# PR #45067 完整报告

vllm-project/vllm

[Bugfix]: Fix Quark gpt-oss weight loading broken by FusedMoe refactor

合并时间：2026-06-11 09:17

原文链接：http://prhub.com.cn/vllm-project/vllm/pull/45067

## 执行摘要

- 一句话：修复 GPT-OSS 权重加载因 MoE 重构导致的崩溃
- 推荐动作：推荐合并。改动极小、风险可控，且解决了用户可复现的模型启动崩溃问题。虽然没有新增测试，但现有 CI 和手动回归可以确保不引入新问题。

## 功能与动机

PR #41184 的 FusedMoe 重构将 expert 相关参数移到了 RoutedExperts 类下，导致 GPT-OSS 模型在加载旧格式 checkpoint 时，因参数路径不匹配而崩溃。Issue 评论中提供了一个可复现的例子：使用 amd/gpt-oss120b-w-mxfp4-a-fp8 模型在 MI355X 上运行 vllm serve 会触发 KeyError，阻止模型启动。

## 实现拆解

1. 定位问题：在 vllm/model_executor/models/gpt_oss.py 的 load_weights 方法中，当 checkpoint 的权重名称包含 experts 时，原有路径（如 mlp.experts.w2_bias）无法匹配重构后的实际参数路径（mlp.experts.routed_experts.w2_bias），导致 params_dict 查询失败。
2. 修复方式：在 fused_name 赋值后、 params_dict 查询前，添加一行 replace 调用，将路径中的 .mlp.experts. 替换为 .mlp.experts.routed_experts.。该修复精确且最小化，仅影响 expert 相关权重的名称解析。
3. 演进过程：初始提交尝试了更复杂的改动，随后逐步简化，最终只使用一行 replace 完成修复，避免对现有逻辑产生额外影响。

关键文件：

- vllm/model_executor/models/gpt_oss.py（模块 模型加载；类别 source；类型 data-contract）：修复的核心文件，在权重加载时增加路径重映射，确保旧 checkpoint 名称能匹配重构后的 RoutedExperts 参数结构。

关键符号：未识别

## 关键源码片段

### vllm/model_executor/models/gpt_oss.py

修复的核心文件，在权重加载时增加路径重映射，确保旧 checkpoint 名称能匹配重构后的 RoutedExperts 参数结构。

```
# vllm/model_executor/models/gpt_oss.py (load_weights method)
# The MoE refactor (#41184) moved expert params under
# `mlp.experts.routed_experts.*`; remap the legacy checkpoint
# name so keys like w2_bias resolve against params_dict.
fused_name = fused_name.replace(
    ".mlp.experts.", ".mlp.experts.routed_experts."
)
```

## 评论区精华

Reviewer bnellnm 指出：“最理想的修复方式是使用 expert_params_mapping 来确保路径正确”，但同时也认可当前改动“就当前情况而言是可以的”。bnellnm 还表示短期内不会再次调整 FusedMoe 对象结构，因此该 hack 是合理且稳定的。最后的 merge 由 AndreasKaratzas 完成。

- 修复方式与 expert_params_mapping 的关系 (design): 当前修复方案被接受，后续若再次调整 FusedMoe 结构需考虑使用更通用的方法。

## 风险与影响

- 风险：风险极低。改动仅一行 replace，且路径替换只在 expert 权重加载分支内生效，不影响其他模型。但是，如果有模型同时使用 mlp.experts. 但不是来自 FusedMoe 重构后的 RoutedExperts，该替换可能导致路径错误。不过考虑到 GPT-OSS 是唯一受影响的模型，且其 checkpoint 格式已明确定义，风险可控。
- 影响：直接影响：修复了 GPT-OSS 模型（特别是 AMD Quark 量化版本，如 amd/gpt-oss120b-w-mxfp4-a-fp8）在 ROCm 平台上的加载崩溃。间接影响：为其他可能受 FusedMoe 重构影响的旧模型提供了修复参考。影响范围较小，仅涉及一个模型文件。
- 风险标记：暂无回归风险，缺少测试覆盖

## 关联脉络

- PR #41184 FusedMoe refactor: 引发本 bug 的根源 PR，将 expert 参数移至 RoutedExperts 子类，导致旧路径失效。
- PR #45037 Fix nemotron accuracy drop introduced by #41184: 同样受 #41184 影响，在相同文件（fused_moe/layer.py）中修复了精度问题，属于同类回归。