华为“智能基座”系列课程

基于Mindspore-Ascend的

llama7b模型推理与微调实验

版本：1.0（武汉理工大学更新）



华为技术有限公司

|  |
| --- |
| 版权所有 © 华为技术有限公司 2023。 保留一切权利。  非经本公司书面许可，任何单位和个人不得擅自摘抄、复制本文档内容的部分或全部，并不得以任何形式传播。  商标声明  HW_POS_RBG_Vertical-150ppi 和其他华为商标均为华为技术有限公司的商标。  本文档提及的其他所有商标或注册商标，由各自的所有人拥有。  注意  您购买的产品、服务或特性等应受华为公司商业合同和条款的约束，本文档中描述的全部或部分产品、服务或特性可能不在您的购买或使用范围之内。除非合同另有约定，华为公司对本文档内容不做任何明示或暗示的声明或保证。  由于产品版本升级或其他原因，本文档内容会不定期进行更新。除非另有约定，本文档仅作为使用指导，本文档中的所有陈述、信息和建议不构成任何明示或暗示的担保。 |

|  |  |
| --- | --- |
| 华为技术有限公司 | |
| 地址： | 深圳市龙岗区坂田华为总部办公楼 邮编：518129 |
| 网址： | http://[e](http://e.huawei.com/).huawei.com |

目录

[1 华为云ModelArts环境创建 2](#_Toc149809483)

[1.1 登录ModelArts 2](#_Toc149809484)

[1.2 创建开发环境 2](#_Toc149809485)

[2 环境版本说明与实验包下载 6](#_Toc149809486)

[2.1 系统及软件要求 6](#_Toc149809487)

[2.2 创建新的ipynb文件 6](#_Toc149809488)

[2.3 实验包下载 7](#_Toc149809489)

[3 Llama7b推理测试 8](#_Toc149809490)

[3.1 推理流程简介 8](#_Toc149809491)

[4 Lora微调训练： 10](#_Toc149809492)

[4.1 数据预处理： 10](#_Toc149809493)

[4.2 修改配置文件（展示） 11](#_Toc149809494)

[4.3 训练 12](#_Toc149809495)

[4.4 使用微调后的模型推理 13](#_Toc149809496)

[5 课后思考题 14](#_Toc149809497)

[5.1 思考题一 14](#_Toc149809498)

[5.2 思考题二 14](#_Toc149809499)

[5.3 思考题三 14](#_Toc149809500)

# 华为云ModelArts环境创建

## 登录ModelArts

登录ModelArts官方链接：<https://console.huaweicloud.com/modelarts/?region=cn-north-9#/dev-container>

## 创建开发环境

在网页左上角，将区域改为：贵阳一， 如下图：



ModelArts栏上点击开发环境，并且点击NoteBook,然后点击创建。如图：



完成以下操作：

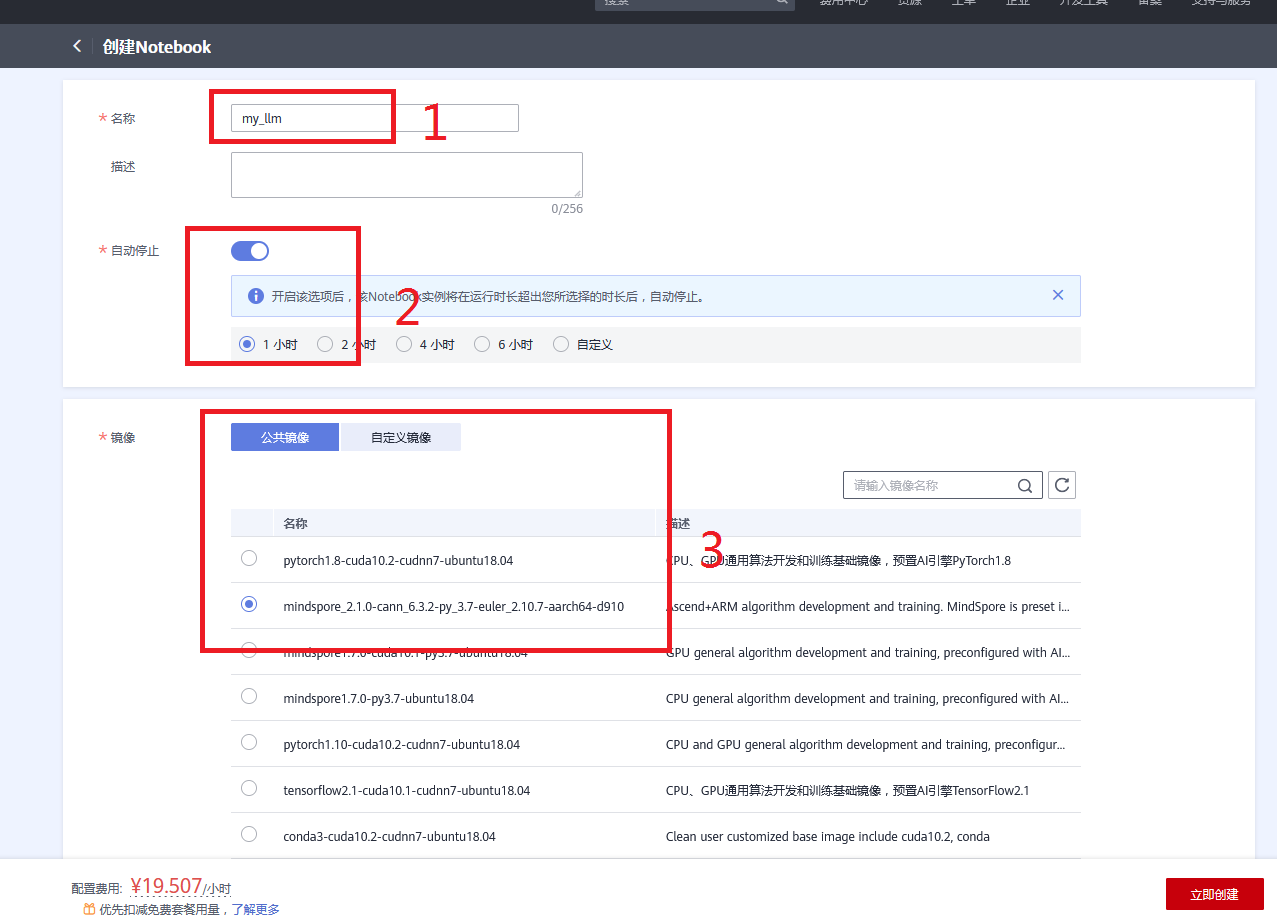
1、修改镜像名字；

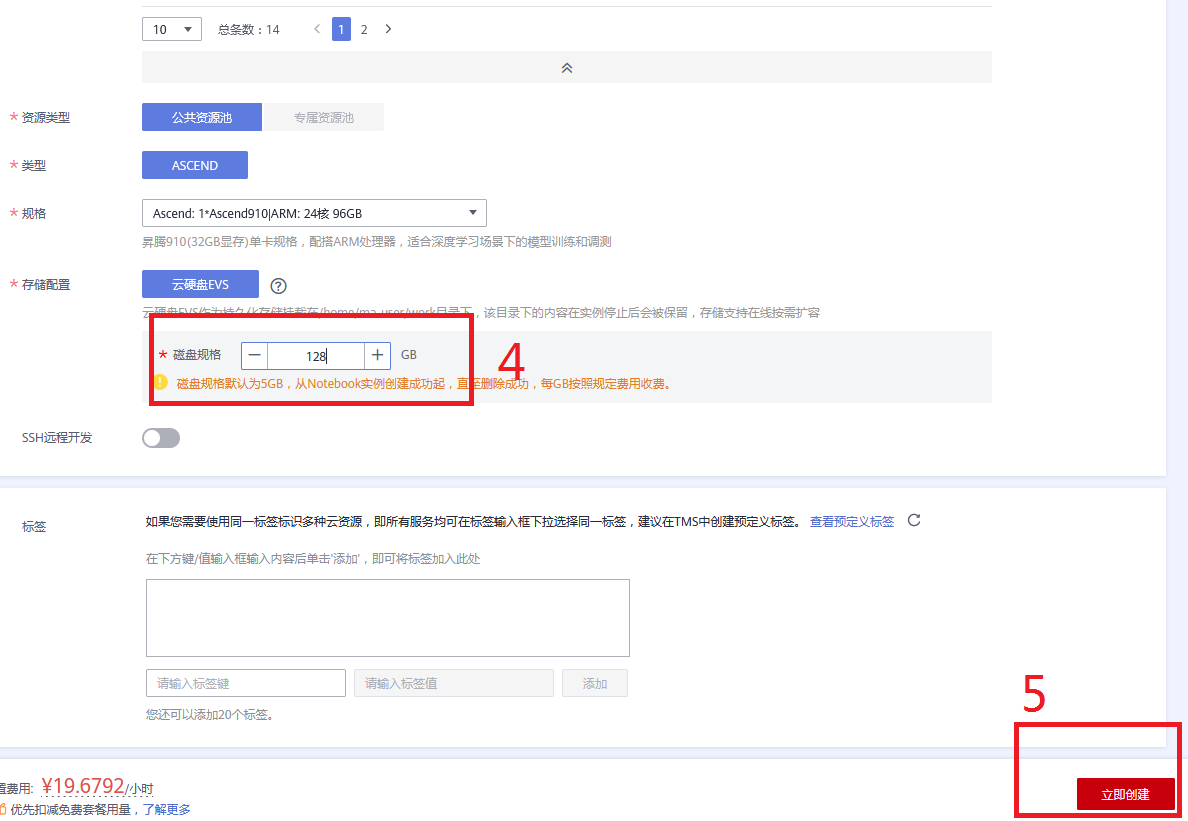
2、开启自动停止，并且选择1小时；

3、选择镜像版本为：mindspore\_2.2.0-cann\_6.3.2-py\_3.7-euler\_2.10.7-aarch64-d910；

4、修改镜像磁盘容量为128g；

5、最后点击“立即创建”，“提交”。





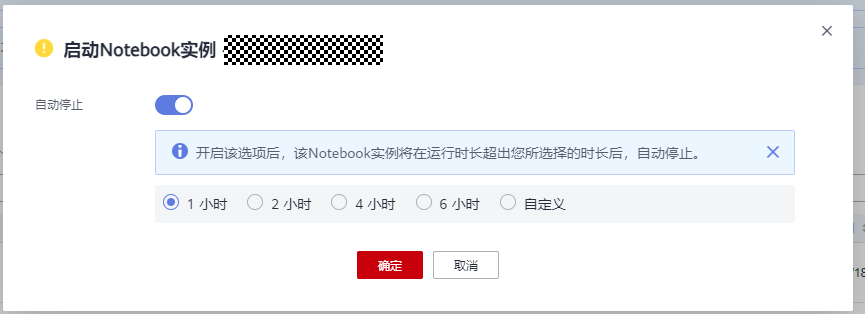
点击“返回”可看见镜像创建中：

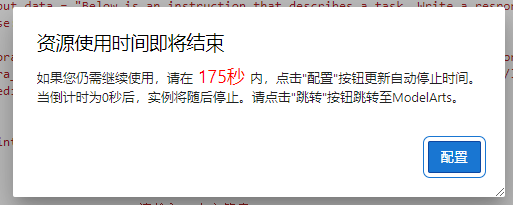


待创建结束后，在操作栏点击“打开”：



* PS：注意！！ 开启自动停止，该Notebook实例将在运行时长超出您所选择的时长后，自动停止；如在使用过程中超过所选的时长，将会有时长更新的提示，按照自己的情况选择更新的时间。





# 环境版本说明与实验包下载

## 系统及软件要求

本实验所使用的系统和软件版本如下表，对于操作界面差异不大的版本没有列出。

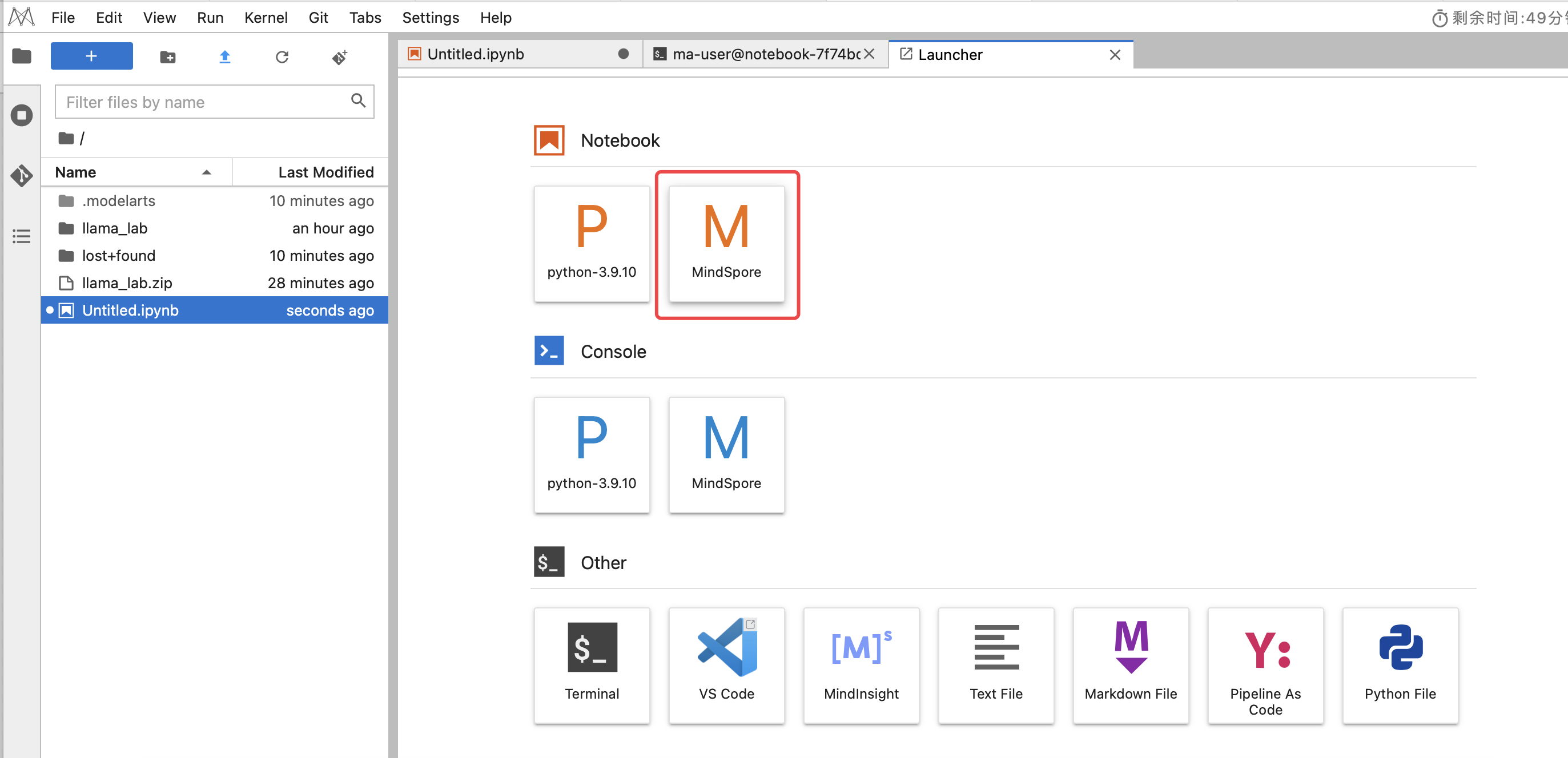
训练环境：

|  |  |  |
| --- | --- | --- |
| 类别 | 版本 | 说明 |
| 硬件 | Ascend 910B | 昇腾NPU |
| MindSpore | MindSpore2.2 | 全场景深度学习框架 |
| MindFormers | r1.0 | 大模型训练、推理和部署全流程开发套件 |
| Python | Python3.9 | 面向对象编程语言 |
| CANN | cann\_7.0 | 昇腾异构计算架构 |
| 系统镜像 | euleros\_2.10.7 | 服务器操作系统 |
| CPU架构 | aarch64 | CPU架构 |
| 存储 | 128G以上 |  |

训练环境系统/软件版本

## 创建新的ipynb文件

点击MindSpore标签按钮，创建的ipynb文件在/home/ma-user/work下：



## 实验包下载

新建cell写入以下代码，并且运行：

%env no\_proxy='a.test.com,127.0.0.1,2.2.2.2'

!wget http://59.69.101.2/llama\_lab.zip

!unzip llama\_lab.zip

* 注意：此包必须下载到/home/ma-user/work下

# Llama7b推理测试

## 推理流程简介

打开终端安装mindformers（不要使用notebook）

pip install tokenizers

pip install pyarrow

cd llama\_lab/mindformers/ && bash build.sh

进入checkpoint\_download/llama下，使用wget 下载以下两个文件：

!cd llama\_lab/mindformers/checkpoint\_download/llama/ && wget https://ascend-repo-modelzoo.obs.cn-east-2.myhuaweicloud.com/XFormer\_for\_mindspore/llama/open\_llama\_7b.ckpt

!cd llama\_lab/mindformers/checkpoint\_download/llama/ && mv open\_llama\_7b.ckpt llama\_7b.ckpt

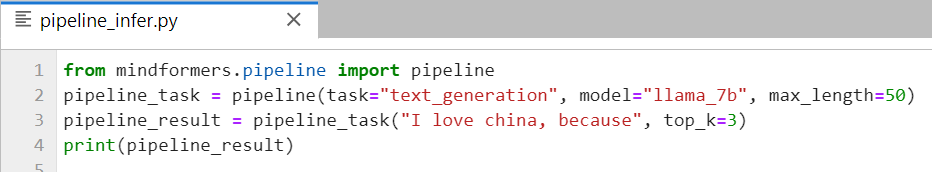
!cd llama\_lab/mindformers/checkpoint\_download/llama/ && wget https://ascend-repo-modelzoo.obs.cn-east-2.myhuaweicloud.com/XFormer\_for\_mindspore/llama/tokenizer.model

下载完模型相关的文件，可以直接使用实验包里面的pipeline\_infer.py推理脚本来推理：

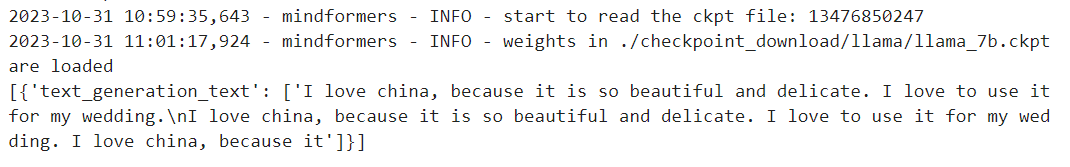
!cd llama\_lab/mindformers && python pipeline\_infer.py

Ps：初次推理时间约为7mins。

pipeline\_infer.py代码如下所示，也可以自己编辑一些感兴趣的问题问llama模型：



输出如下，表示推理成功：



# Lora微调训练：

## 数据预处理：

数据预处理1

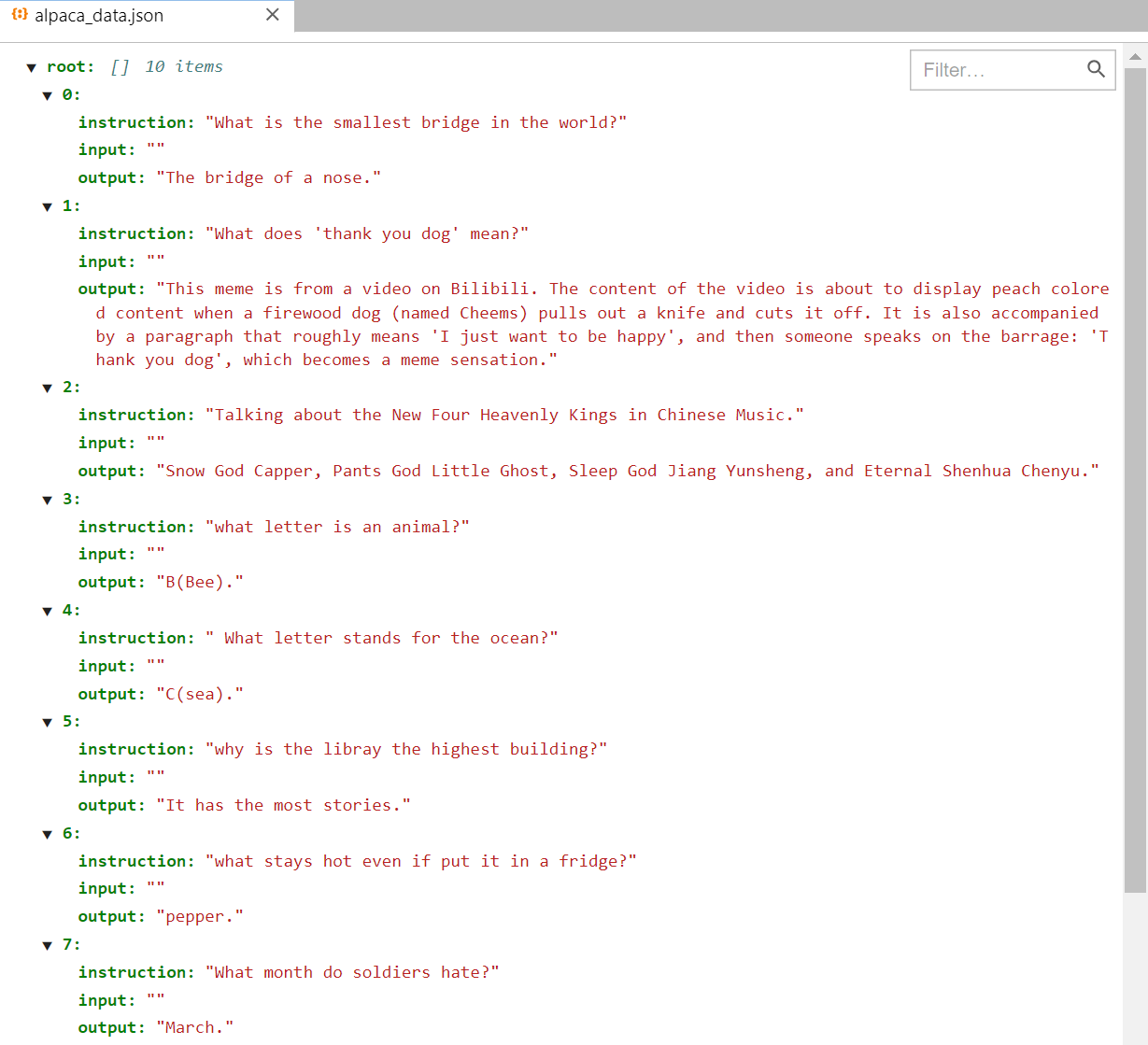
使用alpaca\_converter.py把原数据格式改为指定的问答格式

!cd llama\_lab/mindformers/mindformers/tools/dataset\_preprocess/llama && python alpaca\_converter.py --data\_path /home/ma-user/work/llama\_lab/stanford\_alpaca/alpaca\_data.json --output\_path /home/ma-user/work/llama\_lab/stanford\_alpaca/alpaca-data-conversation.json

参数说明：

* data\_path: 存放alpaca数据的路径
* output\_path: 输出转换后对话格式的数据路径

可以根据alpaca\_data.json的数据格式编写自己的问题，下图为alpaca\_data.json的数据表示：



数据预处理2

使用llama\_preprocess.py把文本数据转为mindspore的数据格式：

!cd llama\_lab/mindformers/mindformers/tools/dataset\_preprocess/llama && python llama\_preprocess.py --dataset\_type qa --input\_glob /home/ma-user/work/llama\_lab/stanford\_alpaca/alpaca-data-conversation.json --model\_file /home/ma-user/work/llama\_lab/mindformers/checkpoint\_download/llama/tokenizer.model --seq\_length 2048 --output\_file /home/ma-user/work/llama\_lab/stanford\_alpaca/alpaca-fastchat2048.mindrecord

参数说明：

* dataset\_type: 数据集类型
* iutput\_glob: 输入文本文件
* model\_file:分词文件
* seq\_length:序列长度
* output\_file:输出数据文件

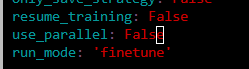
## 修改配置文件（展示）

本实验包对应的配置文件已经修改好，以下介绍需要修改的地方：

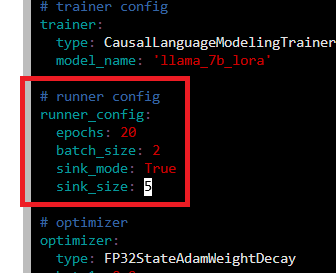
* 打开llama\_lab/mindformers/configs/llama/run\_llama\_7b\_lora.yaml
* 在第3行，设置load\_checkpoint为pipeline\_infer时候自动下载ckpt文件 （/home/ma-user/work/llama\_lab/mindformers/checkpoint\_download/llama/llama\_7b.ckpt）



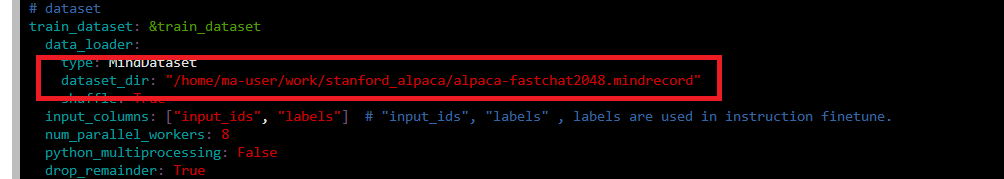
* 在第76行，use\_parallel设置为False



* 在第16行，runner\_config设置为：（以下`设置在少数据情况下，表现优良），如果时间过长可以将epochs调为5或3



* 第41行，修改train\_dataset的dataset\_dir为数据处理部分生产的alpaca-fastchat2048.mindrecord的路径（/home/ma-user/work/llama\_lab/stanford\_alpaca/alpaca-fastchat2048.mindrecord）



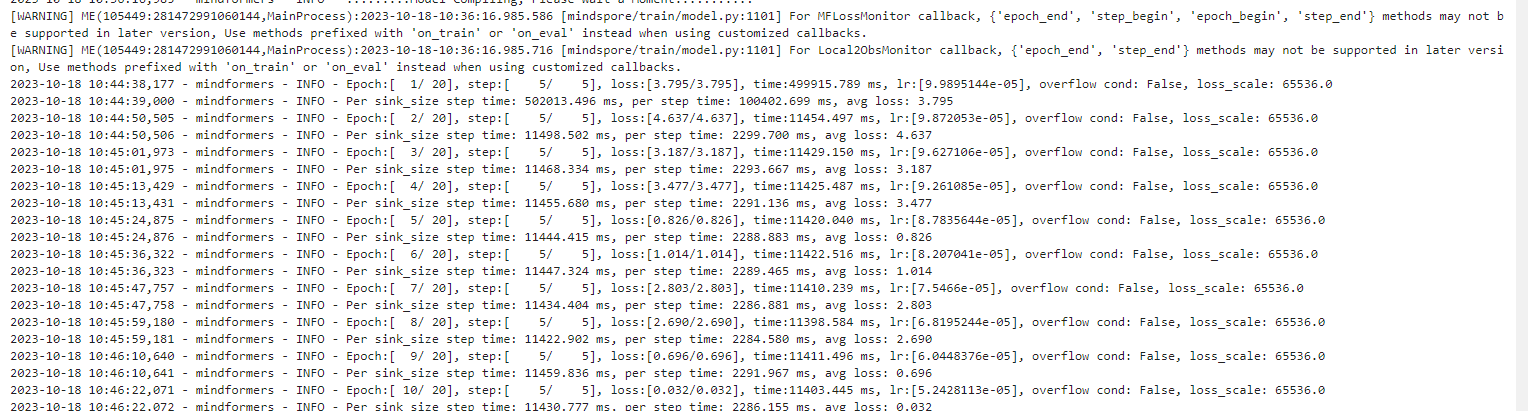
## 训练

使用如下命令开启训练：

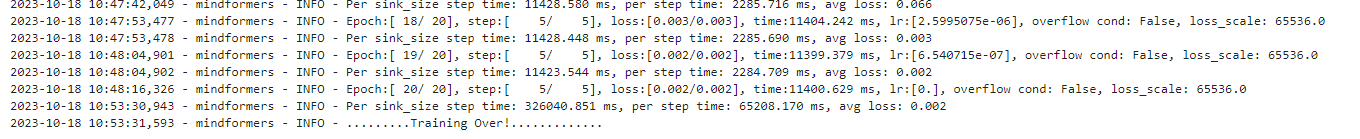
!cd llama\_lab/mindformers/scripts/ && bash run\_standalone.sh ../configs/llama/run\_llama\_7b\_lora.yaml 0 finetune

Ps：训练时间约为20mins

输出如下表明训练正常：



输出如下表示训练结束：



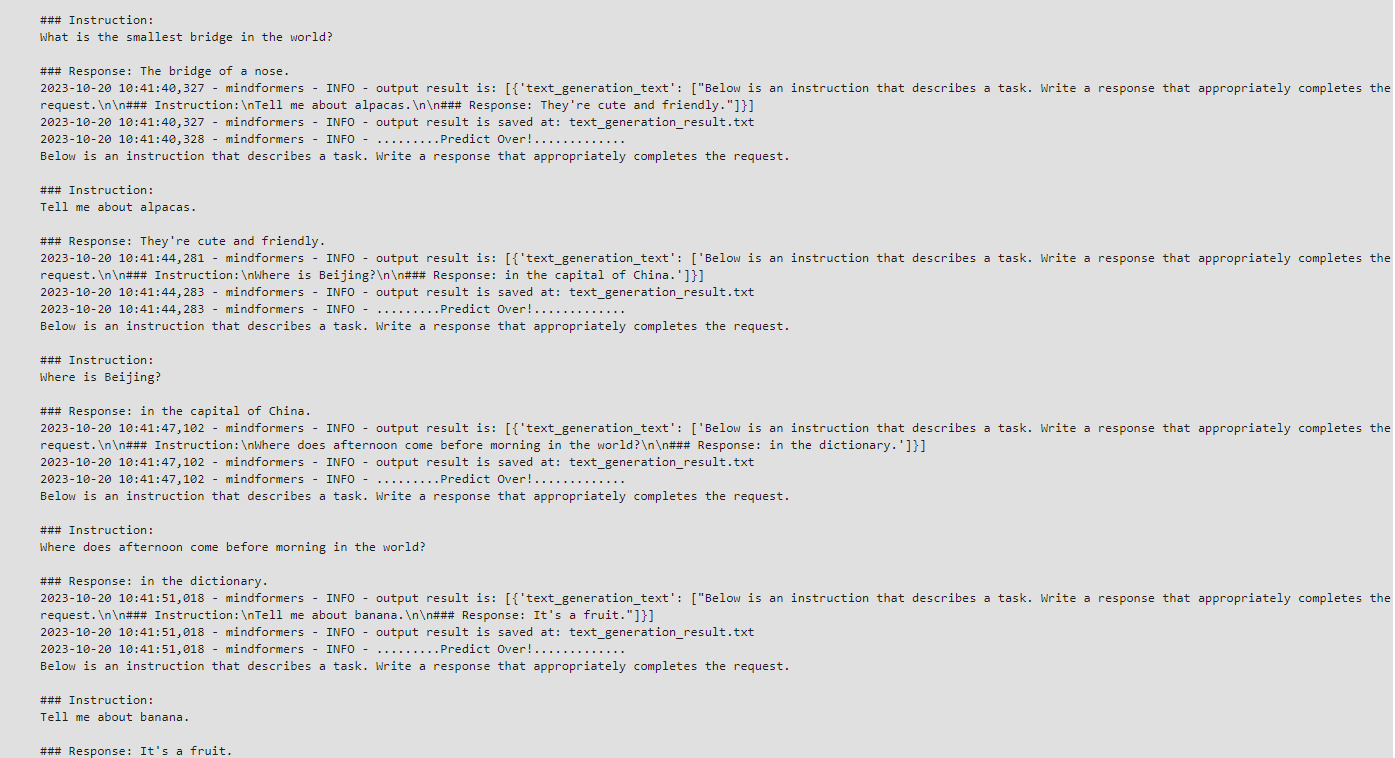
输出模型在llama\_lab/mindformers/scripts/mf\_standalone/output/checkpoint/rank\_0/ llama\_7b\_lora\_rank\_0-20\_5.ckpt。

## 使用微调后的模型推理

使用本实验包准备好的推理脚本（可以根据自己感兴趣的来提问，可修改mindformers下的lora\_infer.py）

!cd llama\_lab/mindformers && python lora\_infer.py

推理结果可显示在学得新的知识以外，也保留了之前的专家级语言能力（之前的能力会有所下降），如图：



# 课后思考题

## 思考题一

根据alpaca\_data.json修改对应的数据，重新微调训练，把llama7b模型做成智慧校园回答机器人。

示例：

Q：如何充值校园卡？

A: 1、打开缴费平台首页,点击校内用户 2、输入学工号、密码、验证码点击登录 3、选择对应的缴费项 4、确认缴费名称和缴费金额，然后点击去结算 5、确认缴费金额,选择支付方式，点击立即支付等等。

## 思考题二

在llama\_lab/mindformers/research文件夹下面有其他主流的大语言模型，可以尝试去运行推理和微调。

## 思考题三

尝试修改yaml文件里的其他参数（如optimizer,lr schedule,etc）,看能否达到其他更好的效果。