PR #38167 完整报告

vllm-project/vllm

[ROCm][CI] Fix wvSplitKrc mock argument order in test_rocm_unquantized_gemm

合并时间：2026-03-26 19:55

原文链接：http://prhub.com.cn/vllm-project/vllm/pull/38167

执行摘要

修复了 ROCm 平台中一个测试函数的 mock 参数顺序错误，解决了因复制粘贴错误导致的 CI 失败，确保测试套件稳定运行。

功能与动机

本 PR 旨在修复 test_rocm_unquantized_gemm_gfx950_wvsplitkrc_path 测试的失败问题。根据 PR body 描述，失败原因是 mock 设置中的复制粘贴错误：wvSplitKrc mock 函数参数顺序与 C++ 实现不匹配（wvSplitK 和 wvSplitKrc 有相反的参数约定），导致返回张量形状错误，使 torch.allclose 断言失败。

实现拆解

变更仅涉及一个文件：tests/model_executor/layers/test_rocm_unquantized_gemm.py。具体修改如下：

- 将 wvsplitkrc_mock 的 lambda 参数顺序从 lambda w, x_view, _,__: x_view @ w.t() 调整为 lambda x_view, w, _,__: x_view @ w.t()。
- 此调整匹配 utils.py 中 ops.wvSplitKrc(x, weight, ...) 的调用方式（activations first），确保 mock 返回正确的张量形状（16x256 而非 256x16）。

评论区精华

review 讨论简洁，主要包括：

- gemini-code-assist[bot]：确认修复，无额外反馈。
- tjtanaa：批准合并，表示 LGTM（Looks Good To Me）。 未出现技术争议或深入讨论，表明修复直接明了且被团队接受。

风险与影响

风险分析：

- 低风险：仅修改测试代码，无生产逻辑变更。
- 潜在风险：mock 参数顺序错误可能隐藏真实问题，但本次修复已纠正。
- 无回归或兼容性问题，因为变更目标明确且单一。

影响评估：

- 对用户：无直接影响，因为是内部测试修复。
- 对系统：提升 CI 测试通过率，确保 ROCm 平台相关功能测试正确性。
- 对团队：维护代码质量，减少 CI flaky 失败，提高开发效率。

关联脉络

从近期历史 PR 分析看，本 PR 是 ROCm 平台测试维护的一部分，与以下 PR 关联：

- PR #38137 和 #38161：同为 ROCm 测试 bugfix，针对状态泄露和 flaky 行为，反映团队在加强测试稳定性。
- PR #38155：添加 ROCm 测试条目，扩展测试覆盖。 这些 PR 共同推动了 ROCm CI 的完善，表明 vLLM 项目在持续优化 AMD GPU 平台的测试套件。