PR #38158 完整报告

vllm-project/vllm

[Bugfix] Fix shared-object aliasing in n>1 streaming with tool calls

合并时间：2026-03-30 18:12

原文链接：http://prhub.com.cn/vllm-project/vllm/pull/38158

执行摘要

修复 vLLM 流式聊天完成中当 n>1 且启用工具调用时，因共享对象导致的工具调用损坏 bug。通过独立初始化每个选择的 token 历史和解析器状态，确保功能正确，并添加测试防止回归。

功能与动机

当使用流式请求（stream=true）且生成多个选择（n>1）时，如果启用了工具调用，所有选择会共享相同的 token 历史和解析器状态，导致工具调用损坏或缺失。PR body 描述："All choices produce corrupted or missing tool calls because token history and parser state are inadvertently shared across choices." 这会导致服务器返回空响应或 JSON 解析错误，影响用户正常使用。

实现拆解

主要改动在 vllm/entrypoints/openai/chat_completion/serving.py 的 chat_completion_stream_generator 函数中：

- 将 all_previous_token_ids = [[]] * num_choices 改为 all_previous_token_ids = [[] for _ in range(num_choices)]
- 将 tool_parsers: list[ToolParser | None] = [self.tool_parser(tokenizer, request.tools)] * num_choices 改为 tool_parsers: list[ToolParser | None] = [self.tool_parser(tokenizer, request.tools) for _ in range(num_choices)] 测试文件 tests/entrypoints/openai/chat_completion/test_serving_chat.py 新增了 test_streaming_n_gt1_independent_tool_parsers 测试，模拟流式生成并验证每个选择有独立解析器。

评论区精华

- gemini-code-assist: "This pull request addresses a potential bug... by correctly initializing lists of mutable objects." 强调了列表初始化中共享对象的陷阱。
- bbrowning: "Great catch... I'd love to see us add a test at some point to prevent regression on this kind of thing." 作者回应已添加测试。
  - sfeng33和 chaunceyjiang批准 PR，显示修复达成共识。

风险与影响

风险：修复本身直接，但若不注意，类似列表初始化可能在其他代码中引入共享状态问题，导致数据污染或错误。测试添加降低了回归风险，但新测试可能因模拟复杂性而维护成本较高。影响：用户能正常使用 n>1 流式工具调用，提升功能完整性；系统解析器状态独立化，避免跨选择干扰；团队通过测试案例增强代码健壮性。

关联脉络

从历史 PR 看，如 #33703 也涉及工具调用解析器的 bugfix，表明工具调用模块是活跃维护区域。此 PR 延续了前端功能的稳定性改进，展示了对 Python 编程细节的关注。