PR #25439 完整报告

sgl-project/sglang

Lift running_batch / running_mbs access — direct + PP-explicit

合并时间：2026-05-16 09:21

原文链接：http://prhub.com.cn/sgl-project/sglang/pull/25439

执行摘要

- 一句话：简化 running_batch 和 running_mbs 的条件守卫
- 推荐动作：建议快速合并。该 PR 是机械重构的一个安全示例，适合作为代码阅读的参考，但不需要精读。

功能与动机

移除冗余的 hasattr 守卫，因为 running_batch 始终在 Scheduler.__init__ 中创建，而 running_mbs 在 PP 模式下通过 init_pp_loop_state 保证在读取前已初始化。hasattr 的存在不仅多余，而且隐式地充当了“是否处于PP模式”的代理，不如直接检查pp_size>1语义清晰。

实现拆解

1. 在 python/sglang/srt/observability/scheduler_metrics_mixin.py 的 update_lora_metrics 方法中，将 PP 模式分支的条件从 if hasattr(self, "running_mbs") and self.running_mbs: 改为 if self.server_args.pp_size > 1:。
2. 将普通模式分支的条件从 elif hasattr(self, "running_batch") and self.running_batch: 改为 elif self.running_batch:。
3. 未引入其他变动，变更仅涉及 2 行删除和 2 行新增。

关键文件：

- python/sglang/srt/observability/scheduler_metrics_mixin.py（模块 观测；类别 source；类型 core-logic；符号 update_lora_metrics）：唯一变更文件，修改了 LoRA 指标采集方法中的条件守卫。

关键符号：update_lora_metrics

关键源码片段

python/sglang/srt/observability/scheduler_metrics_mixin.py

唯一变更文件，修改了 LoRA 指标采集方法中的条件守卫。

```
  def update_lora_metrics(self: Scheduler):
      """Update LoRA pool metrics for monitoring and autoscaling."""
      if not self.enable_lora:
          return
```

```
        try:
            lora_manager = self.tp_worker.model_runner.lora_manager
            if lora_manager is None or lora_manager.memory_pool is None:
                return

            mem_pool = lora_manager.memory_pool
            slots_total = mem_pool.max_loras_per_batch

            active_lora_ids = set()

            # For PP mode, check all running micro batches
            # running_mbs is guaranteed to be initialized by init_pp_loop_state
            # in PP mode before this is ever called.
            if self.server_args.pp_size > 1:
                for batch in self.running_mbs:
                    if batch and hasattr(batch, "reqs"):
                        for req in batch.reqs:
                            if hasattr(req, "lora_id") and req.lora_id is not None:
                                active_lora_ids.add(req.lora_id)
            # For normal mode, check running_batch
            # running_batch is always created in Scheduler.__init__
            elif self.running_batch:
                if hasattr(self.running_batch, "reqs"):
                    for req in self.running_batch.reqs:
                        if hasattr(req, "lora_id") and req.lora_id is not None:
                            active_lora_ids.add(req.lora_id)

            slots_used = len(active_lora_ids)
            utilization = slots_used / slots_total if slots_total > 0 else 0.0

            self.stats.lora_pool_slots_used = slots_used
            self.stats.lora_pool_slots_total = slots_total
            self.stats.lora_pool_utilization = utilization

        except Exception as e:
            logger.warning(f"Failed to update LoRA metrics: {e}")
```

评论区精华

该 PR 无 reviewer 评论，且仅由作者自己合并，表明变更为公认的小型机械重构，无需深入讨论。

- 暂无高价值评论线程

风险与影响

- 风险：风险较低：running_batch 和 running_mbs 的初始化时机在代码结构中已有保证，但若未来有人改变初始化逻辑（例如延迟初始化 running_batch），则直接访问可能导致 AttributeError。当前 hasattr 的移除依赖于 Scheduler.__init__ 和 init_pp_loop_state 的

现有实现，这些实现被视作不变前提。

- 影响：影响范围限于 update_lora_metrics 一处逻辑，无外部可见行为变化。代码语义更清晰，维护性略微提升。
- 风险标记：依赖初始化时机

关联脉络

- 暂无明显关联 PR