PR #42521 完整报告

vllm-project/vllm

[Fix] Weight loading for qwen3_5 using runai_streamer

合并时间：2026-05-14 10:36

原文链接：http://prhub.com.cn/vllm-project/vllm/pull/42521

执行摘要

- 一句话：修复 Qwen3.5 权重加载参数传递问题
- 推荐动作：值得快速合并，修复明确，改动极小。

功能与动机

修复 Qwen3.5 MoE 模型在 TPU 多主机上使用 runai_streamer 加载权重时的 AssertionError，根源是位置参数与下游 hook 的 kwargs 期望不匹配。

实现拆解

1. 在 vllm/model_executor/models/qwen3_5.py 的 load_fused_expert_weights 方法中，将第 265-266 行的 shard_id, expert_id 改为 shard_id=shard_id, expert_id=expert_id 作为关键字参数传递。
2. 仅修改了 2 行代码，无其他文件变更。

关键文件：

- vllm/model_executor/models/qwen3_5.py（模块 模型加载；类别 source；类型 data-contract；符号 load_fused_expert_weights）：修复权重加载参数传递问题，避免 runai_streamer 场景下的 AssertionError。

关键符号：load_fused_expert_weights

关键源码片段

vllm/model_executor/models/qwen3_5.py

修复权重加载参数传递问题，避免 runai_streamer 场景下的 AssertionError。

```
# vllm/model_executor/models/qwen3_5.py
# 修复前：shard_id 和 expert_id 作为位置参数传递
# 修复后：改为关键字参数，确保下游 weight_loader hook
#  （如 maybe_process_weights）能通过 kwargs.get('expert_id') 获取到值
success = weight_loader(
    param,
    curr_expert_weight,
    name,
    shard_id=shard_id, # 原为位置参数，现改为关键字参数
```

```
    expert_id=expert_id, # 同上
    return_success=True,
)
```

评论区精华

无实质性讨论，ZJY0516 直接批准。

- 暂无高价值评论线程

风险与影响

- 风险：风险极低：仅将两个位置参数改为关键字参数，语义等价，不会影响未使用 kwargs 调用的 weight_loader。
- 影响：影响范围：仅影响 Qwen3.5 模型在使用 runai_streamer 加载方式时的权重加载流程，修复了特定场景下的崩溃。对正常加载路径无影响。
- 风险标记：暂无

关联脉络

- 暂无明显关联 PR