

PR #30937 完整报告

sgl-project/sglang

fix: avoid double KV release on disaggregated prefill grammar errors

合并时间: 2026-07-14 20:16

原文链接: <http://prhub.com.cn/sgl-project/sglang/pull/30937>

执行摘要

- 一句话: 修复解耦预填充语法错误导致 KV 双释放
- 推荐动作: 值得精读。该 PR 展示了异步传输与错误处理交织时资源生命周期的细致考量, 是解耦预填充稳定性的重要修复。建议关注 to_finish 升格问题, 后续可能需跟进 PR 完善外部中止保护。

功能与动机

When grammar validation fails after a disaggregated prefill KV transfer has started, the request remains in the inflight queue until the transfer reaches a terminal state. Releasing its KV cache immediately can release the same allocation again during terminal cleanup. Terminal handling can also replace the original grammar error with a transfer result.

实现拆解

实现分为三步:

1. 延迟 KV 释放: 在 `python/sglang/srt/disaggregation/prefill.py` 的语法错误处理中, 移除立即调用 `release_kv_cache` (原第 696 行), 使缓存仅由传输完成回调释放, 避免双释放。
2. 传输成功路径保护: 在 `process_disagg_prefill_inflight_queue` 的 `KVPoll.Success` 分支中, 增加条件判断: 若 `req.finished_reason` 已是 `FINISH_ABORT`, 则不覆盖为 `FINISH_LENGTH`, 保留原始错误原因。
3. 传输失败路径保护: 在 `handle_inflight_transfer_failure` 方法中, 同样添加条件判断: 若 `req.finished_reason` 非 `FINISH_ABORT`, 才调用 `prepare_abort`, 防止内部中止覆盖外部中止原因。配套测试文件 `test_prefill_queue_cleanup.py` 在后续提交中被删除, 可能因测试环境不稳定。

关键文件:

- `python/sglang/srt/disaggregation/prefill.py` (模块 预填充; 类别 source; 类型 core-logic): 唯一变更文件, 核心修复包含延迟 KV 释放、成功 / 失败路径的中止原因保护。

关键符号: `SchedulerDisaggregationPrefillMixin.process_disagg_prefill_inflight_queue`, `SchedulerDisaggregationPrefillMixin.handle_inflight_transfer_failure`

评论区精华

Review 中 gemini-code-assist[bot] 提出两个高优先级问题：

- 在成功路径中，若请求已被外部中止（`req.to_finish = FINISH_ABORT`），直接覆盖为 `FINISH_LENGTH` 会丢失外部中止，建议升格 `to_finish` 到 `finished_reason`。
- 在失败路径中，同样推荐升格 `to_finish` 以保留原始中止原因。PR 作者仅添加了 `isinstance(finished_reason, FINISH_ABORT)` 检查，未升格 `to_finish`，因此外部中止仍可能被静默丢弃。该问题未在本次合并中解决。
- 传输成功路径中外外部中止保护 (correctness): PR 作者仅添加了 `FINISH_ABORT` 检查，未升格 `to_finish`，存在外部中止丢失风险。
- 传输失败路径中外外部中止保护 (correctness): PR 作者添加了 `FINISH_ABORT` 检查，但未升格 `to_finish`，可能丢失外部中止。

风险与影响

- 风险：
 1. 外部中止丢失：由于 `req.to_finish` 未升格为 `finished_reason`，若在升格前发生传输完成或失败，外部中止可能被覆盖，请求被错误视为成功，浪费下游 decode 资源。
 2. 回归风险低：修改仅限于解耦预填充错误路径，不影响正常推理流程和主执行路径。
 3. 测试覆盖缺失：原配套测试文件已被删除，缺少自动化回归覆盖，依赖手动验证。- 影响：影响范围仅限于解耦预填充场景下、语法校验失败且 KV 传输已开始的请求。此修正减少了因双释放导致的内存错误和潜在崩溃，提升了系统可靠性。对用户无感知，正常功能不变。对开发团队而言，该修复是资源生命周期管理的重要补充，但 `to_finish` 升格问题可能需后续跟进。- 风险标记：外部中止可能丢失，测试覆盖不足，依赖内部状态标记

关联脉络

- 暂无明显关联 PR