PR #22312 完整报告

sgl-project/sglang

Make GDN support non-continuous B/A Tensor input to fix the accuracy regression of Qwen3.5-27B

合并时间：2026-04-10 18:58

原文链接：http://prhub.com.cn/sgl-project/sglang/pull/22312

执行摘要

本 PR 修复了 Qwen3.5-27B 模型因 GDN 内核不支持非连续输入而导致的准确性回归问题，通过更新内核步幅处理恢复准确性至正常水平，直接影响模型用户并提升系统可靠性。

功能与动机

动机源于 commit 5bdc07d 引入的优化，该优化导致 Qwen3.5-27B 在 fallback 路径下产生非连续 BA 张量视图。GND Triton 内核假设连续布局，硬编码步幅，引发内存读取错误和准确性严重下降（从 49/50 降至 3/50）。PR body 明确引用 Issue #22311，指出问题根源在于内核未处理 split 操作产生的非连续视图，修复目标是使内核支持非连续输入。

实现拆解

改动按模块拆解如下：

- 核心内核更新：
  - fused_gdn_gating.py：在内核函数中添加 stride_a 和 stride_b 参数，替换硬编码偏移为 a + i_b * stride_a + head_off 和 b + i_b * stride_b + head_off 的加载逻辑。
  - fused_sigmoid_gating_recurrent.py：类似添加 stride_a 参数，并在循环更新中使用 p_a = a + bos * stride_a + i_hv * K + o_k（KDA 路径）或 p_a = a + bos * stride_a + i_hv（GDN 路径）确保正确指针移动。
- 测试覆盖：新增 test_gdn_noncontiguous_stride.py，通过 _make_noncontiguous_ab 函数模拟 Qwen3.5 split 操作，生成非连续张量并对比内核输出与连续版本的差异，验证修复正确性。

评论区精华

Review 过程中无实质性讨论，reviewer 'yizhang2077' 直接批准，表明修改清晰且必要，所有技术细节已在 PR body 和 commit 中阐述。

风险与影响

- 风险：原先内存读取错误已修复，但需确保步幅计算不影响其他模型路径；新增测试覆盖了 Qwen3.5 形状，但未全面验证所有潜在布局变体。

- 影响：直接恢复 Qwen3.5-27B 准确性，提升用户信任；间接增强 GDN 模块对非连续输入的鲁棒性，可能惠及类似模型配置。

关联脉络

与历史 PR #22444 相关，后者也涉及 GDN 模块的性能优化（修改 gdn_backend.py），共同体现了对 GDN 内核的持续改进趋势。本 PR 作为准确性修复，补充了性能优化后的正确性保障。