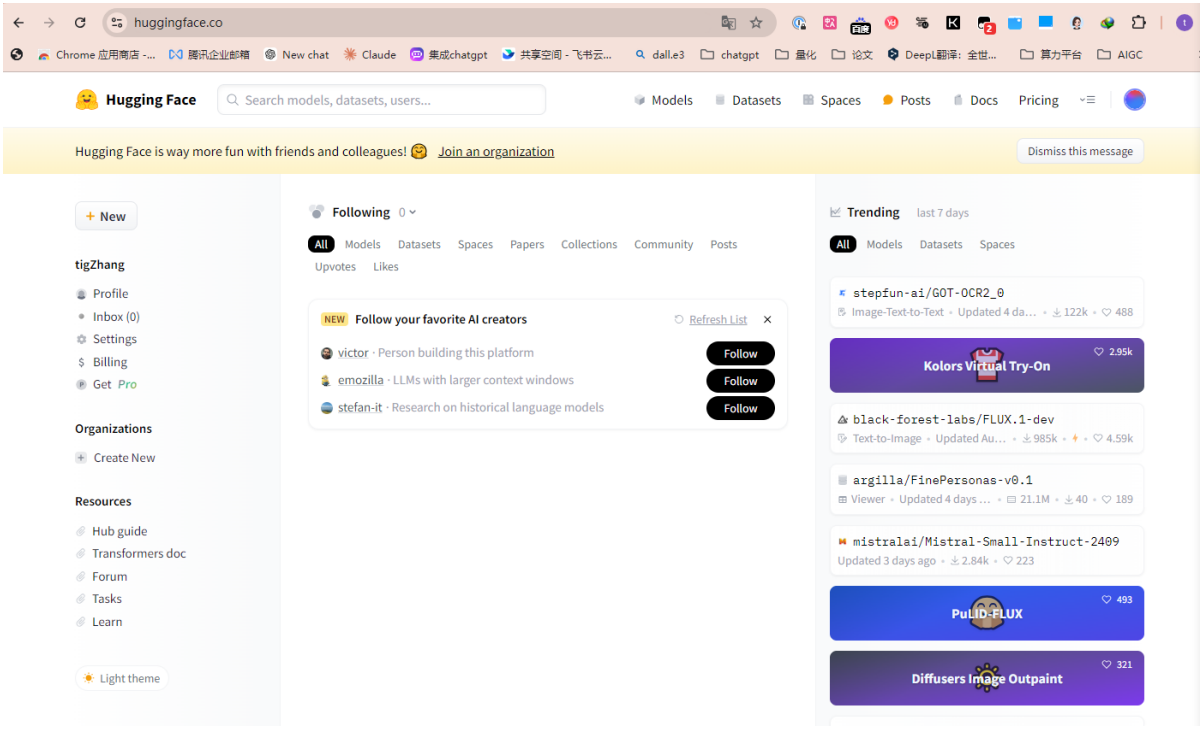
1. Hugging Face

HuggingFace是一家总部在纽约的自然语言处理公司，现在HuggingFace社区可以说是大模型研究中的最重要的开源社区了。

其实HuggingFace最早的定位是希望打造一个可以与人类进行有趣对话的聊天机器人，但是这个事情一直都不温不火，反而后面推出的Transformer库和HuggingFace社区大火起来。目前HuggingFace社区已经共享了超100,000个预训练模型，10,000个数据集，已经变成了LLM的github。

HuggingFace（https://huggingface.co/）有很多开源的NLP工具产品，具体包括：

1. Transformers库：提供了各种预训练的语言模型和NLP架构,如BERT、LLAMA、RoBERTa等。这些模型可以用于各种下游任务,如文本分类、问答系统、文本生成等。
2. Datasets库: 提供了一个标准的、易于使用的接口来访问和共享NLP数据集。它包含了许多常用的数据集,并支持自定义数据集的创建和共享。
3. Tokenizers: 一个高性能的文本标记化库,可以大大加快文本预处理的速度。
4. Accelerate: 可以帮助用户轻松地在不同的硬件(如CPU、GPU、TPU)上训练和微调模型,并支持分布式训练。



1.1. 镜像站

https://hf-mirror.com/

镜像站（Mirror Site）是指与主站点（即原始服务器）保持同步的复制服务器或网站。它的作用是将主站点上的数据（如文件、模型、数据集等）复制到另一个地理位置或服务器上，用户可以通过这个镜像站访问与主站点相同的数据资源。镜像站的核心目的是为了提高访问速度、减少网络延迟，以及分散服务器的负载，特别是在不同地理区域中，镜像站可以显著提升访问体验。

由于某些原因，很多时候我们直接从Huggingface下载模型会非常不稳定，因此就有了HF-Mirror。镜像站提供了快读且稳定的下载渠道。镜像站支持cli及hfd的下载方式。

`HF_ENDPOINT` 是 Huggingface 库中一个被官方支持的环境变量，设置后，这些库在访问 Huggingface Hub 时会使用你指定的 `HF_ENDPOINT` 作为主机名，来替换默认的 `huggingface.co` 域名。这样可以通过镜像站（如国内的加速站）来下载模型和数据集，而不需要修改任何 Python 代码。镜像站支持以下下载：

- `huggingface-cli`
- `snapshot_download`
- `from_pretrained`
- `hf_hub_download`
- `timm.create_model`

1.1.1. 镜像站设置环境变量

要使用镜像站下载，只需要设置环境变量即可。我们可以临时性设置环境变量，也可以把环境变量写入到配置文件中，持久设置。

Linux 及Mac OS

```
export HF_ENDPOINT="https://hf-mirror.com"
```

Linux 写入到`~/.bashrc`中：

```
echo 'export HF_ENDPOINT="https://hf-mirror.com"'~/.bashrc
```

Mac OS 写入到 `~/.zshrc` 中：

```
echo 'export HF_ENDPOINT="https://hf-mirror.com"'~/.zshrc
```

Windows Powershell

```
$env:HF_ENDPOINT = "https://hf-mirror.com"
```

写入到 `~\Documents\WindowsPowerShellMicrosoft.PowerShell_profile.ps1` 中：

```
Add-Content -Path $PROFILE -Value '$env:HF_ENDPOINT = "https://hf-mirror.com"'
```

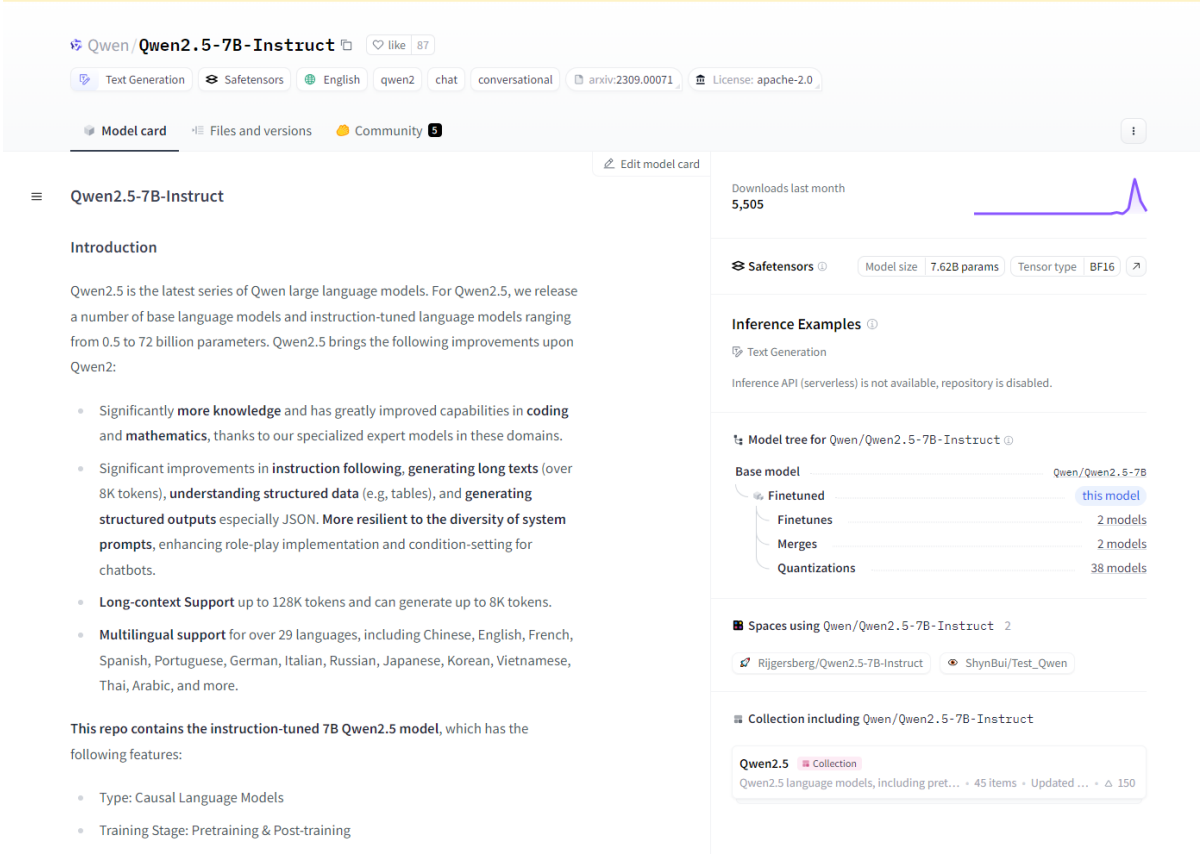
Python脚本

注意`os.environ`得在import huggingface库相关语句之前执行。

```
import os
os.environ['HF_ENDPOINT'] = 'https://hf-mirror.com'
```

1.2. 模型下载

我们可以在HuggingFace上搜索我们需要下载的模型。例如Qwen/Qwen2.5-7B-Instruct

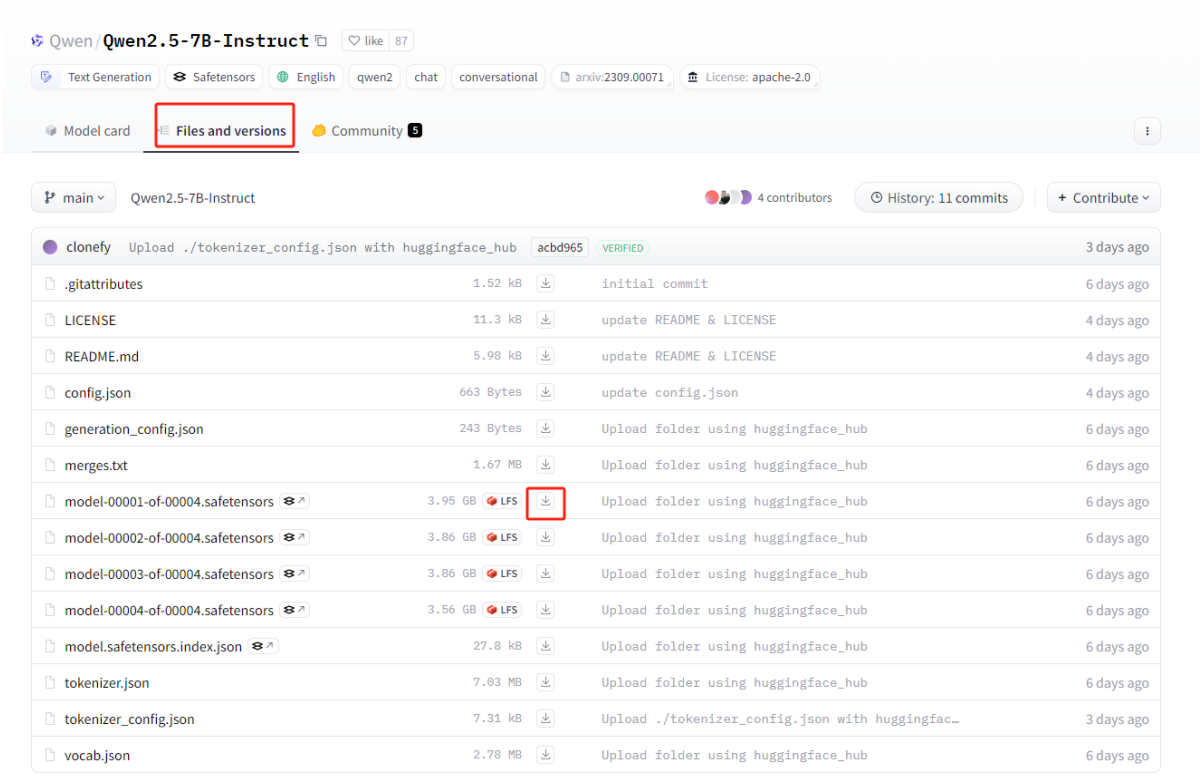


HuggingFace的Model card会介绍模型的有关信息，而Files and version则提供了模型的下阿紫。

要在HuggingFace上下载模型，包含了多种方式，既可以通过直接网页下载，也可以通过下载工具，还可以通过镜像站的下载方式。

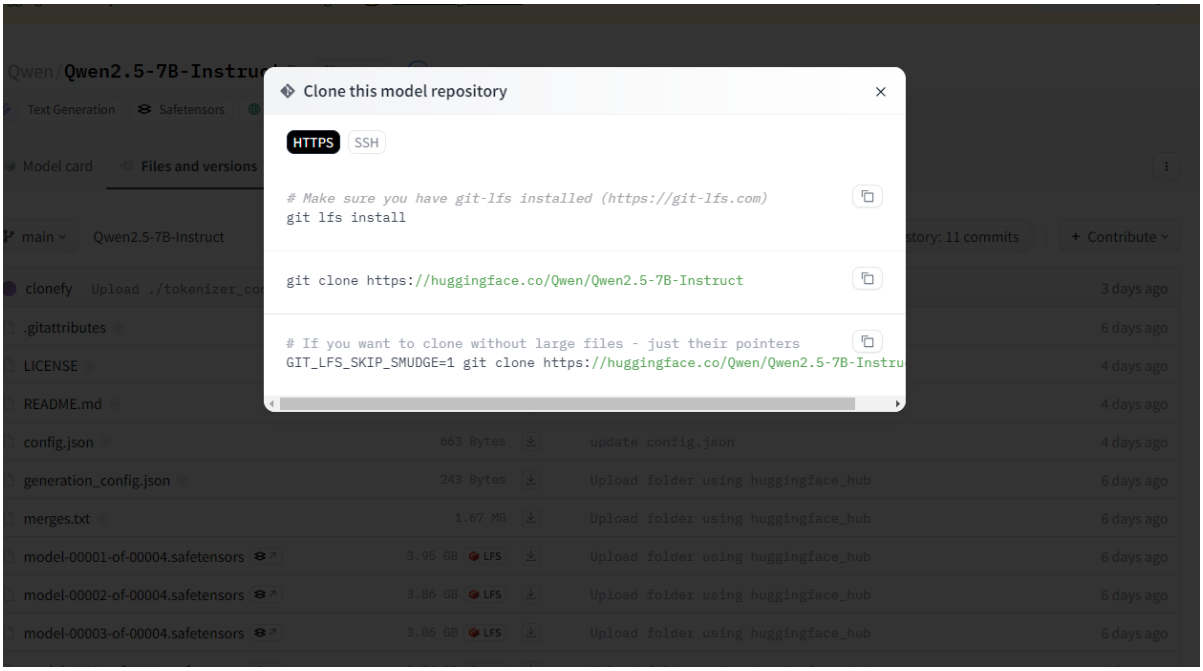
1.2.1. 网页下载

最直接的下载方式，在模型的文件中直接下载选中的文件。这种方式不依赖任何的工具，是最直接的下载方式。如果只需要下载一些小文件还是比较方便的。



1.2.2. Git clone

HuggingFace提供了git clone下载方式，和平时和git下载方式类似，同样比较简单。但值得注意的是这里的git clone由于不支持断点续传，因此中途断了，需要重头下载。此外clone下载会占用更大的磁盘空间（尤其是对于有历史版本的模型来说），所以一般不太推荐这种下载方式，



1.2.3. huggingface-cli

huggingface-cli隶属于huggingface_hub 库，属于官方工具，支持比较好，一般情况下建议使用CLI的下载方式。我们可以使huggingface-cli下载模型以及下载数据集

要使用huggingface-cli，需要先安装 huggingface_hub库：

```
pip install -U huggingface_hub
```

为了获得稳定的下载，使用cli下载时建议替换镜像下载，设置环境变量：

```
export HF_ENDPOINT="https://hf-mirror.com"
```

1.2.3.1. cli下载模型

下载模型：`Qwen/Qwen2.5-0.5B-Instruct`

```
huggingface-cli download --resume-download Qwen/Qwen2.5-0.5B-Instruct --local-dir /mnt/diskb5/zhangbo_private/llm_model/qwen/Qwen2___5-0.5B-Instruct
```

- 这里的模型卡是`Qwen/Qwen2.5-0.5B-Instruct`，指向 Hugging Face Hub 上由用户 `Qwen` 所上传的名为 `Qwen2.5-0.5B-Instruct` 的模型。
- `huggingface-cli download`：这是调用 `huggingface-cli` 工具的 `download` 子命令。此命令的主要功能是从 Hugging Face Hub 下载指定的模型或数据。
- `--resume-download`：这个选项允许下载过程中断后能够继续（而不是中断后再从头开始）。这对于下载大文件时非常有用。
- `--local-dir /mnt/diskb5/zhangbo_private/llm_model/qwen/Qwen2___5-0.5B-Instruct`：这个选项指定了下载文件的本地存储目录。这里使用的是一个绝对路径。这里我们建议建立相应的目录`qwen/Qwen2___5-0.5B-Instruct`。因此cli下载如果指定路径的时候，是直接把所有文件下到你的指定目录

1.2.3.2. cli下载数据集

```
huggingface-cli download --resume-download --repo-type dataset lavita/medical-qa-shared-task-v1-toy
```

- `--repo-type dataset`：这个选项明确指定要下载的资源类型为数据集。`lavita/medical-qa-shared-task-v1-toy`为数据集标识。

1.2.3.3. cli下载的符号链接注意事项

cli下载中，有一个`--local-dir-use-symlinks False`参数可选。

huggingface的工具链默认会使用符号链接来存储下载的文件，因此事实上`--local-dir`指定的目录中都是一些“链接文件”，默认情况下，模型存储在`~/.cache/huggingface`下，如果不喜欢这个可以用`--local-dir-use-symlinks False`取消这个逻辑。

`.cache/huggingface/`会维护一份符号链接，当我们调用模型的时候，我们可以使用模型的绝对路径（model_path = '/mnt/diskb5/zhangbo_private/llm_model/qwen/Qwen2___5-7B-Instruct'），或者模型的标识符加载模型（model_path = 'Qwen/Qwen2.5-1.5B'）。这样的话，无论我们模型下载到什么地方，我们都可以通过符号链接调用，而且避免了忘记下载过某个模型而重新下载。

但是这样做，有个不好的地方就是，不熟悉huggingface的同学，往往下载了模型，但是不知道在哪里找到模型文件

1.2.4. hfd下载

hfd是一种多线性下载工具。是HuggingfaceFace镜像站https://hf-mirror.com/开发的专用下载工具，基于工具`git+aria2`实现稳定下载。

1. 下载hfd

```
wget https://hf-mirror.com/hfd/hfd.sh
chmod a+x hfd.sh
```

2. 安装Aria2（如果没有安装 aria2，则可以默认用 wget：）

```
sudo apt-get install aria2
```

3.设置环境变量
Linux

```
export HF_ENDPOINT="https://hf-mirror.com"
```

Windows Powershell

```
$env:HF_ENDPOINT = "https://hf-mirror.com"
```

4.1 下载模型

```
./hfd.sh Qwen/Qwen2.5-0.5B --tool aria2c -x 4
```

下载的路径是放在当前目录下，新建一个文件夹Qwen2.5-0.5B

4.2 下载数据集

```
./hfd.sh wikitext --dataset --tool aria2c -x 4
```

使用hfd比较直观，多线程并直观看到进度

hfd的完整参数：

```
Usage:
hfd <model_id> [--include include_pattern] [--exclude exclude_pattern] [--
hf_username username] [--hf_token token] [--tool wget|aria2c] [-x threads] [--
dataset]
Description:
使用提供的模型ID从Hugging Face下载模型或数据集。
Parameters:
model_id Hugging Face模型ID，格式为'repo/model_name'。
--include （可选）标志，用于指定要包括在下载中的文件的字符串模式。
--exclude （可选）标志，用于指定要从下载中排除的文件的字符串模式。
exclude_pattern 匹配文件名以排除的模式。
--hf_username （可选）Hugging Face用户名，用于身份验证。
--hf_token （可选）Hugging Face令牌，用于身份验证。
--tool （可选）使用的下载工具。可以是wget（默认）或aria2c。
-x （可选）aria2c的下载线程数。
--dataset （可选）标志，表示下载数据集。
示例：
hfd bigscience/bloom-560m --exclude safetensors
hfd meta-llama/Llama-2-7b --hf_username myuser --hf_token mytoken --tool aria2c -
x 8
hfd lavita/medical-qa-shared-task-v1-toy --dataset
```

1.2.5. snapshot_download

`huggingface_hub` 的 `snapshot_download` 是用于从 Hugging Face 模型库下载特定模型的多功能下载方式。`snapshot_download` 提供包括下载指定版本，使用代理，多线程下载，是否自动恢复下载等多种功能。

```
#下载功能解析
snapshot_download(
    repo_id: str,                    # 仓库ID，如 "bert-base-
uncased"、"facebook/wav2vec2-large-xlsr-53" 等
    revision: str = None,            # 版本号（commit hash、分支名或tag），默认下载最新
版本
```

```
    local_dir: str = None,            # 模型的本地存储目录
    cache_dir: str = None,            # 模型的缓存目录（可选）
    proxies: dict = None,             # 代理设置（如 {"https":
"http://localhost:8080"}）
    resume_download: bool = False,    # 如果下载失败是否自动恢复
    max_workers: int = 1,             # 并发下载线程数，默认为 1
    allow_patterns: list = None,      # 允许下载的文件类型，如 ["*.bin", "*.json"]
    ignore_patterns: list = None,     # 忽略下载的文件类型
    token: str = None,                # Hugging Face 访问令牌（对于私有模型）
    force_download: bool = False      # 是否强制重新下载模型
)
```

示例：

```
import os
os.environ['HF_ENDPOINT'] = 'https://hf-mirror.com'

from huggingface_hub import snapshot_download
snapshot_download(repo_id="Qwen/Qwen2.5-
0.5B",local_dir="/mnt/diskb5/zhangbo_private/llm_model/qwen/Qwen2___5-0.5B",
                  max_workers=8)
```

1.2.6. from_pretrained

from_pretrained是下载模型的常见方式了，用于从 Hugging Face Model Hub 下载预训练模型并加载到内存中。如果模型已经下载了，则直接加载模型。而如果模型还没下载，则先下载再加载。

```
from transformers import AutoModelForCausalLM, AutoTokenizer


def load_model(model_path):
    device = "cuda"  # 将模型加载到 GPU 上
    model = AutoModelForCausalLM.from_pretrained(
        model_path,
        torch_dtype="auto",
        device_map="auto"
    )
    tokenizer = AutoTokenizer.from_pretrained(model_path)
    return device, tokenizer, model

def chat_qwen(device, tokenizer, model, prompt):
    messages = (
        {"role": "system", "content": "You are a helpful assistant."},
        {"role": "user", "content": prompt}
    )
    text = tokenizer.apply_chat_template(messages, tokenize=False,
add_generation_prompt=True)
    model_inputs = tokenizer((text), return_tensors="pt").to(device)
```

```
    generated_ids = model.generate(
        model_inputs.input_ids,
        max_new_tokens=512
    )
    generated_ids = [output_ids[len(input_ids):] for input_ids, output_ids in
zip(model_inputs.input_ids, generated_ids)]
    response = tokenizer.batch_decode(generated_ids, skip_special_tokens=True)[0]
    print(response)

# 载入模型
# model_path = '/mnt/diskb5/zhangbo_private/llm_model/qwen/Qwen2___5-7B-Instruct'
model_path = 'Qwen/Qwen2.5-1.5B'
device, tokenizer, model = load_model(model_path)

# 进行测试
chat_qwen(device, tokenizer, model,'你好，请你介绍一下自己')
```

1.2.7. 模型下载总结

包含镜像站的话，本课程一共介绍7种下载方式

方法	推荐建议	优点	缺点
网页下载	⭐⭐⭐	简单直接	多文件下载麻烦
git clone	⭐⭐	简单	无断点，无多线程，下载速度慢，占用空间多
huggingface cli	⭐⭐⭐⭐	官方工具，简单，支持断点	无多线程
hfd	⭐⭐⭐	支持多线程，支持断点，功能多	配置稍微有点麻烦
snapshot download	⭐⭐⭐⭐	官方工具功能强大，包括代理，多线程，断点均支持	需要简单学习一下参数的使用
from pretrained	⭐⭐⭐	官方工具，功能少，不好管理	简单
镜像站	⭐⭐⭐⭐⭐	建议上述的cli，hfd，snapshot download以及from pretrained都配合镜像站使用	多一行命令，换来速度，你还想怎样?

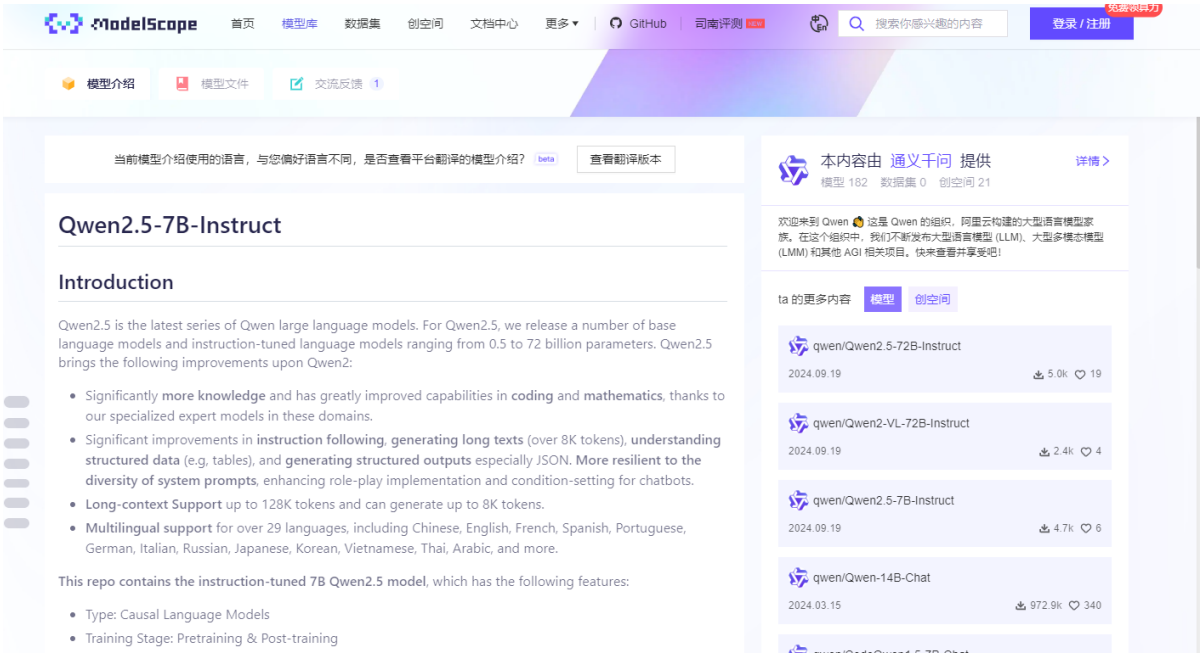
2. ModelScope

ModelScope它是由阿里巴巴通义实验室，联合CCF开源发展委员会，共同作为项目发起创建，提供了一个覆盖多个领域的强大模型库,让开发者能够轻松发现、训练和使用AI模型,可以理解为国产的HuggingFace，

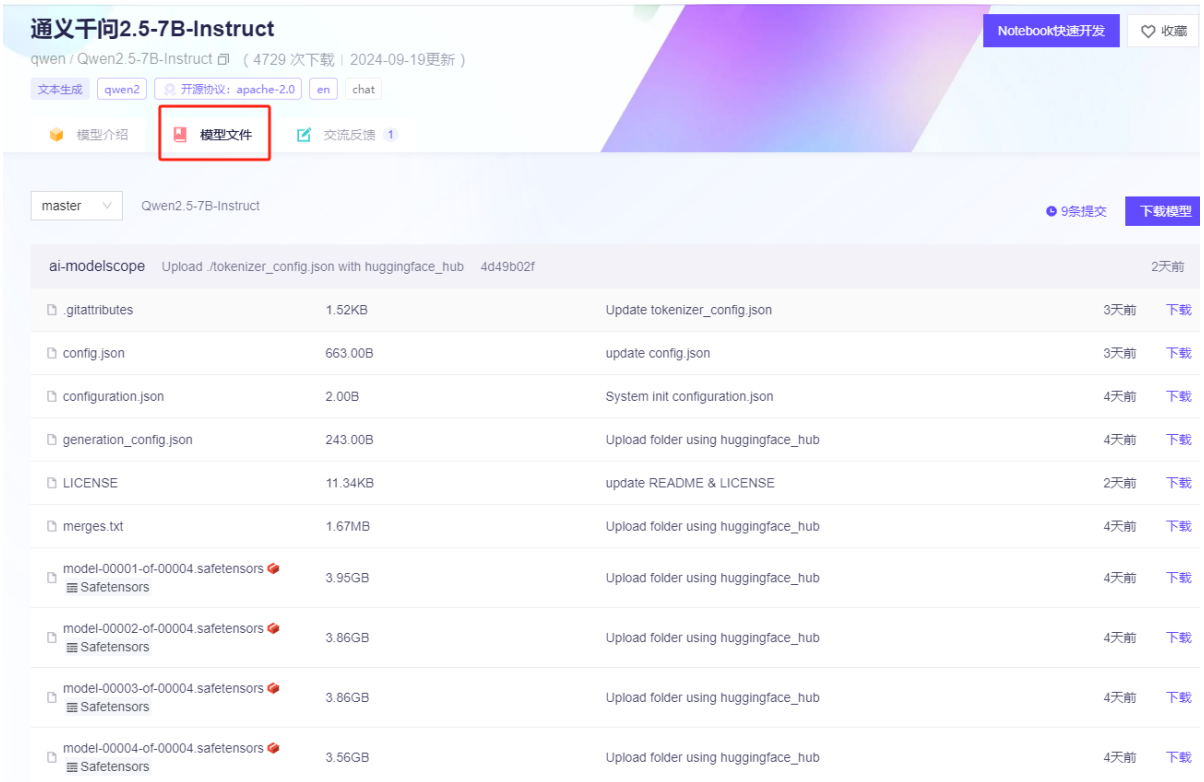
2.1. 模型下载

要在ModelScope中下载模型，可以在ModelScope中查询我们要下载的模型，例如搜索Qwen2.5。

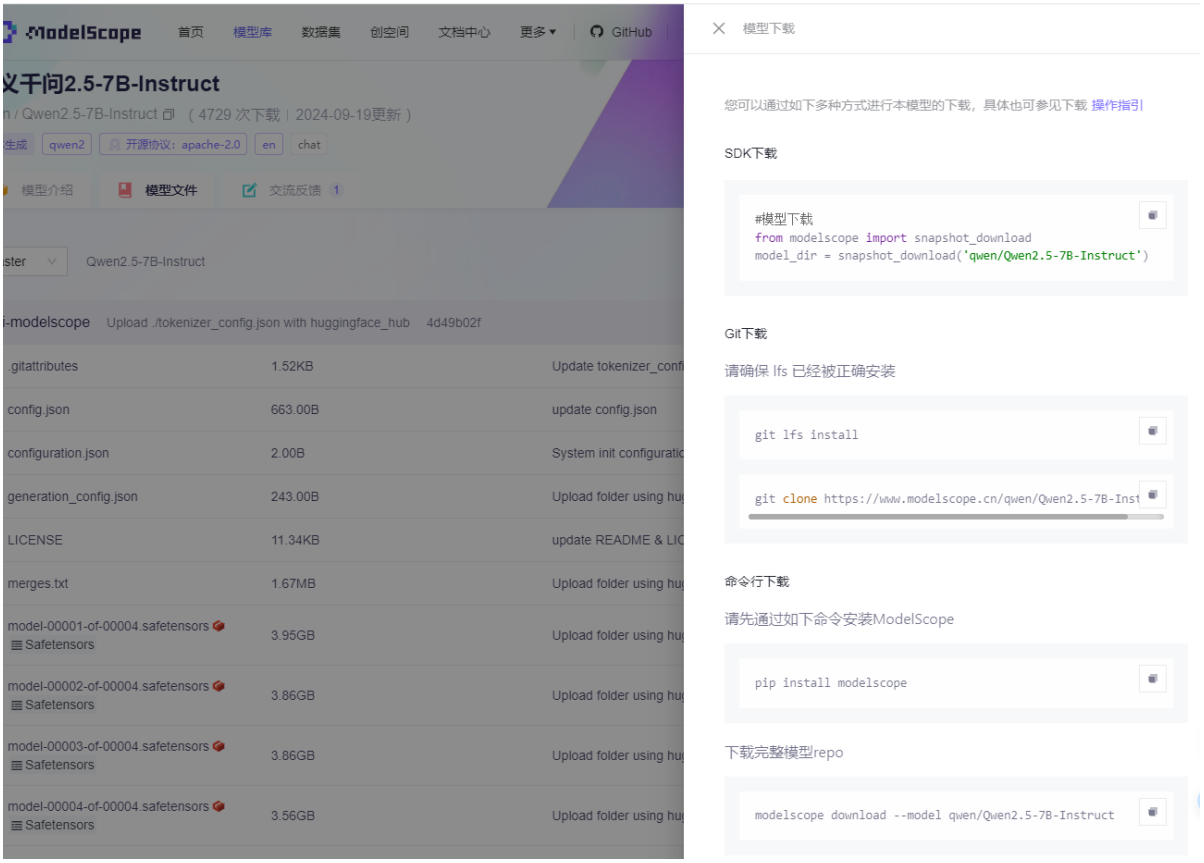
假设我们想下载"通义千问2.5-7B-Instruct"，有链接：https://www.modelscope.cn/models/qwen/Qwen2.5-7B-Instruct/files



如果你想下载模型，可以到导航到模型文件



这个时候，我们可以逐个文件下载。当然也有更加方便的方式。点击右上角的下载模型，modelscope提供了包括sdk，github等多种下载方式。

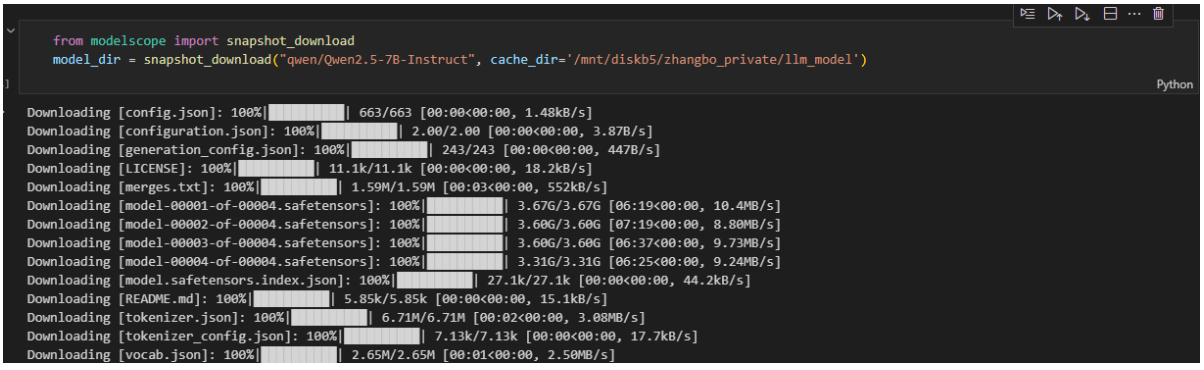


2.1.1. SDK下载

```
from modelscope import snapshot_download
model_dir = snapshot_download("qwen/Qwen2.5-7B-Instruct",
cache_dir='/mnt/diskb5/zhangbo_private/llm_model')
```

会下载在脚本所在文件夹的目录；如果没有cache_dir=''，则会下载在/home/create/.cache/modelscope/hub

如果使用默认下载的话，加载的话可直接引用模型名称就可以了，例如直接qwen/Qwen2.5-7B-Instruct。如果提示找不到路径，加载的话，可以切换回绝对路径。



2.1.2. Git下载

使用git下载，请确保已经安装了lfs

```
git lfs install
git clone https://www.modelscope.cn/qwen/Qwen2.5-7B-Instruct.git
```