PR #42272 完整报告

vllm-project/vllm

[Frontend]Responses API supports chat_template_kwargs

合并时间：2026-05-11 19:59

原文链接：http://prhub.com.cn/vllm-project/vllm/pull/42272

执行摘要

- 一句话：Responses API 支持 chat_template_kwargs 传递
- 推荐动作：该 PR 值得精读，展示了如何为 Responses API 增加参数透传能力的简洁实现，尤其在默认值与请求值合并的设计上值得借鉴。

功能与动机

允许用户通过 Responses API 的请求或服务级默认值向聊天模板渲染器传递额外关键字参数，例如 enable_thinking，以支持模板自定义。

实现拆解

- vllm/entrypoints/openai/responses/protocol.py 的 ResponsesRequest 模型新增 chat_template_kwargs: dict[str, Any] | None 字段，并修改 build_chat_params 将其传入 merge_kwargs 以保留请求级参数。
- vllm/entrypoints/openai/responses/serving.py 的 OpenAIServingResponses.__init__ 新增 default_chat_template_kwargs 参数并存储为 self.chat_template_kwargs；修改 _effective_chat_template_kwargs 通过 with_defaults 合并默认值与请求值；更新 _make_request 和 _render_next_turn 将合并后的 chat_template_kwargs 传递给 preprocess_chat 的 default_template_kwargs。
- vllm/entrypoints/openai/generate/api_router.py 的 init_generate_state 将 args.default_chat_template_kwargs 传递给 OpenAIServingResponses。
- tests/entrypoints/openai/responses/test_function_call.py 的 test_function_calling_with_streaming_types 新增 enable_thinking 参数并传入 extra_body。

关键文件：

- vllm/entrypoints/openai/responses/serving.py（模块 入口层；类别 source；类型 core-logic）：核心实现：新增 default_chat_template_kwargs 初始化，修改 _effective_chat_template_kwargs 合并默认值与请求值，更新渲染调用传递参数。
- vllm/entrypoints/openai/responses/protocol.py（模块 入口层；类别 source；类型 core-logic）：定义 chat_template_kwargs 请求字段，修改 build_chat_params 将其传入 merge_kwargs。

- vllm/entrypoints/openai/generate/api_router.py（模块 入口层；类别 source；类型 entrypoint）：将命令行参数 default_chat_template_kwargs 传递给 OpenAIServingResponses。
- tests/entrypoints/openai/responses/test_function_call.py（模块 功能调用；类别 test；类型 test-coverage）：扩充测试：增加 enable_thinking 参数化，传递 chat_template_kwargs 到请求中。

关键符号：OpenAIServingResponses.init, OpenAIServingResponses._effective_chat_template_kwargs, OpenAIServingResponses._make_request, OpenAIServingResponses._render_next_turn, ResponsesRequest.build_chat_params

关键源码片段

vllm/entrypoints/openai/responses/serving.py

核心实现：新增 default_chat_template_kwargs 初始化，修改 _effective_chat_template_kwargs 合并默认值与请求值，更新渲染调用传递参数。

```
# vllm/entrypoints/openai/responses/serving.py
class OpenAIServingResponses(OpenAIServing):
    def __init__(
        self,
        # ... existing params ...
        default_chat_template_kwargs: dict[str, Any] | None = None, # 新增：服务级默认参数
    ) -> None:
        super().__init__(...)
        self.chat_template_kwargs = default_chat_template_kwargs or {} # 存储默认值
        # ...

    def _effective_chat_template_kwargs(
        self, request: ResponsesRequest
    ) -> dict[str, Any]:
        # 合并请求级与默认值：请求级优先级更高
        return (
            request.build_chat_params(
                self.chat_template,
                self.chat_template_content_format,
            )
            .with_defaults(self.chat_template_kwargs) # 以默认值作为 fallback
            .chat_template_kwargs
        )

    async def _make_request(self, request, ...):
        chat_template_kwargs = self._effective_chat_template_kwargs(request)
        _, engine_inputs = await self.openai_serving_render.preprocess_chat(
            request,
            messages,
            default_template=self.chat_template,
            default_template_content_format=self.chat_template_content_format,
```

```
            default_template_kwargs=chat_template_kwargs, # 传递合并后的参数
            # ...
        )
```

vllm/entrypoints/openai/responses/protocol.py

定义 chat_template_kwargs 请求字段，修改 build_chat_params 将其传入 merge_kwargs。

```
# vllm/entrypoints/openai/responses/protocol.py
# 在 ResponsesExtraParams 区域内新增字段
chat_template_kwargs: dict[str, Any] | None = Field(
    default=None,
    description=(
        "Additional keyword args to pass to the chat template renderer. "
        "Will be accessible by the template."
    ),
)

# build_chat_params 中修改 merge_kwargs 的第一个参数
return ChatParams(
    chat_template=default_template,
    chat_template_content_format=default_template_content_format,
    chat_template_kwargs=merge_kwargs(
        self.chat_template_kwargs, # 原来是 {}，现在使用请求级参数
        dict(
            add_generation_prompt=not continue_final,
            continue_final_message=continue_final,
            reasoning_effort=None if reasoning is None else reasoning.effort,
        ),
    ),
    media_io_kwargs=self.media_io_kwargs,
)
```

评论区精华

gemini-code-assist[bot] 评论指出 default_chat_template_kwargs 未正确传递给 _make_request 和 _render_next_turn 中的 preprocess_chat，会导致服务级默认值被忽略。但最终提交版本已通过 _effective_chat_template_kwargs 获取合并后的值并作为 default_template_kwargs 传入，该问题已解决。

- 默认参数未传递给 preprocess_chat (correctness): 已修复（在最终代码中通过 _effective_chat_template_kwargs 获取合并值并传递）。

风险与影响

- 风险：无显著技术风险。变更仅涉及参数传递和合并逻辑，未修改核心推理路径。需注意 with_defaults 方法的正确性，但已有测试覆盖。
- 影响：对用户：新增 chat_template_kwargs 请求字段和服务级 default_chat_template_kwargs 配置，提供更灵活的模板自定义能力。对系统：无性能或

兼容性影响，为纯参数扩展。

- 风险标记：暂无

关联脉络

- 暂无明显关联 PR