DeepSeek本地部署与应用构建

智灵动力　陈军

目录

DeepSeek 学习手册（持续更新ing..）

学习地址：https://yunyinghui.feishu.cn/wiki/QuK2wYgOXiH3m6k1tADcxuhjnEg

DeepSeek简单介绍与使用

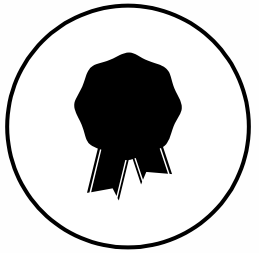
模型简介

推理能力强

DeepSeek R1 推理模型具备强大的推理能力，能够准确理解并回应复杂的对话场景，支持多轮对话，确保用户体验的连贯性和高效性。

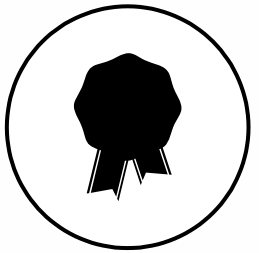
本地化部署 隐私保护

核心亮点在支持完全本地化部署，有效保护用户数据隐私，避免敏感信息泄露，同时提升推理速度与安全性。



多种量化蒸馏模型

提供多种量化蒸馏模型选择，包括8B、32B、70B等，以满足不同应用场景对精度与性能的多样化需求。



开源生态

开源社区支持，并兼容多种开源框架，方便我们二次开发和微调。同时也支持跨平台适配。

应用场景

科研数据分析

DeepSeek R1 在科研领域展现出了巨大潜力，通过高效处理和分析复杂数据，为科研人员提供深入见解，加速科研进展。

自动化工作流

集成DeepSeek R1 于自动化工作流，显著提升流程智能化水平，自动处理数据，实现高效、准确的决策支持。

外挂大脑

日常答疑解惑，信息收集总结类，料汇总，策划分析写周报，写作直播的话术

DeepSeek 使用途径

官网/APP

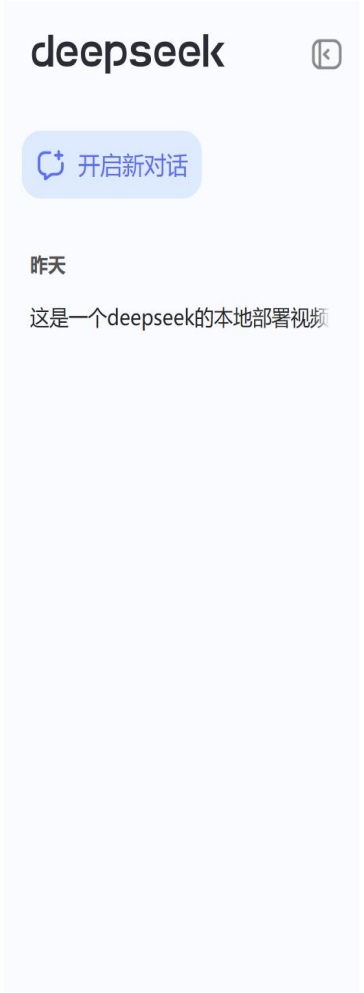
硅基流动

秘塔

cursor

Grok

本地部署/API

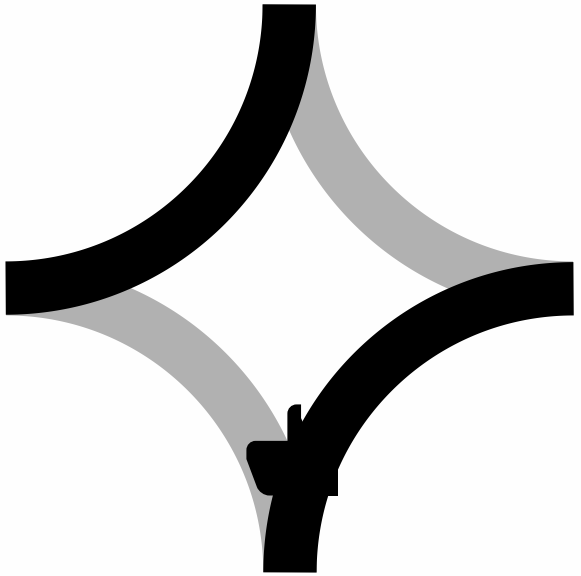


DeepSeek本地部署

本地部署的必要性

数据隐私保护

DeepSeek R1 推理模型本地部署可有效避免敏感数据在推理过程中上传至云端，确保数据隐私安全。



离线使用

即使在网络断开的情况下，用户仍可依赖本地部署的DeepSeek R1 推理模型进行智能分析，保障工作连续性。

自定义模型

数据弱审查

根据特定需求，用户可灵活选择不同量化精度的模型进行本地部署，实现性能与资源利用的最佳平衡。可以弱化审核条件，更加全面的利用大模型能力

性能优化

本地部署DeepSeek R1 推理模型能够充分挖掘并利用本地硬件资源，如CPU、GPU等，实现推理性能的优化提升。

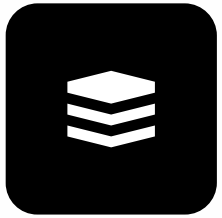
软件要求

操作系统

DeepSeek支持Windows及Linux、mac操作系统，确保了其在不同平台上的兼容性，为用户提供灵活的选择。

安装包

部署DeepSeek需下载其本地部署包，该包已包含所有必要的依赖项和配置文件，确保了安装过程的简便性。



防火墙设置

为确保DeepSeek在断网环境下能够正常运行，需进行防火墙设置，禁止不必要的网络通信，保障数据安全。

模型选择与硬件要求

量化模型选择

DeepSeek支持8B、32B、70B等多种量化模型，官方满配版本是671B。用户可根据实际需求及硬件配置选择合适的模型。

量化模型作用

量化模型的选择直接影响模型的推理速度与精度，用户需根据具体任务权衡利弊，做出最佳选择。

CPU ->1.5B Q8或者 8B Q4

GPU 4G -> 8B Q4 推理

GPU 8G-16G ->32B Q4推理 显存越大，速度越快，达到官方宣传的官方版本的90%能力，效果也不错。

GPU 24G -> 32G Q8或者70B Q2

GPU 40G ->70B Q4 这个效果就非常好了

Ollama方式安装

DeepSeek的安装直接用ollama就能安装，

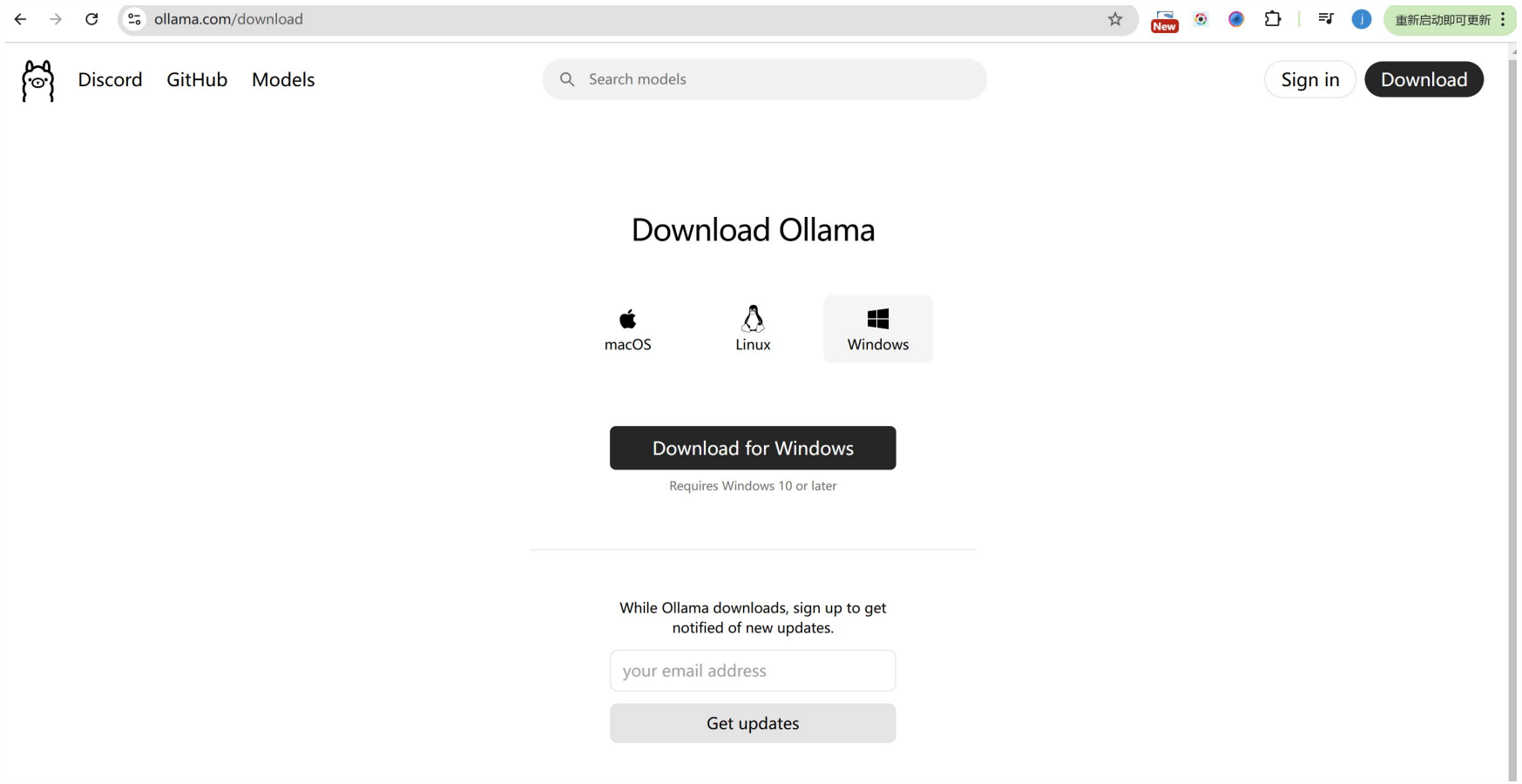
ollama官方地址：https://ollama.com

ollama+chatboxai

ollama+anythinglm

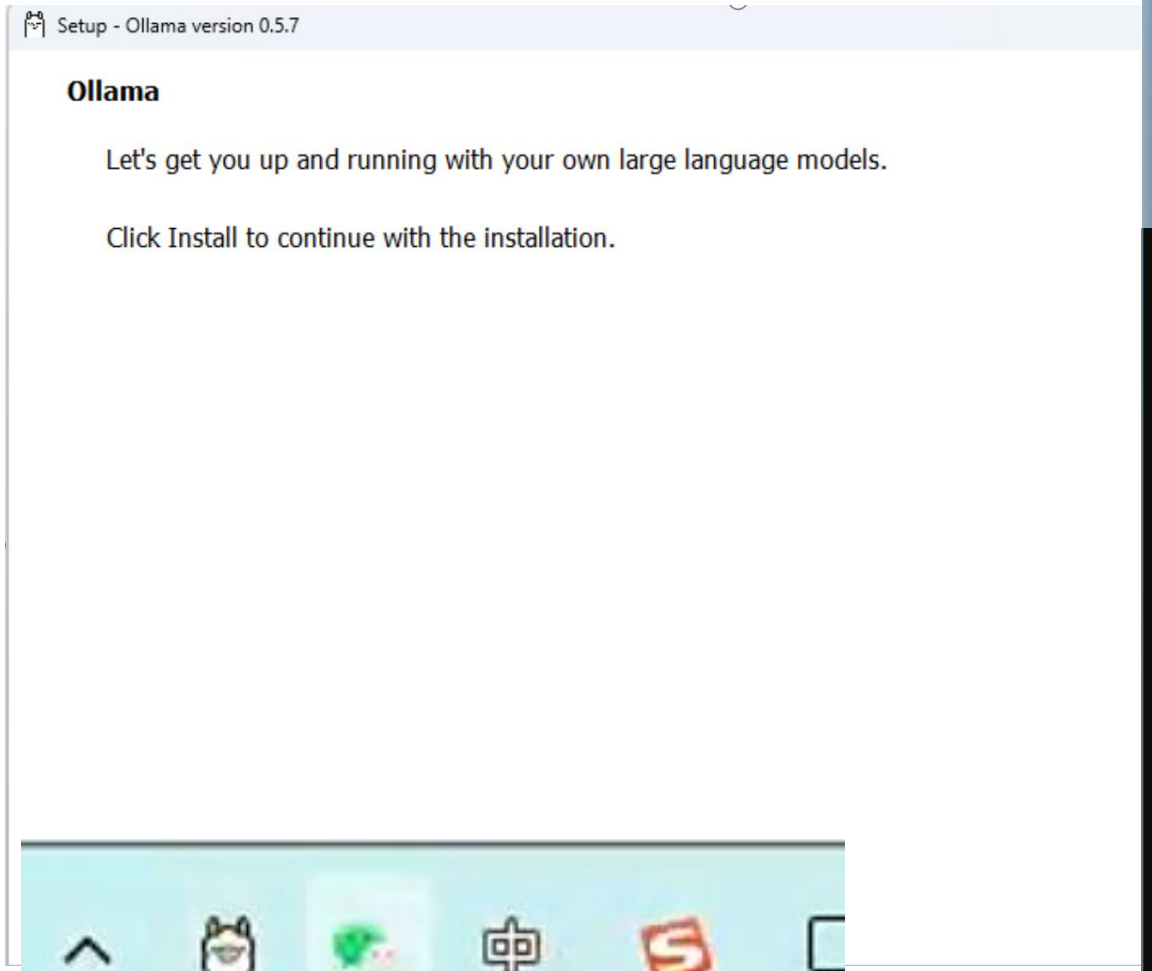
ollama+page assist

Ollama方式安装

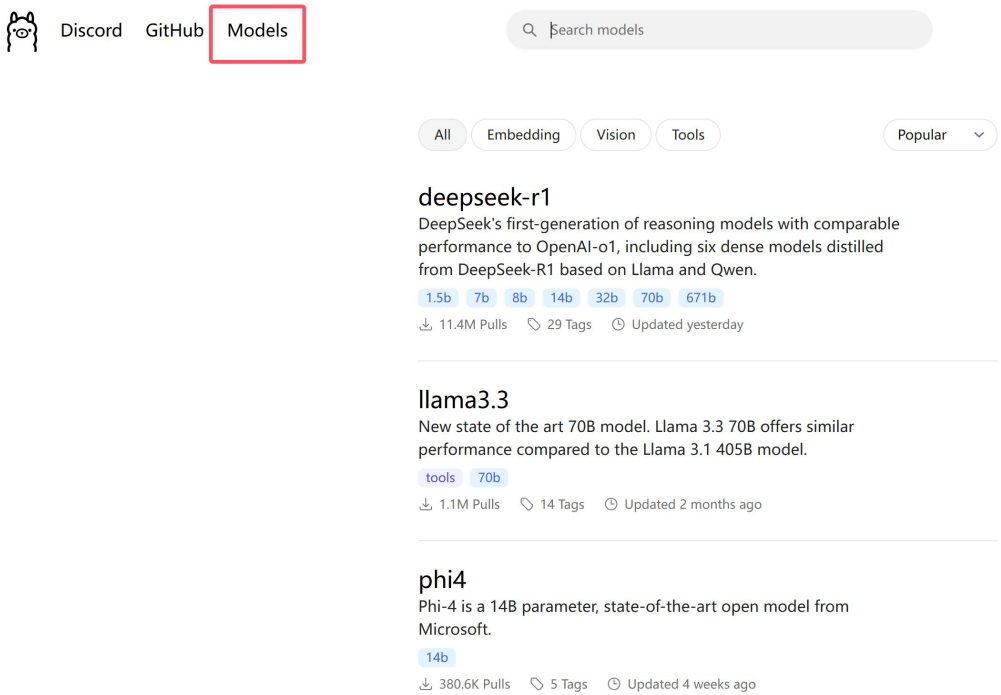

ollama.com/download
重新启动即可更新
Discord
GitHub
Models
Search models
Sign in
Download
Download Ollama
macOS
Linux
Windows
Download for Windows
Requires Windows 10 or later
While Ollama downloads, sign up to get notified of new updates.
your email address
Get updates

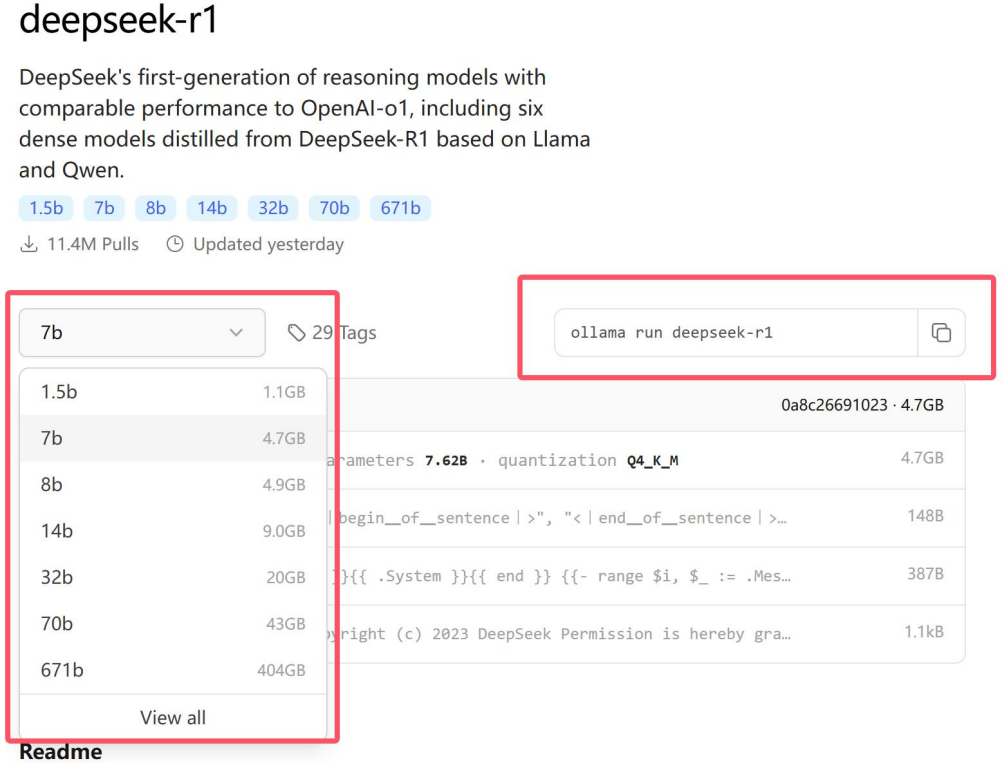
Ollama方式安装

安装软件

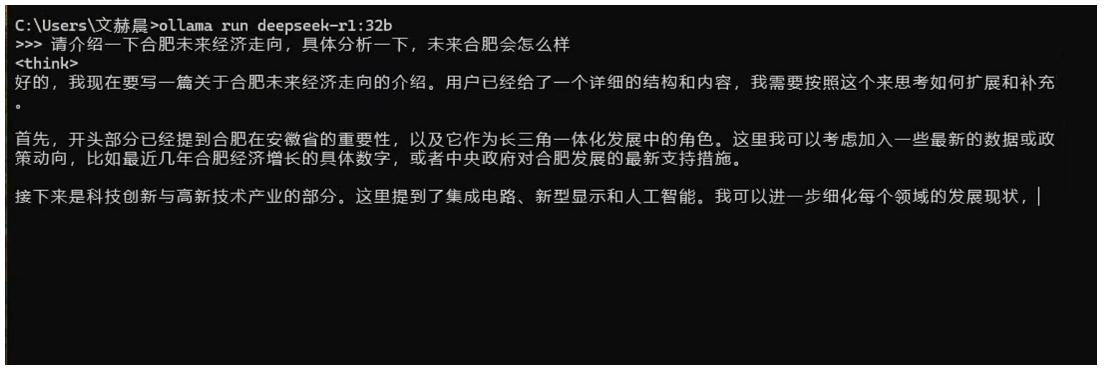


Ollama方式安装


Discord
GitHub
Models
Search models
All
Embedding
Vision
Tools
Popular
deepseek-r1
DeepSeek's first-generation of reasoning models with comparable performance to OpenAI-o1, including six dense models distilled from DeepSeek-R1 based on Llama and Qwen.
1.5b 7b 8b 14b 32b 70b 671b
11.4M Pulls 29 Tags Updated yesterday
llama3.3
New state of the art 70B model. Llama 3.3 70B offers similar performance compared to the Llama 3.1 405B model.
tools 70b
1.1M Pulls 14 Tags Updated 2 months ago
phi4
Phi-4 is a 14B parameter, state-of-the-art open model from Microsoft.
14b
380.6K Pulls 5 Tags Updated 4 weeks ago


deepseek-r1
DeepSeek's first-generation of reasoning models with comparable performance to OpenAI-o1, including six dense models distilled from DeepSeek-R1 based on Llama and Qwen.
1.5b 7b 8b 14b 32b 70b 671b
11.4M Pulls Updated yesterday
7b
29 Tags
ollama run deepseek-r1
1.5b 1.1GB
7b 4.7GB
8b 4.9GB
14b 9.0GB
32b 20GB
70b 43GB
671b 404GB
View all
0a8c26691023 · 4.7GB
arameters 7.62B · quantization Q4_K_M 4.7GB
|begin__of__sentence|>", "<|end__of__sentence|>… 148B
}}{{ .System }}{{ end }} {{- range $i, $_ := .Mes… 387B
yright (c) 2023 DeepSeek Permission is hereby gra… 1.1kB
Readme


C:\Users\文赫晨>ollama run deepseek-r1:8b
pulling manifest
pulling 6340dc3229b0... 23% 1.1 GB/4.9 GB 103 KB/s 10h14m


C:\Users\文赫晨>ollama run deepseek-r1:32b
>>> 请介绍一下合肥未来经济走向，具体分析一下，未来合肥会怎么样
<think>
好的，我现在要写一篇关于合肥未来经济走向的介绍。用户已经给了一个详细的结构和内容，我需要按照这个来思考如何扩展和补充
。
首先，开头部分已经提到合肥在安徽省的重要性，以及它作为长三角一体化发展中的角色。这里我可以考虑加入一些最新的数据或政
策动向，比如最近几年合肥经济增长的具体数字，或者中央政府对合肥发展的最新支持措施。
接下来是科技创新与高新技术产业的部分。这里提到了集成电路、新型显示和人工智能。我可以进一步细化每个领域的发展现状，

Chatbox对话框页面

Chatbox官网： https://chatboxai.app/en

Chatbox AI　首页　下载　定价　帮助　Stars 29k　Chatbox | Product Hunt　文A 中文

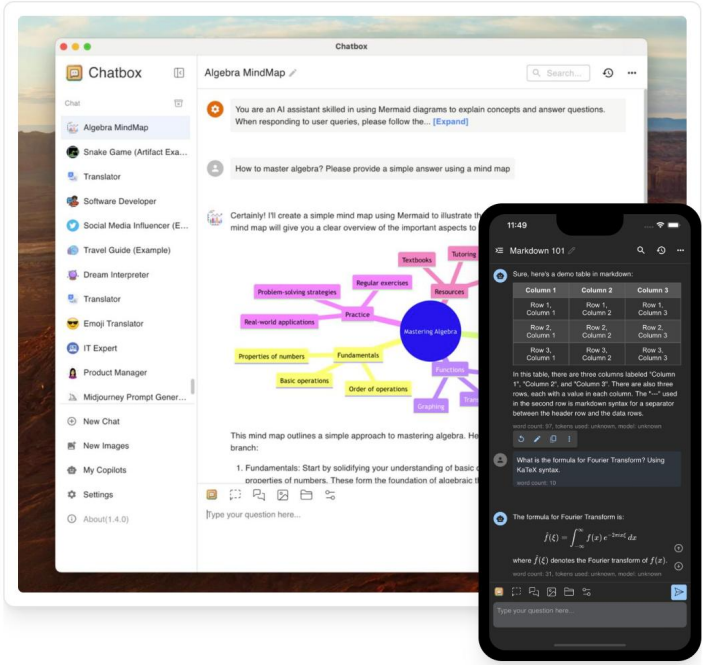
Chatbox AI,
办公学习好助手

Chatbox AI 是一款 AI 客户端应用和智能助手，支持众多先进的 AI 模型和 API，可在 Windows、MacOS、Android、iOS、Linux 和网页版上使用。

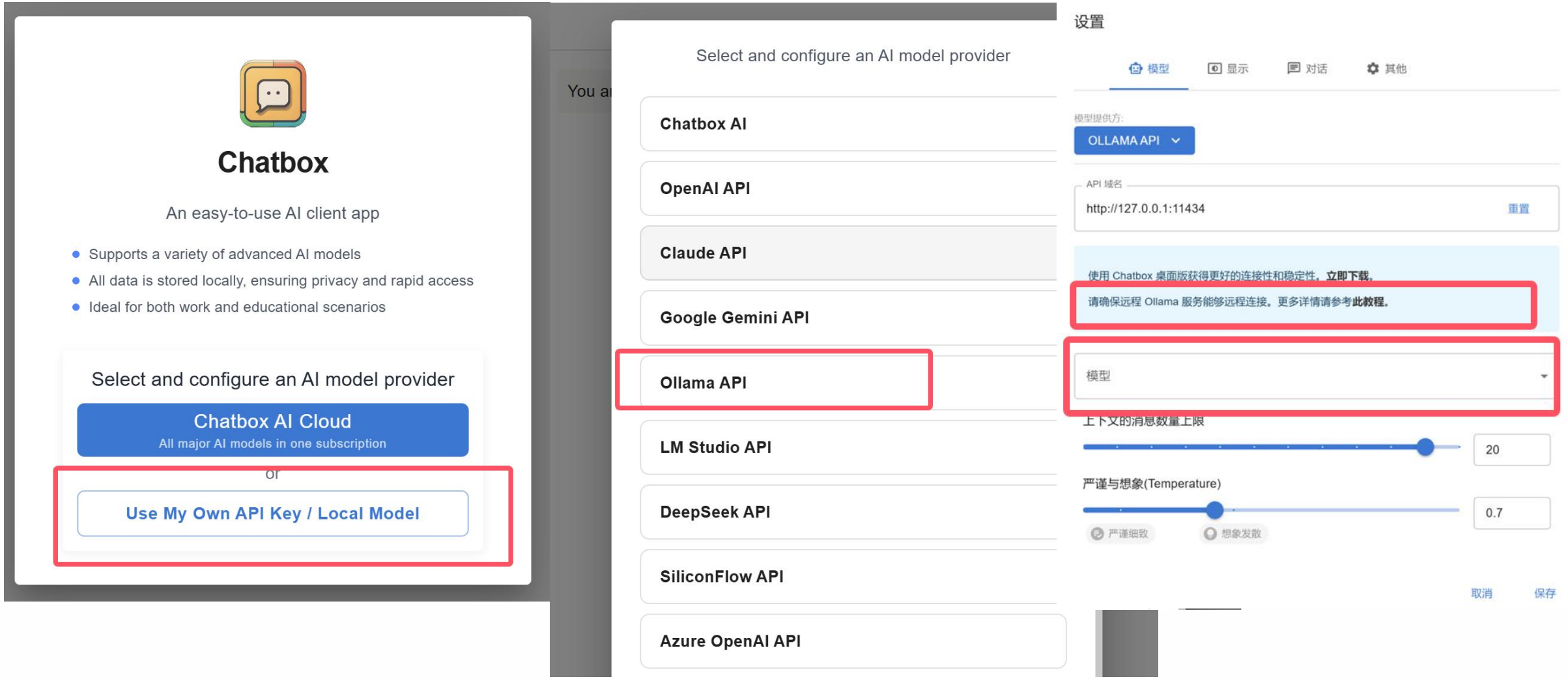
免费下载 (for Windows)　启动网页版

获取移动版

iOS　Android　APK　更多选项



Chatbox对话框页面



Chatbox对话框页面

在 Windows 上配置

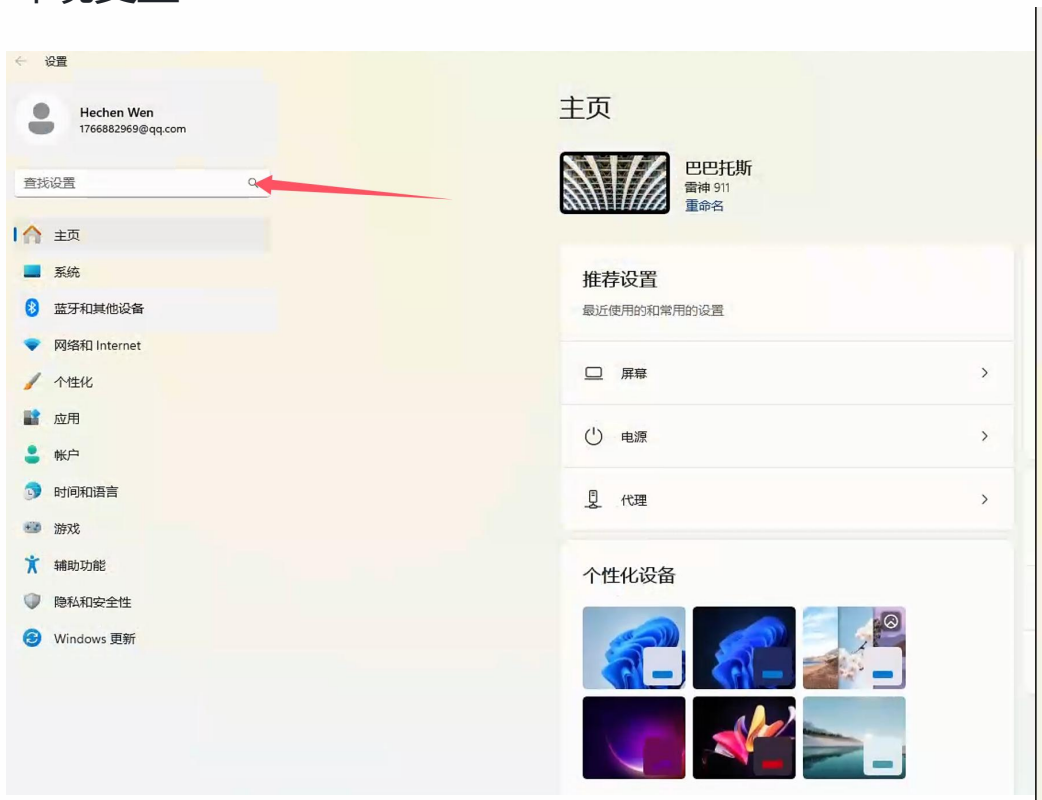
在 Windows 上，Ollama 会继承你的用户和系统环境变量。

1. 通过任务栏退出 Ollama。
2. 打开设置（Windows 11）或控制面板（Windows 10），并搜索“环境变量”。
3. 点击编辑你账户的环境变量。

   为你的用户账户编辑或创建新的变量 **OLLAMA_HOST**，值为 **0.0.0.0**； 为你的用户账户编辑或创建新的变量 **OLLAMA_ORIGINS**，值为 *****。
4. 点击确定/应用以保存设置。
5. 从 Windows 开始菜单启动 Ollama 应用程序。

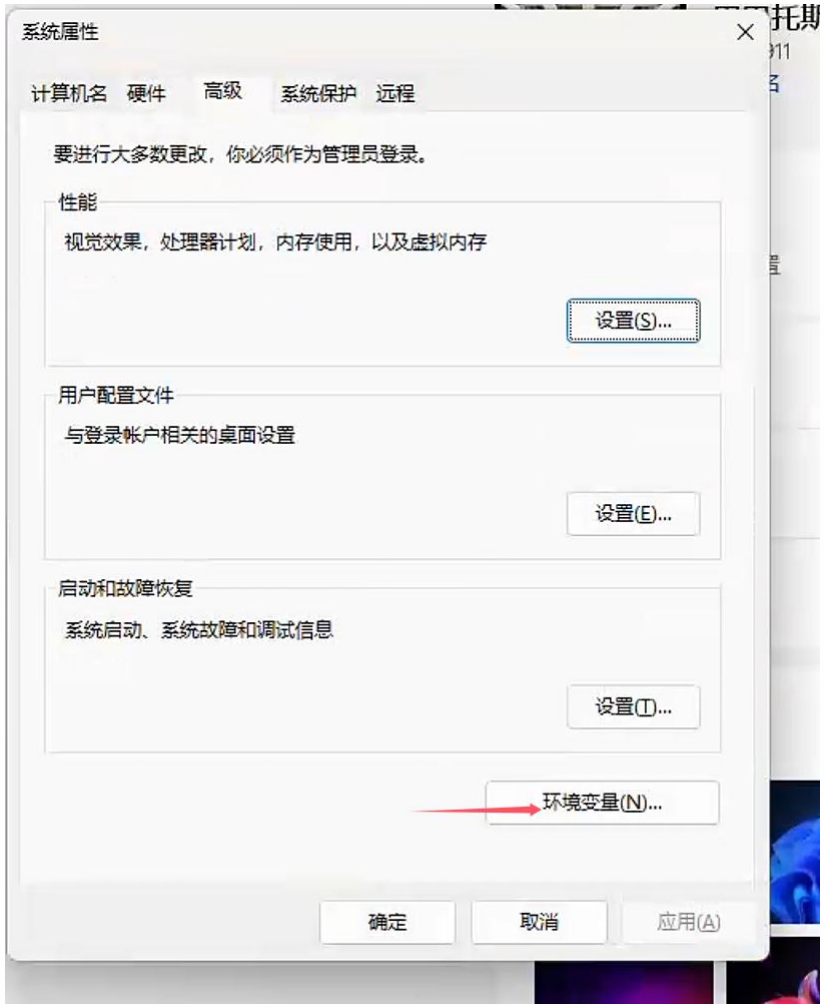
Chatbox对话框页面

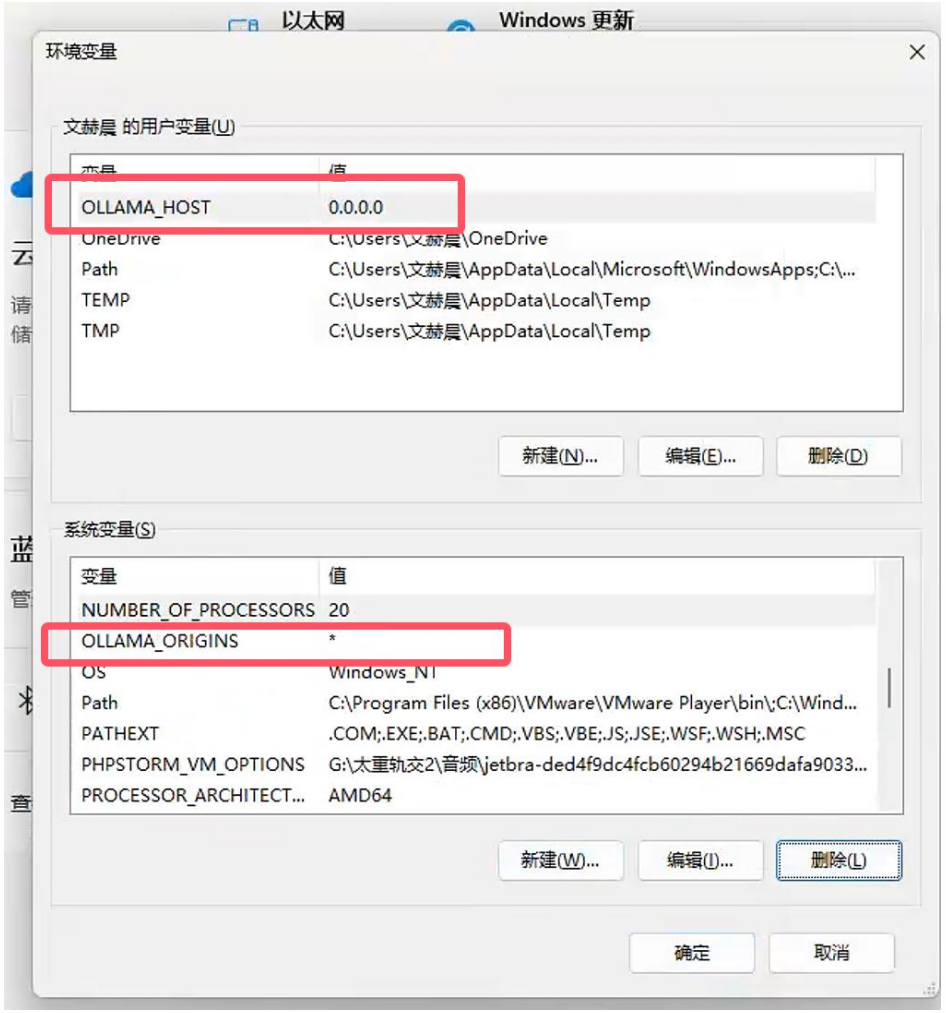
环境变量



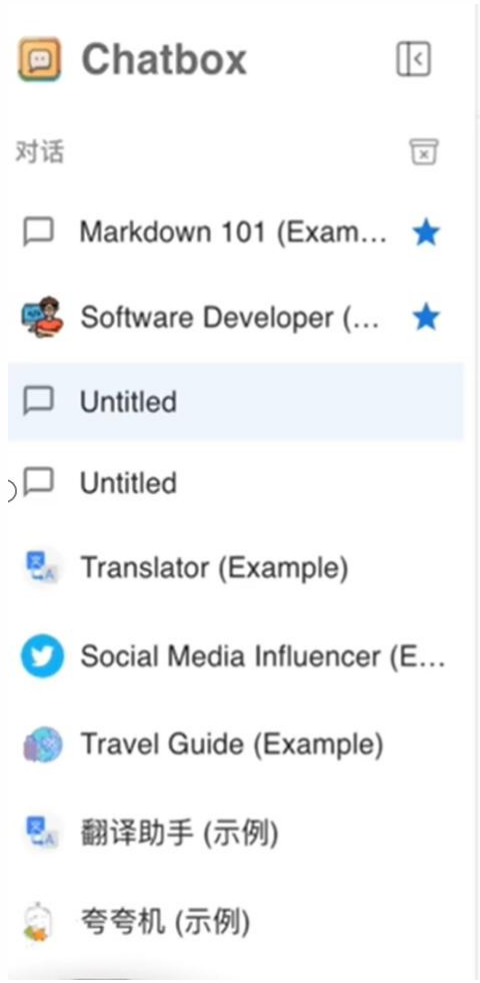
Chatbox对话框页面

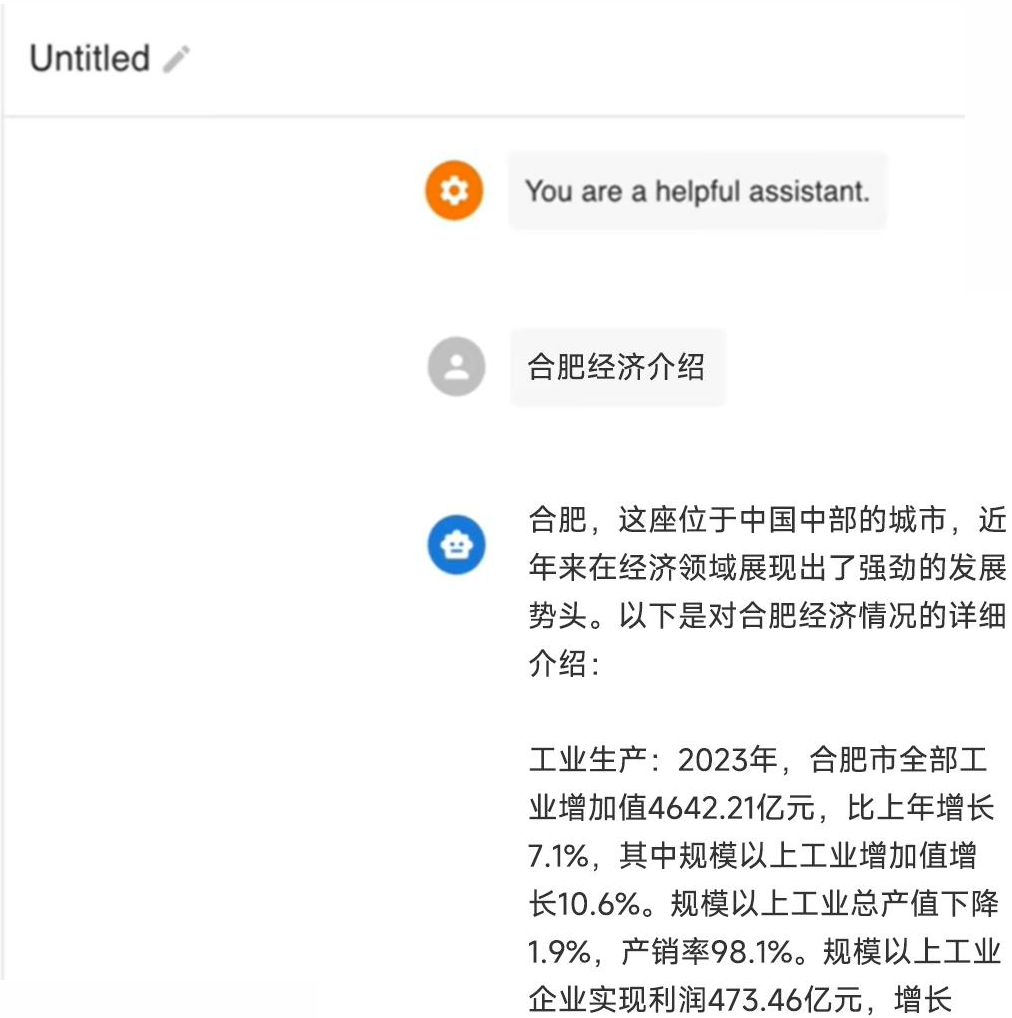
环境变量





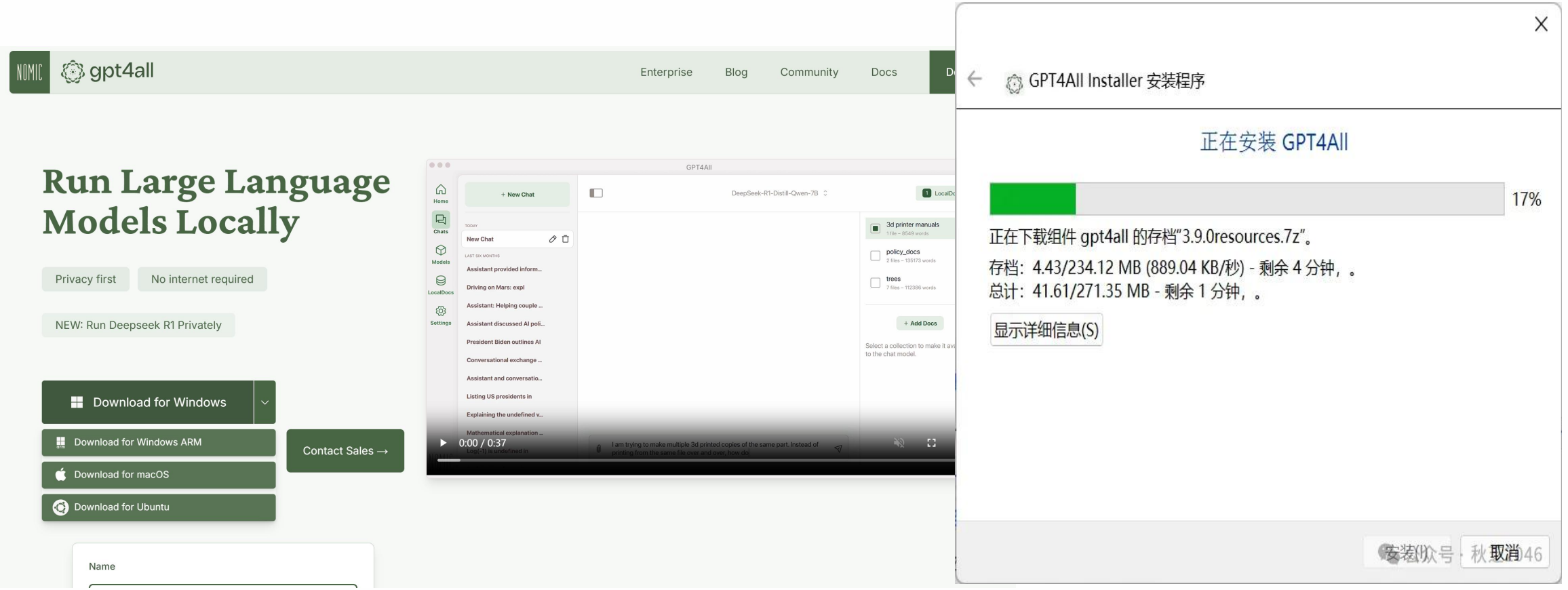
Chatbox对话框页面





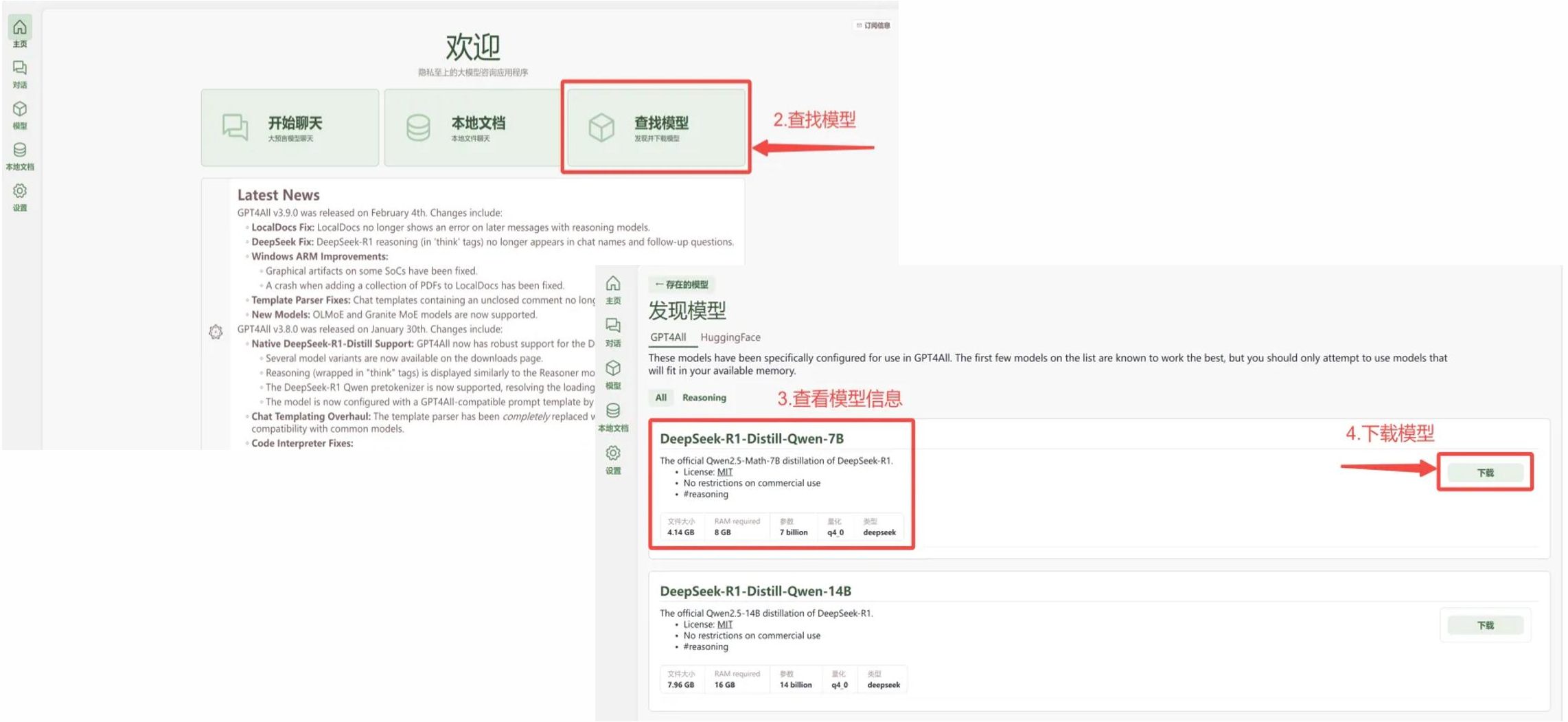
GPT4All方式安装

网址：https://gpt4all.io
第一步：安装 gpt4all
选择适合的系统版本：Windows/macOS/Linux



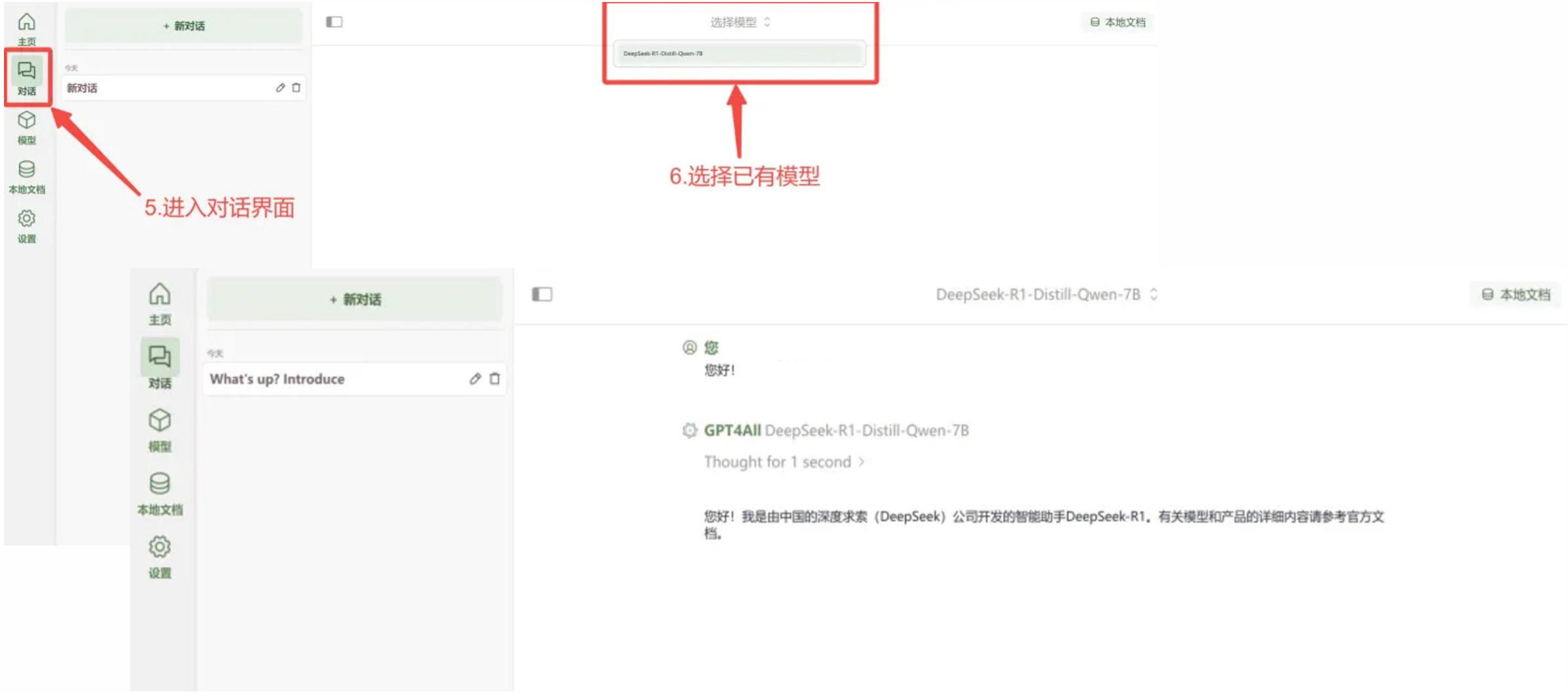
GPT4All方式安装

第二步：下载 DeepSeek 模型



GPT4All方式安装

第三步：开始对话



DeepSeek R1 671B linux完整版本地部署

部署此类大模型的主要瓶颈是内存+显存容量，建议配置如下：

•DeepSeek-R1-UD-IQ1_M：内存 + 显存 ≥ 200 GB

•DeepSeek-R1-Q4_K_M：内存 + 显存 ≥ 500 GB

使用 ollama 部署此模型。ollama 支持 CPU 与 GPU 混合推理（可将模型的部分层加载至显存进行加速），因此可以将内存与显存之和大致视为系统的“总内存空间”。除了模型参数占用的内存+显存空间（158 GB 和 404GB）以外，实际运行时还需额外预留一些内存（显存）空间用于上下文缓存。预留的空间越大，支持的上下文窗口也越大。

此版本主要参考的是李锡涵（Xihan Li）。伦敦大学学院（UCL）计算机系博士研究生的相关论文介绍和截图。

DeepSeek R1 671B 完整版本地部署

1. 下载模型文件从 HuggingFace
官网地址：https://huggingface.co/unsloth/DeepSeek-R1-GGUF

2. 安装 ollama，这个安装刚才讲了，这里是linux的模式。
执行以下命令：

```
curl -fsSL https://ollama.com/install.sh | sh
```

3. 创建 Modelfile 文件，该文件用于指导 ollama 建立模型
文件 DeepSeekQ1_Modelfile（对应于 DeepSeek-R1-UD-IQ1_M）的内容如下：

```
FROM /home/snowkylin/DeepSeek-R1-UD-IQ1_M.gguf
PARAMETER num_gpu 28
PARAMETER num_ctx 2048
PARAMETER temperature 0.6
TEMPLATE "<｜User｜>{{ .Prompt }}<｜Assistant｜>"
```

DeepSeek R1 671B 完整版本地部署

文件 DeepSeekQ4_Modelfile（对应于 DeepSeek-R1-Q4_K_M）的内容如下：

```
FROM /home/snowkylin/DeepSeek-R1-Q4_K_M.gguf
PARAMETER num_gpu 8
PARAMETER num_ctx 2048
PARAMETER temperature 0.6
TEMPLATE "<｜User｜>{{ .Prompt }}<｜Assistant｜>"
```

4. 创建 ollama 模型在第3步建立的模型描述文件所处目录下，执行以下命令：

```
ollama create DeepSeek-R1-UD-IQ1_M -f DeepSeekQ1_Modelfile
```

5. 运行模型，执行以下命令：

```
ollama run DeepSeek-R1-UD-IQ1_M --verbose
```

扩展系统交换空间教程：

```
https://www.digitalocean.com/community/tutorials/how-to-add-swap-space-on-ubuntu-20-04
```

```
journalctl -u ollama --no-pager
```

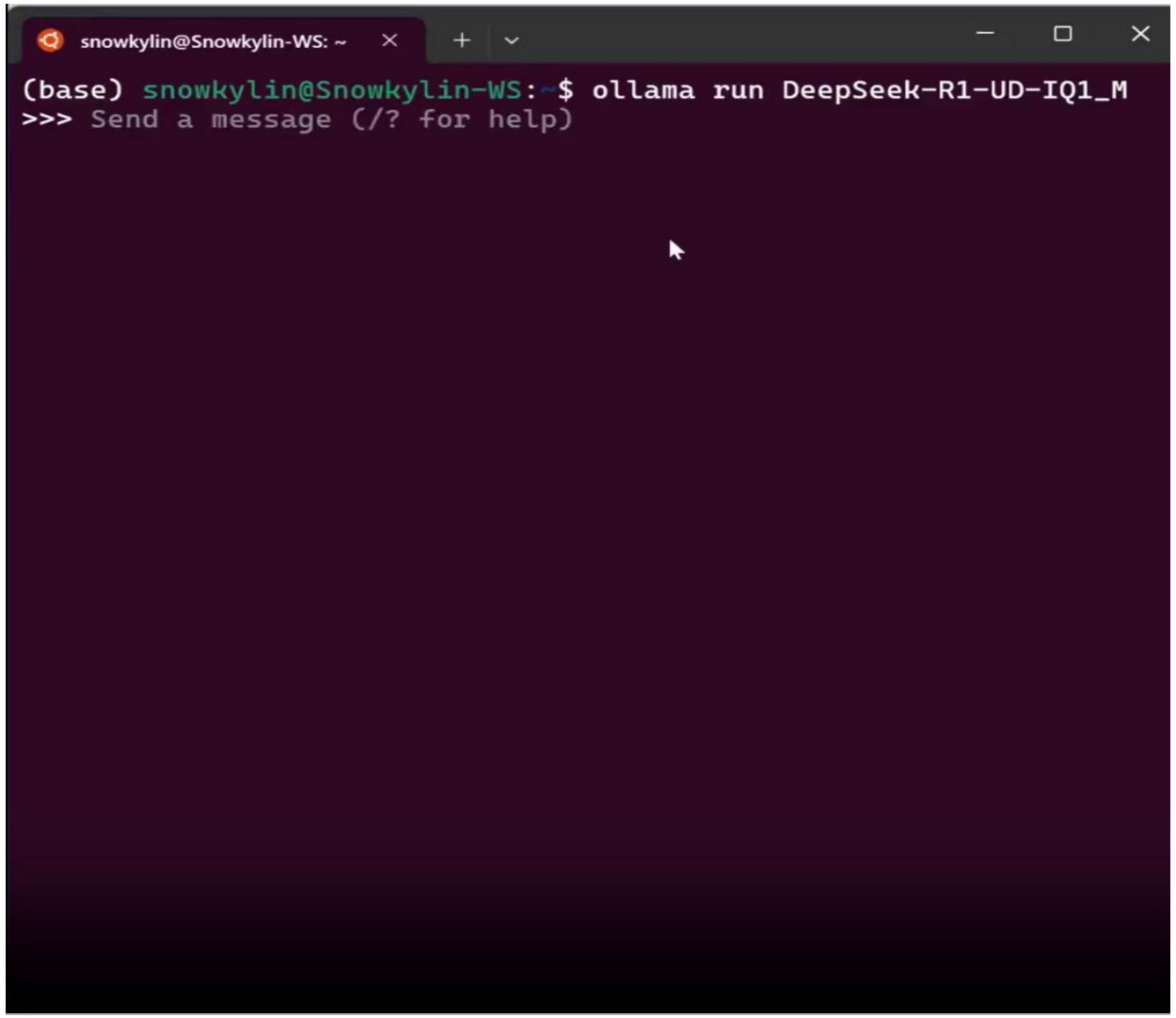
DeepSeek R1 671B 完整版本地部署

6. （可选）安装 Web 界面
使用 Open WebUI：

```
pip install open-webui
open-webui serve
```

DeepSeek R1 671B 完整版本地部署

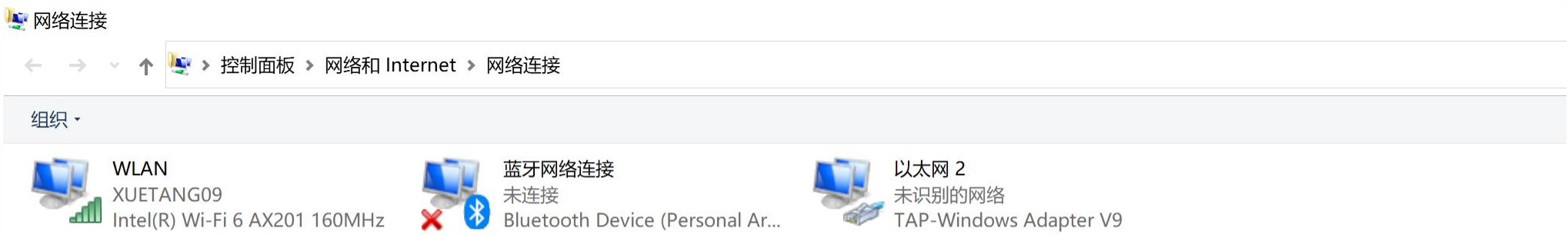
实测观察



本地断网运行设置

虚拟机断网运行

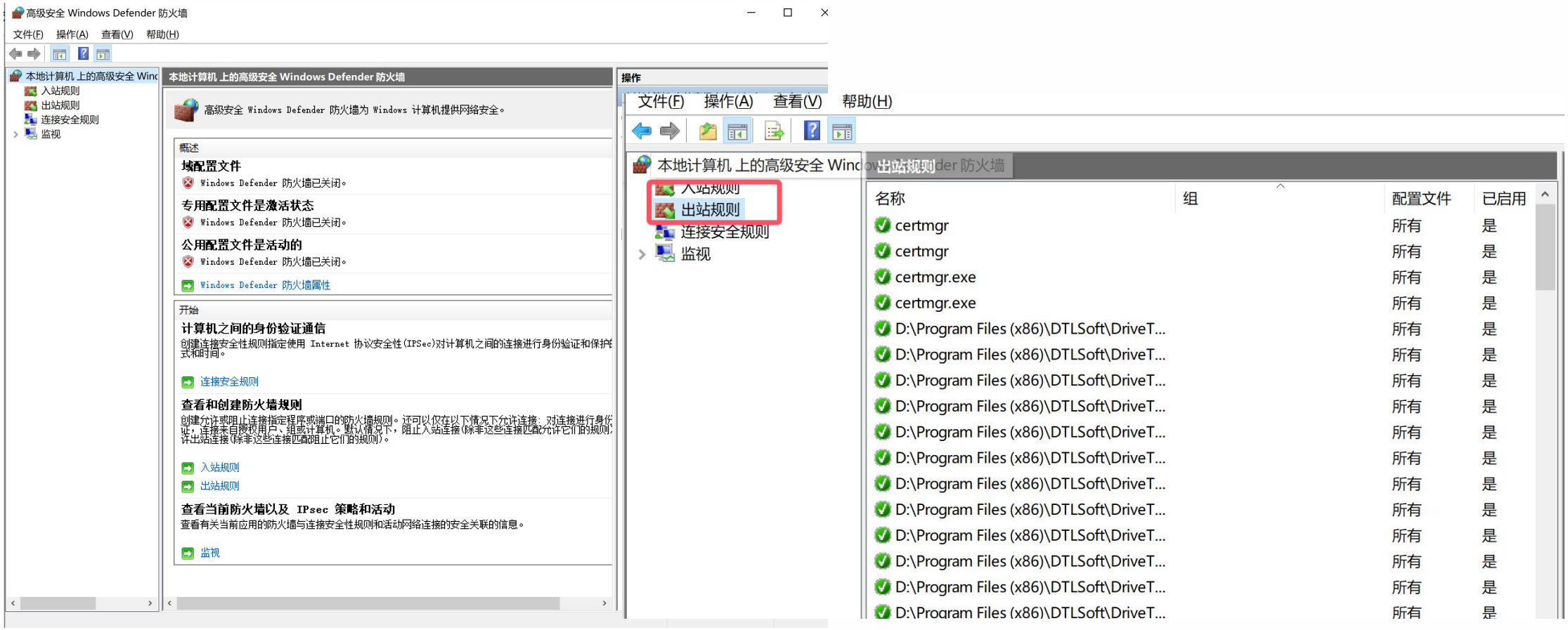
为确保DeepSeek R1在断网环境下运行,我们可以再虚拟机上运行整个程序，然后给虚拟机断网。



本地断网运行设置

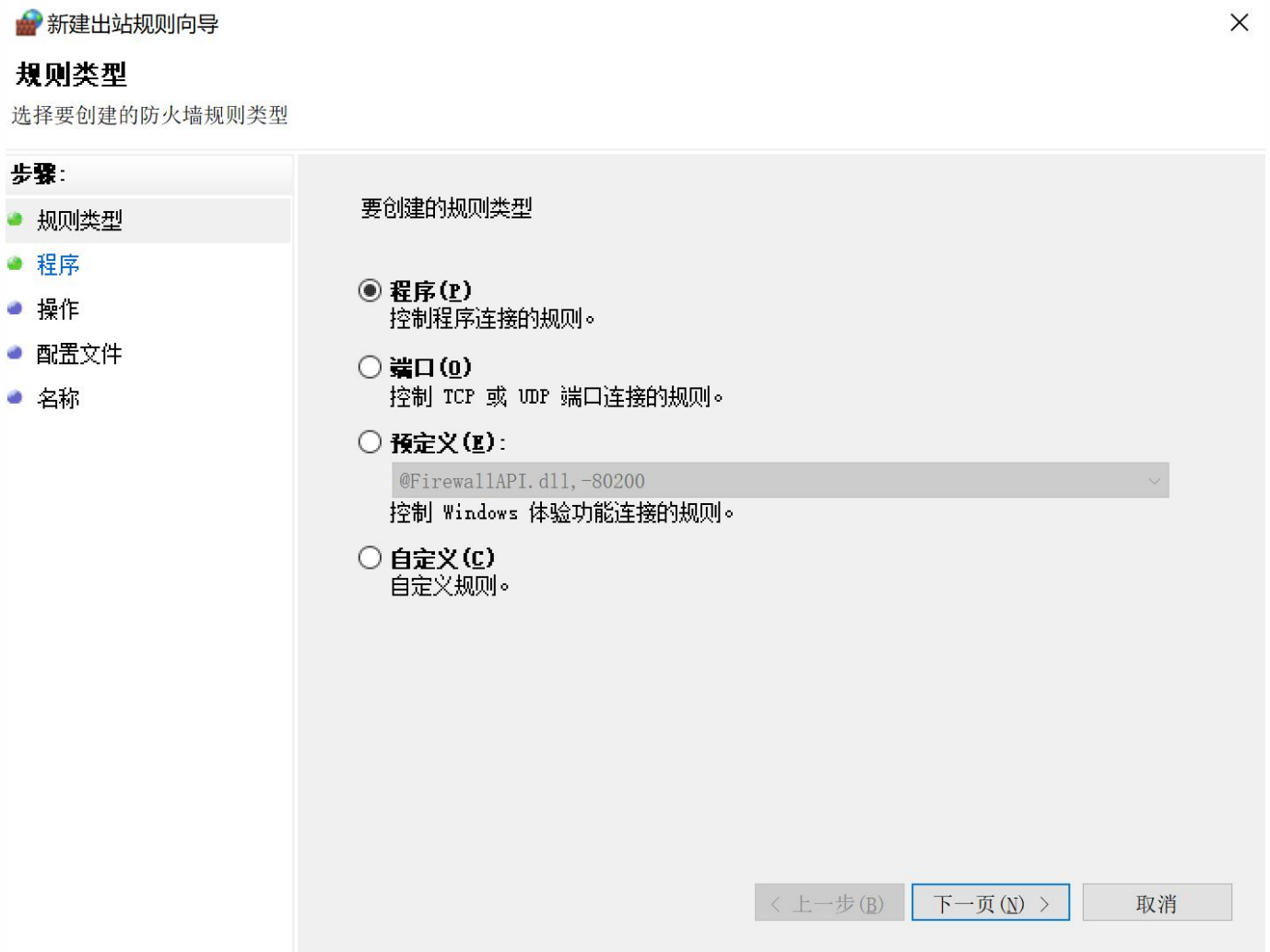
设置防火墙规则

为确保DeepSeek在断网环境下运行，需设置防火墙规则以阻止任何出网流量，从而增强本地部署的安全性。



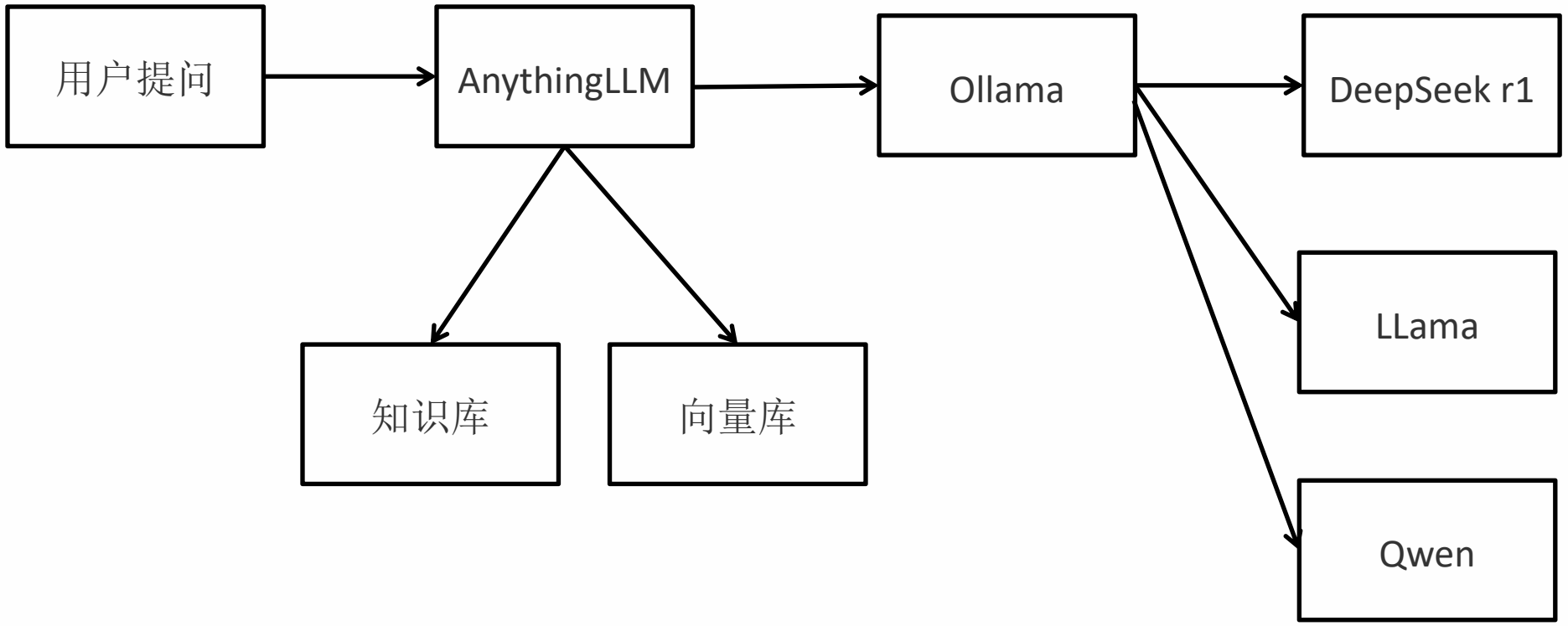
本地断网运行设置

我们在出站规则程序这里添加出站规则



本地知识库系统的搭建

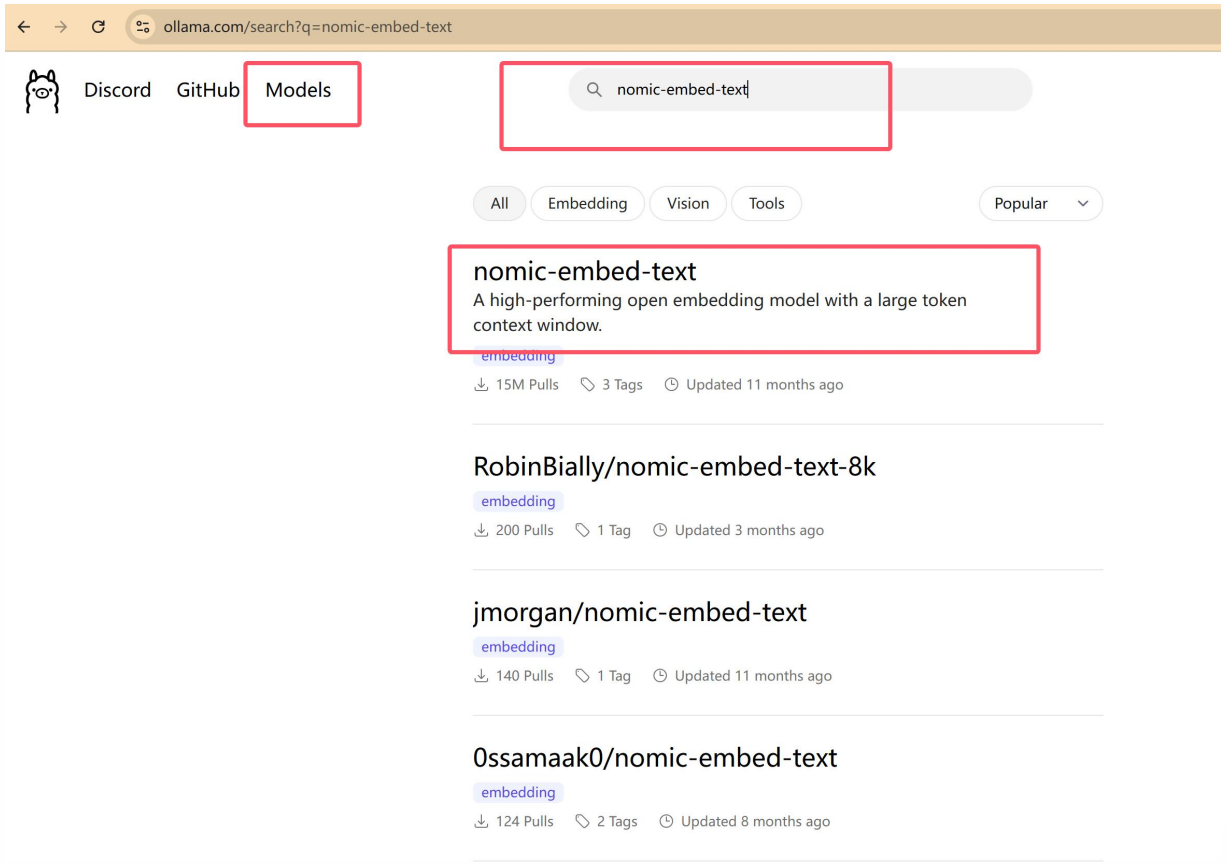
基于AnythingLLM的本地知识库与API搭建

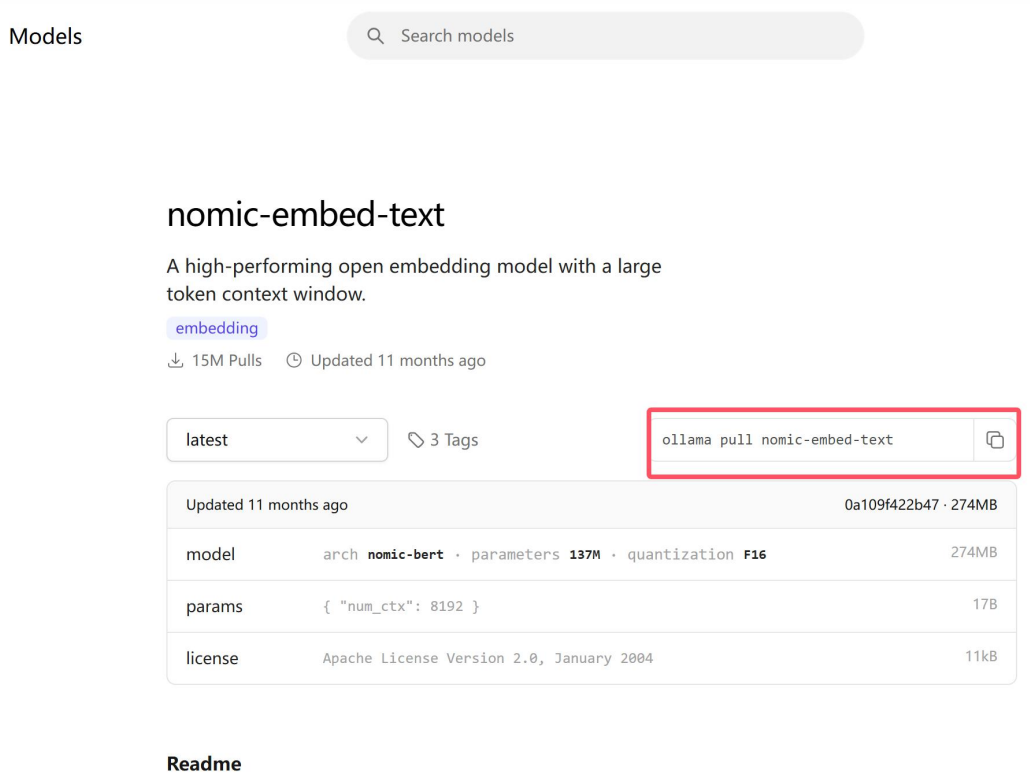


基于AnythingLLM的本地知识库与API搭建

第一步：下载nomic-embed-text

ollama官网->models->nomic-embed-text

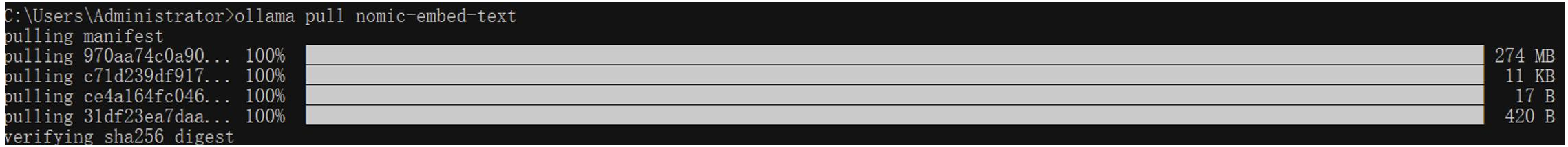




基于AnythingLLM的本地知识库与API搭建

下载nomic-embed-text

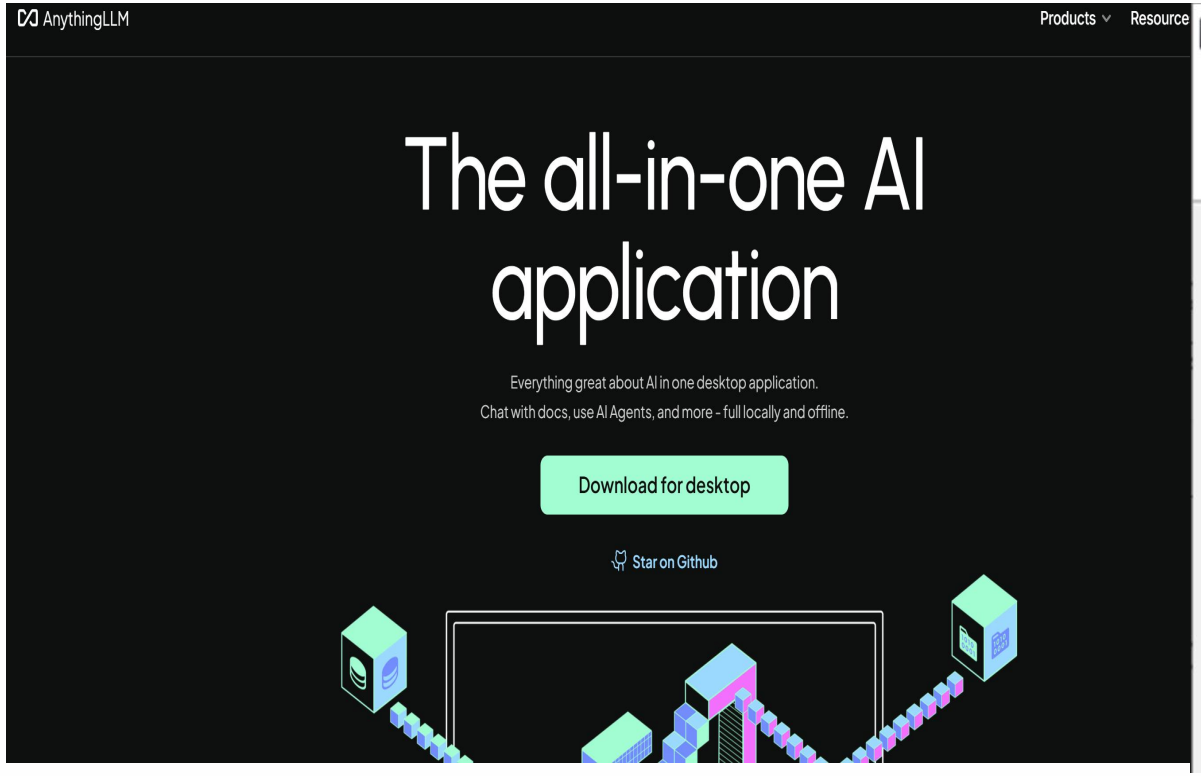
ollama pull nomic-embed-text

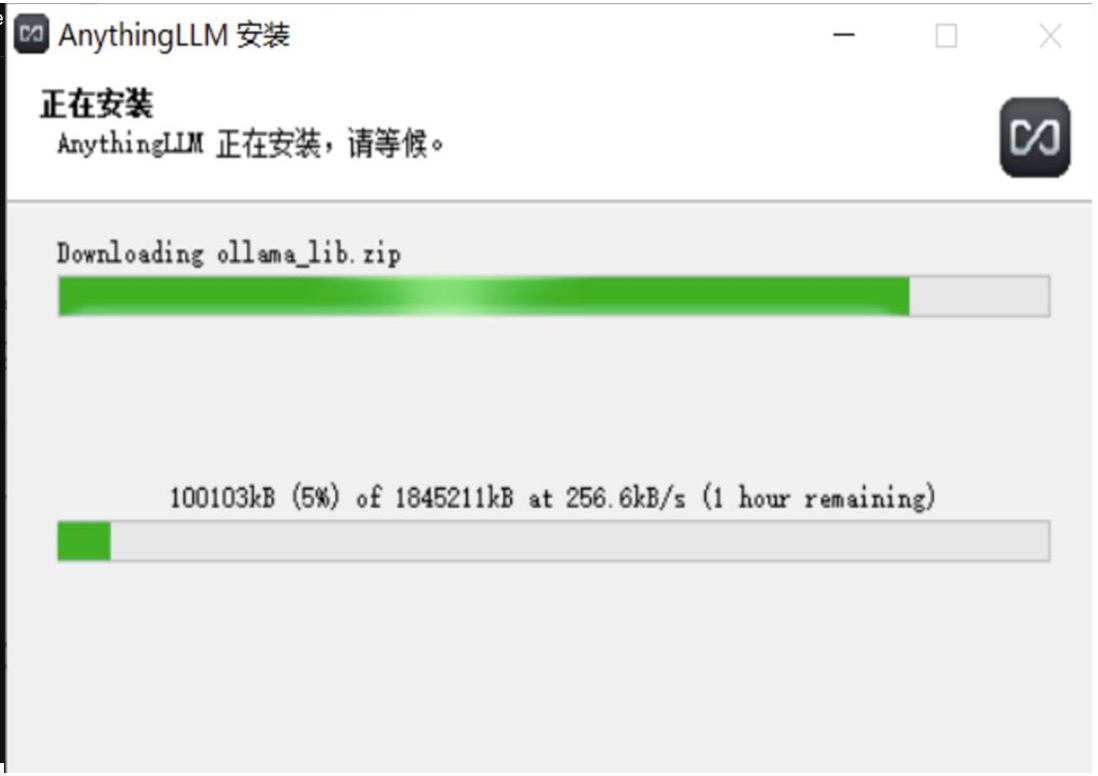


基于AnythingLLM的本地知识库与API搭建

下载AnythingLLM Desktop

官网地址：https://anythingllm.com/



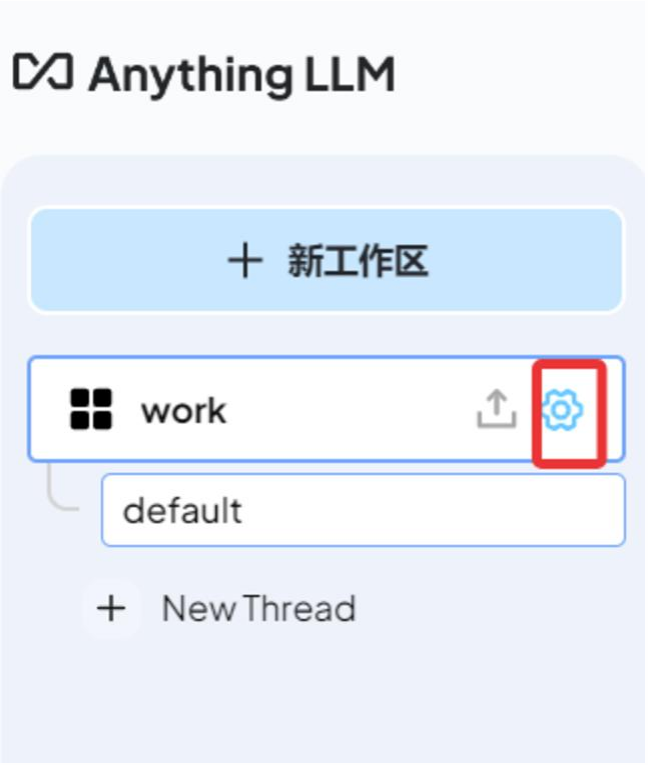


基于AnythingLLM的本地知识库与API搭建

安装完成后

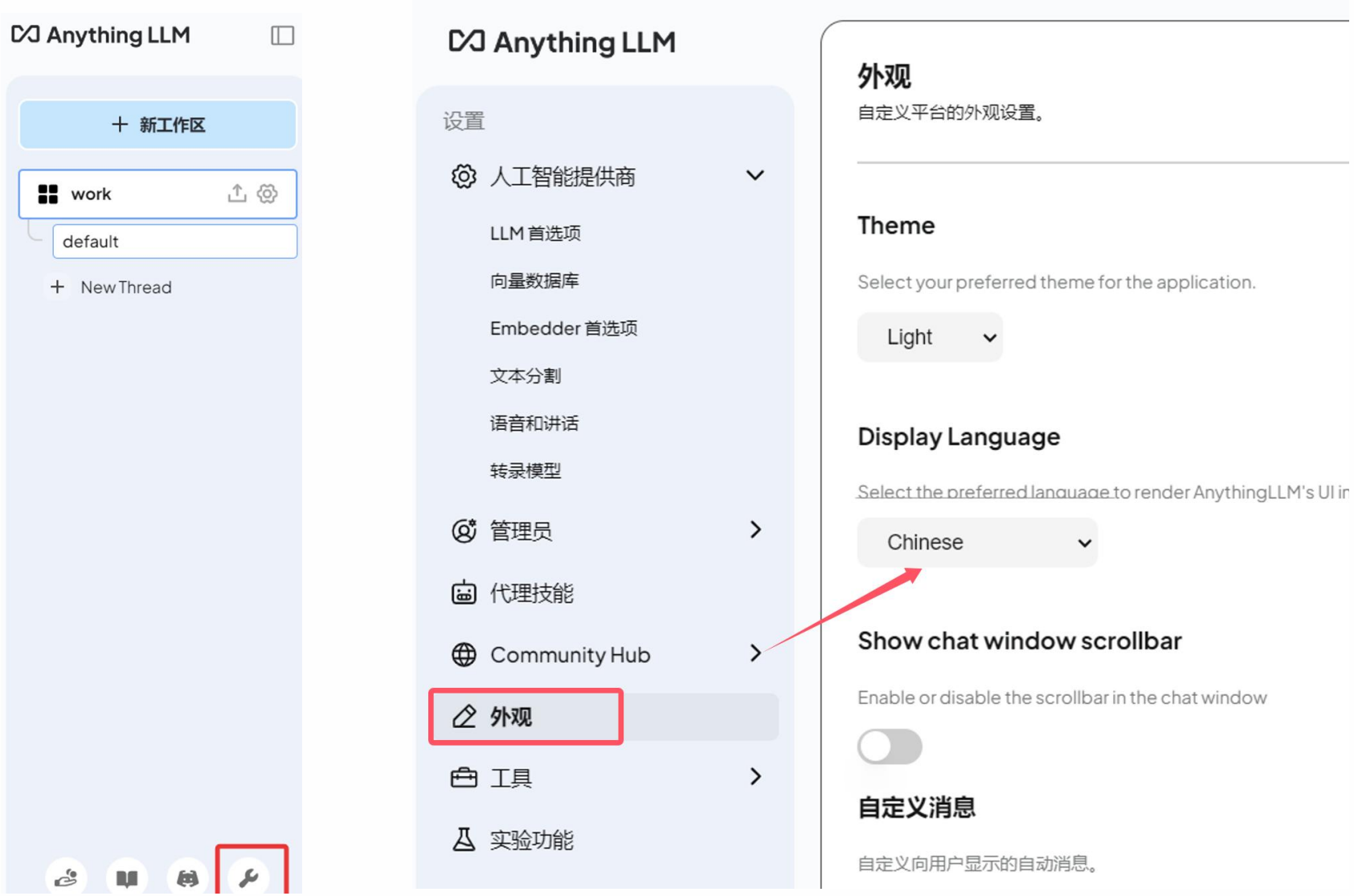
基于AnythingLLM的本地知识库与API搭建

创建工作区，进行设置



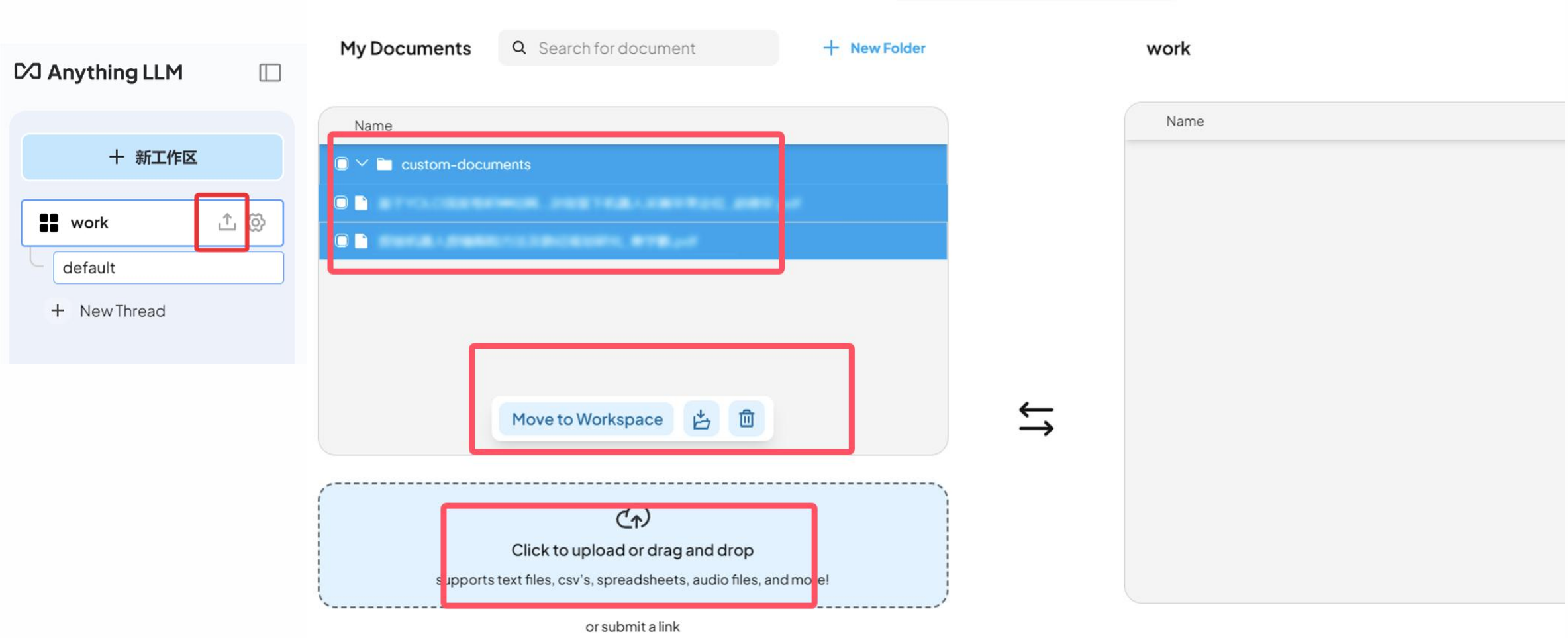
基于AnythingLLM的本地知识库与API搭建

软件设置



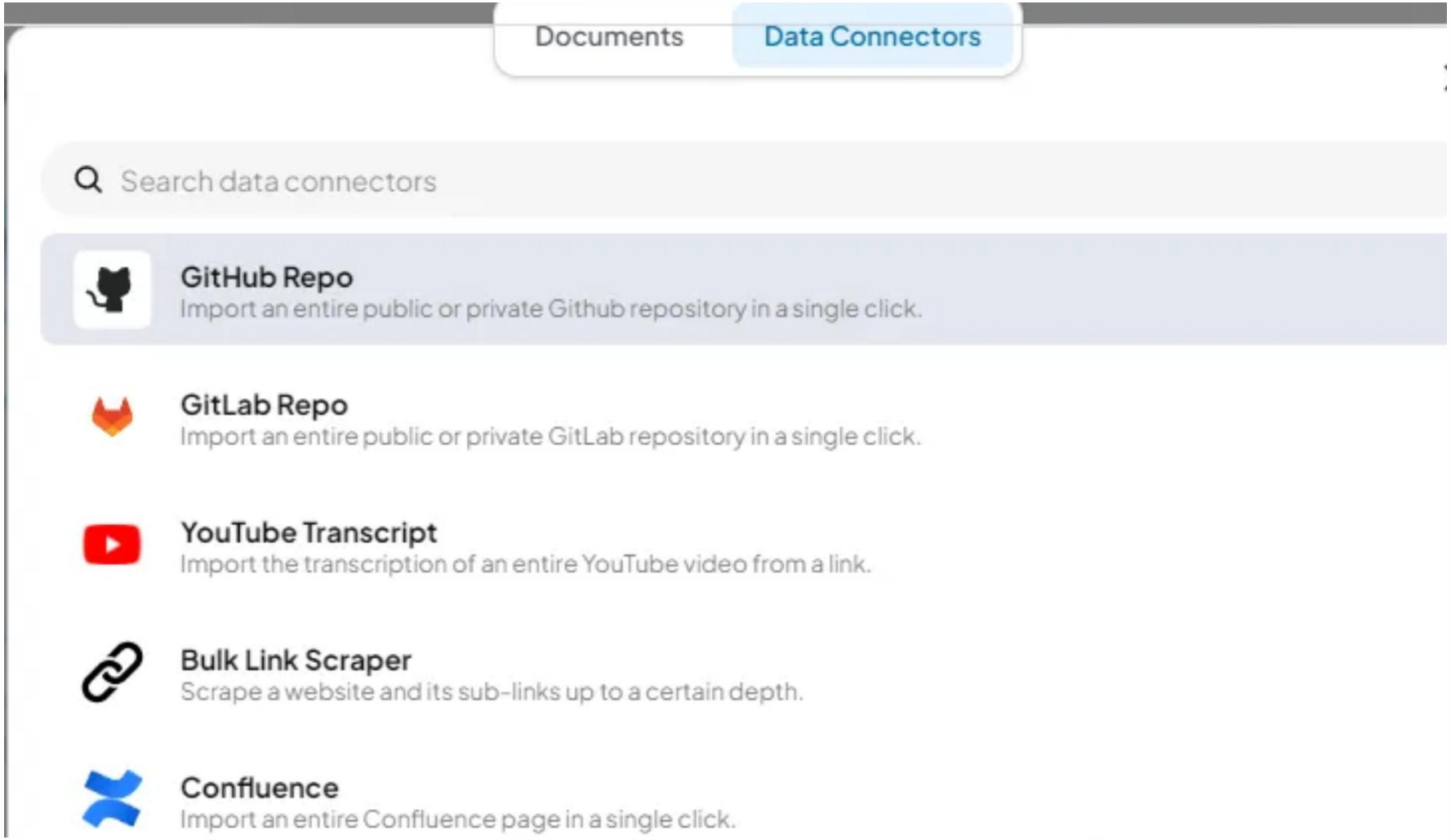
基于AnythingLLM的本地知识库与API搭建

上传文档



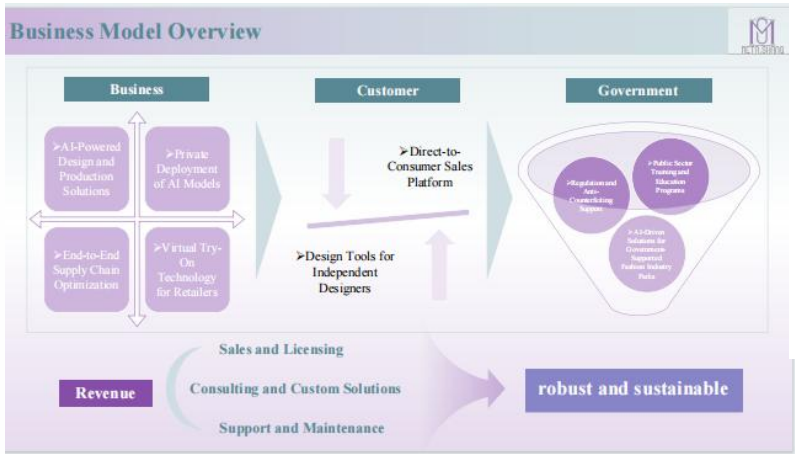
基于AnythingLLM的本地知识库与API搭建

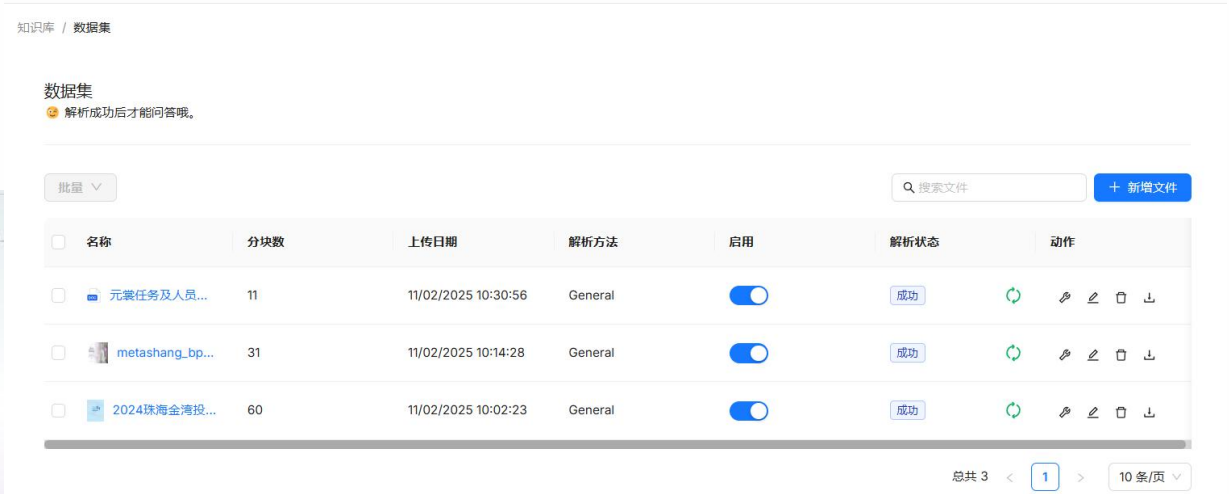
当然你可以配置远程文档，confluence、github都可以。



基于AnythingLLM的本地知识库与API搭建

测试一下





基于AnythingLLM的本地知识库与API搭建

api功能

DeepSeek本地部署完成后能做什么

维度	个人用户	企业用户
核心目标	提升个体工作效率/创造力	降本增效、驱动业务流程变革与数据资产增值
应用场景	写作、学习、娱乐、创意、数据管理	客服、营销、管理、合规、数据分析
模型关注点	轻量化/可移植性	高精度/稳定性/可解释性
隐私与安全	保护个人隐私	确保企业数据安全，符合行业法规
定制化程度	较低，通常直接使用预训练模型	较高，可能需要微调模型以适应特定业务需求

个人典型应用场景（个人助手、生产工具）

- **个人生产力工具**
- **本地资料管理：**将个人文档、笔记或书籍与模型结合，快速检索和总结信息
- **私人助理：**构建一个专属的AI助手，处理日程安排、提醒事项、私人知识库问答等
- **本地任务自动化：**通过自然语言指令完成重复性任务（如文件整理、数据分析）

- **技术实验**
- **模型微调：**根据个人需求对模型进行微调，例如针对特定任务（如写作、翻译）优化模型表现
- **DIY项目：**将模型嵌入到树莓派等小型设备中，打造智能家居助手或语音交互系统

- **隐私保护与数据安全**
- **敏感信息处理：**在本地环境中处理个人财务记录、健康数据或私密文档
- **离线操作：**在网络受限或无网络环境下运行模型，
- **个性化知识库：**将个人笔记、日记或其他私人文档与模型结合，构建专属的知识管理系统

企业典型应用场景（隐私保护、定制化能力、离线操作和高效协作）

- **数据隐私与安全**
- **敏感数据处理：**在本地环境中处理客户信息、财务记录或商业机密，确保数据不离开企业内部网络
- **离线操作：**在网络受限或无网络环境下运行模型，例如在偏远地区或工厂中进行实时分析
- **数据隔离：**将模型部署在完全隔离的环境中，避免数据泄露风险

- **知识管理**
- **企业知识库问答：**将模型与企业内部文档结合，构建专属的知识管理系统，快速检索和总结信息
- **文档自动化：**自动生成会议记录、报告或合同摘要，减少人工工作量
- **员工培训：**针对竞品为新员工生成个性化的培训材料，并提供实时答疑服务。

- **高效协作、监测与分析**
- **业务流程自动化：**合同条款智能审核、客服工单自动分类、报表数据自动生成
- **安全风控：** 内部通讯敏感词监控、代码仓库漏洞检测、 财务异常模式识别
- **决策支持系统：**市场趋势预测、供应链风险预警、客户流失分析、竞品情报自动分析

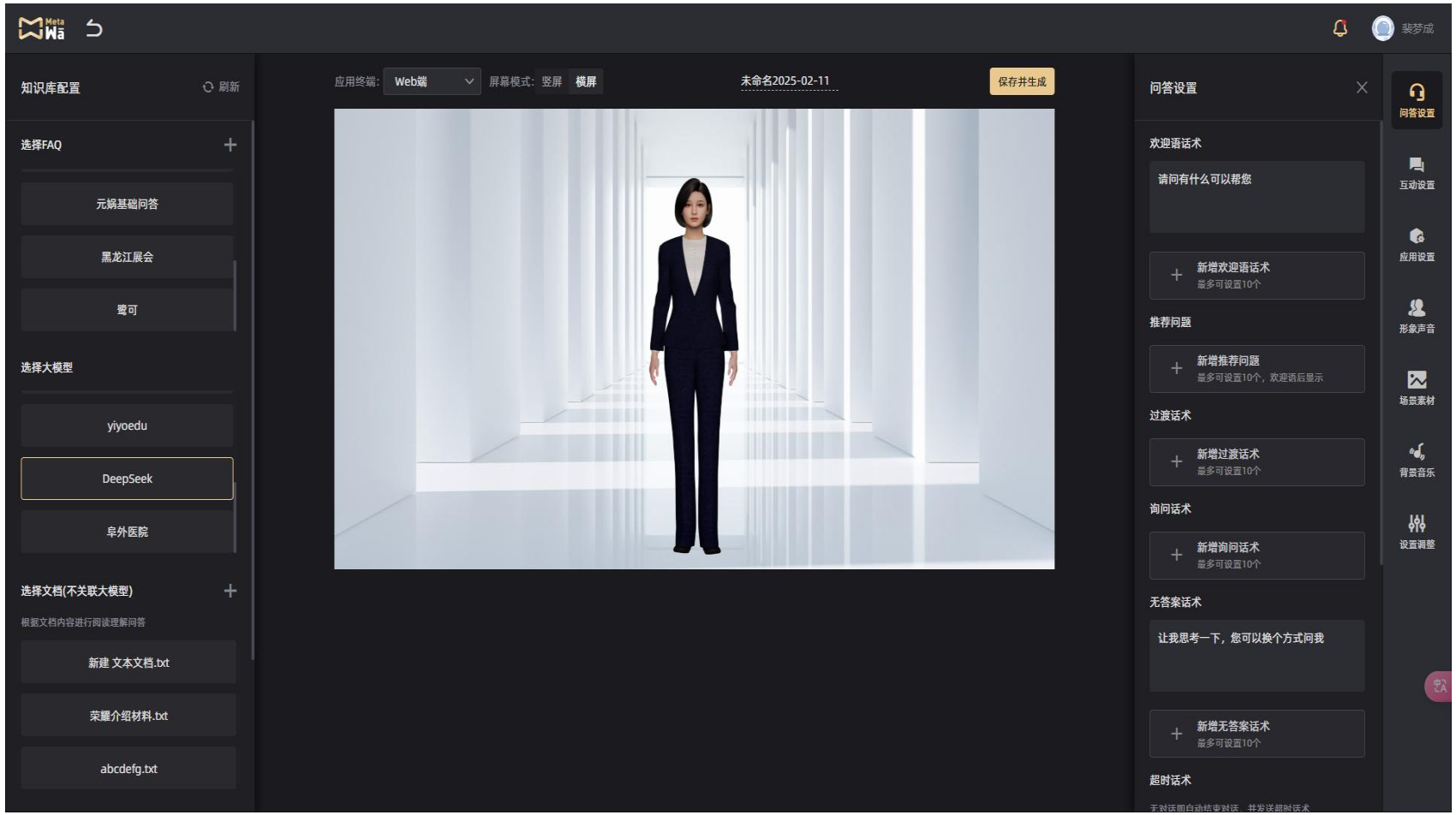
实际应用场景

元娲生态产品介绍

虚拟人定制化平台

满足个性化的虚拟人定制

1.3D超写实虚拟人、3D美型虚拟人、3D卡通虚拟人、2D真身复刻虚拟人多样化的虚拟人满足用户个性化的虚拟人定制需求；
2.多样化的虚拟人应用场景，满足不同客户的场景应用需求。

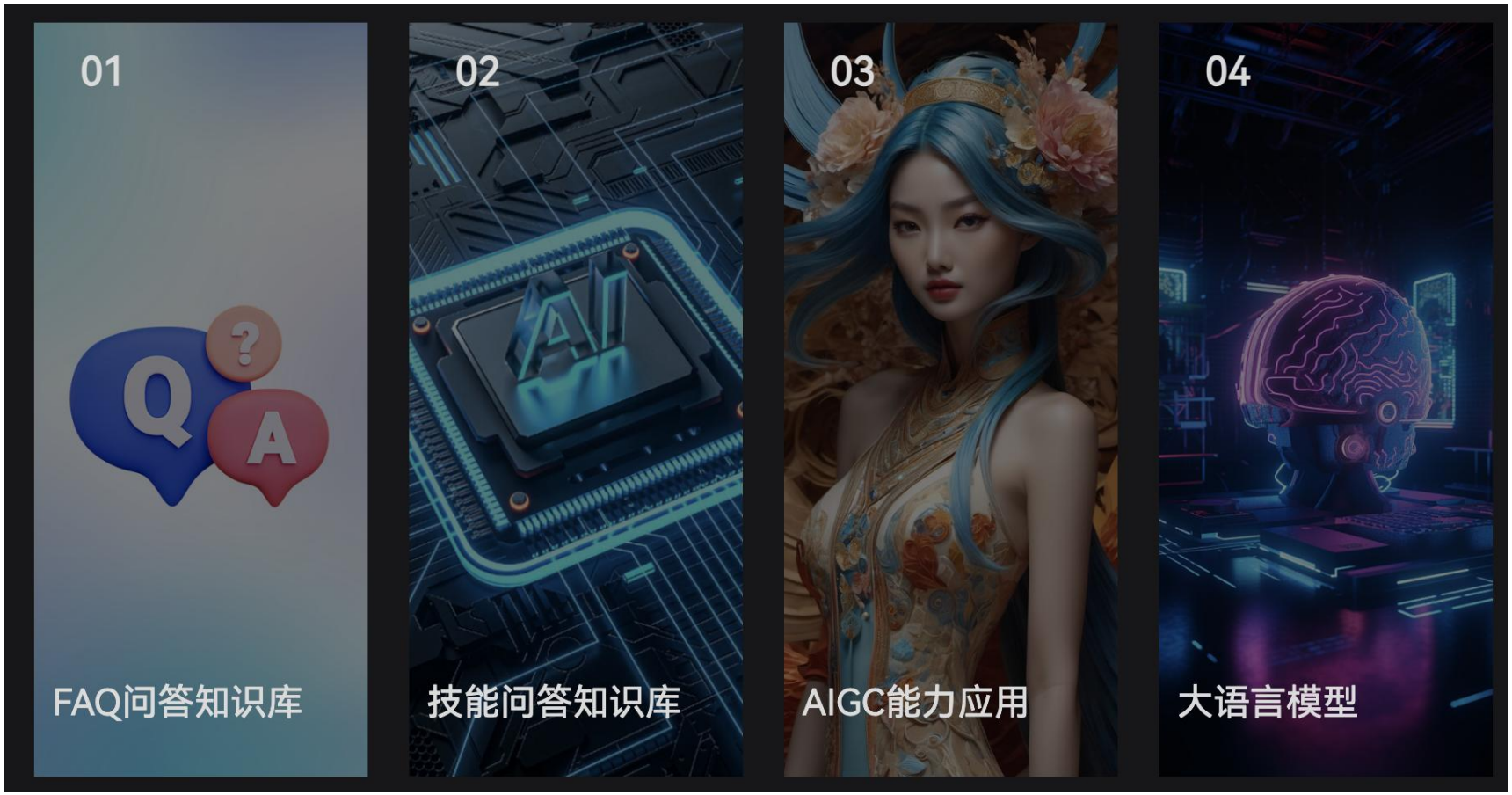


元娲生态产品介绍

元娲智能问答平台

虚拟人智能问答无所不能

1.虚拟人可接入专业的问答知识库；
2.专业的技能知识库（比如查询天气、车票等）；
3.虚拟人全新的AIGC的能力应用（绘画、唱歌、跳舞等）；
4.虚拟人可以接入不同的大语言模型，满足个性化的需求。



元娲生态产品介绍

首批支持国产化数字人平台

元娲平台是国内首批致力于虚拟人领域的AI科技先锋，我们的虚拟人生产力服务平台，全面自主可控，一站式满足虚拟人的创建与驱动需求，无缝适配国产操作系统，坚固安全防线，精准迎合国内市场需求。

产品介绍

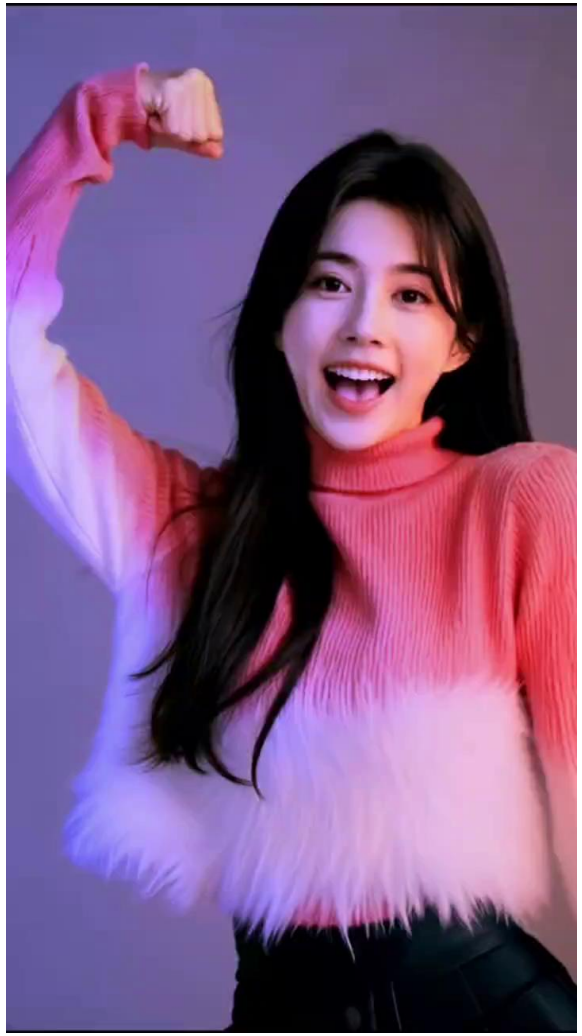
元镜-多模态创意呈现，分镜创作新引擎
一款基于人机快生引擎的AI视频创作系统，从需求提交到成片仅需10分钟，即可输出75分质量的视频。
1.创意视频脚本引擎
从灵感到成品脚本，支持角色定制与创意扩写。

2.多模态创意分镜引擎
支持全方位分镜设计，生成分镜图、视频和音乐，确保风格与情感一致。

3.分镜一键成片引擎
自动合成多分镜视频，智能补全内容，支持字幕与旁白生成，实现快速成片。

案例展示

案例展示



元知：AI综述平台

核心功能

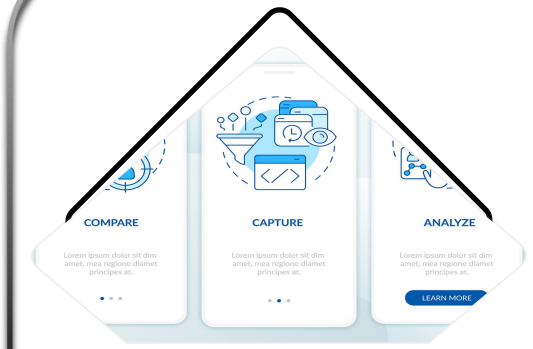
- 自动化整合
- 高质量输出
- 高效助科研
- 高水平综述

语言支持

- 中英文支持
- 国际化综述
- 跨语言便利

智能算法

- 海量文献分析
- 关键信息提取
- 结构内容生成



版本选择

- 基础版（无图）
- 增强版（单图）
- 专业版（单/双图）

THANKS