PR #32022 完整报告

sgl-project/sglang

fix(qwen3.5): restrict MoE weights to local PP layers

合并时间：2026-07-30 03:01

原文链接：http://prhub.com.cn/sgl-project/sglang/pull/32022

执行摘要

- 一句话：限制 Qwen3.5 MoE 权重获取到本地 PP 层
- 推荐动作：值得合并的快速修复。对于维护 PP 或 MoE 模型的工程师，可关注此模式下其他模型是否也存在类似问题。

功能与动机

Qwen3.5 在多 PP rank 场景下，load_fused_expert_weights 中的 _routed_experts_weights_of_layer 使用 enumerate(self.model.layers) 遍历所有层，导致其他 PP 阶段的 PPMissingLayer 占位符被访问（这些占位符没有 mlp 属性），从而引发错误。PR body 明确说明需要“avoid accessing mlp on PPMissingLayer placeholders from other pipeline stages”。

实现拆解

1. 修改 _routed_experts_weights_of_layer Lambda 表达式（文件：python/sglang/srt/models/qwen3_5.py）：
   - 将 for layer_id, layer in enumerate(self.model.layers) 替换为 for layer_id in range(self.model.start_layer, self.model.end_layer)，限制遍历范围仅为当前 PP rank 负责的层。
   - 将 layer.mlp.get_moe_weights() 替换为 self.model.layers[layer_id].mlp.get_moe_weights()，直接通过 layer_id 索引。
2. 无其他文件改动：仅修改了 1 个文件中的 3 行代码（+3/-3），由于该逻辑与 Qwen3.5 前向路径中已使用的局部层范围一致，因此无需额外测试。

关键文件：

- python/sglang/srt/models/qwen3_5.py（模块 模型层；类别 source；类型 data-contract）：唯一变更文件，修复了 MoE 权重收集的 PP 层范围问题。

关键符号：load_fused_expert_weights

关键源码片段

python/sglang/srt/models/qwen3_5.py

唯一变更文件，修复了 MoE 权重收集的 PP 层范围问题。

```
# python/sglang/srt/models/qwen3_5.py (head)

# 之前：遍历所有层，可能访问其他 PP rank 的 PPMissingLayer
# 现在：只遍历当前 rank 的局部层 (start_layer 到 end_layer)
self._routed_experts_weights_of_layer = LazyValue(
    lambda: {
        layer_id: self.model.layers[layer_id].mlp.get_moe_weights()
        for layer_id in range(self.model.start_layer, self.model.end_layer)
        if isinstance(self.model.layers[layer_id].mlp, Qwen2MoeSparseMoeBlock)
    }
)
```

评论区精华

无 review 讨论：该 PR 仅有一位 reviewer（AgainstEntropy）直接批准，无评论。作者在 issue 评论中解释了 CI 失败均与 PR 无关。

- 暂无高价值评论线程

风险与影响

- 风险：风险极低：
  - 改动范围极小（3 行），且已经过 py_compile 语法检查和 CI 测试（失败均为环境问题）。
  - 限制层范围到 start_layer 和 end_layer 是 PP 场景下的标准做法，与其他 PP 模块一致。
  - 不会影响单 rank 或 TP-only 场景，因为单 rank 时 start_layer=0、end_layer=num_hidden_layers。
- 影响：影响范围有限：
  - 仅修复 Qwen3.5 模型在 PP 部署（如 PP2+TP4）时加载权重的正确性。
  - 对单节点或仅 TP 的用户无影响。
  - 显著提升大规模部署场景下 Qwen3.5 的稳定性，避免因跨 PP 访问 PPMissingLayer 导致的 crash。
  - 风险标记：暂无

关联脉络

- 暂无明显关联 PR