Deepseek企业级Agent项目开发实战

Part 3. Ollama REST API - api/chat 接口详解

Ollama 服务启动后会提供一系列原生 `REST API` 端点。通过这些 `Endpoints` 可以在代码环境下与 `ollama` 启动的大模型进行交互、管理模型和获取相关信息。其中两个 `endpoint` 是最重要的，分别是：

- **POST /api/generate**
- **POST /api/chat**

其他端点情况：

- POST /api/create
- POST /api/tags
- POST /api/show
- POST /api/copy
- DELETE /api/delete
- POST /api/pull
- POST /api/push
- POST /api/embed
- GET /api/ps

---

1. /api/chat 接口参数概览

该接口使用提供的模型在聊天中生成下一条消息。与 `/api/generate` 的参数基本一致，但是在请求的参数上会根据聊天场景进行调整。主要调整的是：

- 不再使用 `prompt` 参数，而是使用 `messages` 参数。
- 新增了 `tools` 参数，用于支持工具调用。

其可以使用的具体参数如下所示，

常规参数

参数名	类型	描述
model	*(必需)*	模型名称。
messages	*(必需)*	聊天的消息，用于保持聊天记忆。
tools	*(可选)*	JSON 中的工具列表，供模型使用（如果支持）。

消息对象字段

字段名	描述
role	消息的角色，可以是 `system`、`user`、`assistant` 或 `tool`。
content	消息的内容。
images	*(可选)* 要在消息中包含的图像列表（适用于多模态模型，如 llava）。
tool_calls	*(可选)* 模型希望使用的 JSON 中的工具列表。

高级参数（可选）

参数名	描述
format	返回响应的格式。格式可以是 `json` 或 JSON 模式。
options	文档中列出的其他模型参数，例如 `temperature`。
stream	如果为 `false`，响应将作为单个响应对象返回，而不是对象流。
keep_alive	控制模型在请求后保持加载的时间（默认：5分钟）。

其中，Options参数说明：

参数名	描述	值类型	示例用法
mirostat	启用 Mirostat 采样以控制困惑度。（默认：0，0 = 禁用，1 = Mirostat，2 = Mirostat 2.0）	int	mirostat 0
mirostat_eta	影响算法对生成文本反馈的响应速度。较低的学习率会导致调整较慢，而较高的学习率会使算法更具响应性。（默认：0.1）	float	mirostat_eta 0.1
mirostat_tau	控制输出的连贯性和多样性之间的平衡。较低的值会导致更集中和连贯的文本。（默认：5.0）	float	mirostat_tau 5.0
num_ctx	设置用于生成下一个标记的上下文窗口大小。（默认：2048），影响的是模型可以一次记住的最大 token 数量。	int	num_ctx 4096
repeat_last_n	设置模型回溯的范围以防止重复。（默认：64，0 = 禁用，-1 = num_ctx）	int	repeat_last_n 64
repeat_penalty	设置惩罚重复的强度。较高的值（例如 1.5）会更强烈地惩罚重复，而较低的值（例如 0.9）会更宽松。（默认：1.1）	float	repeat_penalty 1.1
temperature	模型的温度。增加温度会使模型的回答更具创造性。（默认：0.8）	float	temperature 0.7
seed	设置用于生成的随机数种子。将其设置为特定数字将使模型对相同提示生成相同的文本。（默认：0）	int	seed 42
stop	设置使用的停止序列。当遇到此模式时，LLM 将停止生成文本并返回。可以通过在 modelfile 中指定多个单独的停止参数来设置多个停止模式。	string	stop "AI assistant:"
num_predict	生成文本时要预测的最大标记数。（默认：-1，无限生成）,影响模型最大可以生成的 token 数量。	int	num_predict 42
top_k	降低生成无意义文本的概率。较高的值（例如 100）会给出更多样化的答案，而较低的值（例如 10）会更保守。（默认：40）	int	top_k 40
top_p	与 top-k 一起工作。较高的值（例如 0.95）会导致更具多样性的文本，而较低的值（例如 0.5）会生成更集中和保守的文本。（默认：0.9）	float	top_p 0.9
min_p	top_p 的替代方案，旨在确保质量和多样性之间的平衡。参数 p 表示考虑标记的最小概率，相对于最可能标记的概率。例如，p=0.05 时，最可能的标记概率为 0.9，值小于 0.045 的 logits 会被过滤掉。（默认：0.0）	float	min_p 0.05

2. requests 调用方法

`/api/chat` 依然还是可以 `requests` 库进行调用。如下所示：

```
import requests
import json
```

```python
# 设置 API 端点
chat_url = "http://192.168.110.131:11434/api/chat"    # 这里需要根据实际情况进行修改

# 示例数据
chat_payload = {
    "model": "deepseek-r1:32b",    # 这里需要根据实际情况进行修改
    "messages": [
        {
            "role": "user",  # 消息角色，用户发送的消息
            "content": "请生成一个关于人工智能的简短介绍。"  # 用户的消息内容
        }
    ],
    "tools": [],  # 如果有工具可以在这里添加
    "stream": False,  # 默认使用的是True，如果设置为False，则返回的是一个完整的响应，而不是一个流式响应
}

# 调用聊天接口
response_chat = requests.post(chat_url, json=chat_payload)
if response_chat.status_code == 200:
    chat_response = response_chat.json()
    print("生成响应:", json.dumps(chat_response, ensure_ascii=False, indent=2))
else:
    print("生成请求失败:", response_chat.status_code, response_chat.text)
```


```
生成响应: {
  "model": "deepseek-r1:32b",
  "created_at": "2025-02-13T11:18:25.515272041Z",
  "message": {
    "role": "assistant",
    "content": "<think>\n好，我现在需要帮用户生成一个关于人工智能的简短介绍。首先，我得理解用户的需求是什么。他们可能对AI不太了解，想要一个简洁明了的概述。\n\n我应该从基础开始讲起，比如定义。人工智能是模拟人类智能的技术，这点很重要。然后，我可以提到主要应用领域，比如机器学习、自然语言处理和计算机视觉，这样可以让介绍更有针对性。\n\n接下来，我需要说明AI的应用范围，比如在医疗、金融和交通中的作用，这样用户能明白它的实际价值。同时，也不能忽视伦理和社会影响，这部分也是大家关心的点。\n\n还要提到当前的发展阶段，强调它是一个快速发展的领域，这样能展示出未来的潜力。最后，保持整体内容简明扼要，适合快速阅读。\n\n可能用户需要这个介绍用于学习、演讲或者作为参考资料。所以信息要准确，结构清晰，重点突出。我要确保涵盖主要方面，同时不显得冗长。这样用户就能得到一个全面又简洁的人工智能简介了。\n</think>\n\n人工智能（Artificial Intelligence, AI）是指通过模拟人类智能的技术，使计算机系统能够执行如学习、推理、问题解决和自然语言处理等任务。AI技术广泛应用于医疗、金融、交通等领域，帮助提高效率并解决复杂问题。随着算法的进步和数据的增加，人工智能正逐步改变我们的生活方式和社会结构。"
  },
  "done_reason": "stop",
  "done": true,
  "total_duration": 22607165543,
  "load_duration": 9711125924,
  "prompt_eval_count": 13,
  "prompt_eval_duration": 3497000000,
  "eval_count": 286,
  "eval_duration": 9396000000
}
```


返回的响应中包含以下参数，其对应的描述如下：

响应参数

参数名	描述
total_duration	单次响应花费的总时间
load_duration	加载模型花费的时间
prompt_eval_count	提示中的token数
prompt_eval_duration	评估提示所花费的时间（以纳秒为单位）
eval_count	响应中的token数
eval_duration	生成响应的时间（以纳秒为单位）
context	在此响应中使用的对话的编码，可以在下一个请求中发送以保持对话记忆
response	空响应是流的，如果未流式传输，则将包含完整的响应

重点关注以下几个参数：

- **message**

在 `/chat` 接口中，返回的模型响应结果存放在 `message` 中， 同样对于 `DeepSeek-R1` 模型，`response` 字段中包含 标签和正常文本， 标签用于表示模型的思考过程或内部推理，而正常的文本则是模型生成的实际输出内容。注意：非推理类模型的返回结果中没有标识。

```
chat_response["message"]['content']
```

```
'<think>\n好，我现在需要帮用户生成一个关于人工智能的简短介绍。首先，我得理解用户的需求是什么。他们可能对AI不太了解，想要一个简洁明了的概述。\n\n我应该从基础开始讲起，比如定义。人工智能是模拟人类智能的技术，这点很重要。然后，我可以提到主要应用领域，比如机器学习、自然语言处理和计算机视觉，这样可以让介绍更有针对性。\n\n接下来，我需要说明AI的应用范围，比如在医疗、金融和交通中的作用，这样用户能明白它的实际价值。同时，也不能忽视伦理和社会影响，这部分也是大家关心的点。\n\n还要提到当前的发展阶段，强调它是一个快速发展的领域，这样能展示出未来的潜力。最后，保持整体内容简明扼要，适合快速阅读。\n\n可能用户需要这个介绍用于学习、演讲或者作为参考资料。所以信息要准确，结构清晰，重点突出。我要确保涵盖主要方面，同时不显得冗长。这样用户就能得到一个全面又简洁的人工智能简介了。\n</think>\n\n人工智能（Artificial Intelligence, AI）是指通过模拟人类智能的技术，使计算机系统能够执行如学习、推理、问题解决和自然语言处理等任务。AI技术广泛应用于医疗、金融、交通等领域，帮助提高效率并解决复杂问题。随着算法的进步和数据的增加，人工智能正逐步改变我们的生活方式和社会结构。'
```

可以通过简单的字符串操作来分离 标签中的思考内容和正常的文本内容，代码如下：

```
# 提取 <think> 标签中的内容
think_start = chat_response["message"]['content'].find("<think>")
think_end = chat_response["message"]['content'].find("</think>")

if think_start != -1 and think_end != -1:
    think_content = chat_response["message"]['content'][think_start + len("
<think>"):think_end].strip()
else:
    think_content = "No think content found."

# 提取正常的文本内容
```

```
normal_content = chat_response["message"]['content'][think_end + len("
</think>"):].strip()

# 打印结果
print("思考内容:\n", think_content)
print("\n正常内容:\n", normal_content)
```

```
思考内容:
 好，我现在需要帮用户生成一个关于人工智能的简短介绍。首先，我得理解用户的需求是什么。他们可能对AI不太了解，想要一个简洁明了的概述。

我应该从基础开始讲起，比如定义。人工智能是模拟人类智能的技术，这点很重要。然后，我可以提到主要应用领域，比如机器学习、自然语言处理和计算机视觉，这样可以让介绍更有针对性。

接下来，我需要说明AI的应用范围，比如在医疗、金融和交通中的作用，这样用户能明白它的实际价值。同时，也不能忽视伦理和社会影响，这部分也是大家关心的点。

还要提到当前的发展阶段，强调它是一个快速发展的领域，这样能展示出未来的潜力。最后，保持整体内容简明扼要，适合快速阅读。

可能用户需要这个介绍用于学习、演讲或者作为参考资料。所以信息要准确，结构清晰，重点突出。我要确保涵盖主要方面，同时不显得冗长。这样用户就能得到一个全面又简洁的人工智能简介了。

正常内容:
 人工智能（Artificial Intelligence, AI）是指通过模拟人类智能的技术，使计算机系统能够执行如学习、推理、问题解决和自然语言处理等任务。AI技术广泛应用于医疗、金融、交通等领域，帮助提高效率并解决复杂问题。随着算法的进步和数据的增加，人工智能正逐步改变我们的生活方式和社会结构。
```

其他的重点参数和 `/generation` 参数使用方法也保持一致，示例代码如下：

```
import requests # type: ignore
import json

# 设置 API 端点
chat_url = "http://192.168.110.131:11434/api/chat"     # 这里需要根据实际情况进行修改

# 示例数据
chat_payload = {
    "model": "deepseek-r1:32b",    # 这里需要根据实际情况进行修改
    "messages": [
        {
            "role": "user",  # 消息角色，用户发送的消息
            "content": "请生成一个关于人工智能的简短介绍。"  # 用户的消息内容
        }
    ],
    "tools": [],  # 如果有工具可以在这里添加
    "stream": False,  # 默认使用的是True，如果设置为False，则返回的是一个完整的响应，而不是一个流式响应
    "keep_alive": "10m",    # 设置模型在请求后保持加载的时间
    "options":{
        "temperature": 0.7,
        "num_ctx":2048,
        "num_predict": 4096,
    }
}
```

```
# 调用聊天接口
response_chat = requests.post(chat_url, json=chat_payload)
if response_chat.status_code == 200:
    chat_response = response_chat.json()
    print("生成响应:", json.dumps(chat_response, ensure_ascii=False, indent=2))
else:
    print("生成请求失败:", response_chat.status_code, response_chat.text)
```


```
生成响应: {
  "model": "deepseek-r1:32b",
  "created_at": "2025-02-13T11:22:03.848936446Z",
  "message": {
    "role": "assistant",
    "content": "<think>\n好的，用户让我生成一个关于人工智能的简短介绍。首先，我得理解用户的需求是什么。可能是一个学生在做作业，或者是一个职场人士需要快速了解AI的基本概念。\n\n接下来，我应该考虑内容的结构。通常，一个好的简介应该包括定义、关键点和应用领域。这样可以让读者全面了解主题。\n\n然后，思考如何用简单明了的语言解释人工智能，避免太专业的术语，但又要准确。可能还要提到机器学习和深度学习这些核心技术，因为它们是AI的重要组成部分。\n\n再考虑应用场景，比如医疗、金融和交通等，这样可以展示AI的实际影响。最后，加入一些关于伦理和社会影响的内容，让介绍更全面。\n\n现在，把这些点组织成一个连贯的段落，确保逻辑清晰，语言流畅。同时要注意字数控制在简短范围内，大概200字左右。\n</think>\n\n人工智能（Artificial Intelligence, AI）是模拟人类智能的系统或机器，通过学习、推理和自主决策来执行任务。它涵盖多个领域，如机器学习、自然语言处理和计算机视觉，广泛应用于医疗、金融、交通等。AI不仅提高效率，还推动社会进步，但也引发伦理和社会挑战。"
  },
  "done_reason": "stop",
  "done": true,
  "total_duration": 8531725124,
  "load_duration": 62273516,
  "prompt_eval_count": 13,
  "prompt_eval_duration": 87000000,
  "eval_count": 254,
  "eval_duration": 8380000000
}
```


流式输出代码也要针对`/chat`接口的返回响应格式做略微的修改：

```
import requests  # type: ignore
import json

# 设置 API 端点
generate_url = "http://192.168.110.131:11434/api/generate"

# 示例数据
generate_payload = {
    "model": "deepseek-r1:32b",
    "prompt": "请生成一个关于人工智能的简短介绍。",
    "stream": True,  # 启用流式输出
    "options": {
        "temperature": 0.6,
        "keep_alive": "10m"
    }
}

# 调用生成接口
with requests.post(generate_url, json=generate_payload, stream=True) as
response_generate:
    if response_generate.status_code == 200:
```

```
        # 逐行读取流式响应
        for line in response_generate.iter_lines():
            if line:  # 确保行不为空
                # 解析 JSON 响应
                generate_response = json.loads(line)

                # 提取并打印 response 字段
                if "response" in generate_response:
                    print(generate_response["response"], end='')  # end='' 防止换行
                if generate_response.get("done", False):
                    break  # 如果 done 为 True，结束循环
    else:
        print("生成请求失败:", response_generate.status_code, response_generate.text)
```

<think>
好的，我现在需要帮用户生成一个关于人工智能的简短介绍。首先，我得理解用户的需求是什么。看起来他们可能对AI不太了解，所以需要一个简洁明了的解释。

接下来，我要考虑涵盖哪些关键点。人工智能的基本定义是必须有的，比如它是模拟人类智能的技术。然后，可以提到一些主要的应用领域，比如机器学习、自然语言处理和计算机视觉，这样可以让内容更具体。

我还需要强调AI的发展对社会的影响，比如在医疗、教育和交通等方面带来的变化，以及它如何改变我们的生活方式。不过，也要提醒用户注意伦理和社会问题，这显示了全面性。

最后，结构要清晰，每句话简洁有力，避免使用专业术语过多，让不同背景的读者都能理解。同时，控制在50字左右，确保信息精炼。
</think>

人工智能（Artificial Intelligence, AI）是模拟人类智能的技术，通过机器学习、自然语言处理和计算机视觉等方法，使计算机能够执行复杂任务。AI正广泛应用于医疗、教育、交通等领域，推动社会进步，但也需关注伦理和社会影响。