

PR #46095 完整报告

vllm-project/vllm

[Model Runner V2] Fix MRv2 memory leak test

合并时间: 2026-06-19 08:36

原文链接: <http://prhub.com.cn/vllm-project/vllm/pull/46095>

执行摘要

- 一句话: 修复 MRv2 内存泄漏测试, 调大 warmup 轮次
- 推荐动作: 修改合理且安全, 无需深入分析。可快速合并。

功能与动机

MRv2 在第二轮 warmup 时会额外分配约 2MB 显存, 之后保持稳定, 测试误判为内存泄漏。PR 作者在 body 中说明: "MRv2 will allocate additional 2MB in second round warm up, then keep stable, which doesn't have memory leak at all."

实现拆解

仅修改测试配置: 将 `tests/models/multimodal/generation/test_memory_leak.py` 中的 `WARMUP_ROUNDS` 从 1 改为 2。其他所有逻辑、断言和阈值均保持不变。

关键文件:

- `tests/models/multimodal/generation/test_memory_leak.py` (模块 测试; 类别 test; 类型 test-coverage) : 核心变更: 调整 `WARMUP_ROUNDS` 从 1 到 2, 以消除 MRv2 第二轮 warmup 额外分配 2MB 引发的误报。

关键符号: 未识别

评论区精华

无 review 讨论。

- 暂无高价值评论线程

风险与影响

- 风险: 极低风险。仅调整 warmup 轮次, 不影响任何生产代码或测试逻辑的正确性。若未来 MRv2 出现真实内存泄漏, 此修改可能使首次 warmup 后的泄漏被掩盖 (需第二轮 warmup 后才检测), 但测试已包含多轮测量, 实际泄漏仍会被后续 round 的增长检出。
- 影响: 直接影响 MRv2 环境下的内存泄漏测试: 消除误报, 但可能略微降低对首轮后缓慢泄漏的敏感度。对其他模型或 V1 runner 无影响。
- 风险标记: 仅测试配置调整

关联脉络

- PR #44446 [Model Runner V2] Migration to support quantized model by default
[5/N]: MRv2 系列重构之一，本 PR 修复其引发的测试误报。