

# 本地算力部署DeepSeek详细流程

---

讲师：尚硅谷-宋红康（江湖人称：康师傅）

尚硅谷官网：<http://www.atguigu.com>

抖音账号：是康师傅呀(尚硅谷)

配套参考视频：[https://www.bilibili.com/video/BV1uqKGeZEy1/?spm\\_id\\_from=333.337.search-card.all.click&vd\\_source=faf6465655f3477e58a59d1c01d55d1d](https://www.bilibili.com/video/BV1uqKGeZEy1/?spm_id_from=333.337.search-card.all.click&vd_source=faf6465655f3477e58a59d1c01d55d1d)

---

## 1、版本选择

---

本地部署就是自己部署DeepSeek-R1模型，使用本地的算力。

主要瓶颈：内存+显存的大小。

特点：此方案不用联网。

适合：有数据隐私方面担忧的或者保密单位根本就不能上网的。

**使用满血版：**DeepSeek R1 671B 全量模型的文件体积高达720GB，对于绝大部分人而言，本地资源有限，很难达到这个配置。







根据官方及社区的讨论，满血版R1（671B，且不做量化）需要2台8卡 H100，或1台8卡 H20，或1台8卡 H200来实现所有模型参数的内存卸载。如果按这种说法，只有预算至少在200万以上的企业级应用才能用上R1本地化部署。因此，Unsloth.AI社区推出的量化版本R1可以作为使用满血版R1前的“试用装”。——Unsloth：我们探索了如何让更多的本地用户运行它，并设法将 DeepSeek 的 R1 671B 参数模型量化为 131GB，从原来的 720GB 减少了 80%，同时非常实用。

在实际部署中，不同的动态量化+版本的效果不同：

MoE Bits	Type	Disk Size	Accuracy
1.58bit	IQ1_S	131GB	正常
1.73bit	IQ1_M	158GB	好
2.22bit	IQ2_XXS	183GB	更好
2.51bit	Q2_K_XL	212GB	最好

## 使用蒸馏版：

### DeepSeek-R1-Distill Models

Model	Base Model	Download
DeepSeek-R1-Distill-Qwen-1.5B	<a href="#">Qwen2.5-Math-1.5B</a>	 <a href="#">HuggingFace</a>
DeepSeek-R1-Distill-Qwen-7B	<a href="#">Qwen2.5-Math-7B</a>	 <a href="#">HuggingFace</a>
DeepSeek-R1-Distill-Llama-8B	<a href="#">Llama-3.1-8B</a>	 <a href="#">HuggingFace</a>
DeepSeek-R1-Distill-Qwen-14B	<a href="#">Qwen2.5-14B</a>	 <a href="#">HuggingFace</a>
DeepSeek-R1-Distill-Qwen-32B	<a href="#">Qwen2.5-32B</a>	 <a href="#">HuggingFace</a>
DeepSeek-R1-Distill-Llama-70B	<a href="#">Llama-3.3-70B-Instruct</a>	 <a href="#">HuggingFace</a>

蒸馏版本：<https://huggingface.co/deepseek-ai>

开源2+6个模型。R1预览版和正式版的参数高达660B，非一般公司能用。为进一步平权，于是他们就蒸馏出了6个小模型，并开源给社区。最小的为1.5B参数，10G显存可跑。

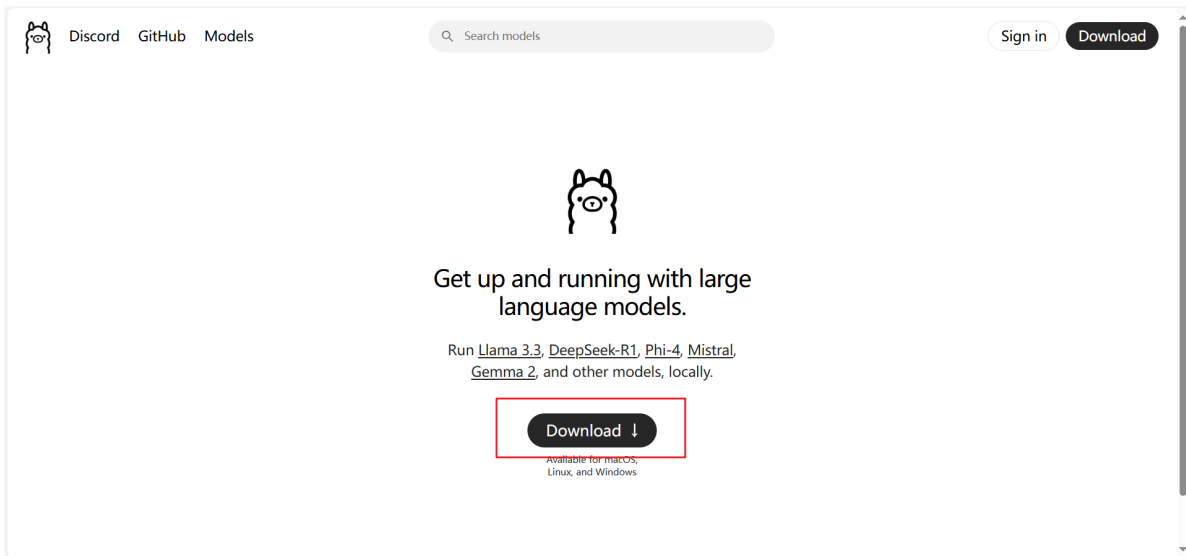
如果你要在个人电脑上部署，一般选择其他架构的蒸馏模型，本质是微调后的Llama或Qwen模型，基本32B以下，并不能完全发挥出DeepSeek R1的实力。

## 2、部署过程

比较流行的是使用ollama: <https://ollama.com/>

Ollama (<https://ollama.com/library>, 可以理解为替换前面的服务器端, 在本地作为服务端, 可以是别的平台) + ChatBox、Cherry Studio等 (<https://chatboxai.app/zh>, 前端, 可以是别的工具如Chrome插件Page Assist或Anything LLM)

## 步骤1: 下载Ollama



## Download Ollama

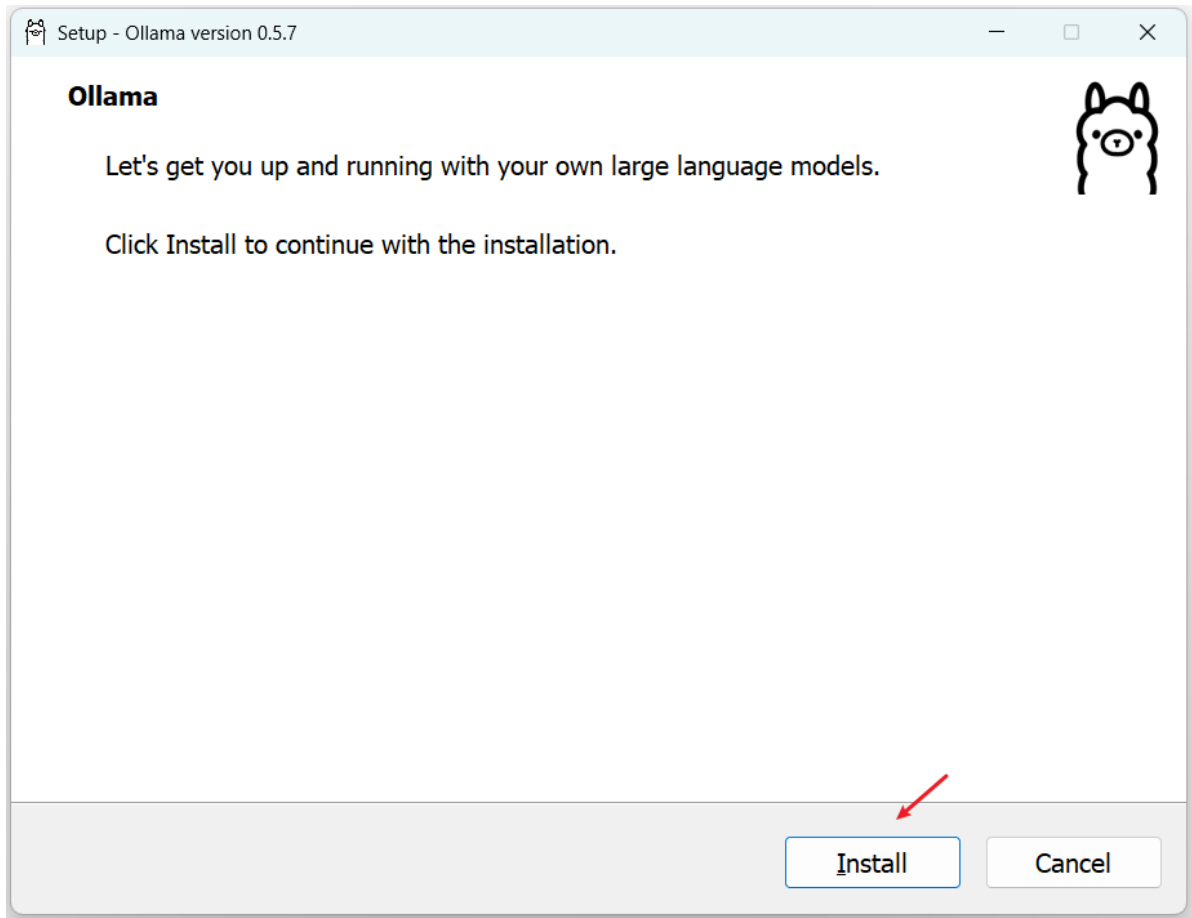


Download for Windows

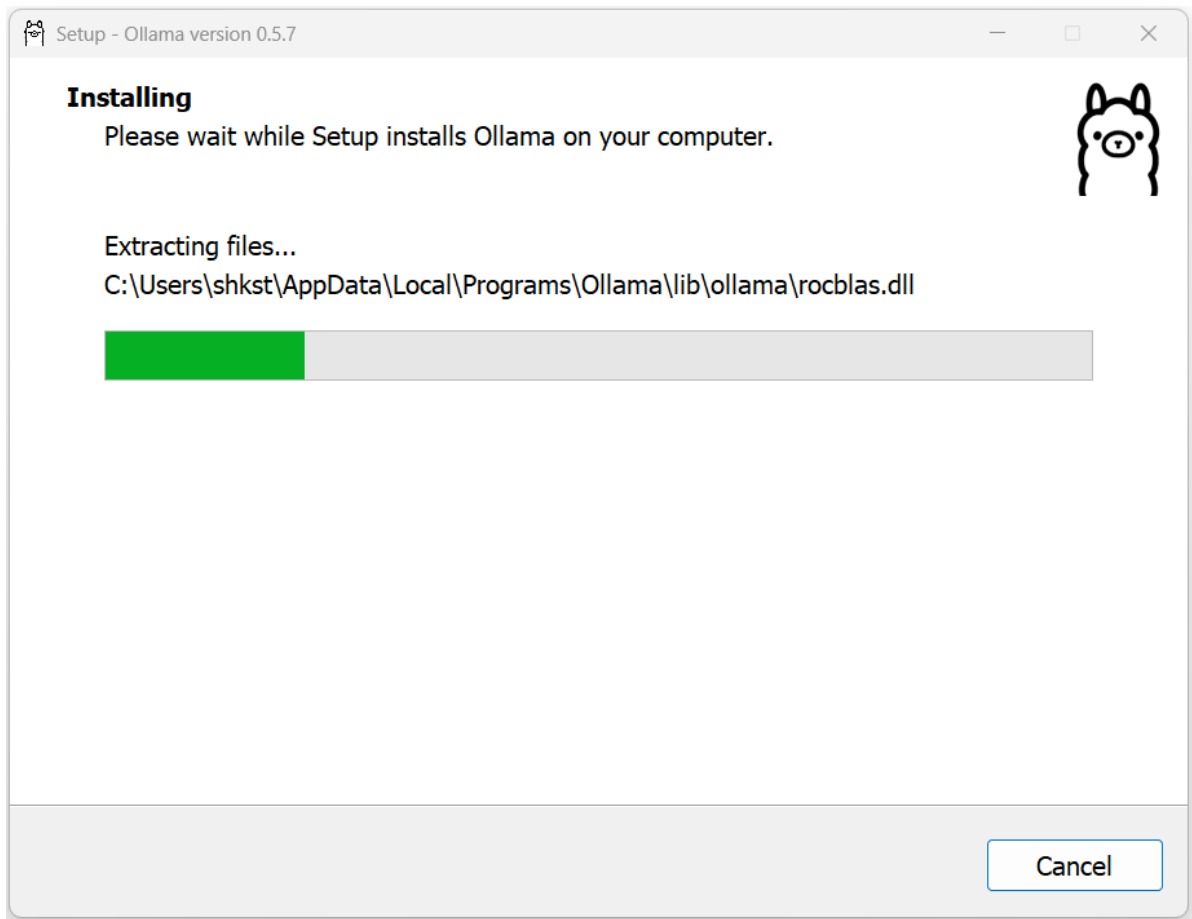
Requires Windows 10 or later

## 步骤2：安装Ollama

傻瓜式安装。过程如下：

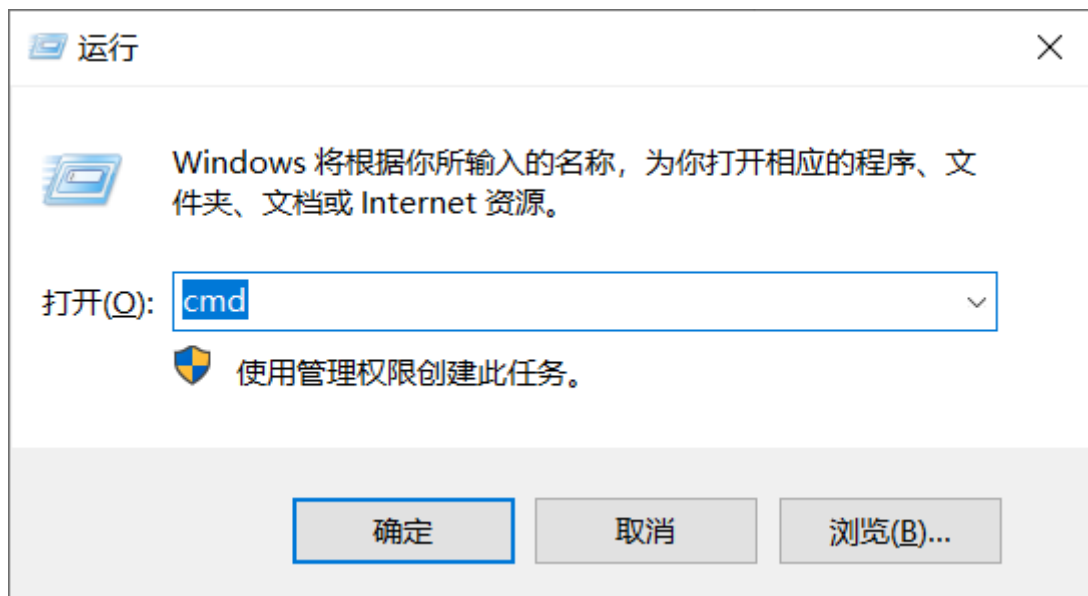


安装过程：



安装完后，验证是否安装成功：

”win+r“输入cmd



命令行输入如下：

```
ollama -v
```

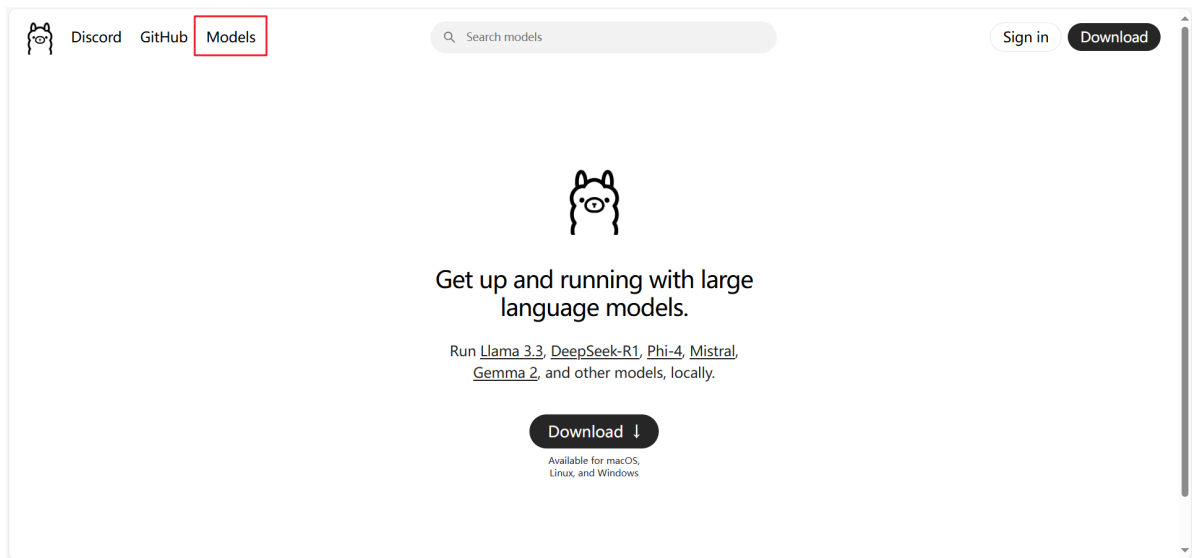
```
C:\WINDOWS\system32\cmd.exe
Microsoft Windows [版本 10.0.26100.2894]
(c) Microsoft Corporation。保留所有权利。

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

C:\Users\shkst>
```

能显示ollama版本说明安装成功。

## 步骤3：选择r1模型



**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

9.3M Pulls 28 Tags Updated 2 weeks ago

**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

# 步骤4：选择版本

**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

9.3M Pulls Updated 2 weeks ago

7b 28 Tags ollama run deepseek-r1

1.5b	1.1GB	0a8c26691023 · 4.7GB
7b	4.7GB	parameters 7.62B · quantization Q4_K_M 4.7GB
8b	4.9GB	begin_of_sentence   >", "<  end_of_sentence   >...
14b	9.0GB	
32b	20GB	}}{{ .System }}{{ end }} {{- range \$i, \$_ := .Mes...
70b	43GB	yright (c) 2023 DeepSeek Permission is hereby gra...
671b	404GB	

View all

Readme

b代表10亿参数量，7b就是70亿参数量。这里的671B是HuggingFace经过4-bit 标准量化的，所以大小是404GB。

ollama 支持 CPU 与 GPU 混合推理。将**内存与显存之和**大致视为系统的“**总内存空间**”。

如果你想运行404GB的671B，建议你的内存+显存能达到500GB以上。

除了模型参数占用的内存+显存空间（比如671B的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。所以根据个人电脑的配置，评估你选择部署哪一个版本。如果你想运行404GB的671B，建议你的内存+显存能达到500GB以上。

这里我们以7B为例，大多数的电脑都能够运行起来。

## 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

↓ 9.4M Pulls    ⌚ Updated 2 weeks ago

点击复制

7b    28 Tags    ollama run deepseek-r1

Updated 2 weeks ago	0a8c26691023 · 4.7GB
model	arch <b>qwen2</b> · parameters <b>7.62B</b> · quantization <b>Q4_K_M</b> 4.7GB
params	{ "stop": [ "< begin_of_sentence >", "< end_of_sentence >..." ] }    148B
template	{{- if .System }}{{ .System }}{{ end }} {{- range \$i, \$_ := .Mes...}}    387B
license	MIT License Copyright (c) 2023 DeepSeek Permission is hereby gra...    1.1kB

## 步骤5：本地运行DeepSeek模型

在命令行中，输入如下命令：

```
ollama run deepseek-r1:7b
```



```
C:\WINDOWS\system32\cmd.exe
Microsoft Windows [版本 10.0.26100.2894]
(c) Microsoft Corporation。保留所有权利。

C:\Users\shkst>ollama run deepseek-r1:7b
```

首次运行会下载对应模型文件：

```
C:\WINDOWS\system32\cmd.exe
Microsoft Windows [版本 10.0.26100.2894]
(c) Microsoft Corporation。保留所有权利。

C:\Users\shkst>ollama run deepseek-r1:7b
pulling manifest
pulling 96c415656d37... 6% |          | 283 MB/4.7 GB 29 MB/s 2m26s|
```

```
C:\WINDOWS\system32\cmd.exe
Microsoft Windows [版本 10.0.26100.2894]
(c) Microsoft Corporation。保留所有权利。

C:\Users\shkst>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)
```

下载支持断点续传，如果下载中速度变慢，可以鼠标点击命令行窗口，然后ctrl+c取消，取消后按方向键“上”，可以找到上一条命令，即“ollama run deepseek-r1:7b”，按下回车会重新链接，按照之前进度接着下载。

如果不想下载，也可以直接使用我们提供的下载好的模型文件，按照后续小节“修改models文件夹路径”的步骤配置好环境变量和对应models文件夹的路径即可。

下载完成后，自动进入模型，直接在命令行输入问题，即可得到回复。

比如：打个招呼！

```
>>> 你好  
<think>  
  
</think>
```

```
你好！很高兴见到你，有什么我可以帮忙的吗？无论是学习、工作还是生活中的问题，都可以告诉我哦！ 😊
```

比如：你好，鲨鱼为什么会溺水呢？

```
C:\WINDOWS\system32\cmd.exe
writing manifest
success
>>> 你好，鲨鱼为什么会溺水呢？
<think>
嗯，鲨鱼为什么会溺水呢？这个问题听起来有点奇怪。鲨鱼是海洋中的掠食者，生活在淡水和咸水中，对吧？那它们怎么会溺水呢？可能是因为我对中国的一些地理知识不太清楚。

让我先回想一下，中国有很多淡水湖泊和河流，比如鄱阳湖、洞庭湖、巢湖等等。这些地方有很深的水域，可能有鱼在浅层游泳，而鲨鱼则喜欢深水区觅食。如果有人在这些地方钓鱼或者划船，可能会遇到鲨鱼。

那为什么鲨鱼会在浅水中呢？通常，鲨鱼喜欢深水，因为它们获得的食物在那里更丰富，而且有更多的庇护所，比如珊瑚礁和岩层。如果他们出现在浅水区，可能是因为某些原因，比如季节性变化、气候变化或者人为活动导致的水域深度变化。

另外，“溺水”这个词在中文里通常指人淹死水中的情况。鲨鱼是海洋生物，它们自己不会像人类那样呼吸水中的空气，所以它们不会直接从水中窒息而亡。但是，如果有人不小心让鲨鱼进入他们的水域，比如在游泳池或其他地方，可能会有危险。

或者，这个问题可能是指“鲨鱼为何会在浅水中出现，导致人们溺水”。这样的话，可能是由于人为因素，比如大型水族馆、人工创造的水域或某些水利工程改变了自然的水深分布，使得鲨鱼出现在更浅的地方。在这种情况下，如果有人在这些地方游泳，可能会遇到鲨鱼攻击的风险。

另外，还有一种可能性是误解，可能问题中的“溺水”并不是指人的溺水，而是指鲨鱼的溺水。但是，作为海洋生物，鲨鱼不会像人类一样在水里窒息而死，除非它们本身有呼吸问题，但这通常是罕见的事件。

总结一下，鲨鱼不会自己溺水，因为它们是适应水深生活的海洋动物。如果有人提到鲨鱼溺水，可能是因为他们误解了鲨鱼出现的位置或环境，或者是描述某种与鲨鱼相关的危险情况。
</think>

根据您的问题，“鲨鱼为什么会溺水呢？”，经过思考和分析，可以得出以下结论：

鲨鱼不会自己“溺水”。它们是海洋中的掠食者，习惯于在深水中生活。如果提到鲨鱼溺水，可能是因为误解或描述与鲨鱼相关的人为风险情况，比如在浅水区活动时遇到危险。

**逐步解释和答案：**

1. **鲨鱼的栖息地**：鲨鱼主要生活在淡水和咸水的深水区，食物丰富且适
```

获取帮助：

/?

```
>>> /?
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.
```

退出对话:

```
/bye
```

```
>>> /bye
C:\Users\shkst>
```

## 步骤6: 查看已有模型

查询已有模型:

```
ollama list
```

```
C:\Users\shkst>ollama list
NAME          ID          SIZE      MODIFIED
deepseek-r1:7b 0a8c26691023 4.7 GB    4 minutes ago
```

后续要运行模型, 仍然使用之前的命令:

```
C:\Users\shkst>ollama run deepseek-r1:7b
>>> |Send a message (/? for help)
```

## 3、使用客户端工具

本地部署好模型之后, 在命令行操作还是不太方便, 我们继续使用一些客户端工具来使用。

## Cherry Studio的下载:

Cherry Studio的下载地址: <https://cherry-ai.com/>



## 其他版本下载

### Windows系统安装包

Cherry-Studio-0.9.19-setup.exe (84.6 MB) - Windows标准安装包

Cherry-Studio-0.9.19-portable.exe (84.2 MB) - Windows便携版

### MacOS系统安装包

Cherry-Studio-0.9.19-x64.dmg (112.5 MB) - Intel芯片Mac

Cherry-Studio-0.9.19-arm64.dmg (105.0 MB) - Apple Silicon芯片Mac

### Linux系统安装包

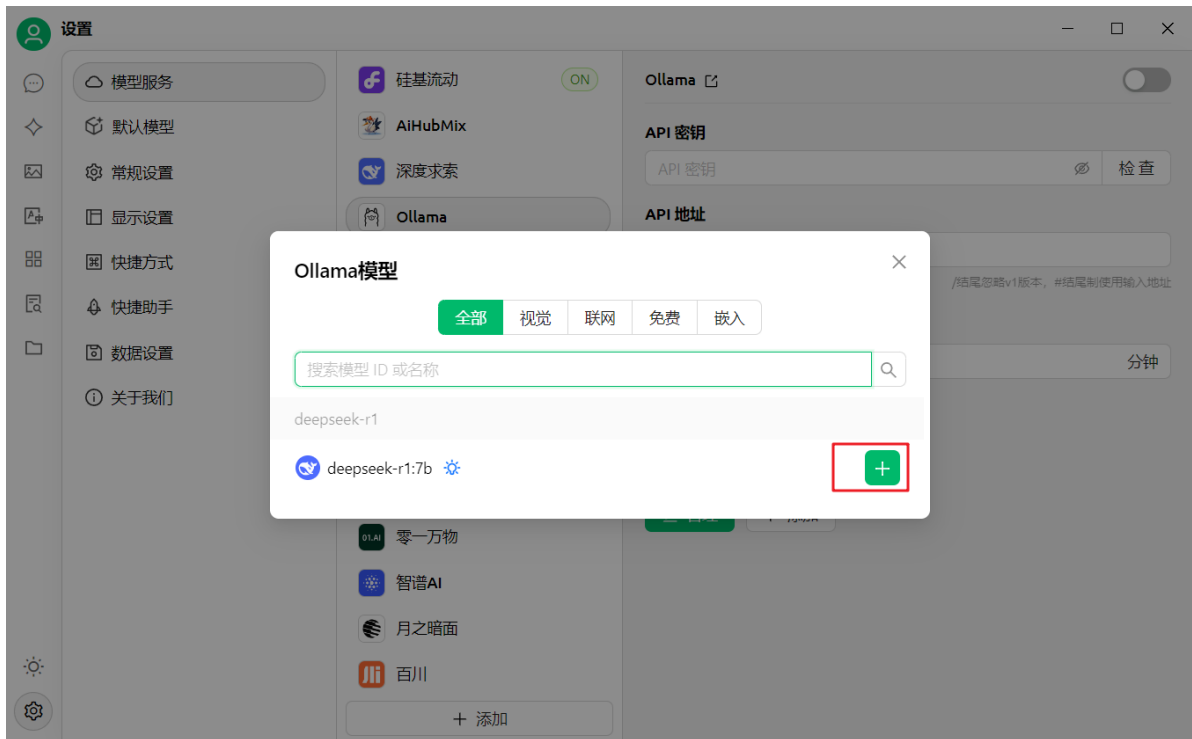
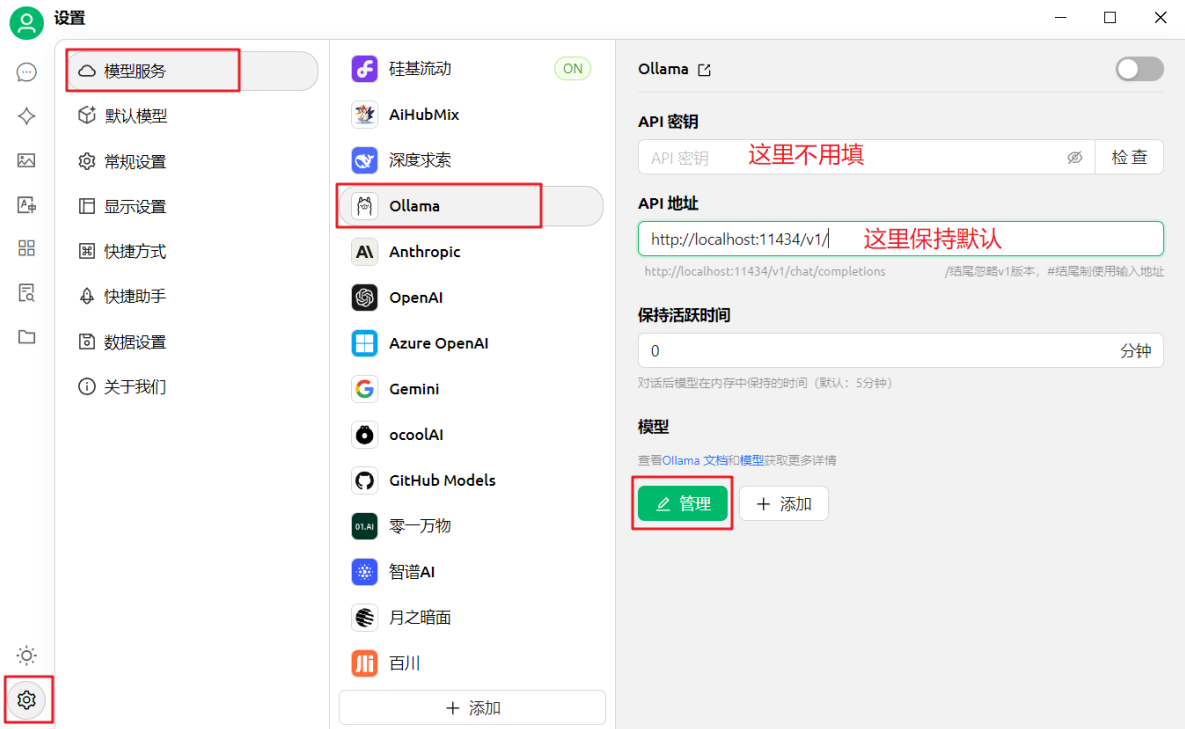
Cherry-Studio-0.9.19-x86\_64.AppImage (120.5 MB) - x86\_64架构

Cherry-Studio-0.9.19-arm64.AppImage (120.3 MB) - ARM架构

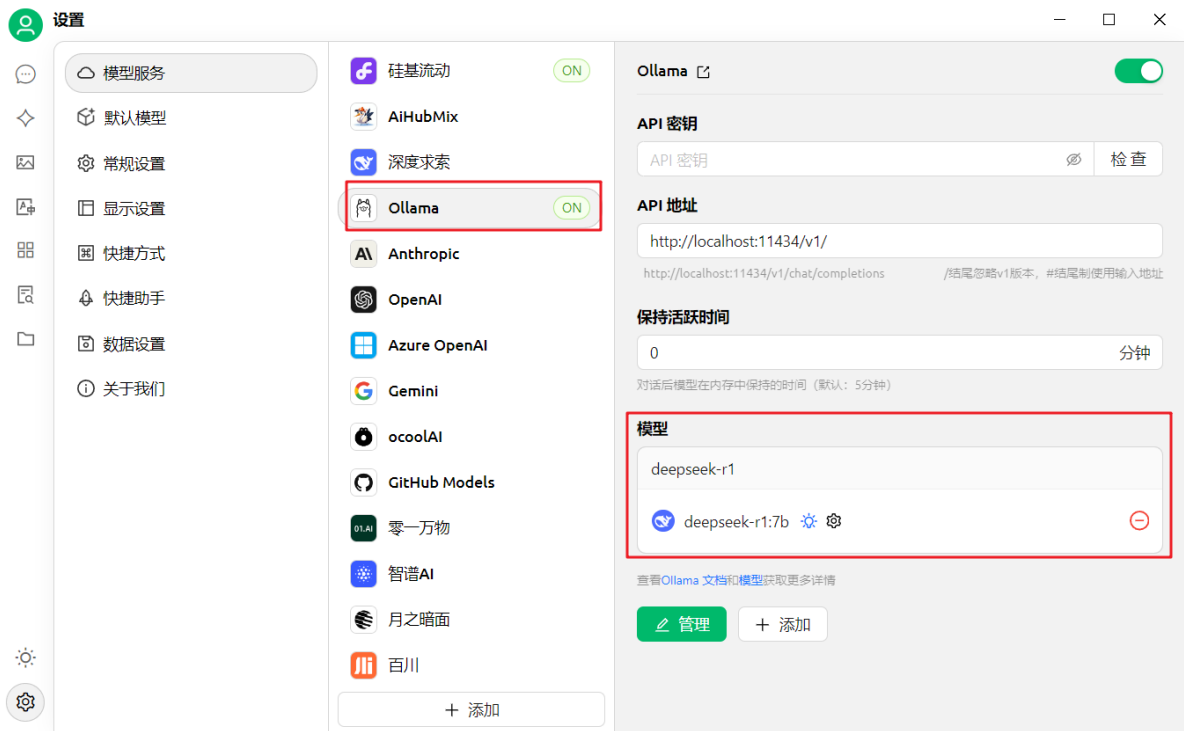
关于Cherry Studio更多功能的使用文档: <https://zhuanlan.zhihu.com/p/10585626732>

**Cherry Studio的安装:** 傻瓜式安装, 这里省略。

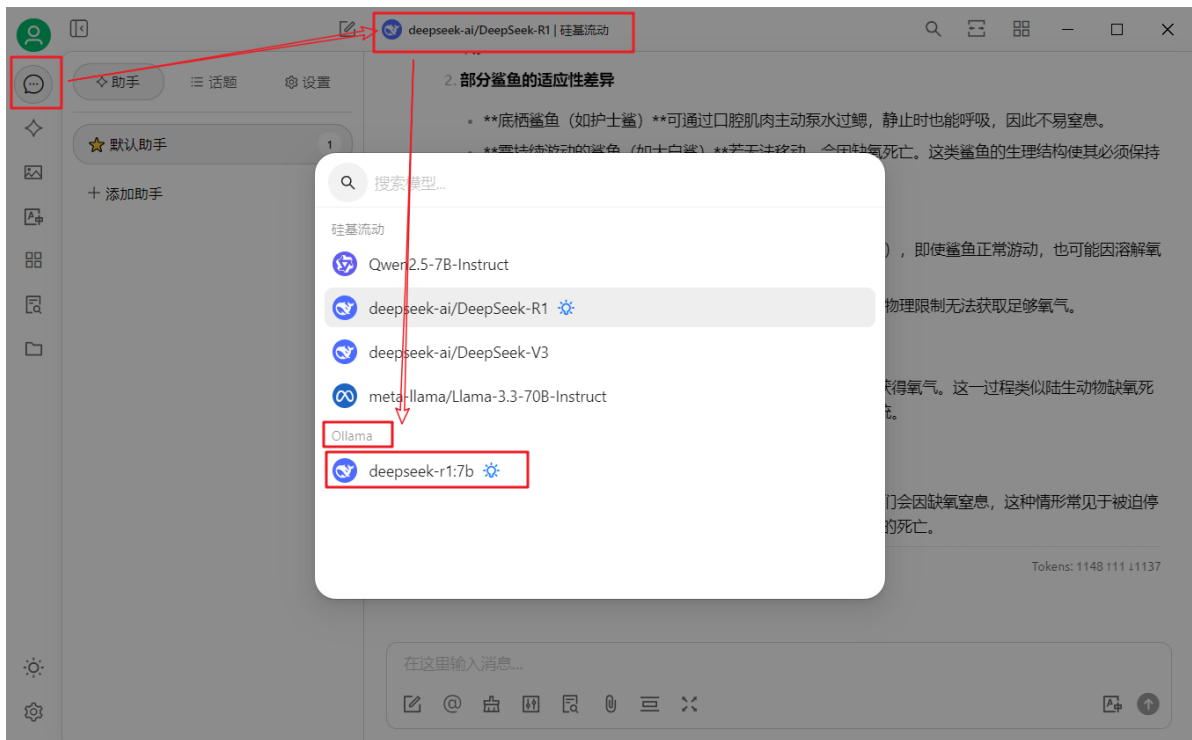
**以Cherry Studio为例访问7b的蒸馏模型:**



如果这里列表没有r1模型，则是之前没有安装好。



## 选择模型



**注意：**使用时要确保ollama客户端已启动



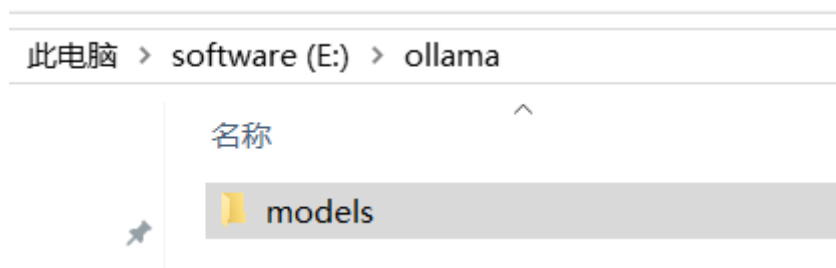


## 4、修改models文件夹路径（可选）

模型默认会下载到：`C:\Users\你的用户名\.ollama` 目录下的models文件夹

如果想修改模型的存放位置，做如下配置：

**步骤1：** 拷贝models文件夹到你指定的目录，比如我剪切到 `E:\ollama` 下



**步骤2：** 添加环境变量

右键“我的电脑”，选择“属性”，按如下方式配置：



## 关于

系统正在监控并保护你的电脑。

[在 Windows 安全中心中查看详细信息](#)

## 设备规格

设备名称

处理器

机带 RAM

设备 ID

产品 ID

系统类型

笔和触控

复制

重命名这台电脑

## Windows 规格

版本

版本号

安装日期

操作系统内部版本

序列号

体验



相关设置

[BitLocker 设置](#)

[设备管理器](#)

[远程桌面](#)

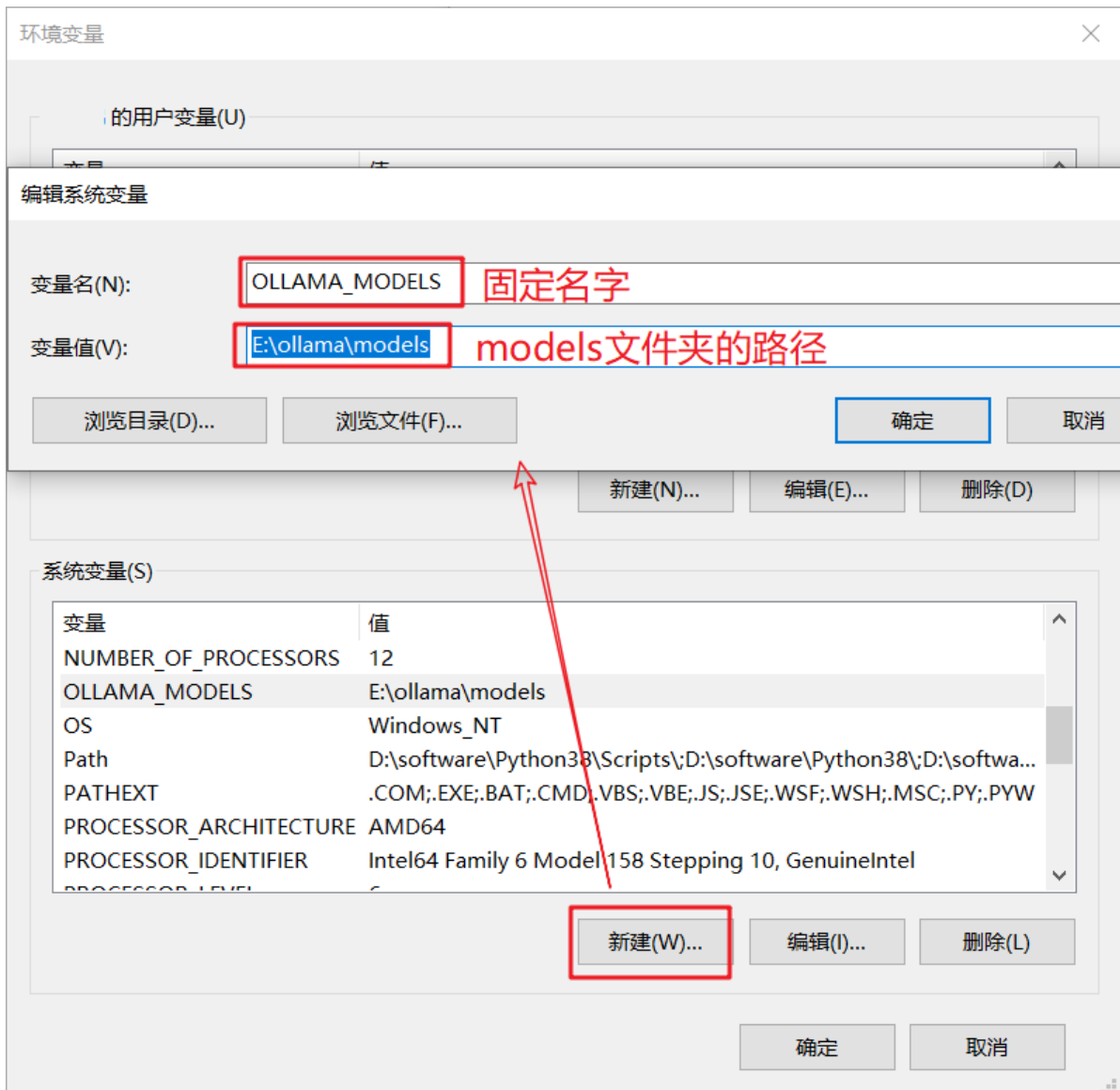
[系统保护](#)

**高级系统设置**

[重命名这台电脑](#)

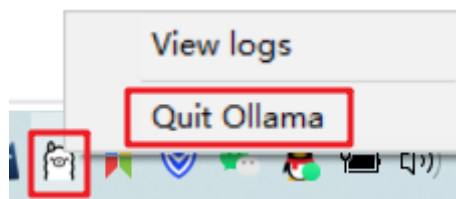
[获取帮助](#)

[提供反馈](#)



**步骤3: 重启Ollama客户端生效**

**注意：**修改完之后，需要重启Ollama客户端，右键图标，选择退出，重新运行Ollama



验证是否生效：重新运行Ollama之后，重新打开命令行，输入命令 `ollama list` 查看：

```
管理员: C:\WINDOWS\System32\cmd.exe
Microsoft Windows [版本 10.0.19045.5371]
(c) Microsoft Corporation。保留所有权利。

E:\ollama\models>ollama list
NAME          ID          SIZE      MODIFIED
deepseek-rl:7b 0a8c26691023 4.7 GB    9 minutes ago
E:\ollama\models>
```

如果list显示为空，则表示操作有问题，确认以上步骤。

## 5、其它方式：服务器部署

在企业中，想要私有化部署满血版DeepSeek-R1，即671B版本，需要有更好的硬件配置。

服务器可以是物理机，也可以是云服务器。

使用Ollama提供的经过量化压缩的671B模型的大小是404GB，建议内存 + 显存  $\geq$  500 GB，举例几种性价比配置如下：

- Mac Studio：配备大容量高带宽的统一内存（比如 X 上的 @awnihannun 使用了两台 192 GB 内存的 Mac Studio 运行 3-bit 量化的版本）
- 高内存带宽的服务器：比如 HuggingFace 上的 alain401 使用了配备了 24×16 GB DDR5 4800 内存的服务器)

- 云 GPU 服务器：配备 2 张或更多的 80GB 显存 GPU（如英伟达的 H100，租赁价格约 2 美元 / 小时 / 卡）

在这些硬件上的运行速度可达到 10+ token / 秒。

**部署流程与个人电脑部署7B的流程没有太大区别，都是以下几个步骤：**

1. 根据服务器的操作系统，下载对应版本的Ollama客户端；
2. 运行Ollama，执行Ollama命令运行671B版本模型；首次执行自动下载模型；
3. 使用客户端工具/自己开发页面/代码调用，对接Ollama的R1模型；