PR #16711 完整报告

sgl-project/sglang

Add `--stream-response-default-include-usage` server flag

合并时间：2026-04-04 12:36

原文链接：http://prhub.com.cn/sgl-project/sglang/pull/16711

执行摘要

本 PR 添加了 --stream-response-default-include-usage 服务器标志，使服务器操作者能强制在流式响应中包含使用信息，便于 token 级别监控；通过提取共享函数重构代码，移除死参数，提升可维护性。变更影响流式处理模块，风险较低，已通过测试验证。

功能与动机

动机：根据 PR body，当流式传输启用时，使用信息默认只在客户端设置 stream_options.include_usage=true 时返回，但服务器操作者需要监控所有请求的 token 使用指标，无法依赖客户端配置。作者在 Issue 评论中补充："为了监控和统计业务 token 指标"，并参考了 VLLM 框架的类似参数。

实现拆解

实现按模块拆解如下：

- 服务器参数模块(python/sglang/srt/server_args.py)：新增 stream_response_default_include_usage 布尔字段和对应 CLI 参数 --stream-response-default-include-usage，帮助文本为"Include usage in every streaming response (even when stream_options is not specified)."。
- 工具函数模块(python/sglang/srt/entrypoints/openai/utils.py)：新增 should_include_usage 函数，统一处理 stream_options 检查和服务器标志，返回 (include_usage, continuous_usage_stats) 元组，代码逻辑为：
- 服务逻辑模块：在 serving_chat.py 和 serving_completions.py 中，修改流式生成函数（如 _generate_chat_stream 和 _generate_completion_stream），调用 should_include_usage 并调整条件检查，移除重复的 stream_options 判断。
- 代码清理：移除 serving_responses.py 中的死参数 enable_force_include_usage 和 http_server.py 中对应的传递，删除 ServerArgs 中未使用的 stream_output 字段。

评论区精华

review 讨论中，hnyls2002提出关键质疑："Why is force?"，作者 syd520zy多次解释：

> "在当前的框架中，是否输出使用信息取决于用户是否主动传入此参数。当用户不传入时，服务器将无法统计此请求的实际使用信息。添加此参数后，我可以在服务器端控制是否强制所有请求输出使用信息，以进行监控和统计分析。"

reviewer 指出这实际上是设置默认值，而非强制，要求重写混淆逻辑。最终，参数被重命名为 stream_response_default_include_usage，逻辑调整为更清晰的默认值方式，并获得批准。

风险与影响

风险：

1. 兼容性风险：新标志仅在不指定 stream_options 时生效，不影响客户端显式设置，但服务器端覆盖可能意外改变行为；测试已更新以包含此属性。
2. 回归风险：流式逻辑变更可能引入错误，但 CI 测试（如 test_serving_chat.py）通过，覆盖基本场景。
3. 性能风险：额外使用信息块可能轻微增加流式数据量，对性能影响可忽略。

影响：

- 对用户：服务器操作者获得监控所有请求 token 使用的能力，无需客户端配合，提升运维便利性。
- 对系统：流式响应可能增加最终使用块，但对整体系统性能影响微小。
- 对团队：代码重构（提取共享函数、移除死代码）提升了可维护性，减少未来维护成本。

关联脉络

从历史 PR 看，PR 22065同样修改了 server_args.py，涉及服务器参数配置的优化（HiSparse 功能限制），显示该模块的持续演进。本 PR 是流式处理功能的一部分，可能与其他 openai 服务改进（如 PR 21342 的重构）间接相关，共同推动代码清晰度和功能扩展。