PR #38791 完整报告

vllm-project/vllm

[Bugfix] Fix test mocks after SM100 restriction in #38730

合并时间：2026-04-03 01:12

原文链接：http://prhub.com.cn/vllm-project/vllm/pull/38791

执行摘要

- 一句话：修复因 #38730 更改 API 导致的 TRT-LLM 注意力测试 mock 失效问题。
- 推荐动作：该 PR 变更简单直接，无需精读。值得关注的是其中揭示的 CI 依赖管理问题（如 #38792 所提），这对测试稳定性和 CI 可靠性有借鉴意义。

功能与动机

PR body 明确指出这是对 #38730 的直接跟进，因为 #38730 将 supports_trtllm_attention() 中的 is_device_capability_family(100) 改为 is_device_capability(100)，但未更新测试 mock，导致测试失败。CI 在 #38730 中未捕获此问题，因为 'Kernels Attention Test' 作业的源文件依赖不包含 vllm/utils/flashinfer.py，仅在生产环境夜间测试中发现。

实现拆解

仅修改一个测试文件 tests/kernels/attention/test_use_trtllm_attention.py，将三个测试函数中的 mock 对象从 vllm.utils.flashinfer.current_platform.is_device_capability_family 替换为 vllm.utils.flashinfer.current_platform.is_device_capability，以匹配 #38730 中生产代码的 API 变更。

关键文件：

- tests/kernels/attention/test_use_trtllm_attention.py（模块 tests/kernels/attention）：唯一被修改的文件，包含三个需要更新 mock 的测试函数，直接修复测试与生产代码 API 不一致问题。

关键符号：test_supports_batch_invariant_disables, test_supports_sm100_with_artifactory, test_supports_non_sm100_platform

评论区精华

review 中无实质性技术讨论，仅包含形式性批准。作者 stecasta 在 Issue 评论中建议关注 #38792，因为它将添加缺失的源文件依赖，避免未来类似测试中断。

- CI 依赖管理缺陷 (testing)：需通过 #38792 添加缺失依赖，避免未来类似测试中断。

风险与影响

- 风险：风险极低：仅修改测试 mock，不涉及生产代码逻辑。但需确保 mock 替换完全正确，未遗漏其他相关测试。由于变更简单且针对性强，回归风险可忽略。
- 影响：影响范围仅限于 TRT-LLM 注意力相关测试的通过性。修复后确保测试能正确模拟生产环境行为，避免因测试失败误导开发或 CI 状态。对用户和系统无直接影响。
- 风险标记：测试 mock 更新

关联脉络

- PR #38730 [Bugfix] Fix test mocks after SM100 restriction in #38730: 本 PR 直接跟进 #38730，修复其引入的测试 mock 不一致问题。
- PR #38773 自动回滚 PR（具体标题未知）：本 PR 替代了自动回滚 PR #38773，说明原变更导致测试失败触发了自动回滚机制。
- PR #38792 添加缺失源文件依赖的 PR（具体标题未知）：作者在 Issue 评论中提及，该 PR 旨在修复 CI 依赖管理缺陷，避免未来类似测试中断。