# PR #45037 完整报告

vllm-project/vllm

[Bugfix] Fix nemotron accuracy drop introduced by #41184

合并时间：2026-06-11 01:16

原文链接：http://prhub.com.cn/vllm-project/vllm/pull/45037

## 执行摘要

- 一句话：修复 Nemotron 模型精度下降问题
- 推荐动作：值得立即合并（已合并）。对于维护 MoE 模型的开发者，建议将 FusedMoE 中类似的一处缩放逻辑写成更清晰的注释或提取为辅助函数，防止未来再次引入双重缩放。

## 功能与动机

PR body 明确指出现象：Nemotron 模型的 routed_scaling_factor 被应用了两次，导致精度下降。修复需检查 apply_routed_scaling_factor_to_output 标志，只在 RoutedExperts 与 MoERunner 之一处缩放。Issue 评论中 GirasoleY 验证了 Kimi-K2.5-NVFP4 模型也受相同问题影响，准确率从 25% 恢复到 92%，表明该 bug 影响范围超出 Nemotron。

## 实现拆解

1. 定位问题：在 vllm/model_executor/layers/fused_moe/layer.py 的 FusedMoE 工厂函数中，调用 RoutedExperts 构造时始终传递了原始的 routed_scaling_factor 参数。而下游 MoERunner 在 apply_routed_scale_to_output=True 时也会应用该因子，导致缩放被应用两次。
2. 修改变更：将传递 routed_scaling_factor 给 RoutedExperts 的逻辑改为条件赋值：当 apply_routed_scale_to_output 为 True 时，传入 1.0（即禁用 RoutedExperts 内部缩放），否则仍传入原始值。
3. 影响范围：仅修改 layer.py 中一行表达式，+3/-1；不改动任何接口或数据流，属于轻量修复。未同步更新测试文件，但 PR body 中附有 GSM8K 验证结果。

关键文件：

- vllm/model_executor/layers/fused_moe/layer.py（模块 MoE 层；类别 source；类型 data-contract）：修复的核心文件：调整 RoutedExperts 初始化时 routed_scaling_factor 的传递逻辑，+3/-1，是唯一变更的文件。

关键符号：FusedMoE

## 关键源码片段

### vllm/model_executor/layers/fused_moe/layer.py

修复的核心文件：调整 RoutedExperts 初始化时 routed_scaling_factor 的传递逻辑，+3/-1，是唯一变更的文件。

```
# vllm/model_executor/layers/fused_moe/layer.py
# 在 FusedMoE 工厂函数中，构造 RoutedExperts 时传递 routed_scaling_factor
# 修复前：始终传递原始 routed_scaling_factor
# 修复后：当 apply_routed_scale_to_output=True 时，RoutedExperts 应接收 1.0
# 因为此时缩放完全由 MoERunner 负责，避免双重缩放
routed_experts = routed_experts_cls(
    layer_name,
    params_dtype,
    moe_config,
    quant_config,
    expert_map_manager=expert_map_manager,
    expert_mapping=expert_mapping,
    renormalize=renormalize,
    use_grouped_topk=use_grouped_topk,
    num_expert_group=num_expert_group,
    topk_group=topk_group,
    custom_routing_function=custom_routing_function,
    scoring_func=scoring_func,
    # 关键修复：当 apply_routed_scale_to_output 为 True 时，
    # 传入 1.0 避免 RoutedExperts 内部重复应用缩放因子
    routed_scaling_factor=routed_scaling_factor
    if not apply_routed_scale_to_output
    else 1.0,
    swiglu_limit=swiglu_limit,
    e_score_correction_bias=e_score_correction_bias,
    apply_router_weight_on_input=apply_router_weight_on_input,
    **routed_experts_args if routed_experts_args is not None else {},
)

# 下游 MoERunner 构造时，当 apply_routed_scale_to_output=True 时仍会应用缩放因子
# 两者配合实现了缩放因子只应用一次的正确行为
runner = runner_cls(
    layer_name=layer_name,
    moe_config=moe_config,
    router=router,
    routed_experts=routed_experts,
    enable_dbo=vllm_config.parallel_config.enable_dbo,
    gate=gate,
    shared_expert_gate=shared_expert_gate,
    shared_experts=shared_experts,
    routed_input_transform=routed_input_transform,
    routed_output_transform=routed_output_transform,
    # 当 apply_routed_scale_to_output 为 True 时，
    # MoERunner 应用缩放因子；否则传入 1.0 使得缩放不起作用
    routed_scaling_factor=routed_scaling_factor
    if apply_routed_scale_to_output
```

```
        else 1.0,
        **runner_args if runner_args is not None else {},
    )
```

## 评论区精华

无 Review 评论。但 Issue 评论中 GirasoleY 提供了额外的复现验证：Kimi-K2.5-NVFP4 模型（与 Nemotron 共享 FusedMoE 路径）在该修复下准确率从 25% 提升到 92%，证实了根因分析的正确性和修复的通用性。

- 暂无高价值评论线程

## 风险与影响

- 风险：风险较低。修改仅控制单个参数在 RoutedExperts 构造函数中的传递条件，逻辑清晰。但需要注意：
  - 如果其他模型自定义了 RoutedExperts 子类并依赖原始 routed_scaling_factor 行为，可能受此变更影响（但 apply_routed_scale_to_output 为 True 时本意就是将缩放权交由 MoERunner，因此此修复符合原始设计意图）。
  - 缺少直接针对该修复的单元测试覆盖。
  - 影响：影响范围：直接修复 Nemotron 模型精度问题；同时修复所有使用 FusedMoE 且 apply_routed_scale_to_output=True 的模型（如 Kimi-K2.5-NVFP4）的相同 bug。对不设置该标志的模型无影响。修复后模型输出与预期一致，精度恢复。
  - 风险标记：缺少测试覆盖

## 关联脉络

- PR #41184 introduce routed_scaling_factor to MoE layer: 该 PR 引入了 routed_scaling_factor 导致本 PR 修复的双重缩放 bug。