PR #20625 完整报告

sgl-project/sglang

[Bug Fix] Fix non-streaming request abort failure when --enable-metrics is enabled

合并时间：2026-03-23 10:58

原文链接：http://prhub.com.cn/sgl-project/sglang/pull/20625

执行摘要

本 PR 修复了启用指标时非流式请求中止失效的 bug，通过引入纯 ASGI 中间件适配器替代 Starlette 的 BaseHTTPMiddleware，恢复 request.is_disconnected() 功能，确保客户端断开时请求能正确中止。新增单元测试验证修复，不影响现有指标收集。

功能与动机

当启用 --enable-metrics 时，非流式请求在客户端断开（如 curl 被 Ctrl+C 取消）后不会中止，导致服务器资源浪费。Issue #20623 详细描述了此问题：根因是 @app.middleware("http") 使用的 BaseHTTPMiddleware 替换了 ASGI receive callable，破坏了 request.is_disconnected()。修复后能恢复中止功能，提升系统资源管理。

实现拆解

- 修补中间件：在 python/sglang/srt/utils/common.py 中添加 patch_app_http_middleware(app) 调用，在指标中间件应用前修补。
- 纯 ASGI 适配器：新增 python/sglang/srt/utils/http_middleware_patch.py，定义 _PureASGIDispatch 类，其 __call__ 方法传递 receive 不变，保持 request.is_disconnected() 工作。python async def __call__(self, scope, receive, send): if scope["type"] != "http": await self.app(scope, receive, send) return request = Request(scope, receive) ...
- 单元测试：新增 test/registered/scheduler/test_abort_with_metrics.py，使用 mock 模拟 ASGI 环境，直接测试 is_disconnected() 行为。

评论区精华

- gemini-code-assist[bot]：建议将 import 移动到文件顶部，遵循 PEP 8 风格规范。

"To adhere to PEP 8 guidelines and improve code clarity, this import should be moved to the top of the file with other imports."

- hnyls2002：建议改进测试策略，从集成测试改为单元测试，提高效率。

"The test in test_abort_with_metrics.py launches a full server with model loading (est ~180s)... This can be tested much faster and more precisely with unittest.mock." 最终测试被重写采纳此建议。

风险与影响

- 技术风险：中间件替换可能干扰其他中间件；依赖 Starlette 内部机制，未来更新可能失效；测试覆盖主要针对中间件层，需确保端到端场景被现有测试覆盖。
- 影响分析：用户端 bug 修复，避免资源浪费；系统恢复中止功能，改善内存管理；团队引入补丁增加维护负担，但提供了针对 Starlette 已知问题的解决方案。

关联脉络

本 PR 直接关联 Issue #20623，是该 issue 的具体实现修复。未提供历史 PR 信息，但从讨论看，测试策略优化体现了团队对测试效率的重视，可能影响后续测试设计。