PR #40314 完整报告

vllm-project/vllm

fix: Do not make function calls when request has no tools for /v1/responses

合并时间：2026-04-20 12:17

原文链接：http://prhub.com.cn/vllm-project/vllm/pull/40314

执行摘要

- 一句话：修复 /v1/responses API 在请求无工具时仍会触发模型幻觉函数调用的问题。
- 推荐动作：该 PR 值得快速浏览，以理解 /v1/responses API 工具调用逻辑的关键修复。关注点在于 construct_tool_dicts 函数中条件判断的调整，这是修复的核心。虽然变更简单，但揭示了 API 层默认行为与工具检查的交互设计，对于维护前端入口点有参考价值。

功能与动机

根据 PR body 描述，/v1/responses API 存在一个行为不一致的问题：当请求中没有提供工具时，由于 construct_tool_dicts 函数只检查 tools 是否为 None，而不检查空列表，且 API 层无条件默认设置 tool_choice="auto"，导致即使请求没有工具，模型仍会尝试进行函数调用，产生幻觉输出。这需要与 /v1/chat/completions 的行为对齐，后者仅在确实提供工具时才设置 tool_choice="auto"。

实现拆解

1. 核心逻辑调整：修改 vllm/entrypoints/openai/responses/utils.py 中的 construct_tool_dicts 函数，将条件判断从 if tools is None or (tool_choice == "none"): 改为 if not tools or (tool_choice == "none"):。这样，当 tools 为空列表或 None 时，都会返回 None，从而禁用工具调用。
2. 影响范围：此变更直接影响 /v1/responses API 的工具调用逻辑，确保无工具请求不会触发模型生成函数调用。它修复了与 /v1/chat/completions 的行为差异，提升 API 一致性。
3. 测试与配置：本次变更仅涉及源码文件修改，没有看到直接对应的测试文件变更或配置调整。这可能意味着需要依赖现有测试覆盖，或后续补充测试以确保回归安全。

关键文件：

- vllm/entrypoints/openai/responses/utils.py（模块 入口点；类别 source；类型 core-logic；符号 construct_tool_dicts）：这是唯一变更的文件，包含修复工具调用逻辑的核心函数 construct_tool_dicts。

关键符号：construct_tool_dicts

关键源码片段

vllm/entrypoints/openai/responses/utils.py

这是唯一变更的文件，包含修复工具调用逻辑的核心函数 construct_tool_dicts。

```
def construct_tool_dicts(
    tools: list[Tool], tool_choice: ToolChoice
) -> list[dict[str, Any]] | None:
    # 修复：使用 not tools 替代 tools is None，以同时处理 None 和空列表情况
    # 当工具列表为空或 tool_choice 为 "none" 时，返回 None 以禁用工具调用
    if not tools or (tool_choice == "none"):
        tool_dicts = None
    else:
        # 否则，将工具列表转换为 API 所需的格式
        tool_dicts = [
            convert_tool_responses_to_completions_format(tool.model_dump())
            for tool in tools
        ]
    return tool_dicts
```

评论区精华

review 评论较少，主要来自自动化工具：

- gemini-code-assist[bot]指出变更使用了更符合 Python 习惯的检查方式（if not tools），没有提供进一步反馈。
- chaunceyjiang简单批准（"Thanks~"），表明变更被接受。 没有出现设计争议或未解决疑虑，变更直接且目标明确。
- 代码风格改进 (style): 变更被接受，无进一步反馈。

风险与影响

- 风险：技术风险：
  - 回归风险：低。变更仅调整条件逻辑，从只检查 None 扩展为检查空列表，不影响其他路径。但若原有代码依赖 tools 为 None 与空列表的不同行为，可能引入意外变化，不过从上下文看，这正是修复目标。
  - 兼容性：无。变更修复行为不一致，不引入新接口或破坏性改动。
  - 测试覆盖：中等。未看到测试文件变更，可能依赖现有测试；若测试未覆盖空工具列表场景，可能存在未检测的边界情况。具体到文件：vllm/entrypoints/openai/responses/utils.py 中的 construct_tool_dicts 函数，修改后需确保所有调用方正确处理返回的 None。
- 影响：用户影响：使用 /v1/responses API 且未提供工具的用户将不再收到模型生成的幻觉函数调用，提升输出质量。系统影响：修复 API 行为不一致，与 /v1/chat/completions 对齐，减少混淆。团队影响：变更微小，易于理解和维护，但需确保相关测试更新以覆盖空工具列表场景。
- 风险标记：边界条件处理，缺少测试覆盖

关联脉络

- PR #39892 [Bugfix][Responses API] Fix streaming tool calls on /v1/responses: 同属 /v1/responses API 的工具调用修复，涉及工具解析器逻辑，可视为相关功能线。
- PR #39352 [Frontend] Preserve structured output special tokens in offline LLM.chat: 同属前端入口点（entrypoints）的变更，关注 API 行为一致性。