# PR #37418 完整报告

vllm-project/vllm

[Bugfix][ROCm] Fix MoRI + AITER FP8 dispatch compatibility for defer_input_quant

合并时间：2026-03-19 17:49

原文链接：http://prhub.com.cn/vllm-project/vllm/pull/37418

## 执行摘要

- 一句话：修复 ROCm 平台 MoRI 与 AITER 后端 FP8 量化分发不兼容的 bug。
- 推荐动作：建议精读此 PR，了解 MoE 架构中 FP8 量化处理的设计权衡，特别是如何通过条件化属性和异常移除实现后端兼容。关注 AiterExperts.expects_unquantized_inputs 的条件逻辑和 MoriPrepareAndFinalize.prepare 中的量化跳过机制。

## 功能与动机

MoE Oracle 重构后（#32414, #32567），AiterExperts.expects_unquantized_inputs 无条件返回 True，导致 defer_input_quant=True 被传递给 MoriPrepareAndFinalize.prepare()，后者引发 NotImplementedError，使得无法在 ROCm 平台上使用 MoRI + AITER 的 FP8 量化模型。PR body 明确指出需解决此兼容性问题。

## 实现拆解

实现涉及两个文件修改：在 rocm_aiter_fused_moe.py 中，将 AiterExperts.expects_unquantized_inputs 属性从无条件返回 True 改为条件返回 not self.moe_config.use_mori_kernels，这样当使用 MoRI 时返回 False；在 mori_prepare_finalize.py 中，移除 defer_input_uant 时的 NotImplementedError，并在 use_fp8_dispatch 且 defer_input_quant=False 时执行 FP8 量化，否则跳过以支持内部量化的后端。

关键文件：

- vllm/model_executor/layers/fused_moe/mori_prepare_finalize.py（模块 fused_moe）：移除 NotImplementedError 并修改 FP8 量化逻辑，核心处理 MoRI 的准备步骤，直接影响 FP8 分发兼容性。
- vllm/model_executor/layers/fused_moe/rocm_aiter_fused_moe.py（模块 fused_moe）：条件化 AiterExperts.expects_unquantized_inputs 属性，协调 MoRI 与 AITER 间的量化责任，是关键控制点。

关键符号：AiterExperts.expects_unquantized_inputs, MoriPrepareAndFinalize.prepare

## 评论区精华

review 评论中，gemini-code-assist[bot] 确认变更逻辑正确，解决了兼容性问题，并指出变更协调了量化步骤；tjtanaa 批准。无争议点或未解决疑虑，讨论简短且一致通过。

- 代码逻辑正确性确认 (correctness): 变更被接受，无异议，reviewer tjtanaa 批准。

## 风险与影响

- 风险：风险包括：修改核心 MoE 量化逻辑可能影响其他配置，尤其在 defer_input_quant=True 时跳过 FP8 分发量化需确保与后端的正确协调；条件化属性可能在不同 MoE 配置下行为不一致，需测试覆盖；FP8 量化跳过可能引入精度问题。具体文件 mori_prepare_finalize.py 中逻辑变更需确保在 defer_input_quant=True 时正确跳过量化。
- 影响：对用户而言，修复了 FP8 量化 MoE 模型在 ROCm 平台上的使用障碍，使 MoRI 能与 AITER 后端协同工作，提升模型部署灵活性。系统层面，增强了 MoE 模块对不同后端的兼容性，优化网络流量减少。团队影响小，但解决了重要 bug，确保 ROCm 平台功能完整性。
- 风险标记：FP8 量化逻辑修改，条件化属性可能引入歧义，缺少广泛测试覆盖

## 关联脉络

- PR #37529 [ROCm] Enable MORI EP for unquantized MoE with AITER backend: 同涉及 ROCm 平台下 MoRI 与 AITER 后端的兼容性问题修复，显示持续优化 MoE 模块。