PR #21705 完整报告

sgl-project/sglang

Fix in-place mode in pause generation

合并时间：2026-04-01 16:36

原文链接：http://prhub.com.cn/sgl-project/sglang/pull/21705

执行摘要

- 一句话：修复调度器中 in-place 暂停模式的内存泄漏问题。
- 推荐动作：建议快速浏览以理解调度器状态管理的设计决策：通过冻结状态而非重复逻辑来避免 bug。这是一个简洁的 bugfix，实现简单但设计值得关注，适合工程师学习状态一致性处理。

功能与动机

PR body 中指出：在 RL 权重同步（如 Miles/Slime）过程中，in-place 模式的 pause_generation 触发内存泄漏消息：'token_to_kv_pool_allocator memory leak detected!'，而其他模式如 abort/retract 正常。目标是通过冻结状态来避免内存泄漏，保持引擎状态不变，由后续事件循环处理。

实现拆解

主要改动在 scheduler.py 的 pause_generation 方法：当 mode 为 'in_place' 时，立即设置 _engine_paused 标志并返回，跳过后续的状态清理逻辑（如处理重叠结果、过滤 batch），从而保持 last_batch、chunked_req 等状态不变。同时新增了 test_scheduler_pause_generation.py 单元测试文件，包含多个测试用例验证 in-place 模式仅设置标志且不修改状态，其他模式行为不变。

关键文件：

- python/sglang/srt/managers/scheduler.py（模块 调度器）：核心逻辑变更，修复 in-place 暂停模式的内存泄漏，修改 pause_generation 方法以冻结状态。
- test/registered/unit/managers/test_scheduler_pause_generation.py（模块 测试）：新增单元测试文件，验证 in-place 模式的行为和其他模式，确保修复正确性和测试覆盖。

关键符号：pause_generation

评论区精华

review 评论仅来自 gemini-code-assist[bot]，聚焦于测试代码优化：一是指出测试中 mock __len__ 方法不必要，因为 pause_generation 检查的是 self.running_batch.reqs 的长度；二是建议初始化 chunked_req 为非 None 值以强化 retract 模式的测试断言。讨论未涉及核心逻辑争议，旨在提升测试准确性和代码清晰度。

- 测试代码中 mock __len__方法的必要性 (testing): 评论旨在优化测试代码，避免不必要的 mock，提升代码清晰度。
- 测试断言的强化 (testing): 评论旨在提升测试的准确性和覆盖率，确保测试断言有效。

风险与影响

- 风险：风险较低：变更只影响 in-place 模式，其他模式逻辑不变。潜在风险包括 in-place 模式与调度器其他部分（如重叠处理、事件循环）的交互未充分测试，但新增的单元测试覆盖了主要场景，且 *e2e* RL 运行已验证修复。代码修改简单，回归风险小。
- 影响：对用户影响：解决了 RL 场景下的内存泄漏，提升系统稳定性和性能。对系统影响：调度器状态管理更一致，避免了不必要的状态修改和潜在账户错误。对团队影响：代码更简洁，减少了重复逻辑，便于维护和后续开发。
- 风险标记：状态管理变更，测试覆盖良好

关联脉络

- 暂无明显关联 PR