

## 3.2 本地化部署Deepseek

探索常用社区资源如Github/Huggingface/Modelscope，如何进行模型下载，并通过VLLM/Ollama/SGLang/Transformer启动模型，进行模型试用。

# 大模型部署实践教程

## 一、GitHub资源使用

### 1.1 GitHub资源使用, 查找Llama-Factory项目为例

1. 访问GitHub官网: <https://github.com>
2. 在搜索栏输入 Llama-Factory
3. 选择星标数高的官方仓库: [hiyouga/LLaMA-Factory](https://github.com/hiyouga/LLaMA-Factory)

### 项目克隆到本地, 以克隆Qwen1.5-0.5B仓库为例

在我们的AutoDL实例中, 打开JupyterLab

在项目文件夹目录下, 打开控制台

初始化git lfs

```
curl -s https://packagecloud.io/install/repositories/github/git-lfs/script.deb.sh | sudo bash && sudo apt-get install git-lfs && git lfs install
```

克隆github仓库的指令

1. git clone

网络监控:

1. 安装iftop

```
sudo apt-get install iftop
```

2. 运行 iftop

```
sudo iftop
```

安装jdk

1. 更新apt库

```
sudo apt update
```

2. 安装jdk

```
apt install openjdk-21-jdk
```

克隆Qwen1.5-0.5B仓库

```
git clone https://www.modelscope.cn/qwen/Qwen1.5-0.5B.git
```

出现如下的结果表示完成克隆

名称	已修改
deepseek-ai	4分钟前
Qwen1.5-0.5B	25秒前
scripts	6小时前
• download(1).ipy...	10分钟前
vllm.log	4小时前

```
Detected apt version as 2.4.12
Running apt-get update... done.
Installing apt-transport-https... done.
Installing /etc/apt/sources.list.d/github_git-lfs.list...done.
Importing packagecloud gpg key... Packagecloud gpg key imported to /etc/apt/keyrings/github_git-lfs-arc
hive-keyring.gpg
done.
Running apt-get update... done.

The repository is setup! You can now install packages.
Reading package lists... Done
Building dependency tree... Done
Reading state information... Done
The following NEW packages will be installed:
  git-lfs
0 upgraded, 1 newly installed, 0 to remove and 72 not upgraded.
Need to get 8489 kB of archives.
After this operation, 18.1 MB of additional disk space will be used.
Get:1 https://packagecloud.io/github/git-lfs/ubuntu jammy/main amd64 git-lfs amd64 3.6.1 [8489 kB]
Fetched 8489 kB in 8s (1121 kB/s)
debconf: delaying package configuration, since apt-utils is not installed
Selecting previously unselected package git-lfs.
(Reading database ... 48144 files and directories currently installed.)
Preparing to unpack .../git-lfs_3.6.1_amd64.deb ...
Unpacking git-lfs (3.6.1) ...
Setting up git-lfs (3.6.1) ...
Git LFS initialized.
Git LFS initialized.
(base) root@autodl-container-6caa4dbf52-eca3d646: # git clone https://www.modelscope.cn/qwen/Qwen1.5-0.5B.git
Cloning into 'Qwen1.5-0.5B'...
remote: Enumerating objects: 31, done.
remote: Counting objects: 100% (31/31), done.
remote: Compressing objects: 100% (30/30), done.
remote: Total 31 (delta 9), reused 0 (delta 0), pack-reused 0
Receiving objects: 100% (31/31), 3.60 MiB | 4.59 MiB/s, done.
Resolving deltas: 100% (9/9), done.
(base) root@autodl-container-6caa4dbf52-eca3d646:~#
```

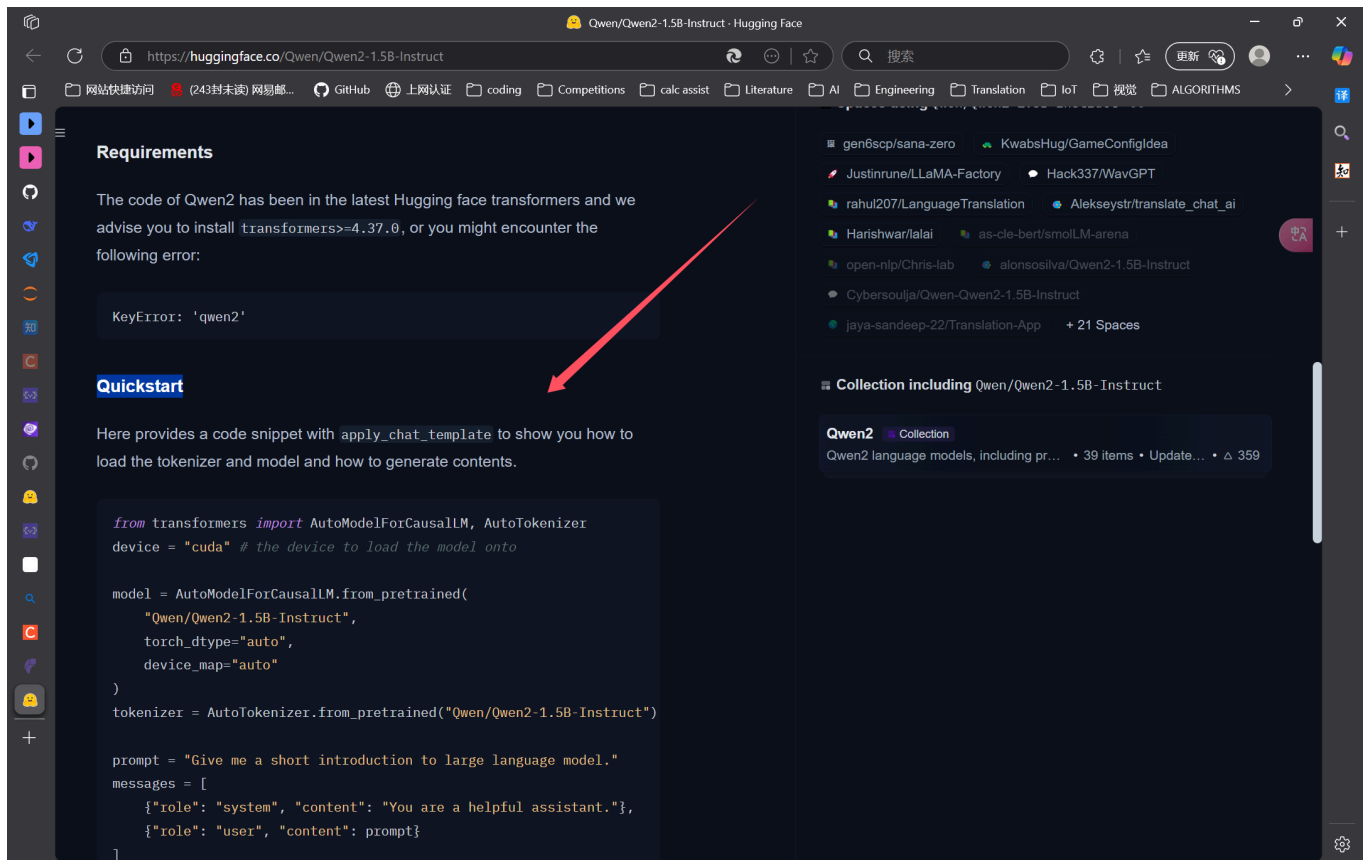
## 2.Hugging Face模型操作

### 搜索下载Qwen2.5模型

1. 访问官网: <https://huggingface.co>
2. 搜索栏输入 Qwen2.5-1.5B-Chat
3. 进入模型页: [Qwen/Qwen2-1.5B-Chat](#)

### 下载方式

在模型页中往下翻, 可以看到**Quickstart**类似的字样



复制使用Python代码下载

```
from transformers import AutoModelForCausalLM, AutoTokenizer
device = "cuda" # the device to load the model onto

model = AutoModelForCausalLM.from_pretrained(
    "Qwen/Qwen2-1.5B-Instruct",
    torch_dtype="auto",
    device_map="auto"
)
tokenizer = AutoTokenizer.from_pretrained("Qwen/Qwen2-1.5B-Instruct")

prompt = "Give me a short introduction to large language model."
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]

```

## HuggingFace查看排行榜

随着开源和闭源机器学习模型数量的爆炸式增长，找到适合你项目的正确模型可能非常困难。这就是HuggingFace启动评估项目的原因

- 开放 LLM 排行榜 评估和排名开源 LLM 和聊天机器人，并提供可重复的分数，将营销宣传与该领域的实际进展区分开来。
- Hub 上的排行榜 旨在收集 Hugging Face Hub 上的机器学习排行榜，并为评估创建者提供支持。

### 排行榜和评估 - Hugging Face 机器学习平台

Spaces

open-llm-leaderboard/open\_llm\_leaderboard


like 12.8k

Running on CPU UPGRADE

App

Files

Community 1135



## Open LLM Leaderboard Archived

Comparing Large Language Models in an open and reproducible way

4576 / 4576
Advanced Filters

Supports strict search and regex • Use semicolons for multiple terms

**Quick Filters**
For Edge Devices · 786
For Consumers · 430
Mid-range · 3185
For the GPU-rich · 165
Only Official Providers · 470

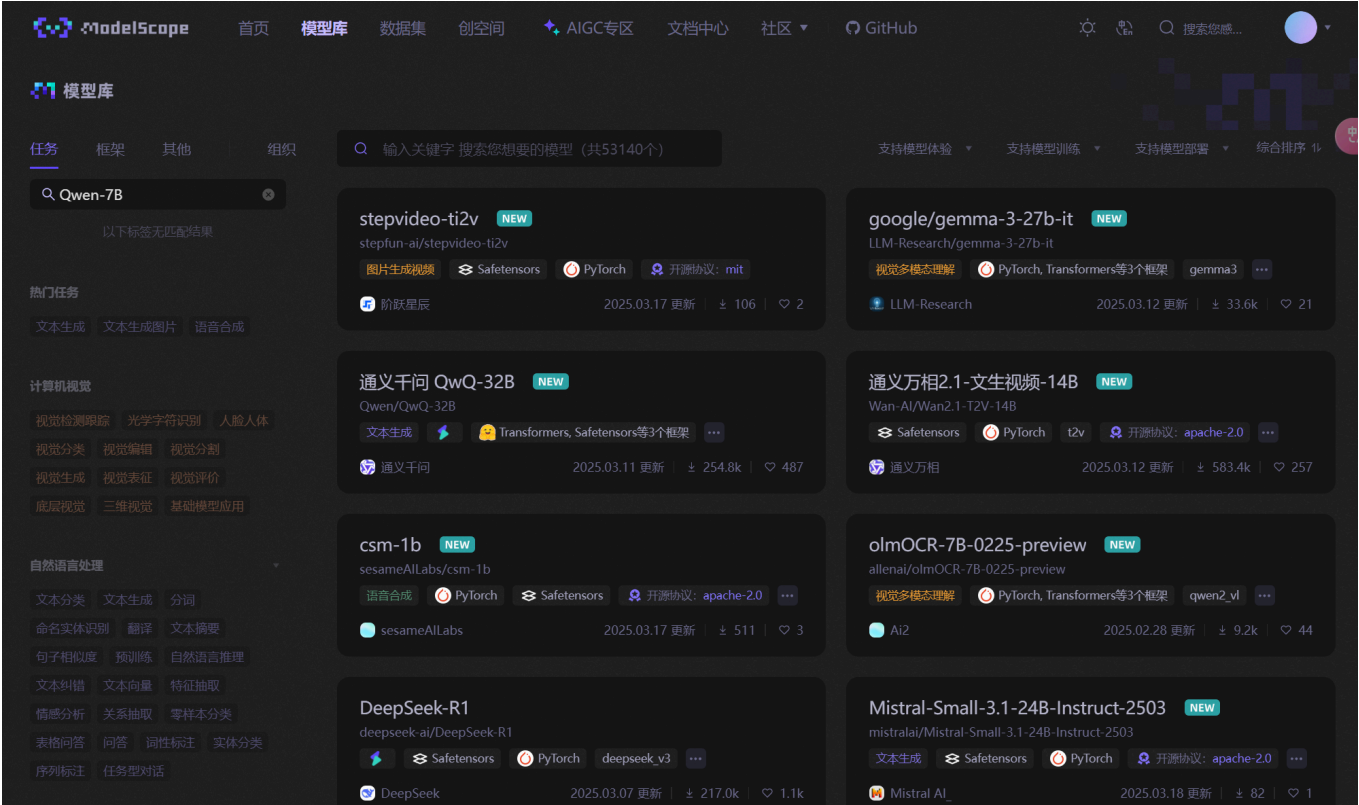
table options
column visibility

	Rank	Type	Model	Average	IFEval	BBH	MATH	GPQA	MUSR
👤	1	🔹	MazyarPanahi/calme-3.2-instruct-78b	52.08 %	80.63 %	62.61 %	40.33 %	20.36 %	38.53 %
👤	2	💬	MazyarPanahi/calme-3.1-instruct-78b	51.29 %	81.36 %	62.41 %	39.27 %	19.46 %	36.50 %
👤	3	💬	dfurman/CalmeRys-78B-Orpo-v0.1	51.23 %	81.63 %	61.92 %	40.63 %	20.02 %	36.37 %
👤	4	💬	MazyarPanahi/calme-2.4-rys-78b	50.77 %	80.11 %	62.16 %	40.71 %	20.36 %	34.57 %
👤	5	🔹	huihui-ai/Qwen2.5-72B-Instruct-abliterated	48.11 %	85.93 %	60.49 %	60.12 %	19.35 %	12.34 %

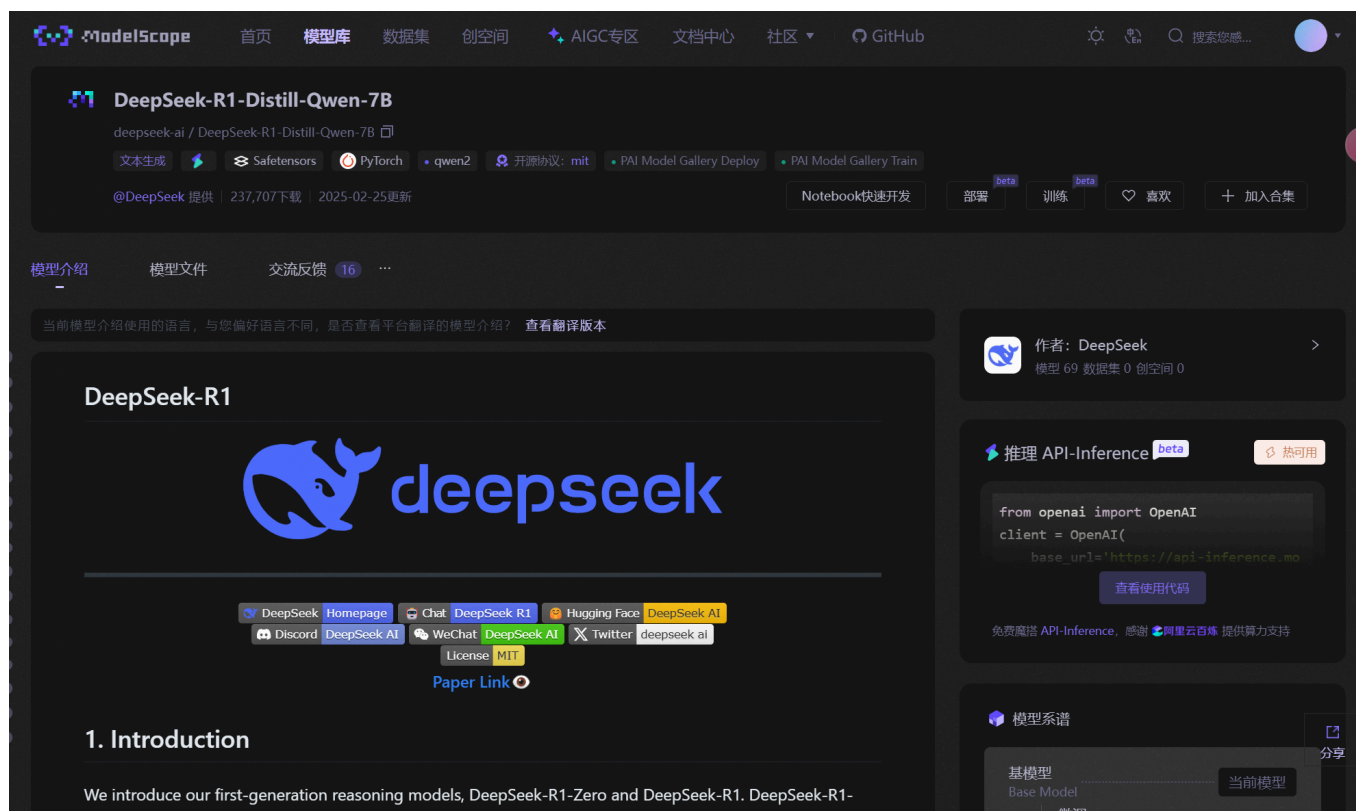
# 3.魔塔社区: 以DeepSeek-R1-Distill-Qwen-7B为例, 搜索模型, 下载模型

进入官网: <https://www.modelscope.cn/>

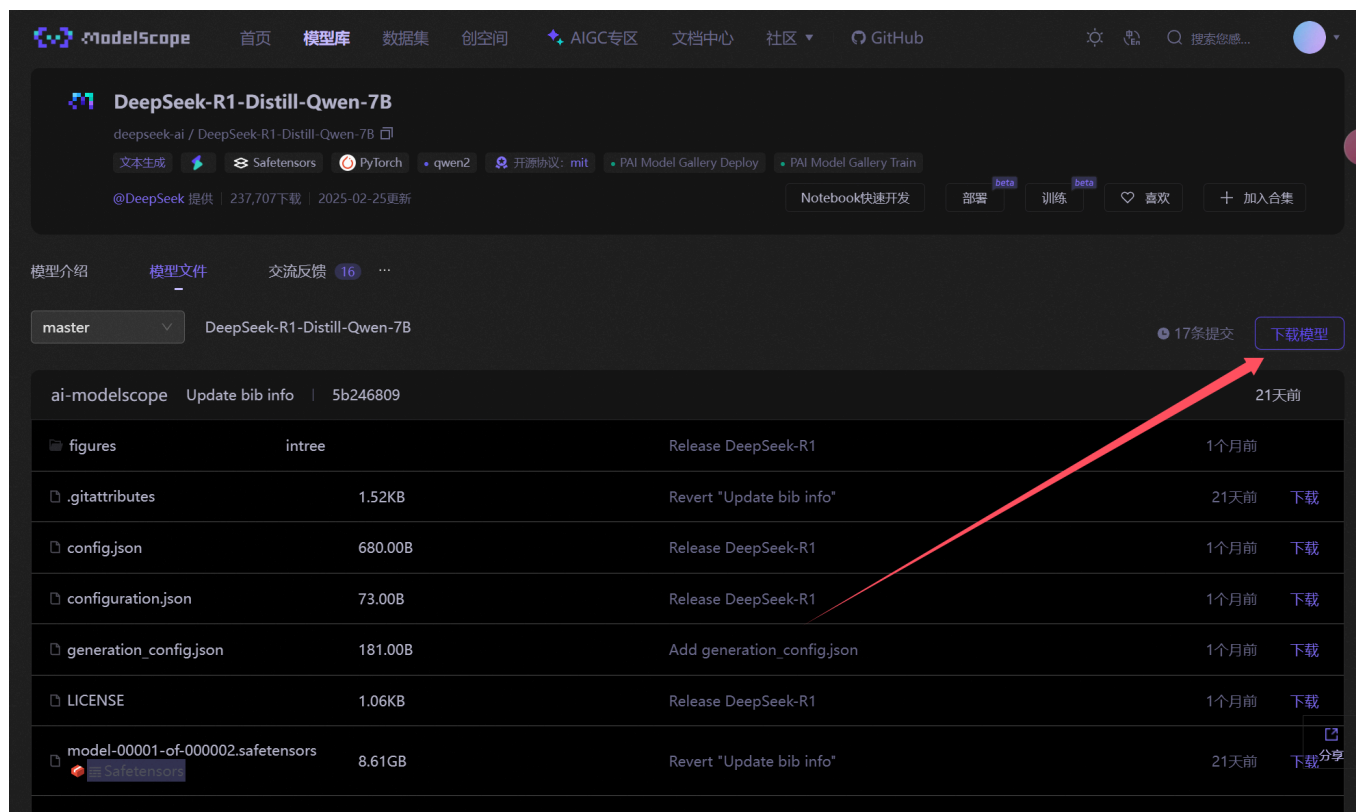
在上方导航栏点击模型库, 搜索 DeepSeek-R1-Distill-Qwen-7B



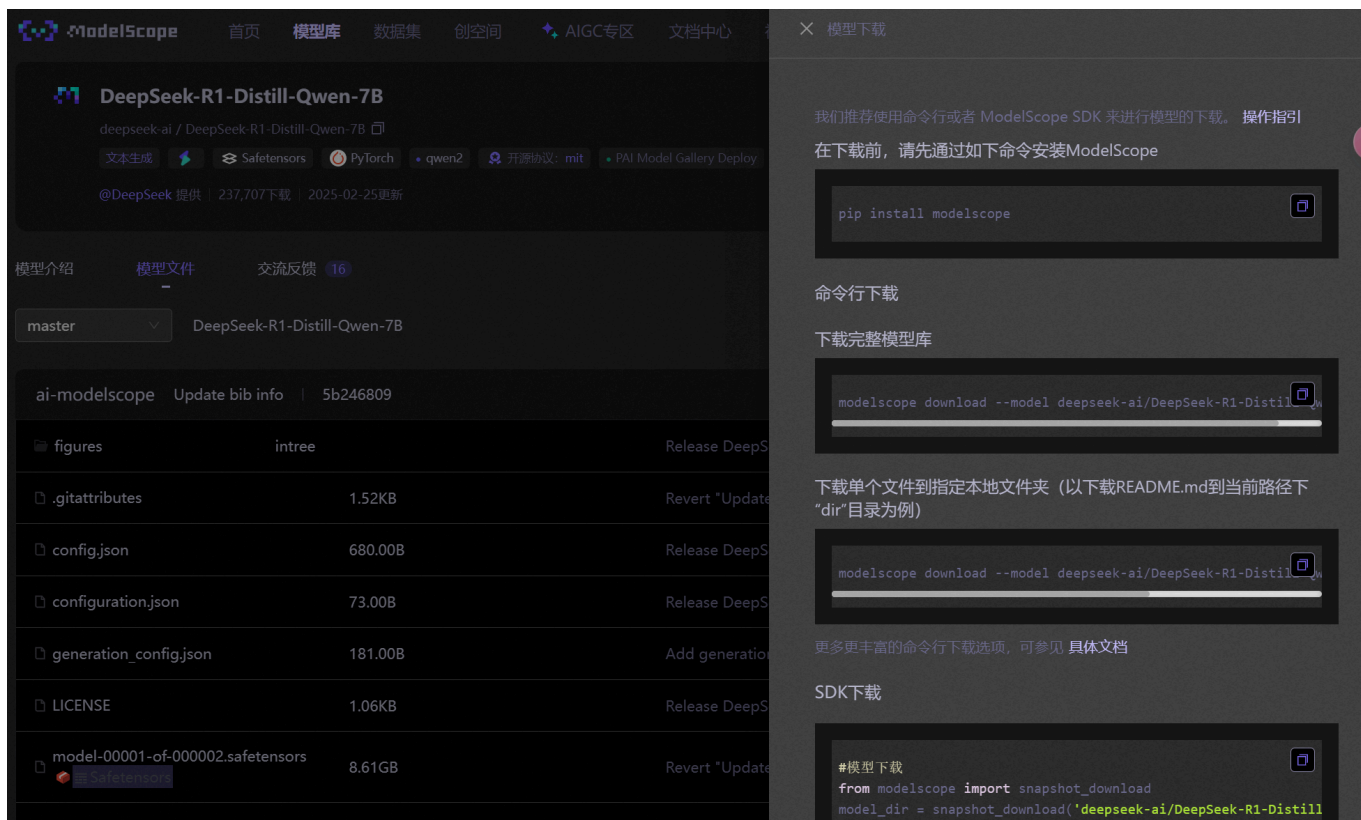
进入模型页



点击**模型文件**, 看到右侧有下载模型选项



选择合适的方式下载模型, 在提供的Jupyter文件中有所说明



在JupyterLab中, 进行如下操作

```
# 安装魔塔sdk
pip install modelscope

# 验证安装
python -c "from modelscope import snapshot_download; print('安装成功')"
```

mkdir model

```
# 安装DeepSeek-R1-Distill-Qwen-1.5B
python - <<EOF
#模型下载
from modelscope import snapshot_download
model_dir = snapshot_download('deepseek-ai/DeepSeek-R1-Distill-Qwen-1.5B',cache_dir = "./model/")
print(f"模型成功下载到{model_dir}")
EOF
```

## 4.VLLM启动DeepSeek-R1-Distill-Qwen-1.5B

在JupyterLab中, 打开进行如下操作

```
# 安装依赖
pip install vllm
```



```
# 使用tmux保持会话
tmux new -s deepseek

# 启动命令（适配3090显存）
python -m vllm.entrypoints.api_server \
--model /home/featurize/data/deepseek-r1-1.5b #!!这里改成刚刚模型下载到的地址\
--tensor-parallel-size 1 \
--gpu-memory-utilization 0.9 \
--max-num-batched-tokens 4096

# 按Ctrl+B D退出tmux会话
```

输出 安装成功 表示成功

## 5.Ollama启动DeepSeek-R1-Distill-Qwen-1.5B

## 6.Transformers启动Qwen2.5-1.5B

细节参考jupyter