

PR #49385 完整报告

vllm-project/vllm

[BugFix][LoRA] Skip marlin-backend gpt-oss LoRA tests on XPU

合并时间: 2026-07-24 09:16

原文链接: <http://prhub.com.cn/vllm-project/vllm/pull/49385>

执行摘要

- 一句话: 跳过 XPU 上 Marlin 后端的 LoRA 测试
- 推荐动作: 这是一个简单、安全的测试修复, 与已有的平台跳过模式一致。值得作为 Intel GPU 持续集成稳定化的参考。建议快速合并。

功能与动机

Marlin MXFP4 kernel 是 CUDA 专属, 未在 XPU 上注册。测试中显式请求 `moe_backend="marlin"` 和 `linear_backend="marlin"` 会导致 XPU 上崩溃, 而自动路径会正确跳过 Marlin。PR 针对此问题增加 XPU 跳过条件以修复测试失败。

实现拆解

在 `tests/lora/test_gptoss_tp.py` 中, 对两个测试函数 `test_gpt_oss_lora` 和 `test_gpt_oss_lora_tp2` 的参数化标记 `mxfp4_use_marlin=True` 的 `pytest.mark.skipif` 条件中, 将 `current_platform.is_rocm()` 改为 `current_platform.is_rocm() or current_platform.is_xpu()`。变更仅涉及两处, 每处增加 2 行 (条件扩展和换行调整)。

关键文件:

- `tests/lora/test_gptoss_tp.py` (模块 LoRA 测试; 类别 test; 类型 test-coverage): 核心变更文件, 修改两处 `skipif` 条件增加 XPU 跳过, 解决 Marlin 后端在 XPU 上崩溃的问题。

关键符号: 未识别

关键源码片段

`tests/lora/test_gptoss_tp.py`

核心变更文件, 修改两处 `skipif` 条件增加 XPU 跳过, 解决 Marlin 后端在 XPU 上崩溃的问题。

```
# 对 test_gpt_oss_lora 的参数化标记修改
pytest.param(
    True,
    marks=pytest.mark.skipif(
        # 原只有 ROCm, 现增加 XPU, 因为 Marlin MXFP4 仅支持 CUDA
        current_platform.is_rocm() or current_platform.is_xpu(),
        reason="marlin not supported",
    ),
```

```
),  
  
# 对 test_gpt_oss_lora_tp2 的参数化标记修改（完全相同）  
pytest.param(  
    True,  
    marks=pytest.mark.skipif(  
        current_platform.is_rocm() or current_platform.is_xpu(),  
        reason="marlin not supported",  
    ),  
),
```

评论区精华

无 review 讨论。PR 获得 jikunshang 批准，无争议。

- 暂无高价值评论线程

风险与影响

- 风险：低风险。仅影响 XPU 平台的测试跳过逻辑，不会影响其他平台或生产代码。由于 XPU 从未支持 Marlin，跳过条件正确且与现有 ROCm 处理方式一致。
- 影响：影响范围仅限于 XPU 上运行 test_gpt_oss_lora 和 test_gpt_oss_lora_tp2 的 CI 任务。修复后，XPU 测试将跳过 Marlin 相关参数化组合（当 mxfp4_use_marlin=True 时），避免崩溃。用户和开发者不受影响。
- 风险标记：暂无

关联脉络

- 暂无明显关联 PR