PR #21875 完整报告

sgl-project/sglang

fix: streaming session race condition + some metrics

合并时间：2026-04-13 09:05

原文链接：http://prhub.com.cn/sgl-project/sglang/pull/21875

PR 分析报告：修复流式会话竞态条件与内存泄漏

执行摘要

本 PR 通过修复流式会话中的竞态条件和多个内存泄漏问题，显著提升了系统的稳定性和可观测性。关键变更包括优化会话生命周期管理、添加推迟关闭机制和引入 Prometheus 指标，影响范围覆盖所有使用流式会话的用户，建议工程师精读以理解复杂的内存管理设计。

功能与动机

为什么做？流式会话在长期运行中常因内存泄漏和竞态条件导致系统不稳定。PR body 明确表示要“稳定流式会话的 KV 生命周期”，修复内存泄漏并增加会话关闭安全性。关联的 Issue #22273 进一步描述了 abort 请求导致的 KV 泄漏问题，需紧急处理以避免崩溃。

实现拆解

按模块拆解关键改动：

模块	关键文件	主要变更
调度器	scheduler.py	修改 handle_generate_request 和 open_session，添加会话关闭时的错误处理，确保 DP-attention 下响应不重复。
会话控制器	session_controller.py	引入 close_on_finish 标志，推迟会话关闭直至请求完成；添加日志记录。代码示例：
``python		
if has_unfinished_request:		
session.close_on_finish = True		

模块	关键文件	主要变更
logger.info("Deferring session close for %s (unfinished request)", session_id)		
return		
``		
内存缓存	session_aware_cache.py	修复 match_prefix 中的 abort-skip 逻辑，避免覆盖会话 KV；优化 cache_finished_req 处理 abort 和 session shrink。
通用缓存	common.py	在 release_kv_cache 中添加 overalloc tail trim，并同步 kv_committed_freed 等标志。
可观测性	metrics_collector.py	新增 num_streaming_sessions 和 streaming_session_held_tokens 指标，受 enable_streaming_session 标志控制。
测试	test_streaming_session_unit.py	添加 7 个单元测试，验证修复点如 overalloc trim 和 release 逻辑。

评论区精华

由于 review 评论为空，讨论亮点主要从 commit 历史中提炼：

- 变量重命名：协作中 hnyls2002 将 pending_close 改为 close_on_finish，提升代码可读性。
- 设计权衡：推迟会话关闭机制避免了 KV 内存损坏，但增加了状态管理的复杂性。

风险与影响

技术风险：

1. 核心路径变更：修改了 session_controller.py 和 session_aware_cache.py 中的关键逻辑，可能引入回归 bug。
2. 内存管理复杂性：新增的 close_on_finish 标志和 KV 释放优化在并发场景下需谨慎测试。
3. 指标依赖：指标仅当 enable_streaming_session 启用时生效，若标志配置错误可能影响监控。

影响分析：

- 对用户：流式会话更稳定，内存泄漏减少，提升使用体验。
- 对系统：添加指标便于实时监控，但可能轻微增加性能开销。

- 对团队：工程师需熟悉新的会话管理逻辑，可能增加维护成本。

关联脉络

本 PR 与近期多个 PR 关联，揭示流式会话功能的持续演进：

- PR #22273：直接合并了 abort 泄漏修复，是本 PR 的基础。
- PR #22213 和 #22597：同样涉及流式会话内存管理和调度优化，形成功能线。 从历史 PR 看，sglang 仓库正加强对流式会话和内存可观测性的投入，未来可能进一步扩展相关功能。