

PR #49690 完整报告

vllm-project/vllm

[CI][ROCm] Fix `test\_ocp\_mx\_wikitext\_correctness` reference value

合并时间: 2026-07-28 05:47

原文链接: <http://prhub.com.cn/vllm-project/vllm/pull/49690>

执行摘要

- 一句话: 调整 MXFP4 测试参考值至 12.45
- 推荐动作: 简单查看即可, 无需深入。这是一个典型的数值回退测试调整, 值得注意的设计决策是更新期望值, 而不是降低容差或禁用测试。这保持了 CI 的可信度。

功能与动机

`test_ocp_mx_wikitext_correctness` 在 AMD CI 上持续失败 (如 Buildkite 构建 #11168), 实际测量值 12.5137 与期望值 12.4 不符。PR #46643 的 `moe_sum` 内核变更引入了无害的数值噪声, 但测试需要更新参考值以适应新的累加精度。

实现拆解

1. 在 `tests/quantization/test_quark.py` 的第 282 行, 将 `excepted_value` 从 12.4 改为 12.45。- 该值位于 `WIKITEXT_ACCURACY_CONFIGS` 列表中, 对应 `AccuracyTestConfig(model_name="fxmarty/qwen_1.5-moe-a2.7b-mxfp4", excepted_value=12.45)`。
2. 新值 12.45 略高于原始值 12.4, 但仍在 `rtol=0.1` 的容差范围内, 确保测试通过。

关键文件:

- `tests/quantization/test_quark.py` (模块 量化测试; 类别 test; 类型 test-coverage) : MXFP4 准确率测试中 Qwen1.5-MoE-A2.7B 配置的期望值从 12.4 改为 12.45, 以适配 `moe_sum` 内核变更后的数值行为。

关键符号: 未识别

关键源码片段

`tests/quantization/test_quark.py`

MXFP4 准确率测试中 Qwen1.5-MoE-A2.7B 配置的期望值从 12.4 改为 12.45, 以适配 `moe_sum` 内核变更后的数值行为。

```
# tests/quantization/test_quark.py, lines 272-284
# WIKITEXT_ACCURACY_CONFIGS 包含三个 MXFP4 量化模型的准确率期望值
WIKITEXT_ACCURACY_CONFIGS = [
    AccuracyTestConfig(
```

```

        model_name="fxmarty/qwen1.5_moe_a2.7b_chat_w_fp4_a_fp6_e2m3",
        excepted_value=11.3,
    ),
    AccuracyTestConfig(
        model_name="fxmarty/qwen1.5_moe_a2.7b_chat_w_fp6_e3m2_a_fp6_e3m2",
        excepted_value=10.6,
    ),
    # PR#49690: 因 moe_sum 内核累加精度改为 float 后数值略升, 期望值从 12.4 → 12.45
    AccuracyTestConfig(
        model_name="fxmarty/qwen_1.5-moe-a2.7b-mxfp4", excepted_value=12.45
    ),
]

```

评论区精华

该 PR 没有 review 评论或公开讨论。审核者 tjanaa 直接批准。

- 暂无高价值评论线程

风险与影响

- 风险：风险极低：仅更改测试期望值，不影响生产代码。数值变化由内核累积精度改进引起，已有独立的 kernel 级测试覆盖（见 tests/kernels/moe/test_moe.py）。
- 影响：影响范围仅限 AMD CI 上的 test_ocp_mx_wikitext_correctness 测试。修复后该测试将稳定通过。
- 风险标记：测试期望值漂移

关联脉络

- PR #46643 [Kernel] Change moe_sum accumulation from bf16 to float for precision: 该 PR 的 moe_sum 内核变更（改为 float 累加再转 bf16）导致了数值变化，是本次测试期望值调整的根本原因。
- PR #46765 [ROCm][Quantization][5/N] Refactor quark_moe w8a8-int8 w/ oracle: 共享 tests/quantization/test_quark.py 文件，属于同一测试基础设施。
- PR #49745 [Refactor] Remove dead code in multiple files: 共享同一测试文件，有代码清理重叠。