

PR #47138 完整报告

vllm-project/vllm

[Bugfix][Tool Parser] PoolsideV1: fix logprobs AttributeError on Responses API

合并时间: 2026-07-01 03:14

原文链接: <http://prhub.com.cn/vllm-project/vllm/pull/47138>

执行摘要

- 一句话: 修复 PoolsideV1 Responses API logprobs AttributeError
- 推荐动作: 该 PR 是典型的低风险 bugfix, 代码改动清晰, 测试覆盖完善, 建议快速合并。可作为工具解析器请求类型兼容性处理的参考模式。

功能与动机

社区用户反馈 (Hugging Face discussion) PoolsideV1 模型在 /v1/responses 端点上因 logprobs 属性缺失而崩溃, 导致工具调用流式解析失败。PR body 明确指出 `extract_tool_calls_streaming` 直接读取 `request.logprobs` 导致 `AttributeError`。

实现拆解

1. 修复核心逻辑 (`vllm/tool_parsers/poolside_v1_tool_parser.py`): 在 `extract_tool_calls_streaming` 方法的第 446 行, 将原直接判断 `if request.logprobs:` 替换为多分支条件: 使用 `getattr(request, "logprobs", None)` 兼容 `ChatCompletionRequest`, 再通过 `isinstance(request, ResponsesRequest)` and `request.is_include_output_logprobs()` 覆盖 `ResponsesRequest`。
2. 新增测试辅助函数 (`tests/tool_parsers/test_poolside_v1_tool_parser.py`):
 - 扩展 `_build_responses_request` 函数, 新增可选参数 `include`, 允许构造带 logprobs 选项的请求。
 - 新增 `_stream_partial_start_token` 辅助函数, 封装流式解析起始 token 的调用逻辑, 减少重复。
3. 新增两个测试用例 (`tests/tool_parsers/test_poolside_v1_tool_parser.py`):
 - `test_streaming_responses_request_without_logprobs`: 验证无 logprobs 时返回 `None`。
 - `test_streaming_responses_request_with_logprobs_emits_empty_delta`: 验证有 logprobs 时返回空 delta 消息, 不崩溃。

关键文件:

- `vllm/tool_parsers/poolside_v1_tool_parser.py` (模块 工具解析器; 类别 source; 类型 core-logic; 符号 `extract_tool_calls_streaming`): 核心修复文件, 修改 `extract_tool_calls_streaming` 方法中 logprobs 判断逻辑, 新增 `ResponsesRequest` 分支。

- tests/tool_parsers/test_poolside_v1_tool_parser.py (模块 工具解析器; 类别 test; 类型 test-coverage; 符号 _build_responses_request, _stream_partial_start_token, test_streaming_responses_request_without_logprobs, test_streaming_responses_request_with_logprobs_emits_empty_delta) : 新增两个流式 Responses 测试用例, 覆盖带 / 不带 logprobs 的场景, 防止回归。

关键符号: extract_tool_calls_streaming, _stream_partial_start_token, _build_responses_request

关键源码片段

vllm/tool_parsers/poolside_v1_tool_parser.py

核心修复文件, 修改 extract_tool_calls_streaming 方法中 logprobs 判断逻辑, 新增 ResponsesRequest 分支。

```
# vllm/tool_parsers/poolside_v1_tool_parser.py (head version, lines 444-453)
tool_calls = list(pending_deltas.values())
if content is None and len(tool_calls) == 0:
    # 修复: 兼容两种请求类型的 logprobs 判断
    # ChatCompletionRequest 使用 getattr 安全读取, ResponsesRequest 调用专用方法
    wants_logprobs = getattr(request, "logprobs", None) or (
        isinstance(request, ResponsesRequest)
        and request.is_include_output_logprobs()
    )
    if wants_logprobs:
        return DeltaMessage(content="")
    return None
return DeltaMessage(content=content, tool_calls=tool_calls)
```

tests/tool_parsers/test_poolside_v1_tool_parser.py

新增两个流式 Responses 测试用例, 覆盖带 / 不带 logprobs 的场景, 防止回归。

```
# tests/tool_parsers/test_poolside_v1_tool_parser.py (head version, lines 80-248 excerpt)
def _build_responses_request(
    *, tool_choice: str | dict[str, Any], include: list[str] | None = None
) -> ResponsesRequest:
    return ResponsesRequest(
        model="poolside-test",
        input=[{"role": "user", "content": "write the file"}],
        tools=[_responses_write_file_tool()],
        tool_choice=tool_choice,
        stream=True,
        max_output_tokens=200,
        include=include, # 新增参数, 用于传递 logprobs 选项
    )

def _stream_partial_start_token(request: ResponsesRequest):
    """辅助函数: 流式解析工具调用起始 token"""
    parser = _make_parser(request)
```

```
delta = parser.tool_call_start_token[0]
return parser.extract_tool_calls_streaming(
    previous_text="",
    current_text=delta,
    delta_text=delta,
    previous_token_ids=[],
    current_token_ids=[],
    delta_token_ids=[],
    request=request,
)
```

```
def test_streaming_responses_request_without_logprobs() -> None:
```

```
    """无 logprobs 时应返回 None"""
```

```
    request = _build_responses_request(tool_choice="auto")
```

```
    assert _stream_partial_start_token(request) is None
```

```
def test_streaming_responses_request_with_logprobs_emits_empty_delta() -> None:
```

```
    """有 logprobs 时应返回空 delta 消息，不崩溃"""
```

```
    request = _build_responses_request(
```

```
        tool_choice="auto", include=["message.output_text.logprobs"]
```

```
    )
```

```
    result = _stream_partial_start_token(request)
```

```
    assert result is not None
```

```
    assert result.content == ""
```

评论区精华

无实质性 review 讨论。sfeng33 直接批准。claude[bot] 因 fork 仓库自动评论指出无法执行自动 review。

- 暂无高价值评论线程

风险与影响

- 风险：变更范围极小（源码修改 6 行，测试新增 32 行），风险低。但需注意：若未来新增请求类型未覆盖，类似 `AttributeError` 可能重现。建议后续关注工具解析器的请求类型扩展。
- 影响：直接影响 `PoolsideV1` 工具解析器在 `Responses API` 流式场景下的稳定性。修复后，使用 `/v1/responses` 端点且开启 `logprobs` 的用户不再遭遇崩溃。对其他模型和 `ChatCompletion API` 无影响。
- 风险标记：影响面小，低风险

关联脉络

- PR #46875 [Parser][Bugfix] Ensure tool call or other special tokens don't leak in non-streaming tool parsing: 同为工具解析器 bugfix，涉及 `Responses API` 和 `logprobs` 相关修正，与本 PR 有相同模块和风险模式。