# PR #27426 完整报告

sgl-project/sglang

Fix flaky test_self_e2e_pd_perturb

合并时间：2026-06-06 19:31

原文链接：http://prhub.com.cn/sgl-project/sglang/pull/27426

## 执行摘要

- 一句话：修复 KV 校验测试因 Radix 缓存去重导致的 flaky 问题
- 推荐动作：值得精读 PR body 中的根因分析，它揭示了 cache_unfinished_req 与 send_kv_chunk 之间的时序竞态如何导致去重后的槽位被错误释放，是理解 PD 架构中 KV 传输、Radix 缓存和 canary 验证三者交互的绝佳案例。

## 功能与动机

test_self_e2e_pd_perturb.py 在 CI 中 flaky：P 侧扰动触发成功，但 D 侧从未报告预期的 verify_real_kv_hash 违例。根因是并行请求共享相同 prompt，Prefill 端 RadixCache.cache_unfinished_req 在 send_kv_chunk 之前去重了 KV 槽位，导致被扰动的非规范副本在传输前就被释放。PR body 详细描述了根因分析过程。

## 实现拆解

1. 在 pd_fixture.py 的 send_parallel_short_requests 方法新增 distinct_prompts 参数：当为 True 时，为每个请求拼接 "{i} {nonce} {_SHORT_PROMPT_BODY}"，其中 nonce 为 uuid.uuid4().hex[:8]；否则保持原有行为。
2. 在 test_self_e2e_pd_perturb.py 的测试方法中，调用 self.send_parallel_short_requests(n=4, distinct_prompts=True) 代替原调用，并添加详细注释说明原因。
3. 回滚了 #27425 中禁用该测试的提交，重新启用 CI 中的测试。

关键文件：

- python/sglang/test/kv_canary/pd_fixture.py（模块 测试夹具；类别 test；类型 test-coverage；符号 send_parallel_short_requests）：新增 distinct_prompts 参数，允许测试用例生成差异化 prompt 以避免 Radix 去重
- test/registered/kv_canary/test_self_e2e_pd_perturb.py（模块 端到端测试；类别 test；类型 test-coverage；符号 _PDPerturbBase.test_p_side_perturb_surfaces_real_kv_hash_violation_on_decode_side）：使用 distinct_prompts=True 调用 send_parallel_short_requests 并添加注释说明根本原因

关键符号：send_parallel_short_requests, test_p_side_perturb_surfaces_real_kv_hash_violation_on_decode_side

## 关键源码片段

### python/sglang/test/kv_canary/pd_fixture.py

新增 distinct_prompts 参数，允许测试用例生成差异化 prompt 以避免 Radix 去重

```
def send_parallel_short_requests(
    self,
    n: int,
    *,
    assert_all_success: bool = True,
    max_new_tokens: int = 100,
    timeout: float = 60.0,
    distinct_prompts: bool = False, # new parameter
) -> list[dict]:
    if distinct_prompts:
        # 每个请求使用不同的前缀（请求索引 + 随机 nonce），
        # 使得请求间除 BOS token 外没有共享前缀，
        # 从而避免 RadixCache.cache_unfinished_req 进行去重。
        nonce = uuid.uuid4().hex[:8]
        prompts = [f"{i} {nonce} {_SHORT_PROMPT_BODY}" for i in range(n)]
    else:
        prompts = [_SHORT_PROMPT_BODY] * n
    results = post_parallel_generate(
        url=self.lb_url + "/generate",
        prompts=prompts,
        max_new_tokens=max_new_tokens,
        timeout=timeout,
    )
    if assert_all_success:
        for result in results:
            self.assertEqual(result.get("status_code"), 200, result)
    return results
```

### test/registered/kv_canary/test_self_e2e_pd_perturb.py

使用 distinct_prompts=True 调用 send_parallel_short_requests 并添加注释说明根本原因

```
def test_p_side_perturb_surfaces_real_kv_hash_violation_on_decode_side(
    self,
) -> None:
    # distinct_prompts 对于可靠触发违例是必须的：
    # 当请求共享同一个 prompt 时，P 端 radix 缓存会在 cache_unfinished_req 中
    # 将每个请求的 req_to_token 行改写为第一个插入的（规范）副本的槽位，
    # 这发生在 send_kv_chunk 快照索引之前。因此，如果扰动落在被去重的副本槽位上，
    # 该槽位会被释放而不会被传输或重新验证，导致两侧均无法报告违例。
    self.send_parallel_short_requests(n=4, distinct_prompts=True)
    # D 端：第一个 decode forward 重新验证已传输的前缀槽位，
    # 因此扰动必须作为 real_kv_hash 违例出现。
    self.assert_per_forward_violation_reported(
        fail_reason="verify_real_kv_hash",
```

```
        target_group=self.target_group,
        side="decode",
        flush_wait_seconds=4.0,
    )
    # P 端：扰动发生在 prefill forward 的 TAIL 之后，
    # 而 PD prefill 不会在 P 上运行另一个 forward 来验证被扰动的槽位，
    # 因此 P 必须保持静默（无假阳性违例）。
    self.assert_no_violation(side="prefill", wait_seconds=0.5)
```

# 评论区精华

无人工 review 评论，仅有 gemini-code-assist 的自动评论，但未提供实质性反馈。PR 作者 fzyzcjy 通过详尽的 PR body 和 commit history 独立完成了根因分析、方案演进和验证。

- 暂无高价值评论线程

# 风险与影响

- 风险：该变更仅影响测试代码，不涉及生产逻辑。使用 distinct_prompts 后，Radix 缓存依然启用（与生产一致），但请求之间不再共享前缀，因此 Radix 去重路径仅在 test_self_e2e_pd_baseline.py 中覆盖（该测试仍使用相同 prompt）。如果未来生产代码的 Radix 去重行为发生变化，该测试可能无法检测到相关回归；但已有基线测试覆盖。
- 影响：影响范围仅限于 test_self_e2e_pd_perturb.py 及其依赖的 pd_fixture.py。修复后测试稳定性显著提升：在 2-GPU H200 上 15/15 轮全文件通过、30/30 轮子类专项通过；在原始 flaky 环境 2-gpu-h100 上 4/5 轮通过（失败为不相关的基础设施问题）。
- 风险标记：测试覆盖调整

# 关联脉络

- PR #27425 Temporarily disable test_self_e2e_pd_perturb in CI: 前序 PR 临时禁用了该测试，本 PR 重新启用并修复了 flaky 问题