PR #22735 完整报告

sgl-project/sglang

Delete dead rematch path in SessionAwareCache.release_session

合并时间：2026-04-14 08:02

原文链接：http://prhub.com.cn/sgl-project/sglang/pull/22735

执行摘要

本 PR 删除了 SessionAwareCache.release_session 中从未执行的 rematch 死代码路径，简化了缓存释放逻辑，并清理了相关单元测试。变更不影响生产行为，属于代码维护性优化。

功能与动机

变更动机源于 #21875 中引入的 rematch 逻辑，该逻辑自引入起即被 TODO 注释禁用，从未在生产中执行。作者指出，radix 树分裂会原地修改 TreeNode 对象，使得 slot.last_node 和 slot.cache_protected_len 保持有效，无需重新匹配前缀。删除死代码可以减少代码复杂性和潜在混淆。

实现拆解

- session_aware_cache.py：删除 _resolve_release_state 方法，将 release_session 方法简化为直接使用 slot.last_node 和 cache_protected_len，并移除 req 参数。
- session_controller.py：更新调用，不再传递 req 参数给 release_session。
- test_streaming_session_unit.py：删除两个专门测试死代码路径的单元测试 test_release_session_recomputes_current_tree_owned_prefix 和 test_release_session_never_grows_tree_owned_prefix。

评论区精华

review 评论为空，但 PR body 中详细阐述了删除理由：rematch 逻辑因 TODO 禁用从未执行，且调用 match_prefix 可能导致分裂副作用，因此信任现有 slot 状态是安全的。

风险与影响

- 风险：极低，删除的是死代码，剩余逻辑经过现有测试验证。需确保调用点更新正确，但变更简单。
- 影响：仅内部缓存管理模块，不改变用户功能。简化代码库，减少维护负担。

关联脉络

本 PR 与 #21875 直接相关，#21875 引入了 rematch 逻辑但禁用了它，导致死代码积累。此次清理是代码卫生的一部分，反映了团队对代码质量的关注。