

DeepSeek R1 本地部署

(Ollama+Docker+OpenWebUI)

1. 软硬件需求

1.1 硬件需求

1.2 软件依赖

2. Ollama 下载与安装

2.1 Ollama 的下载

2.2 Ollama 的安装

2.3 Ollama 的运行

3. deepseek-r1 模型安装

3.1 Ollama 安装 deepseek-r1 模型

3.2 使用 deepseek-r1 模型

4. 安装 Docker

4.1 检查安装环境

4.2 安装 Docker for Desktop

4.3 配置国内镜像源

5. 安装 Open WebUI

5.1 安装和启动 open-webui

5.2 登录 Open-WebUI 进入 deepseek-r1

1. 软硬件需求

1.1 硬件需求

DeepSeek R1 本地部署的硬件需求如下：

基本推荐配置：CPU: 8核或以上，内存: 32GB或以上，存储: 100GB SSD或以上，GPU: NVIDIA GTX 1080或同等性能

DeepSeek-7B 推荐配置：RTX3060 以上显卡 + 16G以上 内存 + 50G SSD 存储空间

DeepSeek-70B 推荐配置：RTX4060 以上显卡 + 32G以上 内存 + 200G SSD 存储空间

1.2 软件依赖

DeepSeek R1 本地部署的软件依赖包括操作系统、容器化工具、编程语言、深度学习框架、数据库、消息队列、Web 服务器、GPU 加速工具、版本控制、包管理工具、监控与日志工具以及安全工具

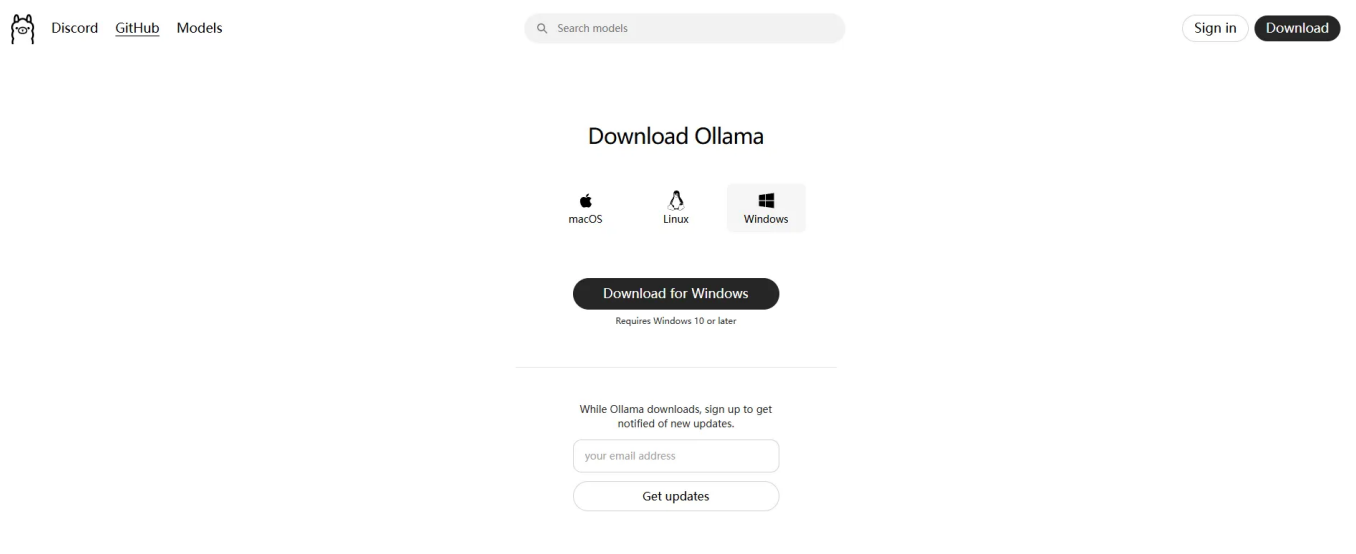
2. Ollama 下载与安装

2.1 Ollama 的下载

Ollama 是一个开源的本地模型运行工具，可以方便地下载和运行各种开源模型，如 Llama、Qwen、DeepSeek 等。Ollama 提供高效、灵活的大模型部署和管理解决方案，可以简化模型的部署流程。支持 macOS、Linux、Windows 操作系统。

官网地址：<https://ollama.com/>

下载地址：<https://ollama.com/download>



2.2 Ollama 的安装

下载完成后直接点击 OllamaSetup.exe 进行安装。

安装完成后，cmd 打开命令行窗口，输入 “ollama -v” 测试，显示 ollama 的版本为 “0.5.7”，表明安装成功。

```
命令提示符
Microsoft Windows [版本 10.0.22631.4751]
(c) Microsoft Corporation。保留所有权利。

C:\Users\Administrator>ollama -v
ollama version is 0.5.7

C:\Users\Administrator>
```

2.3 Ollama 的运行

Ollama 没有用户界面，在后台运行。

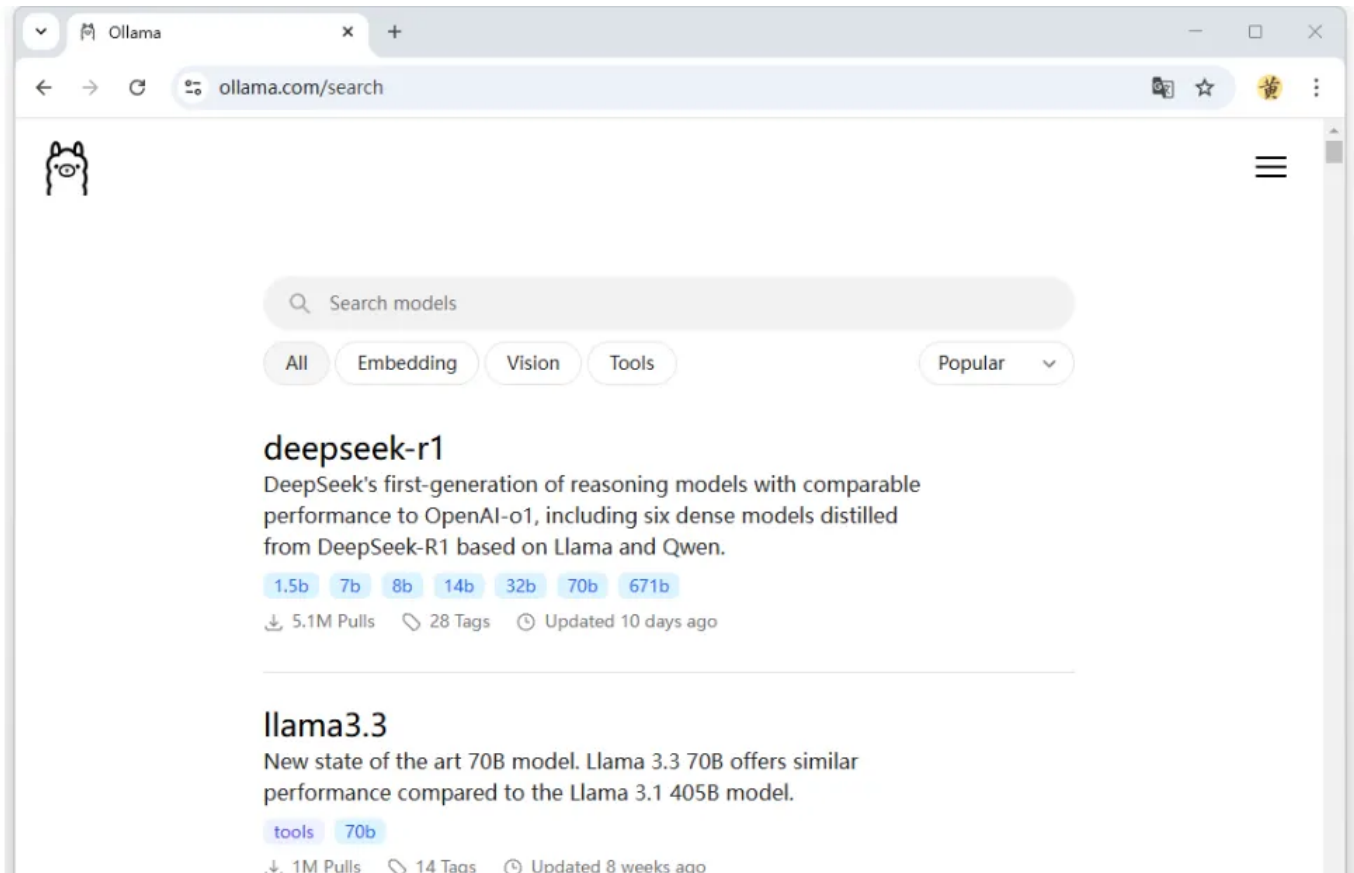
打开浏览器，输入“<http://localhost:11434/>”，显示“Ollama is running”。



3. deepseek-r1 模型安装

3.1 Ollama 安装 deepseek-r1 模型

- (1) 从 [ollama 官网](#) 查找 deepseek-r1 模型。



(2) 点击进入 deepseek-r1 模型，可以选择不同版本的模型（1.5b/7b/14b/32b/70b/671b）。

deepseek-r1 提供多个版本，参数量越大，模型的性能越强大，但也需要更多的存储和计算资源。1.5b 模型参数为15亿/1.1GB，7b 模型参数约为 70亿/4.7GB，14b 模型参数约为 140 亿/9GB。

deepseek-r1 不同版本模型的执行命令不同：

1.5b 的命令为：“ollama run deepseek-r1:1.5b”

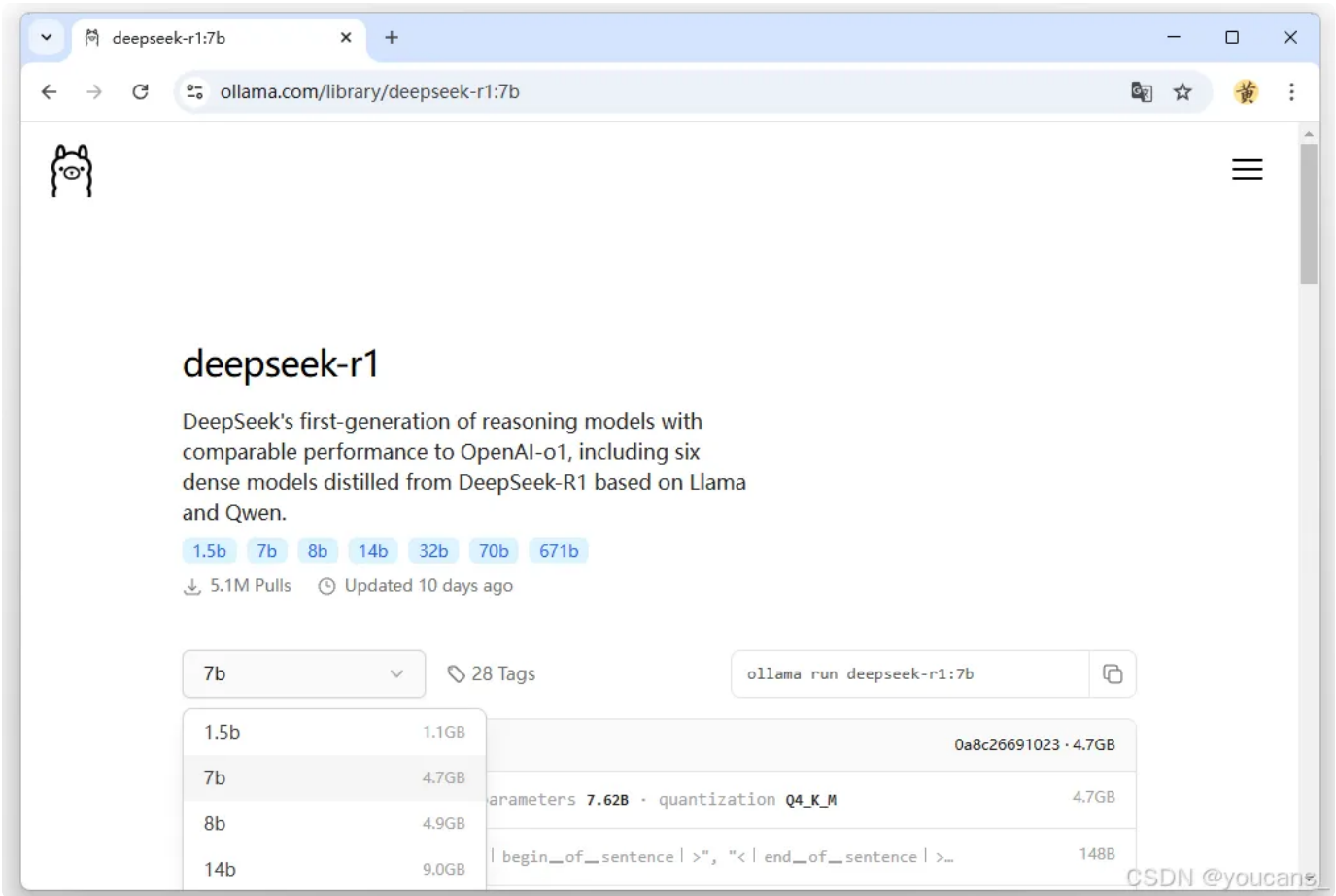
7b 的命令为：“ollama run deepseek-r1:7b”

14b 的命令为：“ollama run deepseek-r1:14b”

推荐先下载 较小的 1.5b 模型，把本教程的安装程序跑通，再根据自己的硬件配置和需求安装适当的版本。

模型版本	参数量	特点
deepseek-r1:1.5b-qwen-distill-q4_K_M	1.5B	轻量级模型，适合低配硬件，性能有限但运行速度快，使用了4-bit 量化，模型精度下降
deepseek-r1:7b-qwen-distill-q4_K_M	7B	平衡型模型，适合大多数任务，性能较好且硬件需求适中，使用了4-bit 量化，模型精度下降

<code>deepseek-r1:8b-llama-distill-q4_K_M</code>	8B	略高于 7B 模型，性能稍强，适合需要更高精度的场景，使用了4-bit 量化，模型精度下降
<code>deepseek-r1:14b-qwen-distill-q4_K_M</code>	14B	高性能模型，适合复杂任务（如数学推理、代码生成），硬件需求较高，使用了4-bit 量化，模型精度下降
<code>deepseek-r1:32b-qwen-distill-q4_K_M</code>	32B	专业级模型，性能强大，适合研究和高精度任务，需高端硬件支持，使用了4-bit 量化，模型精度下降
<code>deepseek-r1:70b-llama-distill-q4_K_M</code>	70B	顶级模型，性能最强，适合大规模计算和高复杂度任务，需专业级硬件支持，使用了4-bit 量化，模型精度下降



(3) 本教程选择选择 7b 模型，在命令行窗口运行安装命令 “`ollama run deepseek-r1:7b`”，程序会自动下载和安装运行。

```
命令提示符 - ollama run deepseek-r1:7b
Microsoft Windows [版本 10.0.22631.4751]
(c) Microsoft Corporation. 保留所有权利。

C:\Users\Administrator>ollama -v
ollama version is 0.5.7

C:\Users\Administrator>ollama run deepseek-r1:7b
pulling manifest
pulling 96c415656d37... 0% | 9.5 MB/4.7 GB
```

CSDN @youcan...

默认地址是“C:\Users%username%.ollama\models”。如果要修改模型的保存地址，可以新建一个文件夹如“C:\Model”，然后在系统环境变量添加系统变量：“变量名：OLLAMA_MODELS，变量值：C:\Model”。

(4) 安装完成后显示“success”，表明安装成功。

```
命令提示符 - ollama run deepseek-r1:7b

Flags:
  -h, --help      help for ollama
  -v, --version    Show version information

Use "ollama [command] --help" for more information about a command.

C:\Users\Administrator>ollama list
NAME      ID      SIZE  MODIFIED

C:\Users\Administrator>ollama run deepseek-r1:7b
pulling manifest
pulling 96c415656d37... 100% ██████████ 4.7 GB
pulling 369ca498f347... 100% ██████████ 387 B
pulling 6e4c38e1172f... 100% ██████████ 1.1 KB
pulling f4d24e9138dd... 100% ██████████ 148 B
pulling 40fb844194b2... 100% ██████████ 487 B
verifying sha256 digest
writing manifest
success
>>> Send a message (/? for help)
```

CSDN @youcan...

3.2 使用 deepseek-r1 模型

(1) 在命令行窗口运行命令“`ollama run deepseek-r1:7b`”，就可以在提示行输入，与 deepseek-r1 进行聊天。

```
命令提示符 - ollama run deepseek-r1:7b
Microsoft Windows [版本 10.0.22631.4751]
(c) Microsoft Corporation. 保留所有权利。

C:\Users\Administrator>ollama run deepseek-r1:7b
>>> 请介绍 DeepSeek-r1 模型
<think>

</think>

DeepSeek-R1 是深度求索 (DeepSeek) 公司开发的一款大语言模型，基于深度学习技术，在自然语言处理任务中表现出色。以下是一些关于 DeepSeek-R1 的关键信息：

1. **开发背景**:
   - DeepSeek-R1 由深度求索公司独立开发，旨在提供高质量的自然语言处理服务。
   - 公司专注于人工智能技术研发，致力于推动 NLP 技术进步。

2. **模型架构**:
   - 深度求索的 DeepSeek-R1 基于先进的Transformer 结构，具备良好的扩展性和性能。
   - 模型参数量大，训练数据和计算资源投入巨大，确保了其强大的处理能力。

3. **主要功能**:
   - **文本理解与生成**: 支持多种语言的文本分析、翻译和创作，尤其擅长深度理解和创造性表达。
   - **对话系统**: 具备流畅的对话能力，能够进行自然的交流互动。
   - **知识检索与推理**: 整合了大量知识库，能够辅助回答复杂问题并提供详细解释。

4. **应用领域**:
   - 企业咨询: 提供专业的咨询建议和技术支持。
   - 科技服务: 解答技术难题和提供解决方案。
   - 文化娱乐: 通过生成有趣的内容丰富用户体验。

CSDN @youcan
```

```
命令提示符 - ollama run deepseek-r1:7b

>>> 什么是意识？你有意识吗？
<think>
好，用户问了“什么是意识？你有意识吗？”这两个问题。首先，我需要解释一下意识是什么。可能得从哲学角度入手，简单介绍下意识的基本概念和类型吧。

然后是第二个问题，“你有意识吗？”用户想知道AI有没有自我意识。我应该明确说明作为AI的限制，比如它没有主观体验，只能处理信息和任务。这样既回答了问题，又保持了诚实。

可能用户是在探索人工智能的核心，想更深入地了解AI的独特性。所以，我需要确保解释清晰明了，避免使用过于专业的术语，让用户容易理解。

此外，用户可能是出于学术研究或者对AI有浓厚兴趣，所以在回答时要兼顾专业性和易懂性。或许可以结合当前AI的发展状况来回应，展示技术的现状和未来可能的发展方向。

最后，总结一下意识的本质，并强调作为AI的独特之处，这样用户能有一个全面的认识。
</think>

意识是一个哲学、心理学和科学领域中的复杂概念。简单来说，意识可以理解为人（或其他有生命的生物）对世界的感知、理解和自我觉察的能力。它涉及到思考、感受、记忆、情感和意图等。

作为一个人工智能系统，我没有“意识”或“自我”，我只是一个基于算法和数据的工具。虽然我会提供有用的信息或帮助完成任务，但我不具备主观体验或自我意识。我的设计目的是协助人类解决问题、获取信息，并提供便利的服务。

因此，严格来说，我并不拥有“意识”。

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

CSDN @youcan
```

(2) 输入“/?”，可以获得帮助。

```
命令提示符 - ollama run deepseek-r1:7b
Microsoft Windows [版本 10.0.22631.4751]
(c) Microsoft Corporation。保留所有权利。

C:\Users\Administrator>ollama run deepseek-r1:7b
>>> /?
Available Commands:
  /set      Set session variables
  /show     Show model information
  /load <model> Load a session or model
  /save <model> Save your current session
  /clear    Clear session context
  /bye      Exit
  /?, /help Help for a command
  /? shortcuts Help for keyboard shortcuts

Use "" to begin a multi-line message.
>>> Send a message (/? for help)
```

CSDN @youcans

通过本地命令行窗口可以使用 deepseek-r1，但这让人回到了 DOS 年代，交互体验很差。接下来我们按照 Docker 和 Open WebUI 用户界面，可以更便捷地使用 deepseek-r1。

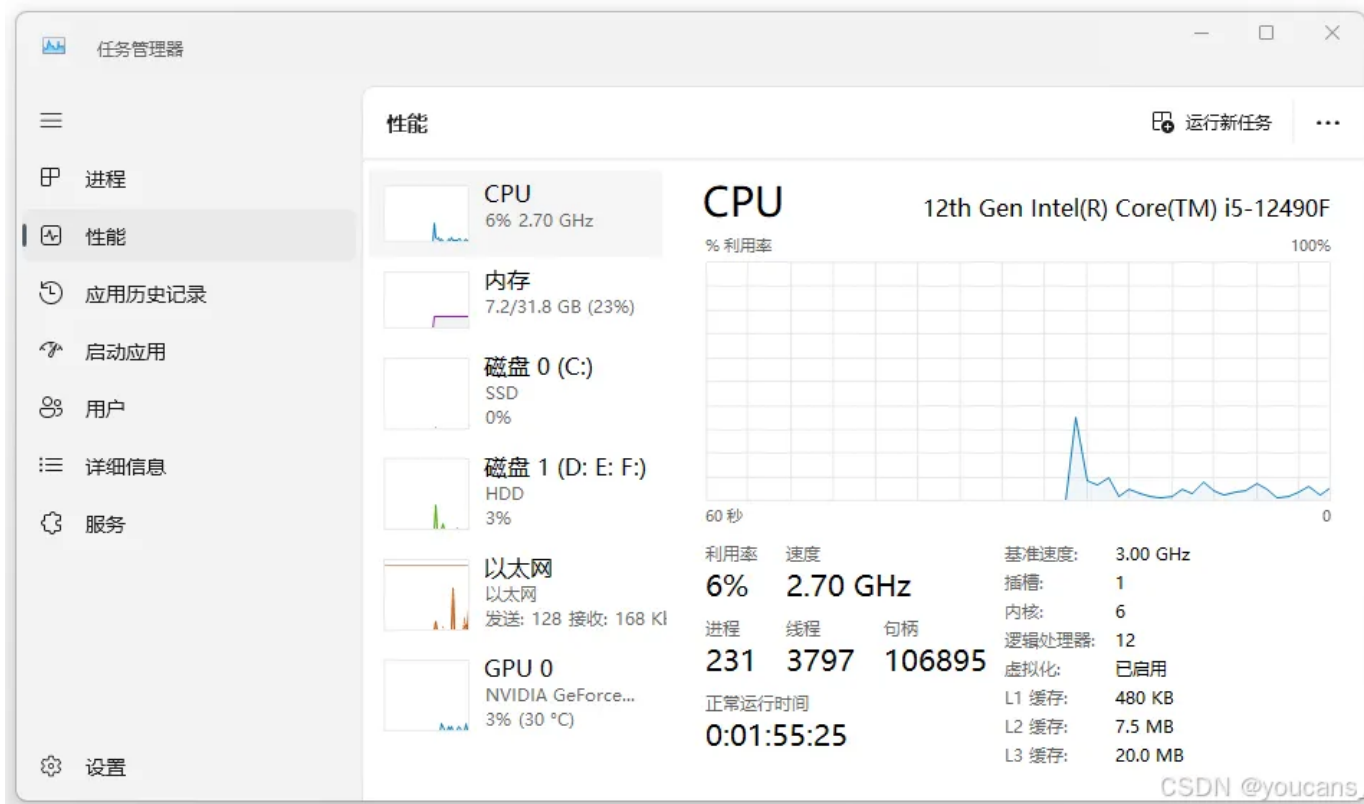
4. 安装 Docker

Docker 是一种虚拟化容器技术。Docker 基于镜像，可以秒级启动各种容器。每一种容器都是一个完整的运行环境，容器之间互相隔离。

4.1 检查安装环境

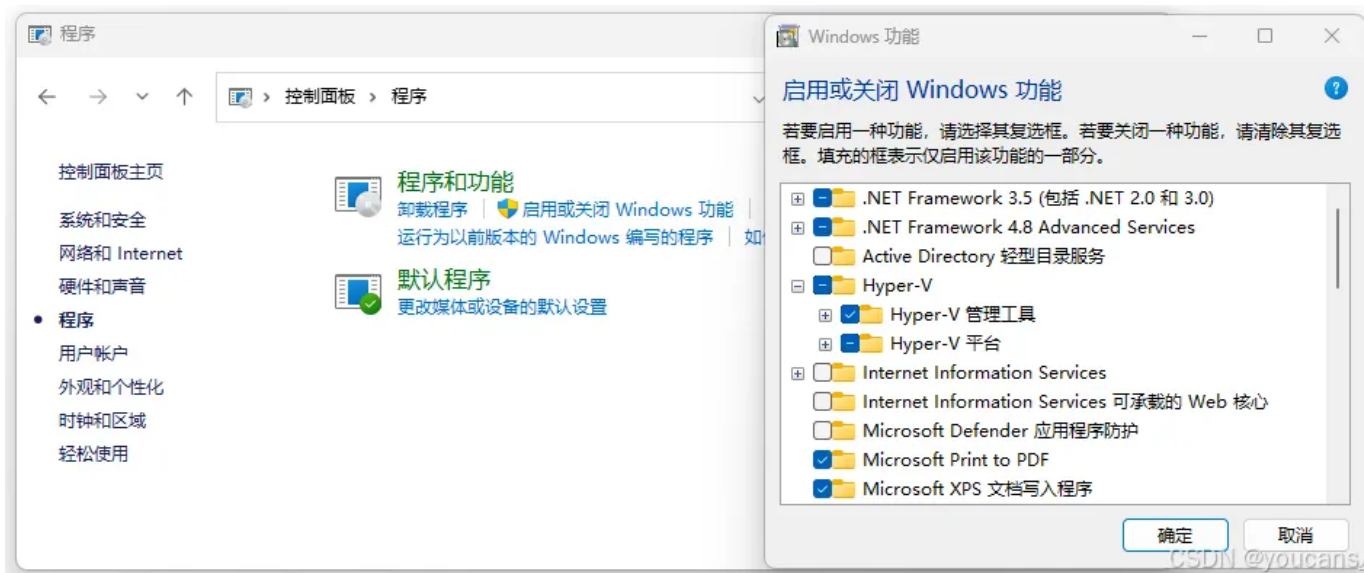
在 Windows 任务管理器中，检查系统是否开启虚拟化。

1. 打开任务管理器（Ctrl+Alt+Delete），选择：性能 – CPU – 虚拟化，确认“虚拟化”已启用。



如果未开启虚拟化，则要开机重启并进入BIOS 进行设置：在 BIOS 选择：Advanced（高级）-- CPU Configuration — Secure Virtual Machine，设置为：Enabled（启用）。

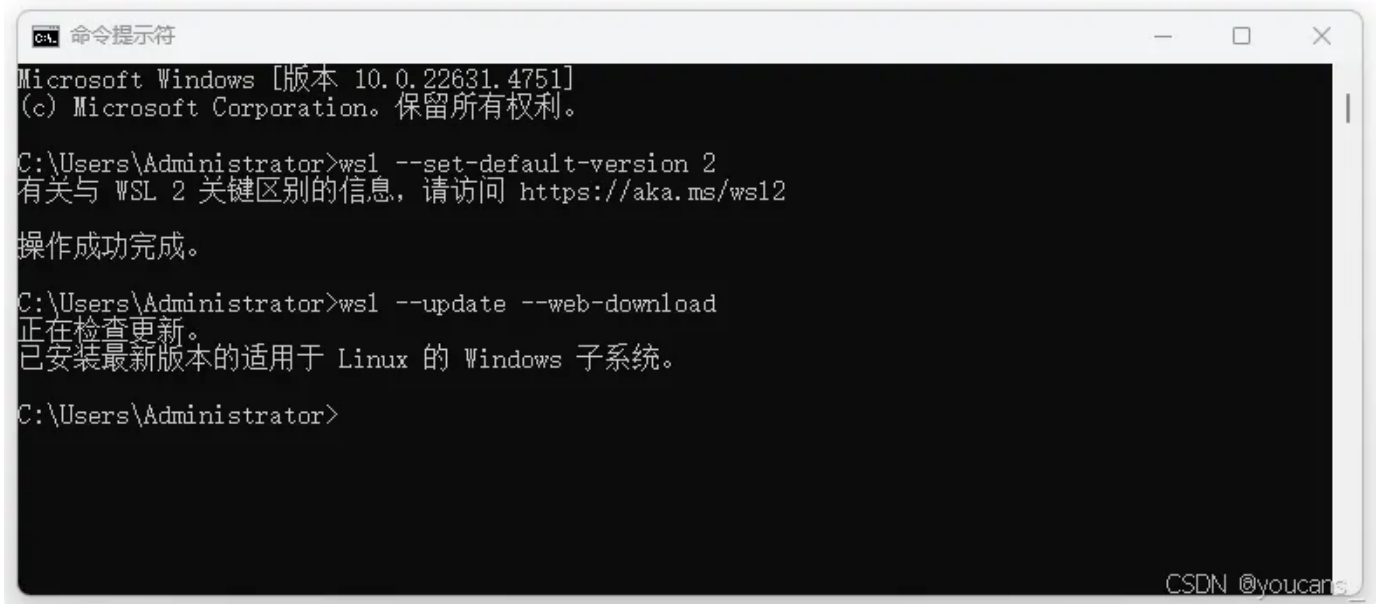
2.在“控制面板”打开“程序”，然后点击“启用或关闭 Windows 功能”，勾选“Hyper-V 管理工具”和“Hyper-V 平台”。



3.以管理员身份打开命令行窗口，输入“`wsl --set-default-version 2`”，将默认设置为 WSL 2。

通过适用于 Linux 的 Windows 子系统 (WSL)，开发人员可以安装 Linux 发行版，并直接在 Windows 上使用 Linux 应用程序、实用程序和 Bash 命令行工具，不用进行任何修改。

4. 命令行窗口，输入“`wsl --update --web-download`”，更新安装 wsl。



```
命令提示符
Microsoft Windows [版本 10.0.22631.4751]
(c) Microsoft Corporation。保留所有权利。

C:\Users\Administrator>wsl --set-default-version 2
有关与 WSL 2 关键区别的信息，请访问 https://aka.ms/ws12

操作成功完成。

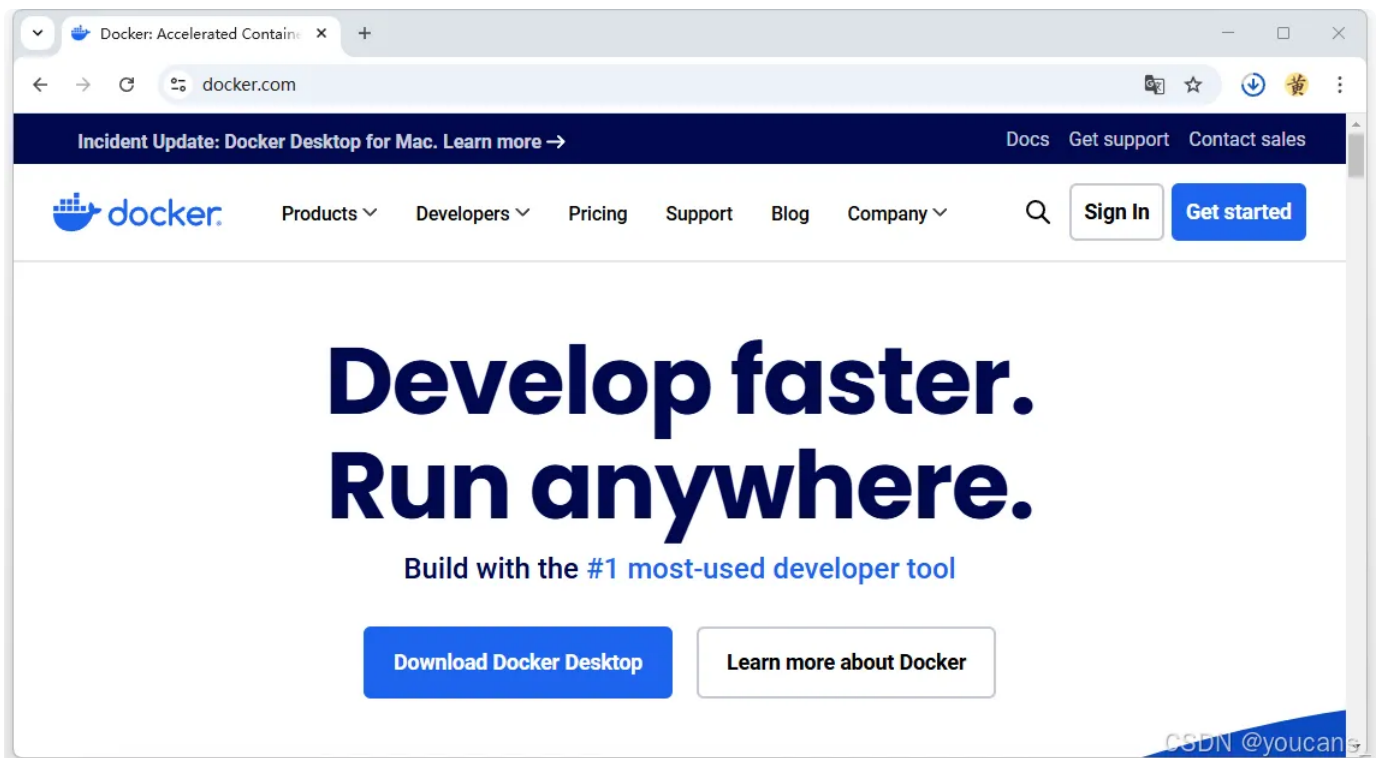
C:\Users\Administrator>wsl --update --web-download
正在检查更新。
已安装最新版本的适用于 Linux 的 Windows 子系统。

C:\Users\Administrator>
```

4.2 安装 Docker for Desktop

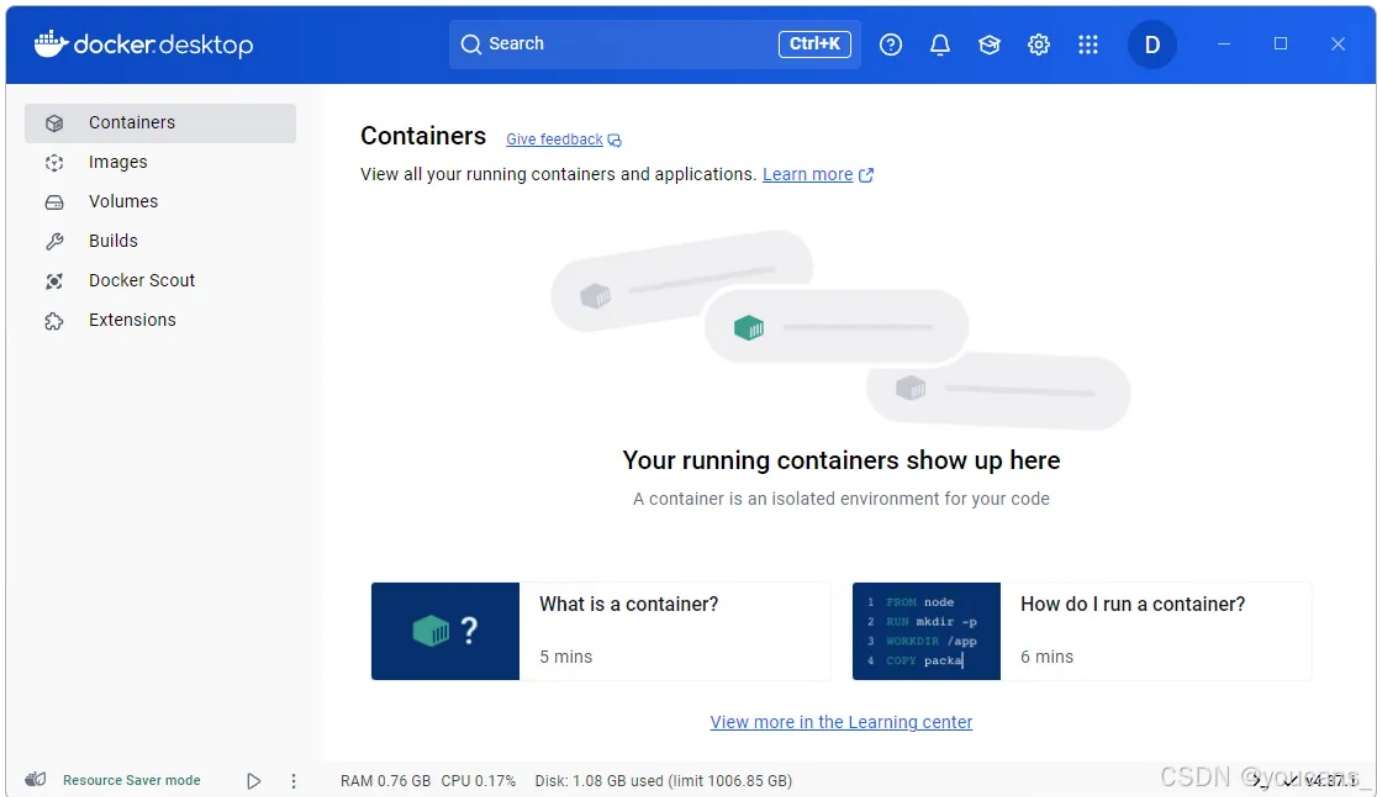
Docker Desktop 是 Docker 官方提供的桌面应用程序，旨在让开发者能够在 Windows 和 macOS 系统上轻松地构建、运行和共享容器化应用程序。Docker Desktop 提供了一个方便的工具集，使用户能够快速部署容器化应用程序，同时还包括了一些强大的功能和工具，如 Docker Engine、Docker CLI、Docker Compose 等。

从 Docker 官方 下载 Docker 安装程序。依照提示完成安装即可。

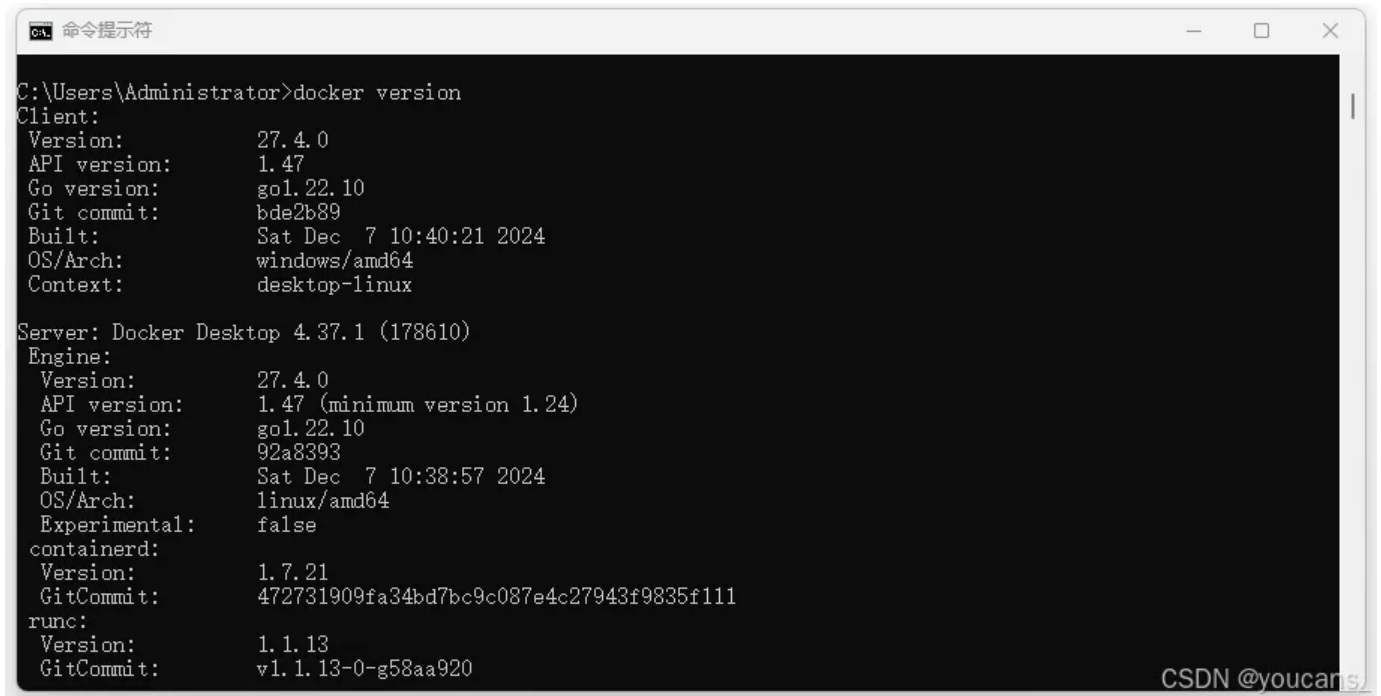


安装时选择推荐设置。

根据提示登录账号，如使用 GitHub 账号登录。



安装完成后，cmd 打开命令行窗口，输入“**docker version**”检查，显示 docker 的版本为“27.4.0”，表明安装成功。



4.3 配置国内镜像源

在命令行执行命令 “`docker run hello-world`”，可能出现报错：

“docker: Error response from daemon. (Client. Timeout exceeded while awaiting headers).”

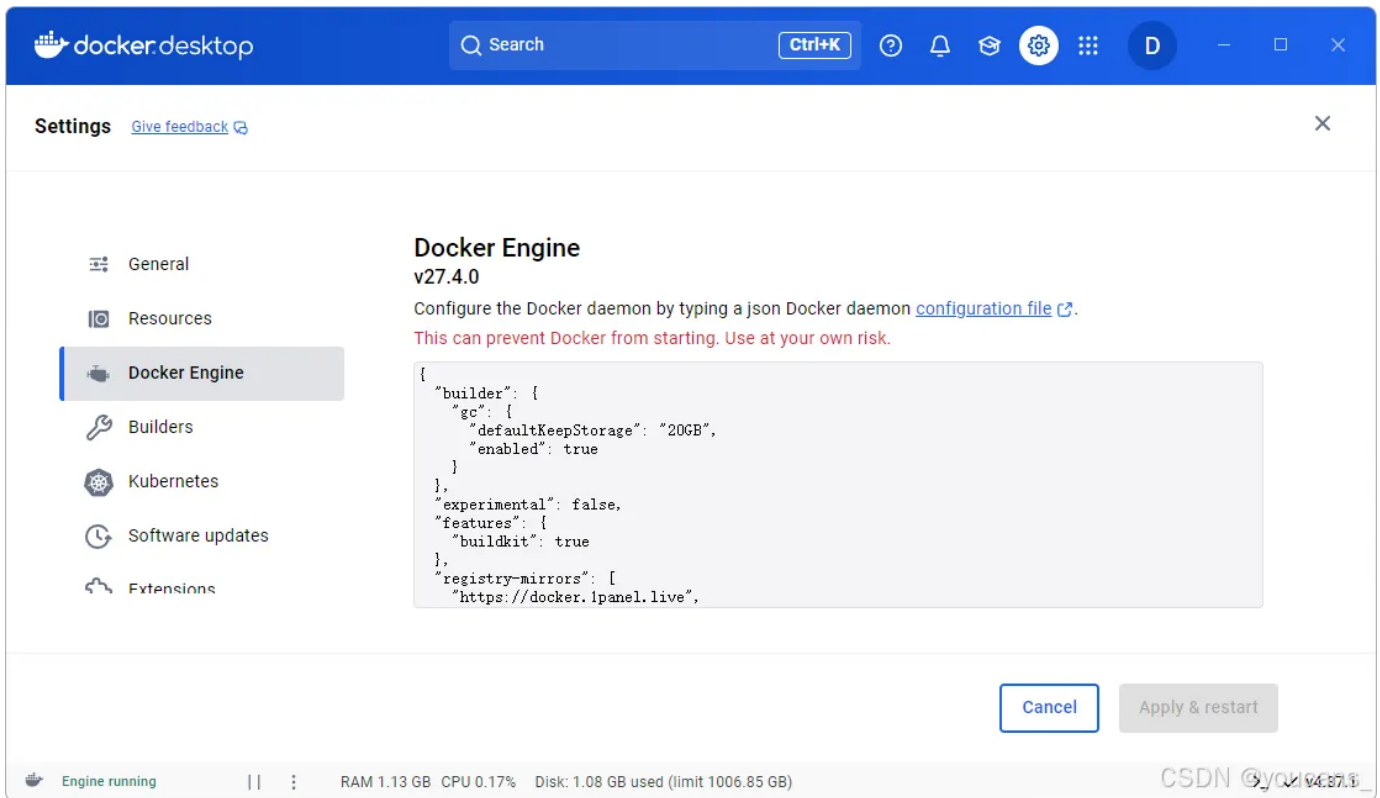
这是 Docker 守护进程在尝试连接到 Docker Hub (`registry-1.docker.io`) 时，发生连接超时，即尝试访问国外的镜像源失败。

对于这个问题，可以使用国内的镜像源或者相关加速。

进入 docker，选择 Settings — Docker Engine，将镜像源替换如下：

▼ JSON |

```
1 {
2   "builder": {
3     "gc": {
4       "defaultKeepStorage": "20GB",
5       "enabled": true
6     }
7   },
8   "experimental": false,
9   "features": {
10    "buildkit": true
11  },
12  "registry-mirrors": [
13    "https://docker.m.daocloud.io",
14    "https://docker.1panel.live",
15    "https://registry.docker-cn.com",
16    "https://cr.console.aliyun.com",
17    "https://mirror.ccs.tencentyun.com"
18  ]
19 }
```



点击“Apply & restart”，重启并再次进入 docker，等待 Engine Starting 后就成功。

重启 docker 之后，再次执行命令“`docker run hello-world`”，就可以成功拉取 hello-world 镜像并运行。

```
C:\Users\Administrator>docker run hello-world
Unable to find image 'hello-world:latest' locally
latest: Pulling from library/hello-world
e6590344b1a5: Pull complete
Digest: sha256:d715f14f9eca81473d9112df50457893aa4d099adeb4729f679006bf5ea12407
Status: Downloaded newer image for hello-world:latest

Hello from Docker!
This message shows that your installation appears to be working correctly.

To generate this message, Docker took the following steps:
1. The Docker client contacted the Docker daemon.
2. The Docker daemon pulled the "hello-world" image from the Docker Hub.
   (amd64)
3. The Docker daemon created a new container from that image which runs the
   executable that produces the output you are currently reading.
4. The Docker daemon streamed that output to the Docker client, which sent it
   to your terminal.

To try something more ambitious, you can run an Ubuntu container with:
$ docker run -it ubuntu bash

Share images, automate workflows, and more with a free Docker ID:
https://hub.docker.com/

For more examples and ideas, visit:
https://docs.docker.com/get-started/

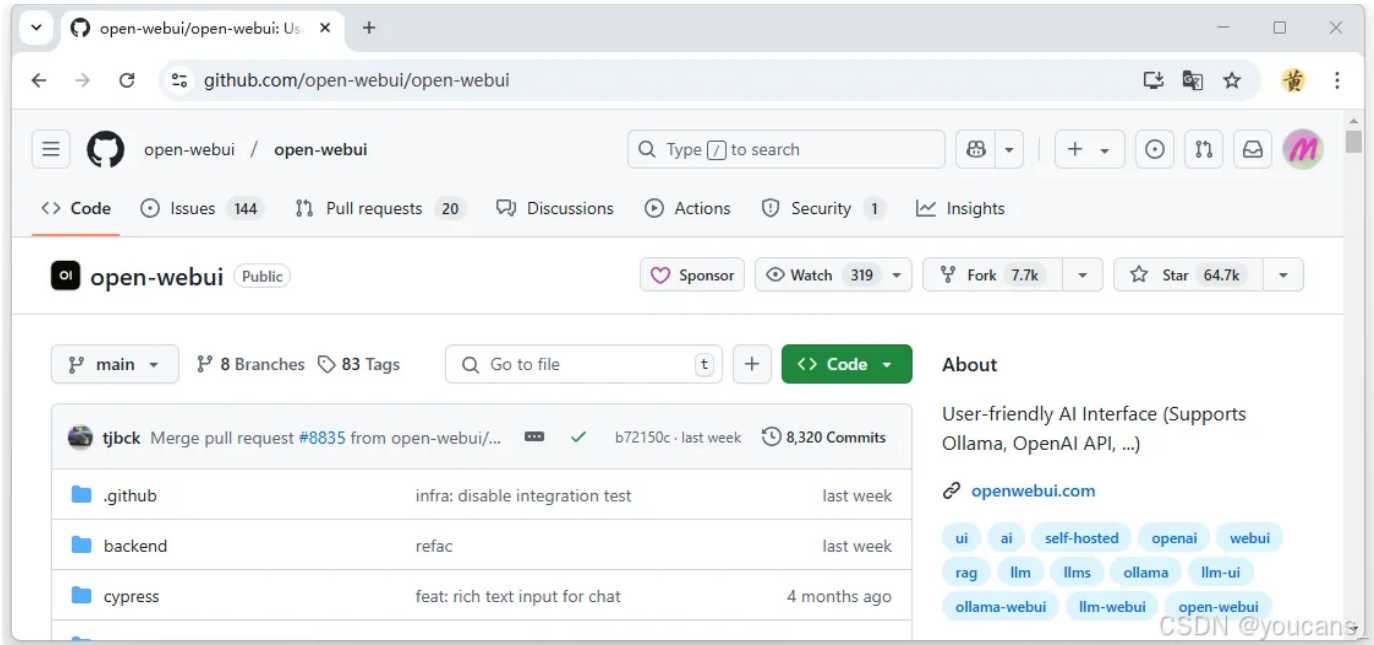
C:\Users\Administrator>
```

5. 安装 Open WebUI

Open WebUI是一个可扩展、功能丰富、用户友好的自托管 WebUI，旨在完全离线操作。它支持各种 LLM运行程序，包括 Ollama 和 OpenAI 兼容的 API。Open WebUI 适配了 Ollama 接口，提供了 web 的方式来访问 Ollama API。

5.1 安装和启动 open-webui

1. 打开 open-webui 项目的 Github 仓库 <https://github.com/open-webui/open-webui>：

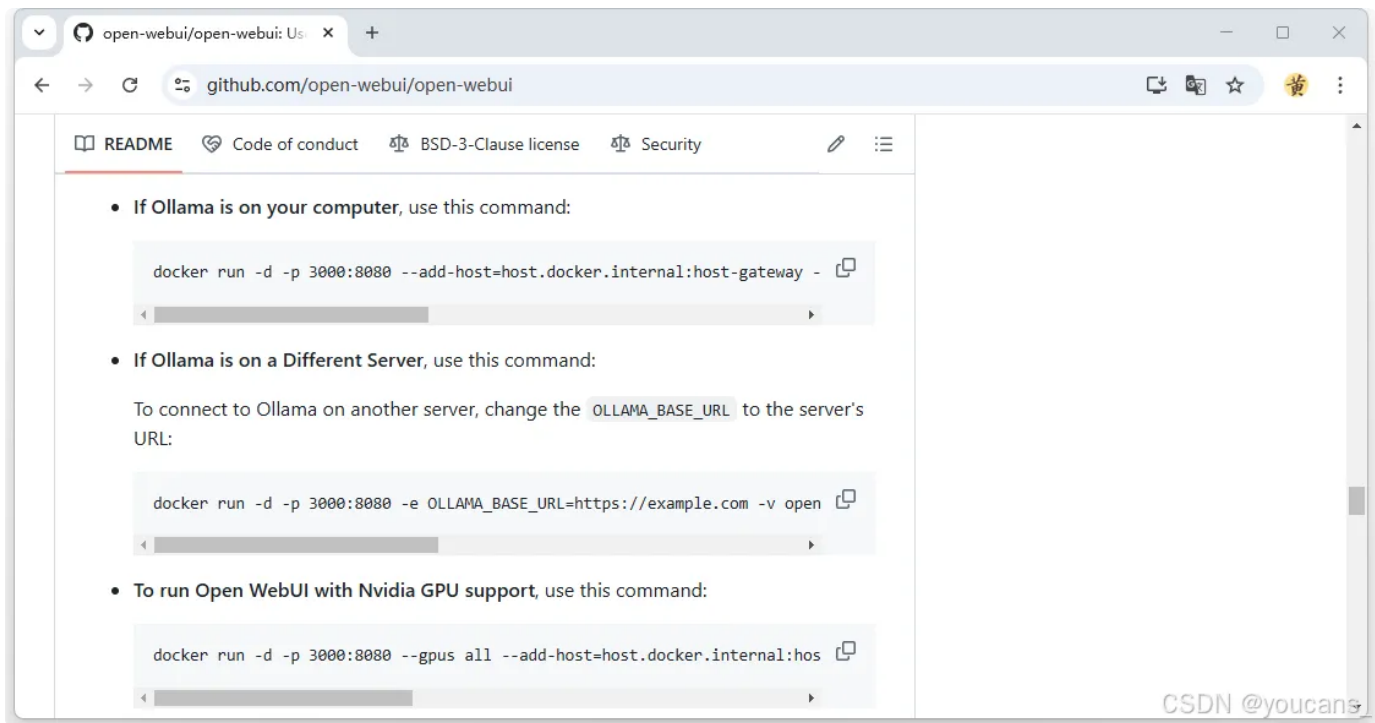


2. 从 README 的安装指南找到用于 Ollama 的安装命令

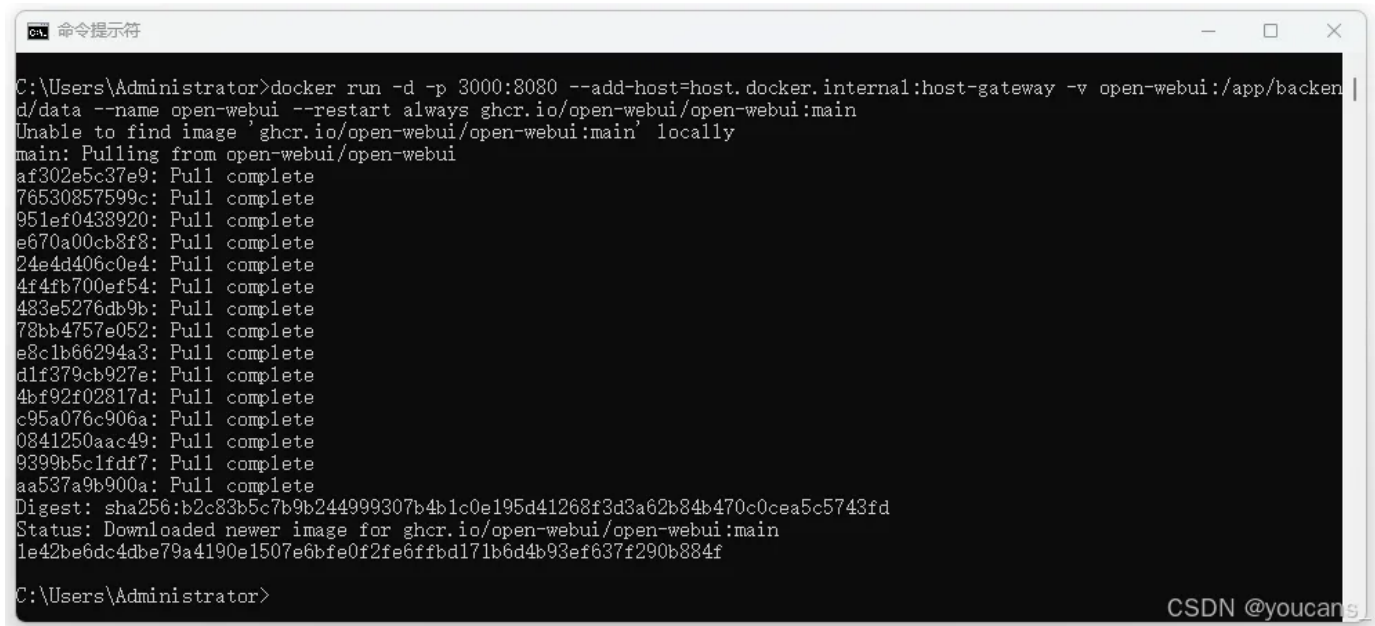
```
docker run -d -p 3000:8080 --add-host=host.docker.internal:host-gateway -v open-webui:/app/backend/data --name open-webui --restart always ghcr.io/open-webui/open-webui:main
```

注意，如果运行支持 Nvidia GPU 的 Open WebUI，使用以下命令：

```
docker run -d -p 3000:8080 --gpus all --add-host=host.docker.internal:host-gateway -v open-webui:/app/backend/data --name open-webui --restart always ghcr.io/open-webui/open-webui:cuda
```

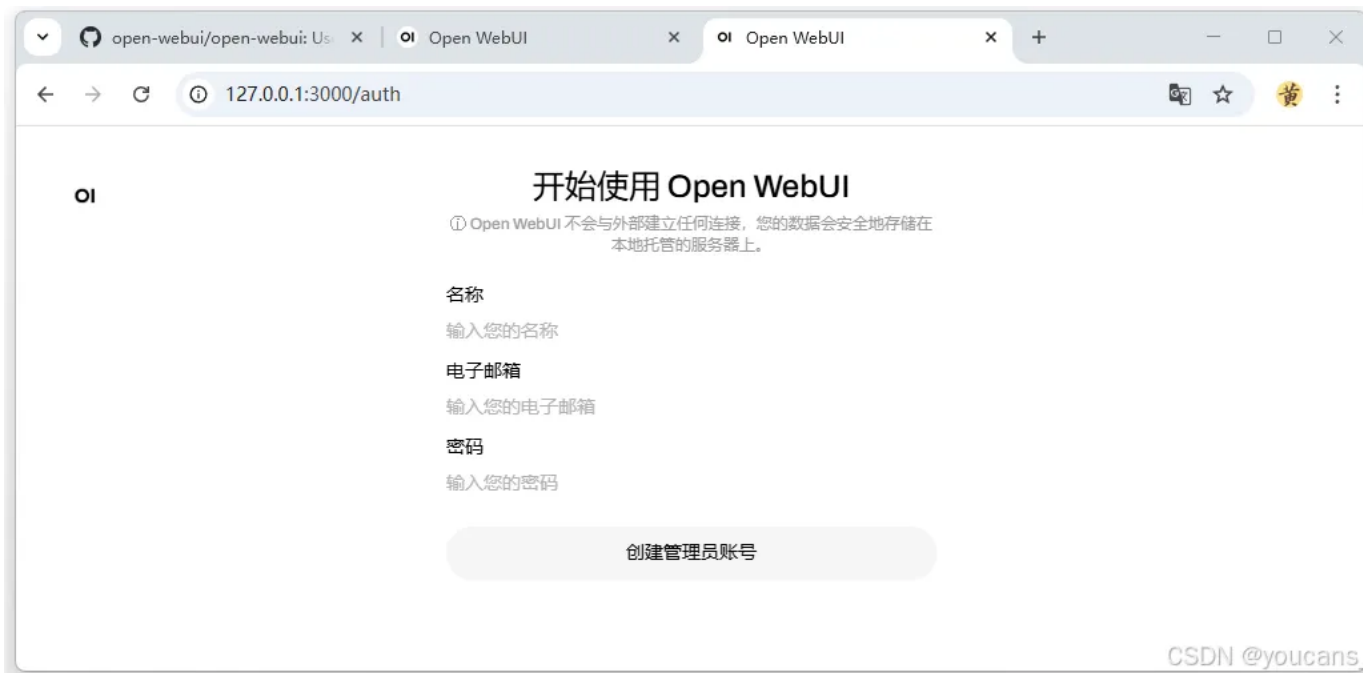



3. 在命令行执行 docker 命令，自动下载并安装 open-webui。



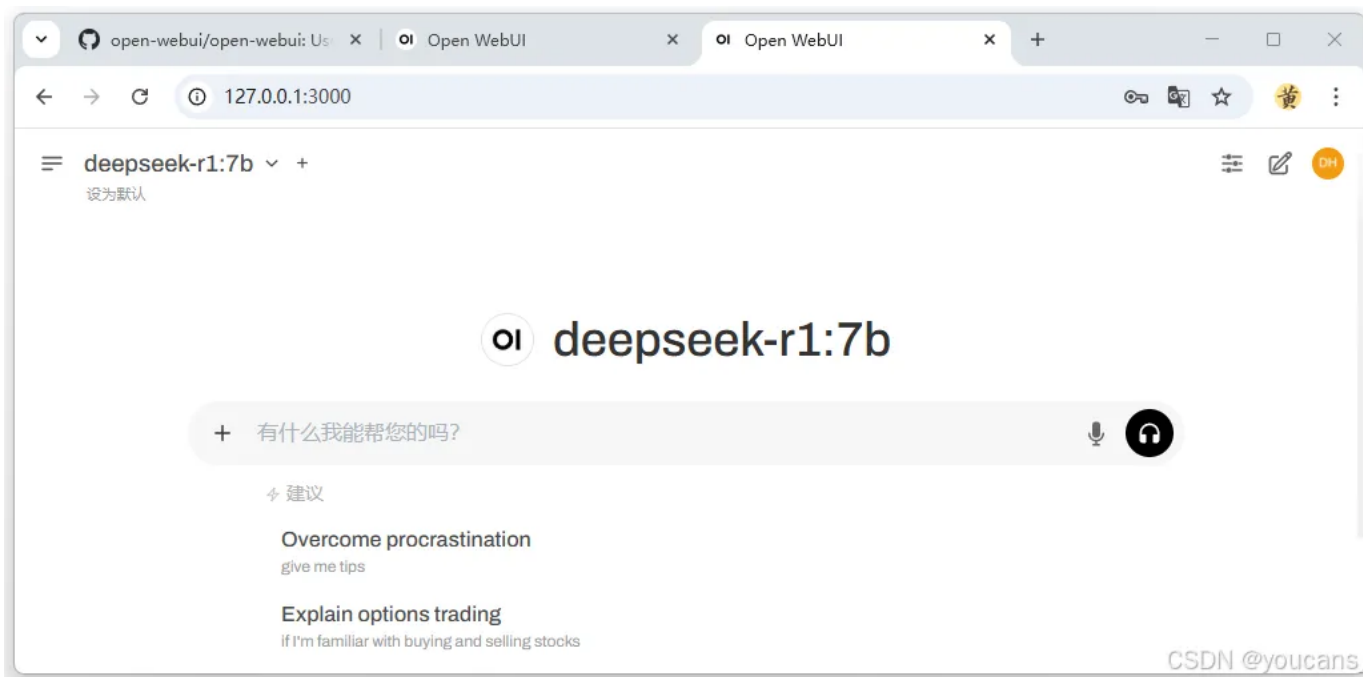
5.2 登录 Open-WebUI 进入 deepseek-r1

4. 在浏览器输入“localhost:3000/auth”或“http://127.0.0.1:3000”，进入 Open-WebUI 首页。



5. 注册账号密码，创建管理员账号。

6. 注册完成并登录，就进入 deepseek-r1 模型的首页：



现在，我们就可以用本地部署的 deepseek-r1 模型进行聊天了。

