ollama 安装和使用

Ollama 安装与使用指南

Ollama 是一个强大的工具，允许你在本地运行各种大语言模型（**LLM**），如 LLaMA、Mistral、Gemma 等。以下是详细的安装和使用教程，涵盖 **Windows**、**macOS**、**Linux** 系统。

1. 安装 Ollama

方法 1：直接下载安装（推荐）

Windows / macOS

1. 访问 **Ollama 官网**，下载对应系统的安装包。
2. 运行安装程序，按照提示完成安装。
3. 安装完成后，打开终端（Windows 用 CMD/PowerShell，macOS 用 Terminal），输入：

   ```
   ollama --version
   ```

   如果显示版本号（如 `v0.1.30`），说明安装成功。

Linux（Ubuntu/Debian/Rocky Linux）

```
# 使用 curl 下载并安装
curl -fsSL https://ollama.com/install.sh | sh
```

安装完成后，运行：

```
ollama serve & # 启动服务（后台运行）
```

方法 2: 使用 Docker 安装（适合开发者）

```
# 拉取 Ollama 镜像
docker pull ollama/ollama


# 运行容器（数据持久化到本地）
docker run -d -v ollama:/root/.ollama -p 11434:11434 --name ollama ollama/ollama
```

- `-p 11434:11434`: Ollama 默认 API 端口。
- `-v ollama:/root/.ollama`: 存储模型数据，避免容器删除后丢失。

---

2. 下载模型

Ollama 支持多种模型，如:

- `llama2`（Meta 开源）
- `mistral`（轻量高效）
- `gemma`（Google 轻量级）
- `phi`（微软小模型）
- `qwen`（阿里通义千问）

下载模型

```
ollama pull llama2  # 下载 LLaMA 2
ollama pull mistral # 下载 Mistral
ollama pull gemma   # 下载 Google Gemma
```

⚠ 如果下载慢，可以使用代理:

```
export HTTP_PROXY=http://127.0.0.1:7890  # 替换为你的代理
export HTTPS_PROXY=http://127.0.0.1:7890
ollama pull llama2
```

---

3. 运行模型

□ 方式 1：命令行交互

```
ollama run llama2  # 运行 LLaMA 2
```

输入问题，如：

```
>>> 用中文写一首关于春天的诗
```

模型会生成回答，按 `Ctrl+D` 退出。

□ 方式 2： API 调用

Ollama 提供 REST API（默认端口 `11434`），可以用 `curl` 或 Python 调用：

```
curl -X POST http://localhost:11434/api/generate -d '{
  "model": "llama2",
  "prompt": "你好，你是谁？"
}'
```

Python 示例：

```
import requests

response = requests.post(
    "http://localhost:11434/api/generate",
    json={"model": "llama2", "prompt": "解释一下人工智能"}
)
print(response.text)
```

---

4. 自定义模型

你可以基于现有模型微调（Fine-tuning）：

1. 创建 `Modelfile`：

   ```
   FROM llama2
   SYSTEM """你是一个专业的中文助手，回答要简洁。"""
   ```

2. 构建自定义模型：

   ```
   ollama create my-llama -f Modelfile
   ```

3. 运行：

   ```
   ollama run my-llama
   ```

---

5. 常见问题

□ □□ 1：`ollama` 命令找不到

- **Windows**：检查是否添加到了 `PATH`（安装时勾选）。
- **Linux/macOS**：手动添加环境变量：

  ```
  export PATH=$PATH:/usr/local/bin
  ```

□□ 问题 2：下载模型太慢

- 方法 **1**：使用代理（见上文）。
- 方法 **2**：手动下载模型文件，放入 `~/.ollama/models`（Linux/macOS）或 `C:\Users\<用户名>\.ollama\models`（Windows）。

□□ 问题 3：端口冲突

如果 `11434` 端口被占用，可以修改端口：

```
OLLAMA_HOST=0.0.0.0:11435 ollama serve  # 改用 11435
```

---

6. 进阶用法

□□ 管理模型

```
ollama list     # 查看已安装的模型
ollama rm llama2  # 删除模型
```

⚙ 配置 GPU 加速

Ollama 支持 **NVIDIA CUDA** 和 **Apple Metal** 加速:

- **Linux (NVIDIA)**: 安装 NVIDIA Container Toolkit, 然后:

  ```
  docker run --gpus all -p 11434:11434 ollama/ollama
  ```

- **macOS (Metal)**: 自动启用，无需额外配置。

---

7. 总结

操作	命令示例
安装 Ollama	`curl -fsSL https://ollama.com/install.sh | sh`
下载模型	`ollama pull llama2`
运行模型	`ollama run mistral`
调用 API	`curl -X POST http://localhost:11434/api/generate`
自定义模型	`ollama create my-model -f Modelfile`
查看已安装模型	`ollama list`

□□□□□□□□□□□□□□□□□□□□ 如需更多模型，可查看 **Ollama** 官方库____。

Revision #2
Created 3 April 2025 14:22:23 by Admin
Updated 9 April 2025 03:10:28 by Admin