PR #43761 完整报告

vllm-project/vllm

[Frontend]Responses API supports chat_template_kwargs

合并时间：2026-05-29 15:58

原文链接：http://prhub.com.cn/vllm-project/vllm/pull/43761

执行摘要

- 一句话：Responses API 支持 chat_template_kwargs 传递
- 推荐动作：修复明确、风险低、影响集中在特定模型配置场景。建议阅读以了解 Responses API 与 chat_template_kwargs 的交互逻辑。无需精读。

功能与动机

PR #42272 引入了 chat_template_kwargs 支持，但在 Responses API 的 parser 调用中遗漏了该参数，导致 DeepSeek-V3.2 等模型的 分割失效。PR body 明确指出 'The parser did not split the model output based on because chat_template_kwargs was not properly passed to the reasoning parser.'

实现拆解

1. 在 vllm/entrypoints/openai/responses/serving.py 的 _make_response_output_items 方法中，实例化 parser 时新增 chat_template_kwargs=self._effective_chat_template_kwargs(request) 参数传递。
2. 在 _process_simple_streaming_events 方法中同样修改 parser 创建逻辑，确保流式场景也能获取到 chat_template_kwargs。
3. 使用已有的 self._effective_chat_template_kwargs(request) 方法获取配置，无新增依赖。

关键文件：

- vllm/entrypoints/openai/responses/serving.py（模块 前端服务；类别 source；类型 core-logic；符号 _make_response_output_items, _process_simple_streaming_events）：唯一修改的文件，包含两处 parser 初始化逻辑变更，是修复核心。

关键符号：_make_response_output_items, _process_simple_streaming_events

关键源码片段

vllm/entrypoints/openai/responses/serving.py

唯一修改的文件，包含两处 parser 初始化逻辑变更，是修复核心。

```
# vllm/entrypoints/openai/responses/serving.py

# 非流式场景：_make_response_output_items 方法中
```

```
if self.parser:
    # 获取请求对应的 chat_template_kwargs，如 reasoning_effort
    chat_template_kwargs = self._effective_chat_template_kwargs(request)
    parser = self.parser(
        tokenizer, request.tools, chat_template_kwargs=chat_template_kwargs
        # ^^ 之前遗漏了 chat_template_kwargs，导致 reasoning parser
        # 无法获知用户配置的 reasoning_effort 等参数
    )
    return parser.extract_response_outputs(...)

# 流式场景：_process_simple_streaming_events 方法中
parser = (
    self.parser(
        tokenizer,
        request.tools,
        chat_template_kwargs=self._effective_chat_template_kwargs(request),
    )
    if self.parser
    else None
)
```

评论区精华

无 review 评论。仅有一位 reviewer 审批通过（LGTM, thanks!）。

- 暂无高价值评论线程

风险与影响

- 风险：变更仅涉及两处 parser 实例化行，逻辑简单且复用已有方法，回滚风险极低。若 _effective_chat_template_kwargs 返回异常值，可能影响所有使用 parser 的 Responses API 调用，但该风险与主 PR 一致，本变更未引入新风险。
- 影响：影响使用 Responses API 且依赖 chat_template_kwargs（如 reasoning_effort、enable_thinking）的模型（如 DeepSeek-V3.2）。修复后，非流式与流式响应均能正确解析推理标记。测试覆盖未显式增加，但 PR body 提供了手动验证方式。
- 风险标记：缺少测试覆盖

关联脉络

- PR #42272 支持 chat_template_kwargs 的初始引入：本 PR 是对 #42272 的后续修复，确保 Responses API 正确应用已引入的 chat_template_kwargs 功能。