PR #24421 完整报告

sgl-project/sglang

[UnifiedRadixTree]: Fix flaky ci

合并时间：2026-05-05 20:22

原文链接：http://prhub.com.cn/sgl-project/sglang/pull/24421

执行摘要

- 一句话：修复 UnifiedRadixCache 测试的稳定性问题
- 推荐动作：可直接合并，属于低风险 CI 稳定性修复。

功能与动机

SWA 模型的 MMLU 评估结果在 CI 环境中不够稳定，导致测试频繁失败；同时多轮 KL 测试中 cached_tokens 计数在解码缓存命中场景下可能略高于预期值，需要放宽断言避免误报。

实现拆解

1. SWA MMLU 测试跳过：在 test/registered/radix_cache/test_unified_radix_cache_kl.py 中，为 TestUnifiedSWARadixCache.test_mmlu 添加 @unittest.skipIf(is_in_ci(), ...) 装饰器，当在 CI 中运行时跳过该测试，并调用 super().test_mmlu() 确保本地运行时仍执行测试逻辑。
2. 放宽 cached_tokens 断言：在 python/sglang/test/kl_multiturn_utils.py 中，将 make_mamba_decode_assert 返回的检查函数从 actual == expected 改为 actual >= expected，以兼容解码缓存命中时 token 计数可能多于期望值的情况。

关键文件：

- test/registered/radix_cache/test_unified_radix_cache_kl.py（模块 测试；类别 test；类型 test-coverage；符号 test_mmlu）：为 SWA 模型的 MMLU 测试添加 CI 跳过逻辑，避免不稳定导致的 CI 失败。
- python/sglang/test/kl_multiturn_utils.py（模块 测试；类别 test；类型 test-coverage）：放宽 cached_tokens 断言，避免解码缓存命中场景下的误报。

关键符号：test_mmlu, make_mamba_decode_assert

关键源码片段

test/registered/radix_cache/test_unified_radix_cache_kl.py

为 SWA 模型的 MMLU 测试添加 CI 跳过逻辑，避免不稳定导致的 CI 失败。

```
from sglang.test.test_utils import (
    DEFAULT_TIMEOUT_FOR_SERVER_LAUNCH,
    DEFAULT_URL_FOR_TEST,
```

```
    CustomTestCase,
    is_in_ci, # 新增导入，用于判断是否在 CI 环境中运行
    popen_launch_server,
)
...
class TestUnifiedSWARadixCache(UnifiedRadixTreeTestMixin, CustomTestCase):
    """SWA hybrid + UnifiedRadixCache."""
    ...
    # SWA 模型的 MMLU 评估在 CI 中不够稳定，跳过以避免误报
    @unittest.skipIf(is_in_ci(), "SWA model mmlu eval not stable enough")
    def test_mmlu(self):
        super().test_mmlu() # 本地运行仍执行父类测试逻辑
```

评论区精华

代码审查机器人指出：

1) is_in_ci 未导入会导致 NameError（已在后续提交修复）； 2) 初始实现使用了 pass 而非调用 super().test_mmlu()，会完全禁用测试（已修复）。

- test_mmlu 实现问题 (correctness): 已修复：添加导入并改为调用 super().test_mmlu()。

风险与影响

- 风险：无显著风险：变更仅涉及测试跳过和断言放宽，不影响核心逻辑。SKIP 装饰器确保本地测试仍会执行，断言放宽仅放宽了下界，不会漏报。
- 影响：降低 UnifiedRadixCache 相关测试的 CI 失败率，避免 SWA MMLU 不稳定导致的误报，同时不牺牲本地测试覆盖。
- 风险标记：暂无

关联脉络

- 暂无明显关联 PR