PR #46388 完整报告

vllm-project/vllm

[CI] Fix CPU-Multi-Modal Model Tests timeout by adding a 4th shard

合并时间：2026-06-23 06:08

原文链接：http://prhub.com.cn/vllm-project/vllm/pull/46388

执行摘要

- 一句话：CPU 多模态测试增加第 4 个分片避免超时
- 推荐动作：该 PR 是典型的 CI 配置优化，逻辑简单且数据支撑充分，值得合并。建议后续关注是否有其他测试套件存在类似的分片不均衡问题。

功能与动机

CPU-Multi-Modal Model Tests 1（shard 0）在慢 CI 节点上由于 45 分钟超时太紧而持续失败。PR body 提供 Build #73220 等证据：shard 0 运行 43+ 分钟仍有 4 个测试未完成，而其他分片仅运行约 22-30 分钟，分片负载严重不均。

实现拆解

1. 定位 YAML 配置：在 .buildkite/hardware_tests/cpu.yaml 中找到 CPU-Multi-Modal Model Tests 的步骤定义，该步骤使用 pytest-shard 根据 BUILDKITE_PARALLEL_JOB_COUNT 和 BUILDKITE_PARALLEL_JOB 分配测试用例。
2. 调整 parallelism 值：将 parallelism 从 3 改为 4，使每个分片承载约 10 个测试（原 3 分片时 shard 0 约 20 个）。
3. 验证效果：基于历史数据推算，慢节点（185s/ 测试）下 4 分片约 31 分钟完成，快节点（126s/ 测试）约 21 分钟完成，均远低于 45 分钟超时阈值。

关键文件：

- .buildkite/hardware_tests/cpu.yaml（模块 CI 配置；类别 infra；类型 test-coverage）：唯一变更文件，将 CPU-Multi-Modal Model Tests 的 parallelism 从 3 改为 4，以均衡分片负载避免超时。

关键符号：未识别

评论区精华

无 review 评论。khluu 直接批准了 PR，说明变更无争议。

- 暂无高价值评论线程

风险与影响

- 风险：低风险。仅修改 CI 并行分片数量，不会影响测试代码或运行时行为。唯一可能的风险是增加了一个 CI 分片，略微增加 CI 资源消耗，但对整体 CI 队列影响很小。
- 影响：影响范围小，仅针对 CPU-Multi-Modal Model Tests 的 CI 分片策略。预期解决该测试套件在慢 CI 节点上的超时问题，提高 CI 稳定性。
- 风险标记：暂无

关联脉络

- 暂无明显关联 PR