PR #45048 完整报告

vllm-project/vllm

[Bugfix] GPT-OSS Autodrop reasoning in Response API and cleanup

合并时间：2026-06-23 21:36

原文链接：http://prhub.com.cn/vllm-project/vllm/pull/45048

执行摘要

- 一句话：修复 Responses API 多轮对话中推理消息未正确丢弃
- 推荐动作：建议关注 render_for_completion 的改动与新增测试，理解如何跨 API 共享逻辑并精确控制库默认行为。该 PR 是一个典型的 bugfix + cleanup 示例，展示了修复多路径不一致问题的策略。

功能与动机

Harmony prompts should drop reasoning from completed turns, but keep reasoning from the current in-progress turn so tool-calling can continue correctly. Responses API previously only relied on Harmony's default filter, which only drops analysis messages when the conversation ends with an assistant final message, leaving stale reasoning from all previous turns when the prompt ends mid-turn on a tool call or tool result. (PR body)

实现拆解

1. 移动自动丢弃逻辑：从 parse_chat_inputs_to_harmony_messages 中移除 auto_drop_analysis_messages 调用，改为在共享的 render_for_completion 函数中调用该函数，使两个 API 路径在渲染 input tokens 前都丢弃已完成轮次的 analysis 消息。
2. 禁用 Harmony 库的默认丢弃：在 render_for_completion 中向 render_conversation_for_completion 传递 RenderConversationConfig(auto_drop_analysis=False)，防止 Harmony 编码器重复丢弃。
3. 简化 Responses 上下文构建：删除 _construct_input_messages_with_harmony 中一个复杂的 slice-delete-reappend 循环（实际是 no-op），避免混淆。
4. 修复构造方法调用：在 responses/harmony.py 中将 Message.from_author_and_contents（不存在的方法）改为 Message(author=..., content=...)。
5. 更新测试：新增 test_completed_turns_drop_reasoning 验证多轮推理正确丢弃；调整 test_serving_chat.py 中多轮测试期望，反映当前轮分析消息被保留。

关键文件：

- tests/entrypoints/openai/parser/test_harmony_render_parity.py（模块 渲染测试；类别 test；类型 test-coverage；符号 test_completed_turns_drop_reasoning）：新增

test_completed_turns_drop_reasoning 测试用例，验证多轮对话中推理消息正确丢弃，是此修复的关键验证。

- vllm/entrypoints/openai/responses/serving.py（模块 服务层；类别 source；类型 core-logic）：主要源码变更，删除无操作的消息循环，简化 _construct_input_messages_with_harmony。
- tests/entrypoints/openai/chat_completion/test_serving_chat.py（模块 聊天测试；类别 test；类型 test-coverage）：更新现有测试以反映行为变化：第二轮的输入消息现在应包含当前轮分析消息而非被丢弃。
- vllm/entrypoints/openai/parser/harmony_utils.py（模块 解析器；类别 source；类型 core-logic；符号 render_for_completion, parse_chat_inputs_to_harmony_messages）：核心逻辑：将 auto_drop_analysis_messages 移到 render_for_completion 并禁用 Harmony 默认丢弃。
- vllm/entrypoints/openai/responses/harmony.py（模块 服务层；类别 source；类型 bugfix；符号 _parse_harmony_format_message）：修复不存在的方法调用 bug。

关键符号：auto_drop_analysis_messages, render_for_completion, _construct_input_messages_with_harmony, _parse_harmony_format_message, parse_chat_inputs_to_harmony_messages, test_completed_turns_drop_reasoning

关键源码片段

tests/entrypoints/openai/parser/test_harmony_render_parity.py

新增 test_completed_turns_drop_reasoning 测试用例，验证多轮对话中推理消息正确丢弃，是此修复的关键验证。

```
def test_completed_turns_drop_reasoning(self):
    """验证已完成轮次的推理被丢弃，而当前正在进行的推理保留。"""
    first_reasoning = "FIRST_TURN_REASONING"
    second_reasoning = "SECOND_TURN_REASONING"

    # Chat Completions 路径
    chat_msgs = self._build_chat_msgs(first_reasoning, second_reasoning)
    chat_tokens = render_for_completion([_system()] + chat_msgs)
    rendered = get_encoding().decode(chat_tokens)
    # 第一轮推理应被丢弃，第二轮推理应保留
    assert first_reasoning not in rendered
    assert second_reasoning in rendered

    # Responses 路径
    resp_msgs = self._build_response_msgs(first_reasoning, second_reasoning)
    resp_tokens = render_for_completion([_system()] + resp_msgs)
    rendered_resp = get_encoding().decode(resp_tokens)
    assert first_reasoning not in rendered_resp
    assert second_reasoning in rendered_resp
```

vllm/entrypoints/openai/parser/harmony_utils.py

核心逻辑：将 auto_drop_analysis_messages 移到 render_for_completion 并禁用 Harmony 默认丢弃。

```
def render_for_completion(messages: list[Message]) -> list[int]:
    # 先应用 vLLM 的自动丢弃逻辑（丢弃已完成轮次的分析消息，保留当前轮）
    messages = auto_drop_analysis_messages(messages)
    conversation = Conversation.from_messages(messages)
    # 禁用 Harmony 编码器自带的自动丢弃，因为我们已经处理
    token_ids = get_encoding().render_conversation_for_completion(
        conversation,
        Role.ASSISTANT,
        config=RenderConversationConfig(auto_drop_analysis=False),
    )
    return token_ids
```

评论区精华

- bug 发现：bbrowning 在 review 中指出 responses/harmony.py 中调用了不存在的方法 Message.from_author_and_contents，作者确认并称是通过 Claude 发现的。
- 测试适配：bbrowning 报告 CI 捕获 test_serving_chat.py 中的失败。作者解释因自动丢弃逻辑移动导致消息序列变化，并更新测试期望以匹配新行为。
  - 不存在的方法调用 (correctness)：已改用正确的 Message(author=..., content=...) 构造修复。
  - CI 测试期望不匹配 (testing)：已更新测试以匹配新行为，测试恢复通过。

风险与影响

- 风险：自动丢弃逻辑移至 render_for_completion 后，所有调用该函数的路径都会应用丢弃，可能影响其他未预期场景，但已有测试覆盖。禁用 Harmony 默认丢弃依赖 RenderConversationConfig 参数，若未来升级 openai-harmony 版本导致该参数行为变化，需重新验证。移除的循环虽注释为 no-op，但考虑其编写时可能另有考虑，不过已有补充测试保障。重构仅针对 GPT-OSS Harmony 路径，其他模型不受影响。
- 影响：直接影响使用 GPT-OSS 模型和 Responses API 的用户，修复了多轮工具调用中推理消息残留导致模型行为异常的问题；Chat Completions 路径行为不变；清理冗余代码降低了维护成本；统一推理丢弃策略使两条 API 路径行为一致。
- 风险标记：核心路径变更，依赖库默认行为关闭，测试覆盖新增

关联脉络

- PR #46030 [Refactor] Responses API parser state into conversation context: 同属 Responses API 重构与修复，改动相同的 responses/serving.py 模块。
- PR #44105 [BugFix] Omit empty tool_calls from OpenAI chat responses: 同为前端工具调用相关 bug 修复，涉及聊天消息处理。
- PR #46441 fix gpt_oss pp>1 with ep: 同为 GPT-OSS 模型修复，属同一功能线。