

PR #31975 完整报告

sgl-project/sglang

Treat partial_json_parser AssertionError as incomplete JSON

合并时间: 2026-07-22 23:25

原文链接: <http://prhub.com.cn/sgl-project/sglang/pull/31975>

执行摘要

- 一句话: 修复 streaming 中 partial JSON 解析断言异常
- 推荐动作: 建议合入此 PR。它是一个典型的小型防御性修复, 虽简单但有效防止了特定输入导致的请求中断。阅读 `_partial_json_loads` 的异常处理结构可获得防御性编程启发。

功能与动机

Issue #31974 报告: 在 streaming function/tool-call 解析中, `_partial_json_loads()` 因 `partial_json_parser.fix_fast()` 对部分输入 (如尾随非空白字符后跟可修复前缀) 抛出 `AssertionError`, 导致请求失败。期望行为是返回 `JSONDecodeError` 以便调用者按“未完成”路径处理。

实现拆解

1. 修改文件: `python/sglang/srt/function_call/utils.py`
2. 核心变更: 在 `_partial_json_loads` 函数的 `except` 块中新增对 `AssertionError` 的捕获。
3. 异常转换: 将捕获到的 `AssertionError` 包装为 `JSONDecodeError` 并重新抛出, 异常消息为“`partial_json_parser assertion (treat as incomplete)`”, 保留原始异常链。
4. 影响: 仅涉及异常处理路径, 不改变其他逻辑; 现有调用者如 `tool_call_parser` 已处理 `JSONDecodeError`, 因此能正确忽略不完整的 JSON 片段。
5. 测试: 无新增测试文件, 但 PR 描述中提及手动复现验证。

关键文件:

- `python/sglang/srt/function_call/utils.py` (模块 工具调用; 类别 source; 类型 core-logic; 符号 `_partial_json_loads`): 核心文件。在 `_partial_json_loads` 函数中新增 `except AssertionError` 处理, 将断言异常转换为 `JSONDecodeError`。

关键符号: `_partial_json_loads`

关键源码片段

`python/sglang/srt/function_call/utils.py`

核心文件。在 `_partial_json_loads` 函数中新增 `except AssertionError` 处理, 将断言异常转换为 `JSONDecodeError`。

```
def _partial_json_loads(input_str: str, flags: Allow) -> Tuple[Any, int]:
    """ ... """
    try:
        return (partial_json_parser.loads(input_str, flags), len(input_str))
    except (JSONDecodeError, IndexError) as e:
        msg = getattr(e, "msg", str(e))
        if "Extra data" in msg or "pop from empty list" in msg:
            start = WHITESPACE.match(input_str, 0).end()
            obj, end = JSONDecoder().raw_decode(input_str, start)
            return obj, end
        raise
    except AssertionError as e:
        # partial_json_parser.fix_fast() 对某些部分 / 模糊输入断言失败
        # （例如在一段本可修复的前缀后带有尾随非空白字符）
        # 而不是表示 "incomplete"。转换为 JSONDecodeError 以便
        # streaming 调用者将其视为未完成（等待更多 token）
        raise JSONDecodeError(
            "partial_json_parser assertion (treat as incomplete)", input_str, 0
        ) from e
```

评论区精华

review 中无争议讨论。PR 作者直接描述了问题与修复方案，维护者 [ispobock](#) 触发了 CI 重新运行测试 [test/registered/unit/function_call/test_function_call_parser.py](#)，结果通过。

- 暂无高价值评论线程

风险与影响

- 风险：低风险。变更仅添加了一个 `except` 子句，不会影响正常解析路径。可能的风险是如果 future 版本的 `partial_json_parser` 修正了断言行为，则此 `catch` 可能变为 `dead code`，但不会导致错误。
- 影响：影响有限但重要：直接影响使用 `streaming tool-call` 功能的用户，修复后此类断言错误不再导致请求失败，提升了流式解析的鲁棒性。不影响非流式场景。
- 风险标记：低风险

关联脉络

- 暂无明显关联 PR