PR #42396 完整报告

vllm-project/vllm

[Feature] Add structured output and effort support to Anthropic Messages API

合并时间：2026-05-28 20:06

原文链接：http://prhub.com.cn/vllm-project/vllm/pull/42396

执行摘要

- 一句话：为 Anthropic API 添加结构化输出和 effort 参数支持
- 推荐动作：该 PR 可以精读以了解如何扩展 Anthropic API 入口，特别关注 _handle_output_config 的转换模式。设计决策方面，注意 effort 被放在 output_config 内而非顶层，与官方规范略有偏离，但保持了内部一致性。测试用例提供了良好的参考。

功能与动机

用户需求（关联 Issue #8907）希望在 vLLM 的 Anthropic 兼容 API 中支持结构化输出和 effort 参数，以更好地控制推理行为和输出格式，参考 Anthropic 官方文档（https://platform.claude.com/docs/en/build-with-claude/structured-outputs 和 https://platform.claude.com/docs/en/build-with-claude/effort）。

实现拆解

1. 协议模型扩展（vllm/entrypoints/anthropic/protocol.py）：新增 AnthropicJsonOutputFormat（包含 json_schema 和 type）和 AnthropicOutputConfig（包含 effort 和 format）两个 BaseModel，并在 AnthropicMessagesRequest 中添加可选字段 output_config: AnthropicOutputConfig | None。
2. 请求转换逻辑（vllm/entrypoints/anthropic/serving.py）：在 _convert_anthropic_to_openai_request 中调用新方法 _handle_output_config。该方法提取 output_config.format 和 output_config.effort，分别映射到 OpenAI 的 ResponseFormat（含 JsonSchemaResponseFormat）和 reasoning_effort；对 AnthropicCountTokensRequest 类型直接返回。
3. 端到端测试（tests/entrypoints/anthropic/test_messages.py）：新增 test_anthropic_structured_output，通过 Anthropic 客户端发送含 output_config 的请求，验证返回 JSON 包含所需字段。

关键文件：

- vllm/entrypoints/anthropic/serving.py（模块 API 入口；类别 source；类型 core-logic；符号 _handle_output_config）：核心业务逻辑，新增 _handle_output_config 方法转换 output_config 和 effort 到 OpenAI 请求格式。
- vllm/entrypoints/anthropic/protocol.py（模块 数据模型；类别 source；类型 core-logic；符号 AnthropicJsonOutputFormat, AnthropicOutputConfig）：新增

AnthropicJsonOutputFormat 和 AnthropicOutputConfig 模型，扩展 AnthropicMessagesRequest 包含 output_config 字段。

- tests/entrypoints/anthropic/test_messages.py（模块 测试；类别 test；类型 test-coverage；符号 test_anthropic_structured_output）：端到端测试验证结构化输出功能，确保配置正确传递并产生预期输出。

关键符号：_handle_output_config, test_anthropic_structured_output

关键源码片段

vllm/entrypoints/anthropic/serving.py

核心业务逻辑，新增 _handle_output_config 方法转换 output_config 和 effort 到 OpenAI 请求格式。

```
# vllm/entrypoints/anthropic/serving.py (partial)

@classmethod
def _handle_output_config(
    cls,
    req: ChatCompletionRequest,
    anthropic_request: AnthropicMessagesRequest | AnthropicCountTokensRequest,
) -> None:
    """处理输出配置，如输出格式（JSON Schema）和推理努力程度"""
    # 计数 token 请求不需要输出配置
    if isinstance(anthropic_request, AnthropicCountTokensRequest):
        return

    output_config: AnthropicOutputConfig | None = anthropic_request.output_config
    # 检查是否有 JSON Schema 格式配置
    if output_config and output_config.format and output_config.format.json_schema:
        req.response_format = ResponseFormat(
            type=output_config.format.type,
            json_schema=JsonSchemaResponseFormat(
                schema=output_config.format.json_schema,
                name=output_config.format.type, # 使用 "json_schema" 作为默认名称
            ),
        )
    # 检查是否有推理努力程度配置
    if output_config and output_config.effort is not None:
        req.reasoning_effort = output_config.effort
```

vllm/entrypoints/anthropic/protocol.py

新增 AnthropicJsonOutputFormat 和 AnthropicOutputConfig 模型，扩展 AnthropicMessagesRequest 包含 output_config 字段。

```
# vllm/entrypoints/anthropic/protocol.py (partial)

class AnthropicJsonOutputFormat(BaseModel):
    """JSON 输出格式配置"""
```

```
    json_schema: dict[str, Any] | None = Field(default=None, alias="schema")
    type: Literal["json_schema"] = "json_schema"


class AnthropicOutputConfig(BaseModel):
    """模型输出配置，包含输出格式和推理努力程度"""
    # 注意：官方 Anthropic API 仅支持 low/medium/high，此处额外包含 xhigh/max
    effort: Literal["low", "medium", "high", "xhigh", "max"] | None = None
    format: AnthropicJsonOutputFormat | None = None


class AnthropicMessagesRequest(BaseModel):
    # ... 其他字段 ...
    output_config: AnthropicOutputConfig | None = None # 新增字段
```

评论区精华

- effort 参数设计争议：gemini-code-assist[bot] 指出 Anthropic 官方 API 中 effort 应为顶层参数而非嵌套于 output_config，且有效值仅为 low/medium/high。作者未采纳建议，合并者 DarkLight1337 批准合并，设计保持不变。
- 功能范围讨论：DarkLight1337 询问 effort 与结构化输出是否属同一功能，作者解释是一并实现的，DarkLight1337 要求更新 PR 标题描述，作者照做。
- 空检查健壮性：gemini-code-assist[bot] 建议使用 `is not None` 而非 `if json_schema:` 以避免空字典误跳过，此建议未被采纳。
  - effort 参数置于 output_config 内部与官方规范不符 (design): 作者未修改，合并者 DarkLight1337 批准，按现有设计合并。
  - reasoning effort 与结构化输出是否为同一功能 (question): 作者解释是一并实现的，DarkLight1337 要求更新 PR 标题描述，作者照做。
  - json_schema 空字典检查应使用 is not None (correctness): 未被采纳。

风险与影响

- 风险：
  - 兼容性风险：effort 支持 xhigh 和 max，但 Anthropic 官方仅定义 low、medium、high，使用非标准值可能导致未来行为不确定。
  - 转换逻辑风险：output_config.format.json_schema 使用 truthy 检查，空字典 {} 会被跳过，可能使用户意图未生效。
  - 交互风险：output_config 与 tool_choice/tools 的交互未经测试，同时使用时可能出现冲突。
  - 影响范围：仅 Anthropic API 入口，不影响核心引擎或其他 API，风险可控。
- 影响：
  - 用户：使用 vLLM Anthropic API 的用户现在可以指定 output_config 参数，包括 format（JSON Schema）和 effort（推理努力程度），扩展了功能性。

- 系统：对核心引擎无影响，仅在 API 入口层进行参数映射，性能开销可忽略。
- 团队：为后续 Anthropic API 的扩展（如其他输出格式）奠定了基础。
- 风险标记：非标准 *effort* 值，空 json_schema 潜在问题，与 tool calls 交互未测试

关联脉络

- 暂无明显关联 PR