

PR #44907 完整报告

vllm-project/vllm

[Cohere] Cohere2 moe parser fix

合并时间: 2026-06-09 21:32

原文链接: <http://prhub.com.cn/vllm-project/vllm/pull/44907>

执行摘要

- 一句话: 修复 Cohere2 tool parser 和 MoE 权重加载异常
- 推荐动作: 推荐阅读, 该 PR 展示了如何处理流式 tool call delta 中的占位符 (null/ 空字符串) 问题, 设计模式 (pending ID 缓存) 值得其他 tool parser 参考。建议后续补充单元测试。

功能与动机

Cohere 模型在流式输出 tool call 时, Melody 框架会产生带空字符串占位符的 delta, 导致客户端收到无效的 tool name 或空 argument 片段, 影响工具调用功能的稳定性。此外, Cohere2 MoE 模型因 FusedMoE API 重构 (PR #41184) 而无法加载权重, 需要紧急修复。

实现拆解

1. 在 `BaseCohereCommandToolParser.__init__` 中新增 pending ID 缓存: 在 `cohere_command_tool_parser.py` 添加 `self._pending_streaming_tool_call_ids: dict[int, str] = {}` 字典, 用于暂存 Melody 提前发出的 tool-call id, 直到真正的 function name delta 出现。
2. 重写 `extract_tool_calls_streaming` 方法: 将原先的直接构造 `DeltaMessage` 改为逐条过滤逻辑——跳过 name 和 arguments 均为 None/ 空字符串的占位符 delta; 当 name 非空时, 从 pending 字典中取出对应 index 的 id 并附带在 delta 中; 对于仅有 arguments 的后续 delta 只发送 index 和 `function.arguments`, 避免重复发送 id。
3. 修复 Cohere2 MoE 权重加载: 在 `cohere2_moe.py` 中将 `FusedMoE.make_expert_params_mapping` 替换为 `fused_moe_make_expert_params_mapping`, 以适配 upstream 的 API 变更。

关键文件:

- `vllm/tool_parsers/cohere_command_tool_parser.py` (模块 工具解析器; 类别 source; 类型 core-logic; 符号 `_pending_streaming_tool_call_ids`, `extract_tool_calls_streaming`): 核心变更文件, 修改了流式 tool call 的 delta 生成逻辑, 新增 pending ID 缓存, 修复占位符泄露问题。

关键符号: `extract_tool_calls_streaming`

关键源码片段

vllm/tool_parsers/cohere_command_tool_parser.py

核心变更文件，修改了流式 tool call 的 delta 生成逻辑，新增 pending ID 缓存，修复占位符泄露问题。

```
class BaseCohereCommandToolParser(ToolParser):
    def __init__(
        self,
        tokenizer: TokenizerLike,
        streaming_opts: PyFilterOptions,
        unary_opts: PyFilterOptions,
    ):
        super().__init__(tokenizer)
        self.melody_streaming = PyFilter(streaming_opts)
        self.melody_unary = PyFilter(unary_opts)
        # Melody can emit the tool-call id before the function name. Keep it
        # until the first real name delta so clients receive both together.
        self._pending_streaming_tool_call_ids: dict[int, str] = {}

    # ... other methods ...

    def extract_tool_calls_streaming(
        self, previous_text, current_text, delta_text, previous_token_ids,
        current_token_ids, delta_token_ids, request
    ) -> DeltaMessage | None:
        r = self.melody_streaming.write_decoded(delta_text)
        if r.content is not None:
            return DeltaMessage(content=r.content)
        if r.reasoning is not None:
            return DeltaMessage(reasoning=r.reasoning)
        if r.tool_calls:
            tool_calls: list[DeltaToolCall] = []
            for tc in r.tool_calls:
                # 缓存首次出现的 tool call id, 按 index 索引
                if tc.id:
                    self._pending_streaming_tool_call_ids[tc.index] = tc.id
                name = tc.name or None
                arguments = tc.arguments or None
                # 跳过 name 和 arguments 均为 None 的占位符 delta
                if name is None and arguments is None:
                    continue

            function_kwargs = {}
            if name is not None:
                function_kwargs["name"] = name
            if arguments is not None:
                function_kwargs["arguments"] = arguments
            tool_call_kwargs = {
                "index": tc.index,
```

```

        "function": DeltaFunctionCall(**function_kwargs),
    }
    # 仅当 name 出现时才附带 id 和 type, 避免重复发送 id
    if name is not None:
        tool_call_id = tc.id or self._pending_streaming_tool_call_ids.pop(
            tc.index, None
        )
        if tool_call_id is not None:
            tool_call_kwargs["id"] = tool_call_id
            tool_call_kwargs["type"] = "function"
        tool_calls.append(DeltaToolCall(**tool_call_kwargs))
    if tool_calls:
        return DeltaMessage(tool_calls=tool_calls)
    return None

```

评论区精华

Reviewer @sfeng33 在 [cohere2_moe.py](#) 第 456 行询问权重加载的改动是否与 tool parser 修复相关。作者 @Terrencezzj 解释该改动是必要的，因为 PR #41184 当天刚刚合并，导致 Cohere2 模型权重加载失效。该讨论无其他争议，reviewer 随后批准了 PR。

- Cohere2 MoE weight loading 改动是否属于本 PR (question): 作者 @Terrencezzj 说明该改动是必需的，因为 PR #41184 当天合并导致 Cohere2 模型权重加载失败，属于同次提交中的附带修复。

风险与影响

- 风险：
 1. 回归风险：流式 tool call delta 的行为变更可能影响其他基于 Cohere 模型或 Melody 框架的 tool parser，但影响范围仅限于 BaseCohereCommandToolParser 的子类。
 2. API 兼容性：改动仅影响 OpenAI 兼容的流式响应中 tool call 字段的输出格式，移除空占位符属于正确性修复，风险低。
 3. 缺少测试覆盖：PR 未包含对应的测试用例，后续回归风险较高。
- 影响：
 1. 用户影响：Cohere Command 系列模型（如 Command-R+、Command-A）的流式 tool calling 功能恢复正常，客户端不再收到无效的 tool name 或空参数片段。
 2. 系统影响：Cohere2 MoE 模型可正常加载权重，避免因上游 API 变更导致的崩溃。
 3. 团队影响：影响范围仅限于 Cohere 模型相关代码，改动量小（+33/-12）。- 风险标记：缺少测试覆盖，核心路径变更

关联脉络

- PR #41184 FusedMoE API 重构（假设）：导致 Cohere2 MoE 权重加载的 FusedMoE.make_expert_params_mapping 变更，本 PR 为此进行了兼容性修复。