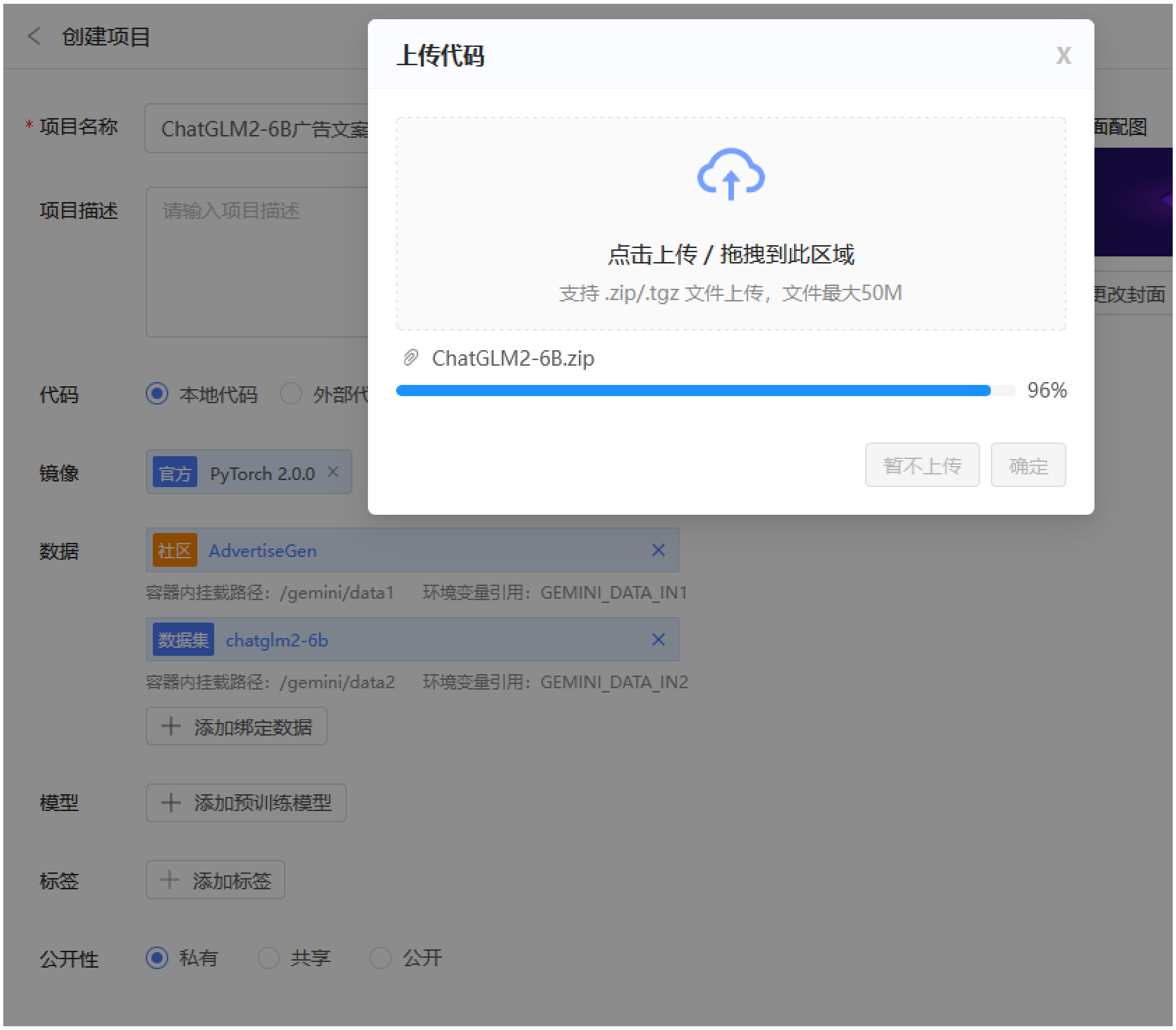
金水Task2笔记

1.首先和Task1流程类似，先配置好环境，数据，和要运行的代码



< 创建项目

* 项目名称 ChatGLM2-6B广告文案

项目描述 请输入项目描述

代码 ● 本地代码 ○ 外部代

镜像 官方 PyTorch 2.0.0 ×

数据 社区 AdvertiseGen ×

容器内挂载路径：/gemini/data1 环境变量引用：GEMINI_DATA_IN1

数据集 chatglm2-6b ×

容器内挂载路径：/gemini/data2 环境变量引用：GEMINI_DATA_IN2

+ 添加绑定数据

模型 + 添加预训练模型

标签 + 添加标签

公开性 ● 私有 ○ 共享 ○ 公开

面配图

更改封面

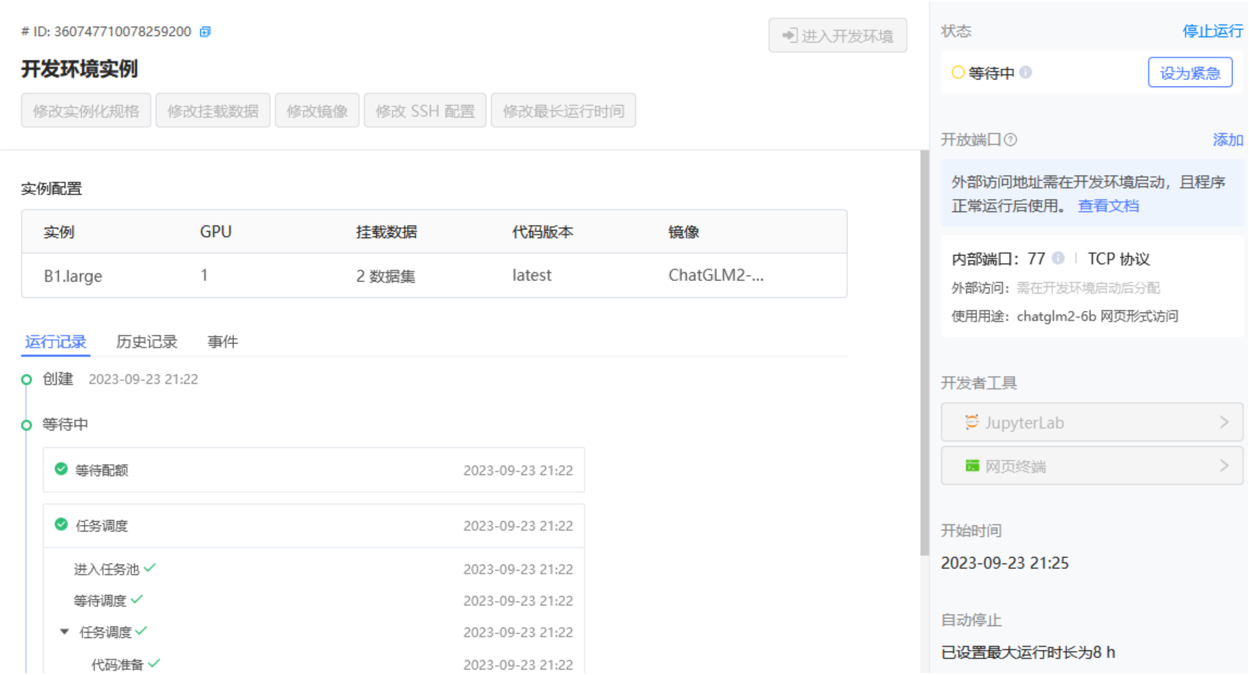
上传代码 X

点击上传 / 拖拽到此区域

支持 .zip/.tgz 文件上传，文件最大50M

ChatGLM2-6B.zip 96%

2.镜像换为chatglm2-6b,并进行端口77的开放



ID: 360747710078259200

开发环境实例

实例配置

实例	GPU	挂载数据	代码版本	镜像
B1.large	1	2 数据集	latest	ChatGLM2-...

创建 2023-09-23 21:22

等待中

等待配额	2023-09-23 21:22
任务调度	2023-09-23 21:22
进入任务池 ✓	2023-09-23 21:22
等待调度 ✓	2023-09-23 21:22
▾ 任务调度 ✓	2023-09-23 21:22
代码准备 ✓	2023-09-23 21:22

状态

○ 等待中

开放端口

外部访问地址需在开发环境启动，且程序正常运行后使用。

内部端口：77 | TCP 协议

外部访问：需在开发环境启动后分配

使用用途：chatglm2-6b 网页形式访问

开发者工具

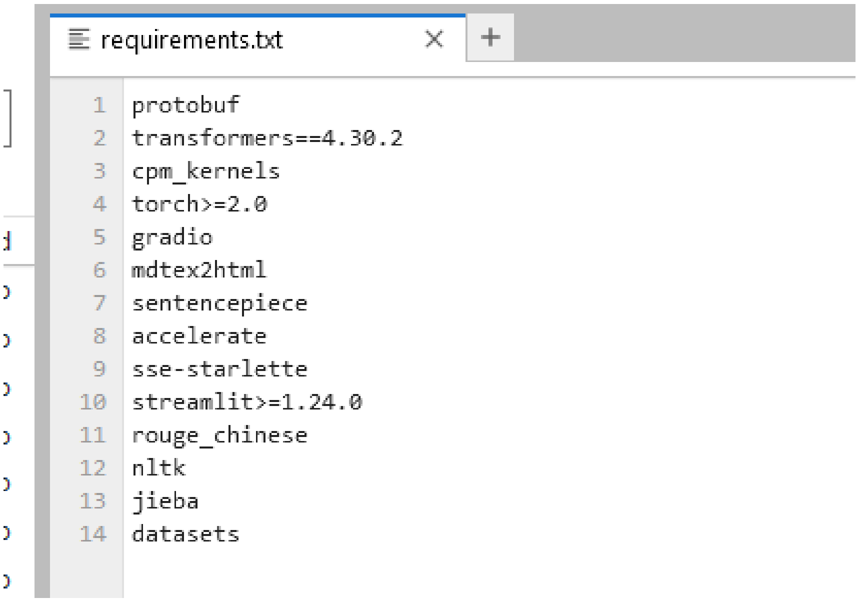
开始时间

2023-09-23 21:25

自动停止

已设置最大运行时长为8 h

3,改下requirements.txt



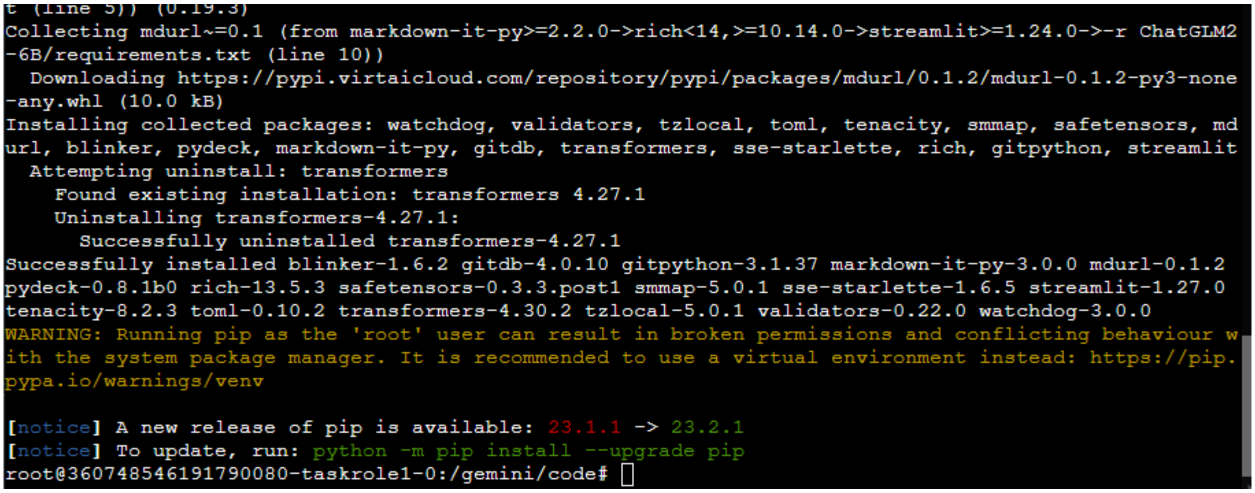
requirements.txt

```
protobuf
transformers==4.30.2
cpm_kernels
torch>=2.0
gradio
mdtex2html
sentencepiece
accelerate
sse-starlette
streamlit>=1.24.0
rouge_chinese
nltk
jieba
datasets
```

4.在终端运行

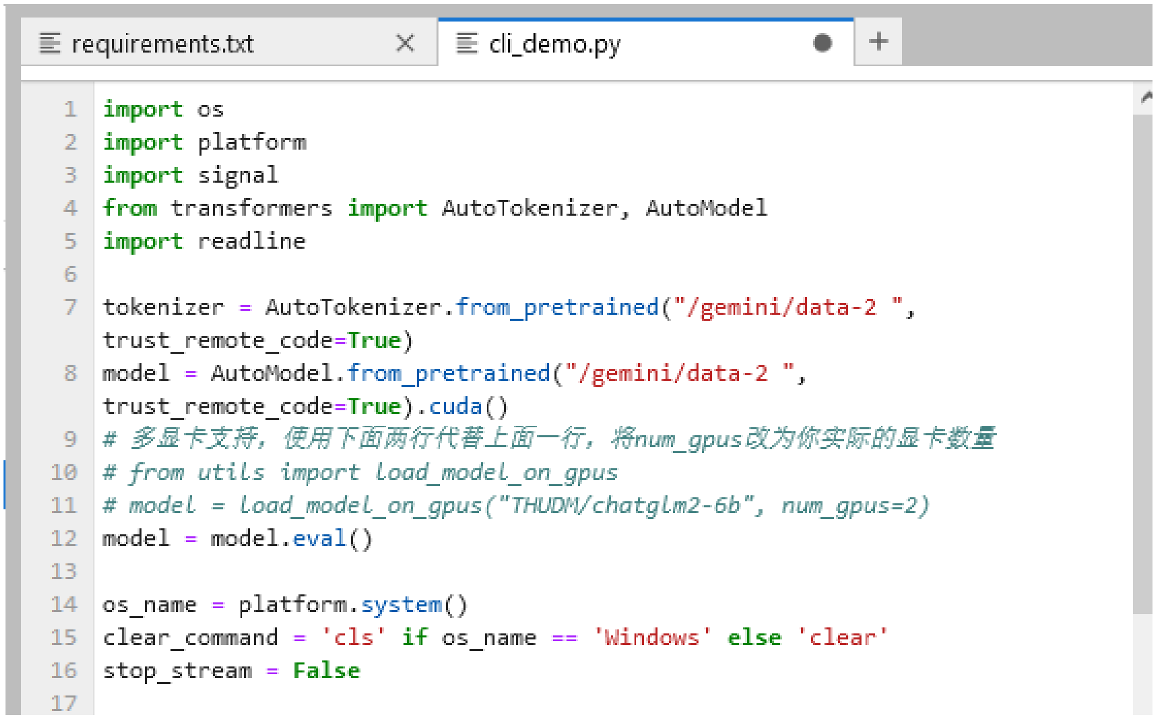
```
pip install -r ChatGLM2-6B/requirements.txt -i https://pypi.virtaicloud.com/repository/pypi/sim
```

安装成功



```
t (line 5)) (0.19.3)
Collecting mdurl~=0.1 (from markdown-it-py>=2.2.0->rich<14,>=10.14.0->streamlit>=1.24.0->-r ChatGLM2
-6B/requirements.txt (line 10))
  Downloading https://pypi.virtaicloud.com/repository/pypi/packages/mdurl/0.1.2/mdurl-0.1.2-py3-none
-any.whl (10.0 kB)
Installing collected packages: watchdog, validators, tzlocal, toml, tenacity, smmap, safetensors, md
url, blinker, pydeck, markdown-it-py, gitdb, transformers, sse-starlette, rich, gitpython, streamlit
  Attempting uninstall: transformers
    Found existing installation: transformers 4.27.1
    Uninstalling transformers-4.27.1:
      Successfully uninstalled transformers-4.27.1
Successfully installed blinker-1.6.2 gitdb-4.0.10 gitpython-3.1.37 markdown-it-py-3.0.0 mdurl-0.1.2
pydeck-0.8.1b0 rich-13.5.3 safetensors-0.3.3.post1 smmap-5.0.1 sse-starlette-1.6.5 streamlit-1.27.0
tenacity-8.2.3 toml-0.10.2 transformers-4.30.2 tzlocal-5.0.1 validators-0.22.0 watchdog-3.0.0
WARNING: Running pip as the 'root' user can result in broken permissions and conflicting behaviour w
ith the system package manager. It is recommended to use a virtual environment instead: https://pip.
pypa.io/warnings/venv

[notice] A new release of pip is available: 23.1.1 -> 23.2.1
[notice] To update, run: python -m pip install --upgrade pip
root@36074854619179008 0-taskrole1-0:/gemini/code# 
```

5.修改 cli_demo.py 中模型的地址，将 `THUDM/chatglm2-6b` 替换为 `/gemini/data-2` （即模型实际挂载到环境的地址）

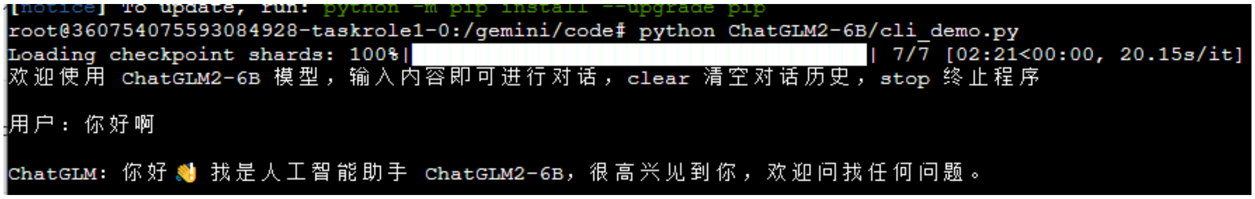


```python
import os
import platform
import signal
from transformers import AutoTokenizer, AutoModel
import readline

tokenizer = AutoTokenizer.from_pretrained("/gemini/data-2 ", 
trust_remote_code=True)
model = AutoModel.from_pretrained("/gemini/data-2 ", 
trust_remote_code=True).cuda()
# 多显卡支持，使用下面两行代替上面一行，将num_gpus改为你实际的显卡数量
# from utils import load_model_on_gpus
# model = load_model_on_gpus("THUDM/chatglm2-6b", num_gpus=2)
model = model.eval()

os_name = platform.system()
clear_command = 'cls' if os_name == 'Windows' else 'clear'
stop_stream = False

```

6.1. 切换至 网页终端。执行如下命令唤醒交互式对话。等待最终 `Loading checkpoint shards: 100%` 且返回 `用户：` 字样。。

```
python ChatGLM2-6B/cli_demo.py
```

成功



```
[notice] To update, run: python -m pip install --upgrade pip
root@360754075593084928-taskrole1-0:/gemini/code# python ChatGLM2-6B/cli_demo.py
Loading checkpoint shards: 100%|██████████████████████████████████| 7/7 [02:21<00:00, 20.15s/it]
欢迎使用 ChatGLM2-6B 模型，输入内容即可进行对话，clear 清空对话历史，stop 终止程序

用户：你好啊

ChatGLM: 你好👋 我是人工智能助手 ChatGLM2-6B，很高兴见到你，欢迎问我任何问题。
```