PR #41401 完整报告

vllm-project/vllm

[Bugfix] Fix RoutedExpertsCapturer for Gemma 4 MoE (top_k_experts)

合并时间：2026-05-01 07:19

原文链接：http://prhub.com.cn/vllm-project/vllm/pull/41401

执行摘要

- 一句话：修复 Gemma 4 MoE 专家捕获器配置键不兼容
- 推荐动作：此 PR 值得快速合并，修复明确且影响范围小。建议未来在模型配置兼容性改进中考虑添加单元测试（如测试 _get_num_experts_per_tok 对不同配置的响应），避免类似回归。

功能与动机

Gemma 4 模型系列使用 top_k_experts 而非 num_experts_per_tok 存储每 token 专家数，直接访问 hf_config.num_experts_per_tok 会导致 AttributeError。PR 描述引用关联 Issue #39066，指出相同根因已在 moe.py 中修复（PR #39067），但未覆盖 routed_experts_capturer.py。

实现拆解

1. 新增辅助函数 _get_num_experts_per_tok：位于 routed_experts_capturer.py 文件顶部，通过 getattr 按 num_experts_per_tok → top_k_experts 顺序回退取值，若两者均不存在则抛出 ValueError。
2. 替换 init_buffer 中的硬编码访问：将第 148 行从 hf_config.num_experts_per_tok 改为 _get_num_experts_per_tok(hf_config)，用于设备缓冲区和共享内存缓冲区的大小计算。
3. 替换 attach_buffer 中的硬编码访问：将第 321 行从 hf_config.num_experts_per_tok 改为 _get_num_experts_per_tok(hf_config)，用于共享内存缓冲区形状定义。
4. 无测试配套改动：PR 未包含直接针对 Gemma 4 或兼容性检查的单元测试或 e2e 测试。

关键文件：

- vllm/model_executor/layers/fused_moe/routed_experts_capturer.py（模块 MoE 层；类别 source；类型 data-contract；符号 _get_num_experts_per_tok）：唯一变更文件，新增辅助函数并替换两处属性访问，是修复核心所在。

关键符号：_get_num_experts_per_tok

关键源码片段

vllm/model_executor/layers/fused_moe/routed_experts_capturer.py

唯一变更文件，新增辅助函数并替换两处属性访问，是修复核心所在。

```
def _get_num_experts_per_tok(hf_config) -> int:
    """Resolve the per-token expert count from the HF config.

    Different model families store this under different attribute names
    (e.g. ``num_experts_per_tok`` for DeepSeek, ``top_k_experts`` for Gemma 4).
    """
    # 优先读取 num_experts_per_tok（DeepSeek, Qwen MoE 等标准模型）
    val = getattr(hf_config, "num_experts_per_tok", None)
    if val is None:
        # 回退读取 top_k_experts（Gemma 4 使用此属性）
        val = getattr(hf_config, "top_k_experts", None)
    if val is None:
        raise ValueError(
            "Cannot determine num_experts_per_tok: HF config has neither "
            "'num_experts_per_tok' nor 'top_k_experts'"
        )
    return val
```

评论区精华

审查人 gemini-code-assist[bot] 确认辅助函数能支持多种属性名，无其他反馈。维护者 ywang96 直接批准，无额外讨论。

- 暂无高价值评论线程

风险与影响

- 风险：低风险。变更集中在一个文件，仅替换属性访问方式，不影响模型加载或推理核心路径。但未添加测试覆盖，若未来有其他模型使用新属性名，可能再次引发类似问题。
- 影响：直接影响：使用 --enable-return-routed-experts 运行 Gemma 4 MoE 模型时，不再因 AttributeError 崩溃，且已路由专家信息能正确返回。不影响使用标准属性 num_experts_per_tok 的模型（如 DeepSeek、Qwen MoE）。
- 风险标记：缺少测试覆盖

关联脉络

- PR #39067 Fix Gemma 4 MoE expert weight remapping: PR #39067 修复了同一根因（top_k_experts vs num_experts_per_tok），但仅限于 moe.py（模型加载路径），未覆盖 routed_experts_capturer.py。
- PR #41206 Fix Gemma4 MoE expert weight remapping: 另一 Gemma 4 MoE 修复，涉及权重重映射，属于同一模型系列兼容性修复。