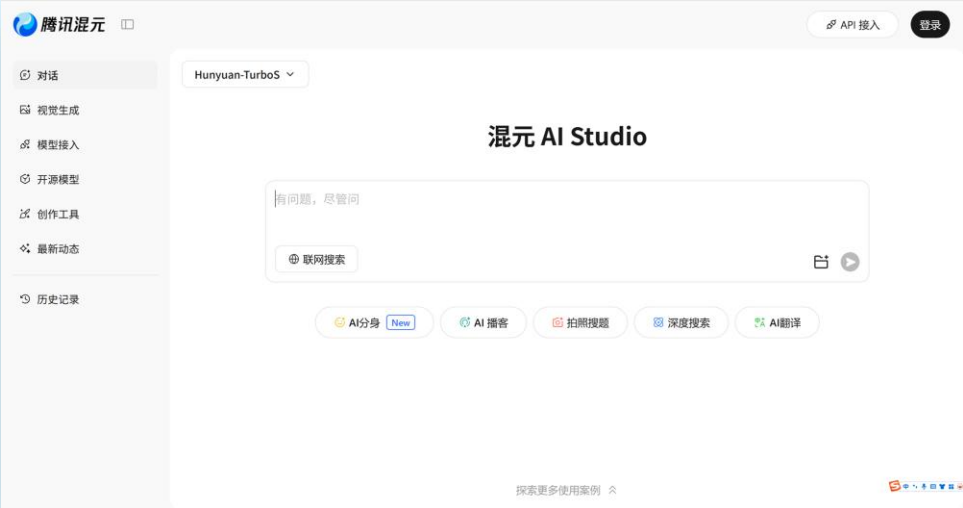
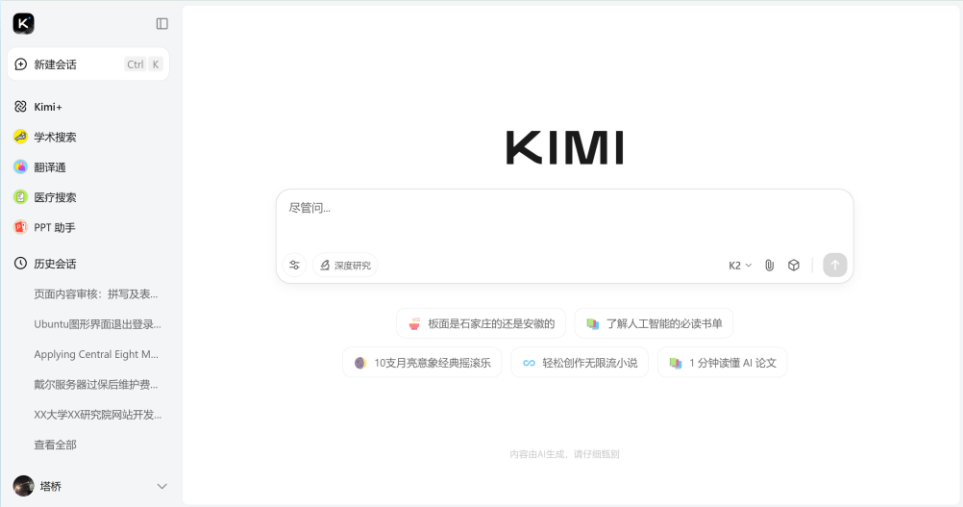




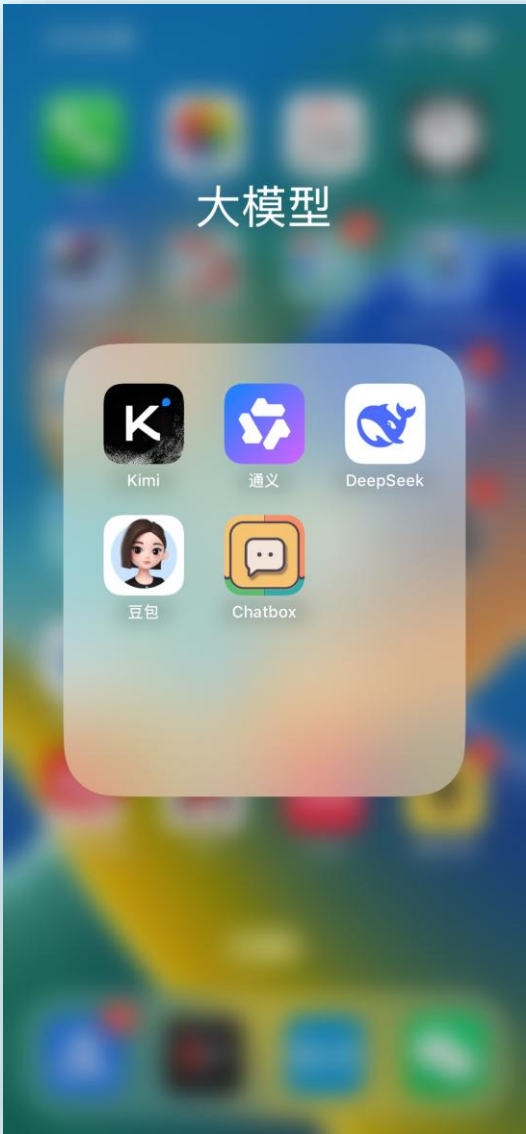
使用大模型：网页端



使用大模型：手机、应用程序



淘宝AI搜索

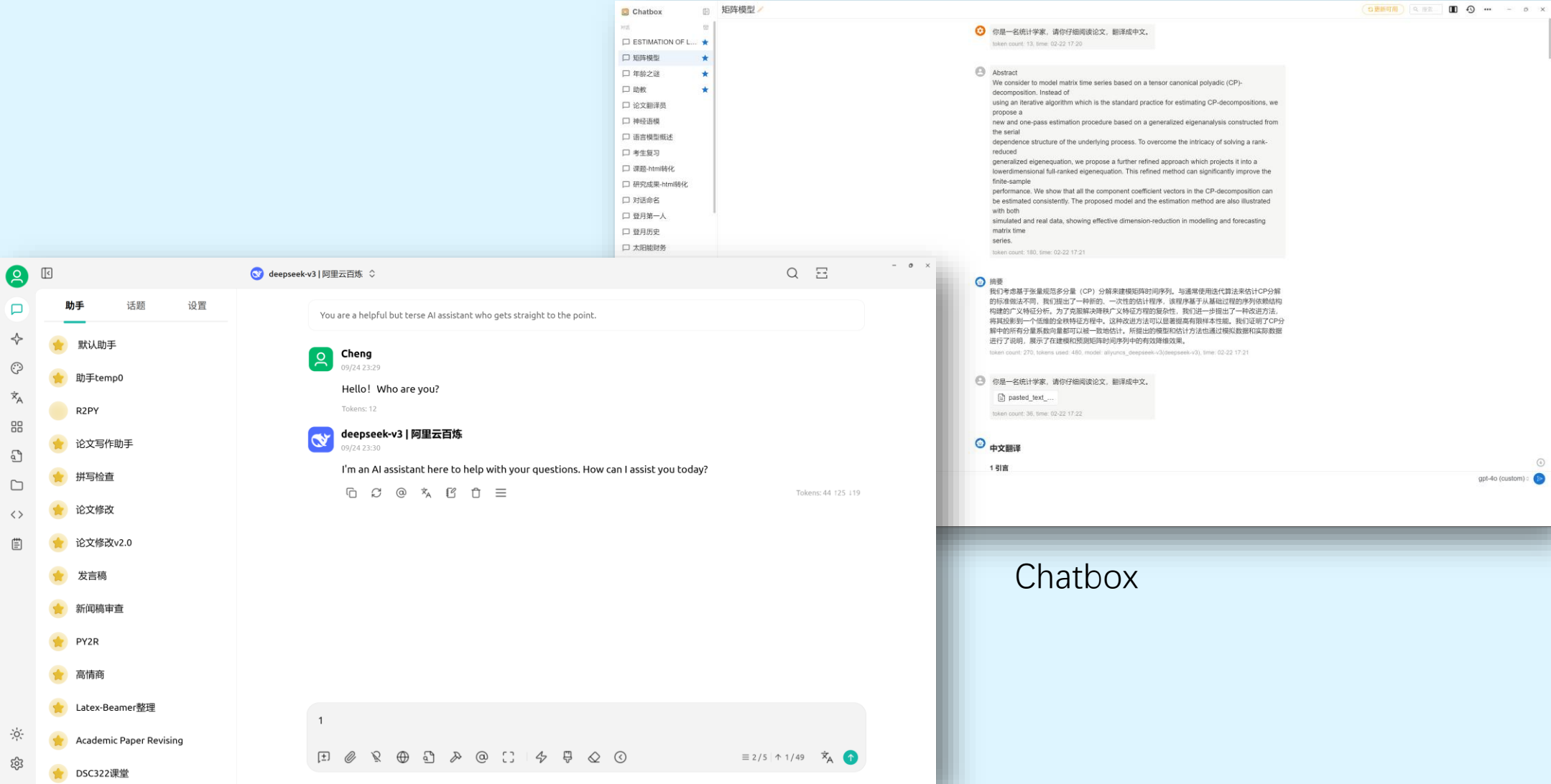


手机app



微信AI搜索

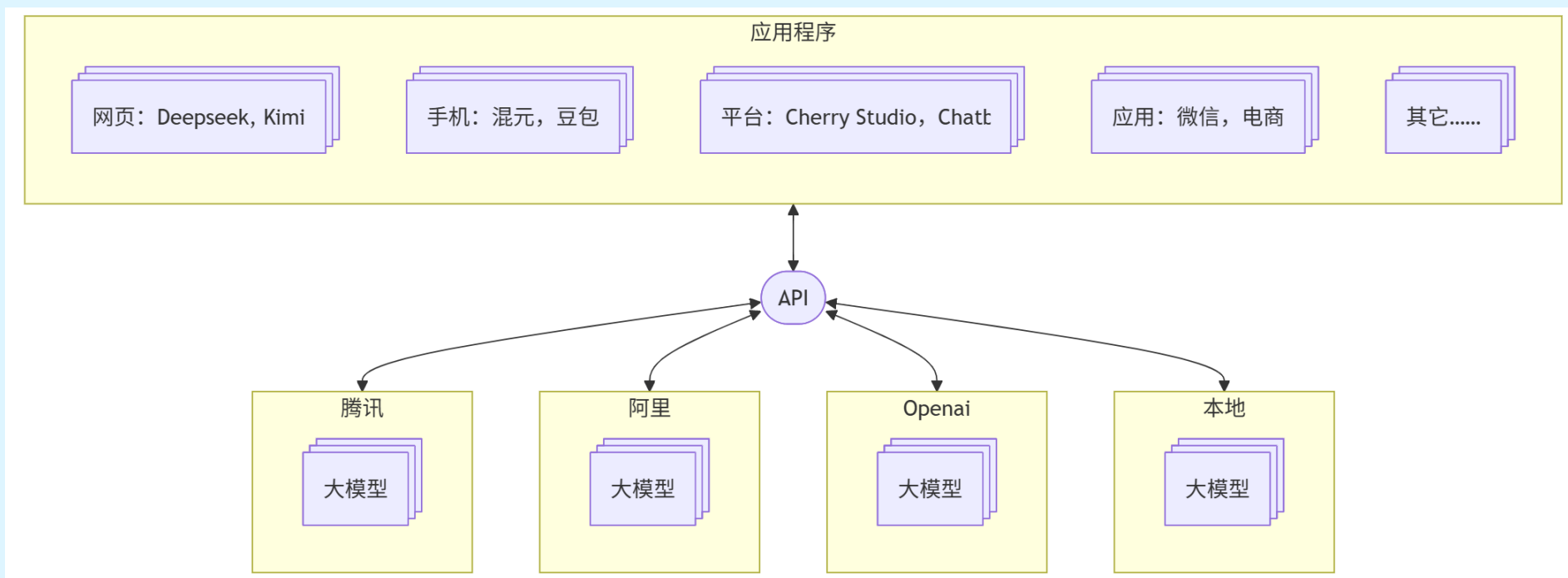
使用大模型：工具平台



Cherry Studio

什么是API

- **API** 的全称是 **应用程序编程接口**（**Application Program Interface**，**API**）。
 - **应用程序**：指任何软件程序，比如你的手机App、网站、电脑软件。
 - **编程**：指开发者通过写代码来与这个接口互动。
 - **接口**：指两个不同系统之间相互通信和交互的“契约”或“连接点”。
- **API** 是一套预先定义好的规则、协议和工具，允许不同的软件应用程序之间相互通信、交换数据和服务，而无需知道对方内部的实现细节。



什么是API

➤ 想象一下你去一家餐厅吃饭：

- **你（顾客）** 相当于一个**软件应用**（比如一个手机App或网站）。
- **厨房** 相当于另一个**软件系统或数据库**（比如微信的服务器、银行的数据库、天气数据的来源）。
厨房里有你需要的所有食材和资源，但你无法直接进入厨房。
- **服务员** 就是 **API**。
- **菜单** 就是 **API的文档**。

➤ 过程是这样的：

1. 你看菜单（**查阅API文档**），决定点什么菜（**发起什么请求**）。
2. 你告诉服务员：“我要一份牛排，七分熟”（**发出请求**）。
3. 服务员把你的要求准确地传达给厨房（**API接收请求并传递给系统**）。
4. 厨房做好牛排后，交给服务员（**系统处理完请求，返回数据**）。
5. 服务员把牛排端到你桌上（**API将数据返回给你的应用**）。

➤ 在这个过程中，你不需要知道厨房是如何运作的，厨师是谁，牛排是从哪个冰箱里拿出来的。你只需要通过服务员（API）这个标准接口，就能获得你想要的食物（数据或功能）。

Python代码实现

```
import OpenAI

with open('API-key.txt', 'r', encoding='utf-8') as file:
    key = file.read()

client = OpenAI(
    api_key = key,
    base_url = "https://dashscope.aliyuncs.com/compatible-mode/v1/"
)

def get_llm_response(prompt):
    """This function takes as input a prompt, which must be a
    string enclosed in quotation marks,
    and passes it to deepseek-v3 model. The function then saves
    the response of the model as
    a string.
    """
    completion = client.chat.completions.create(
        model="deepseek-v3",
        messages=[
            {
                "role": "system",
                "content": "You are a helpful but terse AI
                assistant who gets straight to the point.",
            },
            {"role": "user", "content": prompt},
        ],
        temperature=0.0,
    )
    response = completion.choices[0].message.content
    return response
```

搜索模型平台...

阿里云百炼

Ollama

CherryIN

OpenAI

硅基流动

ON

ON

ON

ON

ON

阿里云百炼

API 密钥

.....

[点击这里获取密钥](#)

API 地址

https://dashscope.aliyuncs.com/compatible-mode/v1/

https://dashscope.aliyuncs.com/compatible-mode/v1/chat/completions

AI Assistant

提示词设置

模型设置

知识库设置

MCP 服务器

默认模型

模型温度 ?

0

0.7

deepseek-v3 | 阿里云百炼

You are a helpful but terse AI assistant who gets straight to the point.

Cheng

09/24 23:29

Hello! Who are you?

Tokens: 12

deepseek-v3 | 阿里云百炼

09/24 23:29

I'm an AI assistant here to help with your questions. How can I assist you today?

🗑️ ↺️ @ ✖️ 📄 🗑️ ☰

Tokens: 44 125 119

Python代码实现

```
import OpenAI

with open('API-key.txt', 'r', encoding='utf-8') as file:
    key = file.read()

client = OpenAI(
    api_key = key,
    base_url = "https://dashscope.aliyuncs.com/compatible-mode/v1/"
)

def get_llm_response(prompt):
    """This function takes as input a prompt, which must be a
    string enclosed in quotation marks,
    and passes it to deepseek-v3 model. The function then saves
    the response of the model as
    a string.
    """
    completion = client.chat.completions.create(
        model="deepseek-v3",
        messages=[
            {
                "role": "system",
                "content": "You are a helpful but terse AI
                assistant who gets straight to the point.",
            },
            {"role": "user", "content": prompt},
        ],
        temperature=0.0,
    )
    response = completion.choices[0].message.content
    return response
```

搜索模型平台...

阿里云百炼 ON

Ollama ON

CherryIN ON

OpenAI ON

硅基流动 ON

阿里云百炼

API 密钥

.....

点击这里获取密钥

API 地址

https://dashscope.aliyuncs.com/compatible-mode/v1/

https://dashscope.aliyuncs.com/compatible-mode/v1/chat/completions

AI Assistant

提示词设置

模型设置

知识库设置

MCP 服务器

默认模型

模型温度 ?

0 0.7

deepseek-v3 | 阿里云百炼

You are a helpful but terse AI assistant who gets straight to the point.

Cheng

09/24 23:29

Hello! Who are you?

Tokens: 12

deepseek-v3 | 阿里云百炼

09/24 23:29

I'm an AI assistant here to help with your questions. How can I assist you today?

Tokens: 44 125 119

Python代码实现

```
import OpenAI

with open('API-key.txt', 'r', encoding='utf-8') as file:
    key = file.read()

client = OpenAI(
    api_key = key,
    base_url = "https://dashscope.aliyuncs.com/compatible-mode/v1/"
)

def get_llm_response(prompt):
    """This function takes as input a prompt, which must be a
    string enclosed in quotation marks,
    and passes it to deepseek-v3 model. The function then saves
    the response of the model as
    a string.
    """
    completion = client.chat.completions.create(
        model="deepseek-v3",
        messages=[
            {
                "role": "system",
                "content": "You are a helpful but terse AI
                assistant who gets straight to the point.",
            },
            {"role": "user", "content": prompt},
        ],
        temperature=0.0,
    )
    response = completion.choices[0].message.content
    return response
```

搜索模型平台...

阿里云百炼 ON

Ollama ON

CherryIN ON

OpenAI ON

硅基流动 ON

阿里云百炼 [API 密钥](#)
.....
[点击这里获取密钥](#)
API 地址

https://dashscope.aliyuncs.com/compatible-mode/v1/

https://dashscope.aliyuncs.com/compatible-mode/v1/chat/completions

提示词设置

模型设置

知识库设置

MCP 服务器

默认模型

模型温度 ?

0 0.7

deepseek-v3 | 阿里云百炼

You are a helpful but terse AI assistant who gets straight to the point.

Cheng
09/24 23:29
Hello! Who are you?
Tokens: 12

deepseek-v3 | 阿里云百炼
09/24 23:29
I'm an AI assistant here to help with your questions. How can I assist you today?

Tokens: 44 125 119

Python代码实现

```
import OpenAI

with open('API-key.txt', 'r', encoding='utf-8') as file:
    key = file.read()

client = OpenAI(
    api_key = key,
    base_url = "https://dashscope.aliyuncs.com/compatible-mode/v1/"
)

def get_llm_response(prompt):
    """This function takes as input a prompt, which must be a
    string enclosed in quotation marks,
    and passes it to deepseek-v3 model. The function then saves
    the response of the model as
    a string.
    """
    completion = client.chat.completions.create(
        model="deepseek-v3",
        messages=[
            {
                "role": "system",
                "content": "You are a helpful but terse AI
                assistant who gets straight to the point.",
            },
            {"role": "user", "content": prompt},
        ],
        temperature=0.0,
    )
    response = completion.choices[0].message.content
    return response
```

搜索模型平台...

阿里云百炼

Ollama

CherryIN

OpenAI

硅基流动

ON

ON

ON

ON

ON

阿里云百炼

API 密钥

.....

点击这里获取密钥

API 地址

https://dashscope.aliyuncs.com/compatible-mode/v1/

https://dashscope.aliyuncs.com/compatible-mode/v1/chat/completions

AI Assistant

提示词设置

模型设置

知识库设置

MCP 服务器

默认模型

模型温度 ?

0

0.7

deepseek-v3 | 阿里云百炼

You are a helpful but terse AI assistant who gets straight to the point.

Cheng

09/24 23:29

Hello! Who are you?

Tokens: 12

deepseek-v3 | 阿里云百炼

09/24 23:29

I'm an AI assistant here to help with your questions. How can I assist you today?

Tokens: 44 125 119

Python代码实现

```
import OpenAI

with open('API-key.txt', 'r', encoding='utf-8') as file:
    key = file.read()

client = OpenAI(
    api_key = key,
    base_url = "https://dashscope.aliyuncs.com/compatible-mode/v1/"
)

def get_llm_response(prompt):
    """This function takes as input a prompt, which must be a
    string enclosed in quotation marks,
    and passes it to deepseek-v3 model. The function then saves
    the response of the model as
    a string.
    """
    completion = client.chat.completions.create(
        model="deepseek-v3",
        messages=[
            {
                "role": "system",
                "content": "You are a helpful but terse AI
                assistant who gets straight to the point.",
            },
            {"role": "user", "content": prompt},
        ],
        temperature=0.0,
    )
    response = completion.choices[0].message.content
    return response
```

搜索模型平台...

阿里云百炼

Ollama

CherryIN

OpenAI

硅基流动

ON

ON

ON

ON

ON

阿里云百炼

API 密钥

.....

点击这里获取密钥

API 地址

https://dashscope.aliyuncs.com/compatible-mode/v1/

https://dashscope.aliyuncs.com/compatible-mode/v1/chat/completions

AI Assistant

提示词设置

默认模型

模型设置

知识库设置

MCP 服务器

模型温度 ?

0

0.7

deepseek-v3 | 阿里云百炼

You are a helpful but terse AI assistant who gets straight to the point.

Cheng

09/24 23:29

Hello! Who are you?

Tokens: 12

deepseek-v3 | 阿里云百炼

09/24 23:29

I'm an AI assistant here to help with your questions. How can I assist you today?

Tokens: 44 125 119

Python代码实现

```
import OpenAI

with open('API-key.txt', 'r', encoding='utf-8') as file:
    key = file.read()

client = OpenAI(
    api_key = key,
    base_url = "https://dashscope.aliyuncs.com/compatible-mode/v1/"
)

def get_llm_response(prompt):
    """This function takes as input a prompt, which must be a
    string enclosed in quotation marks,
    and passes it to deepseek-v3 model. The function then saves
    the response of the model as
    a string.
    """
    completion = client.chat.completions.create(
        model="deepseek-v3",
        messages=[
            {
                "role": "system",
                "content": "You are a helpful but terse AI
                assistant who gets straight to the point.",
            },
            {"role": "user", "content": prompt},
        ],
        temperature=0.0,
    )
    response = completion.choices[0].message.content
    return response
```

搜索模型平台...

阿里云百炼 ON

Ollama ON

CherryIN ON

OpenAI ON

硅基流动 ON

阿里云百炼

API 密钥

.....

点击这里获取密钥

API 地址

https://dashscope.aliyuncs.com/compatible-mode/v1/

https://dashscope.aliyuncs.com/compatible-mode/v1/chat/completions

AI Assistant

提示词设置

模型设置

知识库设置

MCP 服务器

默认模型

模型温度 ?

0 0.7

deepseek-v3 | 阿里云百炼

deepseek-v3 | 阿里云百炼

You are a helpful but terse AI assistant who gets straight to the point.

Cheng

09/24 23:29

Hello! Who are you?

Tokens: 12

deepseek-v3 | 阿里云百炼

09/24 23:29

I'm an AI assistant here to help with your questions. How can I assist you today?

Tokens: 44 125 119

Python代码实现

```
import OpenAI

with open('API-key.txt', 'r', encoding='utf-8') as file:
    key = file.read()

client = OpenAI(
    api_key = key,
    base_url = "https://dashscope.aliyuncs.com/compatible-mode/v1/"
)

def get_llm_response(prompt):
    """This function takes as input a prompt, which must be a
    string enclosed in quotation marks,
    and passes it to deepseek-v3 model. The function then saves
    the response of the model as
    a string.
    """
    completion = client.chat.completions.create(
        model="deepseek-v3",
        messages=[
            {
                "role": "system",
                "content": "You are a helpful but terse AI
                assistant who gets straight to the point.",
            },
            {"role": "user", "content": prompt},
        ],
        temperature=0.0,
    )
    response = completion.choices[0].message.content
    return response
```

搜索模型平台...

阿里云百炼

Ollama

CherryIN

OpenAI

硅基流动

ON

ON

ON

ON

ON

阿里云百炼

API 密钥

.....

点击这里获取密钥

API 地址

https://dashscope.aliyuncs.com/compatible-mode/v1/

https://dashscope.aliyuncs.com/compatible-mode/v1/chat/completions

AI Assistant

提示词设置

模型设置

知识库设置

MCP 服务器

默认模型

模型温度 ?

0

0.7

deepseek-v3 | 阿里云百炼

You are a helpful but terse AI assistant who gets straight to the point.

Cheng

09/24 23:29

Hello! Who are you?

Tokens: 12

deepseek-v3 | 阿里云百炼

09/24 23:29

I'm an AI assistant here to help with your questions. How can I assist you today?

Tokens: 44 125 119

Python代码实现

```
import OpenAI

with open('API-key.txt', 'r', encoding='utf-8') as file:
    key = file.read()

client = OpenAI(
    api_key = key,
    base_url = "https://dashscope.aliyuncs.com/compatible-mode/v1/"
)

def get_llm_response(prompt):
    """This function takes as input a prompt, which must be a
    string enclosed in quotation marks,
    and passes it to deepseek-v3 model. The function then saves
    the response of the model as
    a string.
    """
    completion = client.chat.completions.create(
        model="deepseek-v3",
        messages=[
            {
                "role": "system",
                "content": "You are a helpful but terse AI
                assistant who gets straight to the point.",
            },
            {"role": "user", "content": prompt},
        ],
        temperature=0.0,
    )
    response = completion.choices[0].message.content
    return response
```

搜索模型平台...

阿里云百炼

Ollama

CherryIN

OpenAI

硅基流动

ON

ON

ON

ON

ON

阿里云百炼

API 密钥

.....

[点击这里获取密钥](#)

API 地址

https://dashscope.aliyuncs.com/compatible-mode/v1/

https://dashscope.aliyuncs.com/compatible-mode/v1/chat/completions

AI Assistant

提示词设置

模型设置

知识库设置

MCP 服务器

默认模型

模型温度 ?

0

0.7

deepseek-v3 | 阿里云百炼

You are a helpful but terse AI assistant who gets straight to the point.

Cheng

09/24 23:29

Hello! Who are you?

Tokens: 12

deepseek-v3 | 阿里云百炼

09/24 23:29

I'm an AI assistant here to help with your questions. How can I assist you today?

Tokens: 44 125 119

Python代码实现

```
import OpenAI

with open('API-key.txt', 'r', encoding='utf-8') as file:
    key = file.read()

client = OpenAI(
    api_key = key,
    base_url = "https://dashscope.aliyuncs.com/compatible-mode/v1/"
)

def get_llm_response(prompt):
    """This function takes as input a prompt, which must be a
    string enclosed in quotation marks,
    and passes it to deepseek-v3 model. The function then saves
    the response of the model as
    a string.
    """
    completion = client.chat.completions.create(
        model="deepseek-v3",
        messages=[
            {
                "role": "system",
                "content": "You are a helpful but terse AI
                assistant who gets straight to the point.",
            },
            {"role": "user", "content": prompt},
        ],
        temperature=0.0,
    )
    response = completion.choices[0].message.content
    return response
```

```
prompt = "Hello! Who are you?"
get_llm_response(prompt)
```

```
"I'm an AI assistant here to help with your questions. What do you need?"
```

搜索模型平台...

阿里云百炼 ON

Ollama ON

CherryIN ON

OpenAI ON

硅基流动 ON

阿里云百炼 [API 密钥](#)
.....
[点击这里获取密钥](#)
API 地址
https://dashscope.aliyuncs.com/compatible-mode/v1/
https://dashscope.aliyuncs.com/compatible-mode/v1/chat/completions

AI Assistant

提示词设置

模型设置

知识库设置

MCP 服务器

默认模型

模型温度 ?
0 0.7

deepseek-v3 | 阿里云百炼

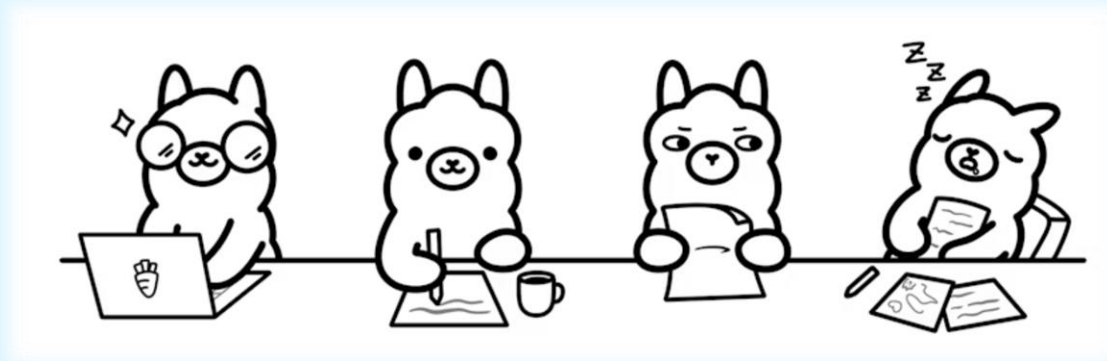
You are a helpful but terse AI assistant who gets straight to the point.

Cheng
09/24 23:29
Hello! Who are you?
Tokens: 12

deepseek-v3 | 阿里云百炼
09/24 23:29
I'm an AI assistant here to help with your questions. How can I assist you today?

Tokens: 44 125 119

本地部署大模型：Ollama



- Ollama 是一个本地大模型运行与管理工具，主要用于在个人电脑上快速部署和调用各种开源大语言模型。它的特点包括：
- ✓ **本地运行：** Ollama 可以让你直接在本地上运行大模型，不需要依赖云端。这样既能提高隐私和安全性，也可以在离线环境中使用。
 - ✓ **简单调用：** 它提供了命令行工具和 API 接口，让用户可以像调用 ChatGPT 一样来运行模型。
 - ✓ **模型管理：** 支持下载和管理多个开源模型（如 Llama 2、Mistral、Deepseek 等）。**可扩展与定制：** 可以加载本地自定义模型或微调模型。与其他开发框架（如 Python、Node.js、前端应用）结合，用于构建 AI 应用。
 - ✓ **跨平台支持：** Ollama 目前支持 **macOS**、**Linux** 和 **Windows**（预览版），安装方式可直接下载安装包。

本地部署大模型：Ollama

下载地址：<https://ollama.com/download>



Cloud models are now available in Ollama

**Chat & build with
open models**

Download

Available for macOS,
Windows, and Linux

Download Ollama



macOS



Linux

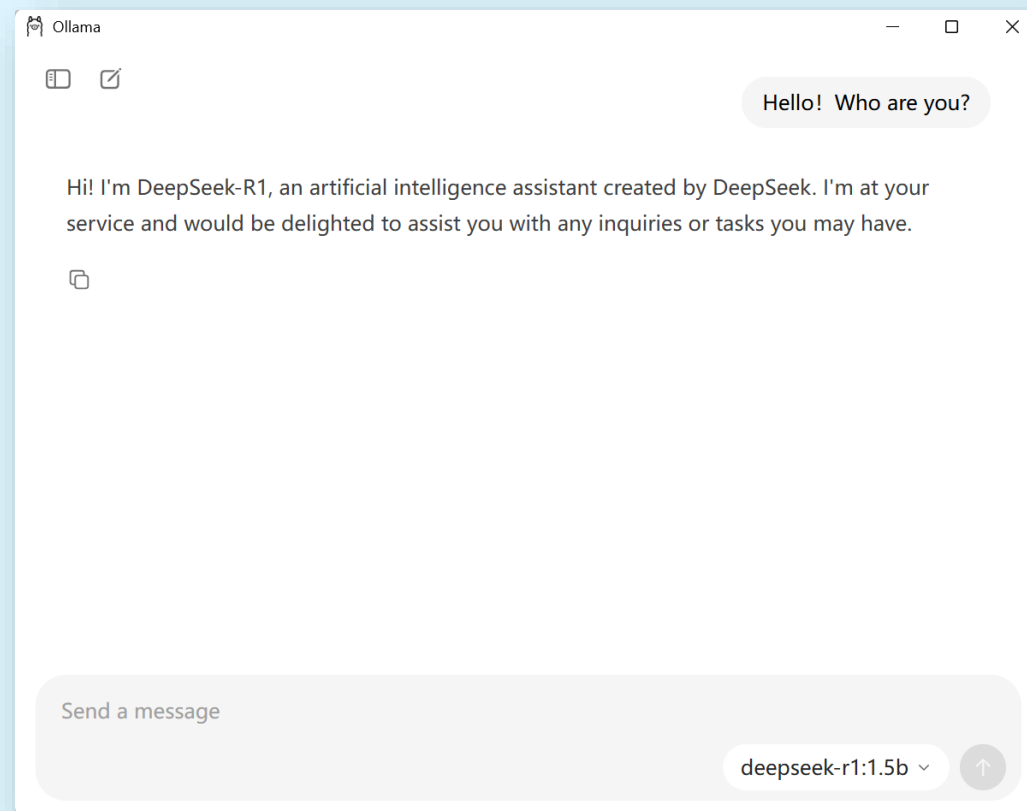
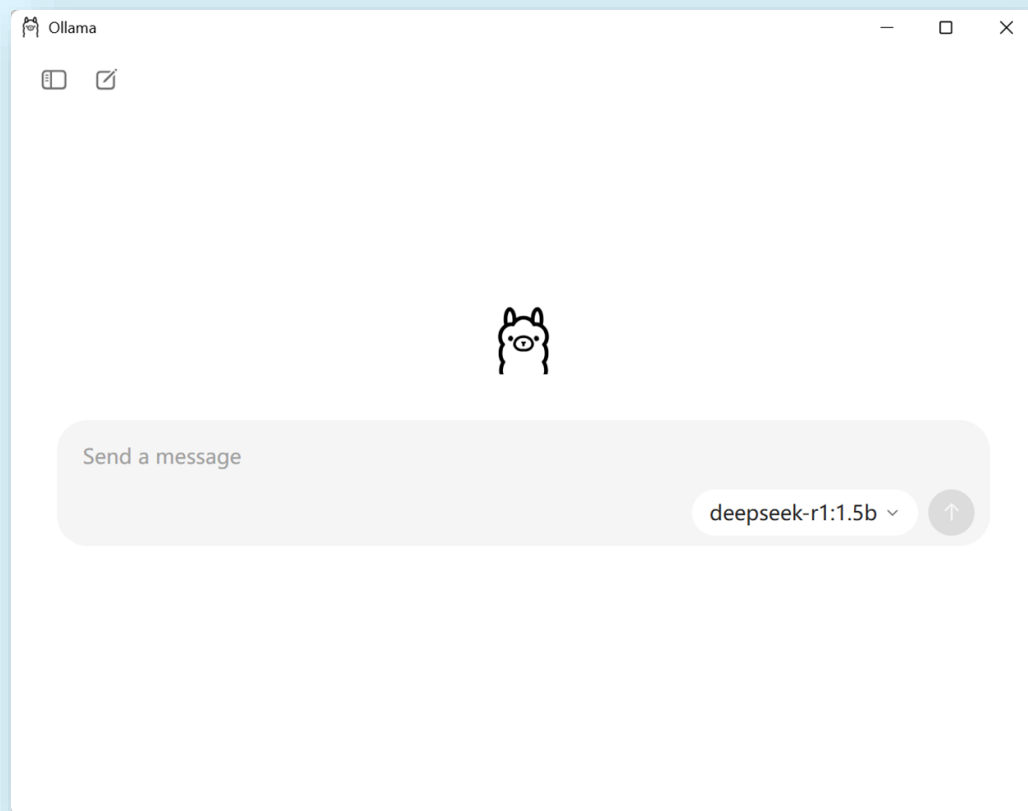


Windows

Download for Windows

Requires Windows 10 or later

本地部署大模型：Ollama



本地部署大模型：Ollama



☰

🔍 Search models

↕

Cloud

Embedding

Vision

Tools

Thinking

gpt-oss

OpenAI's open-weight models designed for powerful reasoning, agentic tasks, and versatile developer use cases.

tools

thinking

cloud

20b

120b

↓ 2.7M

👁 5

🕒 4 days ago

deepseek-r1

DeepSeek-R1 is a family of open reasoning models with performance approaching that of leading models, such as O3 and Gemini 2.5 Pro.

tools

thinking

1.5b

7b

8b

14b

32b

70b

671b

↓ 63.1M

👁 35

🕒 2 months ago

gemma3

The current, most capable model that runs on a single GPU.

vision

270m

1b

4b

12b

27b

↓ 17.9M

👁 26

🕒 1 month ago

embeddinggemma

EmbeddingGemma is a 300M parameter embedding model from Google.

embedding

300m

↓ 40K

👁 5

🕒 2 weeks ago

qwen3

Qwen3 is the latest generation of large language models in Qwen series, offering a comprehensive suite of dense and mixture-of-experts (MoE) models.

tools

thinking

0.6b

1.7b

4b

8b

14b

30b

32b

235b

↓ 8.7M

👁 56

🕒 1 month ago



☰

deepseek-r1

ollama run deepseek-r1

📄

↓ 63.1M

🕒 2 months ago

DeepSeek-R1 is a family of open reasoning models with performance approaching that of leading models, such as O3 and Gemini 2.5 Pro.

tools

thinking

1.5b

7b

8b

14b

32b

70b

671b

Models

View all →

35 models

deepseek-r1:latest

5.2GB · 128K context window · Text · 3 months ago

deepseek-r1:1.5b

1.1GB · 128K context window · Text · 3 months ago

deepseek-r1:7b

4.7GB · 128K context window · Text · 3 months ago

deepseek-r1:8b

latest

5.2GB · 128K context window · Text · 3 months ago

deepseek-r1:14b

9.0GB · 128K context window · Text · 3 months ago

deepseek-r1:32b

20GB · 128K context window · Text · 3 months ago

deepseek-r1:70b

43GB · 128K context window · Text · 3 months ago

deepseek-r1:671b

404GB · 160K context window · Text · 2 months ago

Readme



☰

qwen3

ollama run qwen3

📄

↓ 8.7M

🕒 1 month ago

Qwen3 is the latest generation of large language models in Qwen series, offering a comprehensive suite of dense and mixture-of-experts (MoE) models.

tools

thinking

0.6b

1.7b

4b

8b

14b

30b

32b

235b

Models

View all →

56 models

qwen3:latest

5.2GB · 40K context window · Text · 3 months ago

qwen3:0.6b

523MB · 40K context window · Text · 3 months ago

qwen3:1.7b

1.4GB · 40K context window · Text · 3 months ago

qwen3:4b

2.5GB · 256K context window · Text · 1 month ago

qwen3:8b

latest

5.2GB · 40K context window · Text · 3 months ago

qwen3:14b

9.3GB · 40K context window · Text · 3 months ago

qwen3:30b

19GB · 256K context window · Text · 1 month ago

qwen3:32b

20GB · 40K context window · Text · 3 months ago

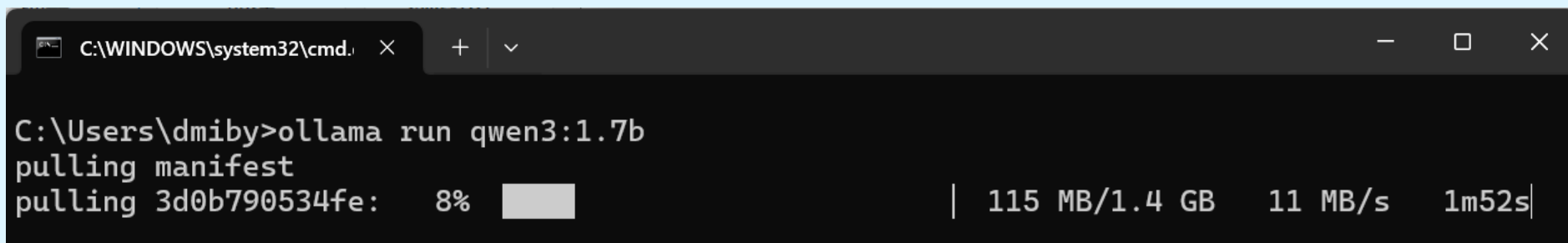
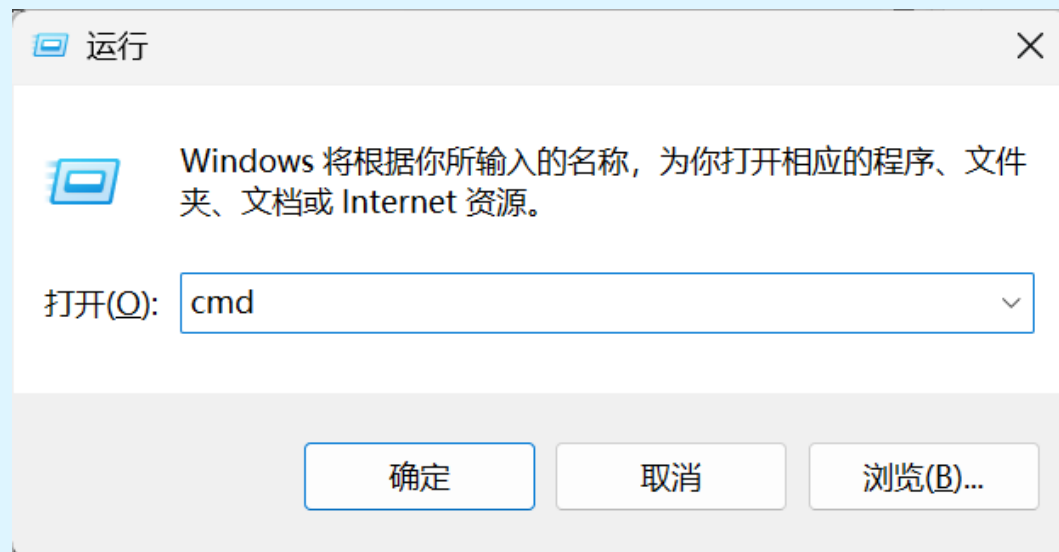
qwen3:235b

142GB · 256K context window · Text · 1 month ago

本地部署大模型：Ollama

以Windows系统为例：

- Win + R：打开“运行”对话框
- 输入 `cmd` 回车，打开命令行窗口
- 输入 `Ollama run qwen3:1.7b`
 - ✓ 首次运行将下载qwen3:1.7b模型
 - ✓ 以后运行，将开启对话

A screenshot of a Windows Command Prompt window. The title bar shows the path 'C:\WINDOWS\system32\cmd.' and standard window controls. The command prompt shows the command 'ollama run qwen3:1.7b' being executed. The output shows 'pulling manifest' and 'pulling 3d0b790534fe: 8%' with a progress bar. On the right side, there is a status bar showing '115 MB/1.4 GB', '11 MB/s', and '1m52s'.

本地部署大模型：Ollama

```
C:\WINDOWS\system32\cmd.exe
C:\Users\dmiby>ollama run qwen3:1.7b
>>> Hello! Who are you?
Thinking...
Okay, the user asked, "Hello! Who are you?" So I need to respond in a friendly and informative way.

First, I should greet them back. Maybe start with "Hello!" to keep it consistent. Then, introduce myself as an AI assistant. Mention that I'm here to help with various tasks like answering questions, providing information, and assisting with tasks.

I should highlight my capabilities, like understanding natural language, helping with different subjects, and supporting various activities. Also, mention that I'm here to help and encourage them to ask questions.

Make sure the tone is friendly and approachable. Avoid technical jargon. Keep it simple and conversational. Maybe add an emoji to keep it light, like 😊.

Check for any possible misunderstandings. Ensure the response is clear and not too long. Summarize key points without being overwhelming. End with a prompt to ask questions so the user feels encouraged to interact.
...done thinking.

Hello! 😊 I'm an AI assistant designed to help with various tasks and answer questions. I can understand and respond in multiple languages, and I'm here to support you in any way I can. Whether you need help with homework, creative ideas, or general knowledge, feel free to ask! Let me know how I can assist you today. 🌟

>>> |Send a message (/? for help)
```

本地部署大模型：Ollama

搜索模型平台...

 阿里云百炼

 Ollama

 CherryIN

 OpenAI

 硅基流动

 AiHubMix

 深度求索

 ocoolAI

 百度云千帆

 LM Studio

 Anthropic

 PPIO 派欧云

ON

ON

ON

ON

ON

ON

ON

ON

ON

ON

ON

Ollama: 3 个模型通过健康检测

API 密钥

API 密钥

检测

API 地址

http://localhost:11434

http://localhost:11434/v1/chat/completions / 结尾忽略 v1 版本, # 结尾强制使用输入地址

模型 3

deepseek-r1

deepseek-r1:1.5b

8.98s

✓

⚙️

—

deepseek-r1:7b

6.92s

✓

⚙️

—

qwen3

qwen3:1.7b

0.67s

✓

⚙️

—

查看 Ollama 文档 和 模型 获取更多详情

本地部署大模型：Ollama

 qwen3:1.7b | Ollama

你好，我是默认助手。你可以立刻开始跟我聊天



Cheng

09/25 00:48

Hello! Who are you?

Tokens: 6



qwen3:1.7b | Ollama

09/25 00:48

Hello! I'm a large language model developed by Alibaba Cloud. I can help with various tasks, such as answering questions, writing texts, coding, and more. How can I assist you today?

Tokens: 63 118 145