

第四次作业：大语言模型部署体验

一、项目目的

本项目的主要目标是评估 Qwen-7B-Chat 和 ChatGLM3-6B 两个模型在处理复杂问答任务中的表现。我们设计了五个问答测试，涵盖了对比句子差异、理解重复语义、复杂句子逻辑分析、具体对话含义解释等多个方面。通过这些测试，我们希望明确两个模型在不同场景下的优劣。

二、环境搭建

启动modelscope社区服务器，打开终端

执行以下命令检测是否联网：

```
pip install -U pip setuptools wheel
```

安装环境依赖：

```
pip install \ "intel-extension-for-transformers==1.4.2" \ "neural-compressor==2.5" \  
"transformers==4.33.3" \ "modelscope==1.9.5" \ "pydantic==1.10.13" \ "sentencepiece"  
\ "tiktoken" \ "einops" \ "transformers_stream_generator" \ "uvicorn" \ "fastapi" \  
"yacs" \ "setuptools_scm"
```

```
pip install fschat--use-pep517
```

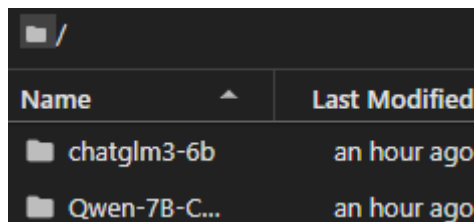
转到数据目录

```
cd /mnt/data
```

下载智普清言和通义千问大模型

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

```
git clone https://www.modelscope.cn/qwen/Qwen-7B-Chat.git
```



Name	Last Modified
chatglm3-6b	an hour ago
Qwen-7B-C...	an hour ago

如图为安装完成

切换到工作目录

```
cd /mnt/workspace
```

编写脚本 run_ChatGLM.py:

```
from transformers import TextStreamer, AutoTokenizer, AutoModelForCausalLM  
model_name = "/mnt/data/chatglm3-6b" # 本地路径  
prompt = "请回答如下三个问题：1.请说说冬天能穿多少穿多少和夏天能穿多少穿多少这两句话的区别？2.请说出以下两句话区别在哪里？单身狗产生的原因有两个，一是谁都看不上，二是谁都看不上。3.明明明明白白白白喜欢他，可她就是不说话。这句话里，明明和白白谁喜欢谁？"  
tokenizer = AutoTokenizer.from_pretrained(  
model_name,  
trust_remote_code=True
```

```

)
model = AutoModelForCausalLM.from_pretrained(
model_name,
trust_remote_code=True,
torch_dtype="auto" # 自动选择 float32/float16 (根据模型配置)
).eval()
inputs = tokenizer(prompt, return_tensors="pt").input_ids
streamer = TextStreamer(tokenizer)
outputs = model.generate(inputs, streamer=streamer, max_new_tokens=300)

```

编写脚本 run_Qwen.py:

```

from transformers import TextStreamer, AutoTokenizer, AutoModelForCausalLM
model_name = "/mnt/data/Qwen-7B-Chat" # 本地路径
prompt = "请回答如下三个问题：1.请说说冬天能穿多少穿多少和夏天能穿多少穿多少这两句话的区别？2.请说出以下两句话区别在哪里？单身狗产生的原因有两个，一是谁都看不上，二是谁都看不上。3.明明明明明白白白喜欢他，可她就是不说。这句话里，明明和白白谁喜欢谁？"
tokenizer = AutoTokenizer.from_pretrained(
model_name,
trust_remote_code=True
)
model = AutoModelForCausalLM.from_pretrained(
model_name,
trust_remote_code=True,
torch_dtype="auto" # 自动选择 float32/float16 (根据模型配置)
).eval()
inputs = tokenizer(prompt, return_tensors="pt").input_ids
streamer = TextStreamer(tokenizer)
outputs = model.generate(inputs, streamer=streamer, max_new_tokens=300)

```

三、运行实例

使用

```
python filename.py
```

指令可以运行python脚本，期中filename处为文件名。

可以看到代码中，我们分别问了两个模型三个问题：

- 1.请说说冬天能穿多少穿多少和夏天能穿多少穿多少这两句话的区别？
- 2.请说出以下两句话区别在哪里？单身狗产生的原因有两个，一是谁都看不上，二是谁都看不上。
- 3.明明明明明白白白喜欢他，可她就是不说。这句话里，明明和白白谁喜欢谁？

chatGLM运行结果：

```

[MacBook-Pro ~]$ python run_qwen.py
请回答如下三个问题：1.请说说冬天能穿多少穿多少和夏天能穿多少穿多少这两句话的区别？2.请说出以下两句话区别在哪里？单身狗产生的原因有两个，一是谁都看不上，二是谁都看不上。3.明明明明明白白白喜欢他，可她就是不说。这句话里，明明和白白谁喜欢谁？
1. 冬天能穿多少穿多少和夏天能穿多少穿多少这两句话的区别在于，冬天强调的是保暖，而夏天强调的是防晒和透气。因此，冬天穿衣要穿得厚一些，多穿一些，以保持身体的温暖；而夏天穿衣要穿得薄一些，少穿一些，以保持身体的透气性和舒适感。
2. 单身狗产生的原因有两个，一是谁都看不上，二是谁都看不上。这两句话的区别在于，第一句“谁都看不上我”强调的是个人情感上的自卑和不安，而第二句“谁都看不上我”则更强调的是个人在社交上的不被认可和接纳。
3. 在这句话里，明明喜欢白白，但白白喜欢的是明明。明明明白白白喜欢他，可她就是不说，这表明明明对白白有感情，但白白并没有对明明有同样的感情。

```

1. 冬天能穿多少穿多少和夏天能穿多少穿多少这两句话的区别在于，冬天强调的是保暖，而夏天强调的是防晒和透气。因此，冬天穿衣服要穿得厚一些，多穿一些，以保持身体的温暖；而夏天穿衣服要穿得薄一些，少穿一些，以保持身体的透气性和舒适感。
2. 单身狗产生的原因有两个，一是谁都看不上，二是谁都看不上。这两句话的区别在于，第一句“谁能看上我”强调的是个人情感上的自卑和不安，而第二句“谁看不上我”则更强调的是个人在社交上的不被认可和接纳。
3. 在这句话里，明明喜欢白白，但白白喜欢的是明明。明明白白白白喜欢他，可她就是不说不说，这表明明明对白白有感情，但白白并没有对明明有同样的感情。

Qwen运行结果：

```
请回答如下三个问题：1. 请说说冬天能穿多少穿多少和夏天能穿多少穿多少这两句话的区别？2. 请说出以下两句话区别在哪里？单身狗产生的原因有两个，一是谁都看不上，二是谁都看不上。3. 明明明明白白白白喜欢他，可她就是不说不说。这句话里，明明和白白谁喜欢谁？<[im_start]>
1. “冬天能穿多少穿多少”和“夏天能穿多少穿多少”的主要区别在于描述的季节不同。前者是针对冬天的情况，后者则是针对夏天的情况。在寒冷的冬季，我们需要穿更多的衣物来保暖，而在炎热的夏季，我们则需要穿更少的衣物来散热。
2. “单身狗产生的原因有两个，一是谁都看不上，二是谁都看不上”这句话的前半句和后半句说的是同样的意思，只是将主语换成了“单身狗”。因此，这句话并没有明显的差别。
3. 在句子“明明明明白白白白喜欢他，可她就是不说不说”中，“明明”是指说话的人，而“白白”则表示喜欢的对象。所以，可以理解为，明明知道自己非常喜欢他，但就是不愿意说出来。<[im_end]>
</endofText>
```

1. "冬天能穿多少穿多少"和"夏天能穿多少穿多少"的主要区别在于描述的季节不同。前者是针对冬天的情况，后者则是针对夏天的情况。在寒冷的冬季，我们需要穿更多的衣物来保暖，而在炎热的夏季，我们则需要穿更少的衣物来散热。
2. "单身狗产生的原因有两个，一是谁都看不上，二是谁都看不上"这句话的前半句和后半句说的是同样的意思，只是将主语换成了"单身狗"。因此，这句话并没有明显的差别。
3. 在句子 "明明明明白白白白喜欢他，可她就是不说不说" 中，"明明" 是指说话的人，而"白白" 则表示喜欢的对象。所以，可以理解为，明明知道自己非常喜欢他，但就是不愿意说出来。

结果分析：

对于第一个问题，两个模型都回答的很准确

对于第二个问题，ChatGLM和Qwen的回答都不对，但是从Qwen的回答中可以看出好像Qwen理解到了题目的意思，即主语不同。

对于第三个问题，两个模型都解答错误了。