PR #22897 完整报告

sgl-project/sglang

streaming session: trim spec v2 overshoot in cache_finished_req

合并时间：2026-04-16 05:15

原文链接：http://prhub.com.cn/sgl-project/sglang/pull/22897

执行摘要

- 一句话：修复流式会话中推测解码超限导致 KV 缓存错误的 bug。
- 推荐动作：值得精读，特别是 _trim_overshoot 和 _free_kv_aligned 的设计，展示了如何处理页面对齐释放和状态修剪，对理解流式会话缓存管理有参考价值。

功能与动机

PR body 指出，推测解码每轮接受多个令牌，当请求完成时，可能超过 max_new_tokens 边界，超限令牌进入 KV 缓存，导致下一个回合的 token/KV 不匹配，引发注意力错误。必须修剪 KV 状态以避免此问题，确保流式会话的正确性。

实现拆解

1. 入口点修改：在 cache_finished_req 方法中计算 finished_len 并调用 _trim_overshoot。
2. 核心修剪逻辑：新增 _trim_overshoot 方法，计算目标位置 target = origin + finished_len，使用 _free_kv_aligned 释放超限 KV，并更新 kv_allocated_len、kv_committed_len 和 output_ids。
3. KV 释放辅助函数：新增 _free_kv_aligned 方法，处理页面对齐的 KV 释放，避免部分页面释放损坏已提交令牌。
4. 重构现有逻辑：修改 _free_tail 方法使用 _free_kv_aligned，统一释放逻辑，提高代码可维护性。

关键文件：

- python/sglang/srt/mem_cache/session_aware_cache.py（模块 会话缓存；类别 source；类型 core-logic；符号 _trim_overshoot, _free_kv_aligned）：核心缓存管理文件，修复流式会话中超限问题，涉及关键状态修剪和 KV 释放逻辑。

关键符号：cache_finished_req, _trim_overshoot, _free_kv_aligned, _free_tail

评论区精华

review 评论指出 _trim_overshoot 方法缺少对 req.swa_evicted_seqlen 的更新，这可能导致滑动窗口注意力（SWA）状态不一致。此问题在后续 PR #22900 中得到修复，体现了设计权衡和状态管理的重要性。

- 缺失 swa_evicted_seqlen 更新 (correctness): 后续 PR #22900 修复了此问题，添加了 swa_evicted_seqlen 的更新。

风险与影响

- 风险：风险包括：1) 回归风险：修改核心缓存路径可能影响其他会话逻辑，需测试覆盖；2) 性能影响：新增页面对齐释放可能增加微小开销，但避免了内存泄漏；3) 兼容性：仅影响使用推测解码的流式会话，其他路径不受影响；4) 状态一致性：未更新 swa_evicted_seqlen 可能导致 SWA 泄漏，但已在 #22900 修复。
- 影响：对用户：修复流式会话中推测解码的正确性，避免输出错误，提升用户体验。对系统：确保 KV 缓存与令牌一致，提高系统可靠性和内存管理效率。对团队：通过重构释放逻辑，减少代码重复，便于未来维护和扩展。
- 风险标记：核心路径变更，状态一致性风险

关联脉络

- PR #22900 trim_overshoot: cap swa_evicted_seqlen + unit test: 修复本 PR 中缺失的 swa_evicted_seqlen 更新，并提供单元测试，是直接后续补丁。
- PR #22862 Streaming session: fix retract tail leak via _free_tail: 同为流式会话正确性修复，涉及类似 KV 泄漏问题，共享 _free_kv_aligned 逻辑。
- PR #22651 streaming session: spec v2 bonus accounting + comprehensive test matrix: 总览 PR，链接到本 PR 作为流式会话正确性工作的一部分，提供上下文和测试覆盖。