PR #23885 完整报告

sgl-project/sglang

[Disagg] Finalize routed_experts_output in process_batch_result_disagg_prefill

合并时间：2026-04-28 07:40

原文链接：http://prhub.com.cn/sgl-project/sglang/pull/23885

执行摘要

- 一句话：修复 PD Disagg 预填充未 finalize 路由专家输出
- 推荐动作：建议合并。该 PR 修复了一个明确的遗漏 bug，修改量小，逻辑清晰，且已本地验证通过。

功能与动机

修复 PD Disagg + --enable-return-routed-experts + overlap scheduling 下路由专家返回全零的 bug。PR body 指出，在启用路由回放的 RL 工作流中，预填充 prompt token 返回全零 topk 行，被训练器解释为 8 次重复选择专家 0，触发断言 Duplicate experts in routing!。

实现拆解

在 python/sglang/srt/disaggregation/prefill.py 的 process_batch_result_disagg_prefill 方法中，在 copy_done.synchronize() 之后，添加了如下代码：

1. 检查 result.routed_experts_output 是否为 None；
2. 若非 None，调用 result.routed_experts_output.finalize() 将 CPU 侧的 tensor 写入 host_cache.buffer；
3. 将 result.routed_experts_output 设为 None，与聚合模式的处理一致。 该修改仅 3 行，逻辑简单但关键。

关键文件：

- python/sglang/srt/disaggregation/prefill.py（模块 调度器；类别 source；类型 core-logic）：修复 PD Disagg 预填充处理器中遗漏的 finalize() 调用，核心 bugfix 所在。

关键符号：未识别

关键源码片段

python/sglang/srt/disaggregation/prefill.py

修复 PD Disagg 预填充处理器中遗漏的 finalize() 调用，核心 bugfix 所在。

```
# python/sglang/srt/disaggregation/prefill.py
# 在 copy_done 同步后，finalize 路由专家输出并清空引用
if copy_done is not None:
    copy_done.synchronize()
```

```
# 新增：确保 host_cache.buffer 被正确写入
if result.routed_experts_output is not None:
    result.routed_experts_output.finalize()
    result.routed_experts_output = None
```

评论区精华

无讨论。

- 暂无高价值评论线程

风险与影响

- 风险：风险极低。修改仅在 result.routed_experts_output is not None 时执行，当 --disable-overlap-schedule 时该字段为 None，因此无影响。但需确保所有调用 process_batch_result_disagg_prefill 的路径都已覆盖。
- 影响：影响范围：修复 PD Disagg + 返回路由专家 + overlap scheduling 下的严重回归，使相关 RL 工作流恢复正常。对不使用该功能的用户无影响。
- 风险标记：核心路径变更

关联脉络

- PR #22911 [perf] support return_routed_experts with overlap scheduling: 本 PR 修复了 PR #22911 引入的遗漏，后者添加了延迟 D2H 路径但未更新 PD Disagg 处理器。
- PR #22916 [perf] mini-lb merge routed experts: PR body 中提及 mini-lb 路由专家合并功能，是验证场景的一部分。