调用 huggingface 下载到本地的模型

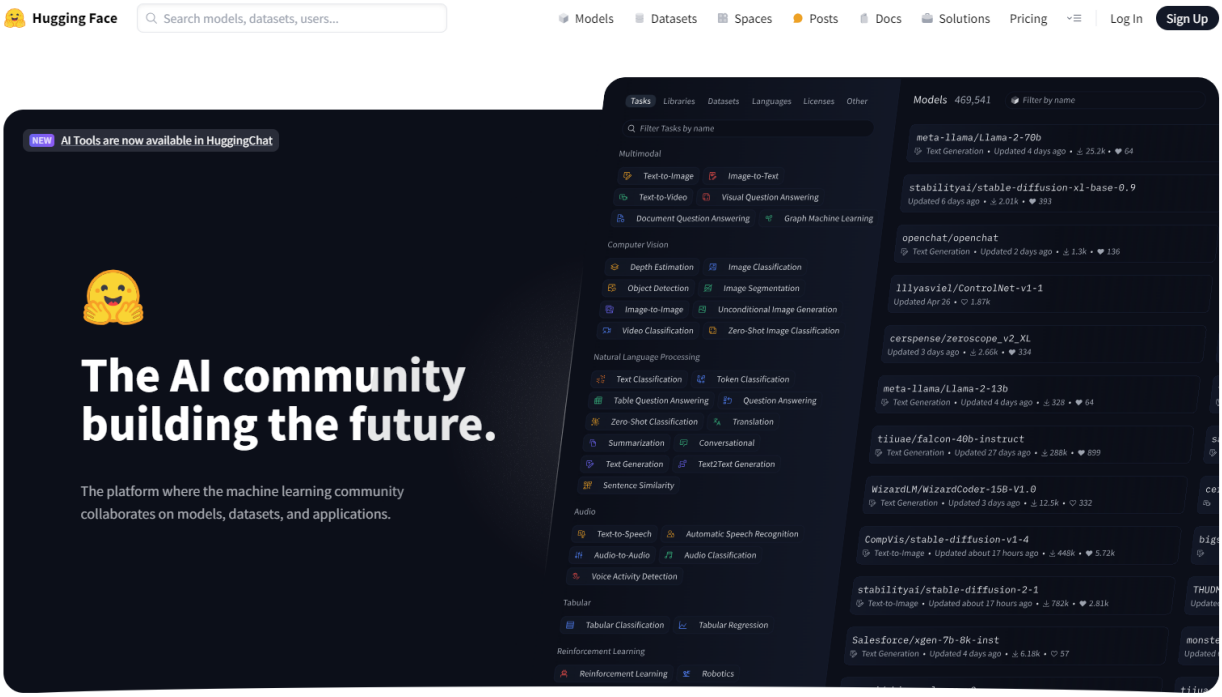
原因

从长远上考虑，后期会对大模型进行个性化调整，什么会使用自己的大模型，所有提前进行测试。

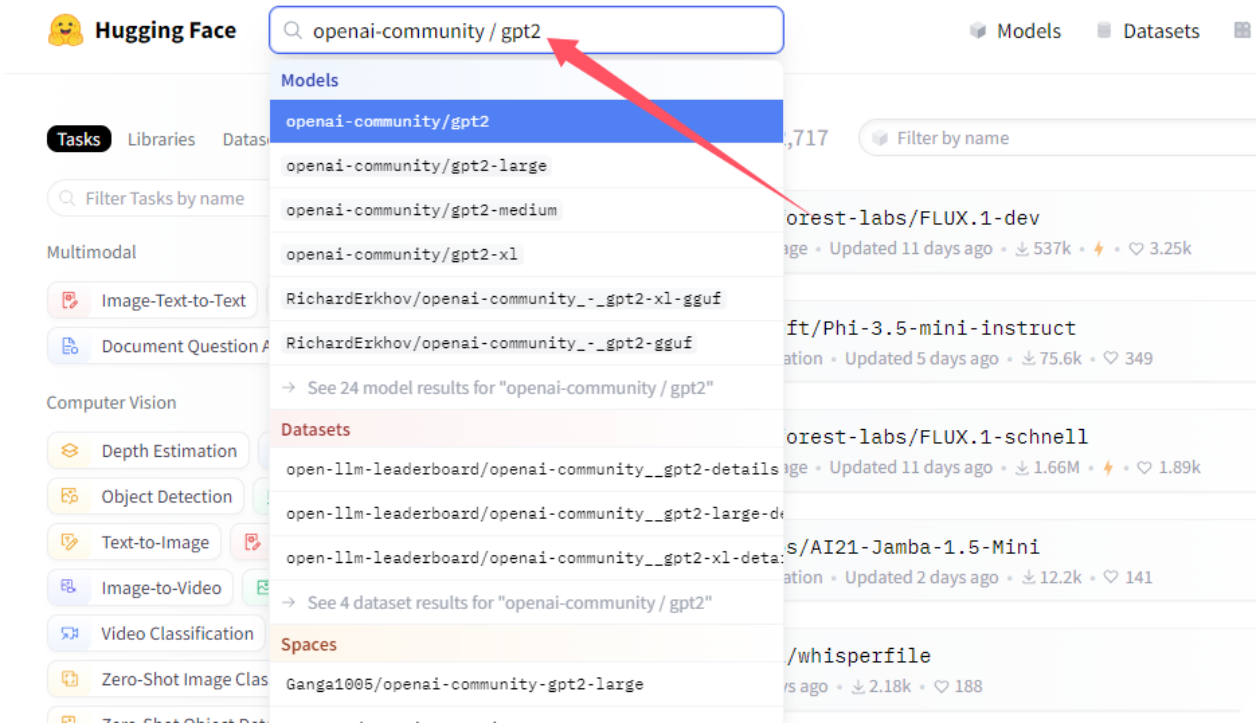
（使用阿里云上的模型不是免费的）

huggingface下载

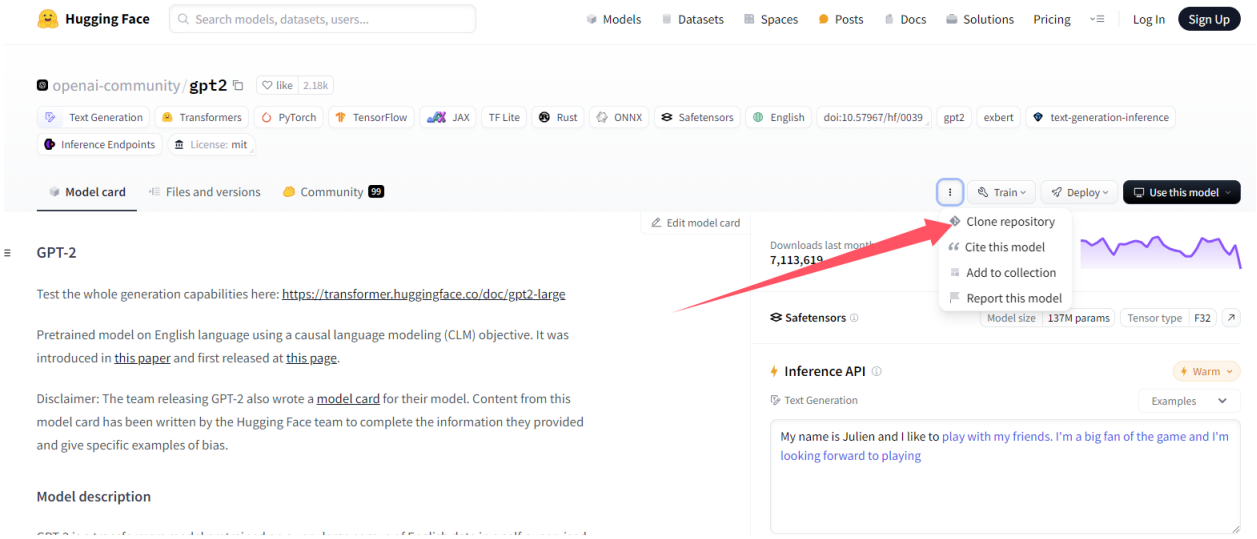
官网：[Hugging Face – The AI community building the future.](https://huggingface.co/)



点击上方的`Models`按钮，进入模型界面



输入想要搜索的模型，这里我用`openai-community / gpt2`这个模型举例子。

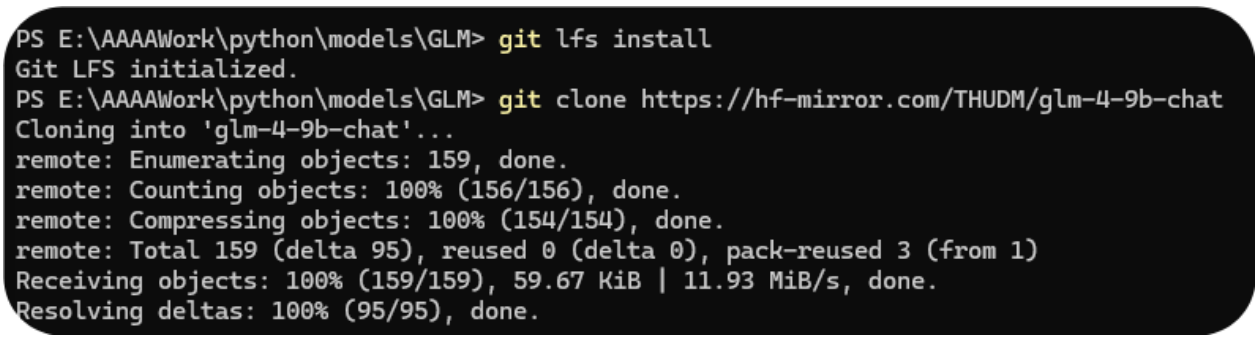


进入之后找到`Clone respository`，并点击

然后在你想要存放模型的文件夹中打开终端，输入前两个指令就可以

```
git lfs install
git clone https://huggingface.co/openai-community/gpt2
```

等待下载完成

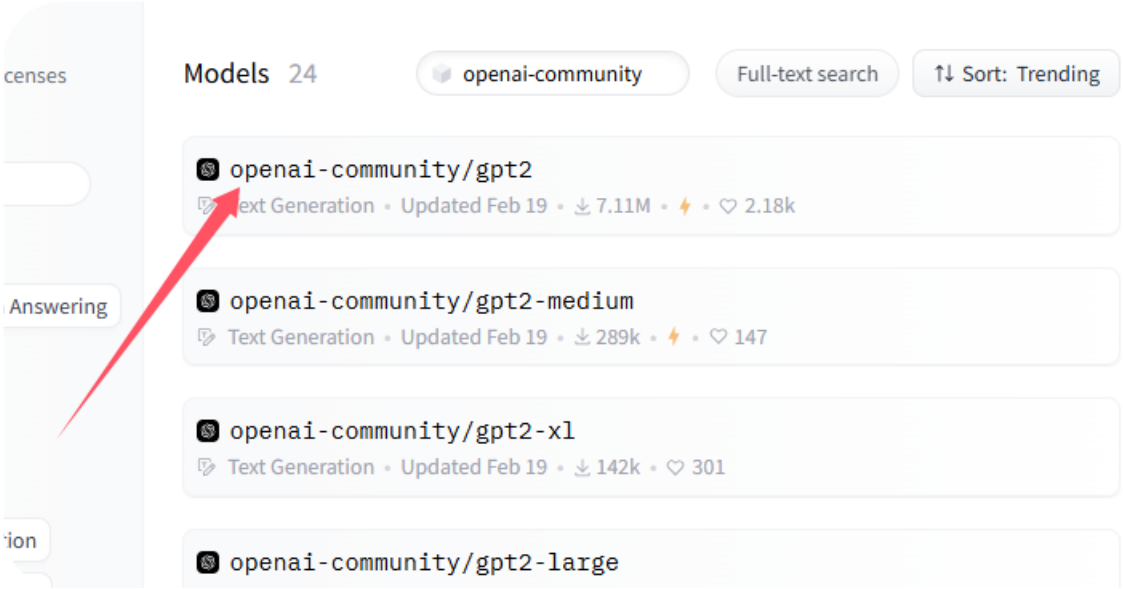


这个地方我使用的是国内huggingface镜像网站下载

注意！！！如果你即使挂了梯子有不能成功将模型拉去下来，那我建议使用国内镜像进行操作

国内huggingface镜像网站：https://hf-mirror.com/

直接搜索



点击进入，然后和上面一样的下载就ok了

使用本地模型

调用的库

```
from langchain_community.llms.huggingface_pipeline import HuggingFacePipeline
```

demo

```
# 本地下载的模型的地址
model_id = "your_path"

# 实例化模型
model = HuggingFacePipeline.from_model_id(
    model_id=model_id,
    task="text-generation", # 参数指定了要执行的任务类型。在这里，任务类型被设置为 "text-generation"，这意味着这个模型将被用于生成文本。
    pipeline_kwargs={"max_new_tokens": 100}, # {"max_new_tokens": 100} 表示在每次调用模型进行文本生成时，最多生成 100 个新令牌（tokens）。
    device=-1, # -1 表示在 CPU 上运行模型。如果将其设置为 0 或其他正数，则模型将在 GPU 上运行，具体取决于设备编号。
    )
```

测试

```
# 模型测试
model('你好') # gpt2 参数很少，训练的很一般，所有如果想要尝试，不建议下载这个模型
# 返回的内容就是一个字符串答案
```