

PR #27473 完整报告

sgl-project/sglang

Fix TRTLLM target verify query metadata

合并时间: 2026-06-07 17:46

原文链接: <http://prhub.com.cn/sgl-project/sglang/pull/27473>

执行摘要

- 一句话: 移除 TRTLLM target-verify 中 query length 的强制覆盖
- 推荐动作: 建议合并, 变更简单且逻辑正确。

功能与动机

修复 target-verify CUDA graph 元数据中 query length 被错误覆盖的问题, 确保其与实际请求形状一致, 避免潜在的 CUDA graph 执行错误。

实现拆解

移除 `trtllm_mha_backend.py` 中 `_apply_cuda_graph_metadata` 方法里 `target_verify` 分支末尾的 `metadata.max_seq_len_q = self.speculative_num_draft_tokens` 赋值语句。该语句在完成所有元数据填充后强制覆盖了 query length, 而其他分支均保留从捕获形状派生的值。

关键文件:

- `python/sglang/srt/layers/attention/trtllm_mha_backend.py` (模块 注意力; 类别 source; 类型 core-logic; 符号 `_apply_cuda_graph_metadata`): 唯一变更文件, 移除 target-verify 分支中多余的 query length 覆盖操作。

关键符号: `_apply_cuda_graph_metadata`

评论区精华

无 review 评论。

- 暂无高价值评论线程

风险与影响

- 风险: 风险极低, 涉及单行删除, 且变更已在 body 中通过 `py_compile` 验证。若上游代码有其他处依赖该覆盖值, 需确认, 但当前代码逻辑表明该赋值多余。
- 影响: 影响范围小, 仅影响 Blackwell 平台下 TRTLLM 后端的 target-verify CUDA graph 元数据。修复后 query length 保持一致, 避免潜在图执行错误。
- 风险标记: 极低风险

关联脉络

- PR #27494 Revert "Fix TRTLLM target verify query metadata (#27473)": 27494 是回滚本 PR 的 revert, CI 状态与本 PR 直接相关。