**人工智能入门指南**

**从基础概念到DeepSeek本地部署**

**知识库对比测评**

提示词（Prompt）：请为的人工智能的初学者写一篇科普文章，首先介绍一些基本概念，例如什么是：人工智能、机器学习、深度学习、神经网络、CNN、AlexNet、EfficientV2、RNN、transformer、pytorch、TensorFlow、ollama、deepseek、AnythingLLM等等。然后重点讲解使用ollama安装deepseek-r1、使用anythingLLM建立本地知识库。

提示词（Prompt）：请用python写一段绘图程序，需要准确表述以下人工智能专业术语之间的关系，不同级别之间使用不同的颜色进行区分。绘图内容至少包括以下专业术语：人工智能、机器学习、深度学习、神经网络、CNN、AlexNet、EfficientV2、RNN、transformer、deepseek、ollama、anythingLLM、pytorch、TensorFlow。

【PS】<think>过程解读、答案解读、AlexNet（80.2%）、马斯克模型（训练日志）、TensorFlow（AlphaGo）解读<think>内部过程、答案修订（有的软件是不具备这个功能的）、提问技巧（越具体越好）

一、下载安装Ollama

Bug1：在官网（ollama.com）无法下载Ollama。

复制网址>迅雷（新建）>立即下载



3、双击OllamaSetup.exe自动安装到了C盘：

C:\Users\Administrator\AppData\Local\Programs\Ollama\ollama app.exe

4、Win+R打开运行窗口，输入cmd启动命令行，输入ollama -v显示Ollama的版本号表示安装成功

 



二、设置环境变量：

Bug2：下载的deepseek模型安装到了C盘，未修改环境变量的地址：

C:\Users\Administrator\.ollama\models\blobs

1、退出ollama，系统默认启动Ollama。

如果不需要自动启动剪切到桌面，后面需要时再粘贴回去，启动的地址是：

C:\Users\Administrator\AppData\Roaming\Microsoft\Windows\Start Menu\Programs\Startup

2、我的电脑右键》高级系统设置》高级》环境变量》系统变量增加：

变量名OLLAMA\_MODELS值F:\DeepSeekV3\models

3、启动ollama

三、安装DeepSeek-R1：

1、概述：DeepSeek-R1模型671B（6710亿）参数量，蒸馏（Distill）出了六个小模型，参数从小到大分别为1.5B、7B、8B、14B、32B以及70B。

|  |  |  |  |
| --- | --- | --- | --- |
| **模型名称** | **推理性能** | **所需内存** | **底精度** |
| DeepSeek-R1-1.5B | GPT4o级 | 4G | 1.1G |
| DeepSeek-R1-7B | 超过GPT4o级 | 14G | 4.7G |
| DeepSeek-R1-8B | 超过GPT4o级 | 14G | 4.9G |
| DeepSeek-R1-14B | 超过GPT4o级 | 24G | 9G |
| DeepSeek-R1-32B | o1min级 | 55G | 20G |
| DeepSeek-R1-70B | o1min级 | 120G | 43G |
| DeepSeek-R1-671B | o1级 | 1000G | 404G |

注：B：模型的参数单位Billion（十亿），参数越大AI越聪明，要求硬件（显卡）越高。

Token：计算机处理语言的计量单位。通常 1 个中文词语、1 个英文单词、1 个数字或 1 个符号计为 1 个 token。

2、效果测试

Bug3：电脑配置低或安装Deepseek的版本不合适。所以先要进行测试，选一个自己最擅长的问题提问，以便测试答案效果。例如：请简述一下润滑原理（我们后面所有的测试都基于这个问题）

（1）笔记本电脑：Dell E7450，CPU是 i7-5600U ，内存16G，无显卡。

测试结果：1.5B运行流畅，答案效果差。7B非常慢，答案不满意。

（2）台式电脑：手动配置：CPU是E5-2696 V3，显卡RTX4060（8G），内存64G。

测试结果：运行7B流畅，但是准度不够（比文档搜索功能强一点）；运行14B慢（比较理想的答案），准度不错（显存比较小8G，去吃内存了）。32B很卡，基本不能用。

3、安装DeepSeek-R1：

Win+R启动命令行，输入cmd打开命令运行窗口，ollama run DeepSeek-R1:7B，安装完成后显示send message（/? For help）表示安装成功。

Bug4：文件太大，下载过程死机解决方案。当cmd窗口死机时，不要关闭，再打开一个cmd窗口，运行同样的命令，下载会继续往下走，再关闭死机的窗口。

4、输入你是谁测试deepseek是否安装成功。如果系统有答复，表示安装成功。部分电脑Ollama运行不正常，需要重启一下电脑。

Bug5：想要终止答复过程和跳出deepseek运行环境，快捷键：Ctrl+C

三、安装Cuda：

Bug6：大师们都默认我们安装了Cuda。

1、去Nidia官网https://www.nvidia.cn/注册账号并下载安装。

2、命令行窗口测试命令nvcc --version 或 nvcc -V

四、安装Chatbox：

1、官网https://chatboxai.app/zh下载安装，配置API

五、搭建本地知识库：

1、检索增强生成RAG技术（Retrieval-Augmented Generation），到以下网站下载AnythingLLM软件：https://anythingllm.com/desktop

2、新建知识库，重点讲文档要求（Markdown>DOCX>PDF(文字)>TXT）。

3、对比默认知识库、本地知识库、官网知识库：

问题1：请介绍一下润滑油铁谱分析

问题2：MOBIL DTE 846的粘度是多少？

4、如果知识聊天，不建立本地知识库，可用ChatBox、Open-WEBUI等

【PS】文档结构层次清晰（段落、标题）、Markdown（无需LLM处理器）、DOCX 样式规范、PDF文字版（扫描的需要OCR）；文档大小（官方1-3万字，默认8000字）

六、结论：

1、大型企业资金充裕，可以建立本地数据库。推理能力强，保密性强。

2、一般企业、个人用API。

七：联系方式：



抖音：[大能讲润滑](https://v.douyin.com/iyYtjYDk/)



好润滑官微 [好润滑技术网网址](https://haorunhua.com/)

八、解决问题

Ollama无法下载

官网？本地部署？第三方？

行业知识深度对比测评