PR #37727 完整报告

vllm-project/vllm

[Bugfix] Fix Responses API instructions leaking through previous_response_id

合并时间：2026-04-13 16:46

原文链接：http://prhub.com.cn/vllm-project/vllm/pull/37727

执行摘要

- 一句话：修复 Responses API 中 previous_response_id 导致 instructions 泄漏的问题
- 推荐动作：该 PR 值得精读，特别是对于处理 OpenAI 兼容 API 的开发者。关注点：
  1. 理解 OpenAI Responses API 中 instructions 参数的设计意图
  2. 学习如何正确处理跨请求的消息历史管理
  3. 参考新增的测试用例，了解如何全面测试此类边界条件

功能与动机

修复 issue #37697 中报告的 bug：当使用 /v1/responses 端点并指定 previous_response_id 时，前一个响应的 instructions 会泄漏到新响应中，即使新请求提供了不同的 instructions。根据 OpenAI Responses API 规范明确说明："When using along with previous_response_id, the instructions from a previous response will not be carried over to the next response."

实现拆解

主要修改集中在两个文件：

1. vllm/entrypoints/openai/responses/utils.py 中的 construct_input_messages 函数：将 messages.extend(prev_msg) 改为 messages.extend(m for m in prev_msg if m.get("role") != "system")，过滤掉历史消息中的系统消息。
2. tests/entrypoints/openai/responses/test_responses_utils.py：新增 4 个单元测试，覆盖旧系统消息被剥离、新请求无 instructions、非系统消息保留、无历史消息等场景。

关键文件：

- vllm/entrypoints/openai/responses/utils.py（模块 frontend/responses-api）：包含核心修复逻辑，修改了 construct_input_messages 函数以过滤历史系统消息
- tests/entrypoints/openai/responses/test_responses_utils.py（模块 tests/responses-api）：新增 4 个单元测试，全面验证修复的正确性和边界情况

关键符号：construct_input_messages

评论区精华

主要讨论围绕 OpenAI 规范的确认：

- chaunceyjiang 询问是否有相关 OpenAI 规范文档链接
- he-yufeng 提供了 OpenAI API Reference 中关于 instructions 参数的明确说明，以及 Text Generation 指南中的相关描述，确认了规范要求 instructions 不应跨响应传递
- 讨论确认了修复方案符合规范要求
- OpenAI 规范确认 (correctness): 确认了规范要求 instructions 不应跨响应传递，修复方案符合规范

风险与影响

- 风险：风险较低：
  1. 回归风险：过滤系统消息可能影响依赖历史系统消息的其他功能，但根据 OpenAI 规范这是正确的行为
  2. 兼容性风险：修复后行为与 OpenAI 规范对齐，但可能影响之前依赖泄漏行为的客户端
  3. 测试覆盖：新增 4 个单元测试充分覆盖了各种场景，降低了回归风险
  4. 核心逻辑变更：仅修改了一行核心逻辑，影响范围可控
- 影响：影响范围：
  1. 对用户：修复后 /v1/responses 端点行为与 OpenAI 规范完全一致，确保 instructions 不会意外泄漏，提升 API 的可靠性和一致性
  2. 对系统：仅影响 Responses API 的消息构建逻辑，不影响其他 API 端点或核心推理引擎
  3. 对团队：修复了一个重要的规范一致性 bug，增强了 vLLM 与 OpenAI API 的兼容性
- 风险标记：规范一致性变更，API 行为调整

关联脉络

- 暂无明显关联 PR