PR #21348 完整报告

sgl-project/sglang

Fix MxInt4 MoE returning wrong output variable

合并时间：2026-03-26 10:57

原文链接：http://prhub.com.cn/sgl-project/sglang/pull/21348

PR 分析报告

执行摘要

修复 MxInt4 MoE 层中 trtllm_mxint4_block_scale_moe 函数返回错误输出变量的 bug，通过修正变量赋值确保输出缓冲区一致性，避免 CombinedOutput 预期不符，提升量化模型推理正确性。

功能与动机

该修复旨在解决 MxInt4 MoE 压缩方案中的输出变量错误。根据 PR body 描述，trtllm_mxint4_block_scale_moe 函数通过 output= 参数将结果写入 symm_output 缓冲区，但其返回值是一个列表，导致 CombinedOutput 无法正确处理，从而引发输出不一致问题。

实现拆解

在 apply_weights 函数中，修改仅涉及两行代码：

- 移除 output = trtllm_mxint4_block_scale_moe(...) 中的赋值，改为直接调用函数并传递 symm_output 缓冲区。
- 将返回语句从 return StandardCombineInput(hidden_states=output) 改为 return StandardCombineInput(hidden_states=symm_output)，确保返回正确的输出缓冲区。

评论区精华

Review 中无具体技术讨论，reviewer ispobock 直接批准，表明修复简单且风险低，无需深入辩论或权衡。

风险与影响

风险较低：变更仅修正变量赋值，不改变底层计算，但需依赖 symm_output 缓冲区的正确性。影响限于使用MxInt4MoE的量化模块，修复后提升模型输出准确性，对系统性能无显著影响。

关联脉络

在提供的近期 PR 历史中，未发现直接修改相同文件或涉及 MxInt4 MoE 量化方案的 PR，表明此修复为独立变更，可能是针对特定 bug 的快速修正。