PR #37958 完整报告

vllm-project/vllm

[Bugfix] Fix IndexError when accessing prev_tool_call_arr in OpenAIToolParser

合并时间：2026-03-25 12:06

原文链接：http://prhub.com.cn/vllm-project/vllm/pull/37958

执行摘要

本 PR 修复了在 OpenAI 工具调用的流式响应中访问 prev_tool_call_arr 时引发的 IndexError，通过调整条件逻辑确保状态正确管理，避免了崩溃问题，提升了前端聊天完成功能的稳定性。

功能与动机

旨在解决 Issue #37937 和 #36849 中报告的崩溃问题。Issue 评论中 sfeng33 指出错误发生在流式路径（serving.py:1133），而 extract_tool_calls() 从未被调用，作者随后在流式路径中添加了处理，如回应“I've added handling for the stream as well”。

实现拆解

主要改动点：

- vllm/entrypoints/openai/chat_completion/serving.py：在 chat_completion_stream_generator 函数中，引入 auto_tools_called 变量检查 tool_parser.prev_tool_call_arr 是否非空，并调整条件逻辑为： python if should_check and tool_parser and auto_tools_called: 避免在数组为空时访问。
- vllm/tool_parsers/openai_tool_parser.py：在 extract_tool_calls 方法中添加代码填充 prev_tool_call_arr，但 review 指出这可能为死代码，仅用于非流式路径。

评论区精华

- gemini-code-assist[bot]指出 openai_tool_parser.py 中的修改可能是死代码，因为 prev_tool_call_arr 主要在流式路径使用。
- sfeng33强调 bug 位于流式路径，作者确认并修复。
- 其他审核者如 yanghui1-arch 和 DarkLight1337 批准，讨论聚焦于正确性和代码清理。

风险与影响

- 风险：条件逻辑变更可能引入边界情况错误，例如在工具调用为空时跳过检查是否正确；死代码增加代码复杂度；未新增测试覆盖可能遗漏回归。
- 影响：修复了流式响应中的崩溃问题，直接影响使用 OpenAI 工具调用的用户，提升系统稳定性；对非流式路径无影响。

关联脉络

与历史 PR 如 #37920（frontend bugfix）和 #37706（结构化输出 bugfix）类似，展示了 vLLM 前端模块的持续优化和 bug 修复趋势，但本 PR 专注于工具调用和流式处理的特定场景。