

PR #31614 完整报告

sgl-project/sglang

[spec decoding] fix multi_layer_eagle rotate_input_ids kernel registration

合并时间: 2026-07-18 06:18

原文链接: <http://prhub.com.cn/sgl-project/sglang/pull/31614>

执行摘要

- 一句话: 修复 rotate_input_ids kernel 注册名
- 推荐动作: 建议精读。该 PR 虽小但涉及 kernel 注册的正确性, 值得快速合并。

功能与动机

PR#27862 中错误地将注册表中的函数名从 rotate_input_ids 改为了 rotate_input_ids_triton, 导致 multi_layer_eagle rotate_input_ids kernel 注册失败。

实现拆解

1. 修改注册表项: 在 python/sglang/kernels/ops/speculative/__init__.py 中将 rotate_input_ids_triton 改回 rotate_input_ids。
2. 同步注释: 在 python/sglang/kernels/ops/speculative/multi_layer_eagle.py 中将注释中的 rotate_input_ids_triton 更新为 rotate_input_ids kernel, 保持一致性。

关键文件:

- python/sglang/kernels/ops/speculative/__init__.py (模块 Kernel 注册; 类别 infra; 类型 infrastructure): 修复 kernel 注册表项, 将错误的函数名改回正确名称。
- python/sglang/kernels/ops/speculative/multi_layer_eagle.py (模块 Kernel 代码; 类别 infra; 类型 infrastructure): 同步更新注释中的函数名, 保持代码可读性。

关键符号: rotate_input_ids

评论区精华

无 review 讨论。

- 暂无高价值评论线程

风险与影响

- 风险: 风险极低, 仅涉及函数名称字符串和注释的修正, 不会引入新 bug。
- 影响: 修复了 multi_layer_eagle 中 rotate_input_ids kernel 的注册问题, 确保该 kernel 能被正确调用; 影响范围仅限于 speculative decoding 中的 multi_layer_eagle 功能。
- 风险标记: 暂无

关联脉络

- PR #27862：该 PR 中错误地修改了注册表项，导致本 PR 需要修复。