PR #27419 完整报告

sgl-project/sglang

fix test_qwen3_next_models flaky

合并时间：2026-06-06 14:26

原文链接：http://prhub.com.cn/sgl-project/sglang/pull/27419

执行摘要

- 一句话：调大 KL 散度阈值修复测试 flaky
- 推荐动作：变更简单直接，无需精读。可作为测试稳定性维护的参考案例。

功能与动机

修复测试 flaky：Qwen3-Next 模型在不同运行时下 KL 散度存在微小波动，原阈值 0.001 过于严格，导致 CI 中测试偶发失败。

实现拆解

1. 在 test/registered/models_e2e/test_qwen3_next_models.py 中，将四个测试类（TestQwen3NextLazyExtraBuffer、TestQwen3NextLazyExtraBufferLargePage、TestQwen3NextLazyExtraBufferAllocFail、TestQwen3NextLazyExtraBufferLargePageAllocFail）的 kl_div_thres 属性从 0.001 修改为 0.002。
2. 其他配置（如模型名、gsm8k_accuracy_thres、cache_chunk_size、启动参数等）保持不变。
3. 无任何源码或配置的修改，纯测试阈值松动。

关键文件：

- test/registered/models_e2e/test_qwen3_next_models.py（模块 测试；类别 test；类型 test-coverage）：唯一的变更文件，修改了 KL 散度阈值以解决 flaky 测试。

关键符号：未识别

关键源码片段

test/registered/models_e2e/test_qwen3_next_models.py

唯一的变更文件，修改了 KL 散度阈值以解决 flaky 测试。

```
# test/registered/models_e2e/test_qwen3_next_models.py
# 关键变更：将 kl_div_thres 从 0.001 调整为 0.002，
# 以容忍 Qwen3-Next 模型 KL 散度的微小波动，消除 CI 中的偶发失败。

class TestQwen3NextLazyExtraBuffer(
    GSM8KMixin, KLDivergenceMixin, PrefixCacheBranchingMixin, DefaultServerBase
```

```
):
    model = QWEN3_NEXT_MODEL
    cache_chunk_size = 64
    gsm8k_accuracy_thres = 0.93
    kl_div_thres = 0.002 # 原为 0.001
    other_args = _make_args(page_size=1, track_interval=2)


class TestQwen3NextLazyExtraBufferLargePage(
    GSM8KMixin, KLDivergenceMixin, PrefixCacheBranchingMixin, DefaultServerBase
):
    model = QWEN3_NEXT_MODEL
    cache_chunk_size = 64
    gsm8k_accuracy_thres = 0.93
    kl_div_thres = 0.002 # 原为 0.001
    other_args = _make_args(page_size=2, track_interval=2)

# 另外两个 skip 测试类也有相同的阈值调整
```

评论区精华

无 review 评论，合并者 ispobock 直接批准合并。PR 创建后 CI 多次 rerun 均通过，表明阈值调整有效消除了 flaky。

- 暂无高价值评论线程

风险与影响

- 风险：风险极低：放宽阈值 0.001 到 0.002 仅为 2 倍，不影响模型质量基准，且 gsm8k_accuracy_thres 保持不变（0.93），不会掩盖真实回归。
- 影响：仅影响测试稳定性，对用户无任何影响，对系统无性能或功能影响。团队 CI 可靠性提升，减少人工 rerun 成本。
- 风险标记：暂无

关联脉络

- PR #27413 Add scripted-runtime unit, core integration, and chunked-prefill tests: 同一仓库近期的测试基础设施增强，但本 PR 独立于该系列。