PR #44348 完整报告

vllm-project/vllm

[Bugfix] Fix unstreamed tool call args dropped in Responses API streaming

合并时间：2026-06-03 18:19

原文链接：http://prhub.com.cn/vllm-project/vllm/pull/44348

执行摘要

- 一句话：修复 Responses API 流式工具调用参数丢失
- 推荐动作：值得快速合并。修复明确，改动量小，风险低。可关注后续是否还有类似遗漏的调用点。

功能与动机

PR body 明确指出：当工具解析器流式增量输出参数时，部分已解析但未发送的参数字段在最终 delta 中未刷新，导致客户端收到截断或空的工具调用参数。Chat Completions 路径已正确传递 finished 参数，但 Responses API 路径未传递，导致调用 _append_unstreamed_tool_args 刷新逻辑未被触发。

实现拆解

1. 修改 parse_delta 签名（vllm/parser/abstract_parser.py）：将 DelegatingParser.parse_delta 的 finished 参数从默认值 False 改为仅关键字参数（*, finished: bool），强制调用者显式传递，避免遗漏。
2. 修复 Responses API 调用点（vllm/entrypoints/openai/responses/serving.py）：在 _process_harmony_streaming_events 方法中，当 parser 存在时，调用 parse_delta 时新增 finished=output.finish_reason is not None 参数，确保最终 delta 传递 finished=True。
3. 同步更新测试（tests/parser/test_streaming.py）：在 stream_text 和 stream_chunks 辅助函数中，调用 parse_delta 时显式传入 finished=False，适配新的签名要求，保证测试通过。

关键文件：

- vllm/entrypoints/openai/responses/serving.py（模块 前端入口；类别 source；类型 core-logic）：修复核心：在 Responses API 流式路径中传递 finished 参数到 parse_delta，触发未发送工具调用参数的刷新。
- vllm/parser/abstract_parser.py（模块 解析器；类别 source；类型 core-logic；符号 parse_delta）：修改 parse_delta 签名，将 finished 参数改为仅关键字参数，强制调用者显式传递。
- tests/parser/test_streaming.py（模块 测试；类别 test；类型 test-coverage）：同步更新测试辅助函数，传递 finished=False 以适配新签名。

关键符号：DelegatingParser.parse_delta, ResponsesServing._process_harmony_streaming_events

关键源码片段

vllm/entrypoints/openai/responses/serving.py

修复核心：在 Responses API 流式路径中传递 finished 参数到 parse_delta，触发未发送工具调用参数的刷新。

```
# vllm/entrypoints/openai/responses/serving.py 第 1405-1412 行
# 关键修复：在流式循环中，调用 parse_delta 时传入 finished 参数
if parser:
    delta_message = parser.parse_delta(
        delta_text=delta_text,
        delta_token_ids=delta_token_ids,
        request=request,
        prompt_token_ids=ctx.last_output.prompt_token_ids,
        finished=output.finish_reason is not None, # 新增，触发最终 delta 的未发送参数刷新
    )
else:
    delta_message = DeltaMessage(content=output.text)
```

vllm/parser/abstract_parser.py

修改 parse_delta 签名，将 finished 参数改为仅关键字参数，强制调用者显式传递。

```
# vllm/parser/abstract_parser.py 第 341-349 行 (DelegatingParser 类 )
# 将 finished 从有默认值改为仅关键字参数，强制调用者显式传递
@abstractmethod
def parse_delta(
    self,
    delta_text: str,
    delta_token_ids: list[int],
    request: ChatCompletionRequest | ResponsesRequest,
    prompt_token_ids: list[int] | None = None,
    *, # 新增，表示后续参数只能通过关键字传递
    finished: bool, # 去掉默认值 False，强制显式指定
) -> DeltaMessage | None:
    """解析单个流式 delta，编排推理和工具调用提取。"""
```

评论区精华

审核者 bbrowning 指出最初对签名变更有些担心（将 finished 改为仅关键字参数），但随即意识到该特性最近才添加，对第三方代码的影响极低。无其他争议讨论。

- parse_delta 签名变更风险 (design): 意识到该特性近期才添加，破坏风险极低，表示认可（Looks good to me）。

风险与影响

- 风险：风险极低。仅增加一个关键字参数并调整一处调用点，行为语义不变。对 Chat Completions 路径无影响（该路径已正确传参）。测试覆盖确保回归安全。
- 影响：仅影响使用 Responses API 流式工具调用的用户。修复后，工具调用参数不再丢失，客户端可收到完整的 JSON 参数。对非流式请求或 Chat Completions 路径无影响。
- 风险标记：接口签名变更

关联脉络

- PR #44346 [Refactor] Suppress SyntaxWarning from ast.literal_eval in tool parsers: 同属工具调用相关的修复 / 清理，涉及 parser 模块。