

PR #41669 完整报告

vllm-project/vllm

[ROCm][CI] Remove benchmarks test group and shard long test groups

合并时间: 2026-05-23 23:31

原文链接: <http://prhub.com.cn/vllm-project/vllm/pull/41669>

执行摘要

- 一句话: 删除 ROCm CI benchmark 组并分片长测试组
- 推荐动作: 纯 CI 配置变更, 无需深入精读; 如需调整 ROCm CI 测试分组和分片策略, 可参考此 PR 的实现方式。

功能与动机

PR body 明确此为 PR#38611 的后续, 目标是删除 benchmarks 测试组并对长时间运行测试组分片, 以优化 ROCm CI 效率。

实现拆解

1. 删除 benchmarks 测试组: 移除了 mi250 和 mi300 两个硬件配置下的 Benchmarks 步骤 (约 20 行), 这些步骤原本运行 benchmarks/ 目录下的 run-benchmarks.sh 脚本。
2. 新增 models 子组并启用分片: 将原来单一的 models 测试组拆分为 'Basic Models Test (Other CPU)', 'Language Models Test' 和 'Multi-Modal Models' 三个子组, 后两个子组通过 BUILDKITE_PARALLEL_JOB 变量进行并行分片。
3. 调整属性: 为 Basic Models Test 添加 no_gpu: true 以在 CPU 节点执行, 新增 torch_nightly 和 optional 标志, 增强测试灵活性。

关键文件:

- .buildkite/test-amd.yaml (模块 CI 配置; 类别 config; 类型 configuration): 唯一变更文件, 实现 PR 全部功能

关键符号: 未识别

评论区精华

PR 通过 fork 提交, 自动化 AI 审查未产生实质讨论, 最终由 tjtanaa 直接批准合并。

- PR 审批流程 (other): 无需修改, 合并

风险与影响

- 风险: 低风险。删除 benchmarks 组可能导致基准测试回归不被及时发现; 分片引入 BUILDKITE_PARALLEL_JOB 变量, 若配置不当可能造成测试遗漏或重复, 但属于 CI 标准

做法，风险可控。

- 影响：正面影响 ROCm CI 效率：减少 2 个 benchmark 步骤，将 models 测试从串行转为并行，显著缩短 CI pipeline 总时长。对用户无影响，对 ROCm 开发者而言 CI 验收更快。
- 风险标记：CI 配置变更，低回归风险，效率改进

关联脉络

- PR #38611 [ROCm][CI] initial benchmarks group: 本 PR 作为 PR#38611 的后续，删除 benchmarks 组并分片长测试组