PR #27495 完整报告

sgl-project/sglang

Fix TRTLLM target verify query metadata

合并时间：2026-06-09 01:58

原文链接：http://prhub.com.cn/sgl-project/sglang/pull/27495

执行摘要

- 一句话：删除 TRTLLM target verify 中错误的 max_seq_len_q 覆盖
- 推荐动作：该 PR 为单行 bugfix，逻辑清晰，建议合并。对于好奇原 bug 细节的开发者，可参考 #27473 和 #27494 了解回滚原因。

功能与动机

在 target_verify 路径中，metadata.max_seq_len_q 被错误地重新赋值为 self.speculative_num_draft_tokens，而该变量已在之前从捕获请求形状正确设置。这种覆盖可能导致与实际请求不匹配，从而引发 bug。PR #27473 曾尝试修复但被回滚，本次为重新提交。

实现拆解

1. 问题定位：在 python/sglang/srt/layers/attention/trtllm_mha_backend.py 的 _apply_cuda_graph_metadata 方法中，target_verify 分支末尾有一行 metadata.max_seq_len_q = self.speculative_num_draft_tokens，该行会覆盖此前基于实际请求设置的 max_seq_len_q。
2. 修复：删除该行赋值语句，保留从捕获请求形状派生的正确值。
3. 测试配套：未添加新测试，但 PR body 提到 CI 已通过 lint.yml 和 pr-test.yml。

关键文件：

- python/sglang/srt/layers/attention/trtllm_mha_backend.py（模块 注意力层；类别 source；类型 core-logic）：唯一修改的文件，在 target_verify 分支中删除一行错误的 max_seq_len_q 赋值。

关键符号：未识别

关键源码片段

python/sglang/srt/layers/attention/trtllm_mha_backend.py

唯一修改的文件，在 target_verify 分支中删除一行错误的 max_seq_len_q 赋值。

```
# python/sglang/srt/layers/attention/trtllm_mha_backend.py
# 在 _apply_cuda_graph_metadata 方法的 target_verify 分支中
# 删除前（第 487 行）： metadata.max_seq_len_q = self.speculative_num_draft_tokens
```

```
# 该赋值将之前从捕获请求形状正确设置的 max_seq_len_q 覆盖为固定 draft token 数，
# 导致 CUDA Graph 元数据与实际请求不匹配。
elif forward_mode.is_target_verify():
    # 仅支持 topk=1
    metadata = self.target_verify_metadata[bs]
    metadata.cache_seqlens_int32.copy_(seq_lens + metadata.max_seq_len_q)
    metadata.max_seq_len_k = seq_lens_cpu.max().item() + metadata.max_seq_len_q
    max_len = seq_lens_cpu.max().item()
    metadata.cu_seqlens_k[1:].copy_(
        torch.cumsum(metadata.cache_seqlens_int32, dim=0, dtype=torch.int32)
    )
    max_seq_pages = (metadata.max_seq_len_k + self.page_size - 1) // self.page_size
    page_indices = self.req_to_token[
        req_pool_indices[:, None],
        self.decode_cuda_graph_metadata["strided_indices"][:max_seq_pages],
    ]
    metadata.page_table[:, :max_seq_pages].copy_(page_indices // self.page_size)
    self._copy_swa_page_table(metadata, page_indices, max_seq_pages)
    # 已删除：metadata.max_seq_len_q = self.speculative_num_draft_tokens
```

评论区精华

无人工 review 讨论。Gemini Code Assist 机器人自动评论表示无反馈。PR 描述为 #27473 的重新提交（原 PR 被回滚），但回滚原因未在此处说明。

- 暂无高价值评论线程

风险与影响

- 风险：
  1. 回归风险：低。删除的是一行刚在之前版本中引入的赋值，且 CI 已通过。但如果没有覆盖，后续逻辑若依赖该值为 speculative_num_draft_tokens 可能出错。从代码看，该值在 target_verify 分支中没有后续使用，默认行为仍稳定。
  2. 性能影响：无。仅删除一行赋值，不影响计算路径。
  3. 兼容性：该变更仅影响 TRTLLM 后端在 Blackwell 等 GPU 上的 target_verify CUDA Graph 路径，其他后端不受影响。
- 影响：
  1. 用户影响：修复了 speculative decoding（EAGLE）中 target_verify 步骤可能出现的错误行为，提升生成正确性。
  2. 系统影响：仅修改单行代码，无部署或配置影响。
  3. 团队影响：低，变更范围极小，易于审查。 - 风险标记：核心路径变更

关联脉络

- PR #27473 Previous attempt (reverted): 该 PR 是 #27473 的重新提交，原 PR 被回滚。
- PR #27494 Revert of #27473: 回滚了之前相同的修复，本次为重新应用。

- PR #27491 Fix SWA pool resolution for EAGLE draft workers: 同样是 speculative decoding + TRTLLM 后端的修复，同一功能领域。