

PR #28349 完整报告

sgl-project/sglang

[Fix] Demote OpenAIServingResponses init failure log to one-line WARNING

合并时间: 2026-06-16 07:58

原文链接: <http://prhub.com.cn/sgl-project/sglang/pull/28349>

执行摘要

- 一句话: 降级 Responses API 初始化失败日志级别
- 推荐动作: 简单直接的日志优化, 值得阅读以了解团队日志策略。

功能与动机

当 OpenAIServingResponses 初始化失败 (如 gpt-oss harmony vocab 下载失败) 时, 异常被捕获且非致命, 服务器仍启动, 仅 /v1/responses 不可用。但原始代码以 WARNING 级别打印了完整 traceback, 使得一个可恢复的条件看起来完全像一个致命的启动错误。

实现拆解

修改 `python/sglang/srt/entrypoints/http_server.py` 中的 `lifespan` 函数内 `try/except` 块:

1. 将 `except Exception:` 改为 `except Exception as e` 以捕获异常对象。
2. 将原单行 `logger.warning(...)` 替换为两条日志:
 - 一条简短的 WARNING, 包含端点名称和异常类型及消息。
 - 一条 DEBUG, 包含完整 traceback。
3. 日志消息更加清晰: 明确说明 /v1/responses 已禁用。

关键文件:

- `python/sglang/srt/entrypoints/http_server.py` (模块 HTTP 服务; 类别 source; 类型 core-logic): 唯一修改的文件, 包含 OpenAIServingResponses 初始化的异常处理逻辑。

关键符号: 未识别

关键源码片段

`python/sglang/srt/entrypoints/http_server.py`

唯一修改的文件, 包含 OpenAIServingResponses 初始化的异常处理逻辑。

```
# python/sglang/srt/entrypoints/http_server.py (lines 383-393)
try:
    from sglang.srt.entrypoints.openai.serving_responses import (
        OpenAIServingResponses,
    )
    fast_api_app.state.openai_serving_responses = OpenAIServingResponses(
```

```
        _global_state.tokenizer_manager,  
        _global_state.template_manager,  
        enable_prompt_tokens_details=True,  
        tool_server=tool_server,  
    )  
except Exception as e:  
    # Optional endpoint; a load failure (e.g. the gpt-oss harmony vocab  
    # download) must not look like a fatal error. One-line WARNING, full  
    # traceback at DEBUG.  
    logger.warning(  
        f"OpenAI Responses API (/v1/responses) disabled: "  
        f"OpenAIServingResponses init failed ({type(e).__name__}: {e})"  
    )  
    logger.debug(  
        f"OpenAIServingResponses init traceback:\n{get_exception_traceback()}"  
    )
```

评论区精华

无 review 评论。仅有一条来自 `gemini-code-assist[bot]` 的自动配额提示。

- 暂无高价值评论线程

风险与影响

- 风险：风险极低。仅改变日志输出方式，不影响控制流或功能逻辑。DEBUG 级别的 `traceback` 在默认配置下不会被输出到生产日志。
- 影响：影响范围仅限运维观察。当 `OpenAIServingResponses` 初始化失败时，日志行从多行 `traceback` 缩短为一行，减少噪音，便于快速识别非致命问题。
- 风险标记：暂无

关联脉络

- 暂无明显关联 PR