

PR #47067 完整报告

vllm-project/vllm

[ROCM][CI] Soft Fail `Spec Decode Ngram + Suffix` and `Entrypoints Integration (LLM)` AMD Mirrors

合并时间: 2026-06-30 05:24

原文链接: <http://prhub.com.cn/vllm-project/vllm/pull/47067>

执行摘要

- 一句话: 为 AMD CI 添加软失败配置
- 推荐动作: 该 PR 价值较低, 主要解决 CI 稳定性问题。如果关注 CI 配置或 AMD 平台状态, 可以了解; 否则无需深入。建议在 Torch 升级后及时移除 soft_fail 并修复底层问题。

功能与动机

Entrypoints Integration (LLM) 和 Spec Decode Ngram + Suffix 测试在使用 AMD 镜像的 CI 中持续失败, 阻塞了流水线。该变更将这两个测试设置为软失败, 允许流水线继续通过, 待 Torch 版本升级后再重新审视。

实现拆解

1. `.buildkite/test_areas/entrypoints.yaml`: 在 Entrypoints Integration (LLM) 步骤的 AMD 镜像配置下, 添加 `# TODO(akaratz): Test after Torch >= 2.12 bump` 注释和 `soft_fail: true`。
2. `.buildkite/test_areas/spec_decode.yaml`: 在 Spec Decode Ngram + Suffix 步骤的 AMD 镜像配置下, 添加相同的 TODO 注释和 `soft_fail: true`。
3. 两个变更均涉及 CI 配置文件的修改, 无源码变更。

关键文件:

- `.buildkite/test_areas/entrypoints.yaml` (模块 CI 配置; 类别 config; 类型 configuration) : 为 Entrypoints Integration (LLM) 步骤的 AMD 镜像添加了 `soft_fail: true` 和 TODO 注释, 使其软失败。
- `.buildkite/test_areas/spec_decode.yaml` (模块 CI 配置; 类别 config; 类型 configuration) : 为 Spec Decode Ngram + Suffix 步骤的 AMD 镜像添加了 `soft_fail: true` 和 TODO 注释, 使其软失败。

关键符号: 未识别

关键源码片段

`.buildkite/test_areas/entrypoints.yaml`

为 **Entrypoints Integration (LLM)** 步骤的 AMD 镜像添加了 `soft_fail: true` 和 TODO 注释, 使其软失败。

```
# .buildkite/test_areas/entrypoints.yaml 片段
mirror:
  amd:
    device: mi325_1
    # TODO(akaratz): Test after Torch >= 2.12 bump
    soft_fail: true # 允许该步骤失败而不阻塞流水线
    depends_on:
      - image-build-amd
```

.buildkite/test_areas/spec_decode.yaml

为 **Spec Decode Ngram + Suffix** 步骤的 AMD 镜像添加了 `soft_fail: true` 和 TODO 注释, 使其软失败。

```
# .buildkite/test_areas/spec_decode.yaml 片段
mirror:
  amd:
    device: mi325_1
    timeout_in_minutes: 65
    # TODO(akaratz): Test after Torch >= 2.12 bump
    soft_fail: true # 允许该步骤失败而不阻塞流水线
    depends_on:
      - image-build-amd
```

评论区精华

该 PR 没有 review 评论, 仅有一条来自 AndreasKaratzas 的审批意见, 内容为“LGTM let's keep a close eye on these and test them again after torch version bump”, 表达了同意并提醒密切关注, 待 Torch 升级后重新测试。

- 软失败配置的审批与后续跟进 (other): 同意合并, 同时计划在 Torch 升级后重新测试并可能移除软失败。

风险与影响

- 风险: 风险极低: 仅修改 CI 配置, 将两个测试步骤标记为软失败, 不会影响任何核心逻辑。但需要关注这些测试在 Torch 升级后的状态, 避免长期忽略导致回归。
- 影响:
 - 用户: 无影响。
 - 系统: CI 流水线对于 AMD 镜像不再因这两个测试失败而阻塞, 提高了 CI 的稳定性。
 - 团队: 减少 CI 噪音, 让开发者更专注于其他平台的失败; 但需后续跟进修复。
 - 风险标记: 临时软失败配置

关联脉络

- 暂无明显关联 PR