

PR #51980 完整报告

vllm-project/vllm

[Bugfix][ROCm][MoE] Update AITER MXFP4 W4A16 tests to the renamed expert_mask

合并时间: 2026-08-13 01:45

原文链接: <http://prhub.com.cn/vllm-project/vllm/pull/51980>

执行摘要

- 一句话: 测试适配 expert_mask 重命名, 修复 gfx950 报错
- 推荐动作: 建议快速合入。该 PR 本身价值不大, 但可作为 '接口重命名需同步测试调用点' 的典型示例, 提醒团队在改 API 时检查所有调用方 (包括测试)。同时值得确认 PR #49758 的重命名是否还有其他遗漏。

功能与动机

PR body 明确指出: `rocm_aiter_fused_experts() got an unexpected keyword argument 'expert_map'`, 根源是 #49758 重命名了关键字但未同步测试。同时, 两个测试都被 `on_gfx950()` 守卫, 导致 #49758 的 CI 未触发这些测试, 问题被隐藏。

实现拆解

1. 定位问题: 在 `tests/kernels/moe/test_rocm_aiter_moe.py` 的 `test_aiter_fused_moe_mi350_mxfp4_w4a16_accuracy` 和 `test_aiter_fused_moe_mi350_mxfp4_w4a16_determinism` 中, `rocm_aiter_fused_experts` 的调用使用了旧关键字 `expert_map`。
2. 修正调用: 将两处 `expert_map=None` 改为 `expert_mask=None`, 两个位置传参值均为 `None`, 只改关键字名, 不改变行为。
3. 验证: 在 MI355 (gfx950) 上运行 `pytest -v -s tests/kernels/moe/test_rocm_aiter_moe.py`, 修复前两个测试报 `TypeError`, 修复后整文件 31 个测试全部通过。
4. 未涉及生产源码、配置或文档变更, 属纯测试同步。

关键文件:

- `tests/kernels/moe/test_rocm_aiter_moe.py` (模块 MoE 测试; 类别 test; 类型 test-coverage; 符号 `test_aiter_fused_moe_mi350_mxfp4_w4a16_accuracy`, `test_aiter_fused_moe_mi350_mxfp4_w4a16_determinism`): 本次唯一改动文件, 修复两个 gfx950 测试对 `rocm_aiter_fused_experts` 接口的调用, 使其与 PR #49758 的重命名对齐。

关键符号: `test_aiter_fused_moe_mi350_mxfp4_w4a16_accuracy`,
`test_aiter_fused_moe_mi350_mxfp4_w4a16_determinism`

关键源码片段

tests/kernels/moe/test_rocm_aiter_moe.py

本次唯一改动文件，修复两个 gfx950 测试对 `rocm_aiter_fused_experts` 接口的调用，使其与 PR #49758 的重命名对齐。

```
# tests/kernels/moe/test_rocm_aiter_moe.py
# 本测试验证 gfx950 上 AITER MXFP4 W4A16 融合 MoE 的精度。
# PR #49758 将 rocm_aiter_fused_experts 的关键字 expert_map 重命名为 expert_mask,
# 但遗漏了这里的两个测试调用点，导致 TypeError。本 PR 修复这两处。
```

```
def test_aiter_fused_moe_mi350_mxfp4_w4a16_accuracy():
    from vllm.model_executor.layers.fused_moe.activation import MoEActivation
    from vllm.model_executor.layers.fused_moe.experts.rocm_aiter_moe import (
        rocm_aiter_fused_experts,
    )

    _assert_aiter_supported()
    case = _make_aiter_mxfp4_moe_case(
        num_tokens=32,
        hidden_dim=512,
        intermediate_dim=1024,
        num_experts=4,
        topk=2,
        seed=11,
    )
    ref_out = ref_moe_forward(
        case.hidden_states,
        case.w1_ref,
        case.w2_ref,
        case.topk_weights,
        case.topk_ids,
    )
    out = rocm_aiter_fused_experts(
        hidden_states=case.hidden_states,
        w1=case.w1_kernel,
        w2=case.w2_kernel,
        topk_weights=case.topk_weights,
        topk_ids=case.topk_ids,
        activation=MoEActivation.SWIGLUOAI,
        quant_config=case.quant_config,
        moe_config=case.moe_config,
        expert_mask=None, # 原为 expert_map, PR #49758 后必须使用 expert_mask
    )

    assert out.shape == case.hidden_states.shape
    _assert_close_budget(
        out.float(),
        ref_out.float(),
        label="mi350_mxfp4_w4a16_accuracy",
```

```
atol=0.1,  
rtol=0.1,  
pass_rate=0.99,  
max_violation_factor=2.0,  
)
```

评论区精华

该 PR 没有实质性的技术讨论。维护者 tjtananaa 和 AndreasKaratzas 直接批准 (LGTM) ; claude[bot] 因 fork 来源提示自动 review 已禁用。PR body 中作者特别指出 #49758 CI 未捕获该问题的原因：两个测试都被 `@pytest.mark.skipif(not on_gfx950(), ...)` 守卫，而 CI 环境并非 gfx950。

- Fork 自动 review 状态 (other): 维护者 tjtananaa 与 AndreasKaratzas 直接批准，无需额外 review。

风险与影响

- 风险：风险极低：改动仅限测试文件中的两个关键字，且传参值均为 None，不改变测试语义。潜在问题在于若 rocm_aiter_fused_experts 未来再次改接口，类似遗漏仍可能发生，建议在重构公共 API 时对测试目录同样做全局搜索。此外，由于测试仅在有 gfx950 的硬件上运行，普通 CI 无法覆盖，需依赖 ROCm 专属 CI。
- 影响：影响范围仅限 ROCm gfx950 平台上的 AITER MoE 测试。修复后恢复 MXFP4 W4A16 精度与确定性测试的有效性，避免测试套件报错。对生产推理路径无任何影响。
- 风险标记：接口重命名遗漏，gfx950 测试覆盖不足

关联脉络

- PR #49758 Rename expert_map to expert_mask in AITER MoE: PR body 引用 #49758 重命名了 rocm_aiter_fused_experts 的关键字，本 PR 是其遗漏测试调用点的补丁。