PR #21343 完整报告

sgl-project/sglang

[Fix] Fix trtllm fp4 moe kernel not found error

合并时间：2026-03-25 07:38

原文链接：http://prhub.com.cn/sgl-project/sglang/pull/21343

执行摘要

修复 FP4 量化混合专家（MoE）内核的导入错误，通过调整代码逻辑确保 CI 测试通过，避免因依赖问题导致的测试失败。

功能与动机

此 PR 旨在解决在 PR 21330 的 CI 测试中发现的 trtllm fp4 moe kernel not found 错误。错误日志链接显示测试失败，需修复以支持 FP4 量化功能的正常运行，确保内核能正确导入。

实现拆解

变更集中于文件 python/sglang/srt/layers/moe/fused_moe_triton/layer.py：

- 删除冗余导入检查：移除全局导入尝试，包括 trtllm_fp4_block_scale_moe 的导入条件逻辑（原第 79-85 行）。
- 动态导入：将 import trtllm_fp4_block_scale_moe 移动到 forward_impl 函数内部（第 1321 行），改为运行时导入，并添加 assert 语句验证量化方法。

评论区精华

无 review 讨论。变更由作者直接合并，表明问题简单且无争议。

风险与影响

- 风险：导入时机变化可能导致每次调用 forward_impl 时重复导入，带来轻微性能开销；若导入失败，assert 可能引发运行时错误，但场景罕见。
- 影响：修复 CI 测试失败，对用户无感知影响，但确保系统稳定性和 FP4 量化支持的完整性。

关联脉络

与 PR 21330 紧密相关，因其 CI 测试错误触发此修复。同时，作为 FP4 量化支持的一部分，可能与历史 PR 如 20137（支持 NVFP4 量化）有间接关联，但当前 PR 主要聚焦于 CI 错误修复。