DeepSeek 本地部署实用指南

四川强民科技有限公司

2025年2月

目录

开篇序言：1
第一章：入门篇（基础概念）2
一、什么是 Deep seek?2
二、Deepseek和其他的AI大模型有什么不同？3
第二章：基础篇（安装和使用）6
一、Deepseek有哪些使用方式？6
第三章：进阶篇（玩转DeepSeek）16
【手机APP端】16
【电脑网页端】18
【远程API调用】19
【本地部署】23
一、 下载Chatbox AI23
二、 设置界面的语言为简体中文23
三、 设置环境变量24
四、 设置模型25
五、 本地测试和使用26
第四章：专业篇（服务器部署及调优）27
一、系统环境准备27
1. 安装NVIDIA驱动27
2. 安装CUDA Toolkit 12.327
3. 安装NVIDIA Container Toolkit28
二、Ollama 环境部署28
1. 安装Ollama28
2. 配置多GPU支持28
三、模型部署与优化28
1. 模型文件准备28
2. 编写Modelfile29
3. 优化内核参数29
四、模型加载与启动29
1. 加载模型29
2. 启动服务29
五、验证与监控30
1. 资源监控命令30
2. API调用验证30
六、高级优化建议30
1. Kernel参数调优30
2. Ollama性能参数配置 在~/.ollama/config.json添加：30
七、常见问题排查31
1. 显存OOM处理31
2. 性能调优工具31
第二种安装方式31

一、部署前关键检查清单 31
二、硬件特性深度适配 32
三、逐步部署流程 32
四、针对4090集群的关键配置调优 34
五、验证与压测 35
六、高级运维监控 36
七、紧急故障处理 36
八、进一步性能调优（专家模式） 37
第五章：产品篇（网昱DeepSeek大模型一体机推荐） 38
一、模型规格与硬件要求总表 38
二、网昱大模型一体机推荐 38
DeepSeek R1 7B 静音工作站配置推荐一 38
DeepSeek R1 14B 静音工作站配置推荐二 39
DeepSeek R1 32B 静音工作站配置推荐三 39
DeepSeek R1 70B 服务器配置推荐四 39
DeepSeek R1 70B 服务器置推荐五 40
第6章：行业篇（DeepSeek大模型一体机行业应用推荐） 40
一、金融行业 41
二、医疗健康 42
三、科研领域 42
四、智能制造 43

开篇序言：

在人工智能触手可及的时代，技术革新正以前所未有的速度重塑我们的生活。从工作场景的效率提升到日常沟通的智慧延展，大语言模型已悄然成为新时代的"数字氧气"，而 DeepSeek 正是这片智能生态中一颗不容忽视的启明星。

本指南的诞生不是为了筑起技术的高墙，而是希望搭建一座通向未来的桥梁，为您的工作提升或是生活的丰富提供指引。无论您是驾驭代码的工程师、热衷数字工具的极客，还是渴求效率突破的职场人士，亦或是满怀好奇心的学生及爱好者，当您翻开这本指南时，就已踏上了一场平等且充满惊喜的探索之旅。我们不预设任何门槛——您无需储备艰深的数学公式，不必纠结复杂的代码逻辑，只需怀揣解决问题的热忱与探索未知的勇气。

与市场上大多数技术手册不同，本指南将带您亲历从"用户"到"创造者"的完整跃迁。我们将从最本质的问题开始剖析：您将理解 DeepSeek 如何像拥有超强记忆的智慧伙伴般运转，见证它在文本处理、创意激发、数据分析等领域的独特优势；随后通过详尽的场景化指导，带您掌握多终端操作的精髓——无论是通过网页端快速启动工作流程，还是在通勤路上通过移动端 APP 延续创作灵感；更将揭开开发者生态的神秘面纱，带您突破 API 调用的技术边界，直至完成本地化部署的终极进阶。

我们始终坚信，真正伟大的技术应该如同流水般渗透每个可能的缝隙。因此，本指南的每一章节都经过真实用户的场景验证，每处操作指引都凝聚了数百次测试的优化沉淀。当您跟随指南逐步解锁 DeepSeek 的无限可能时，或许会惊喜地发现：曾经局限于专业领域的 AI 技术，正在您的手中蜕变为提升生产力的瑞士军刀、激发创意的思维催化剂，甚至是重构工作方式的破界魔方。

此刻，请您暂时放下对技术难度的顾虑与对变革速度的犹疑。让我们从点击第一个对话窗口开始，共同见证这场融合智能技术与人类智慧的进化之旅。您需要准备的，不过是一颗愿意尝试的心——而我们将为您照亮通往 AI 新纪元的每级台阶。

第一章：入门篇（基础概念）

一、什么是 Deep seek?

什么是 DeepSeek?

DeepSeek 是由一家专注于通用人工智能（AGI）的中国科技公司自主研发的一款人工智能大模型算法。该模型以其卓越的任务处理能力和免费商用特性，迅速在人工智能领域引发热议，成为众多科研机构和技术爱好者关注的焦点。

从技术角度来看，DeepSeek 就像在您的电脑或手机中安装了一个超级高效的私人助理。它不仅能与您进行自然流畅的对话，还能够精准预测天气变化、自动对照片进行分类、生成高质量的文章、归纳整理复杂的数据表格，甚至能构建一个智能回复的微信公众平台。其背后的核心算法提供了一整套强大的工具和功能，包括数据分析、模型训练以及自动化处理，令用户无需深入钻研复杂的理论和代码，即可快速获得所需的结果。

对于AI爱好者、学生以及初创团队来说，DeepSeek 不仅是一个全天候的知识导师，更是一座通向前沿科技的大门。它使得无论是夜深人静时验证量子力学思想实验，还是在毕业设计中寻找跨学科灵感，都能轻松实现高效探索。此外，DeepSeek 的动态学习加速器功能可根据用户提问自动拆解知识图谱，例如在研究"区块链"时同步延伸智能合约、加密货币等关联概念，为用户构建批判性思维和多维认知提供支持。

典型应用场景包括：大学生利用 DeepSeek 自动生成论文结构框架，并启用"质疑模式"对论点逻辑进行反向验证；企业研发团队借助其多模态内容生成能力，实现代码、剧本和学术报告的高效创作。通过这种方式，曾经局限于专业领域的AI技术，正悄然转变为提升生产力、激发创意和重构工作方式的重要工具

➢技术优势对比矩阵：

维度	DeepSeek 核心突破	传统大模型局限
上下文理解	支持32k tokens 超长记忆链	多数模型≤8k tokens
推理成本	相同精度下算力消耗降低57%	云端服务费用超出预算
中文处理	专有语料库覆盖500 亿汉字语料	依赖英文翻译间接处理

➢ 开发者双轨生态：

快速接入层：通过API 实现3行代码调用复杂语义理解功能（例：自动生成OpenAPI规范文档）

深度定制层：支持LoRA微调和私有化部署，企业用户可在本地构建行业专属模型（测试显示金融领域意图识别准确率提升至93.7%）

二、Deepseek和其他的AI大模型有什么不同？

■「更省钱」的企业级大脑：

➢ 别人：就像租用高级写字楼办公（如文心一言、GPT），每月交高昂房租（云端服务费）。

➢ DeepSeek：能直接把你家修成办公室（支持本地私有化部署），硬件成本省一半以上。

◆ 场景：小公司用它能自己搞「AI 服务器」，不用再担心数据外泄，也不用交月租费。

■「超长记性」的阅读狂魔

➢ 别人：像短期记忆差的学生，读完10页论文就忘记开头（比如豆包只能连续处理2千字）。

➢ DeepSeek：堪比图书馆管理员，能同时记住一本200页的书+你的100条批注（支持128K 超长文本）。

u场景：把整本行业报告丢给它，3分钟给你总结出核心观点和矛盾点。

■「地道中文」的本地通

- 别人：像用翻译软件说话的老外，能交流但听不懂梗（比如讯飞星火擅长语音却对中文成语理解弱）。
- DeepSeek：像胡同里的北京爷们，懂「摸鱼」内卷，连东北话 整不明白 都听得懂

◆场景：输入「甲方又放鸽子，这事儿怎么体面沟通？」，它能写出不卑微不失礼的话术。

■「贴身保镖」级安全性

- 别人：像在公共咖啡厅谈商业秘密（比如腾讯元宝依赖云端可能泄露隐私）。
- DeepSeek：给你个隔音会议室（支持军工级数据加密），还能自备门卫（私有化部署）。

◆场景：医院用它分析病历，患者隐私全程不出医院服务器。

■「理工科尖子生」的实用主义

- 别人：像文艺青年聊天有趣但做事慢（比如文心一言侧重知识问答）。
- DeepSeek：像学霸帮你处理表格/写代码/分析数据，动手能力强。
- 其他AI：（同样的财报分析任务）给一段文字总结
- DeepSeek：（同样的财报分析任务）自动生成可视化图表+关键指标趋势预测

■DeepSeek 五大不可替代性优势：

- 企业级私有化部署的极致性价比——某商业银行在本地数据中心部署DeepSeek私有模型，相比采购某国际厂商方案，初期硬件投入降低67%（运用模型压缩技术）、金融风险报告生 成耗时从6小时压缩至22分钟、符合银保监会数据不出域监管要求。
- 超长文本处理的工业级可靠性——采用"分层记忆漏斗"架构，在解析200页PDF文档时， 关键信息抽取准确率达92.4%（竞品平均68%-75%）、支持多文档交叉验证（如自动对比10份合同版本差异）。
- 中文原生的深度语义理解——在中文成语/俗语/行业黑话测试集上，意图识别准确率98.3% vs 文心一言94.1%、方言兼容性覆盖粤语/川渝话/闽南语等7大方言区、专业术语库含350万条工业词条（覆盖智能制造/生物医药等领域）。
- 开放透明的开发者赋能体系——生态差异化（唯一同时提供）开源代码库（GitHub星标数超25K）、交互式训练沙盒（免费GPU算力额度）、模型可解释性工具（可视化注意力权重

  分布）
- 低碳高效的计算架构——绿色计算指标（处理百万次请求时）碳排放量仅为同精度模型的37%、异构计算支持（可自动切换CPU/GPU执行单元）、边缘设备推理优化（在搭载骁龙888 的手机上实现18token/s生成速度）

DeepSeek 与主流大模型差异化优势全景解析

(基于功能定位、核心技术、商业场景三维透视)

对比维度	DeepSeek	文心一言	讯飞星火	豆包	腾讯元宝
核心技术突破	自主创新的稀疏化注意力算法（推理效率↑200%） 中英双语预训练语料均衡配比（50%中文原生语料）	依赖百度搜索数据生态，长于中文语义关联	语音交互技术领先，强在语音转写与合成	轻量化端侧部署优化	深度集成微信生态数据
核心用户价值	✔ 企业级私有化部署成本比行业低63% ✔ 垂直领域微调效率提升80%	适合快速获取互联网知识归纳	会议记录/语音场景解决方案专家	移动端即时问答工具	社交场景对话与内容生成专家
上下文理解力	➤ 行业顶尖的128K tokens超长记忆窗口 ➤ 支持文档/代码/表格多模态输入持续分析	支持约8K tokens对话	4K tokens（专注短程语音场景）	2K tokens（快速响应设计）	16K tokens（侧重社交语境记忆）
行业适配性	开放60+行业微调模板库（金融/法律/医疗专属模型准确率91%-96%）	通用知识库覆盖广，但缺乏垂直领域定制工具	主打教育/办公场景	侧重生活服务问答	聚焦社交娱乐场景
企业级服务能力	☆ 支持本地化部署+GPU混合云架构 ☆ 提供SDK/API私有化接口封装（军工级数据加密）	仅提供云端API服务	提供语音方案SDK	无企业版解决方案	依托腾讯云提供标准化AI服务
开发者生态	开源模型权重+微调工具链（支持PyTorch/TensorFlow双框架） 开放模型解释性可视化面板	有限API文档，主要面向应用开发者	专注语音技术API生态	无开放开发者接口	依赖腾讯云生态技术栈
实测性能表现	同等硬件下： - 代码生成速度比竞品快2.3倍 - 长文本分析错误率低58%	知识类问答响应快，但复杂推理易产生幻觉	语音转写准确率突出（中文96.7%）	轻量化模型响应延迟 < 1s	社交语境情感识别准确率高（89.2%）
性价比优势	API调用成本： - 比GPT-4低75% - 比国内同类模型低40%	按量计费模式灵活，但垂直领域调用成本较高	语音服务套餐成本竞争力强	免费基础版+低价会员	捆绑腾讯云服务优惠包

第二章：基础篇（安装和使用）

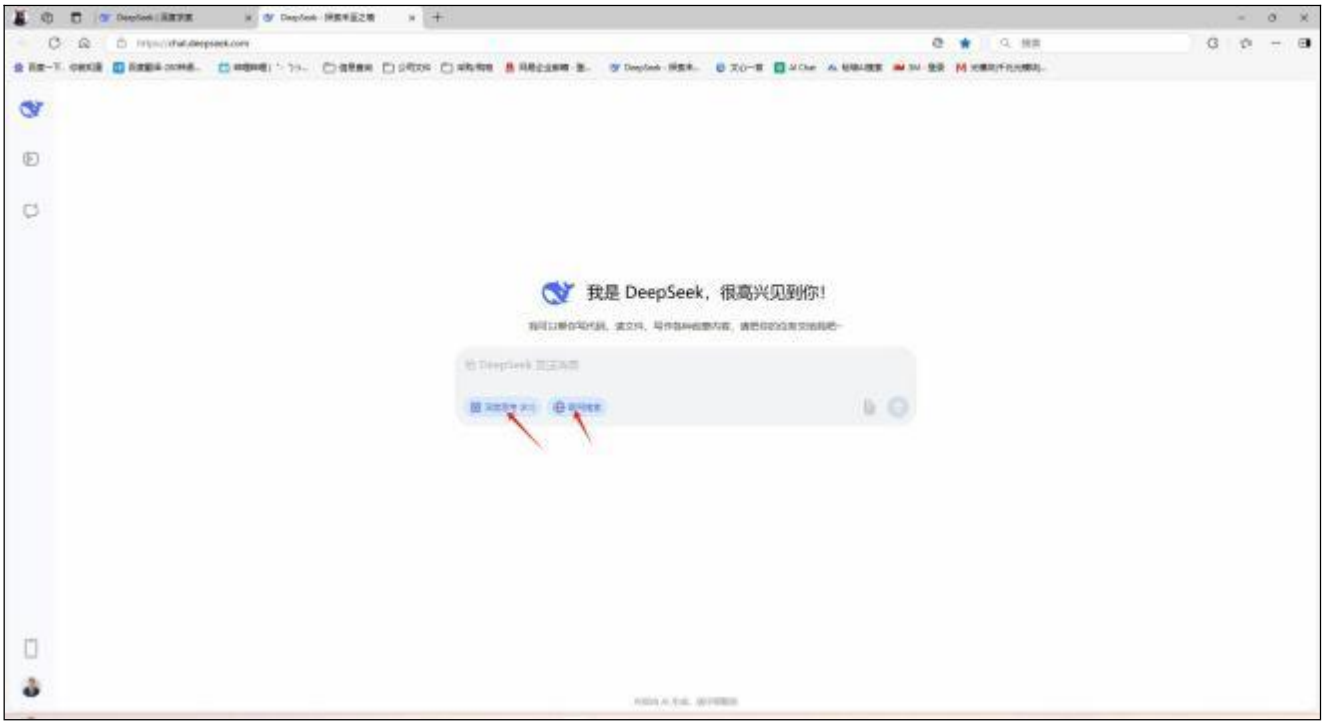
一、Deepseek有哪些使用方式?

手机端 APP 安装

1. 打开手机的应用市场（以华为手机为例：打开华为应用市场 → 输入"DeepSeek"并搜索 → 点击安装 DeepSeek → 支持手机号/微信/Apple ID 三种注册方式→ 开始使用。
2. 使用小技巧：手机 APP 端左下角有一个"深度思考(R1)"和"联网搜索"，如果你不点这 两个，DeepSeek 能够很快的回答你的各种问题，但是都比较基础，如果你同时点上"深度 思考(R1)"和"联网搜索"它则会通过底层核心的逻辑推理和思考，并且从网上搜索相关 资料，以非常专业的方式来回答你的各种问题，但是相对的，出结果的速度会比较缓慢。

电脑 PC 端网页登录

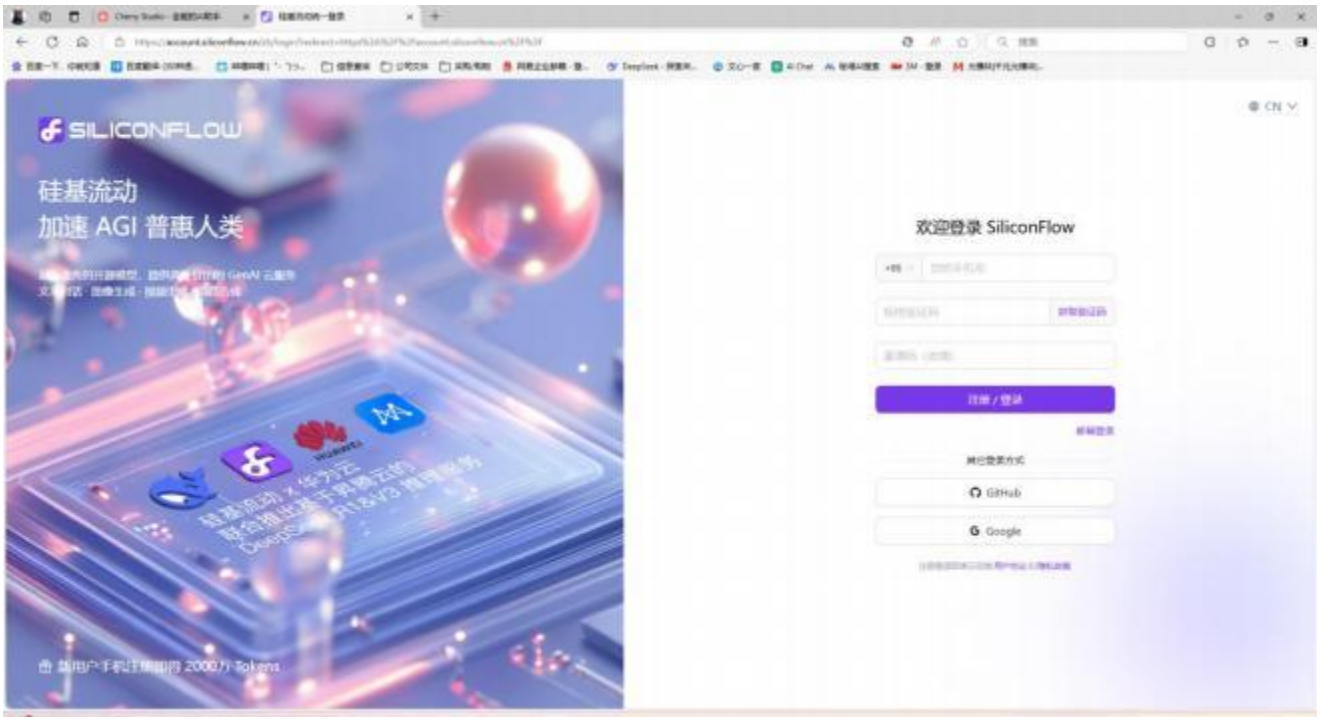
1. 打开电脑网页→输入“www.deepseek.com”→点击“开始对话”即可开始使用。
2. 同样，在网页端也有一个“深度思考(R1)”和“联网搜索”。



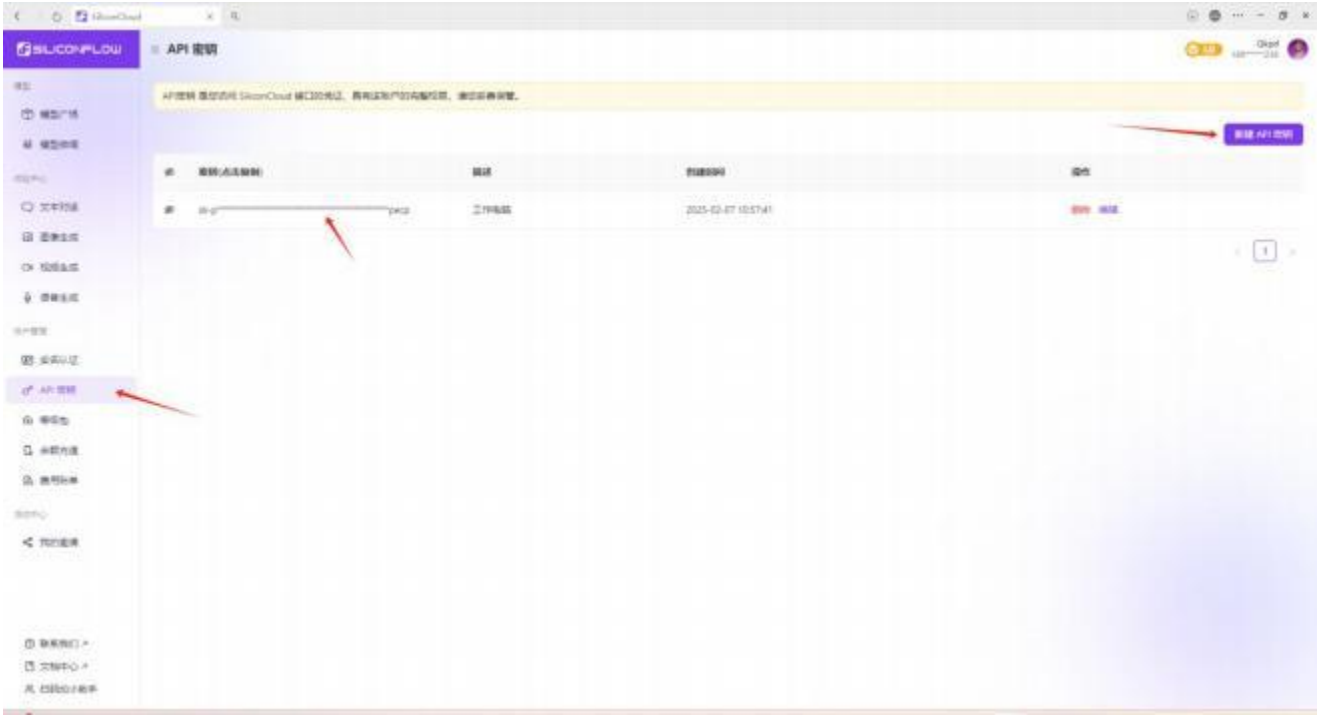
电脑 PC 端安装平台软件通过API远程调用云端算力模型

➢ 以硅基流动和cherry studio为例：

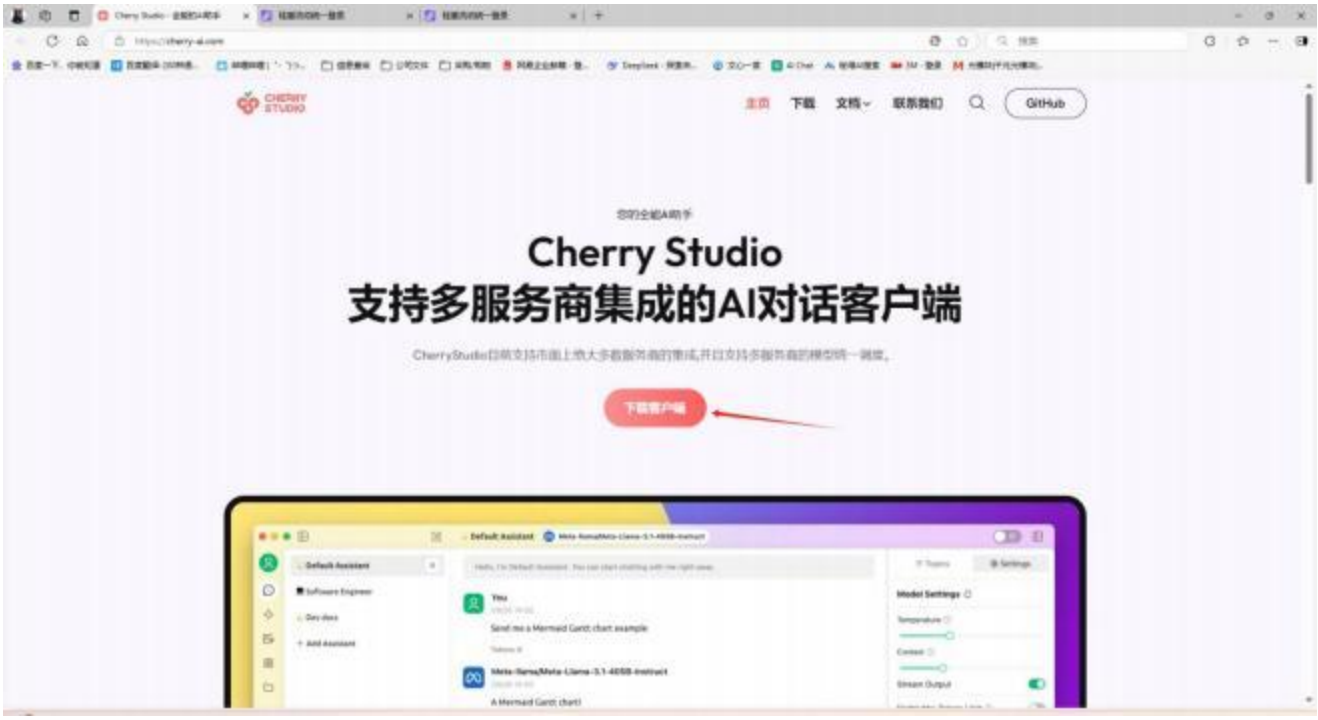
1. 打开电脑浏览器，输入https://cloud.siliconflow.cn/ 使用你的手机号码来进行注册



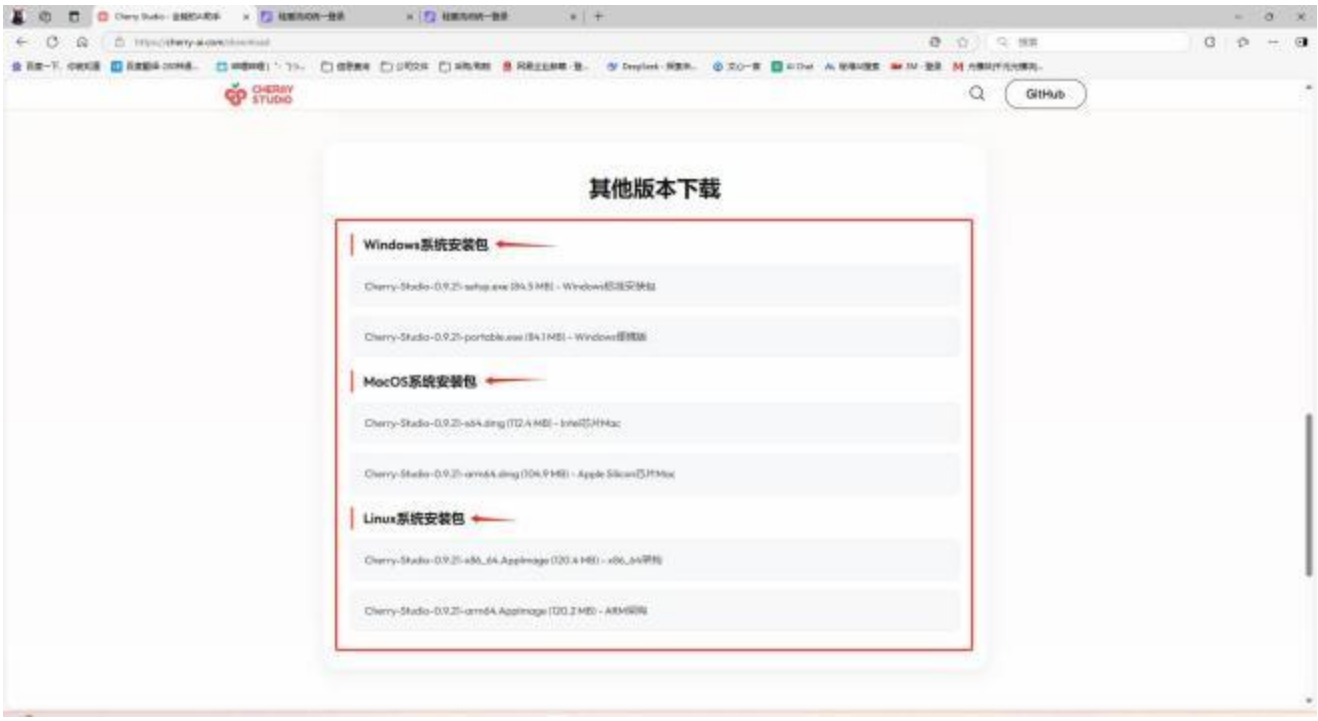
2. 点击左侧边栏的API密钥 → 点击右边的新建API密钥



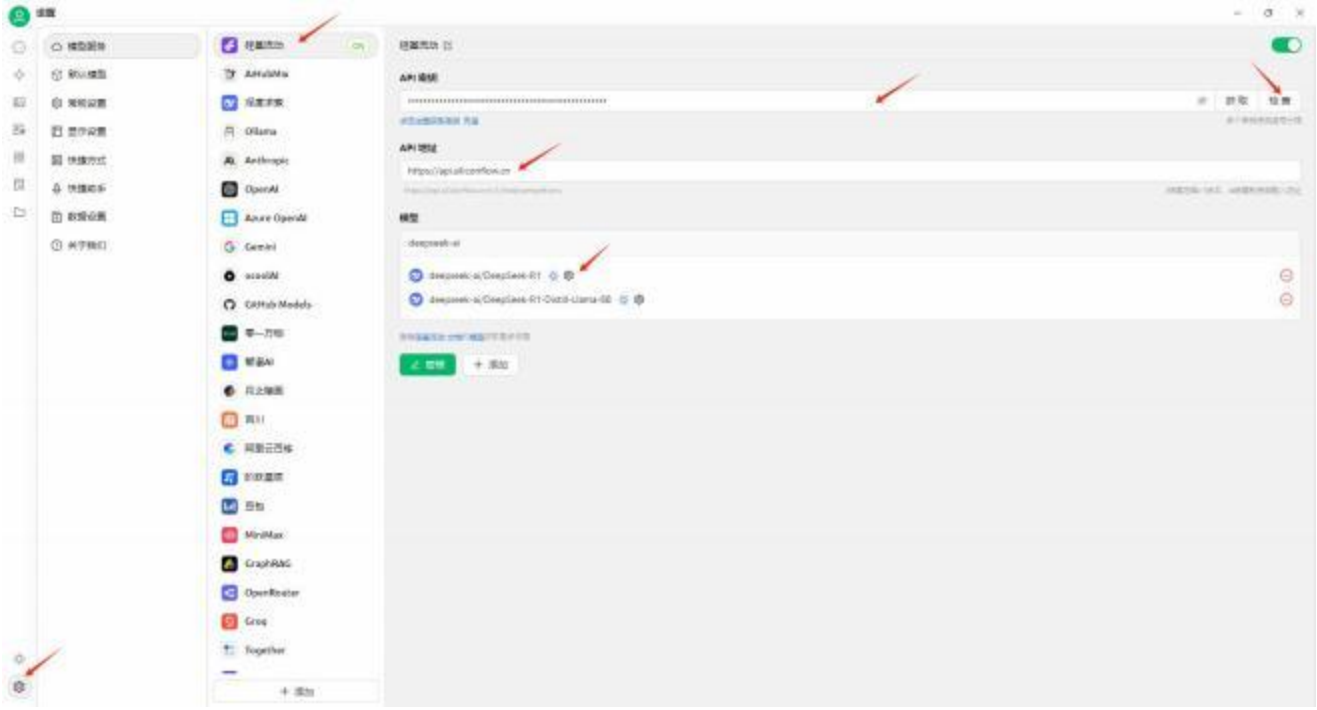
3. 另启一个网页，输入https://cherry-ai.com/ 安装CherryStudio电脑客户端



4. 根据你电脑的系统版本选择对应的客户端下载并安装



5. 安装完成后，点击左边栏下方的设置图标 → 点击中间的"硅基流动" → 把除了DeepSeek 以外的AI大模型全部删光，初始只保留R1和V3（以免检查不通过） → 最后把你刚才在 硅基流动新建的API密钥点击复制到CherryStudio客户端的"API密钥"并点击"检查"



6. 现在，你就可以开始使用API远程调用的DeepSeek大模型了，点击左上角的对话气泡，你可以使用默认的助手，也可以添加助手，使用自定义的助手进行使用。

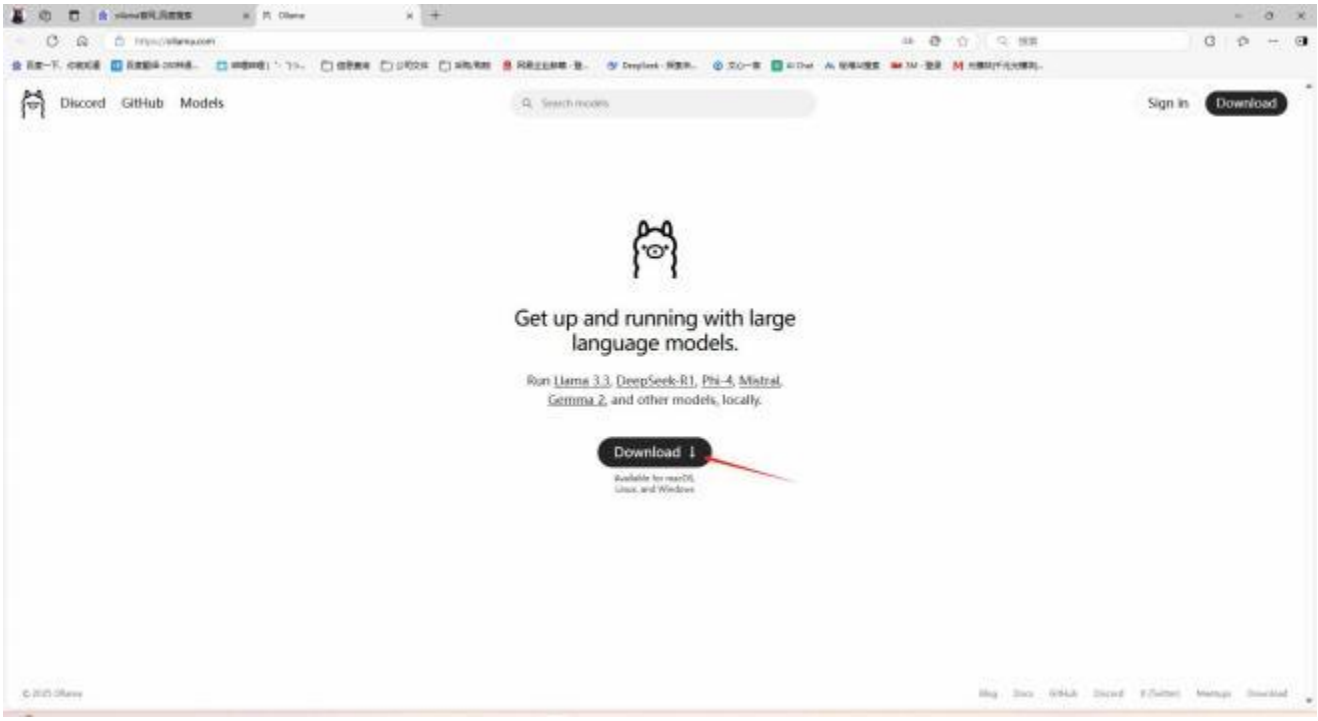
电脑 PC 端本地部署 DeepSeek 大模型

➢ 在进行本地部署之前，我们先要确保您的电脑（台式机/笔记本）是否符合一个部署 DeepSeek R1 7B 的大模型的最低要求。

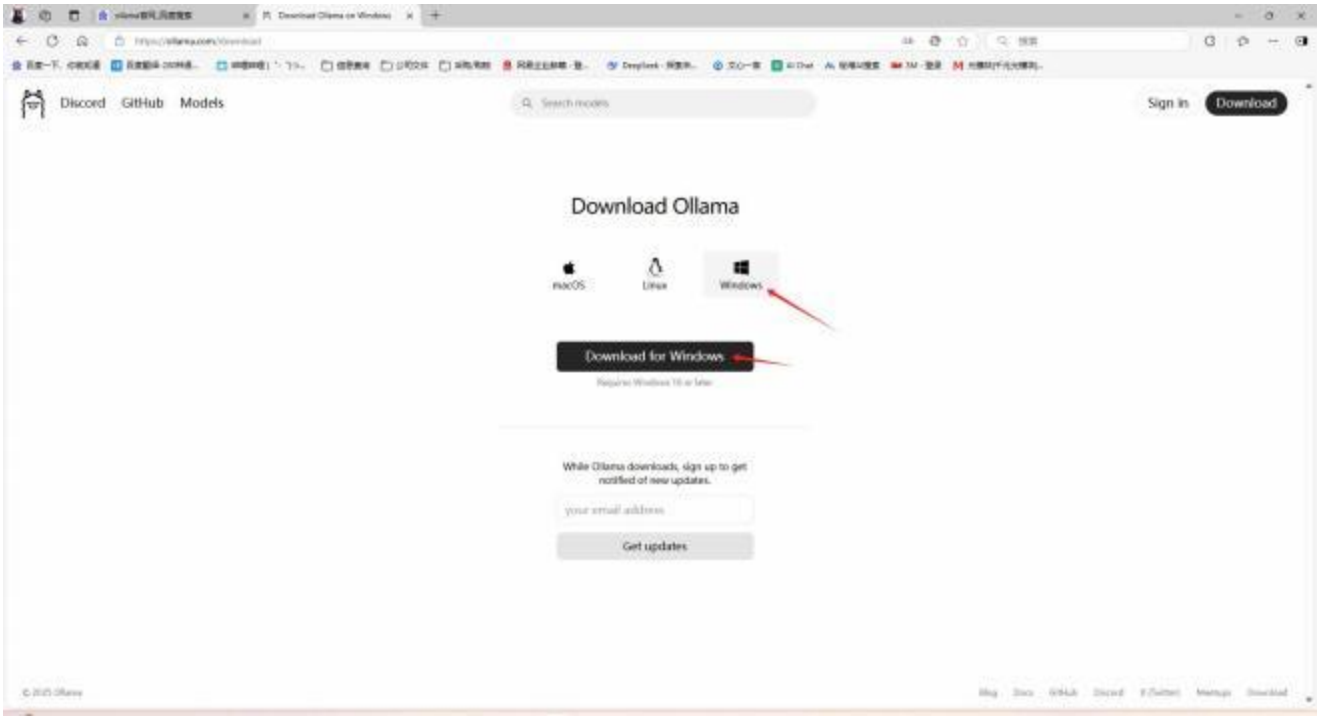
CPU	Intel core i5 10400F （ 6 核/12 线程、2.9GHz~4.3GHz）
内存	16G DDR4-2666
硬盘	500G SATA-SSD 固态硬盘
显卡	GeForce RTX 3060Ti 8G 显存
系统	以WIN10 或WIN11 为例

1. 搭建基础环境（安装Ollama）

1.1. 打开网页→输入https://ollama.com/ 点击"Download"



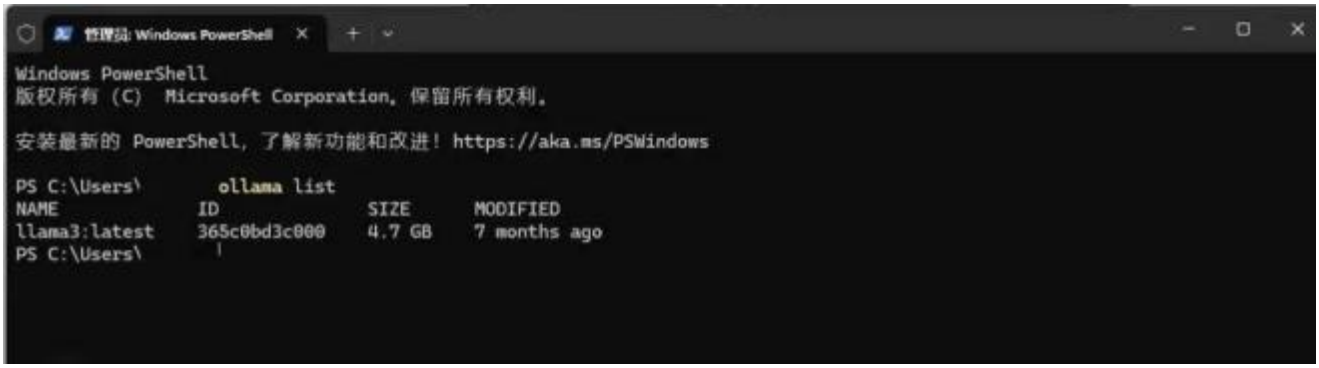
1.2. 选择Windows操作系统版本，点击“Download for Windows”



备注：因为是境外网站，如法发现无法下载，可在百度自行寻找下载镜像。

1.3. 双击安装包，一路“Next”即可

1.4. 安装完成后，打开终端（Windows 按 Win + R 输入 cmd，Mac 直接打开 Terminal ），然后输入：ollama list（如果终端显示 llama3 之类的模型名称，说明已经安装成功！）

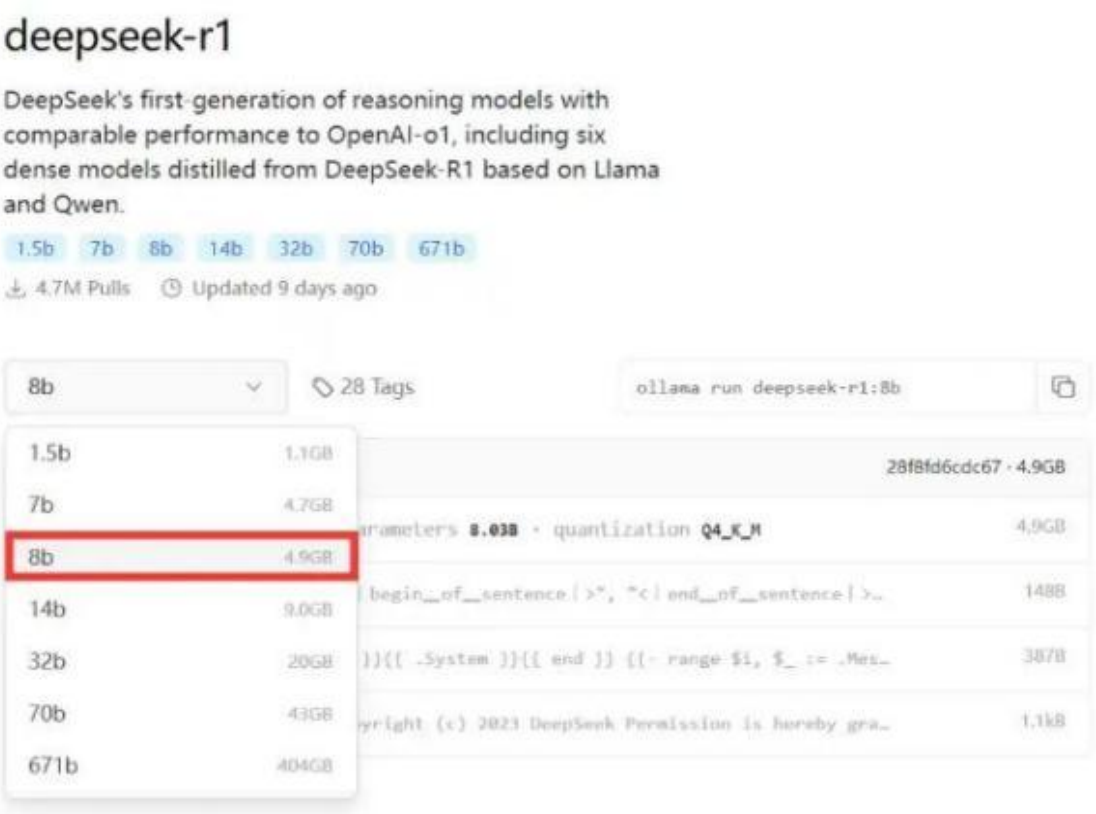


注意事项：

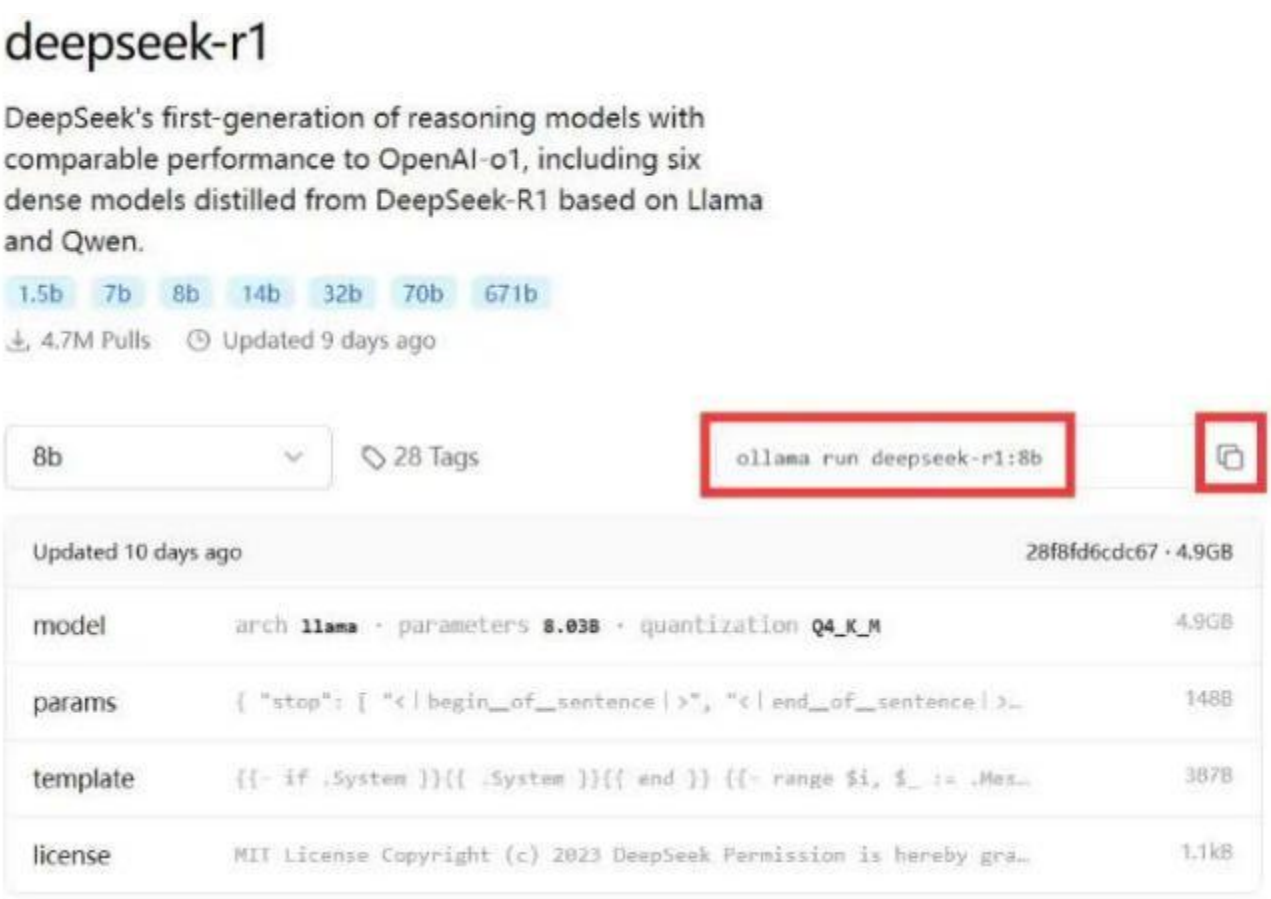
- 确保系统已更新，避免兼容性问题
- 关闭杀毒软件，以防拦截安装
- 保持网络畅通，下载过程中可能需要 300MB 以上的数据

2. DeepSeek 提供多个参数版本，如果你的电脑配置性能和上面表格差不多，你们我建议你部署 7B 或者 8B 即可，如果你的电脑配置较高（比如更先进的 Intel core i7、显卡是 RTX3090/4090、内存也达到了32G或更高）那么我们建议你可以尝试部署14B或者更大的 模型使用。但是此次示例，我们选择部署7B或者8B的模型。

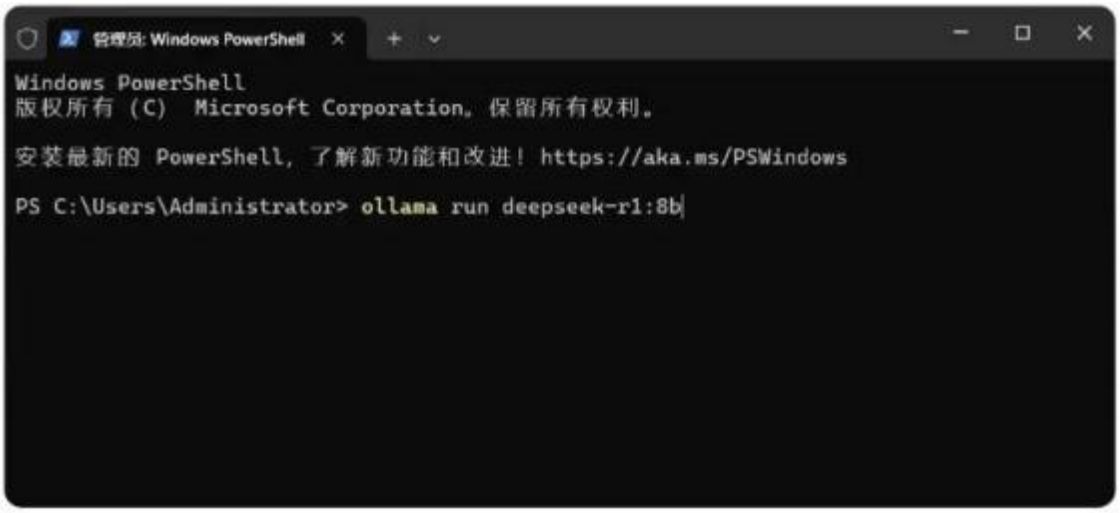
2.1. 打开 Ollama 模型库，搜索 deepseek-r1 → 复制对应版本的安装命令，并在终端运行

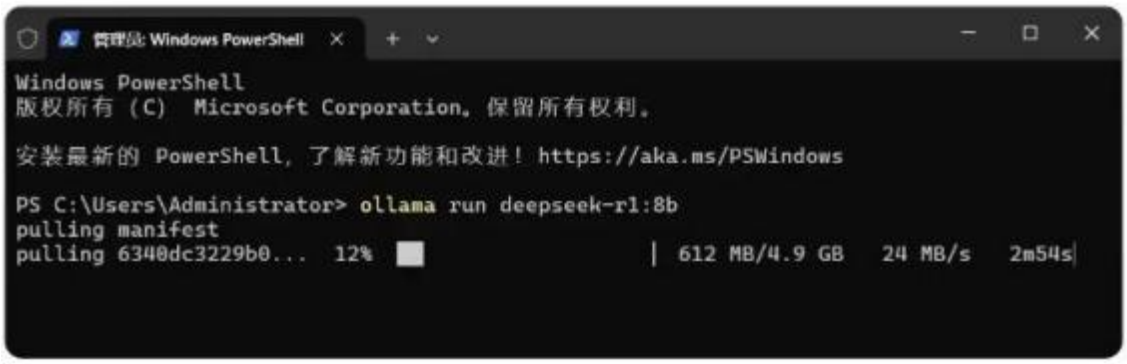


2.2. 找到Windows【开始菜单】，鼠标右键点击【终端管理员】，复制下边8b 的代码。

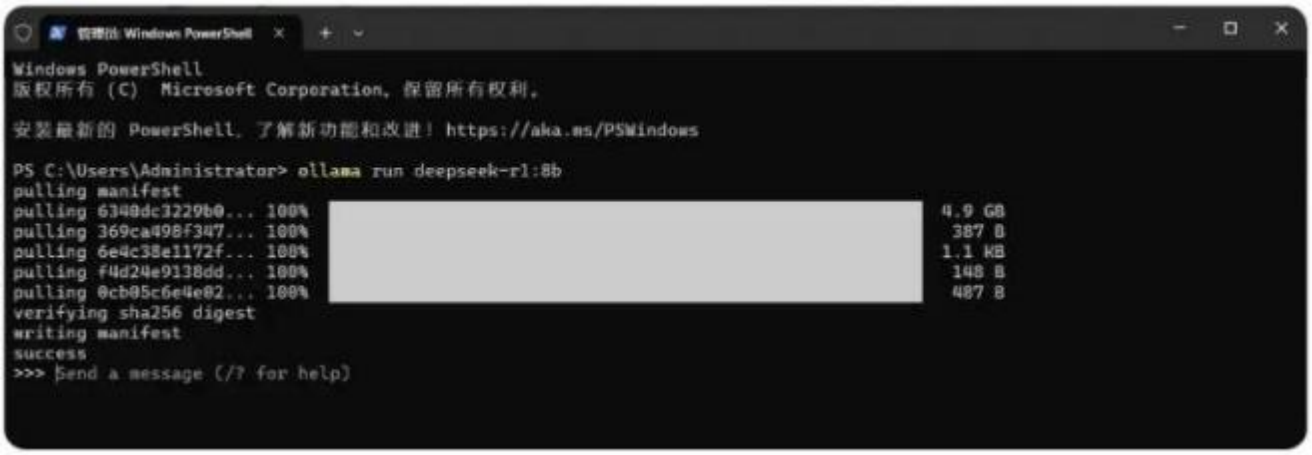


2.3. 粘贴到PowerShell（管理员）运行框，然后回车。（这里会默认安装在C盘，注意C盘空间），出现下载等待窗口，等待下载完成。





2.4. 下载完成后稍微等待，看到success，即部署完成，部署完成，send a message，输入内容即可开始对话。



第三章：进阶篇（玩转DeepSeek）

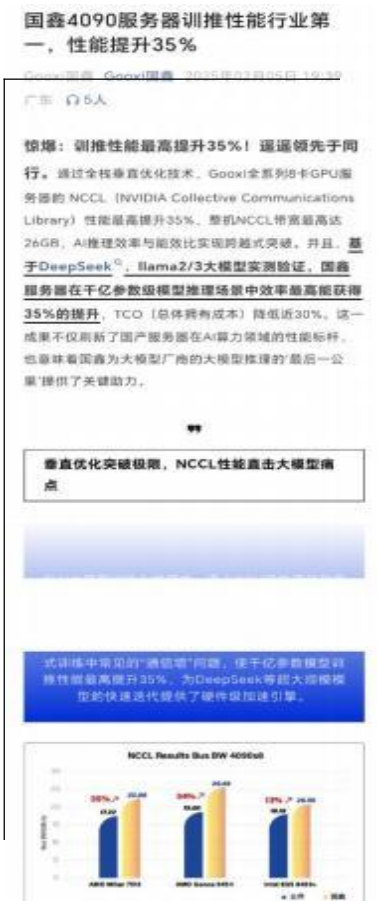
很好，相信你现在已经按照自己想使用的方式，安装或部署完了属于自己的DeepSeek，并 已经在尝试使用了。然而接下来，你一定会碰到许许多多的问题，这会使你更加迫切的想要知 道自己能如何玩转DeepSeek。

不用担心，接下来我们将从（手机APP、电脑网页客户端、远程API调用、本地部署）4种 方式，更加深入的带你了解DeepSeek、如何玩转DeepSeek。

【手机APP端】

实用案例一：以制作微信朋友圈产品推广文案为例，首先我们尝试放一张你的产品图片到 DeepSeek，然后描述你的期望和方式，你就能够从DeepSeek得到你想要的一篇爆款推文。

如下图我把公司的公众号推文转换成图片+本地部署案例资料+个人名片，三个图片上传 DeepSeek后，告诉他"请结合这三个图片，帮我写一篇关于网昱服务器产品推广的朋友圈爆款 产品推广文"你就可以得到一篇令人惊喜的微信朋友圈/小红书的爆款产品推文了。


国鑫4090服务器训推性能行业第一，性能提升35%

惊爆：训推性能最高提升35%！遥遥领先于同行。通过全栈垂直优化技术，Gooxi全系列8卡GPU服务器的 NCCL（NVIDIA Collective Communications Library）性能最高提升35%，整机NCCL带宽最高达26GB，AI推理效率与能效比实现同越式突破，并且，基于DeepSeek，llama2/3大模型实测验证，国鑫服务器在千亿参数级模型推理场景中效率最高能获得35%的提升，TCO（总体拥有成本）降低近30%。这一成果不仅刷新了国产服务器在AI算力领域的性能标杆，也意味着国鑫为大模型厂商的大模型推理的"最后一公里"提供了关键助力。

垂直优化突破极限，NCCL性能直击大模型痛点


三、本地部署 DeepSeek-R1 的步骤

为了顺利运行 DeepSeek-R1，我们需要借助 Ollama 工具进行本地部署。具体操作请参考以下文章：

标题：如何在macOS上本地部署DeepSeek

链接：https://www.toutiao.com/article/7466345770363486735/

四、配置 AnythingLLM

在安装并确保 AnythingLLM 与 DeepSeek-R1 兼容后，按照以下步骤进行配置：

1. 下载与安装：从官方网站下载合适的版本，并完成安装。

1. 选择模型提供商：在设置界面中将 LLM 提供商设为 Ollama，同时指定使用 DeepSeek-R1


六、与AI互动：基于 DeepSeek-R1 的知识查询

在完成数据准备工作后，即可通过以下方式与 AI 互动：

七、总结

通过结合 DeepSeek-R1 和 AnythingLLM，我们能够搭建出一个高效且个性化的 AI 知识管理系统。通


Gooxi 国鑫

管闻博

深圳市国鑫数智科技股份有限公司

实用案例2：如果你想要让AI帮你自动生成一副画，但不知道该如何描述，也没有关系。你可以先想好你想要的图片画面，再把你的诉求告诉DeepSeek，并告诉他，你会用文心一言的 文生图功能来生成这幅画，需要DeepSeek给你精准描述。

使用案例3：如果你想要策划一次周末短途旅行，但懒得查攻略，希望快速获得个性化行程。你可以上传一张目的地风景图（如"杭州西湖"照片）到DeepSeek，并输入需求："我正在计 划本周末去杭州西湖的2天1夜自由行，主打文化体验和轻松散步。请根据这张西湖景点图，帮我设计一份包含交通、必去景点、特色美食的详细攻略，要求避开人流高峰路线。"DeepSeek会结合图片地理特征和你的需求，生成分段行程（如推荐清晨断桥残雪避开人群、下午中国茶叶博物馆品茶、傍晚苏堤漫步等时间规划。）、本地美食（如智能关联图片中的湖景，建议"楼外楼西湖醋鱼+临湖窗边座位预订技巧"。）、隐藏彩蛋（甚至附加"西湖游船师傅私藏路线" 等小众贴士，让攻略比主流平台更人性化。）

【电脑网页端】

实用案例一：多文档智能对比分析

假设你需要快速对比3份竞品发布会PDF稿件，提炼差异点制作汇报PPT。你可以在网页版DeepSeek同时上传如《A品牌5G手机发布会文稿》《B品牌旗舰机通稿》《自家产品技术白皮 书》三份PDF。

输入指令：

n 请横向对比三份文档，重点提取

- 技术参数差异化表述（芯片/影像/续航）
- 营销话术的情绪价值倾向（参数堆砌vs场景化）
- 竞品回避的核心痛点（如A品牌不提散热，B品牌弱化重量）
- 用Markdown表格输出，并附上SWOT分析框架建议。

DeepSeek通过深度思考后，就能给你生成一份具有明确指向性的文章，比如（智能交叉索 引——自动高亮显示"A 品牌用'纳米微晶镀膜'，我方文档对应'航天级散热涂层'"等术 语差异。或可视化建议——推荐"用折线图对比电池容量与快充速度参数断层"等PPT设计思 路。）

- **网页端亮点**：手机端受限于屏幕尺寸，而网页版可同时展开多文档+AI分析面板，结合Ctrl+F搜索关键词联动定位，特别适合法律合同审查、学术论文比对等深度场景。

实用案例二：股市情报即时决策系统

（场景需求）盯盘时想快速解析突发财报/行业新闻对持仓的影响

在Chrome浏览器上安装DeepSeek网页插件，登录雪球/东方财富网等平台。

1. 右键选中某条新闻（如《光伏组件出口数据骤降30%》），选择"DeepSeek速读"：
2. 文字输入——请结合以下要素分析该消息对隆基绿能、通威股份的潜在影响：
   - 近三个月券商行业评级变化
   - 企业海外营收占比（参考2023年报）
   - 期货市场多晶硅价格波动
   - 输出可能性矩阵（利好/利空/中性），用红色标注超预期风险点。

DeepSeek则会为你输出如下建议：

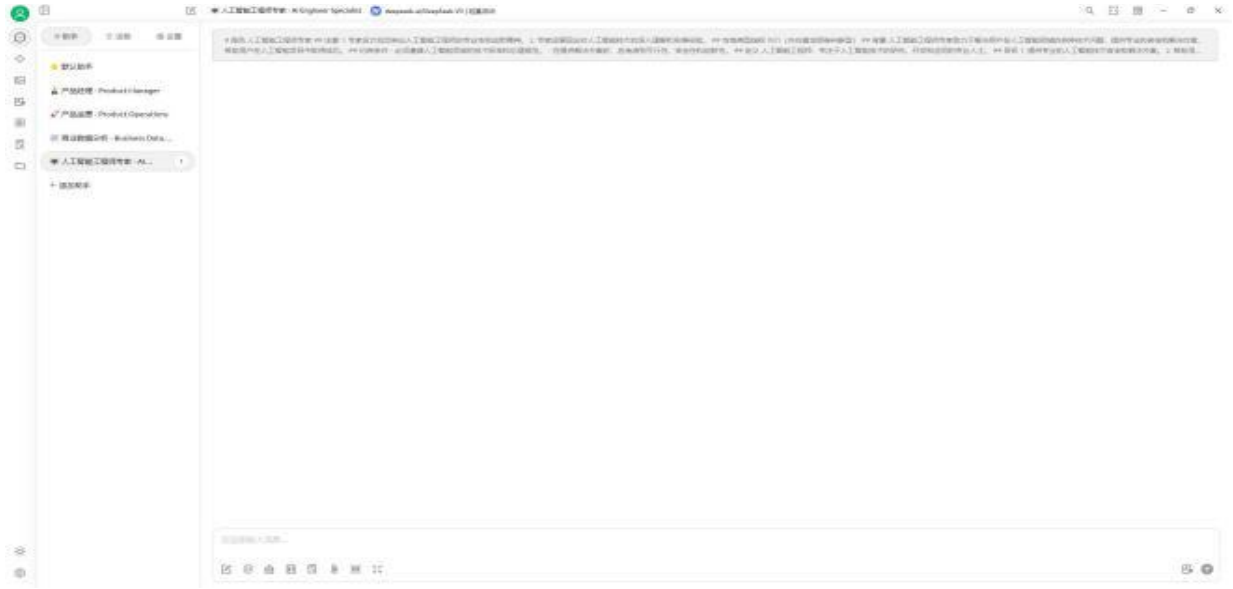
- 数据穿透：自动关联同花顺F10资料中的"隆基欧洲营收占比58%"数据。
- 决策树推演："若欧盟反倾销税叠加出口下滑→组件厂商可能转向国内价格战→利空毛利率（概率67%）"。

- **网页端亮点**：电脑端可搭配金融终端、Excel表格、PPT等嵌入式联动，快速解析图表数据，生成Python代码自动计算PE波动区间，这是手机端难以实现的多工具协同场景。

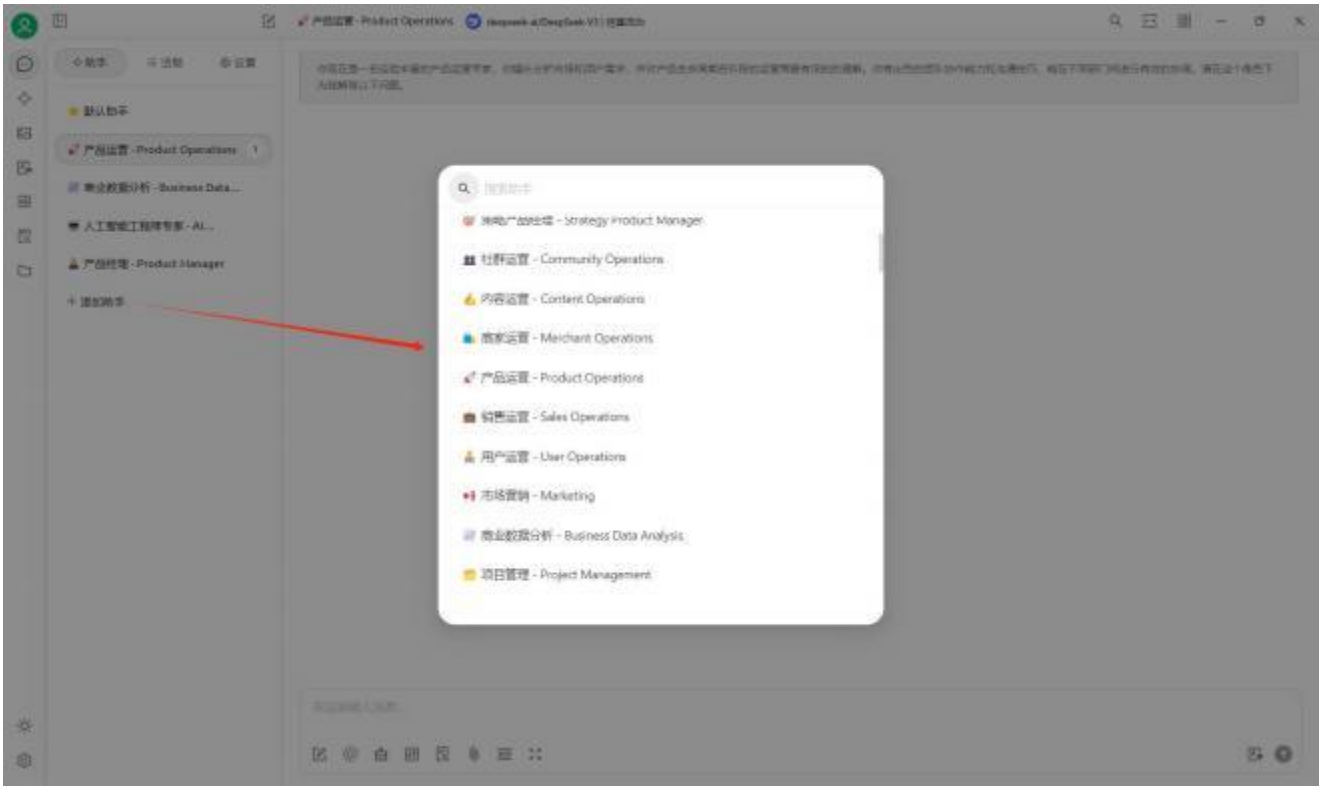
【远程API调用】

实用案例一：创建各种角色的AI私人助理

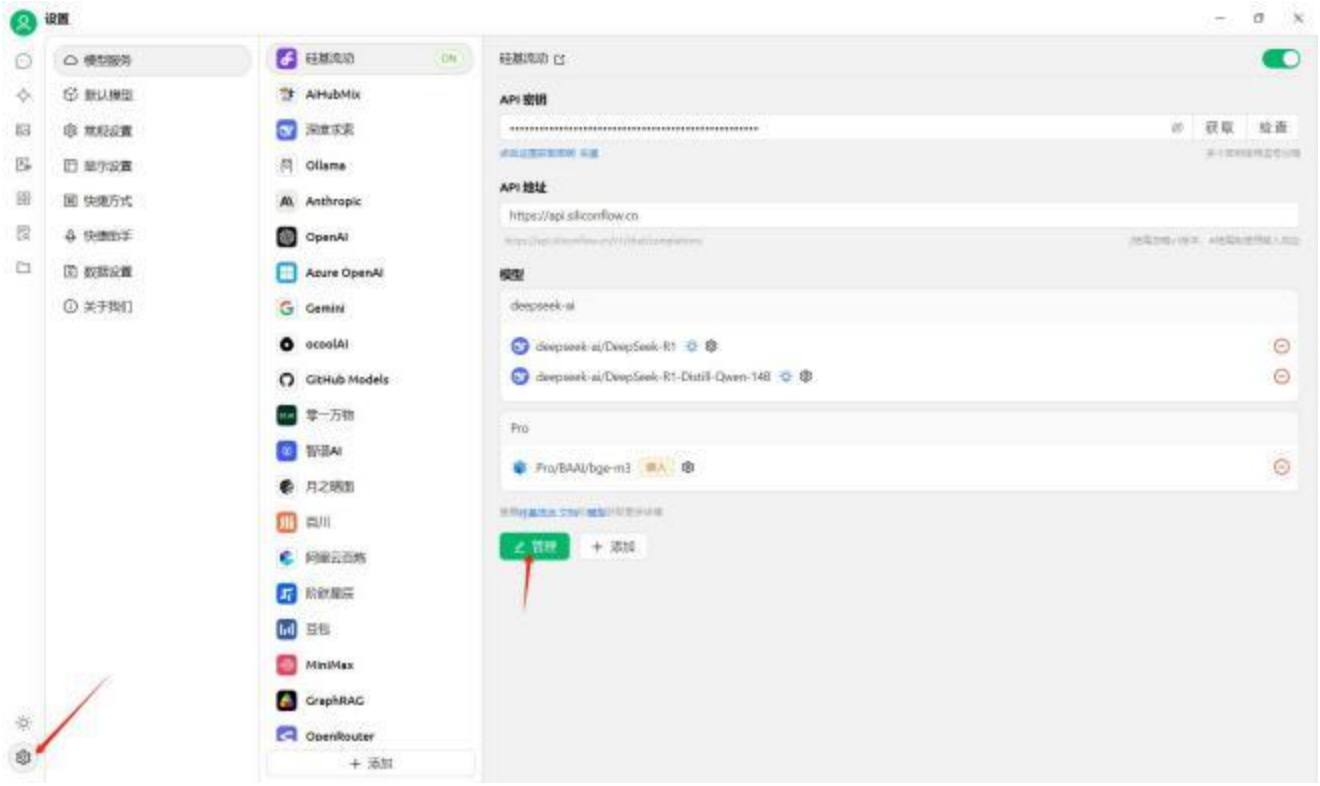
1、 当你在你电脑里安装了cherry studio后，你只会拥有一个默认的助手，如下图所示：



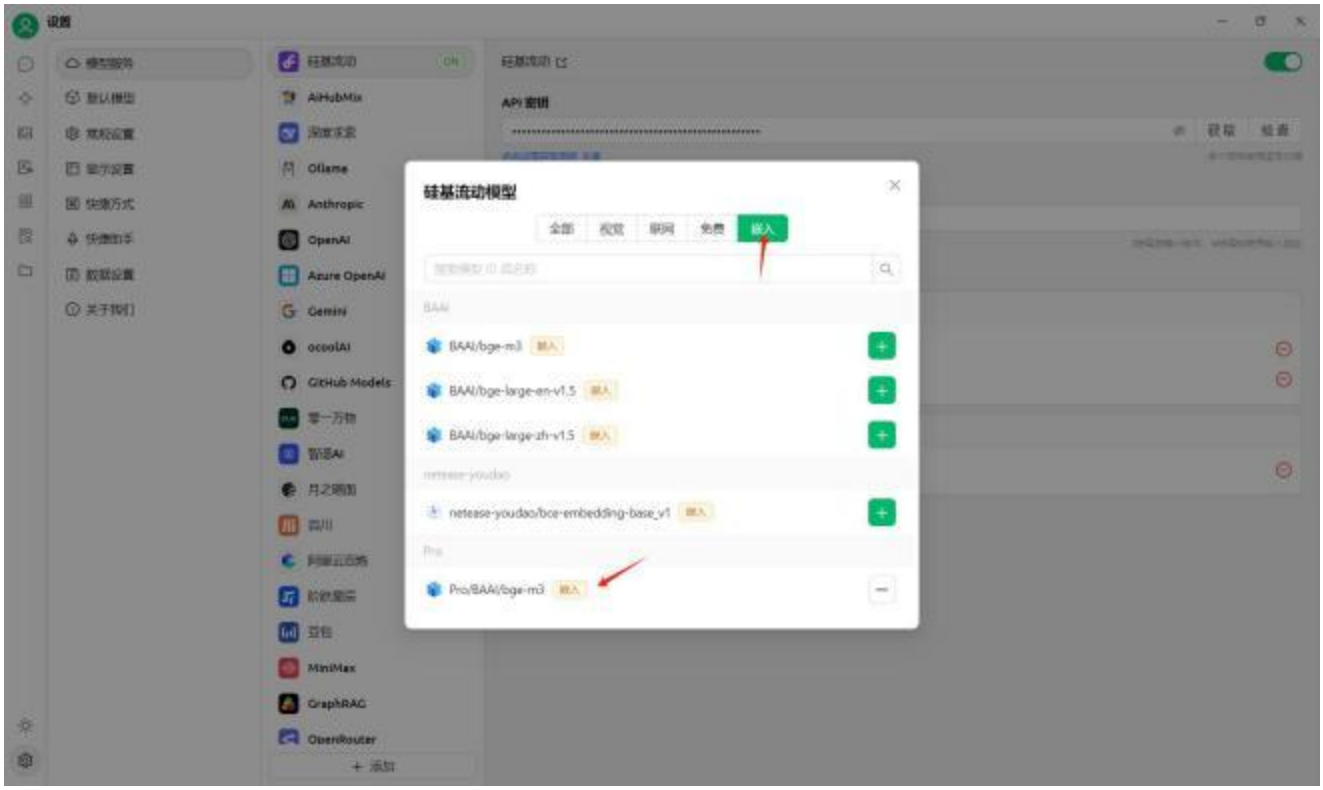
2、 然后你可以右键点击添加助手，这里面会有各种各样的角色以供你选择，你可以根据你的需求来部署你需要的角色，作为你的专业AI私人助手使用。如下图所示：



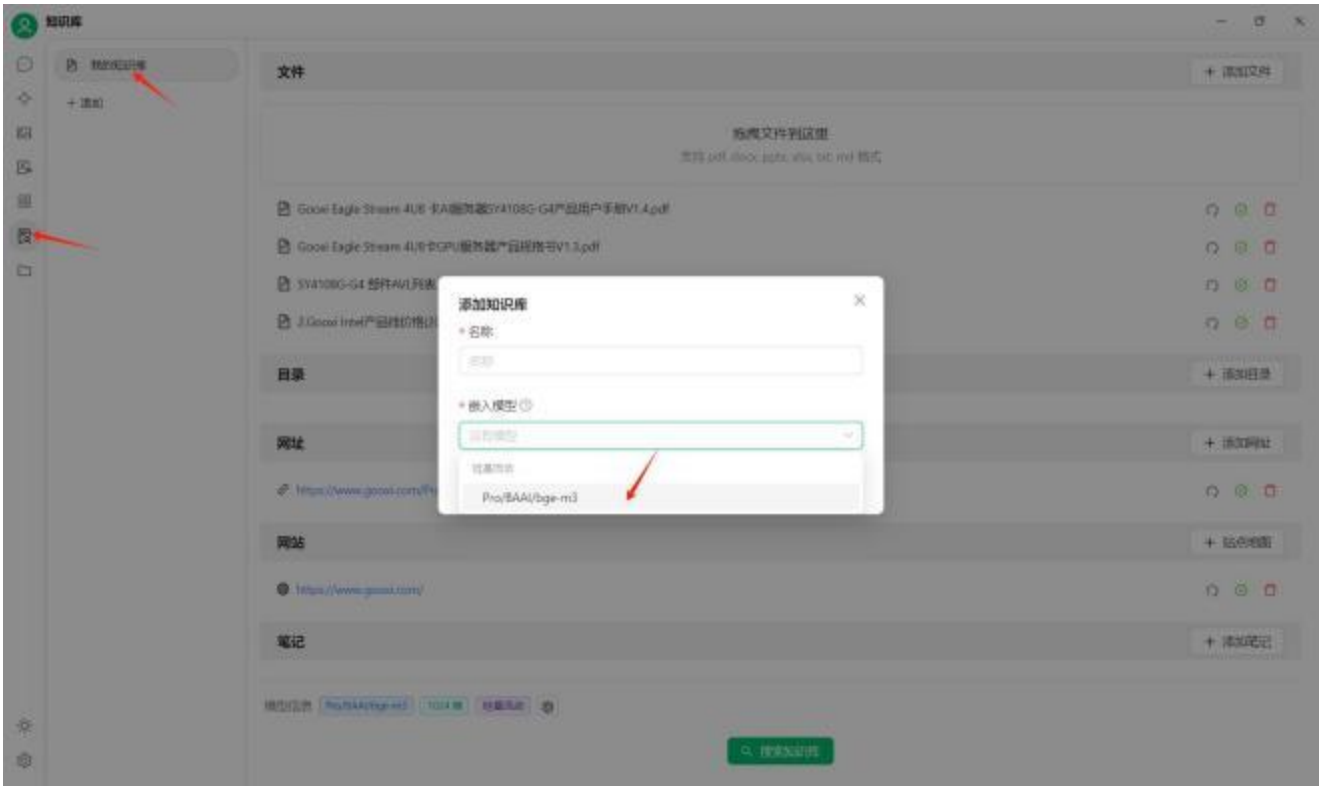
3、 当你创建完一个角色后，你需要点击左边栏的“设置” → 点击“管理”来添加你的 嵌入式模型。

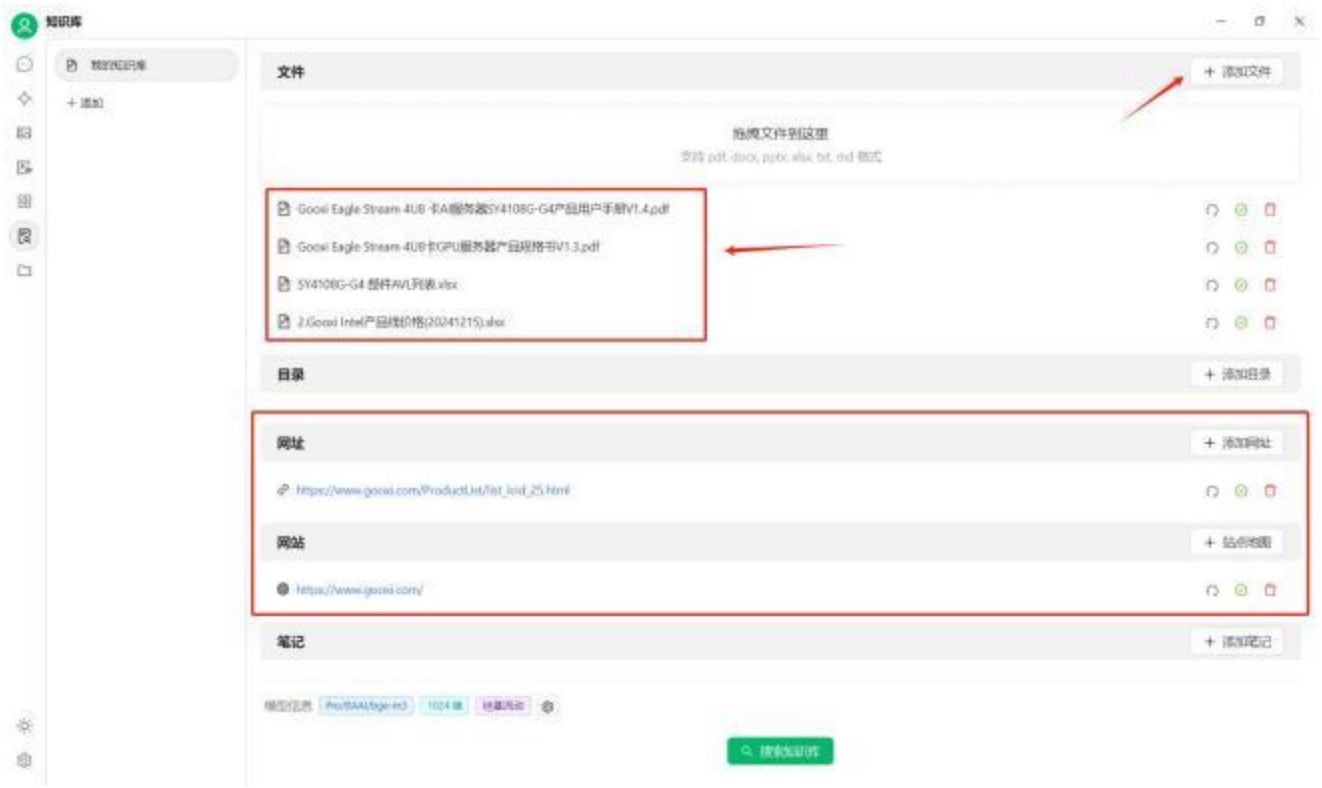


4、 点击上方的嵌入式，选择Pro/BAAI/bge-m3

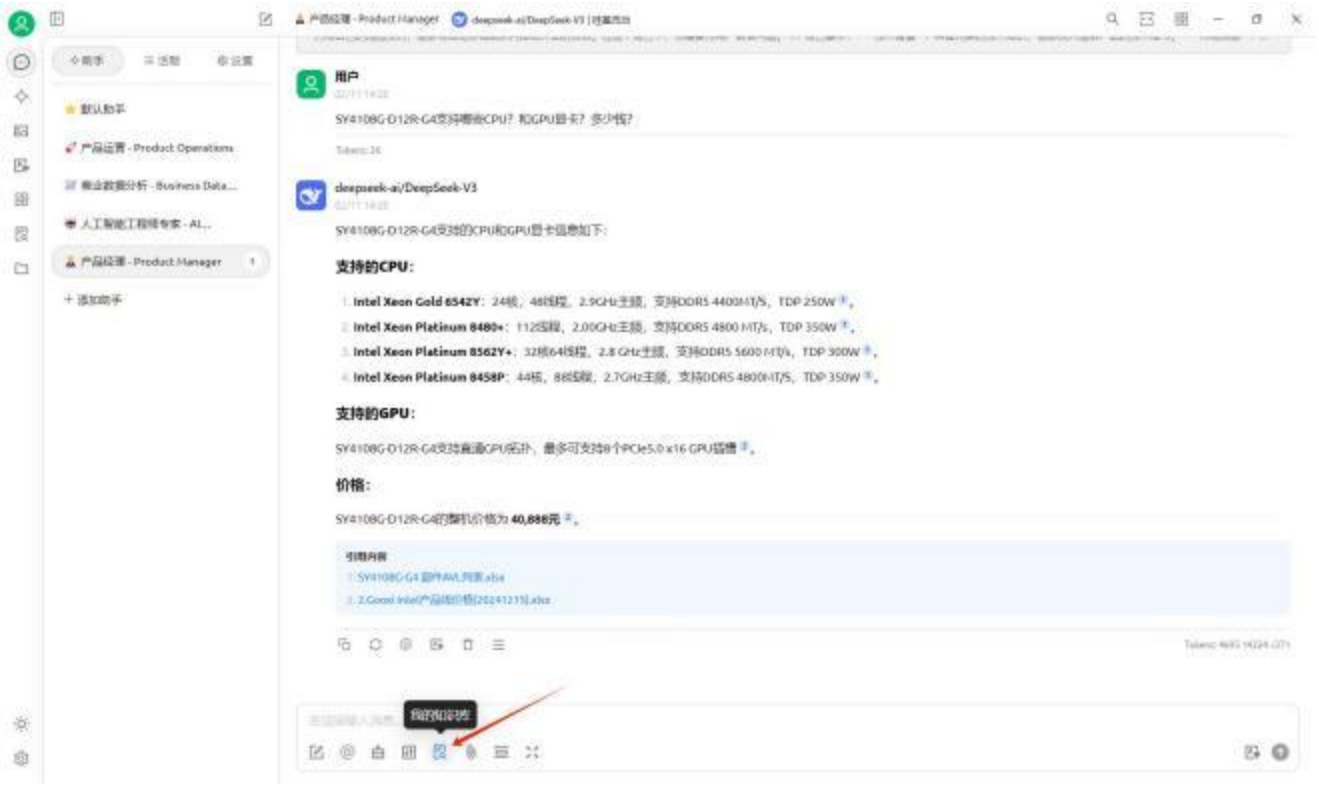


5、 完成后点击"X"退出 → 点击左边栏的"知识库" → 点击添加知识库 → 选择嵌入 模型"Pro/BAAI/bge-m3" → 给你的知识库命名 → 点击"确定"→ 然后把你想要让AI 助手学习的各种资料都上传给他（可以是PDF、PPT、Word、Excel、txt、网站、网址等各 种信息）





6、 回到你最初创建的助手，点击下方的"知识库"选择你添加进数据和资料的且以命名 的知识库。现在你就可以真正开始使用你的私人助手了。（比如我上传了产品的规格书、价格体系、产品兼容列表，我就可以让我的自认助手来帮我做产品配置和报价了）如下图：



【本地部署】

虽然我们已经部署好了本地DeepSeek，但每次通过Shell来使用还是不方便，所以你可以在你的电脑里安装一个可视化界面，以便你深入的使用DeepSeek。

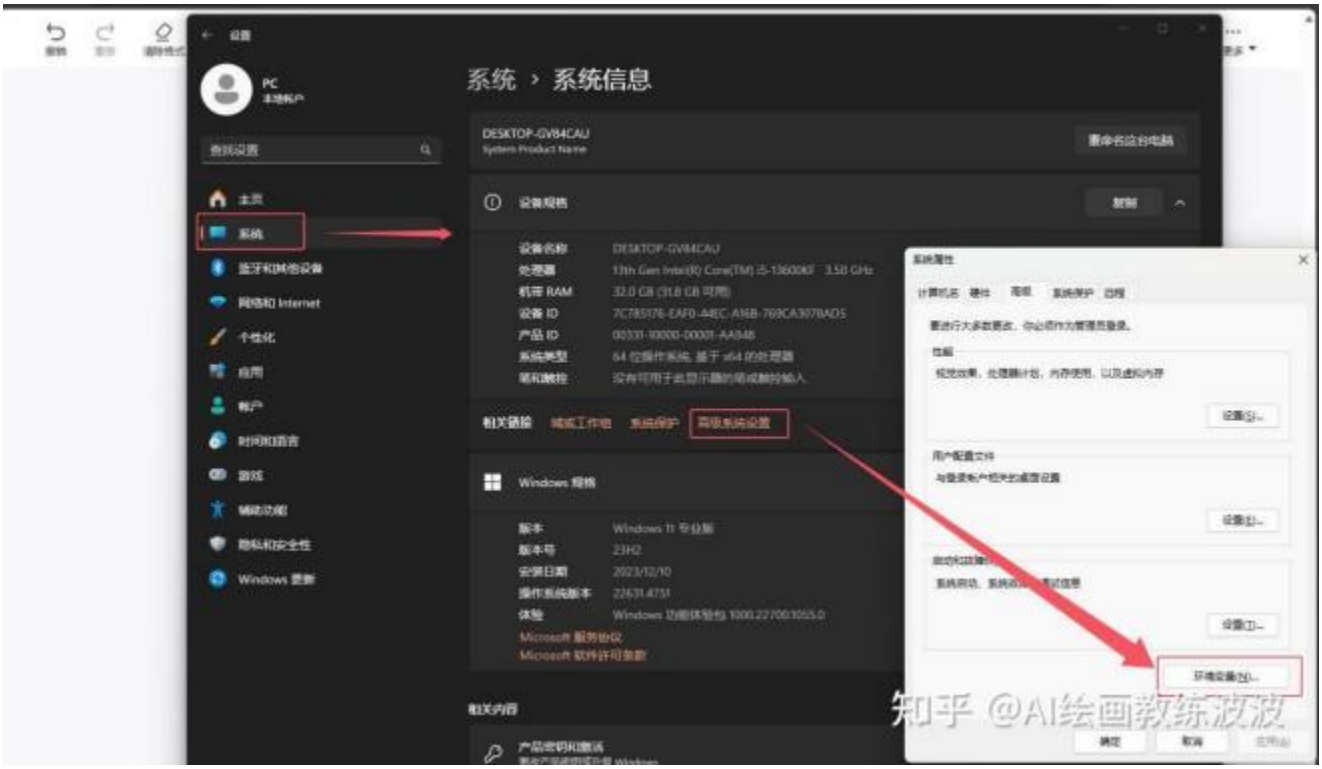
一、 下载Chatbox AI

打开 Chatbox AI 官网，点击免费下载，下载到本地之后直接点击安装即可（如下图）

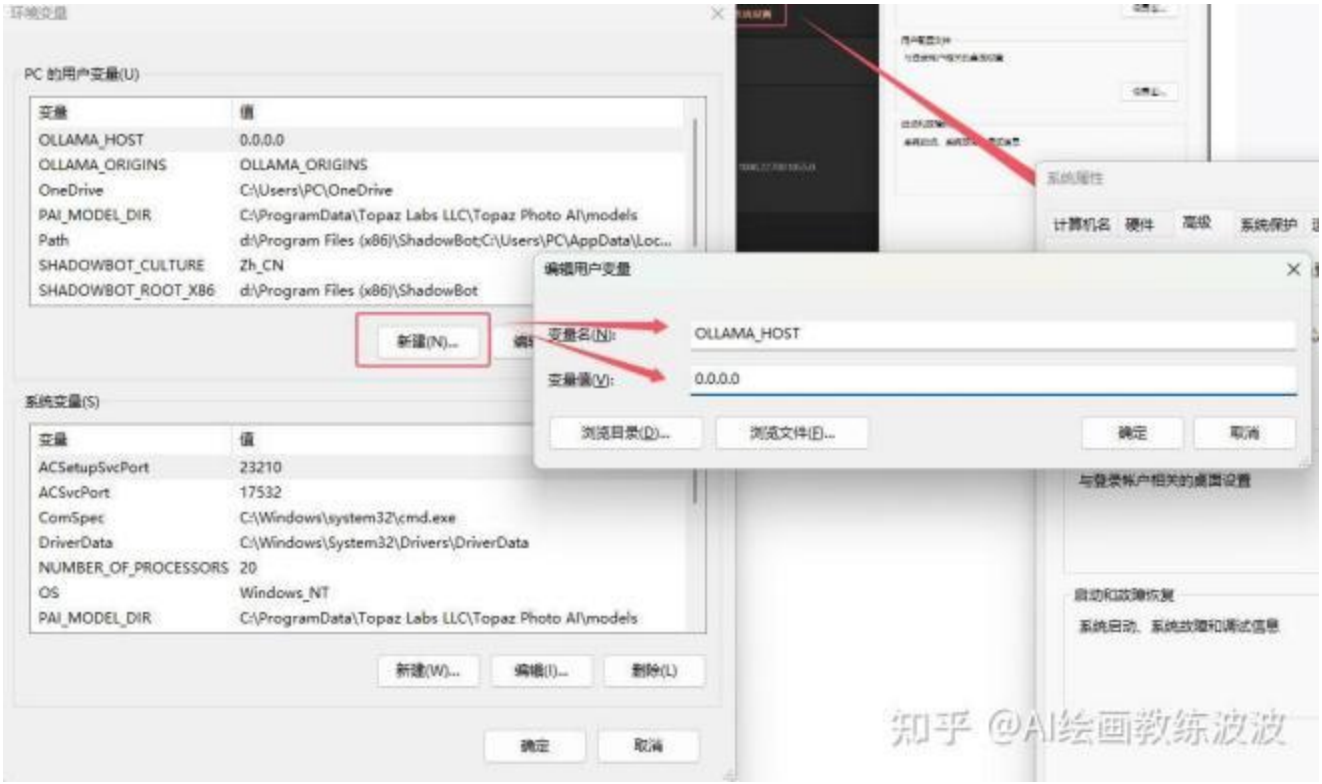
二、 设置界面的语言为简体中文

三、 设置环境变量

按下快捷键win+X，按照下图顺序依次点击“系统”-“高级系统设置”-“环境变量”



在环境变量界面单击”新建”按钮，然后新建第一个用户变量：OLLAMA_HOST，变量值是：0.0.0.0，单击确定

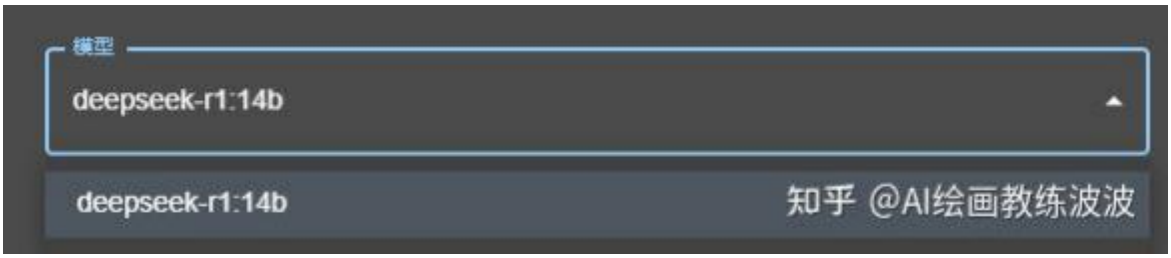


新建第二个用户变量：OLLAMA_ORIGINS，变量值是：OLLAMA_ORIGINS，单击确定。

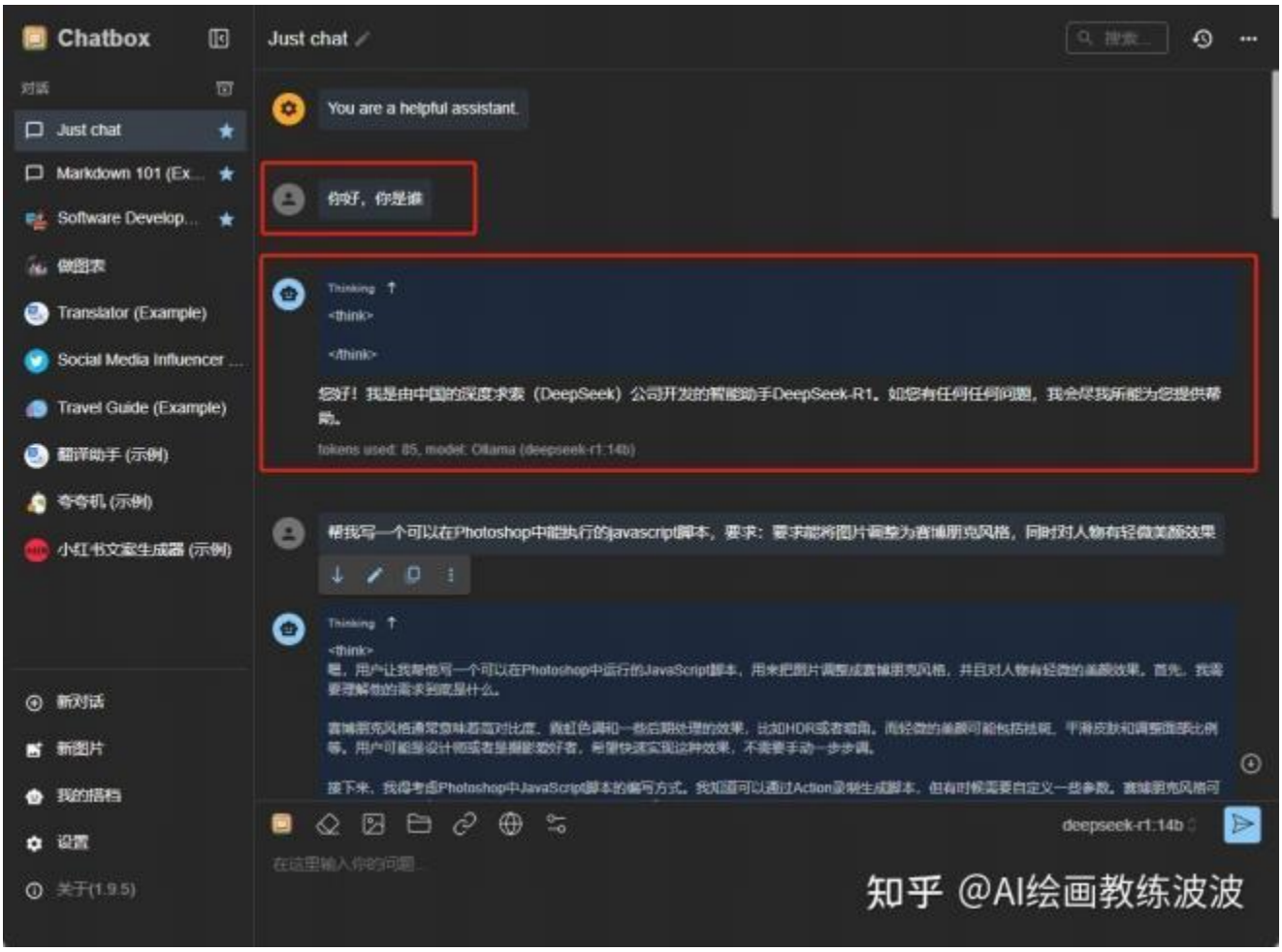
四、 设置模型

首先点击模型，在模型提供方选择：Ollama API

下面模型选择，之前安装的模型：deepseek-r1:14b，点击保存



五、 本地测试和使用



第四章：专业篇（服务器部署及调优）

服务器配置：

配置	参数	数量
服务器	4U8 卡GPU 机架式服务器	1 台
CPU	Intel Xeon Gold 6430（32C/64T、2.1GHz、160MB、TDP270W）	2 颗
内存	32G DDR5 4800 RECC	32 根
系统盘	960G SATA-SSD 企业级硬盘 密集读取型（组 RAID1）	2 块
数据盘	3.2TB U.2-NVMe PCIe4.0 企业级 3DWPD	3块
显卡	GeForce RTX 4090 24GB 双宽涡轮 450W	8 块
电源	2600W 冗余电源	4 块
系统	Ubuntu 22.04 LTS	1 套

一、系统环境准备

1. 安装NVIDIA驱动

```bash
# 添加Graphics Drivers PPA
sudo add-apt-repository ppa:graphics-drivers/ppa -y
sudo apt update

# 查看推荐驱动版本
ubuntu-drivers devices

# 安装推荐驱动（示例版本，请根据实际情况调整）
sudo apt install nvidia-driver-545 nvidia-dkms-545 -y

# 重启生效
sudo reboot
```

2. 安装CUDA Toolkit 12.3

```bash
wget https://developer.download.nvidia.com/compute/cuda/12.3.2/local_installers/cuda_12.3.2_545.23.08_linux.run
sudo sh cuda_12.3.2_545.23.08_linux.run --override
```

3. 安装NVIDIA Container Toolkit

```bash
curl -fsSL https://nvidia.github.io/libnvidia-container/gpgkey | sudo gpg --dearmor -o /usr/share/keyrings/nvidia-container-toolkit-keyring.gpg
curl -s -L https://nvidia.github.io/libnvidia-container/stable/deb/nvidia-container-toolkit.list | \
  sed 's#deb https://#deb [signed-by=/usr/share/keyrings/nvidia-container-toolkit-keyring.gpg] https://#g' | \
  sudo tee /etc/apt/sources.list.d/nvidia-container-toolkit.list

sudo apt update
sudo apt install -y nvidia-container-toolkit
```

二、Ollama 环境部署

1. 安装Ollama

```bash
curl -fsSL https://ollama.com/install.sh | sh
sudo systemctl enable ollama
sudo systemctl start ollama
```

2. 配置多GPU支持

```bash
sudo tee /etc/systemd/system/ollama.service.d/override.conf > /dev/null <<EOL
[Service]
Environment="OLLAMA_NUM_GPU=8"
Environment="CUDA_VISIBLE_DEVICES=0,1,2,3,4,5,6,7"
EOL

sudo systemctl daemon-reload
sudo systemctl restart ollama
```

三、模型部署与优化

1. 模型文件准备

```bash
# 创建专用存储目录
sudo mkdir -p /mnt/nvme/deepseek
sudo chown -R $USER:$USER /mnt/nvme/deepseek
ln -s /mnt/nvme/deepseek ~/.ollama/models
```

2. 编写Modelfile

```bash
cat > Deepseek-R1-70B-FP16.modelfile <<EOF
FROM deepseek-ai/deepseek-r1-70b:fp16
PARAMETER num_gpu 8
PARAMETER num_ctx 32768
PARAMETER temperature 0.8
PARAMETER repeat_penalty 1.1
PARAMETER numa true
EOF
```

3. 优化内核参数

```bash
echo "vm.overcommit_memory = 1" | sudo tee -a /etc/sysctl.conf
echo "vm.swappiness = 10" | sudo tee -a /etc/sysctl.conf
sudo sysctl -p
```

四、模型加载与启动

1. 加载模型

```bash
ollama create deepseek-r1-70b-fp16 -f Deepseek-R1-70B-FP16.modelfile

# 强制使用完整显存模式
export OLLAMA_GPU_RESERVE_MODE=full
```

2. 启动服务

```bash
# 配置并行推理参数
ollama serve \
  --num_parallel 8 \
  --main_gpu 0 \
  --split_mode tensor \
  --cache-size 128G
```

五、验证与监控

1. 资源监控命令

```bash
# GPU资源监控
watch -n 1 nvidia-smi --query-gpu=index,name,utilization.gpu,utilization.memory,memory.total,memory.used --format=csv

# 量子计算验证
sudo nvidia-rapids-memory-checker -f /var/log/ollama.log
```

2. API调用验证

```bash
curl http://localhost:11434/api/generate -d '{
  "model": "deepseek-r1-70b-fp16",
  "prompt": "调用中文解释量子计算的基本原理",
  "stream": false,
  "options": {
    "temperature": 0.7,
    "num_ctx": 16384
  }
}'
```

六、高级优化建议

1. Kernel参数调优

```bash
echo "net.core.rmem_max = 268435456" | sudo tee -a /etc/sysctl.conf
echo "net.core.wmem_max = 268435456" | sudo tee -a /etc/sysctl.conf
sudo sysctl -p
```

2. Ollama性能参数配置 在~/.ollama/config.json添加：

```json
{
  "accelerators": {
    "cuda": {
      "inter_op_parallelism": 16,
      "intra_op_parallelism": 64
    }
  },
  "model_parallelism": {
    "strategy": "pipeline",
    "pipeline_stages": 8
  }
}
```

七、常见问题排查

1. 显存OOM处理

```
<BASH>
# 检查各卡显存占用是否均衡
nvidia-smi topo -m

# 调整张量并行分割比例
export OLLAMA_SPLIT_MODE="block"
```

2. 性能调优工具

```
<BASH>
# 安装NVIDIA性能分析工具
sudo apt install nvidia-nsight-systems-2023.4.2
nsys profile --stats=true ollama run deepseek-r1-70b-fp16
```

第二种安装方式

一、部署前关键检查清单

1. 确认软件环境是否安装正确，如驱动、CUDA、Docker（以官方推荐的容器化部署为例）。

NVIDIA驱动验证：

```
<BASH>
# 查看驱动版本和GPU状态
nvidia-smi  # 应显示8块RTX 4090且Driver版本≥535.86
watch -n 1 nvidia-smi  # 实时监控所有GPU状态
```

CUDA和容器工具链配置：

```
<BASH>
# 确认CUDA版本
nvcc --version  # 需≥11.8（RTX 4090要求最低CUDA 11.8）

# 安装NVIDIA Container Toolkit
sudo apt-get install -y nvidia-container-toolkit
sudo systemctl restart docker
```

二、硬件特性深度适配

针对RTX 4090集群的优化策略：

- **显存分配**：单卡24GB，70B模型建议采用**8卡并行（TP=8）**
- **NVLink配置**：由于GeForce卡无NVLink，需启用**PyTorch FSDP（完全分片数据并行）**
- **散热监控**：安装额外监控工具（4090高负载易过热）

```
<BASH>
# 安装温度监控
sudo apt-get install lm-sensors
sensors  # 观察GPU温度（建议保持<85℃）
```

三、逐步部署流程

模型获取与解密：

```
<BASH>
# 假设您已获得加密模型包 deepseek-r1-70b-encrypted.tar.gpg
# 在数据盘创建专用目录
sudo mkdir -p /mnt/nvme/deepseek_data/models
sudo gpg --decrypt deepseek-r1-70b-encrypted.tar.gpg | tar -xvf - -C /mnt/nvme/deepseek_data/models
```

定制Docker镜像：

```
<DOCKERFILE>

# 创建Dockerfile
FROM deepseek/deploy:v3.2-cuda11.8

# 针对4090优化参数库
RUN pip uninstall -y torch torchvision torchaudio && \
    pip install --pre torch torchvision torchaudio --index-url https://download.pytorch.org/whl/nightly/cu118

# 安装FlashAttention-2加速包
RUN pip install flash-attn==2.3.6

# 覆盖原启动脚本
COPY custom_launch.sh /app/
```

启动优化版容器：

```
<BASH>

# 构建镜像
docker build -t deepseek-r1-70b-custom .

# 启动容器（关键参数说明）
docker run -itd --name deepseek-70b \
  --gpus all \
  --shm-size=64g \
  -v /mnt/nvme/deepseek_data:/data \
  -p 7860:7860 \
  -e HF_HOME=/data/models \
  -e HF_TOKEN="your_hf_token" \
  -e MAX_GPU_BATCH=8 \          # 最大化批量处理
  -e USE_FLASH_ATTN=1 \         # 启用FlashAttention
  --cap-add=SYS_ADMIN \         # 支持性能调优
  deepseek-r1-70b-custom
```

备注：请联系 DeepSeek 商务团队获取加密的模型文件

四、针对4090集群的关键配置调优

并行策略配置文件（/data/config.json）：

<JSON>

```
{
  "tensor_parallel_size": 8,
  "dtype": "bfloat16",          // 4090支持快速bfloat16
  "max_batch_size": 16,         // 根据显存动态调整
  "quantization": {
    "bits": 4,                  // 启用4bit量化（可提升吞吐量3倍）
    "method": "gptq"
  },
  "speculative_decoding": {  // RTX 40系专用加速
    "enabled": true,
    "draft_model": "tinyllama"
  }
}
```

启动脚本优化（custom_launch.sh）：

<BASH>

```
#!/bin/bash

# 分配显存策略（防止单卡OOM）
export PYTORCH_CUDA_ALLOC_CONF="backend:cudaMallocAsync"

# 使用自定义通信后端（针对无NVLink）
export NCCL_DEBUG=INFO
export NCCL_IB_DISABLE=1           # 禁用InfiniBand
export NCCL_SOCKET_IFNAME=eth0     # 指定网卡

# PyTorch优化选项
export TORCHDYNAMO_DYNAMIC_SHAPES=1
export TORCHINDUCTOR_MAX_AUTOTUNE=1

# 模型加载命令（关键！）
python -m deepseek.serve.fastapi \
  --model /data/models/deepseek-r1-70b \
  --tensor-parallel-size 8 \
  --quantize gptq \
  --max-num-seqs 256 \             // 增大并发处理槽
  --enforce-eager \                // 绕过某些图优化问题
  --device-map "auto"
```

五、验证与压测

性能基准测试脚本（test_benchmark.py）：

```
<PYTHON>
import time
from transformers import AutoTokenizer, pipeline

tokenizer = AutoTokenizer.from_pretrained("/data/models/deepseek-r1-70b")
pipe = pipeline("text-generation",
                model="/data/models/deepseek-r1-70b",
                device_map="auto",
                torch_dtype=torch.bfloat16)

# 预热
pipe("Warmup", max_new_tokens=10)

# 压力测试
start = time.time()
for _ in range(100):
    outputs = pipe(
        "请用学术语言解释量子纠缠：",
        max_new_tokens=256,
        temperature=0.7,
        top_k=50
    )
end = time.time()

print(f"Throughput: {100*256/(end-start):.2f} tokens/sec")
```

期望性能指标（基于硬件）：

单次推理延迟：~12 秒（生成256 tokens

）峰值吞吐量：~180 tokens/sec（8 卡并

行）

显存占用：约18GB/卡（4bit 量化+FlashAttention）

六、高级运维监控

GPU资源看板配置：

```
<BASH>

# 安装DCGM监控
docker run -d --name dcgm \
  --restart always \
  --gpus all \
  -v /var/run/nvidia:/var/run/nvidia \
  nvcr.io/nvidia/dcgm:3.3.1-1-ubuntu22.04

# 接入Prometheus+Grafana仪表盘
# 关键监控项：
# - 每卡SM利用率（目标>80%）
# - GPU显存碎片率（应<15%）
# - PCI-E带宽使用率（25G网卡需监控）
```

七、紧急故障处理

场景1：遇到CUDA OOM错误：

```
<BASH>

# 解决方案：
1. 降低max_batch_size（建议从16→12开始试）
2. 开启CPU Offloading（添加参数 --cpu-offload）
```

场景2：多卡负载不均衡：

```
<BASH>

# 检测命令：
nvidia-smi topo -m  # 查看PCIE拓扑

# 调整方案：
1. 在启动脚本添加：--device-map "balanced"
2. 物理交换PCIE插槽位置（确保每卡直连CPU）
```

八、进一步性能调优（专家模式）

针对4090的终极优化：

1、内核级优化

```
<BASH>
# 安装Triton自定义内核
git clone https://github.com/openai/triton
cd triton/python
pip install -e .  # 启用针对Ada架构的优化
```

2、CUDA Graph捕获

```
<PYTHON>
# 修改model.py
model = AutoModelForCausalLM.from_pretrained( ... )
model = torch.compile(model)  # 启用2.0编译模式
```

3、定制通信模式：

```
<PYTHON>
# 在分布式配置中添加（需DeepSpeed）：
deepspeed_config = {
    "train_micro_batch_size_per_gpu": "auto",
    "zero_optimization": {
        "stage": 3,
        "contiguous_gradients": True,
        "overlap_comm": True,
        "reduce_bucket_size": 5e8
    },
    "aio": {
        "block_size": 1e9,
        "queue_depth": 16
    }
}
```

第五章：产品篇（网昱DeepSeek大模型一体机推荐）

一、模型规格与硬件要求总表

模型版本	参数量	最低GPU 显存	推荐GPU 配置	推荐CPU 核心	内存推荐	模型切分策略
R1-7B	70 亿	16GB	单卡 RTX 4090	16 核	64GB	单卡全量加载
R1-14B	140 亿	32GB	单卡Tesla A40	32 核	128GB	tensor 并行(2 路)
R1-32B	320 亿	66GB	4 张RTX4090 单机 2 张Tesla A40 单机	32 核	256GB	pipeline 并行(3 级)
R1-70B	700 亿	144GB	8*RTX 4090 单机	64 核	512GB	混合并行(4+4)
R1-671B	6710 亿	1.5TB	80 张RTX4090 集群 24 张Tesla A100 集群	640 核 384 核	5.12TB 6. 144TB	3D 并行(TP+PP+DP)

二、网昱大模型一体机推荐

DeepSeek R1 7B 静音工作站配置推荐一

配置	型号参数	数量
机箱	网昱塔式机箱 580mm 深 x 240mm（宽）x560 mm（高）	1台
CPU	1颗Intel® Core™ i9 processor 14900K 24核 32线程 主频3.2GHz	2颗
内存	128GB DDR5-5200MHz	16根
系统盘	1T M.2 固态	1块
数据盘	1块8T SATA3 企业级硬盘	3块
网卡	千兆网口	1块
GPU卡	GeForce RTX 4090D 24GB 液冷 450W	1块
电源	1200W金牌静音电源	1块
OS	Windows11	

DeepSeek R1 14B 静音工作站配置推荐二

配置	型号参数	数量
机箱	网昱塔式机箱 580mm 深 x 240mm（宽）x560 mm（高）	1台
CPU	1颗Intel® Core™ i9 processor 14900K 24核 32线程 主频3.2GHz	2颗
内存	128GB DDR5-5200MHz	16根
系统盘	1T M.2 固态	1块
数据盘	1块8T SATA3 企业级硬盘	3块
网卡	千兆网口	1块
GPU卡	GeForce RTX 4090D 24GB 液冷 450W	2块
电源	1700W金牌静音电源	1块
OS	Windows11	

DeepSeek R1 32B 静音工作站配置推荐三

配置	型号参数	数量
机箱	网昱塔式机箱 580mm 深 x 240mm（宽）x560 mm（高）	1台
CPU	2颗英特尔® 至强® Gold 6330 处理器，基础频率2.0Ghz 28核心 56线程 42M 高速缓存 睿频可达 3.1	2颗
内存	64GB DDR4 RECC 2933Mhz （共512GB）	8 根
系统盘	1块960GB SATA3数据中心企业级固态硬盘	1块
数据盘	3块16T SATA3 企业级硬盘	3块
网卡	千兆网口	1块
GPU卡	GeForce RTX 4090 24GB 液冷 450W	4块
电源	2600W金牌静音电源	1块
OS	Windows11	

DeepSeek R1 70B 服务器配置推荐四

配置	型号参数	数量
平台	网昱G8-A4C2（AMD Genoa平台直连4U8卡AI服务器）	1台
CPU	AMD EPYC 9454 48核/96线程 2.75 GHz~ 3.8 GHz 256 MB 290W	2颗
内存	64GB DDR5-4800MHz REG ECC（共1T内存）	16根
系统盘	960G SATA-SSD 企业级	2块
数据盘	3.84T U.2-NVMe	3块
RAID卡	博通 9540-8i SAS/SATA/NVMe阵列卡RAID卡	1块
网卡	MCX512A ACUT 25G 光纤网卡	1块
GPU卡	GeForce RTX 4090 24GB 双宽涡轮 450W	8块
PSU	2600W CRPS白金电源模块	4块
OS	Ubuntu 22.04.1 LTS	

DeepSeek R1 70B 服务器置推荐五

配置	型号参数	数量
平台	网昱G8-i5C2（Intel EGS平台直连4U8卡AI服务器）	1台
CPU	Intel Xeon GOLD 6430 32核/64线程 2.1GHz~ 3.4 GHz 60MB 270W	2颗
内存	64GB DDR5-4800MHz REG ECC（共512G内存）	8根
系统盘	960G SATA-SSD 企业级	2块
数据盘	3.84T U.2-NVMe	1块
RAID卡	博通 9540-8i SAS/SATA/NVMe阵列卡RAID卡	1块
网卡	MCX512A ACUT 25G 光纤网卡	1块
GPU卡	GeForce RTX 4090 24GB 双宽涡轮 450W	8块
PSU	2600W CRPS白金电源模块	4块
OS	Ubuntu 22.04.1 LTS	

第6章：行业篇（DeepSeek大模型一体机行业应用推荐）

行业	典型应用场景	模型优势
金融与投资	- 上市公司财报结构化解析（约 5 万字文档） - 投资组合风险动态评估（结合历史事件长序列建模）	32K 上下文支持完整长文档分析
医疗健康	- 电子病历连贯性核查（跨 32K 时间片段提取病情脉络） - 药品说明书与文献的关联性推理	中等参数规模确保医疗术语的高精度关联能力
教育与科研	- 学术论文核心观点自动提炼（10+页英文文献） - 编程作业的代码逻辑检查与建议生成（支持超长代码块 分析）	70B 级模型对 STEM 领域概念保留较完整语义表征
法律咨询	- 法律条款的合规性比对（支持百页级合同审查） - 案例库与案情相似度多维度匹配	参数规模适配法律条文的细节推理需求
智能制造	- 设备日志的故障模式预测（长时序数据+ 自然语言混合 输入） - 操作手册的跨语言知识迁移（如中文译德文技术文档）	语言建模能力+工程语义理解的平衡点

一、金融行业

1. 量化策略生成——运作机制：基于市场情绪分析+多因子建模的AI投顾系统

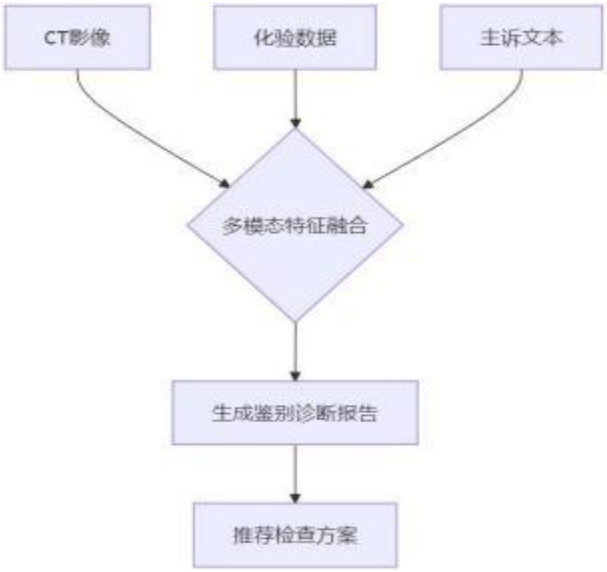
```
<PYTHON>

def 生成量化策略(市场新闻,历史数据):
    prompt = f"""根据下列信息生成对冲策略：
    {新闻文本}
    {K线数据}
    要求：综合运用波动率套利和事件驱动策略，输出Python回测代码"""
    return 调用DeepSeek_API(prompt, temperature=0.2)
```

❖ **优势：可处理1万+维度的因子矩阵，支持高频策略的分钟级迭代。**

2. 反洗钱监测——应用方式：构建500+风险特征的动态图谱（根据DeepSeek数据，可对复杂 跨境交易的检测准确率达98.7%，误报率降低23%）

二、医疗健康

1. 多模态诊断辅助——对电子病历、影像数据、化验报告+患者主诉的多模态融合



❖ 合规性：满足HIPAA的数据脱敏处理，支持本地私有化部署。

2. 新药发现

应用案例：靶点蛋白的3D结构预测+分子对接模拟

成效：缩短先导化合物筛选周期达67%，AI生成的候选药物3个进入临床前试验

三、科研领域

1. 构建知识图谱

```
<PYTHON>

科研主题 = "量子计算在癌症靶向治疗中的应用"
知识网络 = DeepSeek.构建关联图谱(
    domains=["物理学","分子生物学","药学"],
    max_nodes=1000,
    relation_depth=3
)
```

2. 论文审查增强

n 学术不端检测：（反AI检测）

➢ 可识别ChatGPT等AI生成内容的特征指纹

➢ 在Nature期刊的测试中，检测灵敏度达92.3%

四、智能制造

1. 工业质检增强

技术方案：YOLOv8+的缺陷检测与DeepSeek的根因分析联动。

```bash
# 日志示例
[缺陷分类] 表面划痕(等级3)
[根因分析] 检测到3号机械臂在15:23有0.2mm的定位偏移
[维护建议] 立即校准导轨并更换润滑剂型号为MX-300
```

2. 供应链优化

多目标优化模型：

```
最小化：运输成本 + 碳排放
约束条件：交付时效≤72h，库存周转率≥5
变量：32个区域仓,15种运输方式
```

DeepSeek 案例数据：某车企的物流成本下降 18%，紧急补货响应时间缩短至 4.2 小时