PR #22203 完整报告

sgl-project/sglang

[Spec][Ngram] Support multiple SAMs with dynamic HTTP API

合并时间：2026-04-07 09:49

原文链接：http://prhub.com.cn/sgl-project/sglang/pull/22203

执行摘要

此 PR 为 Ngram 推测解码引入了多 SAM（后缀自动机）动态 HTTP API 支持，允许用户在运行时通过新增的 HTTP 端点（添加、移除、列出）管理外部语料库，无需重启服务器。实现包括 C++ 层多 SAM 存储、非阻塞后台加载和预算平均分配，保持向后兼容。这是一个有意义的改进，增强了系统灵活性，但需关注并发限制和预算分配风险。

功能与动机

动机源于 Ngram 推测解码的灵活性不足问题：之前仅支持通过启动参数 --speculative-ngram-external-corpus-path 静态加载单个外部 SAM，用户无法在运行时动态调整语料库。根据 PR body，此变更是 Ngram 重构系列（Issue #21052）的一部分，跟随 PR #21425，旨在提升用户体验和系统适应性。关键表述："Users may want to add/remove corpora at runtime without restarting the server."

实现拆解

实现按模块拆解如下：

模块	关键改动	说明
C++ 核心层(ngram.cpp)	替换 sam_ 为 sams_ 映射；新增 staging_sam_；修改 batchMatch 分配预算	核心逻辑变更，支持多 SAM 存储和匹配
FFI 接口(ngram_corpus_ffi.cpp)	添加 remove_external_corpus、list_external_corpora 方法	暴露 C++ 功能给 Python 层
Python 层(ngram_corpus.py)	更新 NgramCorpus 类，支持命名语料库操作	封装底层调用，提供高级 API
HTTP 服务器(http_server.py)	新增 /add_external_corpus、/remove_external_corpus、/list_external_corpora 端点	提供外部管理接口，支持文件或文档输入
调度器(scheduler.py)	集成 ExternalCorpusManager，添加请求处理逻辑	协调异步加载和请求转发

模块	关键改动	说明
令牌管理器(tokenizer_communicator_mixin.py)	新增通信器处理语料库操作	沿用现有模式，确保数据流一致
测试(test_ngram_corpus.py)	扩展单元测试，覆盖多 SAM 功能和 HTTP API	验证正确性和回归防护

关键代码片段示例（来自 ngram.cpp）：

```
void Ngram::batchMatch(...) {
    // 计算预算分配
    const size_t num_sams = sams_.size();
    const size_t total_sam_budget = num_sams > 0 ? std::min(param_.external_sam_budget,
    total_draft_token_num) : size_t{0};
    const size_t per_sam_budget = num_sams > 0 ? total_sam_budget / num_sams : size_t{0};
    const size_t trie_budget = total_draft_token_num - (per_sam_budget * num_sams);
    // 后续合并各SAM结果
}
```

评论区精华

Review 讨论由 gemini-code-assist[bot] 发起，重点包括：

- 异常处理 bug：在 ngram_corpus.py 中，异常时调用 clear_external_corpus() 会清除所有语料库，reviewer 指出这是错误行为，建议专用清理。结论：添加 FIXME 注释，但未完全解决。
- 预算分配计算：reviewer 发现 trie_budget 计算可能错误，整数除法余数被分配给 trie。提交中已调整，但分配逻辑仍需关注。引用原话："The difference (total_sam_budget % num_sams) is currently given to the trie, which might not be the intended behavior."
- Mutex 保护注释：reviewer 建议恢复注释以明确保护范围。结论：已更新注释，增强可维护性。
- 字符串连接性能：reviewer 指出 list_external_corpora 中循环连接字符串可能低效。结论：改用换行符分隔，但性能优化未实施。

风险与影响

风险：

1. 并发加载限制：C++ 代码中 FIXME 标记单 staging slot，不支持并发加载，可能导致竞争或阻塞。
2. 预算分配不均：SAM 预算平均分配可能忽略语料库特性，影响草案生成质量。
3. HTTP API 安全：端点依赖可选认证，配置不当可能引发未授权访问。
4. 回归风险：核心 batchMatch 逻辑变更可能影响推测解码正确性和性能。

影响：

- 用户：获得动态管理能力，提升使用灵活性。

- 系统：非阻塞加载减少干扰，但多 SAM 增加内存开销；HTTP API 扩展服务器功能。
- 团队：新增模块增加维护负担，但设计模式借鉴现有组件，降低学习成本。

关联脉络

此 PR 是 Ngram 推测解码演进的关键一步：

- 直接关联：PR #21425 引入了外部语料库加载基础，此 PR 在其上扩展为动态多 SAM 管理。
- 系列关联：PR #22180、#22199 等同属 Ngram 优化系列，聚焦性能改进和测试，反映团队持续投入推测解码功能增强。
- 趋势：从静态加载到动态 API，显示系统向更灵活、可运维方向演进，支持运行时调整以适应多变场景。