## 模型选用

一下几个模型都是跑过demo的，最终用了 LLaMA3-8B-Chinese-Chat 在 GJB5000B数据集上做了微调。

| **模型名** | **模型大小** | **Template** |
| --- | --- | --- |
| [LLaMA-3](https://gitee.com/link?target=https%3A%2F%2Fhuggingface.co%2Fmeta-llama) | 8B/70B | llama3 |
| LLaMA3-8B-Chinese-Chat | 8B | llama3 |
| [Qwen](https://gitee.com/link?target=https%3A%2F%2Fhuggingface.co%2FQwen) | 1.8B/7B/14B/72B | qwen |
| [Qwen1.5 (Code/MoE)](https://gitee.com/link?target=https%3A%2F%2Fhuggingface.co%2FQwen) 没下载 | 0.5B/1.8B/4B/7B/14B/32B/72B/110B | qwen |
| [Qwen2 (MoE)](https://gitee.com/link?target=https%3A%2F%2Fhuggingface.co%2FQwen) 没下载 | 0.5B/1.5B/7B/57B/72B | qwen |

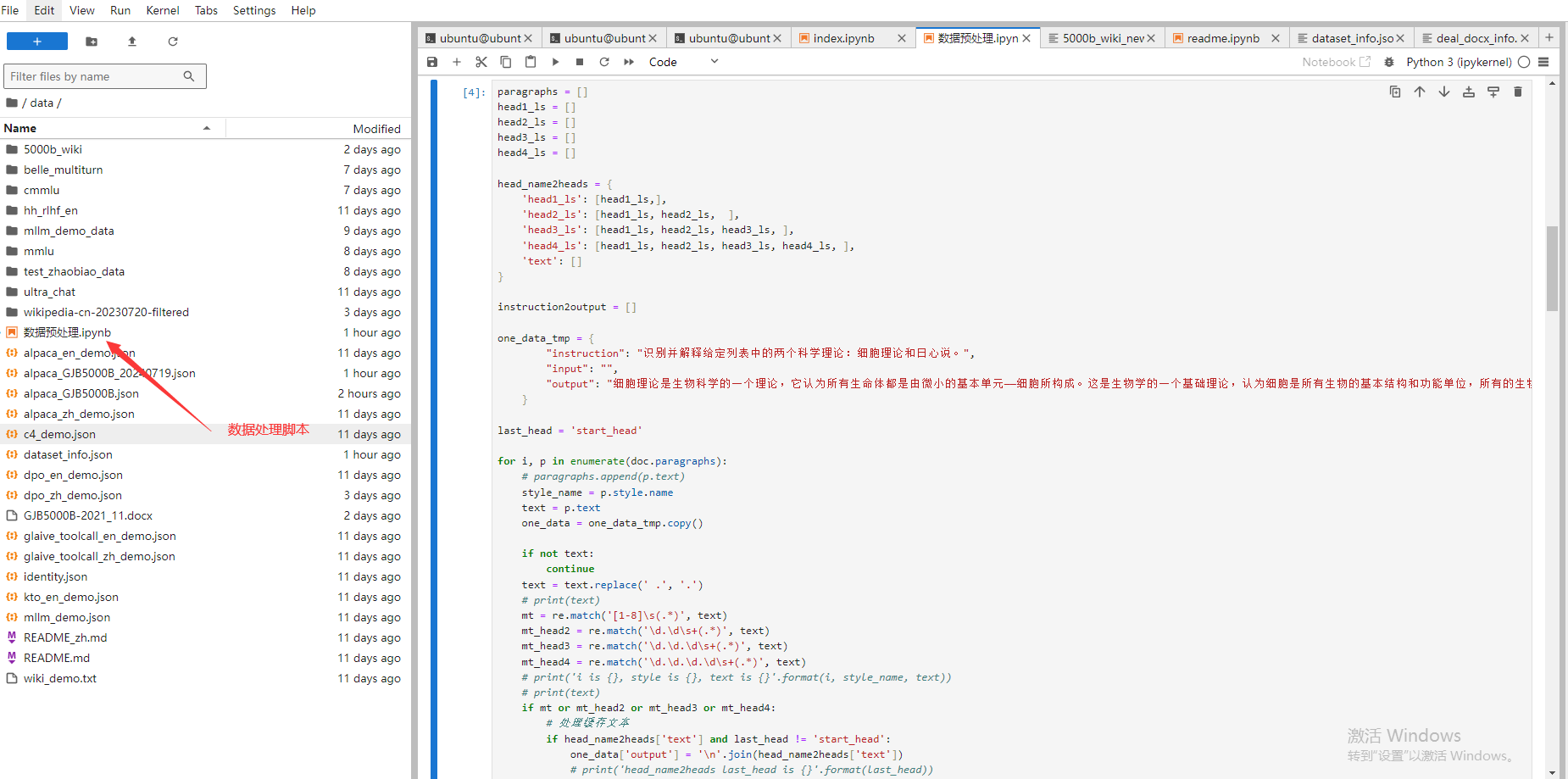
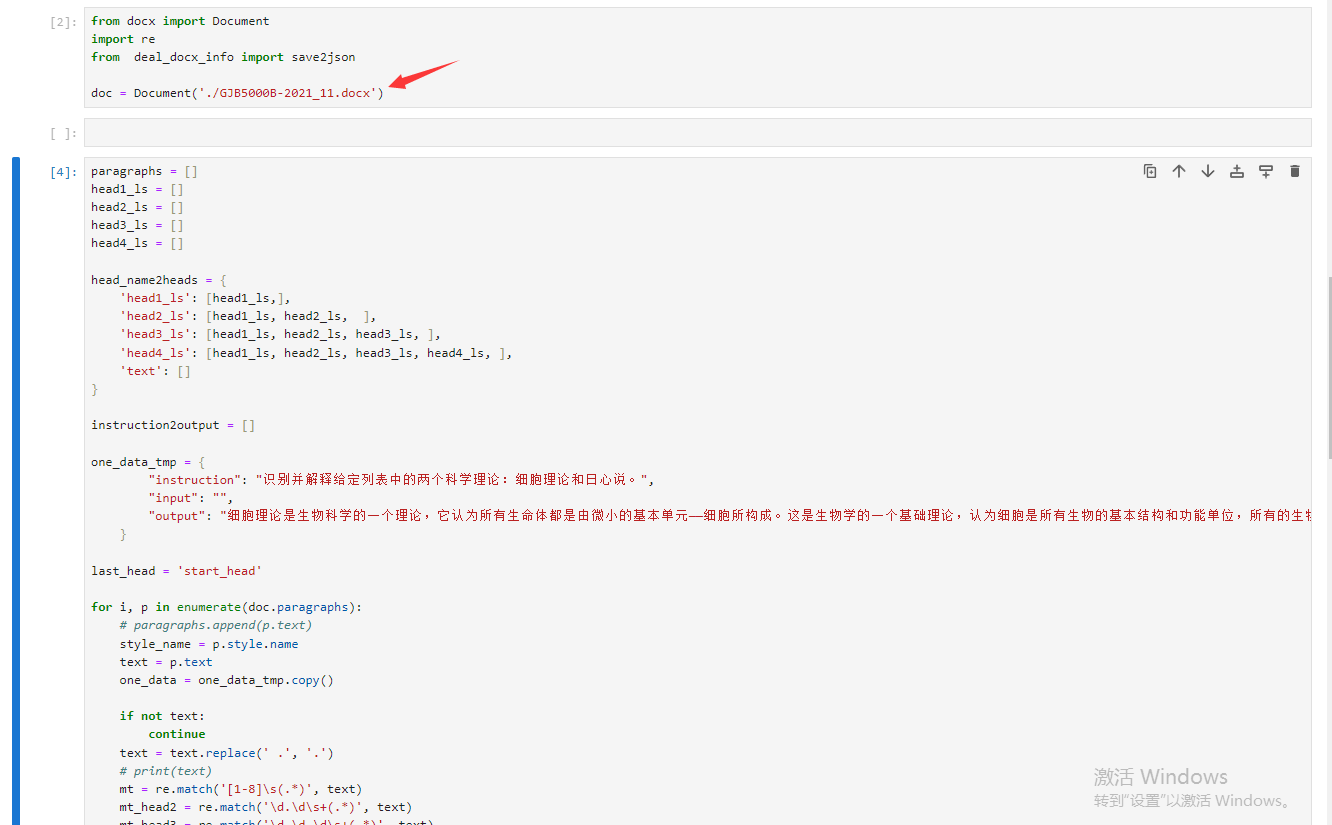
## 微调方法

| 方法 | LoRA | QLoRA | 微调模型 | 微调结果路径 |
| --- | --- | --- | --- | --- |
| Pre-Trainnig | 是 | 没有 | LLaMA3-8B-Chinese-Chat |  |
| Supervised Fine-Tuning | 是 | 没有 | LLaMA3-8B-Chinese-Chat |  |
| Pre-Trainnig | 是 | 没有 | Qwen-7B-Chat |  |
| Supervised Fine-Tuning | 是 | 没有 | Qwen-7B-Chat |  |
|  |  |  |  |  |

## 数据集处理

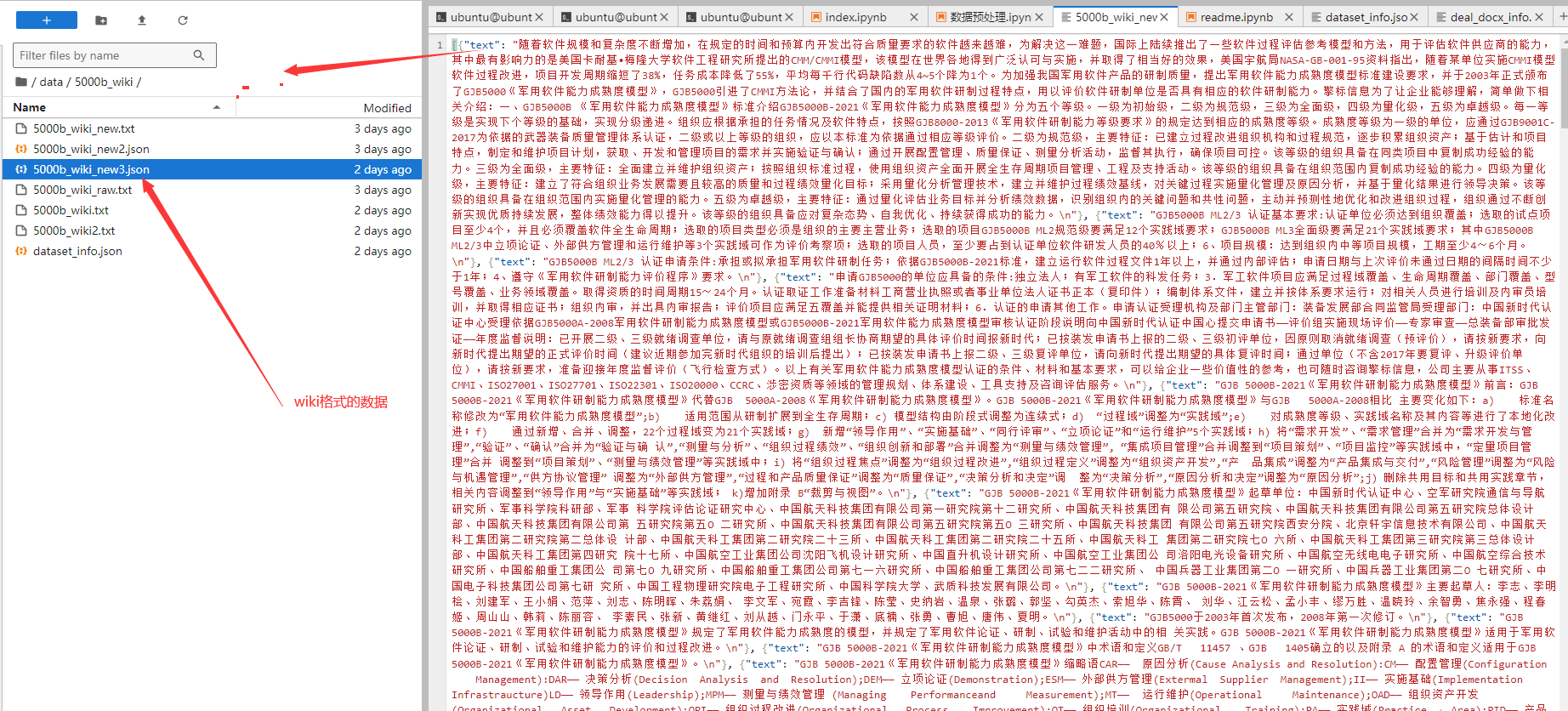
通过解析word文档，组织QA对，

数据预处理脚本

### Wiki 格式的数据

给 Pre-Trainnig 使用的数据集



### alpaca 格式数据

该数据集给 Supervised Fine-Tuning 使用

5000B材料整理后的数据集，



## 使用指定数据集微调

# 激活虚拟环境，，

conda activate llmfactory2

# 端口随意指定

export GRADIO\_SERVER\_PORT=7871

# 模型下载仓库

export USE\_MODELSCOPE\_HUB=1

# 指定第二个显卡

export CUDA\_VISIBLE\_DEVICES=1

# LoRA **微调**

llamafactory-cli train examples/lora\_single\_gpu/llama3\_lora\_sft.yaml

# 测试推理demo

llamafactory-cli chat examples/inference/llama3\_lora\_sft.yaml

# 不用yaml文件的微调

llamafactory-cli train \

--stage sft \

--do\_train True \

--model\_name\_or\_path ./models/Llama3-8B-Chinese-Chat \

--preprocessing\_num\_workers 16 \

--finetuning\_type lora \

--quantization\_method bitsandbytes \

--template llama3 \

--flash\_attn auto \

--dataset\_dir data \

--dataset alpaca\_GJB5000B\_20240719 \

--cutoff\_len 1024 \

--learning\_rate 5e-05 \

--num\_train\_epochs 5.0 \

--max\_samples 1000 \

--per\_device\_train\_batch\_size 4 \

--gradient\_accumulation\_steps 4 \

--lr\_scheduler\_type cosine \

--max\_grad\_norm 1.0 \

--logging\_steps 5 \

--save\_steps 100 \

--warmup\_steps 0 \

--optim adamw\_torch \

--packing False \

--report\_to none \

--output\_dir saves/LLaMA3-8B-Chinese-Chat/lora/train\_alpaca\_GJB5000B\_20240719 \

--bf16 True \

--plot\_loss True \

--ddp\_timeout 180000000 \

--include\_num\_input\_tokens\_seen True \

--lora\_rank 8 \

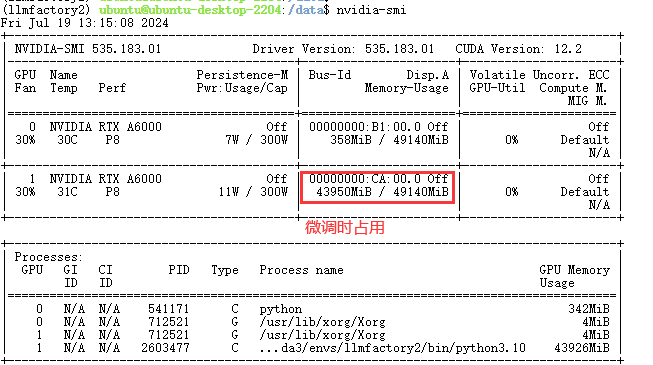
--lora\_alpha 16 \

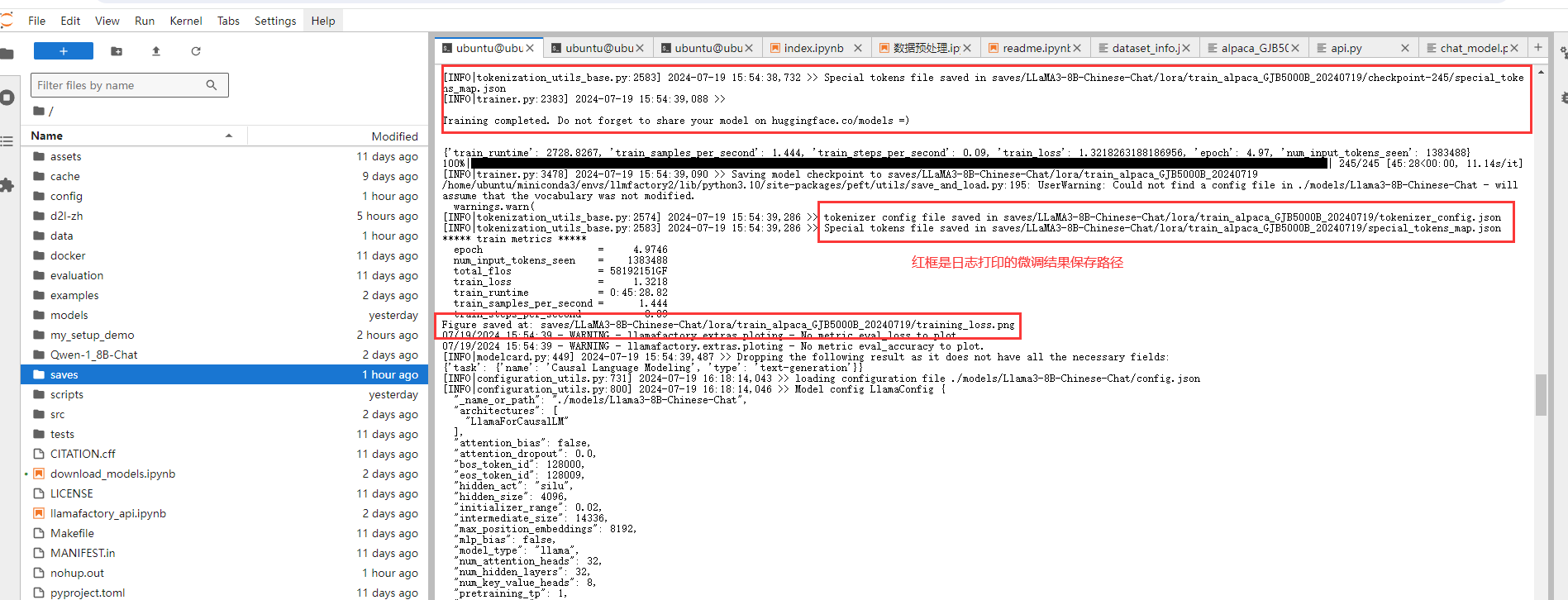
--lora\_dropout 0 \

--use\_dora True \

--lora\_target all

微调时占用





## 微调前后对比





## 微调接口开放

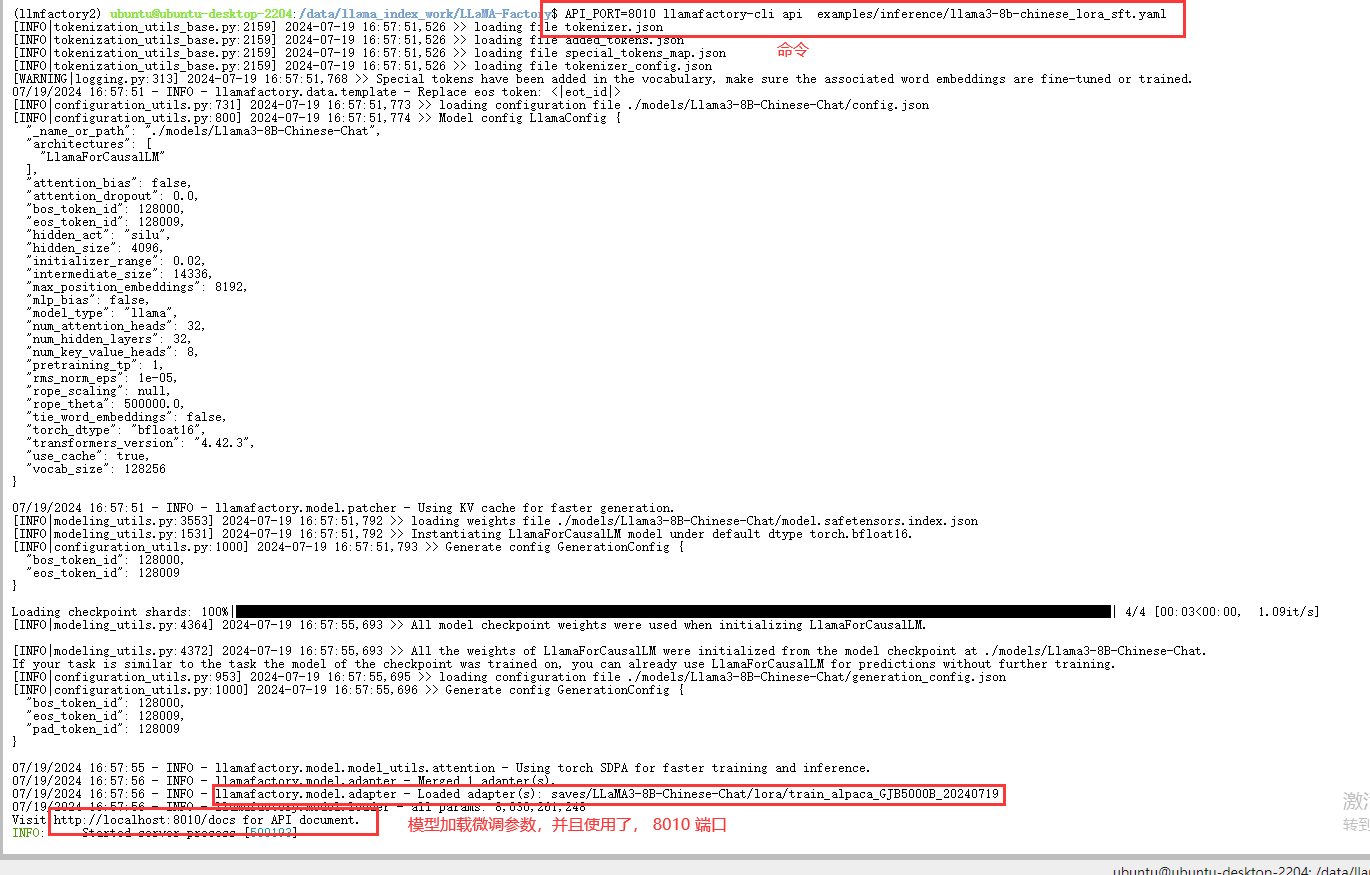
# 指定第二个显卡

export CUDA\_VISIBLE\_DEVICES=1

export API\_PORT=8010

# 还有些其他参数，

llamafactory-cli api examples/inference/llama3\_vllm.yaml



### 大模型微调相关接口

#### 接口说明：GJB 5000B微调模型对话接口

接口功能：GJB 5000B

接口请求地址：

http://10.0.1.40:8010/v1/chat/completions

**请求头 :**

| 请求头 | 请求内容 | 说明 |
| --- | --- | --- |
| Content-Type | application/json | 请求方式 |
| **accept** | **application/json** |  |

**请求方式:  POST**

**参数类型** ：**JSON**

#### 请求示例：

**curl -X 'POST' \**

**'http://10.0.1.40:8010/v1/chat/completions' \**

**-H 'accept: application/json' \**

**-H 'Content-Type: application/json' \**

**-d '{**

**"model": "./models/Qwen-7B-Chat",**

**"messages": [**

**{**

**"role": "user",**

**"content": "GJB 5000B开发库"**

**}**

**],**

**"do\_sample": true,**

**"temperature": 0.95,**

**"top\_p": 0.1,**

**"n": 1,**

**"max\_tokens": 1024,**

**"stream": false**

**}'**

#### 请求参数说明

| 字段名 | 字段说明 | 字段类型 | 是否必填 |
| --- | --- | --- | --- |
| model | 模型名称 | string | 是 |
| messages | 信息 | list | 是 |
| role | 角色, user 和 system, | string | 是 |
| **content** | 用户输入 | string | 是 |
| **do\_sample** | 这个字段一般不动 | bool | 是 |
| **temperature** | 这个字段一般不动 | float | 是 |
| **top\_p** | 这个字段一般不动 | float | 是 |
| **n** | 这个字段一般不动 | float | 是 |
| **max\_tokens** | 这个字段一般不动,简单理解为最大输出长度 | float | 是 |
| **stream** | 这个字段一般不动 | bool | 是 |
|  |  |  |  |

#### 响应示例

**{**

**"id": "chatcmpl-0779827da4584342b63fa3611e1f8b52",**

**"object": "chat.completion",**

**"created": 1721208830,**

**"model": "./models/Qwen-7B-Chat",**

**"choices": [**

**{**

**"index": 0,**

**"message": {**

**"role": "assistant",**

**"content": "GJB 5000B是中国军用软件工程标准，是军用软件开发、测试、维护和保障的通用标准。它规定了军用软件的生命周期管理过程，包括需求分析、设计、编码、测试、维护和保障等阶段。GJB 5000B还规定了军用软件的质量保证和软件工程管理的要求。",**

**"tool\_calls": null**

**},**

**"finish\_reason": "stop"**

**}**

**],**

**"usage": {**

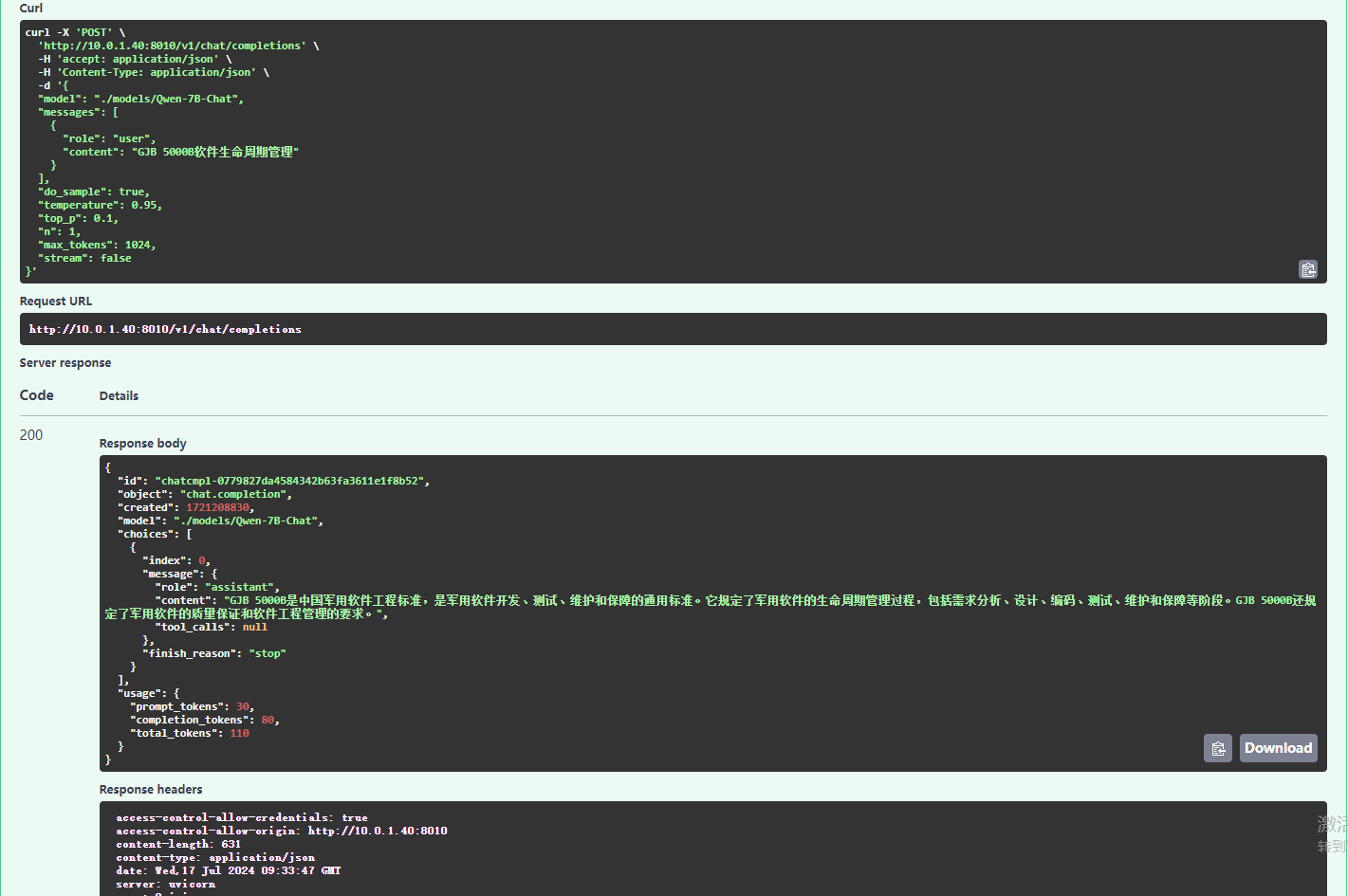
**"prompt\_tokens": 30,**

**"completion\_tokens": 80,**

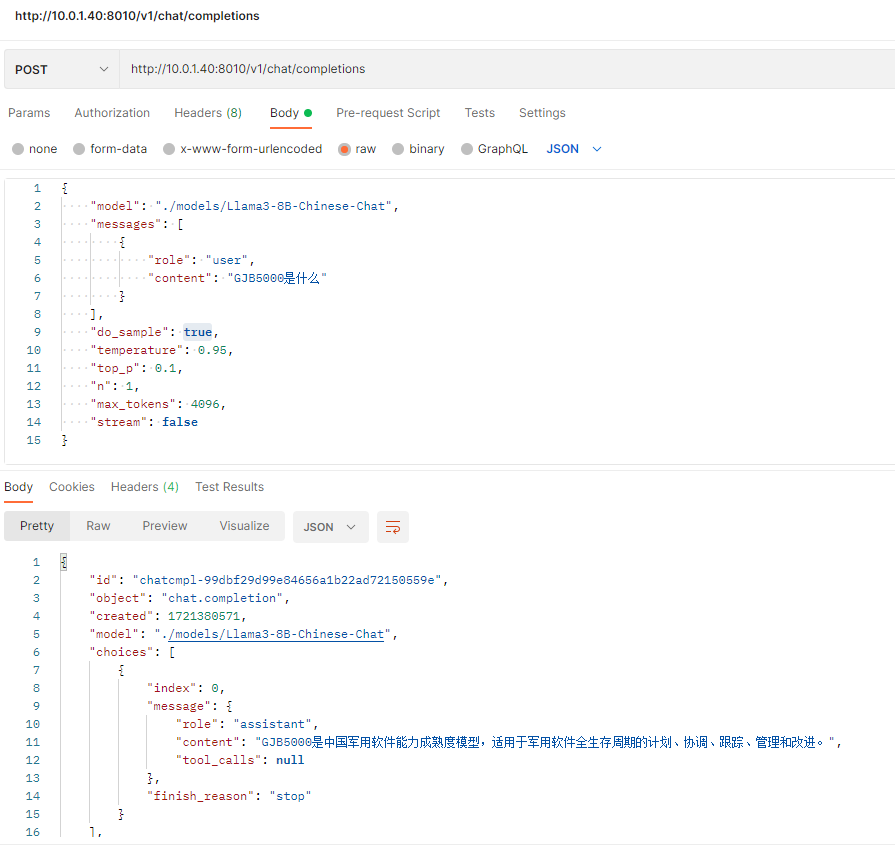
**"total\_tokens": 110**

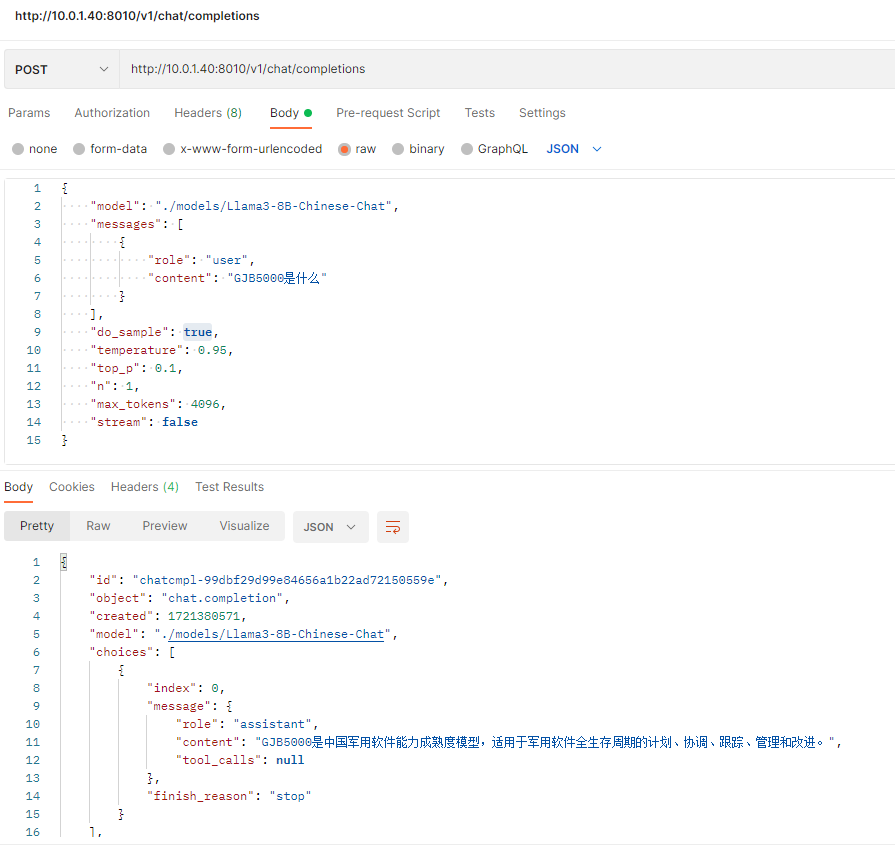
**}**

**}**



使用postman请求





#### 接口说明：llama3微调模型对话接口

接口功能：llama3 原始接口

接口请求地址：

**http://10.0.1.40:7575/original\_llama3**

**请求头 :**

| 请求头 | 请求内容 | 说明 |
| --- | --- | --- |
| Content-Type | application/json | 请求方式 |
| **accept** | **application/json** |  |

**请求方式:  POST**

**参数类型** ：**JSON**

#### 请求示例：

**curl -X 'POST' \**

**'http://10.0.1.40:7575/original\_llama3' \**

**-H 'accept: application/json' \**

**-H 'Content-Type: application/json' \**

**-d '{**

**"model": "./Llama3-8B-Chinese-Chat/",**

**"question": "你好，什么是傻瓜"**

**}'**

#### 请求参数说明

| 字段名 | 字段说明 | 字段类型 | 是否必填 |
| --- | --- | --- | --- |
| model | 模型名称 | string | 是 |
| question | 问题 | string | 是 |

#### 响应示例

**{**

**"response": "在中文里，“傻瓜”这个词的含义通常是指没有智力，或者说非常愚蠢。但是，在不同的文化和历史背景下，“傻瓜”这个词也可能有不同的含义。\n\n在技术领域，像“傻瓜机”（Dumb Terminal）这样的术语则有一个不同含义。傻瓜机指的是一个不具备处理功能的电脑终端，只能显示信息和接受用户的输入，而所有的处理工作都在远程服务器上完成。这种设备主要用于远程访问和控制大型计算机系统。\n\n如果您有任何特定的工作问题或需要进一步的帮助，请告诉我，我会尽力为您提供帮助。"**

**}**

