PR #22471 完整报告

sgl-project/sglang

[Spec][Ngram] Return token counts in list_external_corpora API

合并时间：2026-04-11 12:50

原文链接：http://prhub.com.cn/sgl-project/sglang/pull/22471

执行摘要

本 PR 将 list_external_corpora API 的返回值从语料库 ID 列表扩展为包含 token 计数的字典，为管理员提供更详细的语料库洞察，是 Ngram 推测解码系列的一部分，影响范围从 C++ 内核到 HTTP 接口。

功能与动机

动机源于 Issue #21052 的 Ngram 推测解码路线图，管理员需要监控外部语料库的 token 使用情况。PR body 明确指出："Admin will manage the external corpora themselves, we need to give them more insights into their corpora (token counts of each corpus)"，这直接驱动了 API 的增强需求。

实现拆解

变更涉及全栈更新：

- C++ 层：在 SuffixAutomaton 类新增 tokenCount() 方法返回 pos_ 计数器；Ngram::listExternalCorpora() 改为返回 std::pair<std::string, int64_t> 向量。cpp // 示例代码：suffix_automaton.h int64_t tokenCount() const { return pos_; }
- FFI 层：ngram_corpus_ffi.cpp 中更新字符串编码，使用制表符分隔 ID 和计数（例如 "id\t123"）。
- Python 层：从 jit_kernel/ngram_corpus.py 到 srt/speculative/ 的多个模块，将返回值类型从 List[str] 更新为 Dict[str, int]。
- HTTP 层：http_server.py 中响应字段从 corpus_ids 重命名为 corpus_token_counts。
- 测试层：test/registered/unit/spec/test_ngram_corpus.py 更新验证字典返回和计数正确性。

评论区精华

由于 review 评论为空，未发生具体技术讨论，PR 的迭代主要通过提交历史体现，如分隔符从逗号改为制表符以规避 corpus ID 包含逗号的风险。

风险与影响

风险：

- API 变更可能破坏现有客户端代码，但通过全栈同步更新缓解。

- 分隔符选择（制表符）避免了 corpus ID 包含逗号的问题，但若 ID 包含制表符仍可能解析错误（概率较低）。
- 测试覆盖了基本功能，但边缘案例（如特殊字符 ID）需进一步确认。

影响：

- 用户：管理员获得 token 计数数据，便于资源优化。
- 系统：API 响应格式变更，需客户端适配；对内部性能无显著影响。
- 团队：工程师需关注新 API 结构，但变更范围有限且已全栈同步。

关联脉络

此 PR 是 Issue #21052 系列工作的一部分，近期 PR 如 #22487 同样涉及 Ngram 模块清理，显示了团队在增强推测解码功能（尤其是外部语料库支持）上的持续投入，预计未来将有更多相关 PR 跟进。