PR #25497 完整报告

sgl-project/sglang

Update kl_div_thres to 0.02 in swa_radix_cache

合并时间：2026-05-17 07:43

原文链接：http://prhub.com.cn/sgl-project/sglang/pull/25497

执行摘要

- 一句话：调整 KL 散度测试阈值为 0.02
- 推荐动作：建议合并，但需要添加明确的动机说明（例如引用具体的 CI 失败链接或 KL 散度分布数据）。此外，删除冗余注释后代码更简洁。

功能与动机

现有的 KL 散度阈值 0.002 过于严格，导致测试在正常模型输出波动下频繁失败。通过提高阈值至 0.02，可以容忍合理的变化幅度，避免误报回归。但 PR body 未明确说明具体失败案例或基准输出变化，动机不够充分。

实现拆解

仅修改了一个测试文件 test/registered/radix_cache/test_swa_radix_cache_kl.py 中 TestSWARadixCacheKL 类的 kl_div_thres 属性，从 0.002 改为 0.02，并删除了多余的注释。

关键文件：

- test/registered/radix_cache/test_swa_radix_cache_kl.py（模块 Radix Cache；类别 test；类型 test-coverage）：该文件是变更的唯一文件，修改了 KL 散度测试阈值，直接影响测试通过标准。

关键符号：未识别

关键源码片段

test/registered/radix_cache/test_swa_radix_cache_kl.py

该文件是变更的唯一文件，修改了 KL 散度测试阈值，直接影响测试通过标准。

```
# 文件：test/registered/radix_cache/test_swa_radix_cache_kl.py
class TestSWARadixCacheKL(KLDivergenceMixin, DefaultServerBase):
    model = MODEL
    kl_div_thres = 0.02 # 阈值从 0.002 提升至 0.02，以容忍输出波动
    kl_div_decode_max_new_tokens = 2048
    other_args = [
        "--tp-size", "1",
        "--mem-fraction-static", "0.70",
        "--disable-piecewise-cuda-graph",
```

]

评论区精华

Review 评论中，gemini-code-assist[bot] 指出内联注释 # it was 0.002 是冗余的（版本控制系统已记录历史），并要求提供增加阈值的动机，以确保不是掩盖回归。

- 冗余注释与动机缺失 (other): 作者删除了注释，但未补充动机说明。

风险与影响

- 风险：风险较低：仅改动测试阈值，不影响任何生产逻辑。但若阈值提升过大，可能掩盖真正的输出回归，需要结合具体 KL 散度值判断。
- 影响：影响范围小，仅影响 TestSWARadixCacheKL 这一个测试用例的判定宽松度。CI 中该测试通过率将提高，但可能降低回归检测灵敏度。
- 风险标记：阈值调整可能掩盖回归

关联脉络

- PR #25477 [BugFix]: Fix DeepSeek V4 HiCache layer count logic: 同一目录下的 HiCache 相关测试文件，反映 SWA Radix Cache 测试维护活动。