PR #42954 完整报告

vllm-project/vllm

[XPU][CI] Temporarily skip test_moe_lora_align_block_size_mixed_base_and_lora[1] in Intel GPU CI

合并时间：2026-05-18 20:34

原文链接：http://prhub.com.cn/vllm-project/vllm/pull/42954

执行摘要

- 一句话：暂时跳过 Intel GPU CI 中一个不稳定的 MoE LoRA 测试
- 推荐动作：这是一次临时性的、低风险的 CI 稳定性应急措施，不建议精读。但应提醒团队在后续尽快修复被跳过的测试用例，并恢复执行。

功能与动机

Intel GPU CI 中的一个 MoE LoRA 测试用例 test_moe_lora_align_block_size_mixed_base_and_lora[1] 存在不稳定因素，导致 CI 失败。为了不阻塞其他变更的合并，暂时跳过该测试，待后续修复后再恢复。

实现拆解

1. 修改 .buildkite/intel_jobs/lora_intel.yaml 文件，在第 52 行的 pytest 命令中添加 --deselect 选项，排除指定测试用例。
2. 该文件是 Intel GPU CI 的 Buildkite 配置，位于 .buildkite/intel_jobs/ 模块，属于配置类变更。

关键文件：

- .buildkite/intel_jobs/lora_intel.yaml（模块 CI 配置；类别 config；类型 configuration）：这是本次 PR 唯一变更的文件，通过添加 --deselect 参数跳过不稳定的测试用例，直接影响 Intel GPU CI 中 LoRA MoE 测试的执行行为。

关键符号：未识别

关键源码片段

.buildkite/intel_jobs/lora_intel.yaml

这是本次 PR 唯一变更的文件，通过添加 --deselect 参数跳过不稳定的测试用例，直接影响 Intel GPU CI 中 LoRA MoE 测试的执行行为。

```
# .buildkite/intel_jobs/lora_intel.yaml (LoRA Fused/MoE Kernels job)
    commands:
      - >-
        bash .buildkite/scripts/hardware_ci/run-intel-test.sh
        'cd tests &&
```

```
export VLLM_WORKER_MULTIPROC_METHOD=spawn &&
pytest -v -s lora/test_fused_moe_lora_kernel.py &&
# 跳过不稳定的测试用例，防止 CI 阻塞，后续需修复后恢复
pytest -v -s lora/test_moe_lora_align_sum.py --deselect="tests/lora/test_moe_lora_align_
sum.py::test_moe_lora_align_block_size_mixed_base_and_lora[1]"'
```

评论区精华

无实质讨论。只有自动机器人和审批人 jikunshang 的评论，审批人已批准。

- 暂无高价值评论线程

风险与影响

- 风险：风险极低：仅临时跳过一个测试用例，不影响任何生产代码。但需要关注该测试用例的根本原因，并在后续 PR 中修复后恢复。
- 影响：仅影响 Intel GPU CI 管道：该测试用例不再执行，其他测试不受影响，CI 稳定性得到提升。
- 风险标记：临时跳过测试可能掩盖根本问题

关联脉络

- PR #40131 [Bugfix] moe lora align kernel grid: 关联同一测试文件 test_moe_lora_align_sum.py，该 PR 修复了 MoE LoRA 对齐内核的 grid 越界问题，可能与本测试的不稳定性有关。