

PR #52021 完整报告

vllm-project/vllm

[Bugfix] Preserve Anthropic disable_parallel_tool_use

合并时间: 2026-08-13 20:25

原文链接: <http://prhub.com.cn/vllm-project/vllm/pull/52021>

执行摘要

- 一句话: 修复 Anthropic disable_parallel_tool_use 被静默丢弃的问题
- 推荐动作: 值得快速阅读: 改动虽小, 但展示了协议兼容层中“保留源字段 + 取反映射”的典型做法。建议关注 _convert_tool_choice 中先设置 parallel_tool_calls 再按 type 分支的顺序, 以及 None 缺省值如何在不改动其他路径的前提下维持向后兼容。

功能与动机

PR body 明确说明: Anthropic 的 disable_parallel_tool_use 字段被静默丢弃, 转换后的 OpenAI 请求总是 parallel_tool_calls=True。reproducer 表明传入 'disable_parallel_tool_use': True 时, 转换后 parallel_tool_calls 仍为 True, 与用户意图相反, 因此需要保留并映射到 OpenAI 的对应字段。

实现拆解

1. 协议模型补字段: 在 vllm/entrypoints/anthropic/protocol.py 的 AnthropicToolChoice 类中新增 disable_parallel_tool_use: bool | None = None, 使请求体中的该字段能被 Pydantic 解析并保留。
2. 转换逻辑映射: 在 vllm/entrypoints/anthropic/serving.py 的 _convert_tool_choice 类方法中, 在按 tool_choice.type 分支之前设置 req.parallel_tool_calls = not anthropic_request.tool_choice.disable_parallel_tool_use, 利用两个协议字段语义相反的关系完成取反映射。
3. 验证与配套: PR 未新增测试文件, 通过 tests/entrypoints/anthropic/test_anthropic_messages_conversion.py 全部 51 个用例验证; CI 在 Buildkite 上经过 2 次 retry 后通过。

关键文件:

- vllm/entrypoints/anthropic/serving.py (模块 请求转换; 类别 source; 类型 core-logic; 符号 _convert_tool_choice): 核心转换逻辑所在: _convert_tool_choice 中新增 parallel_tool_calls 取反映射, 直接修复字段被丢弃的问题。
- vllm/entrypoints/anthropic/protocol.py (模块 协议模型; 类别 source; 类型 core-logic; 符号 AnthropicToolChoice): 协议模型定义处: 为 AnthropicToolChoice 补充 disable_parallel_tool_use 字段, 使该入参不再被忽略。

关键符号: _convert_tool_choice, AnthropicToolChoice

关键源码片段

vllm/entrypoints/anthropic/serving.py

核心转换逻辑所在：_convert_tool_choice 中新增 parallel_tool_calls 取反映射，直接修复字段被丢弃的问题。

```
@classmethod
def _convert_tool_choice(
    cls,
    anthropic_request: AnthropicMessagesRequest | AnthropicCountTokensRequest,
    req: ChatCompletionRequest,
) -> None:
    """Convert Anthropic tool_choice to OpenAI format"""
    if anthropic_request.tool_choice is None:
        req.tool_choice = None
        return

    # 关键修复：此前该字段被静默丢弃， parallel_tool_calls 恒为 True
    # Anthropic 的 disable_parallel_tool_use 与 OpenAI 的 parallel_tool_calls 语义相反，取反映射
    # 缺省值 None 取反后为 True，与历史行为保持一致
    req.parallel_tool_calls = (
        not anthropic_request.tool_choice.disable_parallel_tool_use
    )

    tool_choice_type = anthropic_request.tool_choice.type
    if tool_choice_type == "auto":
        req.tool_choice = "auto"
    elif tool_choice_type == "any":
        req.tool_choice = "required"
    elif tool_choice_type == "none":
        req.tool_choice = "none"
    elif tool_choice_type == "tool":
        req.tool_choice = ChatCompletionNamedToolChoiceParam.model_validate(
            {
                "type": "function",
                "function": {"name": anthropic_request.tool_choice.name},
            }
        )
```

vllm/entrypoints/anthropic/protocol.py

协议模型定义处：为 AnthropicToolChoice 补充 disable_parallel_tool_use 字段，使该入参不再被忽略。

```
class AnthropicToolChoice(BaseModel):
    """Tool Choice definition"""

    type: Literal["auto", "any", "tool", "none"]
    name: str | None = None

    # 新增字段：Anthropic 允许在 tool_choice 中显式禁用并行工具调用
    # 若不声明此字段，Pydantic 会忽略请求体中的该键，导致后续转换丢失信息
```

```
disable_parallel_tool_use: bool | None = None
```

```
@model_validator(mode="after")
def validate_name_required_for_tool(self) -> "AnthropicToolChoice":
    if self.type == "tool" and not self.name:
        raise ValueError("tool_choice.name is required when type is 'tool'")
    return self
```

评论区精华

review 过程非常简短：维护者 yewentao256 给出 APPROVED，评论为 'LGTM, thanks for the work!'. Claude bot 说明 fork 的 PR 默认不进行自动 review，可由维护者手动触发。没有其他实质讨论；缺省 `None` 经 `not` 取反后为 `True` 与旧行为一致的兼容性决策未在评论中显式提及，但属于该实现的关键设计点。

- 暂无高价值评论线程

风险与影响

- 风险：主要风险点： 1) `disable_parallel_tool_use` 缺省为 `None`，`not None` 为 `True`，保证了向后兼容，但若未来 Anthropic 改变字段默认值或引入三态枚举，此处取反逻辑可能产生语义偏差； 2) `tool_choice` 为 `None` 时方法提前 `return`，不设置 `parallel_tool_calls`，此时依赖 `ChatCompletionRequest` 自身的默认值（通常为 `True`），该路径行为未变； 3) 未新增专门测试，若后续重构 `_convert_tool_choice` 分支顺序可能破坏映射。
- 影响：影响面限于 Anthropic 兼容层的请求转换（`/v1/messages` 端点）。用户现在可以通过 `disable_parallel_tool_use` 显式控制是否并行调用工具，对既有未传该字段的请求行为完全不变。对团队而言，该改动提供了一个简洁的协议字段取反映射示例，后续 Anthropic 其他字段的转换可参考此模式。
- 风险标记：缺省值依赖隐式取反，未新增测试文件

关联脉络

- 暂无明显关联 PR