大语言模型部署项目文档

本次项目需要进行大语言模型部署实验。项目包括以下操作：

（1）登录并使用魔搭平台，注册时需要关联阿里云账号来获得免费的CPU云计算资源；

（2）通过Jupyter Notebook或者相应环境镜像进入相应的项目部署环境，根据相应模型的部署文档完成模型的部署；

（3）针对2个不同的模型进行一些应用场景的测试，并开展不同模型之间的横向对比。

项目链接在<https://github.com/yunhaishang/LLM-examination>

本次项目比较的模型有Qwen-7B和ChatGLM3-6B。

对于模型进行以下的问答测试：

（1）请说出以下两句话区别在哪里？ 1、冬天：能穿多少穿多少 2、夏天：能穿多少穿多少

（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）他知道我知道你知道他不知道吗？ 这句话里，到底谁不知道

（4）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？

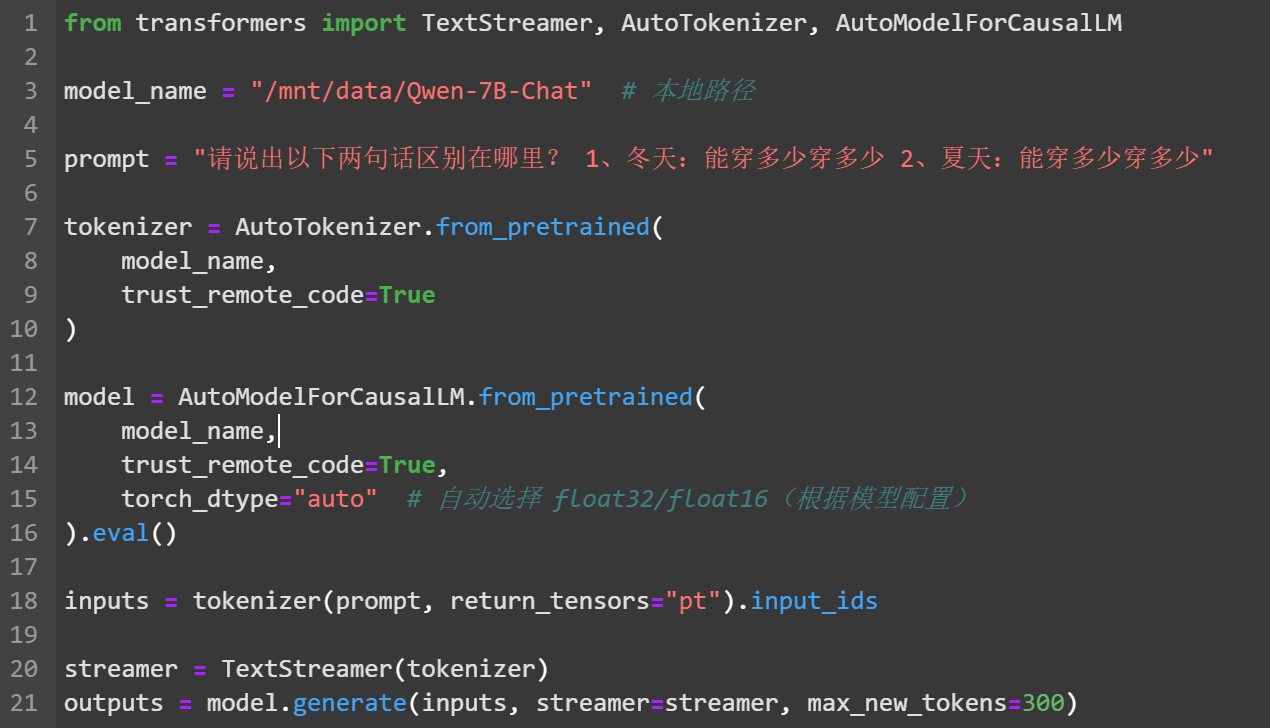
（5）领导：你这是什么意思？ 小明：没什么意思。意思意思。 领导：你这就不够意思了。 小明：小意思，小意思。领导：你这人真有意思。 小明：其实也没有别的意思。 领导：那我就不好意思了。 小明：是我不好意思。请问：以上“意思”分别是什么意思。

通义千问Qwen-7B-Chat

安装模型

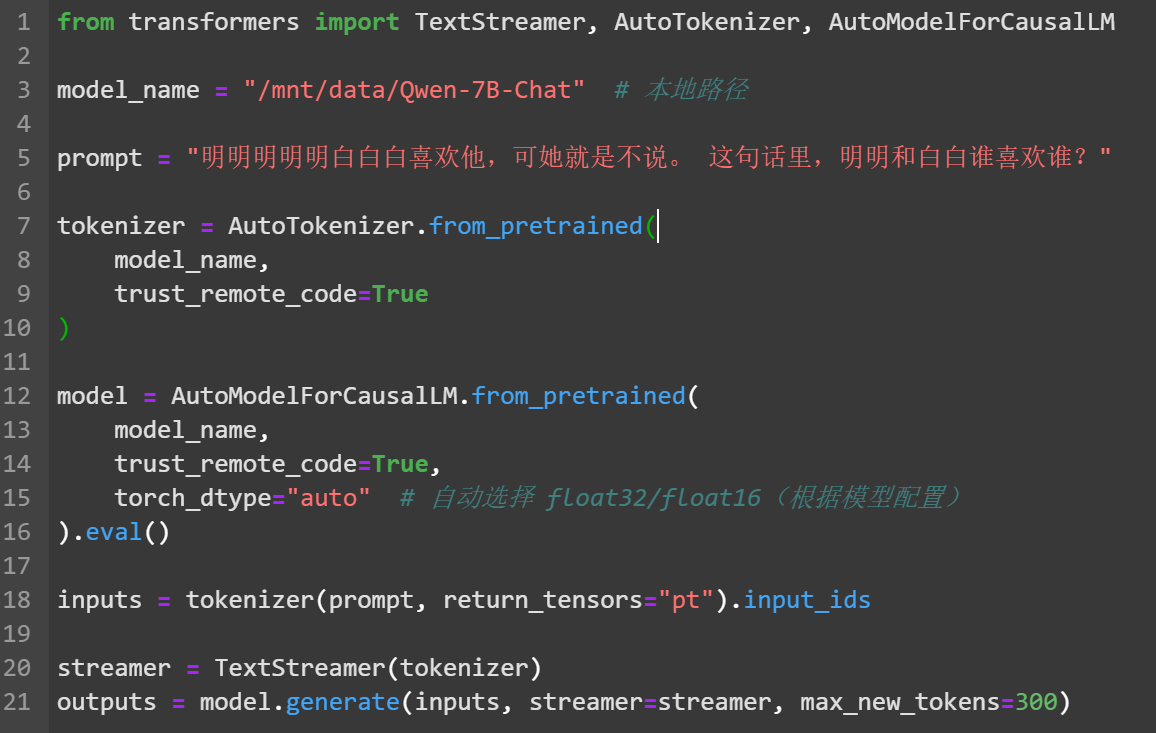


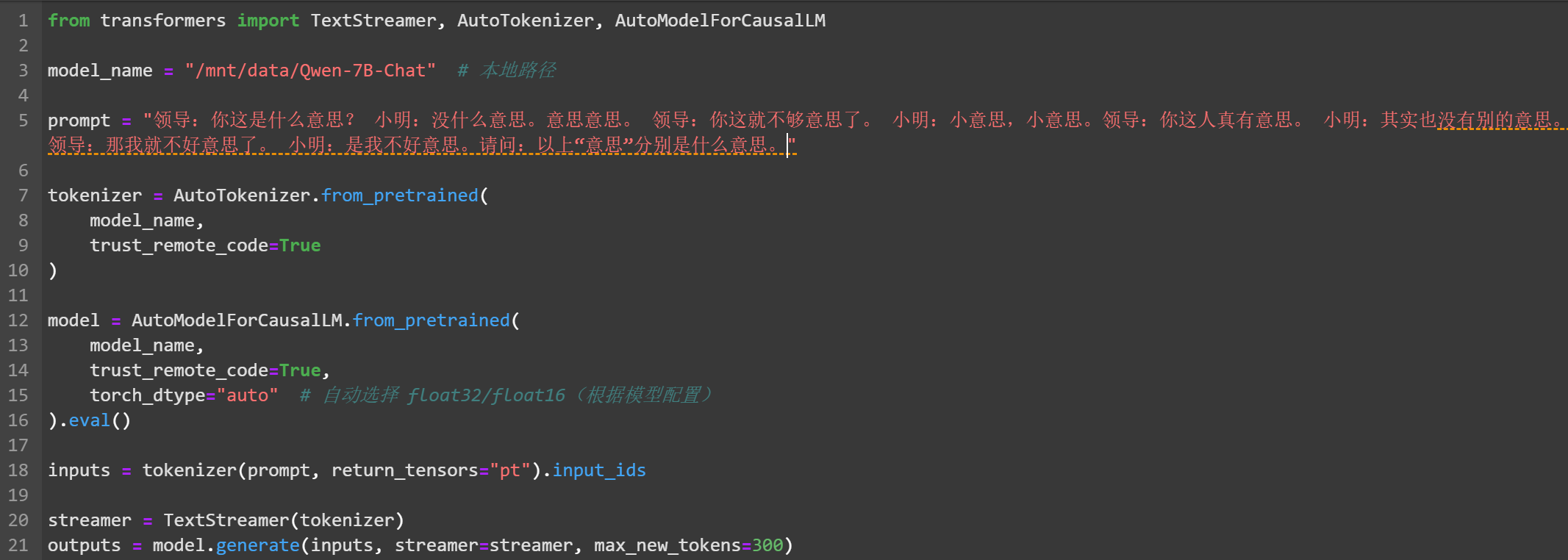
编写python脚本，修改 prompt即可进行不同的问答，脚本如下：





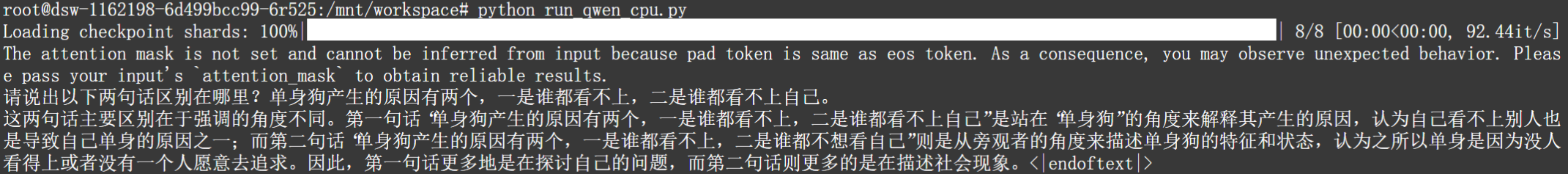


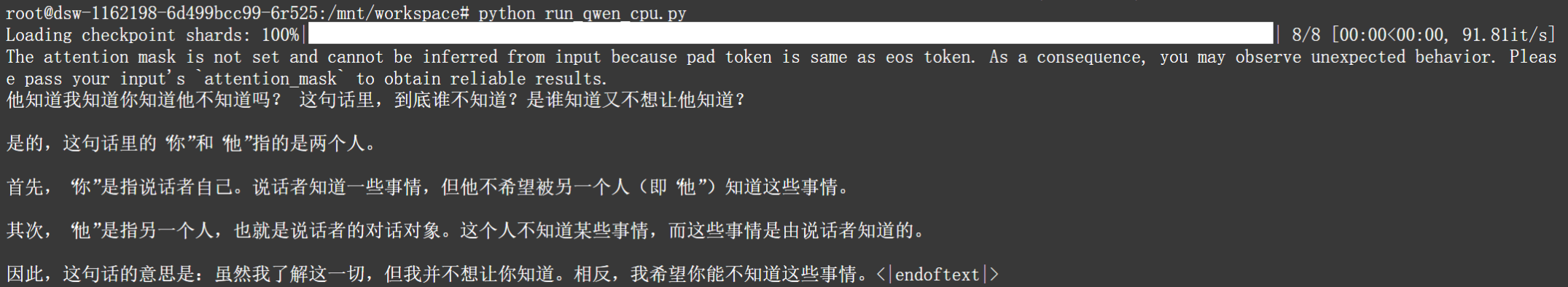




在终端输入python run\_qwen\_cpu.py，运行结果如下：





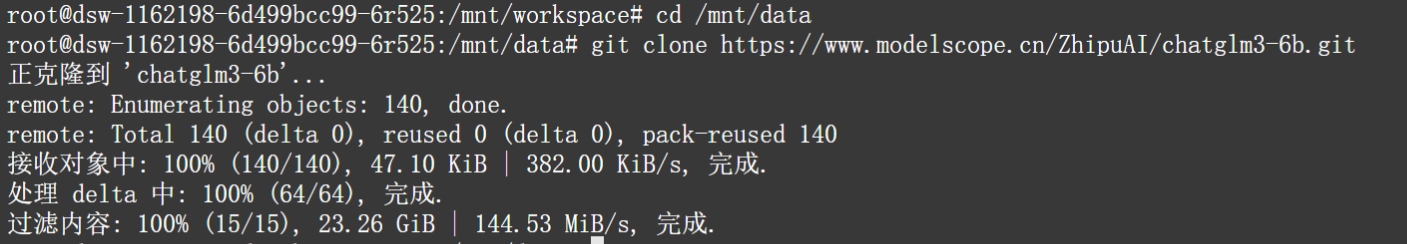




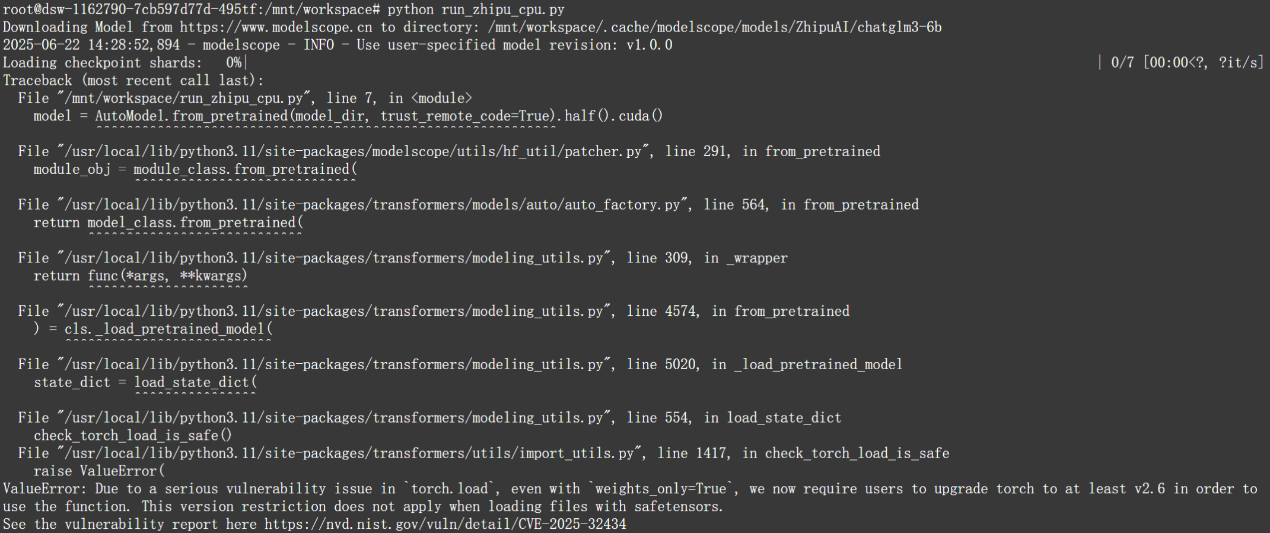


智谱清言ChatGLM3-6B

安装模型



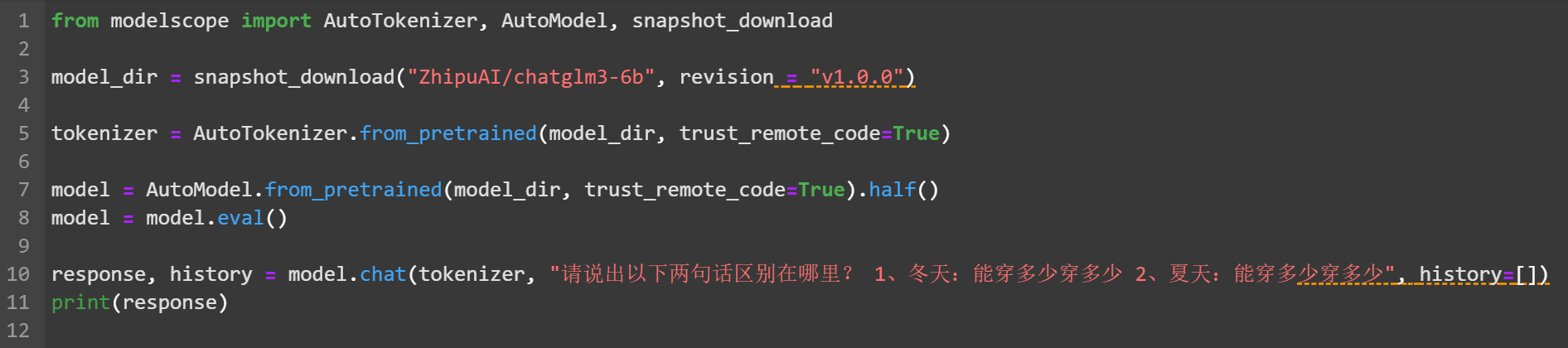
若直接使用，会显示版本问题



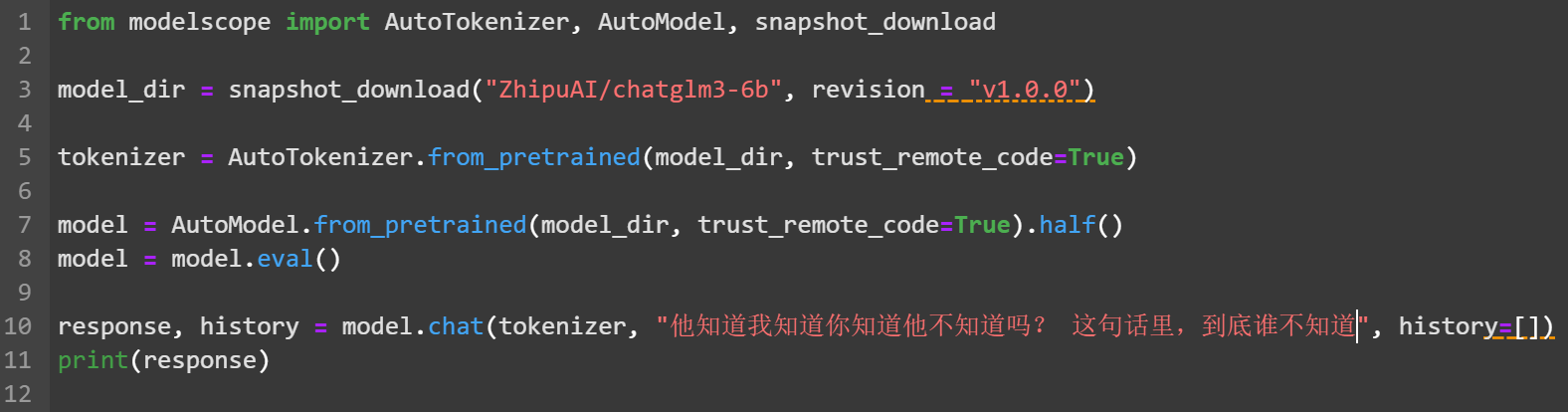
解决方法是下载transformers旧版本



编写python脚本，修改 model.chat 语句即可进行不同的问答，脚本如下：



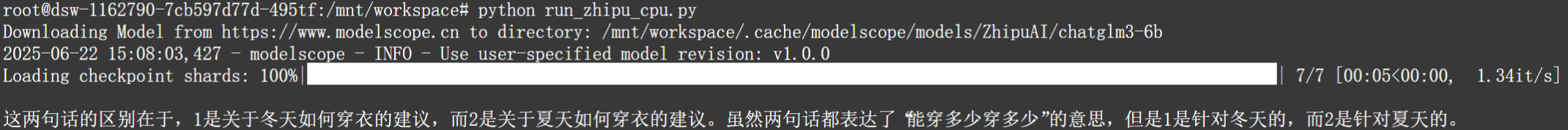


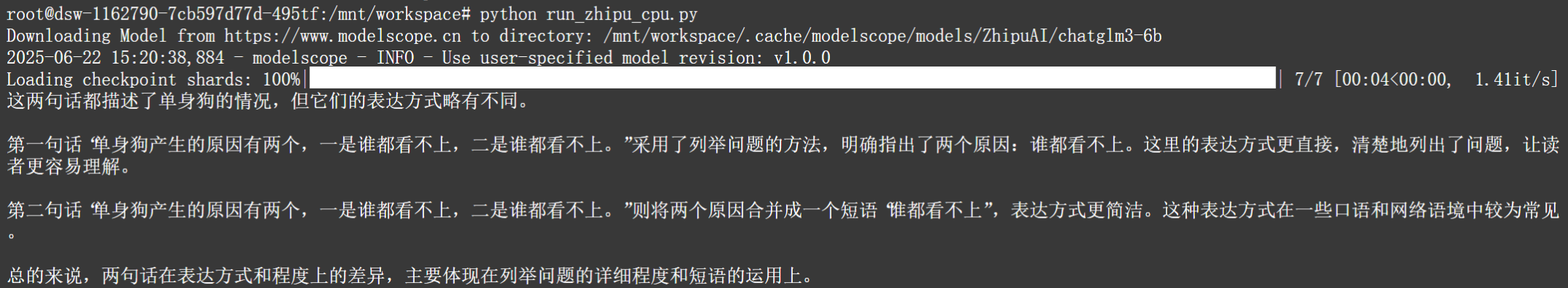


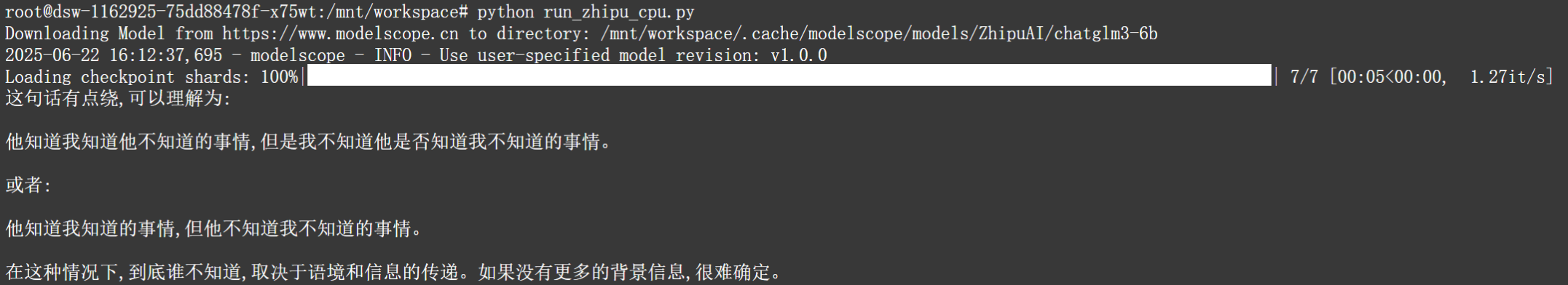


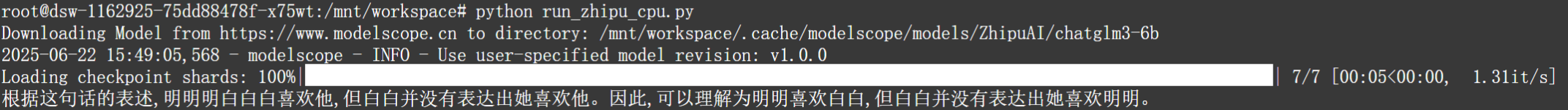


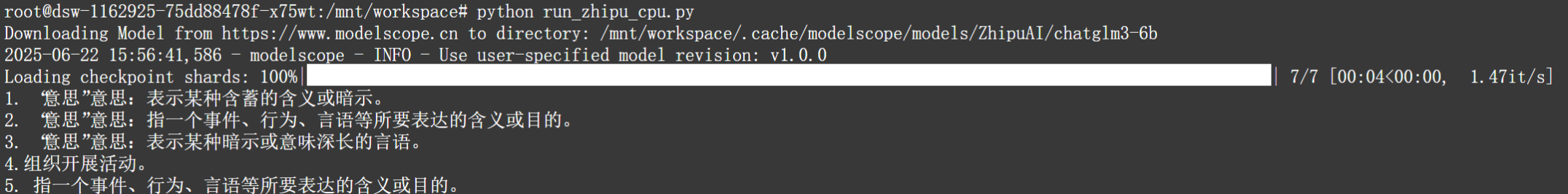
在终端输入python run\_zhipu\_cpu.py，运行结果如下：











对比分析

Qwen-7B会出现修改prompt的现象，比如把

“他知道我知道你知道他不知道吗？ 这句话里，到底谁不知道”  
变成

他知道我知道你知道他不知道吗？ 这句话里，到底谁不知道？是谁知道又不想让他知道？”，

这极大地影响了回答质量；但是对于没有修改prompt的情况，能答到“能穿多少穿多少“等的双关意，答不到“意思”的隐含意。

ChatGLM3-6B不会出现修改prompt的情况，但是它的分析比较答不到重点，体会不到双关意和隐含意，证明对于问题的理解力不够充足。

分析来看，Qwen-7B修改prompt的情况可能和脚本程序的编写有关，如果更换脚本程序的编写方法可能可以规避这个问题，而当没有出现修改prompt的问题时Qwen-7B可以回答出双关意，但是答不出隐藏意。ChatGLM3-6B不会出现修改prompt的问题，但是总体的回答比较答不到重点。