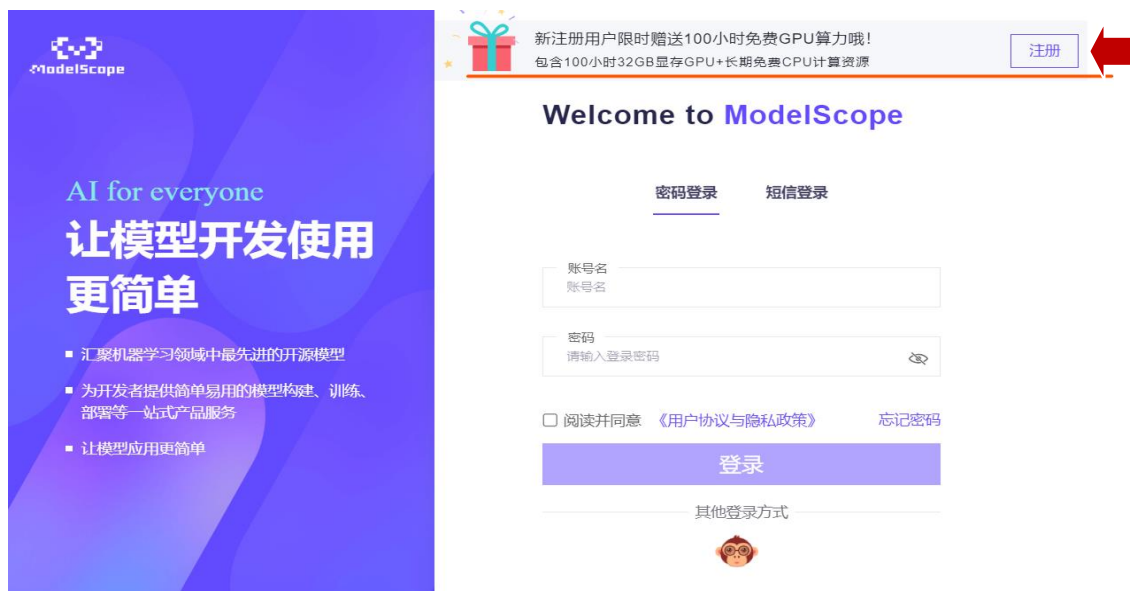
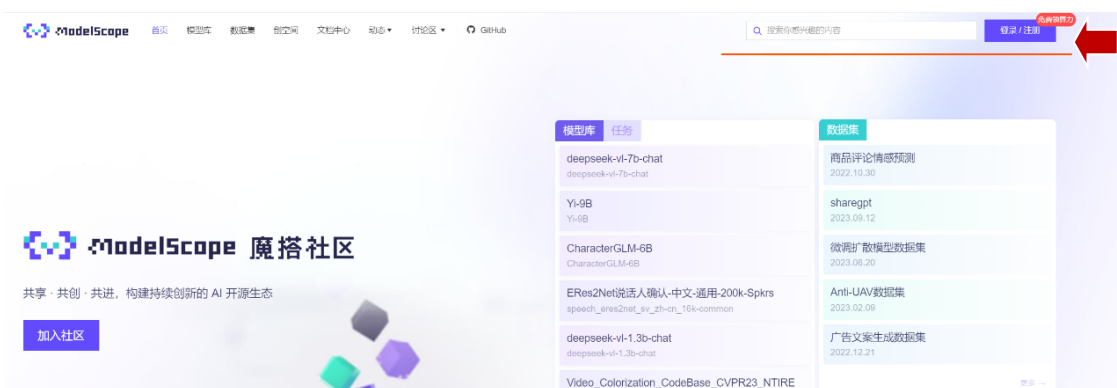


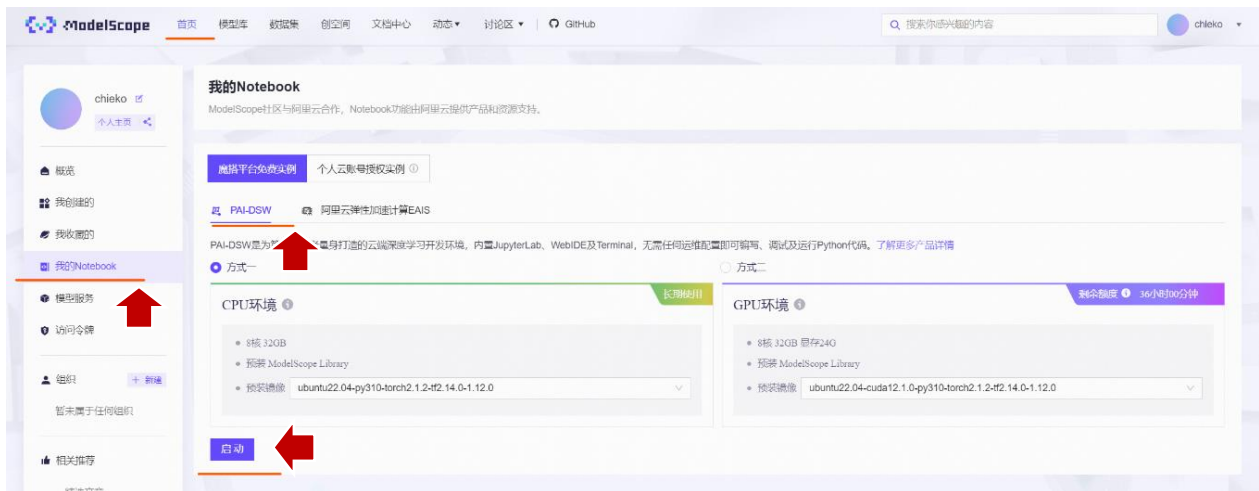
动手实验

一、注册&登录 ModelScope

1. 进入 ModelScope (<https://www.modelscope.cn/home>)，右上角点击完成新用户的注册；



2. 注册完成后，登录 modelscope，进入首页，请确定已经绑定阿里云账号，并已获得免费的配套云计算资源。启动 CPU 服务器：

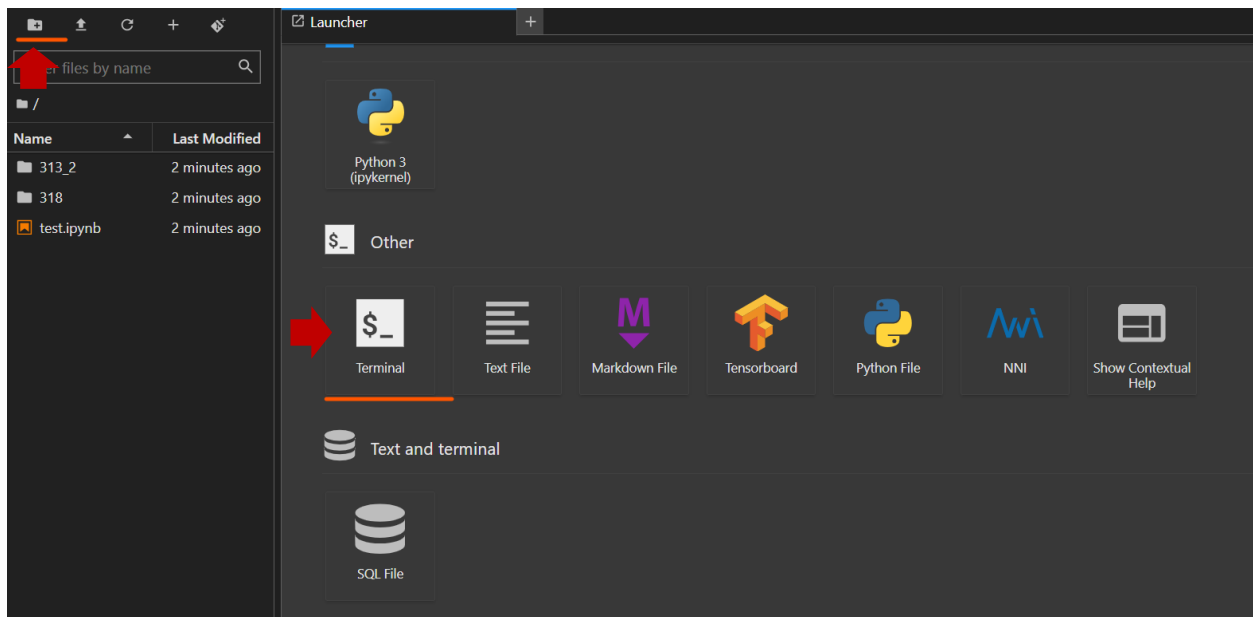


3. 启动成功后，点击“查看”



二、环境配置

1. 打开 terminal



2. 环境搭建:

进入环境目录/opt/conda/envs, 新建文件夹 itrex:

```
cd /opt/conda/envs  
mkdir itrex
```

拷贝镜像文件至 itrex 目录(任选一个下载即可):

```
下载源 1:  
wget https://idz-ai.oss-cn-hangzhou.aliyuncs.com/LLM/itrex.tar.gz  
下载源 2:  
wget https://filerepo.idzcn.com/LLM/itrex.tar.gz
```

解压文件:

```
tar -zxvf itrex.tar.gz -C itrex/
```

激活环境:

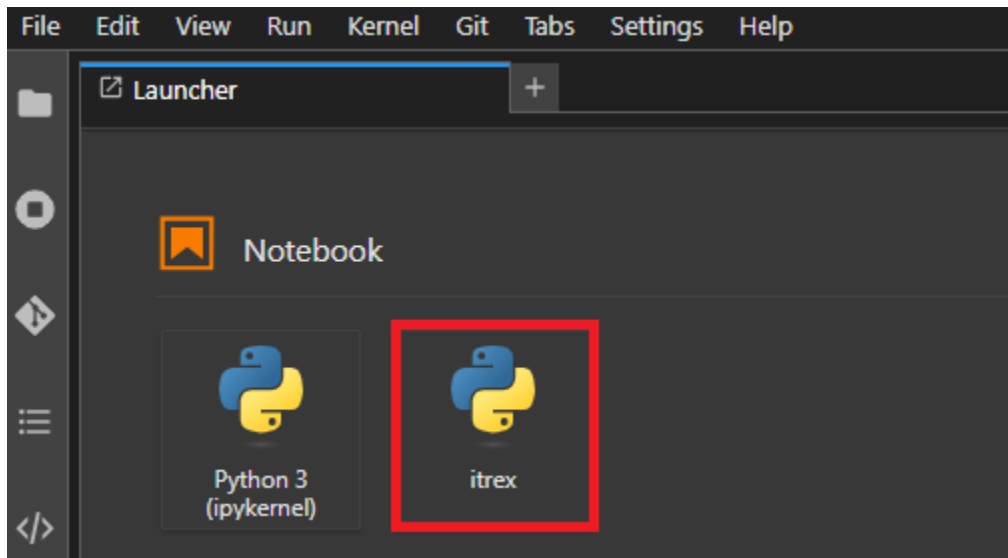
```
conda activate itrex
```

安装对应的 kernel:

```
python -m ipykernel install --name itrex
```

三、创建 notebook

1. 基于 itrex kernel 新建 notebook



2. 下载模型

新建 cell, 下载中文大模型:

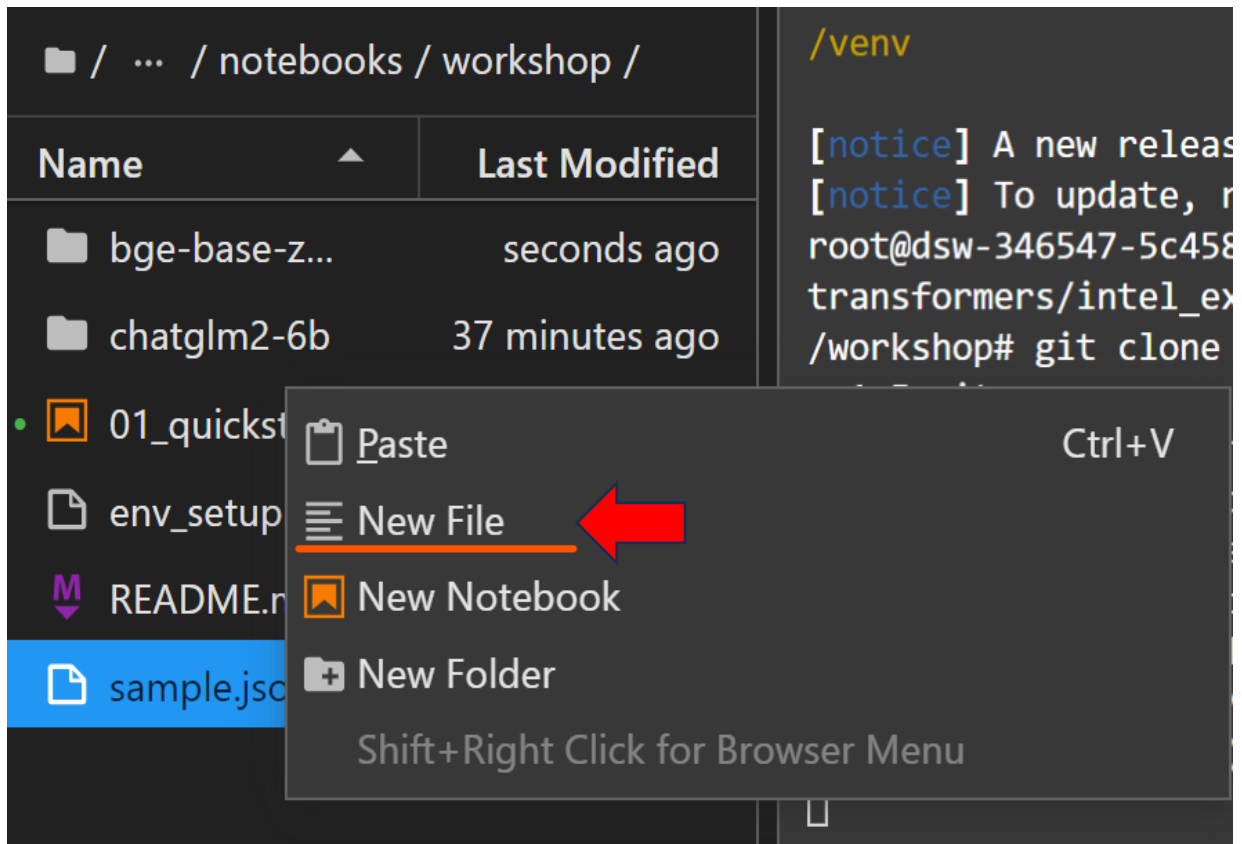
```
! git clone https://www.modelscope.cn/ZhipuAI/chatglm3-6b.git
```

新建 cell, 下载 embedding 模型:

```
! git clone https://www.modelscope.cn/AI-ModelScope/bge-base-zh-v1.5.git
```


准备知识库文件 sample.jsonl

1) 新建文件 sample.jsonl: 右键->New File



2) 双击文件，在文件里面输入如下内容，保存文件。

```
{"content": "cnvrg.io 网站由 Yochay Ettun 和 Leah Forkosh Kolben 创建.", "link": 0}
```

回到 notebook，新建 cell，添加以下代码构建 chatbot，点击  运行：

```

from intel_extension_for_transformers.neural_chat import
PipelineConfig
from intel_extension_for_transformers.neural_chat import
build_chatbot
from intel_extension_for_transformers.neural_chat import plugins
from intel_extension_for_transformers.transformers import RtnConfig

plugins.retrieval.enable=True
plugins.retrieval.args['embedding_model'] = "./bge-base-zh-v1.5"
plugins.retrieval.args["input_path"]="./sample.jsonl"
config = PipelineConfig(model_name_or_path='./chatglm3-6b',
                        plugins=plugins,
                        optimization_config=RtnConfig(compute_dtype="int8",
weight_dtype="int4_fullrange"))

chatbot = build_chatbot(config)

```

新建 cell，添加以下代码 disable retrieval，点击  运行：

```

plugins.retrieval.enable=False # disable retrieval
response = chatbot.predict(query="cnvrg.io 网站是由谁创建的? ")
print(response)

```

新建 cell，添加以下代码 enable retrieval，点击  运行：

```

plugins.retrieval.enable=True # enable retrieval
response = chatbot.predict(query="cnvrg.io 网站是由谁创建的? ")
print(response)

```