PR #20232 完整报告

sgl-project/sglang

[fix] qwen3.5 fuse_moe_triton_tune bug

合并时间：2026-03-28 07:23

原文链接：http://prhub.com.cn/sgl-project/sglang/pull/20232

执行摘要

此 PR 修复了 Qwen3.5 模型在 fused MoE Triton tuning benchmark 中的架构解析错误，通过调整配置处理逻辑确保兼容性，影响范围仅限于该 benchmark 的使用，属于常规维护性修复。

功能与动机

在运行 benchmark/kernels/fused_moe_triton/tuning_fused_moe_triton.py 时，Qwen3.5 模型因 config.architectures[0] 访问错误而崩溃。PR body 中提供了具体错误堆栈，指出 architectures 不是通过 text_configs 定义，导致 IndexError。修复动机是确保 benchmark 能正常支持 Qwen3.5 模型。

实现拆解

修改了 benchmark/kernels/fused_moe_triton/common_utils.py 文件中的 get_model_config 函数。关键变更如下：

- 将 architecture = config.architectures[0] 代码行从函数末尾移至检查 text_config 之前。
- 调整条件逻辑顺序，确保先获取 block_shape 再处理 text_config。 这避免了在 Qwen3.5 等模型中因 text_config 不存在而导致的配置解析失败。

评论区精华

Review 中没有实质性讨论，仅 b8zhong 批准了 PR，无争议或深入交流点。

风险与影响

- 风险：修改逻辑顺序可能引入回归，影响其他模型类型（如 encoder-decoder 模型）的配置解析。建议通过测试验证兼容性。
- 影响：修复了特定 benchmark 错误，提升了对 Qwen3.5 模型的支持，不影响系统其他部分，属于局部维护性修复。

关联脉络

此 PR 与历史 PR #19059（标题："[jit_kernel] Add fused_qknorm_rope JIT kernel"）相关，后者修改了 python/sglang/srt/models/qwen3_moe.py，涉及 Qwen MoE 模型的优化。这

表明项目在持续改进 MoE 相关功能，本 PR 是其中针对 benchmark 的小幅修复。