PR #23179 完整报告

sgl-project/sglang

[LoRA] add lora chunked req test and fix

合并时间：2026-06-02 07:25

原文链接：http://prhub.com.cn/sgl-project/sglang/pull/23179

执行摘要

- 一句话：修复 LoRA 分块请求槽位遗漏
- 推荐动作：值得阅读以了解 LoRA 调度中的分块请求处理陷阱。虽然代码改动极小，但反映了状态同步容易遗漏的典型场景。

功能与动机

当前 LoRA 准入代码未考虑分块请求，导致分块请求的 LoRA 槽位未被正确标记为已占用，进而可能引发同一 LoRA 被重复调度。

实现拆解

1. 定位问题：在 scheduler.py 的 _get_new_batch_prefill_raw 方法中，LoRA 门控代码只从 self.running_batch.reqs 收集 LoRA ID，忽略了已通过 add_chunked_req 加入 adder.can_run_list 的分块请求。
2. 核心修复：在 running_loras 集合初始化后，增加一行 running_loras.update(req.lora_id for req in adder.can_run_list)，确保分块请求占用的 LoRA 也被计入已占用槽位。
3. 涉及文件：仅修改了 python/sglang/srt/managers/scheduler.py，新增 2 行代码（一行注释加一行逻辑）。
4. 测试调整：最初添加的专用测试文件在 review 中被认为不必要，最终被删除。

关键文件：

- python/sglang/srt/managers/scheduler.py（模块 调度器；类别 source；类型 core-logic；符号 _get_new_batch_prefill_raw）：核心修复文件，在 LoRA 门控逻辑中增加了对 adder.can_run_list 中请求的 LoRA ID 的计数。

关键符号：_get_new_batch_prefill_raw

关键源码片段

python/sglang/srt/managers/scheduler.py

核心修复文件，在 LoRA 门控逻辑中增加了对 adder.can_run_list 中请求的 LoRA ID 的计数。

```
# In _get_new_batch_prefill_raw method of Scheduler class
if self.enable_lora:
```

```
    # 从当前运行批次收集 LoRA ID
    running_loras = {req.lora_id for req in self.running_batch.reqs}
    # 关键修复：加入已在 adder 中的请求（如 chunked requests）的 LoRA ID
    running_loras.update(req.lora_id for req in adder.can_run_list)

    if self.lora_drainer:
        self.lora_drainer.update_draining_state(
            self.waiting_queue,
            self.running_batch.reqs,
        )

# 后续循环使用 running_loras 检查新请求是否可调度
for req in self.waiting_queue:
    if self.enable_lora and not self._can_schedule_lora_req(req, running_loras):
        continue
    # ...
```

评论区精华

Reviewer Fridge003 对新增的测试文件评论 "No need to add this test"，随后作者删除了该测试文件。其他评论均为 CI 重跑命令，未涉及代码逻辑讨论。

- 是否需要专用的 LoRA 分块请求测试 (testing): 作者删除了测试文件，只保留核心源代码修改。

风险与影响

- 风险：风险极低。修改仅追加一个集合更新操作，本质是修正计数逻辑，不影响其他代码路径。需确保 adder.can_run_list 中请求的 lora_id 有效，且 set.update 天然去重，不会导致重复计数。
- 影响：影响范围仅限于使用 LoRA 且启用 chunked prefill 的场景。修复后这些场景下 LoRA 槽位计算正确，避免因计数遗漏导致的调度错误。无需用户配置变更。
- 风险标记：缺少直接测试覆盖

关联脉络

- 暂无明显关联 PR