PR #22194 完整报告

sgl-project/sglang

[Qwen3-Specv2]: Fix flaky ci

合并时间：2026-04-07 00:40

原文链接：http://prhub.com.cn/sgl-project/sglang/pull/22194

执行摘要

- 一句话：提高 Qwen3 Next MTP V2 测试的 KL 散度阈值以修复 CI 不稳定性。
- 推荐动作：此 PR 无需精读，除非您负责 Qwen3 Next MTP 测试维护。关注点：阈值调整是否合理（从 0.0025 到 0.0035 的增量是否基于数据驱动决策）。

功能与动机

根据 PR 标题和 Issue 评论，此变更旨在修复 CI 测试的间歇性失败（flaky ci）。评论中 ispobock 指出这是为了修复一个特定的 CI 运行失败（链接指向一个失败的测试步骤）。

实现拆解

仅修改了一个测试文件：将 test/registered/4-gpu-models/test_qwen3_next_models_mtp.py 中 TestQwen3NextMTPV2 类的 kl_div_thres 属性从 0.0025 调整为 0.0035。

关键文件：

- test/registered/4-gpu-models/test_qwen3_next_models_mtp.py（模块 测试）：唯一被修改的文件，包含 Qwen3 Next MTP V2 测试的 KL 散度阈值调整。

关键符号：未识别

评论区精华

review 讨论非常有限。gemini-code-assist[bot] 的评论仅描述了变更内容（“放宽了模型评估的 KL 散度阈值”），没有提出任何问题或建议。ispobock 直接批准，没有额外评论。

- KL 散度阈值调整 (testing): 变更被接受，无争议。

风险与影响

- 风险：风险极低：
  1. 仅修改测试阈值，不涉及生产代码，无回归风险。
  2. 提高阈值可能掩盖模型输出的微小退化，但这是测试策略权衡，而非技术风险。
  3. 变更范围极小（单行修改），易于验证。
- 影响：影响有限：
  1. 对用户：无直接影响。

2. 对系统：提高测试通过率，减少 CI 噪声。
3. 对团队：简化维护，但需注意阈值调整可能降低测试严格度。

- 风险标记：测试严格度降低

关联脉络

- PR #22190 Update test coverage report: 同属测试 /CI 优化类别，关注测试稳定性和报告改进。
- PR #22176 Fix ut module importing: 同属修复测试问题的 PR，涉及测试环境依赖和导入机制。
- PR #22170 fix hisparse LRU policy: 同属修复 CI 问题的 PR（标签包含 run-ci），但涉及核心 JIT 内核 bugfix，而本 PR 仅调整测试阈值。