llama3本地部署、微调及量化

1、llama3.1本地部署

Ollama是一个基于Go语言开发的本地大语言模型运行框架，它允许用户在本地环境中运行和定制大型语言模型。Ollama提供了一个简单易用的命令行界面和服务器，使得用户可以轻松下载、运行和管理各种开源的大型语言模型（LLM）。它支持多种操作系统，包括macOS、Windows、Linux，以及通过Docker容器在几乎任何支持Docker的环境中运行。

Ollama的特点包括：

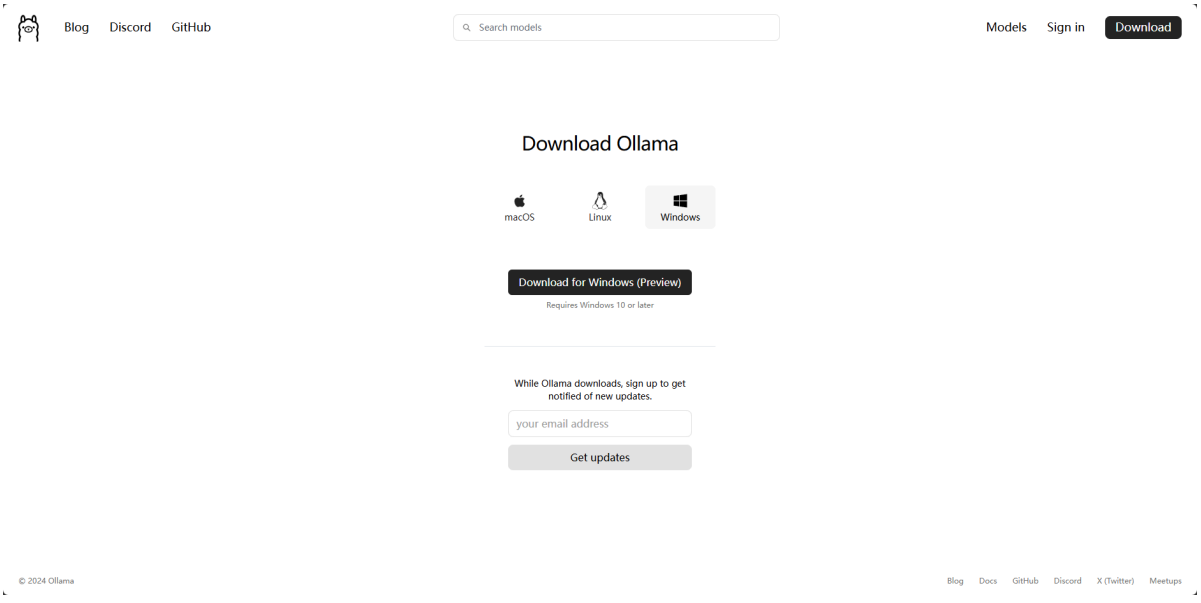
- 开源免费，允许自由使用、修改和分发。
- 简单易用，无需复杂配置，只需简单命令即可启动和运行。
- 模型丰富，支持Llama 3、Mistral、Qwen2等众多热门开源LLM。
- 资源占用低，即使在普通笔记本电脑上也能流畅运行。
- 社区活跃，提供帮助、分享经验和参与模型开发的空间。

使用Ollama的基本步骤：

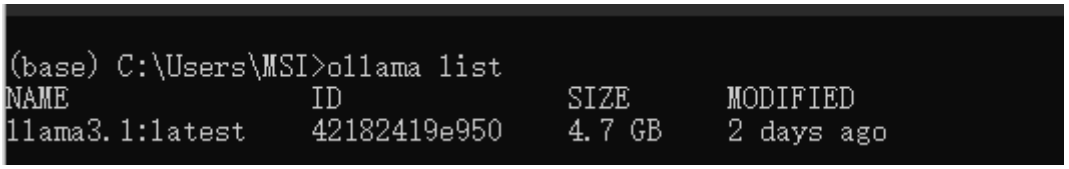
- **安装ollama**

ollama下载链接：

Download Ollama on Windows



- **在cmd中运行Ollama**



在cmd中运行

```
ollma list
```

即可查看当前已下载的模型

- **下载并运行模型**

预先查找想要下载的模型名字，比如我这里想下载llama3.1，运行以下命令即可

```
ollma run llama3.1
```

在第一次运行的时候会对模型进行下载，如果是已经下载过的模型，那么会直接运行这个模型，运行成功后可以与该语言模型进行交互

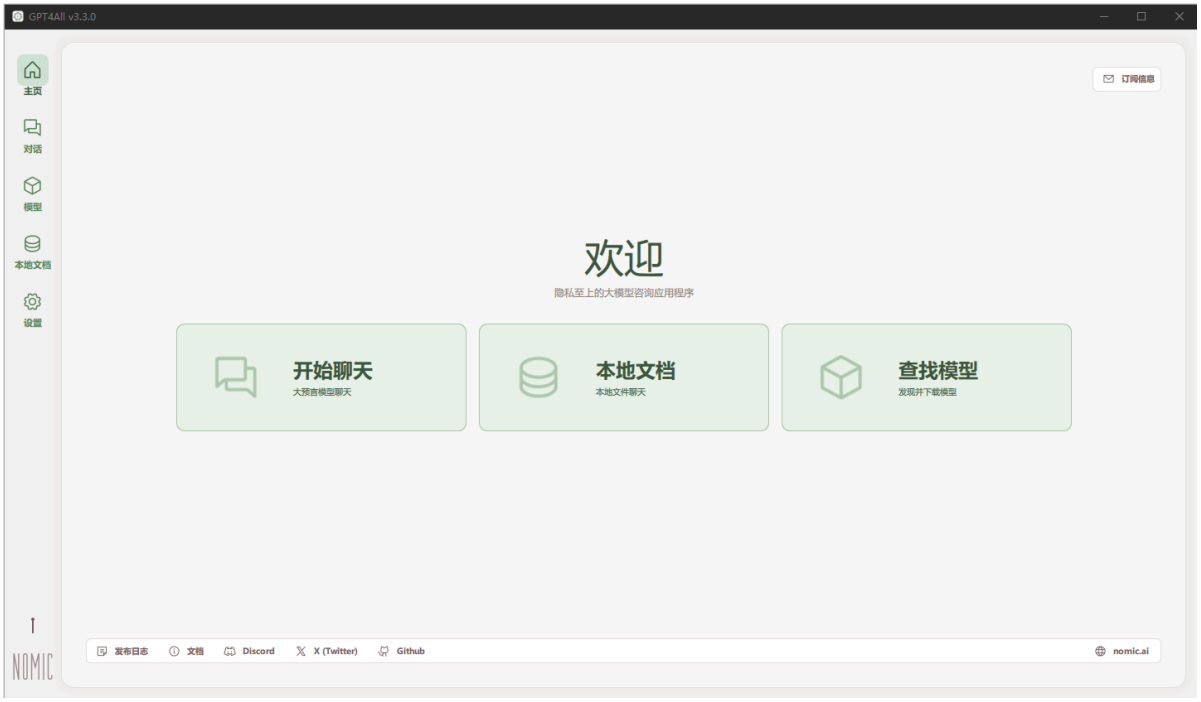
```
(base) C:\Users\MSI>ollama run llama3.1
>>> hello,what's your name
I don't have a personal name. I'm an AI designed to assist and communicate with users, so I'm often referred to as
a "chatbot" or a "language model." You can think of me as a helpful conversational companion! How are you doing
today?

>>> Send a message (/? for help)
```

想要查看有哪些可供下载的模型可以查看该链接[library (ollama.com)]

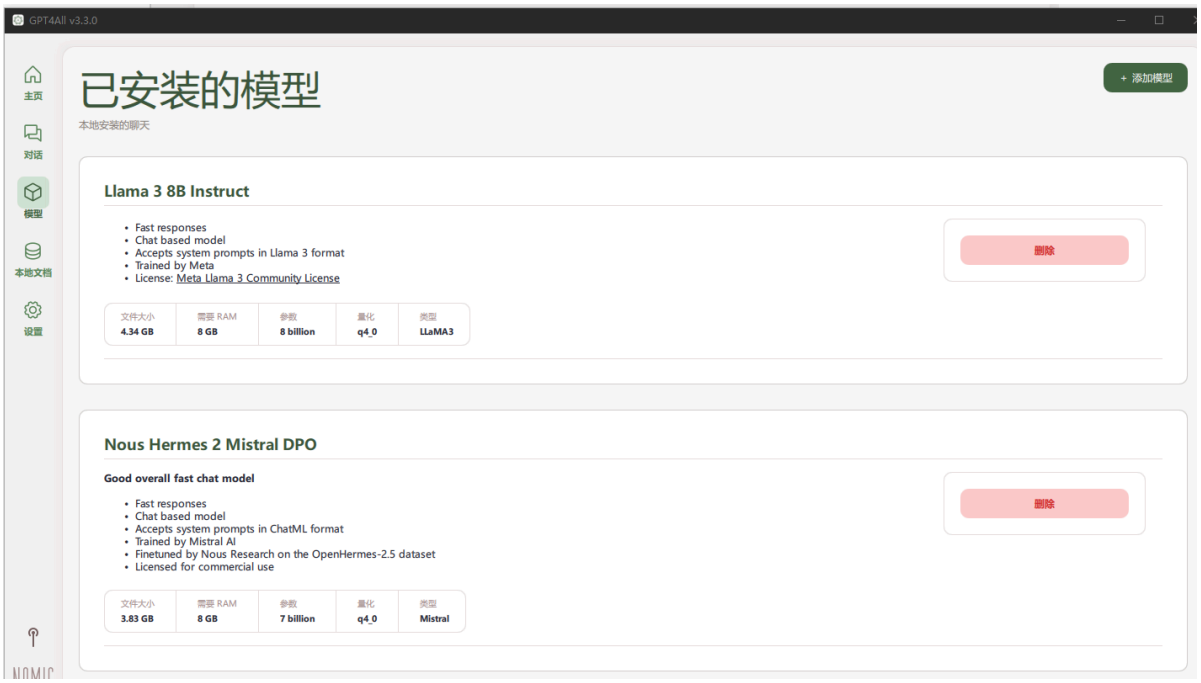
2、GPT4ALL可视化界面

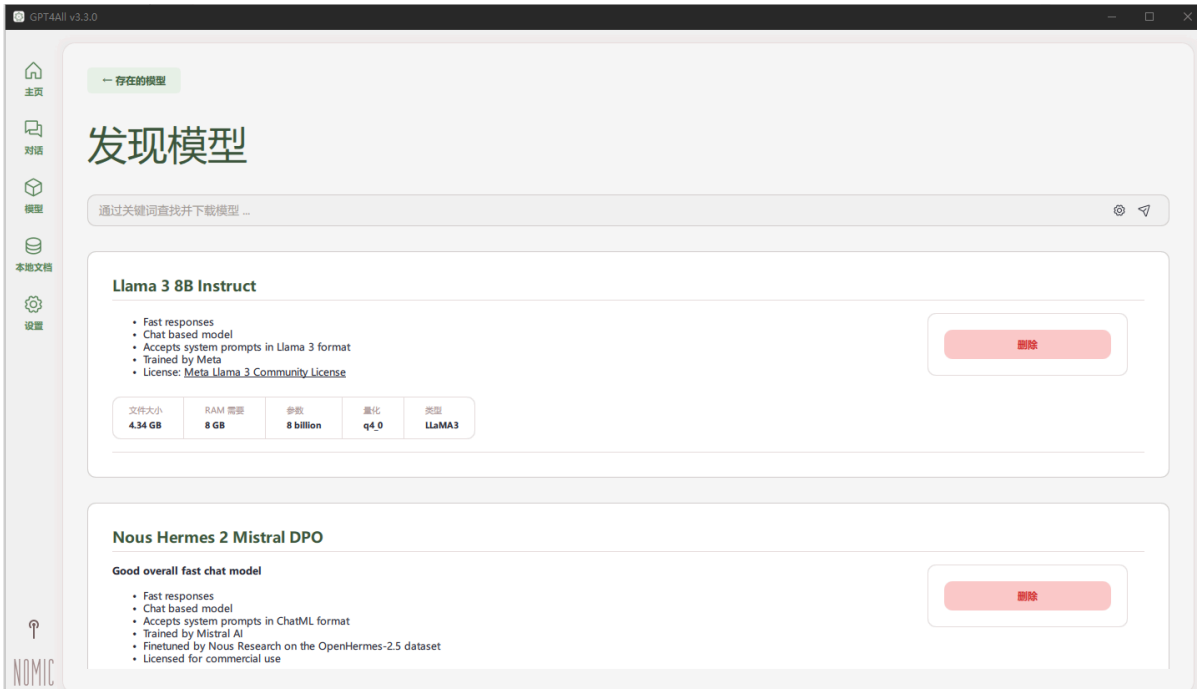
鉴于现在的ollama并没有可视化ui界面，这里提供一个可选的大模型本地部署gui软件GPT4ALL



下载链接：[nomic-ai/gpt4all: GPT4All: Run Local LLMs on Any Device. Open-source and available for commercial use. (github.com)]

利用gpt4all软件，我们可以直接从它的模型库中下载所有的开源大模型，只需要点进模型界面，然后点击添加模型即可下载想要的大模型





下载好想要的模型之后，我们便可以在对话界面对模型进行加载，加载完成后便可以像在线的chatgpt那样进行交互

不过该软件目前的缺点就是仍在开发中，所以在加载模型的时候经常会遇到闪退的现象，这个可以关注后续版本的更新

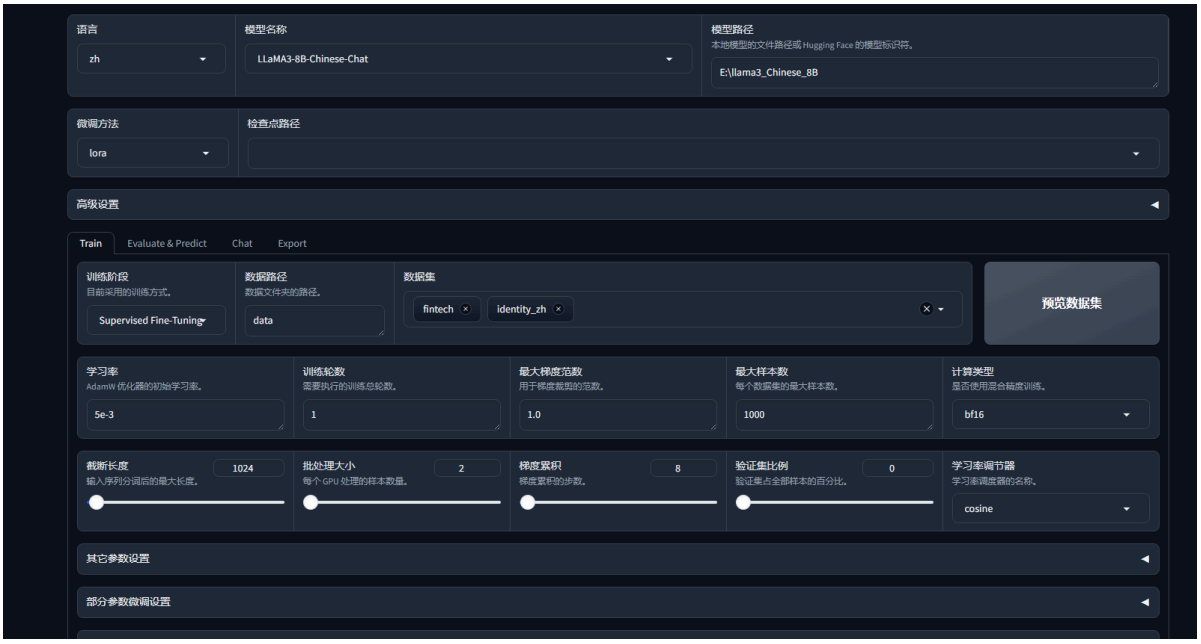
3、大模型微调

上面的ollama只能让我们对大模型进行本地部署并进行日常使用，但是往往别人预训练好的模型与我们实际的应用场景并不会那么相符，比如原生的llama3在训练的时候超过95%的数据集用的都是英文数据，中文数据只占很小的比例，所以我们如果想要直接把llama3应用到我们的各种应用场景中就会遇到很多问题。

因此我们需要对大模型进行微调操作，使其更符合我们当下的需要，比如我们更希望它能运用中文与我们进行交互，那么我们就要用自己的中文数据集对大模型进行微调（已经有比较好的微调llama3中文模型了，llama3-Chinese-chat），下面我将介绍如何对llama3-8B-Chinese-chat这个预训练语言模型进行微调

llama-factory

下载链接：[hiyouga/LLaMA-Factory: Efficiently Fine-Tune 100+ LLMs in WebUI (ACL 2024) (github.com)](https://github.com/hiyouga/LLaMA-Factory)



LLaMA Factory 是一个旨在简化大型语言模型（LLM）训练和微调过程的平台。它支持多种预训练模型，并提供了一系列工具和算法来帮助用户在本地或云端环境中高效地微调模型。以下是LLaMA Factory的一些关键特性：

1. **模型种类**：支持包括LLaMA、LLaVA、Mistral、Mixtral-MoE、Qwen、Yi、Gemma、Baichuan、ChatGLM、Phi 等多种模型。
2. **训练算法**：提供多种训练算法，包括增量预训练、多模态指令监督微调、奖励模型训练、PPO、DPO、KTO、ORPO 等。
3. **运算精度**：支持16比特全参数微调、冻结微调、LoRA微调，以及基于AQLM/AWQ/GPTQ/LLM.int8/HQQ/EETQ的2/3/4/5/6/8比特QLoRA微调。
4. **优化算法**：包括GaLore、BAdam、DoRA、LongLoRA、LLaMA Pro、Mixture-of-Depths、LoRA+、LoftQ和PiSSA等。
5. **加速算子**：支持FlashAttention-2和Unsloth等加速算子。
6. **推理引擎**：支持Transformers和vLLM推理引擎。
7. **实验面板**：提供LlamaBoard、TensorBoard、Wandb、MLflow等多种实验面板工具。

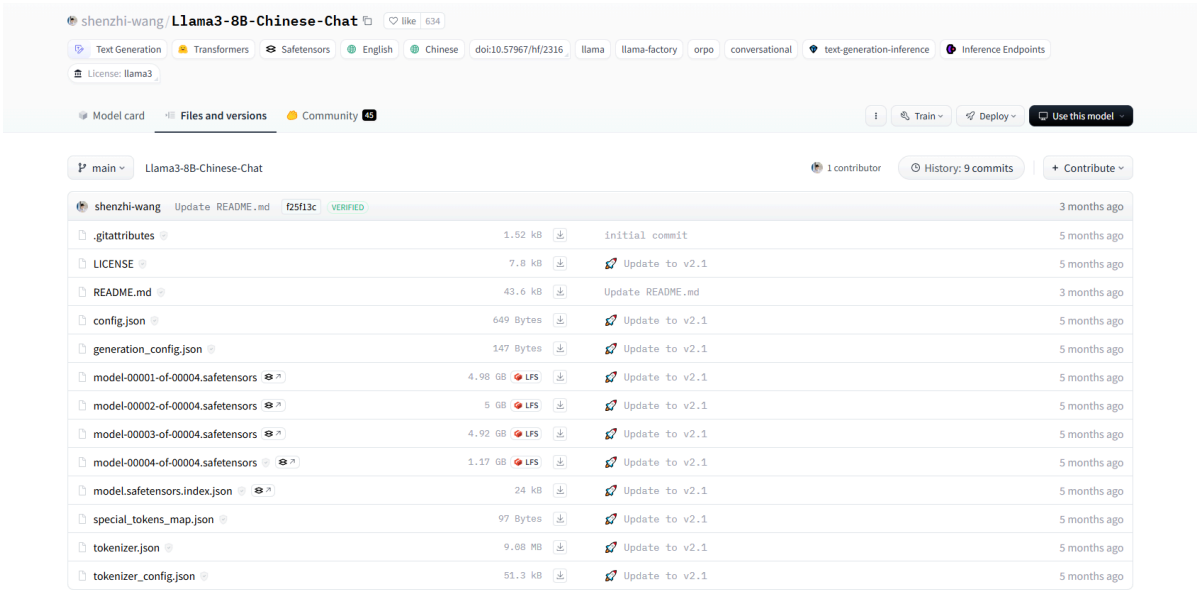
LLaMA Factory 通过提供高层次的抽象接口，使得开发者即使在没有深入技术细节的情况下也能进行模型的微调。此外，LLaMA Factory 还提供了基于gradio的网页版工作台，方便初学者快速上手操作，开发出自己的第一个模型。

微调操作演示

因为原生llama3模型所用以训练的数据集中英文数据占比超过95%，所以我们用该模型进行微调的效果会比较差，因此我们采用国人已经预训练好的llama3-8B-Chinese-chat模型来进行微调，以取得更好的效果

llama3-8B-Chinese-chat模型下载链接：[shenzhi-wang/Llama3-8B-Chinese-Chat · Hugging Face](https://huggingface.co/shenzhi-wang/Llama3-8B-Chinese-Chat)

在下载的时候如果用huggingface命令行去下载，因为国内网络的原因，很有可能会频繁中断（中断了3次...），所以这里推荐在files and versions里面手动下载，逐个点击下载然后放到同一个文件夹即可



LORA是一种微调方法，相关链接：[大模型高效微调-LoRA原理详解和训练过程深入分析-CSDN博客]

什么是量化，相关链接：[为啥大模型需要量化？如何量化 (qq.com)]

量化两大作用：

- 降低显存需要
- 提升推理性能

接下来推荐一个b站视频，里面详细地解释了如何利用llama3进行微调和量化：[【大模型微调】使用Llama Factory实现中文llama3微调*哔哩哔哩*bilibili]

当然对于大模型的部署和微调对于显存的要求还是挺高的，部署还好一点，微调估计只能在咱们的实验室服务器上面跑