# PR #44603 完整报告

vllm-project/vllm

fix: pad dummy run query_start_loc

合并时间：2026-06-05 15:43

原文链接：http://prhub.com.cn/vllm-project/vllm/pull/44603

## 执行摘要

- 一句话：修复 dummy run 中 query_start_loc 填充不足
- 推荐动作：值得快速合入，但建议补充单元测试覆盖 dummy run 的 query_start_loc 填充行为，防止回归。

## 功能与动机

decoder 实例在 GLM-5.1-FP8 的 disaggregation 模式下崩溃，CUDA coredump 显示 repeat 操作中 repeat >= 0 断言失败。通过日志定位到 repeat_interleave 调用中 repeats 出现负数，原因是 dummy run 中 query_start_loc 不是单调递增序列（实际数据为 [0, 3]，但 padding 区域未被填充，导致下游计算异常）。

## 实现拆解

1. 在 vllm/v1/worker/gpu_model_runner.py 的 _dummy_run 方法中，填充完 query_start_loc 的有效部分后，对 padding 区域（索引 num_reqs + 1 到 num_reqs_padded）使用 cum_num_tokens[-1] 进行填充。
2. 这样确保整个 query_start_loc 数组单调非递减，使得下游 repeat_interleave 等操作不会因负值而崩溃。
3. 仅 3 行新增代码，无其他文件修改。

关键文件：

- vllm/v1/worker/gpu_model_runner.py（模块 模型运行器；类别 source；类型 data-contract）：修复的单一文件，在 _dummy_run 方法中新增 padding 填充逻辑。

关键符号：未识别

## 关键源码片段

### vllm/v1/worker/gpu_model_runner.py

修复的单一文件，在 _dummy_run 方法中新增 padding 填充逻辑。

```
# vllm/v1/worker/gpu_model_runner.py ( 简化片段 )
# 在 _dummy_run 方法中，填充完有效区间后对 padding 区域进行补全
cum_num_tokens = self._get_cumsum_and_arange(
    num_scheduled_tokens, self.query_pos.np
```

```
)
# 设置有效请求的 query_start_loc（0 到 num_reqs）
self.query_start_loc.np[1 : num_reqs + 1] = cum_num_tokens
# 关键修复：padding 区域（num_reqs+1 到 num_reqs_padded）
# 必须用最后一个有效值填充，保证单调非递减
self.query_start_loc.np[num_reqs + 1 : num_reqs_padded + 1].fill(
    cum_num_tokens[-1]
)
self.query_start_loc.copy_to_gpu()
```

## 评论区精华

Review 较少：Claude bot 无法自动审查 fork PR；WoosukKwon 快速批准（LGTM）。无深入技术讨论。

- 暂无高价值评论线程

## 风险与影响

- 风险：风险较低。变更仅影响 _dummy_run 中的 query_start_loc padding 逻辑，属于一处遗漏的边界填充。不影响 execute_model 的正常路径，不改变模型推理核心逻辑。但缺少测试覆盖，需要验证其他场景（如混合 batch、profile seq lens）是否同样受影响。
- 影响：直接影响使用 dummy run 的 CUDA graph 或 disaggregation 模式的解码器，修复 GLM-5.1-FP8 的崩溃。对不涉及 dummy run 的场景无影响。团队收益：避免 decoder 在特定模型和部署方式下崩溃，提高稳定性。
- 风险标记：缺少测试覆盖

## 关联脉络

- PR #43720 [KVConnector][1/N] PP-aware handshake aggregation and intermediate-PP output plumbing: 同样涉及 vllm/v1/worker/gpu_worker.py 等 GPU worker 相关文件，且与 disaggregation 模式相关。
- PR #44569 [DSV4] Refactor DeepseekV4Attention: DeepSeek V4 相关的注意力重构，与 GLM-5.1-FP8 问题可能同属 attention 路径。