

大语言模型如何下载

1. modelscope使用说明

ModelScope是阿里巴巴推出的开源模型即服务（MaaS）平台，它集成了众多领先的预训练模型，旨在为AI开发者提供一站式、易用且低成本的模型服务解决方案。

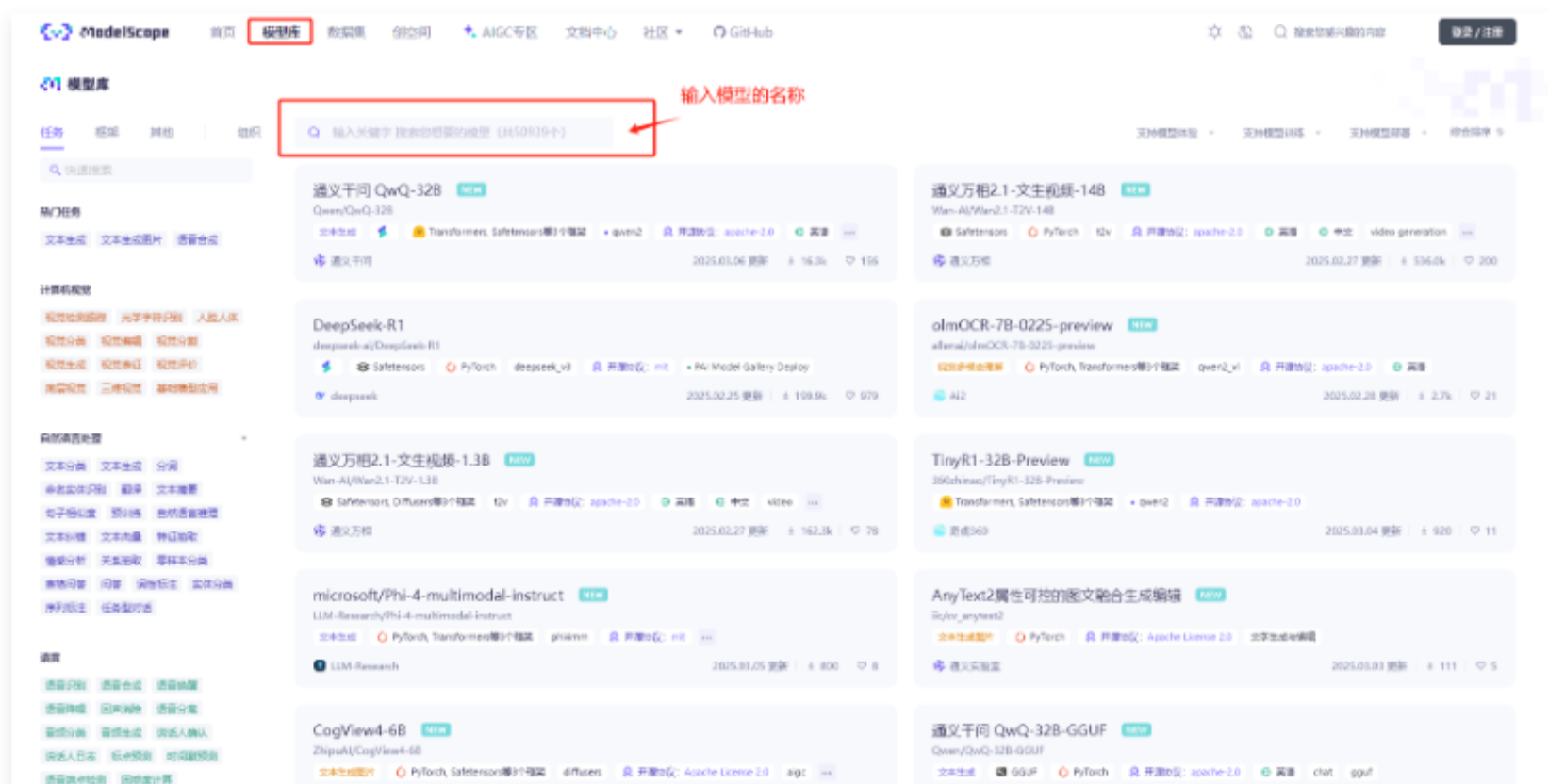
ModelScope的主要功能包括以下几个方面：

- 模型库：**ModelScope汇集了各领域最先进最新的机器学习模型，包括但不限于自然语言处理、计算机视觉、语音识别等。这些模型已经过训练和优化，开发者可以直接使用，无需从头开始训练，从而节省了大量时间和资源。
- 数据集：**平台提供了方便共享及访问的数据集合，可用于算法训练、测试、验证。这些数据集通常以表格形式出现，按照模态可划分为文本、图像、音频、视频、多模态等，为开发者提供了丰富的数据资源，有助于提升模型的性能和准确性。
- 模型管理与优化：**ModelScope支持模型的版本控制、缓存管理等功能，使开发者能够高效地管理自己的模型。此外，平台还提供了模型性能的监控与分析工具，帮助开发者了解模型权重的分布情况以及模型的预测准确度等关键指标。通过这些信息，开发者可以对模型进行调优和优化，提升模型的效率和准确性。
- 模型应用可视化私域空间：**ModelScope的创空间（Studio）为开发者提供了一个模型应用可视化的私域空间。在这里，开发者可以基于平台上的模型原子能力，自行搭建与展示不同的AI应用，包括自定义的模型输入输出、多模型的组合以及可视化交互展现形式等。这为开发者提供了一个灵活且强大的工具，用于探索和实现各种创新的AI应用。
- 开源与共享：**作为一个开源项目，ModelScope鼓励开发者共享自己的模型和代码，促进社区内的交流与合作。通过参与ModelScope社区，开发者可以获取最新的AI技术动态，学习他人的优秀实践，并共同推动AI技术的发展。

- <https://www.modelscope.cn/home>

(1) 模型下载

- <https://www.modelscope.cn/models>



ai modelscope Update README.md 39693e97			0更新
attributes	1.52KB	Update config.json	9个月前 下载
config.json	663.00B	Update config.json	9个月前 下载
configuration.json	40.00B	System init configuration.json	9个月前 下载
generation_config.json	243.00B	upload qwen2	9个月前 下载
LICENSE	11.34KB	Update config.json	9个月前 下载
merges.txt	1.67MB	upload qwen2	9个月前 下载
model-00001-of-00004.safetensors	3.95GB	upload qwen2	9个月前 下载
model-00002-of-00004.safetensors	3.86GB	upload qwen2	9个月前 下载
model-00003-of-00004.safetensors	3.86GB	upload qwen2	9个月前 下载
model-00004-of-00004.safetensors	3.56GB	upload qwen2	9个月前 下载
model.safetensors.index.json	27.75KB	upload qwen2	9个月前 下载

模型下载

我们推荐使用命令行或者 ModelScope SDK 来进行模型的下载。 [操作指引](#)

在下载前，请先通过如下命令安装ModelScope

```
pip install modelscope
```

命令行下载

下载完整模型库

```
modelscope download --model qwen/Qwen2-7B-Instruct
```

下载单个文件到指定本地文件夹（以下载README.md到当前路径下“dir”目录为例）

```
modelscope download --model qwen/Qwen2-7B-Instruct README.md
```

更多更丰富的命令行下载选项，可参见 [具体文档](#)

SDK下载

```
#模型下载
from modelscope import snapshot_download
model_dir = snapshot_download('qwen/Qwen2-7B-Instruct')
```

Git下载

请确保 lfs 已经被正确安装

建议下载到指定路径, 这个路径为 `model_path`

调用

```
In [8]: from modelscope import AutoModelForCausalLM, AutoTokenizer, GenerationConfig
```

```
In [9]: model_path = './data/llm_app/llm/Qwen2-7B-Instruct/'
        model = AutoModelForCausalLM.from_pretrained(model_path,
                                                    device_map="auto")
        tokenizer = AutoTokenizer.from_pretrained(model_path)
```

```
gen_config = GenerationConfig.from_pretrained(model_path)
```

```
# pip install modelscope
```

```
conda activate llm
```

```
modelscope download --model qwen/Qwen2-7B-Instruct README.md --local_dir ./data/llm_app/llm/Qwen2-7B-Instruct
```

Qwen2-7B-Instruct目录里所有文件都下载

config.json	663.00B
configuration.json	48.00B
generation_config.json	243.00B
LICENSE	11.34KB
merges.txt	1.67MB
model-00001-of-00004.safetensors	3.95GB
model-00002-of-00004.safetensors	3.86GB
model-00003-of-00004.safetensors	3.86GB
model-00004-of-00004.safetensors	3.56GB

model.safetensors.index.json	27.75KB
README.md	6.58KB
tokenizer.json	7.03MB
tokenizer_config.json	1.29KB
vocab.json	2.78MB

有git环境的建议采用git来下载

```
# git 下载的路径 为 程序中的model_path
# linux环境可以centos: yum install git
# windows 环境可以下载git window客户端, 执行git bash
git lfs install
git lfs clone https://www.modelscope.cn/qwen/Qwen2-7B-Instruct.git
```



需要等等下载结束, 等光标结束 (下面的状态为正在下载)

```
MINGW64:/d
@DESKTOP-77V8KSU MINGW64 ~
$ cd /d/

@DESKTOP-77V8KSU MINGW64 /d
$ git lfs install
Git LFS initialized.

@DESKTOP-77V8KSU MINGW64 /d
$ git lfs clone https://www.modelscope.cn/qwen/Qwen2-7B-Instruct.git
WARNING: `git lfs clone` is deprecated and will not be updated
with new flags from `git clone`

`git clone` has been updated in upstream Git to have comparable
speeds to `git lfs clone`.
Cloning into 'Qwen2-7B-Instruct'...
remote: Enumerating objects: 39, done.
remote: Counting objects: 100% (39/39), done.
remote: Compressing objects: 100% (24/24), done.
remote: Total 39 (delta 13), reused 39 (delta 13), pack-reused 0
Receiving objects: 100% (39/39), 3.60 MiB | 916.00 KiB/s, done.
Resolving deltas: 100% (13/13), done.
Downloading LFS objects: 0% (0/4), 3.4 MB | 686 KB/s
```

此时 `model_path=r'D://Qwen2-7B-Instruct'`

(2) 课程中使用到的大语言模型下载

- Qwen2-7B-Instruct: <https://www.modelscope.cn/models/qwen/Qwen2-7B-Instruct/files>
- chatglm3-6b-32k: <https://www.modelscope.cn/models/ZhipuAI/chatglm3-6b-32k/files>

其他的模型可以自行在modelscope上下载, 如qwen2.5 和 deepseek r1 等

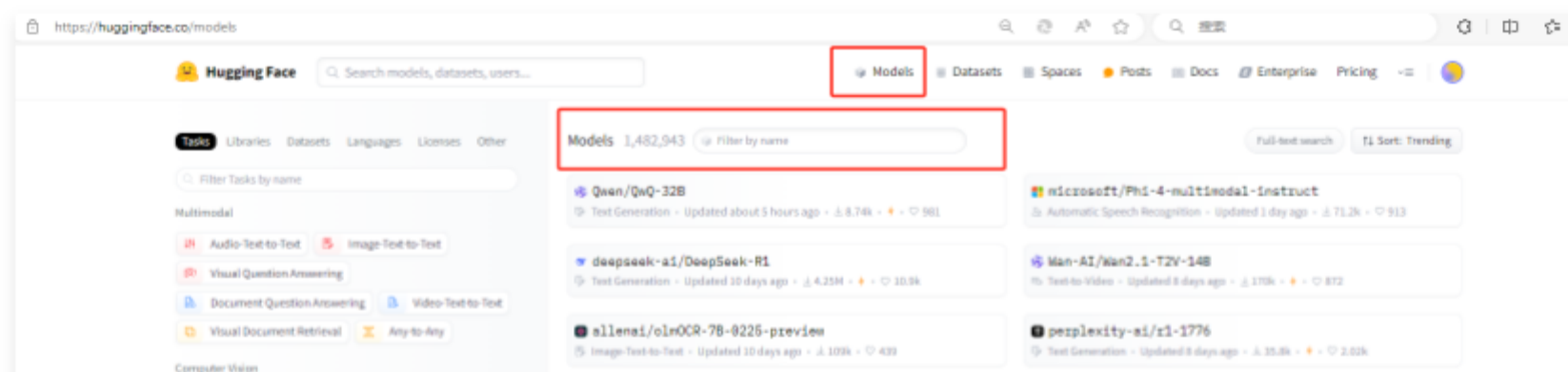
qwen2.5 72b:

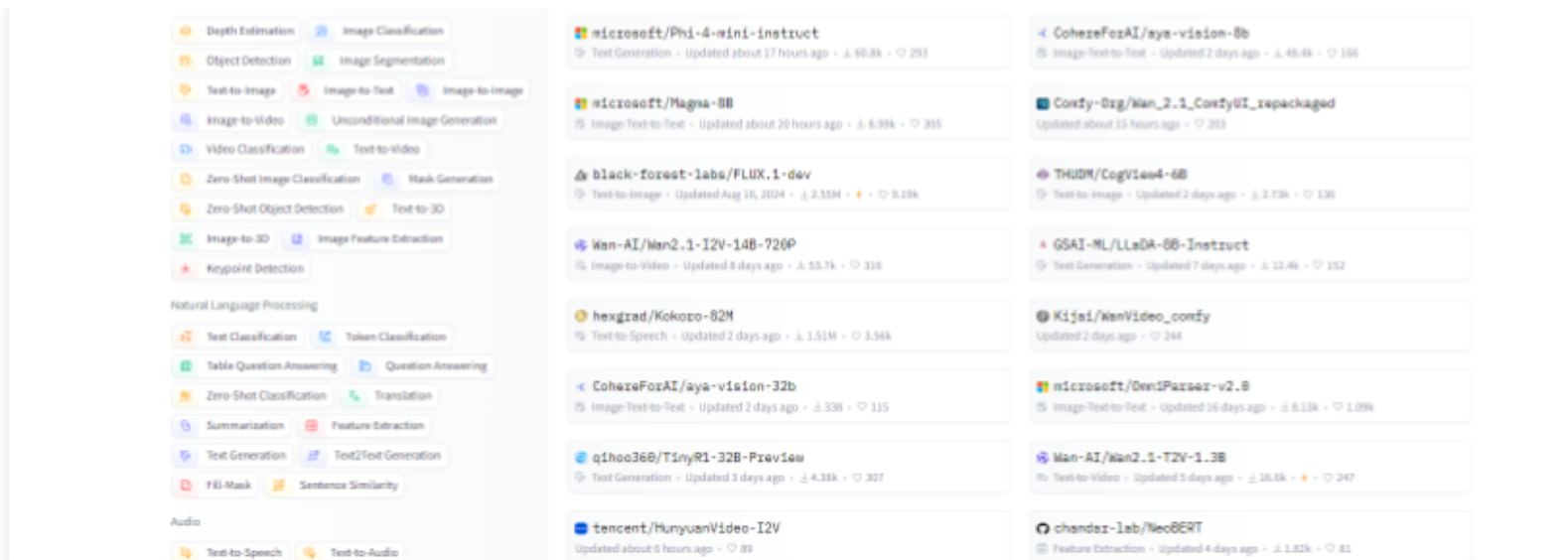
- <https://www.modelscope.cn/models/Qwen/Qwen2.5-7B-Instruct>
- <https://www.modelscope.cn/models/Qwen/Qwen2.5-72B-Instruct>

deepseek r1 (deepseek r1 建议采用ollama和vllm部署) :

- <https://www.modelscope.cn/models/deepseek-ai/DeepSeek-R1-Distill-Qwen-7B>
- <https://www.modelscope.cn/models/deepseek-ai/DeepSeek-R1-Distill-Qwen-32B>

2. huggingface



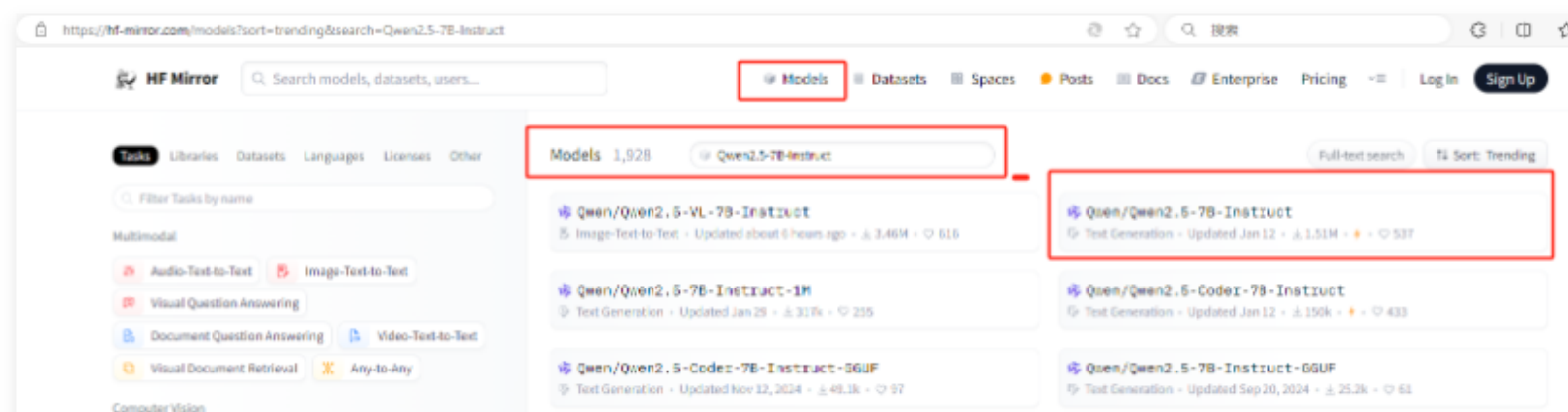


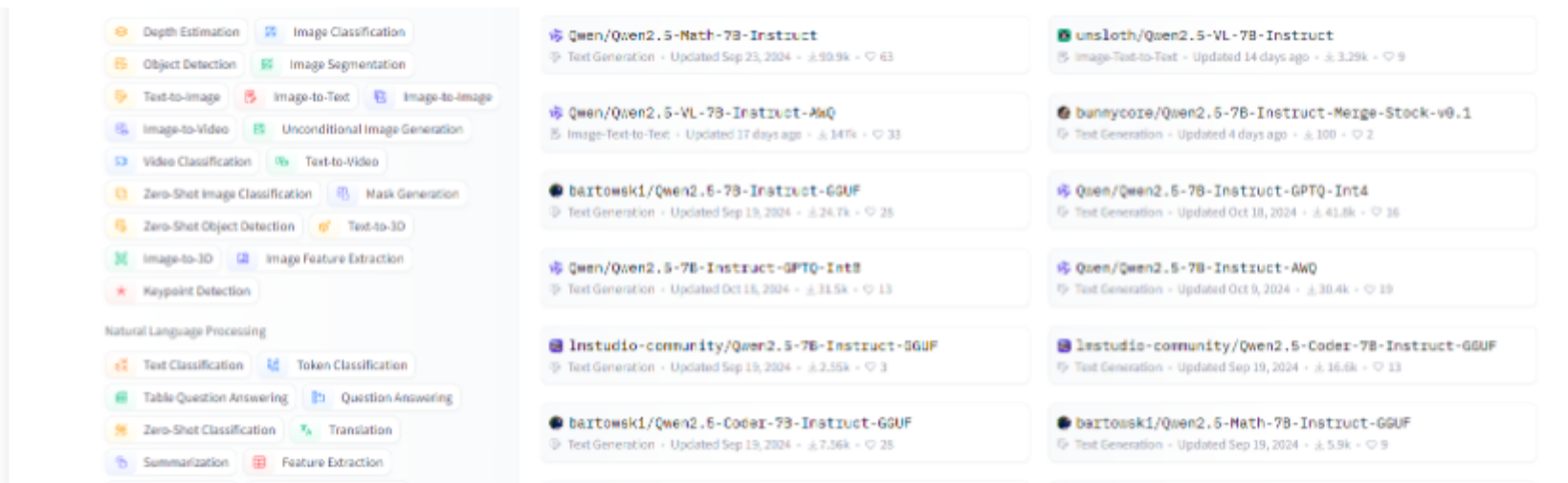
- <https://huggingface.co/models>

Hugging Face是一个机器学习（ML）和数据科学平台及社区，和modelscope功能类型，但目前国内无法访问，可以通过多镜像站来下载模型和数据集。

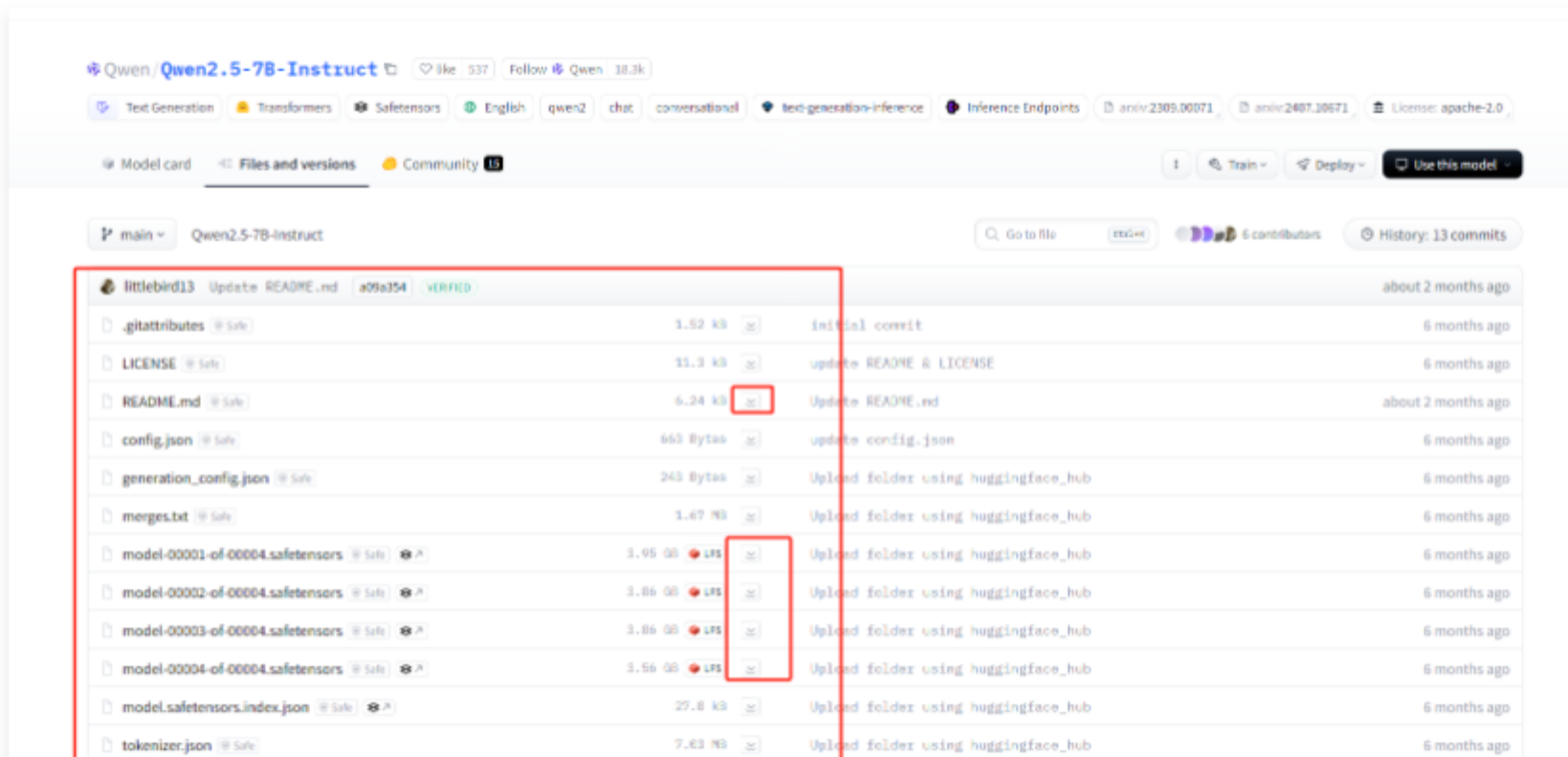
镜像站为：


- <https://hf-mirror.com/models>





网页直接下载



 tokenizer_config.json @ Safe	7.31 KB		Upload ./tokenizer_config.json with huggingface_hub	6 months ago
 vocab.json @ Safe	2.78 MB		Upload folder using huggingface_hub	6 months ago

huggingface cli工具

方法二: huggingface-cli

`huggingface-cli` 是 Hugging Face 官方提供的命令行工具, 自带完善的下载功能。

1. 安装依赖

```
pip install -U huggingface_hub
```

[Copy](#)

2. 设置环境变量

Linux

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

[Copy](#)

Windows Powershell

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

[Copy](#)

建议将上面这一行写入 `~/.bashrc`。

3.1 下载模型

```
huggingface-cli download --resume-download gpt2 --local-dir gpt2
```

[Copy](#)

3.2 下载数据集

```
huggingface-cli download --repo-type dataset --resume-download wikitext --local-dir wikitext
```

[Copy](#)

可以添加 `--local-dir-use-symlinks False` 参数禁用文件软链接, 这样下载路径下所见即所得, 详细解释请见上面提到的教程。

```
conda activate llm
```

```
pip install -U huggingface_hub
```

```
# Linux
```

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

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

# model_path 为 Qwen2.5-7B-Instruct
huggingface-cli download --resume-download Qwen2.5-7B-Instruct --local-dir Qwen2.5-7B-Instruct
```

课程中使用到的大语言模型下载

- Qwen2-7B-Instruct: <https://hf-mirror.com/Qwen/Qwen2-7B-Instruct/tree/main>
- chatglm3-6b-32k: <https://hf-mirror.com/THUDM/chatglm3-6b-32k>

这里的模型和modelscope的模型权重都是一样，没有必要重复下载，国内建议在modelscope下载

3. 模型的使用

上面1和2 下载的都是模型的配合和权重参数，比如你下载的路径为 `D://Qwen2-7B-Instruct`

再次强调：将这个路径替换掉代码中 `model_path` 即可，使用采用transformers和modelscope调用大语言模型的方式。

调用

```
In [8]: from modelscope import AutoModelForCausalLM, AutoTokenizer, GenerationConfig

In [9]: model_path = './data/llm_app/llm/Qwen2-7B-Instruct/'

        model = AutoModelForCausalLM.from_pretrained(model_path,
                                                    device_map="auto")
        tokenizer = AutoTokenizer.from_pretrained(model_path)
        gen_config = GenerationConfig.from_pretrained(model_path)
```

- ollama和api的使用详情在其他文档说明。

下一节：【文档】星火大模型API使用

下一节