PR #22548 完整报告

sgl-project/sglang

[tokenizer] lazy text accumulation + use deltas directly for streaming

合并时间：2026-04-11 12:26

原文链接：http://prhub.com.cn/sgl-project/sglang/pull/22548

执行摘要

本 PR 通过统一懒文本积累并直接使用增量数据流，优化了 tokenizer 管理器的流式处理性能，消除了 O(N) 的字符串切片复制，提升长输出场景下的吞吐量。

功能与动机

基于 #20310 的 tokenizer-2 优化，本 PR 旨在进一步统一流式和非流式路径的文本积累逻辑。动机是解决流式输出中每令牌的 O(N) 切片问题，PR body 指出：“eliminates O(N) slice per token on both text and output_ids”，基准测试显示在输出长度 1024、4096、16384 时性能有显著提升。

实现拆解

主要变更集中在 python/sglang/srt/managers/tokenizer_manager.py：

- ReqState 类：移除 buffer_text 和 last_text_offset 字段，引入 text_chunks 列表用于懒积累文本。append_text 方法现在始终将 chunk 添加到 text_chunks，get_text 方法缓存 materialized 前缀并清除 chunks。 ```python def append_text(self, chunk: str): if chunk: self.text_chunks.append(chunk)

def get_text(self) -> str: if self.text_chunks: self.text += "".join(self.text_chunks) self.text_chunks.clear() return self.text `` - **_handle_batch_output 函数**：直接使用 delta_text和delta_output_ids，避免切片操作。例如，在流式增量模式下，output_text = delta_text而不是text[state.last_text_offset :]。 - **清理死代码**：移除make_req_state` 工厂函数及相关字段。

测试文件 test/manual/test_tokenizer_manager.py 相应更新，移除 buffer_text 相关测试，验证新懒积累逻辑。

评论区精华

Review 过程仅有 hnyls2002 的批准，无实质性讨论。这表明变更被快速接受，但缺乏技术深度交锋。

风险与影响

风险：

- 懒积累缓存逻辑在 get_text() 中可能引入错误，如 text_chunks 管理不当导致数据丢失。
- 直接使用 delta 依赖上游数据正确性，需确保 recv_obj 提供准确的增量。
- 移除 buffer_text 可能影响非流式请求的边缘情况，但测试覆盖。

影响：

- 对用户：流式输出性能提升，减少内存复制，尤其在生成长文本时。
- 对系统：提高吞吐量，降低延迟。
- 对团队：代码简化，移除冗余，便于维护。

关联脉络

本 PR stacked on #20310（tokenizer-2），是 tokenizer 优化系列的一部分。从近期历史 PR 看，类似性能优化和重构常见，如 PR #22517 使用 reshape 避免内存复制，表明团队持续关注核心路径性能。