

PR #36721 完整报告

sgl-project/sglang

[mem_cache] Add `free\_kv\_row` to release a request's kv row by row range

合并时间: 2026-09-01 16:14

原文链接: <http://prhub.com.cn/sgl-project/sglang/pull/36721>

执行摘要

- 一句话:
- 推荐动作:

功能与动机

实现拆解

1. 核心源码路径 - `python/sglang/srt/mem_cache/common.py` (core-logic) : 源码主路径 ; 涉及符号 `free_kv_row_segments`; 包含 控制流调整、配置键调整、符号定义调整; +40/-4; 关键符号: `free_kv_row_segments` - `python/sglang/srt/session/streaming_session.py` (core-logic) : 源码主路径; 涉及符号 `_free_kv_aligned`; 包含 控制流调整、配置键调整、符号定义调整; +14/-16; 关键符号: `_free_kv_aligned` - `python/sglang/srt/mem_cache/base_prefix_cache.py` (core-logic) : 源码主路径; 涉及符号 `free_kv_row`; 包含 导入关系调整、配置键调整、符号定义调整; +13/-0; 关键符号: `free_kv_row`
2. 测试与验证 - `test/registered/unit/mem_cache/test_swa_unittest.py` (test-coverage) : 测试配套; 涉及符号 `_SinglePoolAllocator`, `init`, `clear`, `alloc`; 包含 测试覆盖调整、导入关系调整、控制流调整; +121/-1; 关键符号: `_SinglePoolAllocator`, `init`, `clear`, `alloc` - `test/registered/unit/mem_cache/test_streaming_session_unit.py` (test-coverage) : 测试配套; 涉及符号 `_FakeAllocator`, `init`, `clear`, `alloc`; 包含测试覆盖调整、导入关系调整、控制流调整; +56/-6; 关键符号: `_FakeAllocator`, `init`, `clear`, `alloc` - `test/registered/unit/disaggregation/test_specv2_kvcache_offloading.py` (test-coverage) : 测试配套; 涉及符号 `_RecordingAllocator`, `init`, `clear`, `alloc`; 包含 测试覆盖调整、导入关系调整、控制流调整; +34/-6; 关键符号: `_RecordingAllocator`, `init`, `clear`, `alloc`

关键文件:

- `test/registered/unit/mem_cache/test_swa_unittest.py` (模块 `swa/unittest`; 类别 `test`; 类型 `test-coverage`; 符号 `_SinglePoolAllocator`, `init`, `clear`, `alloc`) : 测试配套; 涉及符号 `_SinglePoolAllocator`, `init`, `clear`, `alloc`; 包含 测试覆盖调整、导入关系调整、控制流调整; +121/-1
- `python/sglang/srt/mem_cache/common.py` (模块 `sglang/srt`; 类别 `source`; 类型 `core-logic`; 符号 `free_kv_row_segments`) : 源码主路径; 涉及符号

free_kv_row_segments; 包含 控制流调整、配置键调整、符号定义调整; +40/-4

- test/registered/unit/mem_cache/test_streaming_session_unit.py (模块 streaming/session/unit; 类别 test; 类型 test-coverage; 符号 _FakeAllocator, init, clear, alloc) : 测试配套; 涉及符号 _FakeAllocator, init, clear, alloc; 包含 测试覆盖调整、导入关系调整、控制流调整; +56/-6
- python/sglang/srt/session/streaming_session.py (模块 sglang/srt; 类别 source; 类型 core-logic; 符号 _free_kv_aligned) : 源码主路径; 涉及符号 _free_kv_aligned; 包含 控制流调整、配置键调整、符号定义调整; +14/-16
- test/registered/unit/disaggregation/test_specv2_kvcache_offloading.py (模块 specv2/kvcache/offloading; 类别 test; 类型 test-coverage; 符号 _RecordingAllocator, init, clear, alloc) : 测试配套; 涉及符号 _RecordingAllocator, init, clear, alloc; 包含 测试覆盖调整、导入关系调整、控制流调整; +34/-6

关键符号: _SinglePoolAllocator, init, clear, alloc, free, _RowCache, TestFreeKvRow, setUp, free_kv_row_segments, _FakeAllocator

评论区精华

- 暂无高价值评论线程

风险与影响

- 风险:
- 影响:
 - 风险标记: 暂无

关联脉络

- 暂无明显关联 PR