

DeepSeek本地部署与应用构建

智灵动力 陈军

目录

- 1、DeepSeek简单介绍与使用
- 2、DeepSeek本地部署
- 3、本地知识库搭建
- 4、实际应用场景

DeepSeek 学习手册（持续更新ing...）

学习地址：<https://yunyi nghui . fei shu. cn/wi ki /QuK2wYg0Xi H3m6k1tADcxuhj nEg>

高帆文化 > ... > AI对话能力学习 > Deepseek 学习手册（持续更新ing...） 外部 平

最近修改：2 小时前

分享 +78

Deepseek 学习手册（持续更新ing...）

大国 | 91356 | 132242 | 372 | 73



扫码分享文档

更新日志

Deepseek 交流群

- 一、Deepseek 简介
- 二、使用方式
- 三、本地部署
- 四、使用技巧
- 五、提示词技能
- 六、实用案例
- 七、赚钱案例
- 八、研究报告
- 九、学习文档
- 十、名人观点及媒体报道

附录：官方材料

统一鸣谢

文件汇总（敬请期待...）

适用人群：关注国内 AI 大模型发展及应用人群

首发时间：2025 年 2 月 6 日

更新时间：2025 年 2 月 18 日

内容出品人：大国

内容出品方：玩赚 AI 实验室

使用建议：如果需要快速定位到精确内容，可以使用快捷键 Ctrl +F/Command +F 的形式，搜索关键字/词，查找你想要的内容。

持续更新 Deepseek 的相关介绍及动态，研究报告，实用场景，赚钱案例等 10 个板块内容，永久免费在线查看，欢迎收藏转发支持，文末有惊喜。

更新日志

DeepSeek简单介绍与使用

模型简介



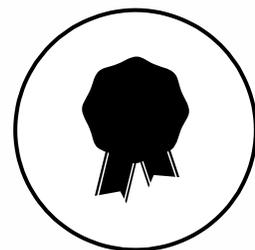
推理能力强

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。



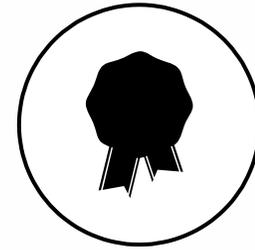
本地化部署 隐私保护

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 workflow

集成 DeepSeek R1 于自动化 workflow，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

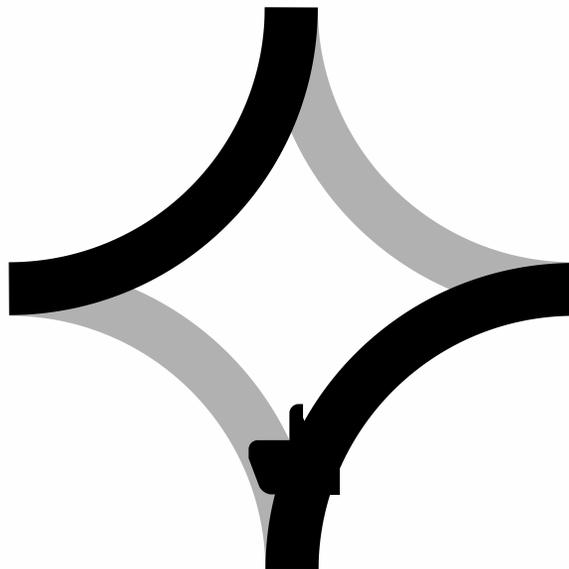
本地部署的必要性

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。

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



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。

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



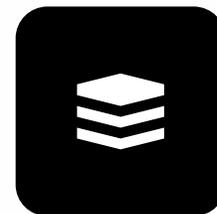
操作系统

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

The screenshot shows the Ollama website's download page. At the top, the browser address bar displays 'ollama.com/download'. The navigation bar includes a logo, links for 'Discord', 'GitHub', and 'Models', a search bar for models, and buttons for 'Sign in' and 'Download'. The main heading is 'Download Ollama', followed by three platform options: 'macOS', 'Linux', and 'Windows'. The 'Windows' option is highlighted with a dark button labeled 'Download for Windows', which includes the note 'Requires Windows 10 or later'. Below this, a section prompts users to sign up for updates, featuring an input field for an email address and a 'Get updates' button.

ollama.com/download

Discord GitHub Models

Search models

Sign in Download

Download Ollama

macOS Linux Windows

Download for Windows

Requires Windows 10 or later

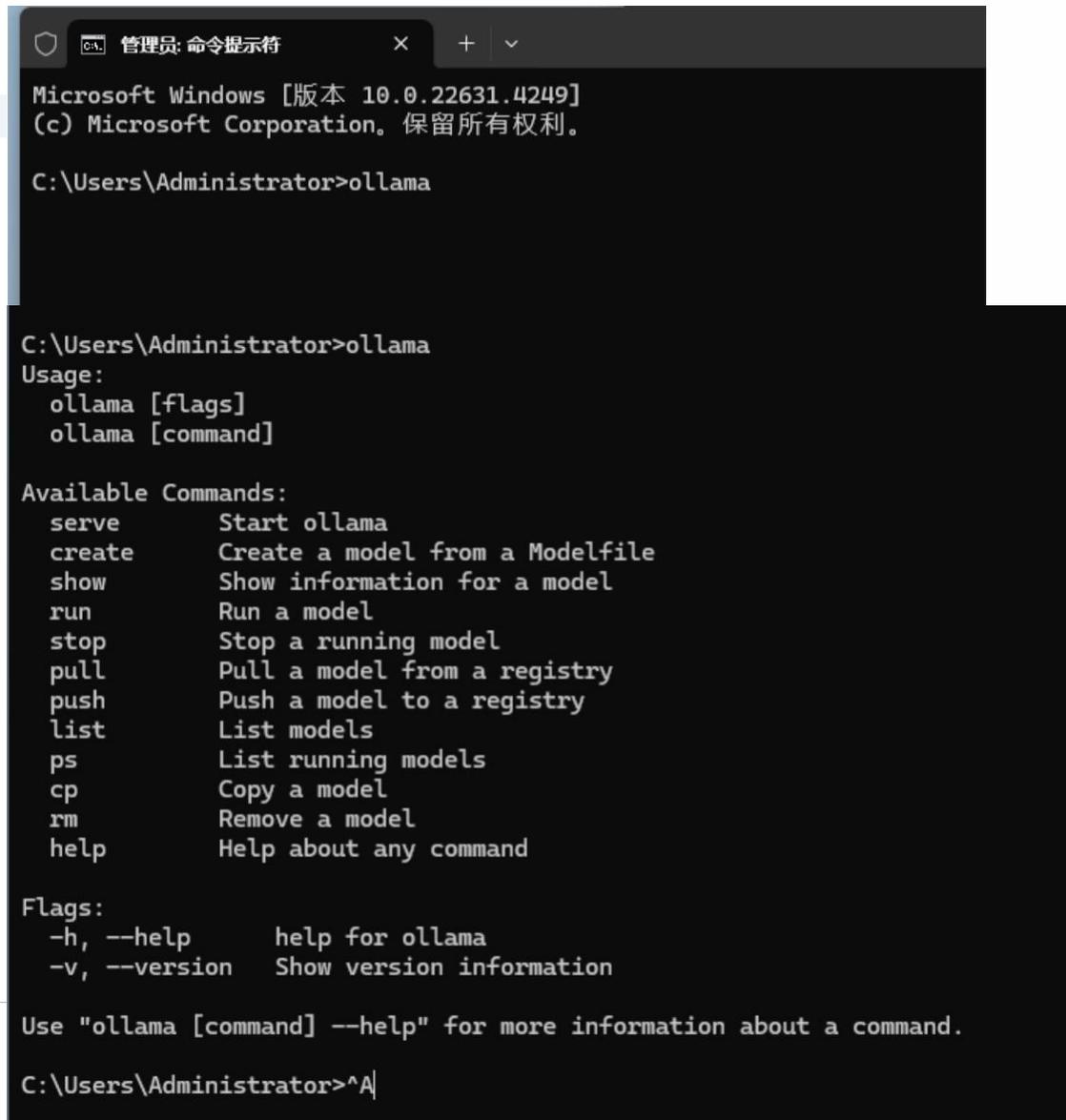
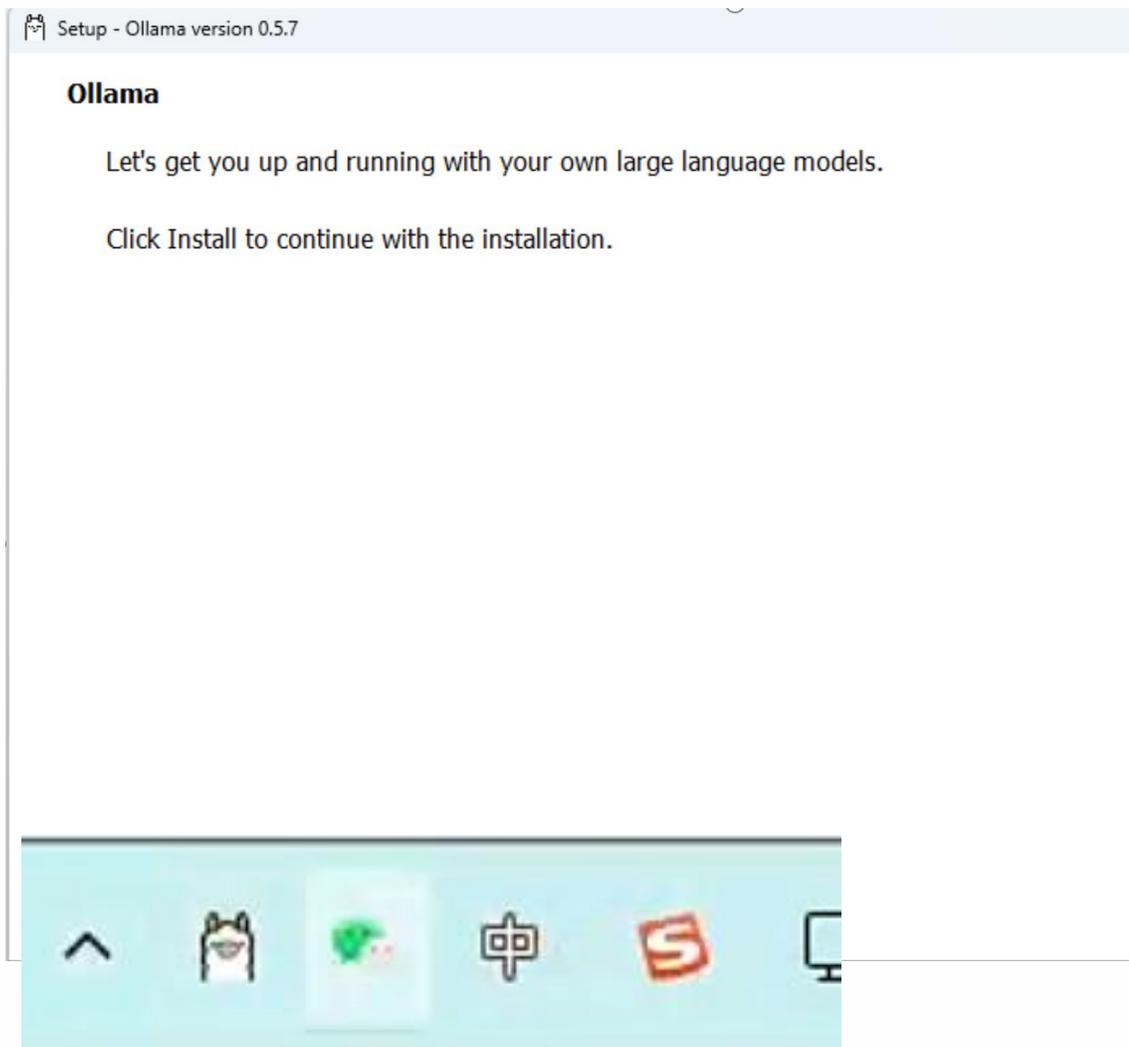
While Ollama downloads, sign up to get notified of new updates.

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
1.5b	1.1GB
7b	4.7GB
8b	4.9GB
14b	9.0GB
32b	20GB
70b	43GB
671b	404GB
View all	

ollama run deepseek-r1

0a8c26691023 · 4.7GB

parameters 7.62B · quantization Q4_K_M 4.7GB

begin_of_sentence | >", "< | end_of__sentence | >...

148B

}}{{ .System }}{{ end }} {{- range \$i, \$_ := .Mes...

387B

right (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>



首页 下载 定价 帮助 Stars 29k

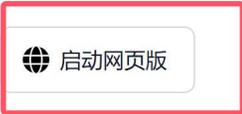
Chatbox | Product Hunt



Chatbox AI, 办公学习好助手

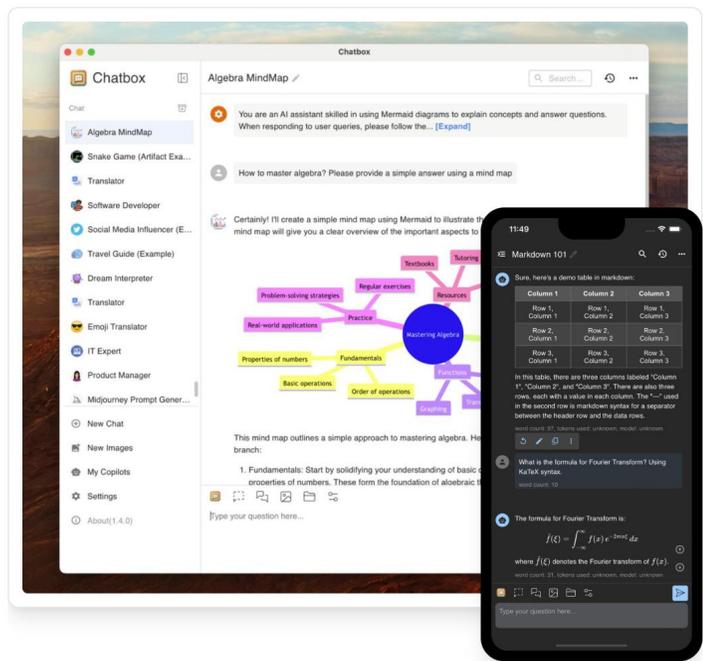
Chatbox AI 是一款 AI 客户端应用和智能助手，支持众多先进的 AI 模型和 API，可在 Windows、MacOS、Android、iOS、Linux 和网页版上使用。

↓ 免费下载 (for Windows)



获取移动版

- iOS
- Android
- APK
- + 更多选项



Chatbox对话框页面



Chatbox

An easy-to-use AI client app

- Supports a variety of advanced AI models
- All data is stored locally, ensuring privacy and rapid access
- Ideal for both work and educational scenarios

Select and configure an AI model provider

Chatbox AI Cloud
All major AI models in one subscription

OR

Use My Own API Key / Local Model

You are

Select and configure an AI model provider

- Chatbox AI
- OpenAI API
- Claude API
- Google Gemini API
- Ollama API**
- LM Studio API
- DeepSeek API
- SiliconFlow API
- Azure OpenAI API

设置

模型 显示 对话 其他

模型提供方:

Ollama API

API 域名

http://127.0.0.1:11434

重置

使用 Chatbox 桌面版获得更好的连接性和稳定性。立即下载。

请确保远程 Ollama 服务能够远程连接。更多详情请参考此教程。

模型

上下文的消息数量上限

20

严谨与想象(Temperature)

0.7

严谨细致

想象发散

取消

保存

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。

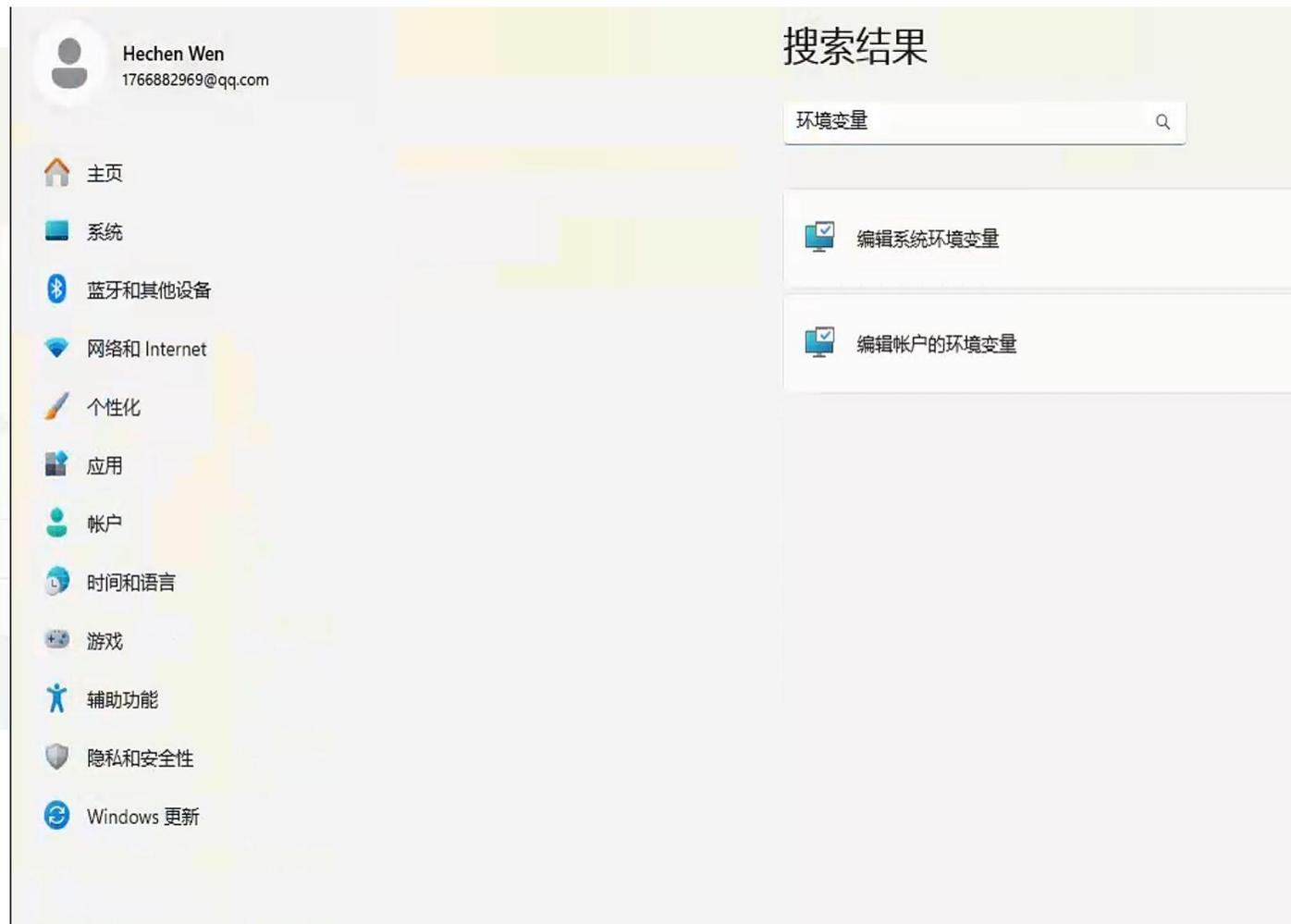
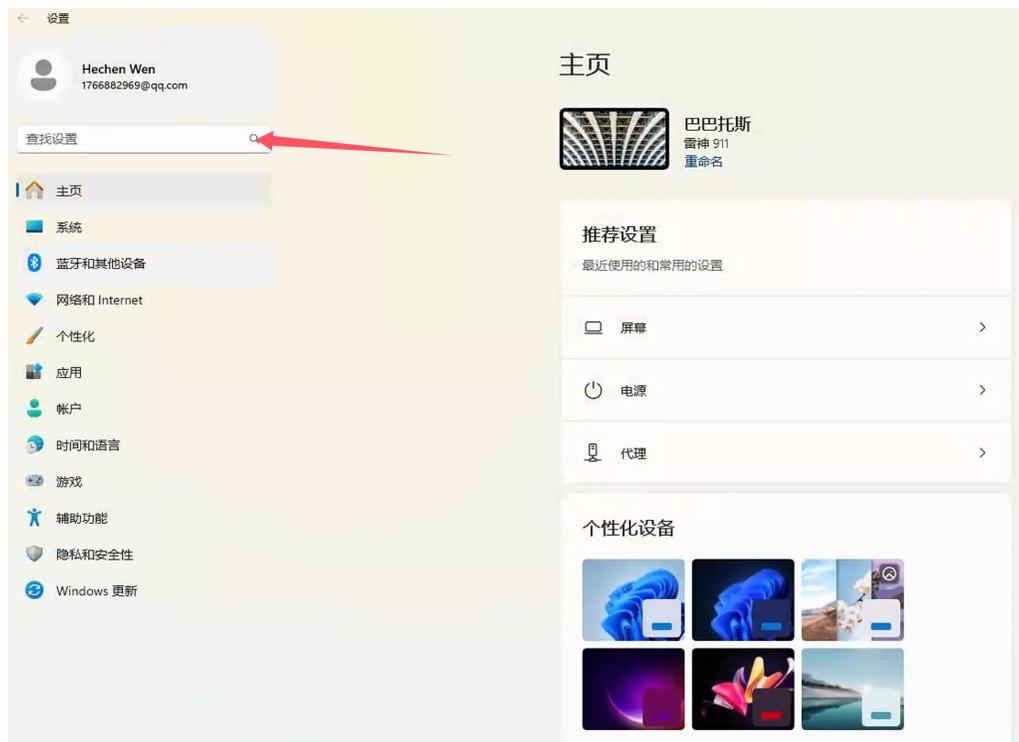
1. 通过任务栏退出 Ollama。
2. 打开设置（Windows 11）或控制面板（Windows 10），并搜索“环境变量”。
3. 点击编辑你账户的环境变量。

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。

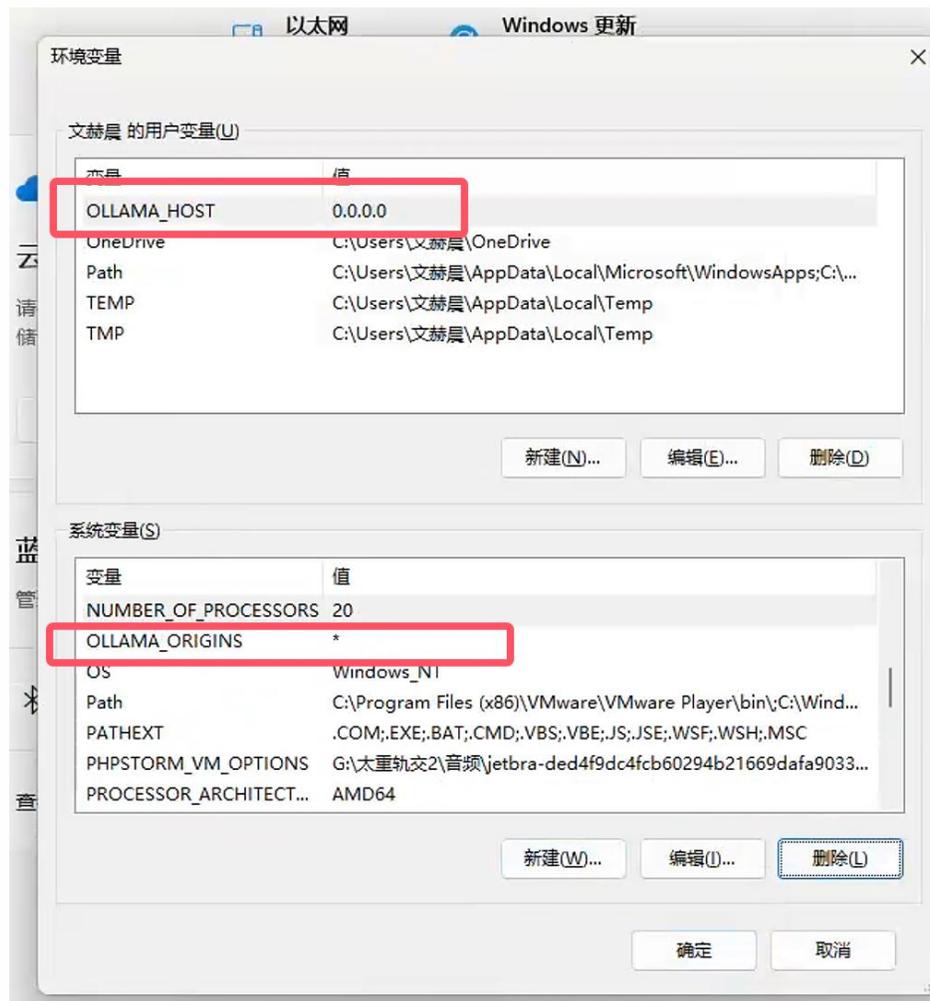
Chatbox对话框页面

环境变量



Chatbox对话框页面

环境变量



Chatbox对话框页面

设置

模型 显示 对话 其他

模型提供方:
OLLAMA API

API 域名
http://127.0.0.1:11434 重置

使用 Chatbox 桌面版获得更好的连接性和稳定性。立即下载。
请确保远程 Ollama 服务能够远程连接。更多详情请参考[此教程](#)。

模型

上下文的消息数量上限
20

严谨与想象(Temperature)
0.7
严谨细致 想象发散

取消 保存

- Chatbox
- 对话
- Markdown 101 (Exam... ★
 - Software Developer (... ★
 - Untitled
 - Untitled
 - Translator (Example)
 - Social Media Influencer (E...
 - Travel Guide (Example)
 - 翻译助手 (示例)
 - 夸夸机 (示例)

Untitled

You are a helpful assistant.

合肥经济介绍

合肥，这座位于中国中部的城市，近年来在经济领域展现出了强劲的发展势头。以下是对合肥经济情况的详细介绍：

工业生产：2023年，合肥市全部工业增加值4642.21亿元，比上年增长7.1%，其中规模以上工业增加值增长10.6%。规模以上工业总产值下降1.9%，产销率98.1%。规模以上工业企业实现利润473.46亿元，增长

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

The image shows a composite view of the gpt4all website and its installer. On the left, the website header includes the 'NOMIC gpt4all' logo and navigation links for 'Enterprise', 'Blog', 'Community', and 'Docs'. The main content area features the heading 'Run Large Language Models Locally' and buttons for 'Download for Windows', 'Download for Windows ARM', 'Download for macOS', and 'Download for Ubuntu'. A 'Contact Sales' button is also visible. In the center, a preview window shows the GPT4All application interface with a chat history on the left and a document viewer on the right. On the right side, a separate window titled 'GPT4All Installer 安装程序' is shown. It displays the progress of the installation, which is at 17%. The text in the installer window reads: '正在安装 GPT4All', '正在下载组件 gpt4all 的存档“3.9.0resources.7z”', '存档: 4.43/234.12 MB (889.04 KB/秒) - 剩余 4 分钟, 。', and '总计: 41.61/271.35 MB - 剩余 1 分钟, 。'. There is a '显示详细信息(S)' button and a progress bar at the bottom with '安装(0)号' and '取消 46'.

GPT4All方式安装

第二步：下载 DeepSeek 模型

The screenshot shows the GPT4All web interface. The main navigation bar includes '开始聊天' (Start Chat), '本地文档' (Local Docs), and '查找模型' (Find Models). The '查找模型' button is highlighted with a red box and labeled '2. 查找模型'. Below this, the 'Latest News' section provides updates on model releases and fixes. The '发现模型' (Discover Models) section is highlighted with a red box and labeled '3. 查看模型信息'. It lists available models, with 'DeepSeek-R1-Distill-Qwen-7B' highlighted by a red box and labeled '4. 下载模型'. The '下载' (Download) button for this model is also highlighted with a red box and labeled '4. 下载模型'.

欢迎
隐私至上的大模型咨询应用程序

开始聊天
大语言模型聊天

本地文档
本地文件聊天

查找模型
发现并下载模型

2. 查找模型

Latest News

GPT4All v3.9.0 was released on February 4th. Changes include:

- **LocalDocs Fix:** LocalDocs no longer shows an error on later messages with reasoning models.
- **DeepSeek Fix:** DeepSeek-R1 reasoning (in 'think' tags) no longer appears in chat names and follow-up questions.
- **Windows ARM Improvements:**
 - Graphical artifacts on some SoCs have been fixed.
 - A crash when adding a collection of PDFs to LocalDocs has been fixed.
- **Template Parser Fixes:** Chat templates containing an unclosed comment no longer
- **New Models:** OLMoE and Granite MoE models are now supported.

GPT4All v3.8.0 was released on January 30th. Changes include:

- **Native DeepSeek-R1-Distill Support:** GPT4All now has robust support for the D
 - Several model variants are now available on the downloads page.
 - Reasoning (wrapped in "think" tags) is displayed similarly to the Reasoner mo
 - The DeepSeek-R1 Qwen pretokenizer is now supported, resolving the loading
 - The model is now configured with a GPT4All-compatible prompt template by
- **Chat Templating Overhaul:** The template parser has been *completely* replaced v compatibility with common models.
- **Code Interpreter Fixes:**

存在的模型

发现模型

GPT4All HuggingFace

These models have been specifically configured for use in GPT4All. The first few models on the list are known to work the best, but you should only attempt to use models that will fit in your available memory.

All Reasoning

3. 查看模型信息

DeepSeek-R1-Distill-Qwen-7B

The official Qwen2.5-Math-7B distillation of DeepSeek-R1.

- License: MIT
- No restrictions on commercial use
- #reasoning

文件大小	RAM required	参数	量化	类型
4.14 GB	8 GB	7 billion	q4_0	deepseek

4. 下载模型

DeepSeek-R1-Distill-Qwen-14B

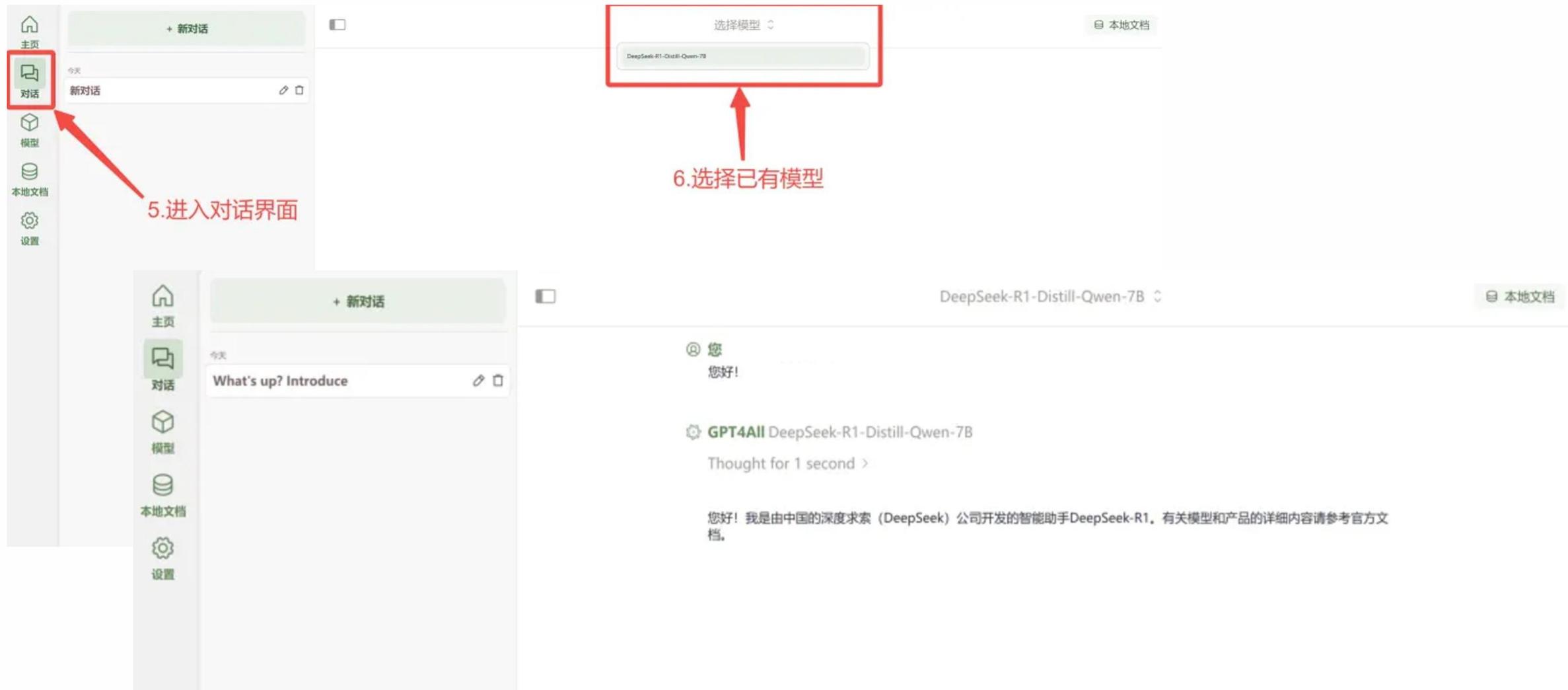
The official Qwen2.5-14B distillation of DeepSeek-R1.

- License: MIT
- No restrictions on commercial use
- #reasoning

文件大小	RAM required	参数	量化	类型
7.96 GB	16 GB	14 billion	q4_0	deepseek

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 \geq 200 GB
- DeepSeek-R1-Q4_K_M: 内存 + 显存 \geq 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

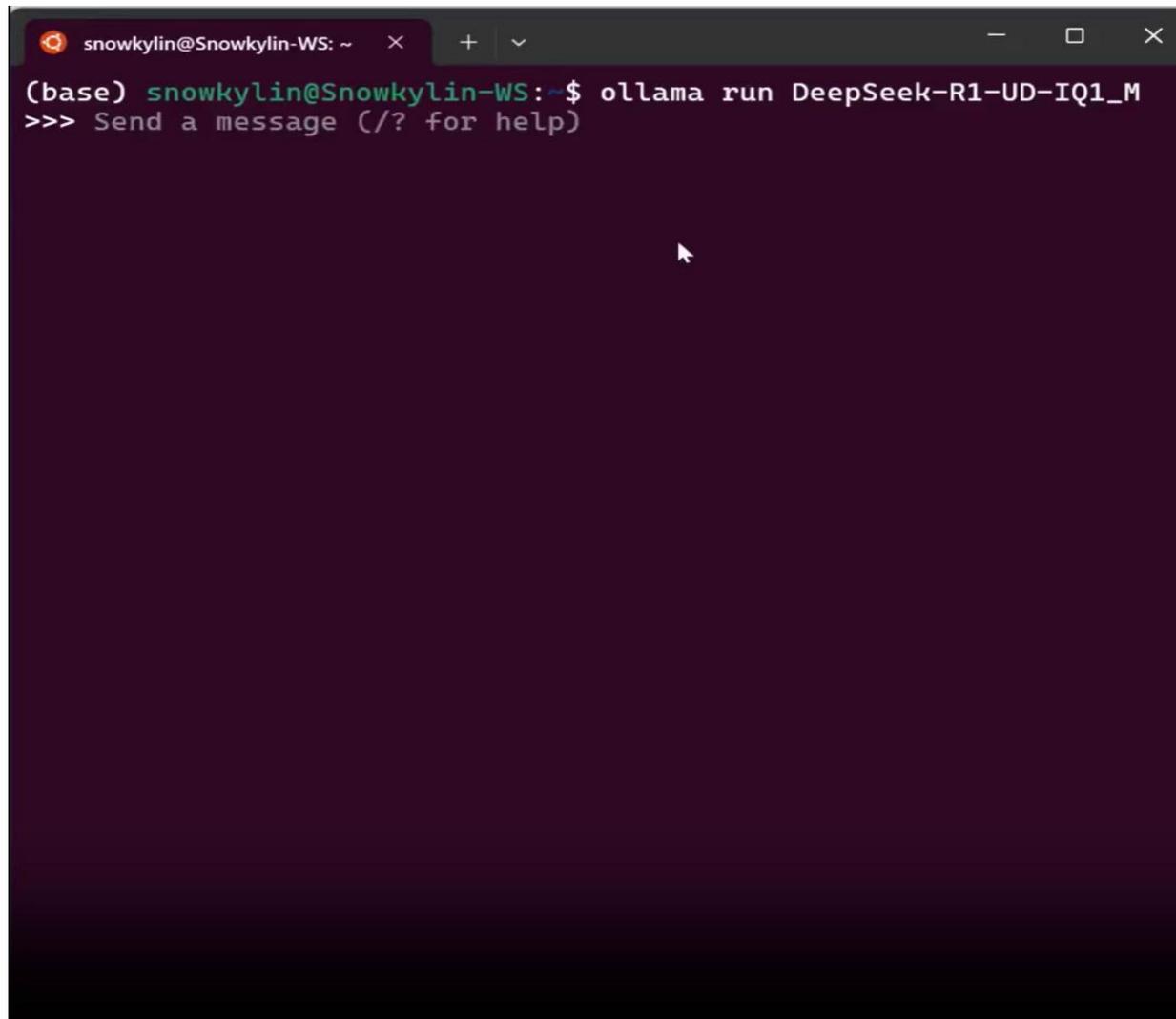
6. (可选) 安装 Web 界面

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

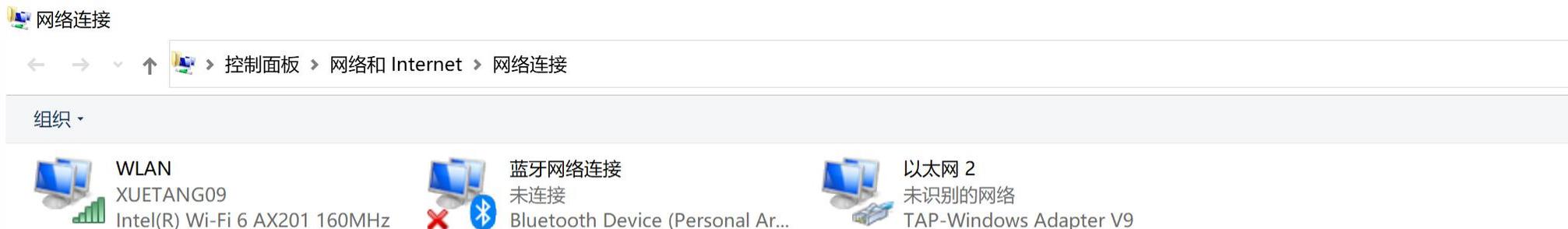


```
snowkylin@Snowkylin-WS: ~  
(base) snowkylin@Snowkylin-WS:~$ ollama run DeepSeek-R1-UD-IQ1_M  
>>> Send a message (/? for help)
```

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

我们在出站规则程序这里添加出站规则

新建出站规则向导 ×

规则类型

选择要创建的防火墙规则类型

步骤:

- 规则类型
- 程序
- 操作
- 配置文件
- 名称

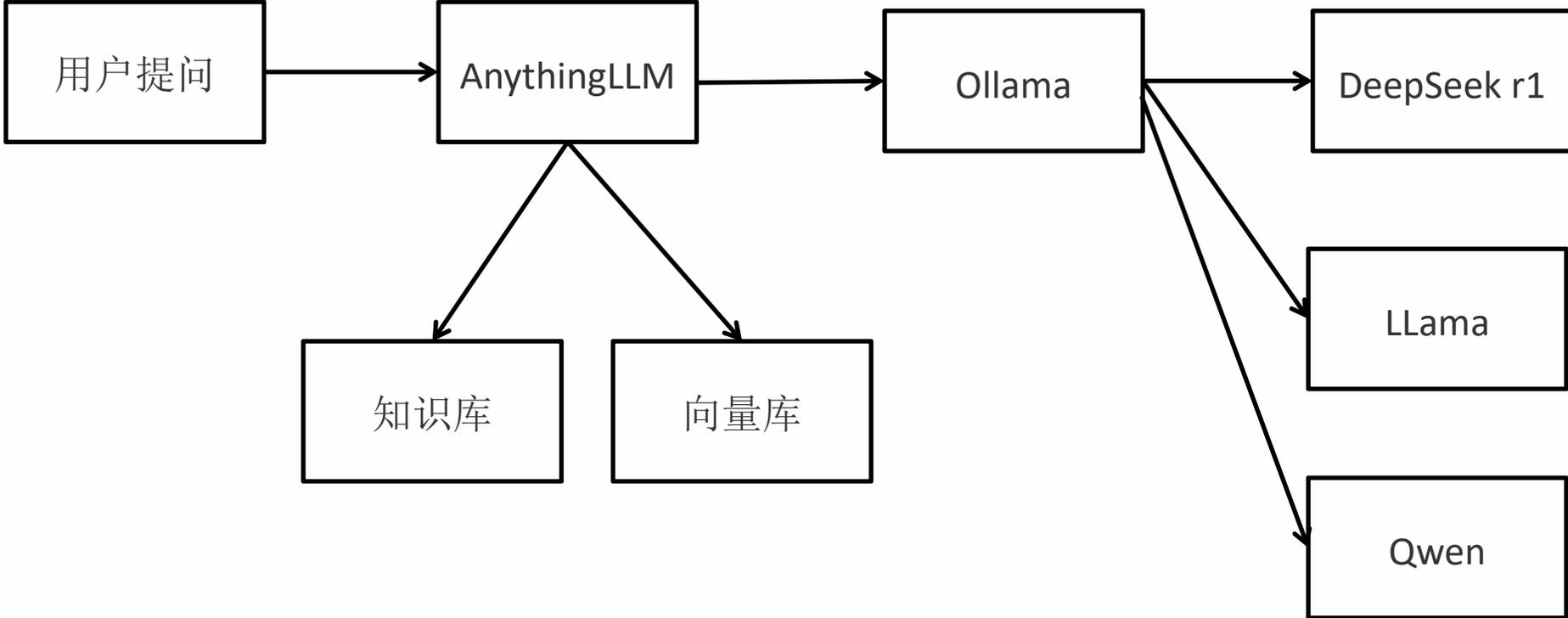
要创建的规则类型

- 程序 (P)**
控制程序连接的规则。
- 端口 (O)**
控制 TCP 或 UDP 端口连接的规则。
- 预定义 (E):**
@FirewallAPI.dll, -80200
控制 Windows 体验功能连接的规则。
- 自定义 (C)**
自定义规则。

< 上一步 (B) 下一页 (N) > 取消

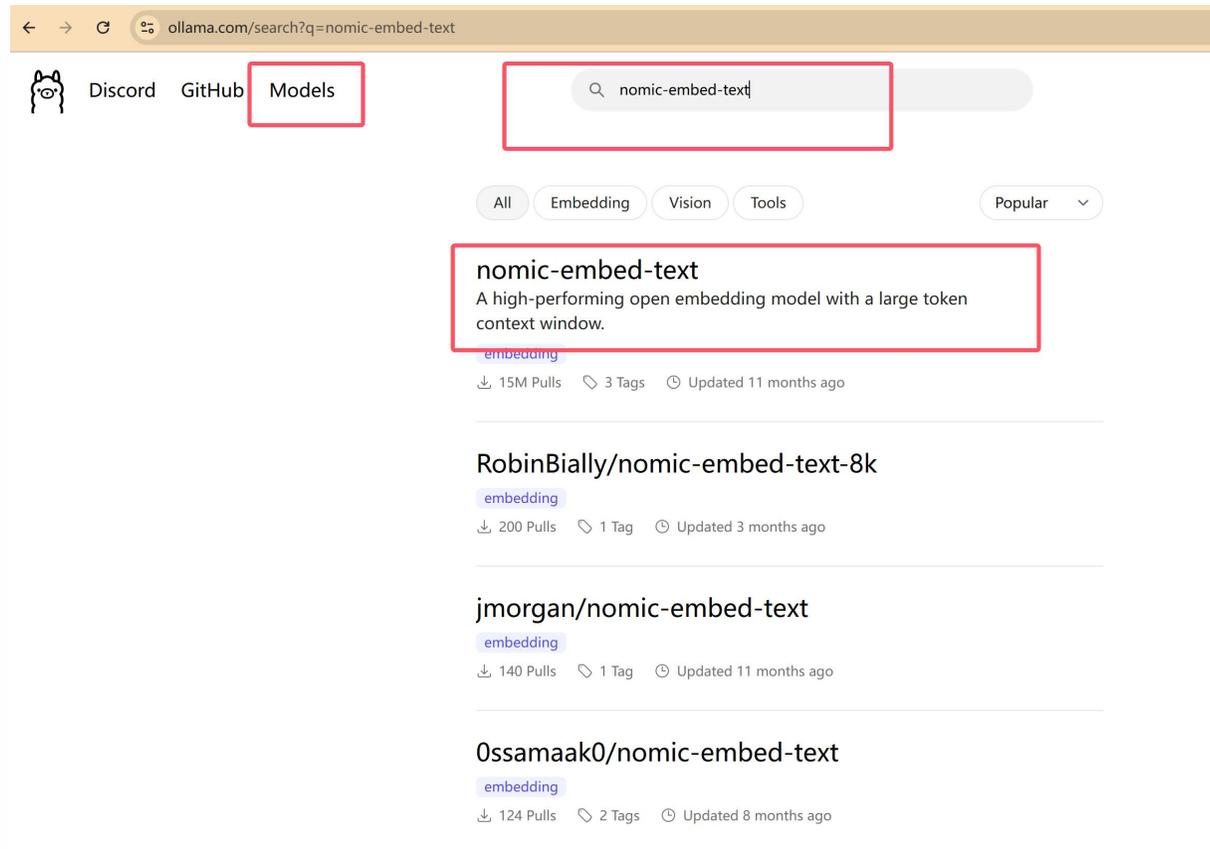
本地知识库系统的搭建

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



Models

Search models

nomic-embed-text

A high-performing open embedding model with a large token context window.

embedding

↓ 15M Pulls Updated 11 months ago

latest

3 Tags

ollama pull nomic-embed-text



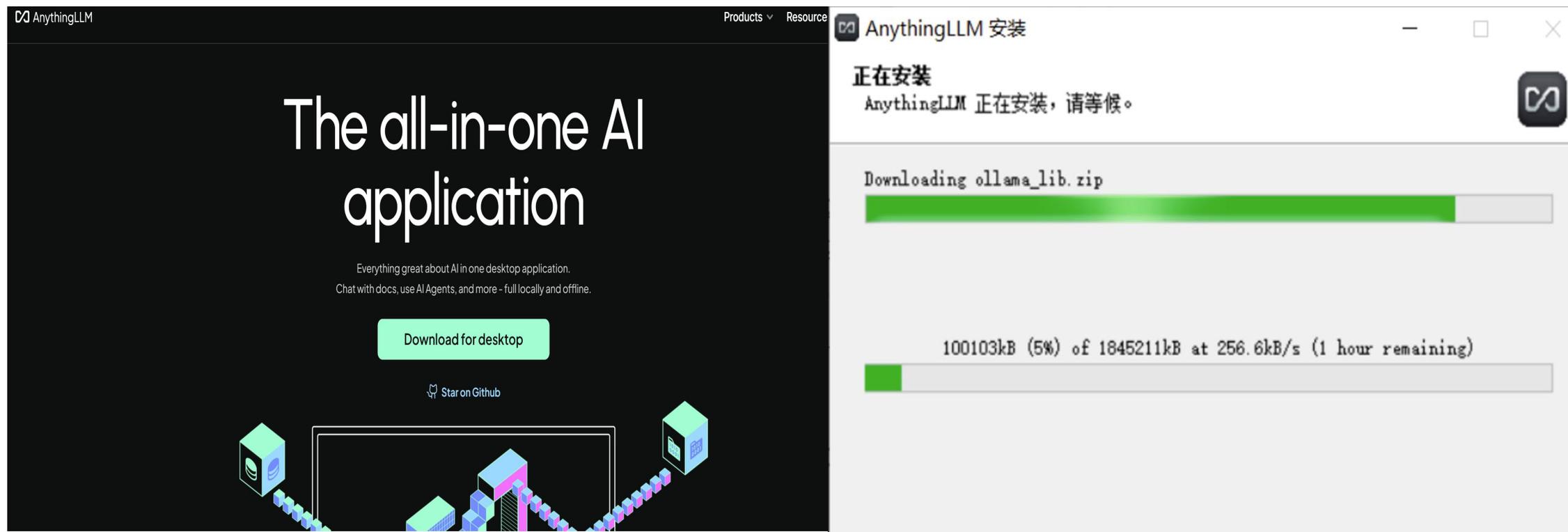
Updated 11 months ago	0a109f422b47 · 274MB
model	arch nomic-bert · parameters 137M · quantization F16 274MB
params	{ "num_ctx": 8192 } 17B
license	Apache License Version 2.0, January 2004 11kB

Readme

基于AnythingLLM的本地知识库与API搭建

下载AnythingLLM Desktop

官网地址: <https://anythingllm.com/>



The image shows two overlapping windows. The background window is the AnythingLLM website, which features a dark theme and the text "The all-in-one AI application". Below this, it says "Everything great about AI in one desktop application. Chat with docs, use AI Agents, and more - full locally and offline." There is a prominent green button labeled "Download for desktop" and a "Star on Github" link. The foreground window is the "AnythingLLM 安装" (AnythingLLM Installation) window. It displays the status "正在安装" (Installing) and "AnythingLLM 正在安装, 请等候。" (AnythingLLM is installing, please wait). A progress bar shows the download of "ollama_lib.zip" is approximately 5% complete. Below this, another progress bar indicates the download of a large file: "100103kB (5%) of 1845211kB at 256.6kB/s (1 hour remaining)".

基于AnythingLLM的本地知识库与API搭建

安装完成后



The screenshot displays the AnythingLLM web application interface. On the left is a sidebar with a header 'Anything LLM' and a toggle switch. Below the header is a blue button labeled '+ 新工作区' (New Workspace). Underneath that is a search bar containing the text 'work'. The main content area on the right features a chat interface with a welcome message from a bot icon, followed by three informational messages from the bot, a button to '在 Github 上创建问题' (Create issue on Github), a user question '我该如何开始?!', a bot response explaining the workspace concept, a button to '+ 创建您的第一个工作区' (Create your first workspace), and a final user question '这像是一个 AI Dropbox 吗? 那么聊天呢? 它是一个聊天机器人, 不是吗?'.

Anything LLM

+ 新工作区

work

欢迎使用 AnythingLLM, 这是由 Mintplex Labs 开发的开源 AI 工具, 可以将任何东西转换为您可以查询和聊天的训练有素的聊天人。AnythingLLM 是一款 BYOK (自带密钥) 软件, 因此除了您想使用的服务外, 此软件不收取订阅费、费用或其他费用。

AnythingLLM 是将强大的 AI 产品 (如 OpenAi、GPT-4、LangChain、PineconeDB、ChromaDB 等) 整合在一个整洁的包中并繁琐操作的最简单方法, 可以将您的生产力提高 100 倍。

AnythingLLM 可以完全在您的本地计算机上运行, 几乎没有开销, 您甚至不会注意到它的存在! 无需 GPU。也可以进行云端和本装。AI 工具生态系统每天都在变得更强大。AnythingLLM 使其易于使用。

在 Github 上创建问题

我该如何开始?!

很简单。所有集合都组织成我们称之为“工作区”的桶。工作区是文件、文档、图像、PDF 和其他文件的存储桶, 这些文件将被转: LLM 可以理解和在对话中使用的内容。您可以随时添加和删除文件。

+ 创建您的第一个工作区

这像是一个 AI Dropbox 吗? 那么聊天呢? 它是一个聊天机器人, 不是吗?

基于AnythingLLM的本地知识库与API搭建

创建工作区，进行设置

Anything LLM

+ 新工作区

work



default

+ New Thread



通用设置

聊天设置



向量数据库



代理配置

工作区 LLM 提供者

将用于此工作区的特定 LLM 提供商和模型。默认情况下，它使用系统 LLM 提供程序和设置。



Ollama

Run LLMs locally on your own machine.



工作区聊天模型

将用于此工作区的特定聊天模型。如果为空，将使用系统 LLM 首选项。

deepseek-r1:1.5b



聊天模式

聊天

查询

聊天 将提供 LLM 的一般知识 **和** 找到的文档上下文的答案。

基于AnythingLLM的本地知识库与API搭建

软件设置

The image displays the 'AnythingLLM' software settings interface, divided into three main sections:

- Left Panel (Main Interface):** Shows the workspace management area. It includes a '+ 新工作区' (New Workspace) button, a workspace named 'work' with a settings icon, and a thread named 'default'. Below these is a '+ New Thread' button. At the bottom, there is a navigation bar with icons for home, chat, community, and settings (the settings icon is highlighted with a red box).
- Middle Panel (Settings Menu):** Titled '设置' (Settings), it lists various configuration options:
 - 人工智能提供商 (AI Provider)
 - LLM 首选项 (LLM Preferences)
 - 向量数据库 (Vector Database)
 - Embedder 首选项 (Embedder Preferences)
 - 文本分割 (Text Splitting)
 - 语音和讲话 (Voice and Speech)
 - 转录模型 (Transcription Model)
 - 管理员 (Admin)
 - 代理技能 (Agent Skills)
 - Community Hub
 - 外观 (Appearance)** (highlighted with a red box)
 - 工具 (Tools)
 - 实验功能 (Experimental Features)
- Right Panel (Appearance Settings):** Titled '外观' (Appearance), it provides options to customize the platform's look:
 - 自定义平台的外观设置。 (Customize the appearance of the platform.)
 - Theme:** Select your preferred theme for the application. A dropdown menu shows 'Light'.
 - Display Language:** Select the preferred language to render AnythingLLM's UI in. A dropdown menu shows 'Chinese' (highlighted with a red arrow).
 - Show chat window scrollbar:** Enable or disable the scrollbar in the chat window. A toggle switch is currently turned off.
 - 自定义消息 (Custom Messages):** 自定义向用户显示的自动消息。 (Customize automatic messages displayed to users.)

基于AnythingLLM的本地知识库与API搭建

上传文档

The image displays the AnythingLLM interface for document management. On the left, a sidebar shows the 'work' workspace selected, with an upload icon highlighted. The main area, titled 'My Documents', contains a search bar, a '+ New Folder' button, and a list of documents. A red box highlights the 'custom-documents' folder. Below the list, a 'Move to Workspace' button is also highlighted. At the bottom, a dashed box contains an upload icon and the text 'Click to upload or drag and drop' and 'or submit a link'. On the right, a 'work' workspace is shown as an empty list with a 'Name' header. A double-headed arrow indicates the relationship between the document list and the workspace.

AnythingLLM

+ 新工作区

work

default

+ NewThread

My Documents

Search for document

+ New Folder

Name

custom-documents

Move to Workspace

Click to upload or drag and drop

supports text files, csv's, spreadsheets, audio files, and more!

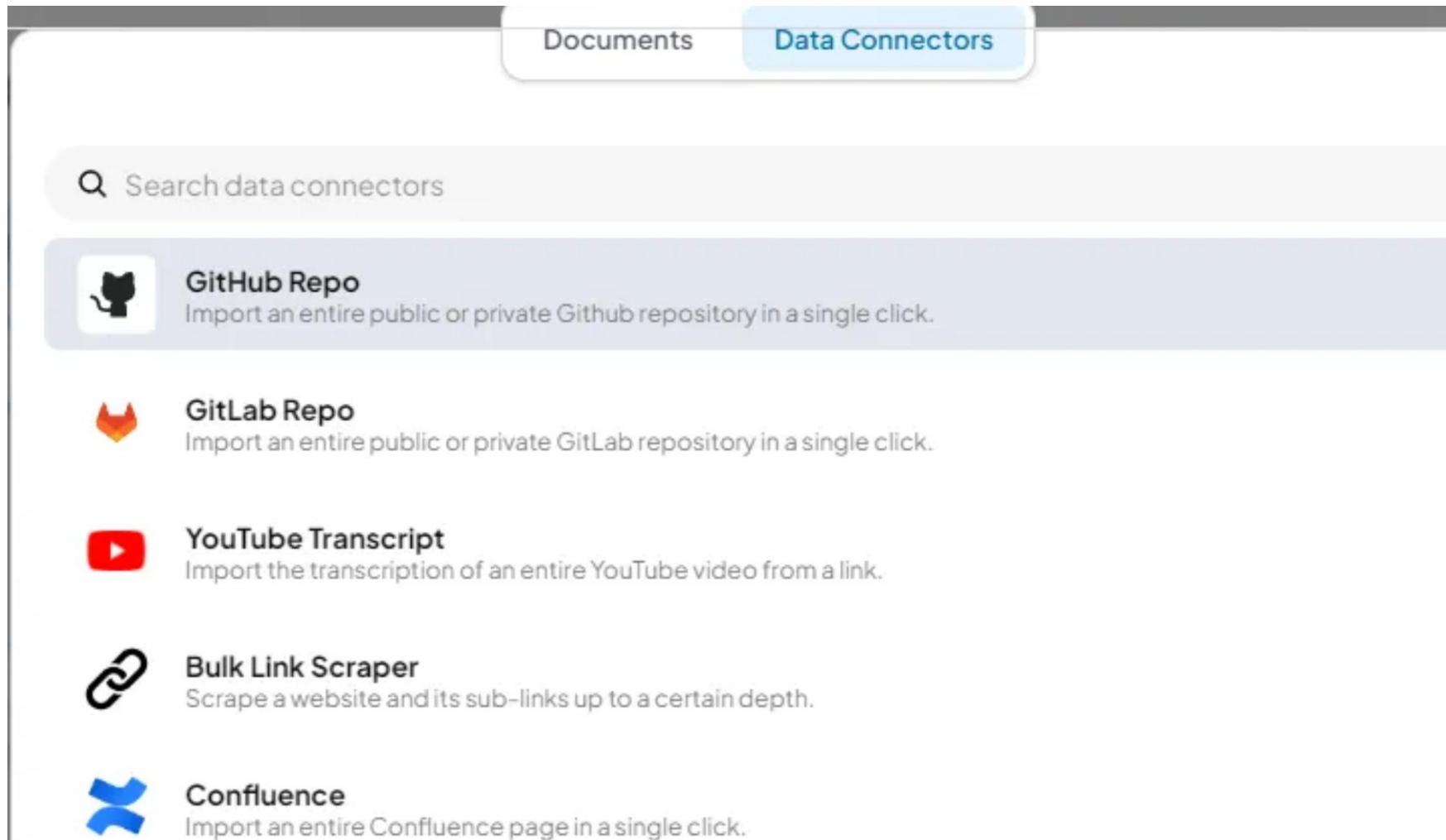
or submit a link

work

Name

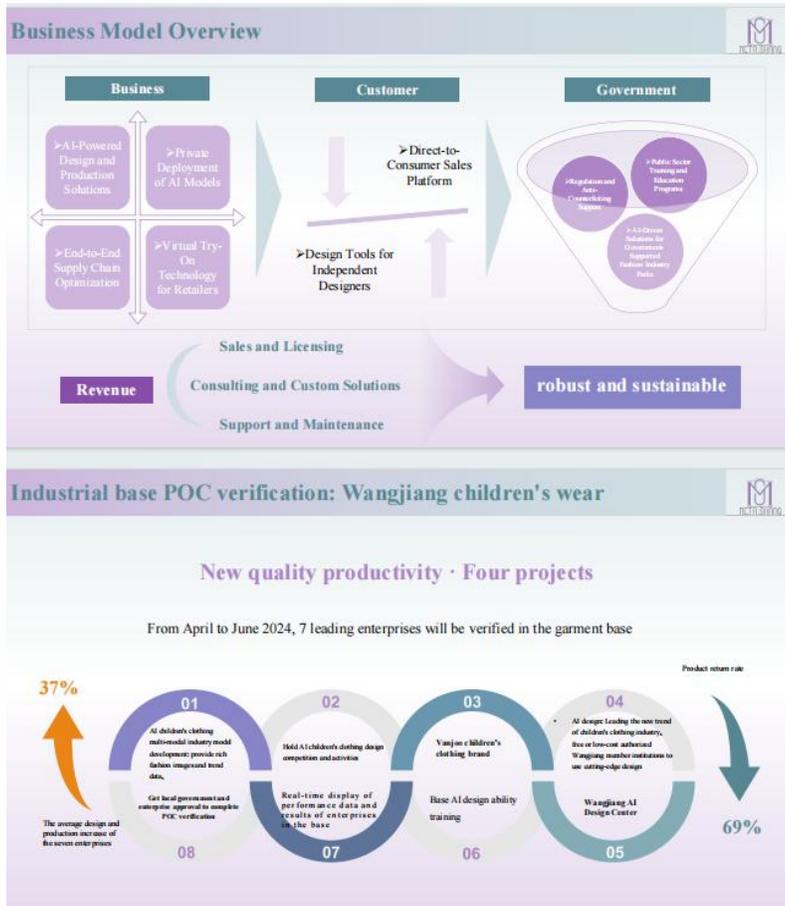
基于AnythingLLM的本地知识库与API搭建

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



知识库 / 数据集

数据集

🔔 解析成功后才能问答哦。

批量 ▾

🔍 搜索文件

+ 新增文件

<input type="checkbox"/>	名称	分块数	上传日期	解析方法	启用	解析状态	动作
<input type="checkbox"/>	元鉴任务及人员...	11	11/02/2025 10:30:56	General	<input checked="" type="checkbox"/>	成功	🔄 ✎ 🗑️ ⬇️
<input type="checkbox"/>	metashang_bp...	31	11/02/2025 10:14:28	General	<input checked="" type="checkbox"/>	成功	🔄 ✎ 🗑️ ⬇️
<input type="checkbox"/>	2024珠海金湾投...	60	11/02/2025 10:02:23	General	<input checked="" type="checkbox"/>	成功	🔄 ✎ 🗑️ ⬇️

总共 3 < 1 > 10 条/页 ▾

帮我分析下投资价值

在集成电路产业方面，奕达科技和量旺电子是区域内的优质项目，并且计划推进半导体材料研发。对于MetaShang来说，如果他们有相关业务，这可能是一个很好的合作机会。此外，先进装备制造业也值得关注，比如格力电器在该区的大规模投资和生产线布局。经济数据方面，2022年的新材料产业产值达到829亿元，而集成电路产业产值为55.7亿元。这些数字显示了金湾区在高端制造业的强劲发展势头。如果MetaShang能找到与当地产业链相契合的点，可以有效利用现有的资源和市场需求。政策支持也是一个重要因素，作为国家级开发区，当地政府通常会提供税收优惠、土地补贴等激励措施，这对于降低企业运营成本非常有帮助。MetaShang可以考虑如何利用这些政策来提高盈利能力。另外，金湾区的交通网络包括港澳大桥、珠海国际机场和临港铁路，这些完善的物流设施能够有效降低运输成本，加快产品上市速度。这对于依赖供应链效率的企业来说非常重要，可以提升整体竞争力。综合以上分析，我认为MetaShang在金湾区投资有以下几个潜力的方面：
产业契合：如果MetaShang涉及新能源、储能或集成电路相关业务，金湾区现有的产业基础和资源将为其提供良好的发展环境。
市场需求：高产值的新材料和IC产业显示出强劲的市场需求，MetaShang可以针对这些领域开发产品或服务，以满足当地及周边地区的需求。
政策优势：利用国家级开发区提供的税收优惠和其他激励措施，可以降低运营成本，提高投资回报率。
物流便利：完善的交通网络有助于MetaShang在供应链管理、市场拓展方面更具竞争力，加快产品推向国内外的速度。当然，在做出最终决定之前，还需要进一步了解MetaShang具体业务是什么，是否与金湾区已有的产业相契合。此外，也应评估当地的人才资源、技术支持和潜在合作伙伴，以确保投资的可行性和成功率。
基于以上分析，以下是针对MetaShang公司在珠海金湾区投资潜力的详细建议：
明确业务定位：首先需要了解MetaShang的核心业务是什么，是新能源、储能、高端制造还是集成电路等领域。根据具体业务选择最适合的合作伙伴和市场机会。
利用产业资源：如果涉及新能源或储能，寻求与珠海银隆、新宙邦等企业的合作，共享技术研发和供应链优势。在集成电路领域，与奕达科技、量旺电子建立战略合作关系，推动半导体材料和产品的创新。
申请政策支持：积极争取国家级开发区提供的税收优惠、土地补贴等政策措施，以降低初期投资成本，提高项目可行性。
优化物流与供应链：利用港澳大桥、珠海国际机场和临港铁路等完善交通网络，加快产品运输速度，扩展市场覆盖范围。建立高效的供应链管理系统，以应对快速变化的市场需求。
评估人才与技术支持：考察当地的人才资源是否能满足企业发展需要，如是否有足够的专业技能人员和研发能力。利用金湾区内的科技园区、创新平台等资源，促进技术交流和产品创新。
市场调研与需求分析：

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
- **安全风险控：**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

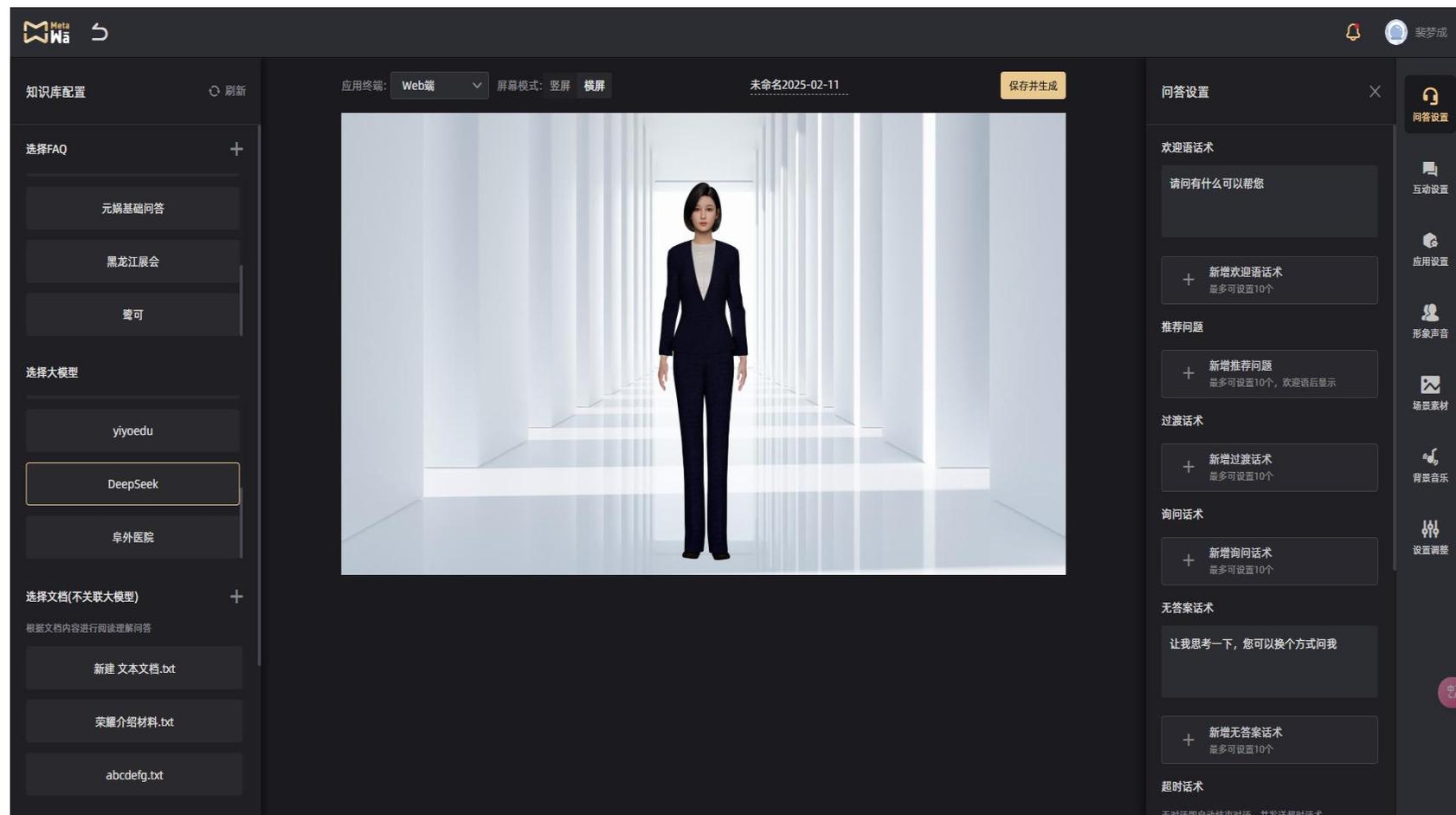
实际应用场景

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。

01



FAQ问答知识库

02



技能问答知识库

03



AIGC能力应用

04



大语言模型

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。

元镜

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