PR #43129 完整报告

vllm-project/vllm

[ci] Move language models tests (hybrid) back to L4

合并时间：2026-05-20 02:51

原文链接：http://prhub.com.cn/vllm-project/vllm/pull/43129

执行摘要

- 一句话：将混合语言模型测试从 H200 迁移回 L4
- 推荐动作：该 PR 属于运维性质的小调整，不建议精读代码。但值得关注的是，持续在 H200 上失败可能暗示更深层的兼容性或配置问题，建议后续跟进 H200 MIG 的根因。

功能与动机

根据 PR 描述，H200 MIG 上该测试持续失败（"It starts failing more on H200 MIG"），因此需要将设备切回之前稳定的 L4 以避免 CI 阻塞。

实现拆解

1. 定位配置文件：编辑 .buildkite/test_areas/models_language.yaml，该文件定义了 Buildkite CI 中语言模型测试的步骤。
2. 移除设备指定：删除 Language Models Tests (Hybrid) 步骤中的 device: h200_35gb 行（第 37 行），共 1 行删除，无新增行。
3. 效果：移除显式设备指定后，Buildkite 会为该步骤分配默认的 agent 类型（通常是 L4），从而避免使用不稳定的 H200 MIG 环境。

关键文件：

- .buildkite/test_areas/models_language.yaml（模块 CI 配置；类别 config；类型 configuration）：唯一变更文件，删除了 'Language Models Tests (Hybrid)' 步骤的 device: h200_35gb 配置，使测试回退到默认设备 L4。

关键符号：未识别

评论区精华

无实质性讨论：PR 无 review 评论，仅有一条来自 gemini-code-assist 的自动评论表示无反馈。

- 暂无高价值评论线程

风险与影响

- 风险：风险极低：仅涉及 CI 配置文件，无代码逻辑变更。回退到 L4 可能导致混合模型测试执行时间略有变化，但功能不受影响。如果 L4 资源紧张，可能造成队列等待，但不会引发

构建失败。

- 影响：
  - 对用户：无直接影响，用户不感知 CI 配置变更。
  - 对系统：缓解 H200 MIG 不稳定导致的 CI 失败，提高 CI 可靠性；L4 资源消耗增加。
  - 对团队：减少因基础设施问题导致的 PR 合入阻塞，提升开发效率。
  - 风险标记：基础设施变更

关联脉络

- 暂无明显关联 PR