本数可能延长测试时间，但并行度提高可部分缓解；并行数增加可能导致资源使用上升，需在 CI 环境中验证。无其他技术风险，因未修改生产代码。
- 影响分析：对用户无影响；对系统，测试更稳定，减少 CI flaky 失败；对团队，降低维护开销，提升测试信心。影响范围小，仅限 hicache 测试模块。

关联脉络

从历史 PR 分析中，关联 PR 如 #21745 和 #21751 都涉及测试稳定性改进，显示团队在持续优化 CI 流程以减少 flaky 问题。此 PR 是这一趋势的一部分，体现了对测试可靠性的重视。未

来可关注类似调整是否在其他测试模块中推广。