PR #43571 完整报告

vllm-project/vllm

[BugFix][Platform] Fix import vllm.platforms.rocm error on non-CUDA test_gpt_oss.py

合并时间：2026-05-30 14:16

原文链接：http://prhub.com.cn/vllm-project/vllm/pull/43571

执行摘要

- 一句话：修复非 ROCm 平台导入 rocm 模块异常
- 推荐动作：值得阅读。该 PR 展示了处理平台特定导入的最佳实践：避免在模块导入时执行可能失败的硬件检测，使用条件导入和本地 fallback 函数。对于测试文件的平台兼容性处理有参考价值。设计决策清晰，讨论聚焦。

功能与动机

在非 CUDA torch 构建（如 XPU/CPU）下，pytest 在收集 tests/models/quantization/test_gpt_oss.py 测试时会因导入 vllm.platforms.rocm 而触发内部 AssertionError: Torch not compiled with CUDA enabled，导致整个测试无法运行。需要修复使得测试文件能够被收集并正常跳过。

实现拆解

1. 问题定位：vllm/platforms/rocm.py 在模块导入级别调用了 torch.cuda.get_device_properties，该函数仅在 CUDA-enabled 的 PyTorch 中可用。
2. 方案讨论：最初尝试在 rocm.py 中增加 try/except 包裹调用返回空字符串，但 reviewer 建议在测试文件中避免直接导入该模块，以免引入外部依赖。
3. 最终实现：在 tests/models/quantization/test_gpt_oss.py 中，将无条件导入替换为条件导入——通过 from vllm.platforms import current_platform 检查平台类型，若为 ROCm 则从 rocm 模块导入 on_gfx950，否则定义一个始终返回 False 的本地函数。
4. 结果：非 ROCm 构建下，测试文件收集不再崩溃，会被 pytest skipif 守卫跳过，输出 '8 skipped'。

关键文件：

- tests/models/quantization/test_gpt_oss.py（模块 GPT-OSS 测试；类别 test；类型 test-coverage；符号 on_gfx950）：唯一变更文件，通过条件导入修复了非 ROCm 平台下导入崩溃的问题。

关键符号：on_gfx950

关键源码片段

tests/models/quantization/test_gpt_oss.py

唯一变更文件，通过条件导入修复了非 ROCm 平台下导入崩溃的问题。

```
from vllm.platforms import current_platform

# 仅在当前平台为 ROCm 时从 rocm 模块导入 on_gfx950，
# 否则定义一个返回 False 的本地函数，避免非 CUDA 构建下的导入崩溃。
if current_platform.is_rocm():
    from vllm.platforms.rocm import on_gfx950
else:
    def on_gfx950() -> bool:
        return False
```

评论区精华

- jikunshang 最初建议不在 rocm.py 中添加 fallback，而是避免在测试文件中直接导入：'I think we should avoid call from vllm.platforms.rocm import on_gfx950 in test_gpt_oss.py instead of adding falling back.'
- AndreasKaratzas 提出了使用 current_platform.is_rocm() 的条件导入方式，并道歉之前他引入了这个疏忽：'I would probably write it like this: if current_platform.is_rocm(): ... else: ... Also I apologize for the oversight there.'
- 最终采用条件导入方案，未修改 rocm.py。
- 使用条件导入而非在 rocm.py 增加 fallback (design): 采纳了条件导入方案，在测试文件中根据平台条件导入 on_gfx950 或定义本地 fallback。

风险与影响

- 风险：风险极低。变更仅限于测试文件，且使用已封装的 current_platform API 进行判断，不涉及生产逻辑。主要风险在于如果未来 current_platform.is_rocm() 的行为发生变化（如平台检测逻辑修改），可能导致测试执行条件改变，但仅影响该测试文件的跳过行为，不影响其他。
- 影响：影响范围小，仅影响 tests/models/quantization/test_gpt_oss.py 文件的收集和执行。对非 CUDA 构建（XPU/CPU）的用户，该测试不再崩溃，增加测试可靠性。对 ROCm 用户无影响。对其他组件无影响。
- 风险标记：测试兼容性修复，条件导入

关联脉络

- 暂无明显关联 PR