

PR #45345 完整报告

vllm-project/vllm

[CI][BugFix] Fix broken `test\_mamba\_prefix\_cache.py` due to stale mock

合并时间: 2026-06-12 11:26

原文链接: <http://prhub.com.cn/vllm-project/vllm/pull/45345>

执行摘要

- 一句话: 修复测试 mock 与新版 KVCacheManager 签名不匹配
- 推荐动作: 此类伴随重构的测试更新非常重要, 建议 CI 中增加对涉及变更模块的相关测试的自动触发机制, 避免遗漏。

功能与动机

PR #44594 变更了 KVCacheManager.allocate_slots() 的方法签名, 新增 has_scheduled_reqs 参数, 但相应的测试 mock 未同步更新, 导致 test_mamba_prefix_cache.py 因签名不匹配而失败。由于该测试未被触发运行, 问题未被及时发现。

实现拆解

1. 在 tests/v1/e2e/general/test_mamba_prefix_cache.py 的 fake_allocate_slots_fn mock 函数形参列表中新增 has_scheduled_reqs: bool = True 参数。
2. 在调用 original_allocate_slots_fn 时将 has_scheduled_reqs 参数透传。此改动仅涉及测试代码, 不修改任何生产逻辑。

关键文件:

- tests/v1/e2e/general/test_mamba_prefix_cache.py (模块 Mamba 前缀缓存; 类别 test; 类型 test-coverage): 测试 mock 函数签名与生产代码不同步, 导致测试失败; 本 PR 修复该问题。

关键符号: 未识别

关键源码片段

tests/v1/e2e/general/test_mamba_prefix_cache.py

测试 mock 函数签名与生产代码不同步, 导致测试失败; 本 PR 修复该问题。

```
# 修复后的 mock 函数: 补全了 `has_scheduled_reqs` 参数
# 与 KVCacheManager.allocate_slots 的最新签名保持一致
def get_fake_allocate_slots_fn(original_allocate_slots_fn: Callable):
    def fake_allocate_slots_fn(
        self: KVCacheManager,
        request: Request,
```

```

num_new_tokens: int,
num_new_computed_tokens: int = 0,
new_computed_blocks: KVCacheBlocks | None = None,
num_lookahead_tokens: int = 0,
num_external_computed_tokens: int = 0,
delay_cache_blocks: bool = False,
num_encoder_tokens: int = 0,
full_sequence_must_fit: bool = False,
reserved_blocks: int = 0,
has_scheduled_reqs: bool = True, # 新增参数, 默认 True
):
    ret = original_allocate_slots_fn(
        self,
        request,
        num_new_tokens,
        num_new_computed_tokens,
        new_computed_blocks,
        num_lookahead_tokens,
        num_external_computed_tokens,
        delay_cache_blocks,
        num_encoder_tokens,
        full_sequence_must_fit,
        reserved_blocks,
        has_scheduled_reqs, # 透传参数
    )
    # 其余验证逻辑不变
    # ...
    return ret

return fake_allocate_slots_fn

```

评论区精华

无 review 讨论，已获得批准。

- 暂无高价值评论线程

风险与影响

- 风险：无风险。仅修改测试 mock，与生产代码完全解耦，且变更极小。
- 影响：影响范围仅限于修复 `test_mamba_prefix_cache.py` 测试，使其能通过且正确验证 `KVCacheManager.allocate_slots` 的行为。
- 风险标记：测试遗漏

关联脉络

- PR #44594 [Core] Add kvcache watermark to reduce preemptions: 该 PR 修改了 `KVCacheManager.allocate_slots()` 的方法签名，新增 `has_scheduled_reqs` 参数，导致本

PR 中的测试 mock 需要同步更新。