PR #27377 完整报告

sgl-project/sglang

fix: add missing guard for use_jit_ep_activation

合并时间：2026-06-19 03:51

原文链接：http://prhub.com.cn/sgl-project/sglang/pull/27377

执行摘要

- 一句话：修复 JIT EP 激活缺失的尺寸校验
- 推荐动作：建议尽快合并以修复 Qwen 3.5 35B 在 DEP2 模式下的 CUDA 崩溃问题。后续需跟进 swiglu_limit 相关冲突，避免从 CUDA crash 转为 Python assertion。可考虑添加针对不同专家数、分组尺寸组合的回归测试。

功能与动机

修复 Qwen 3.5 35B 在 DEP2（Expert Parallelism=2）模式下的崩溃。该模型配置为 num_experts=256、moe_intermediate_size=512，导致 D // 8 = 64、E = 128，满足 64 < 128 条件，触发 CUDA 内核断言失败。

实现拆解

在 python/sglang/srt/layers/moe/moe_runner/deep_gemm.py 的 _varlen_deep_gemm_silu_mul_quant 函数中，将 use_jit_ep_activation 的禁用条件从 N % 4 != 0 or G % 4 != 0 扩展为 N % 4 != 0 or G % 4 != 0 or D // 8 < E。当每个专家的输出维度 D 除以 8 小于本地专家数 E 时，JIT EP 激活路径无法满足底层 CUDA 核的要求，因此回退到标准 FP8 量化流程。

关键文件：

- python/sglang/srt/layers/moe/moe_runner/deep_gemm.py（模块 MoE 内核；类别 source；类型 core-logic；符号 _varlen_deep_gemm_silu_mul_quant）：MoE 量化核心函数所在文件，修改了 JIT EP 激活的守卫条件。

关键符号：_varlen_deep_gemm_silu_mul_quant

关键源码片段

python/sglang/srt/layers/moe/moe_runner/deep_gemm.py

MoE 量化核心函数所在文件，修改了 JIT EP 激活的守卫条件。

```
# python/sglang/srt/layers/moe/moe_runner/deep_gemm.py
# _varlen_deep_gemm_silu_mul_quant 函数中的守卫条件
use_jit_ep_activation = envs.SGLANG_OPT_USE_JIT_EP_ACTIVATION.get()
# 当 N（token 数）或 G（分组数）不对齐 4，或 D // 8（每组输出块数）小于 E（本地专家数）时
# 禁用 JIT EP 激活，避免底层 CUDA 核断言（如 Qwen 3.5 35B D=512, E=128）
```

```
if N % 4 != 0 or G % 4 != 0 or D // 8 < E:
    use_jit_ep_activation = False
```

评论区精华

gemini-code-assist[bot] 指出禁用 use_jit_ep_activation 后，swiglu_limit 不为 None（如 Qwen 3.5 35B）会在 else 分支触发新的 Python assertion。建议手动应用 _apply_swiglu_limit 并重置 swiglu_limit 为 None。author lawrence-harmonic 回复 "This is unrelated"，认为该问题与当前 PR 无关，可能属于已有或待处理的独立问题。该讨论被关闭，未进一步处理。

- 禁用 JIT EP 后可能触发 Python assertion (correctness): author lawrence-harmonic 回复 "This is unrelated"，认为该问题不属于本 PR 范围，讨论未解决。

风险与影响

- 风险：当前修改仅解决 CUDA assertion 崩溃，但当 use_jit_ep_activation 被禁用且 swiglu_limit 非 None 时，会触发 Python 层 assertion（如 assert swiglu_limit is None）。这可能导致 Qwen 3.5 35B 等模型从 CUDA crash 转为 Python crash。此外，此次修改未添加单元测试覆盖 D // 8 < E 条件。
- 影响：直接影响使用 deep_gemm 内核且专家数量较大的 MoE 模型（如 Qwen 3.5 35B）在 DEP2 模式下的运行。回退路径不会影响模型推理结果，但可能因缺少 swiglu_limit 处理而引入 Python 断言失败。影响范围仅限于启用了 SGLANG_OPT_USE_JIT_EP_ACTIVATION 环境变量的用户。
- 风险标记：潜在 Python assertion, 缺少测试覆盖，核心路径变更

关联脉络

- PR #28649 Pass quant_config to attention gate projection: 同为 Laguna 系列模型修复，属于 MoE 量化配置传递问题。
- PR #28559 fix: speculative draft worker clobbering target attention backend: 同为修复 JIT 内核与 MoE 交互的 bug。