PR #21900 完整报告

sgl-project/sglang

Return HTTP 400 for streaming validation errors

合并时间：2026-04-02 12:58

原文链接：http://prhub.com.cn/sgl-project/sglang/pull/21900

执行摘要

该 PR 修复了 SGLang OpenAI 兼容 API 中流式请求验证错误返回状态码不一致的 bug。当输入令牌超过上下文长度时，流式请求（stream=true）现在会正确返回 HTTP 400，而非之前的 HTTP 200 附带 SSE 错误负载。这一变更统一了与 vLLM 的行为，提升了 API 的兼容性和一致性。实现通过预启动流式生成器来捕获验证错误，确保在 HTTP 响应发送前返回正确状态码。

功能与动机

根据 Issue #19996，用户报告在输入令牌数超过模型上下文长度时，SGLang 的 /v1/chat/completions 端点行为不一致：非流式请求正确返回 HTTP 400，而流式请求返回 HTTP 200 并附带错误负载（error.code=400）。这与 vLLM 的行为不符，vLLM 在两种情况下均返回 HTTP 400。PR 的目标是修复此不一致性，使流式请求在验证失败时也返回 HTTP 400，从而提升 API 的标准化和客户端兼容性。

实现拆解

实现涉及三个关键文件，按模块拆解如下：

1. OpenAI 服务器模块：
   - serving_chat.py 和 serving_completions.py 中的 _handle_streaming_request 方法被修改，返回类型从 StreamingResponse 扩展为 Union[StreamingResponse, ErrorResponse]。
   - 新增预启动逻辑：通过 await generator.__anext__() 触发验证，如果捕获 ValueError 则直接返回 create_error_response（HTTP 400）。
   - 代码示例：
2. 流生成逻辑：
   - 在 _generate_chat_stream 和 _generate_completion_stream 方法中添加 stream_started 标志，初始为 False。
   - 在首次 yield 后设置 stream_started = True，确保验证错误在流开始前能通过 raise 传播到调用方。
   - 修改 try/except 块，仅在 stream_started 为 False 时重新抛出验证错误。
3. 测试模块：

- test_request_length_validation.py 新增 test_input_length_longer_than_context_length_streaming 测试，验证流式请求在上下文超限时抛出 openai.BadRequestError（对应 HTTP 400）。

评论区精华

由于 review 评论为空，无具体讨论内容可提炼。但提交历史显示作者通过三次提交逐步完善解决方案：

- 首次提交引入预启动生成器机制。
- 第二次提交添加测试覆盖。
- 第三次提交修复生成器内部错误吞没问题，通过 stream_started 标志确保验证错误正确传播。

风险与影响

- 技术风险：预启动生成器可能轻微增加流式请求的延迟，但仅在验证失败时有额外开销；修改核心错误处理路径需谨慎测试以避免回归。
- 兼容性影响：API 行为从返回 HTTP 200 改为 HTTP 400，可能破坏依赖旧行为的客户端，需在更新日志中明确说明。
- 测试覆盖：新增测试覆盖了上下文长度超限场景，但建议扩展测试以覆盖其他验证错误（如令牌数超限）。
- 性能影响：对正常流式请求影响极小，因为预启动仅涉及一次异步调用。

关联脉络

- 与 Issue #19996 直接关联，解决了该 bug 报告中描述的不一致性问题。
- 在近期历史 PR 中，PR #21463（迁移 API 端点）同样关注 API 表面的一致性，但本 PR 更专注于错误处理逻辑。
- 本 PR 是 OpenAI 兼容 API 维护的一部分，反映了团队对标准化和兼容性的持续投入。