

PR #46431 完整报告

vllm-project/vllm

[ROCm][CI] Skip Quark mxfp4 tests unless Quark version is compatible with Torch version

合并时间: 2026-06-23 22:26

原文链接: <http://prhub.com.cn/vllm-project/vllm/pull/46431>

执行摘要

- 一句话: 跳过 Quark MXFP4 测试在不兼容版本上
- 推荐动作: 这是一次典型的“测试基础设施兼容性”修补, 设计模式值得学习: 利用模块级别常量进行版本兼容性检查, 避免在每个测试函数中重复判断。维护者应关注 Quark 0.12.0 发布并安排后续 PR 移除该临时检查。建议未来将这种版本守卫抽取为公共工具函数 (如 `vllm/utils/compat.py`), 避免在多个测试文件间复制相同逻辑。

功能与动机

Quark 在 Torch 2.11 上导入失败, 导致 CI 测试全部崩溃。根因是 Torch 2.11 移除了 `torch.ao.quantization.pt2e` 模块, 而旧版本 Quark 仍然引用它。PR 描述中引用了 [AMD Quark issue #34](#), 该问题将在 Quark 0.12.0 中修复。社区团队选择在测试层面前置兼容性检查, 临时跳过不兼容的测试用例。

实现拆解

1. 在 `tests/evals/gsm8k/test_gsm8k_correctness.py` 中: 引入 `importlib.metadata`, `find_spec`, `torch`, `packaging.version` 等新导入; 新增模块级常量 `QUARK_MXFP4_TORCH_COMPATIBLE`, 其逻辑为: 若 Quark 包未安装则为 `False`, 若 Torch 版本 ≥ 2.11 则要求 Quark $\geq 0.12.0$, 否则 (Torch < 2.11) 直接判定为 `True` (兼容旧 Quark)。将原有 Qwen3.5 跳过条件从精确匹配 `"Qwen3.5-35B-A3B-MXFP4-AITER-TP2"` 放宽为仅匹配 `"Qwen3.5-35B-A3B-MXFP4"`, 然后根据是否含 `"AITER-TP2"` 分叉: 不含该后缀时, 若 `QUARK_MXFP4_TORCH_COMPATIBLE` 为假则跳过测试。
2. 在 `tests/kernels/moe/test_ocp_mx_moe.py` 中: 删除旧的 `QUARK_MXFP4_AVAILABLE` 常量 (仅检查 Quark $\geq 0.8.99$, 不区分 Torch 版本), 替换为与上面完全相同的 `QUARK_MXFP4_TORCH_COMPATIBLE` 常量定义; 将 `test_mxfp4_loading_and_execution_moe` 的 `skipif` 装饰器条件从旧常量切换为新常量, 跳过信息更新为 `"MXFP4 via quark requires amd-quark  $\geq 0.12$  on torch  $\geq 2.11$ "`。
3. 两个文件的改动模式一致: 均前置一个跨 Torch 版本的条件检查逻辑, 确保 `QUARK_MXFP4_TORCH_COMPATIBLE` 在模块加载时即可求值 (不需要运行时推理), 且在后续的所有 MXFP4 相关测试路径中应用该 `guard`。

关键文件:

- tests/evals/gsm8k/test_gsm8k_correctness.py (模块测试; 类别 test; 类型 test-coverage; 符号 QUARK_MXFP4_TORCH_COMPATIBLE, test_gsm8k_correctness) : 新增模块级版本兼容性常量 QUARK_MXFP4_TORCH_COMPATIBLE, 并修改 Qwen3.5 跳过条件逻辑, 是变更的核心文件之一。
- tests/kernels/moe/test_ocp_mx_moe.py (模块测试; 类别 test; 类型 test-coverage; 符号 QUARK_MXFP4_TORCH_COMPATIBLE, QUARK_MXFP4_AVAILABLE, test_mxfp4_loading_and_execution_moe) : 替换旧的 QUARK_MXFP4_AVAILABLE 常量为新兼容性常量, 调整 skipif 装饰器条件, 是另一个核心测试文件。

关键符号: test_gsm8k_correctness, test_mxfp4_loading_and_execution_moe

关键源码片段

tests/evals/gsm8k/test_gsm8k_correctness.py

新增模块级版本兼容性常量 QUARK_MXFP4_TORCH_COMPATIBLE, 并修改 Qwen3.5 跳过条件逻辑, 是变更的核心文件之一。

tests/evals/gsm8k/test_gsm8k_correctness.py (head 版本)

```
import importlib.metadata
from importlib.util import find_spec
import torch
from packaging import version
```

```
# MXFP4 via quark 要求: 若 torch >= 2.11 则必须 amd-quark >= 0.12
# 否则 (torch < 2.11) 旧版 quark 仍可用
# 参考 https://github.com/amd/Quark/issues/34
# TODO: 待 amd-quark >= 0.12.0 发布后移除
QUARK_MXFP4_TORCH_COMPATIBLE = find_spec("quark") is not None and (
    version.parse(importlib.metadata.version("amd-quark")) >= version.parse("0.12.0")
    if version.parse(torch.__version__.split("+")[0]) >= version.parse("2.11")
    else True
)
```

```
def test_gsm8k_correctness(config_filename):
    # ... 省略前面的检查 ...
    if current_platform.is_rocm() and ("Qwen3.5-35B-A3B-MXFP4" in config_filename.name):
        from vllm.platforms.rocm import on_gfx950
        # 仅对 AITER-TP2 后缀检查 GFX950 要求
        if not on_gfx950() and "AITER-TP2" in config_filename.name:
            pytest.skip(
                "Skipping Qwen3.5-35B-A3B-MXFP4-AITER-TP2 on non-GFX950 platforms. "
                "The quantization scheme is not supported on non-GFX950 platforms."
            )
        # 所有 Qwen3.5 MXFP4 配置均需检查 Quark/Torch 兼容性
        if not QUARK_MXFP4_TORCH_COMPATIBLE:
            pytest.skip(
                "Skipping Qwen3.5-35B-A3B-MXFP4: amd-quark >= 0.12 is required "
```

```
        "on torch >= 2.11."
    )
    # ... 后续保持不变 ...
```

tests/kernels/moe/test_ocp_mx_moe.py

替换旧的 QUARK_MXFP4_AVAILABLE 常量为新兼容性常量，调整 skipif 装饰器条件，是另一个核心测试文件。

```
# tests/kernels/moe/test_ocp_mx_moe.py (head 版本)

# MXFP4 via quark 要求: 若 torch >= 2.11 则必须 amd-quark >= 0.12
# 否则 (torch < 2.11) 旧版 quark 仍可用
# 参考 https://github.com/amd/Quark/issues/34
# TODO: 待 amd-quark >= 0.12.0 发布后移除
QUARK_MXFP4_TORCH_COMPATIBLE = find_spec("quark") is not None and (
    version.parse(importlib.metadata.version("amd-quark")) >= version.parse("0.12.0")
    if version.parse(torch.__version__.split("+")[0]) >= version.parse("2.11")
    else True
)

@pytest.mark.skipif(
    not QUARK_MXFP4_TORCH_COMPATIBLE,
    reason="MXFP4 via quark requires amd-quark >= 0.12 on torch >= 2.11.",
)

def test_mxfp4_loading_and_execution_moe(vllm_runner, model_case: ModelCase):
    # 测试体不变 ...
```

评论区精华

Review 中无实质性讨论线程。审核人 Andreaskaratzas 直接 Approve，仅留下 "LGTM"。社区对这次变更加入了一个 TODO 标记，提示将来 Quark 0.12.0 发布后应移除该兼容性检查。

- 暂无高价值评论线程

风险与影响

- 风险：风险很低。变更仅局限在测试文件内部，不涉及任何生产代码路径。主要技术风险：
 - 若后续 Quark 0.12.0 修复了 Torch 2.11 兼容性，但忘记移除该版本检查，将导致本应在新版 Quark 上运行的测试被错误跳过（CI 覆盖率漏报）。
 - 模块级 QUARK_MXFP4_TORCH_COMPATIBLE 在非 ROCm 平台也可能被求值，但额外开销可忽略；在当前 CUDA 平台上 Quark 不被安装，find_spec("quark") 直接返回 None，因此检查立即短路，不产生副作用。
 - 放宽 Qwen3.5 文件名匹配条件可能让一些意料之外的配置名也被跳过，但该文件下的 MXFP4 配置都依赖 Quark，所以实质安全性较高。
- 影响：影响范围：仅影响 ROCm CI 上所有依赖 Quark 的 MXFP4 测试，包括 GSM8K 评估测试和 MoE 加载执行测试。在 Torch 2.11 + Quark < 0.12.0 环境下，这些测试将静默跳过，替代原有的 ModuleNotFoundError 崩溃。对非 ROCm 平台、不使用

Quark 的测试无影响。影响程度：中等——解决了 CI 阻塞性问题，但降低了 ROCm MXFP4 测试覆盖率，直到 Quark 新版本发布。

- 风险标记：临时兼容性检查可能被遗忘移除，测试覆盖率暂时降低

关联脉络

- 暂无明显关联 PR