PR #45818 完整报告

vllm-project/vllm

[Bugfix]: Fix unquantized gpt-oss weight loading broken by FusedMoE r···

合并时间：2026-06-24 00:56

原文链接：http://prhub.com.cn/vllm-project/vllm/pull/45818

执行摘要

- 一句话：修复非量化 GPT-OSS 检查点权重加载 KeyError
- 推荐动作：值得合并。修复明确，风险低。建议后续添加测试覆盖未量化 GPT-OSS 权重加载路径。

功能与动机

修复 Issue #45830：FusedMoE 重构（PR#41184）将专家权重参数从 mlp.experts.<w> 重命名为 mlp.experts.routed_experts.<w>，但 _load_weights_other 未更新，导致加载标准（非 mxpf4/ 非 quark）GPT-OSS 检查点时抛出 KeyError。

实现拆解

1. 在 vllm/model_executor/models/gpt_oss.py 的 _load_weights_other 方法中，将迭代 weights 的循环改为调用 remap_moe_expert_weights(weights, params_dict)。该函数仅重新映射已知的专家参数，并在嵌套名称存在于 params_dict 时进行替换，保持向后兼容。
2. 移除之前可能存在的本地字符串替换逻辑（patch 中未体现，但原代码采用直接遍历）。

关键文件：

- vllm/model_executor/models/gpt_oss.py（模块 模型加载；类别 source；类型 data-contract）：核心修复文件，修改 _load_weights_other 方法中的循环迭代逻辑。

关键符号：_load_weights_other, remap_moe_expert_weights

关键源码片段

vllm/model_executor/models/gpt_oss.py

核心修复文件，修改 _load_weights_other 方法中的循环迭代逻辑。

```
# vllm/model_executor/models/gpt_oss.py ( 部分 )
def _load_weights_other(self, ep_rank_end, ep_rank_start, heads_per_rank, head_start, weights,
stacked_params_mapping):
    params_dict = dict(self.named_parameters())
    loaded_params: set[str] = set()
    # ... 省略切片计算 ...
    # 修复：使用 remap_moe_expert_weights 重新映射专家权重参数名
    # 原代码为 for name, weight in weights:
```

```
for name, weight in remap_moe_expert_weights(weights, params_dict):
    if is_pp_missing_parameter(name, self):
        continue
    # ... 原有权重处理逻辑 ...
```

评论区精华

无 review 讨论。PR 提交后获得 4 个 approve，无争议。

- 暂无高价值评论线程

风险与影响

- 风险：风险低。变更仅涉及一行循环迭代的替换，且 remap_moe_expert_weights 已在 Quark 量化路径中经过测试。但缺少针对未量化 GPT-OSS 检查点的回归测试，若未来参数命名规则再次变更可能遗漏。
- 影响：影响范围小，仅修复未量化 GPT-OSS 模型的权重加载。对量化路径无影响。
- 风险标记：缺少测试覆盖

关联脉络

- PR #41184 FusedMoE inversion refactor: 是本 PR 修复的根因重构，将专家权重移入 routed_experts 子模块。