

PR #50329 完整报告

vllm-project/vllm

[CI Bugfix] Temp disable Humming wNa8 INT8 H100 CI

合并时间: 2026-07-30 07:14

原文链接: <http://prhub.com.cn/vllm-project/vllm/pull/50329>

执行摘要

- 一句话: 临时禁用 Humming INT8 H100 CI 测试
- 推荐动作: 这是一个临时修复, 建议关联待解决问题以便后续重新启用 INT8 测试。

功能与动机

Humming 模型在 Hopper GPU 上使用 INT8 激活时有已知的测试失败, 该 PR 旨在禁用该测试, 直到问题得到修复。参考: PR body 中的 [nightly failure](#) 链接。

实现拆解

在 `.buildkite/test_areas/lm_eval.yaml` 文件中, 将 H100 设备的 `config-act-int8.txt` 测试命令注释掉。

其他设备 (B200) 的 INT8 测试保持不变。

关键文件:

- `.buildkite/test_areas/lm_eval.yaml` (模块 CI 配置; 类别 config; 类型 configuration) : 仅修改此文件, 注释掉 H100 设备的 INT8 测试命令。

关键符号: 未识别

评论区精华

- 暂无高价值评论线程

风险与影响

- 风险: 风险极低: 仅临时禁用一个 CI 测试步骤, 不会影响任何运行时逻辑。
- 影响: 影响有限: H100 CI 将不再运行 Humming INT8 测试, 避免已知失败干扰, 但可能延迟发现问题。
- 风险标记: 暂无

关联脉络

- 暂无明显关联 PR