**2大语言模型部署体验报告**

1. 搭建环境

魔搭社区平台，阿里云服务器，Python3.1，Jupyter Notebook

1. 搭建方法
2. 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；
3. 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；
4. 搭建模型

千问模型，ChatGLM3-b6两种大语言模型。

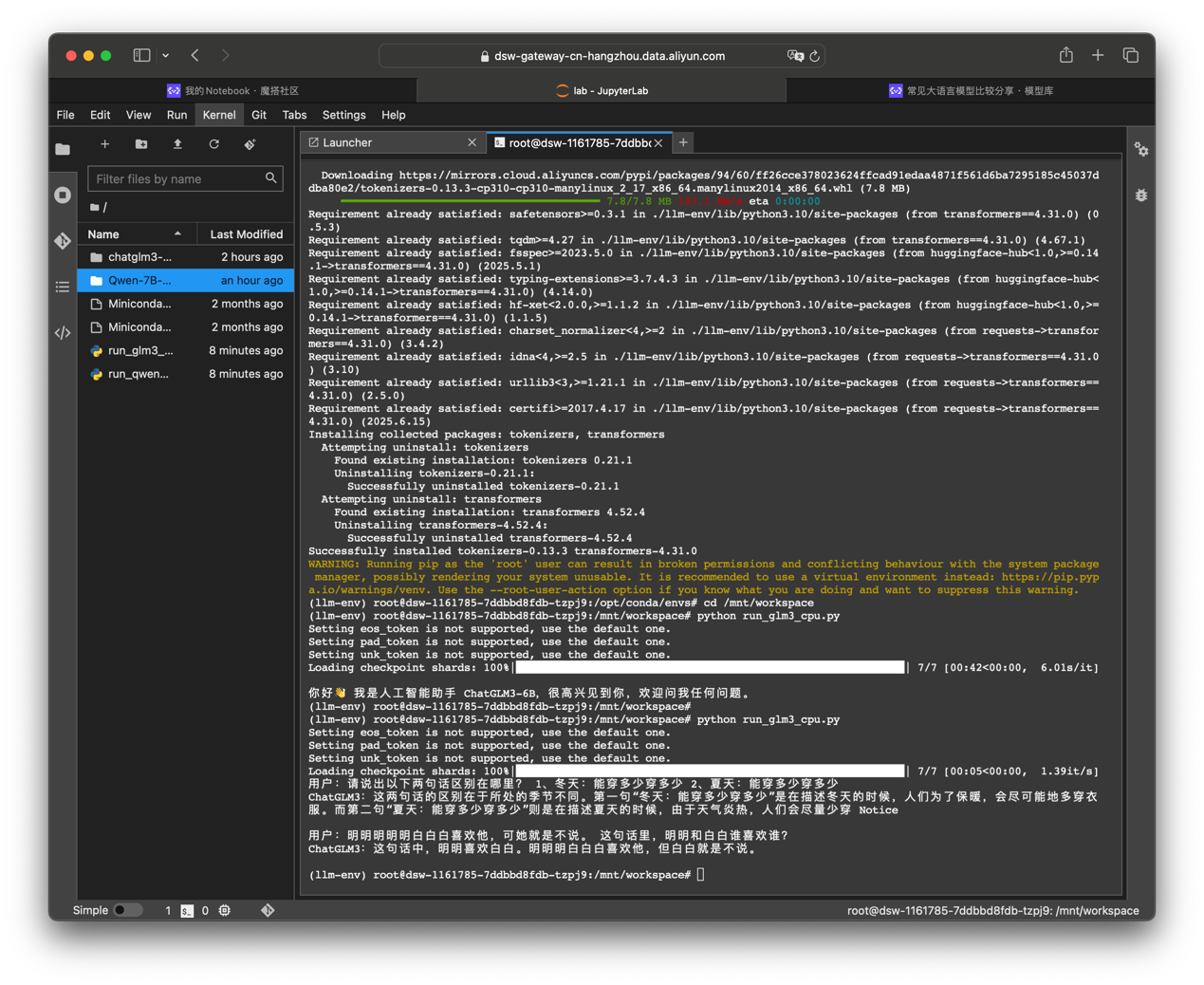
1. 部署截图



千问大模型部署截图



ChatGLM3-b6大模型部署截图



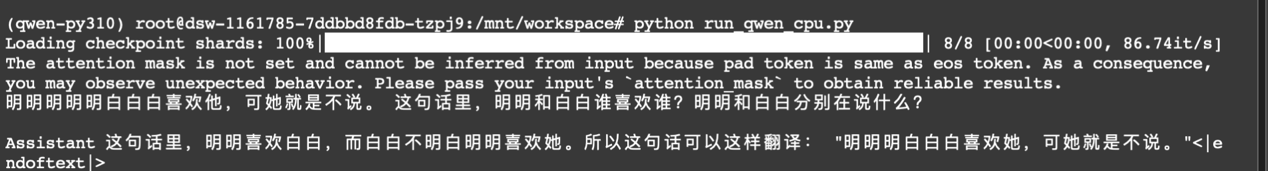
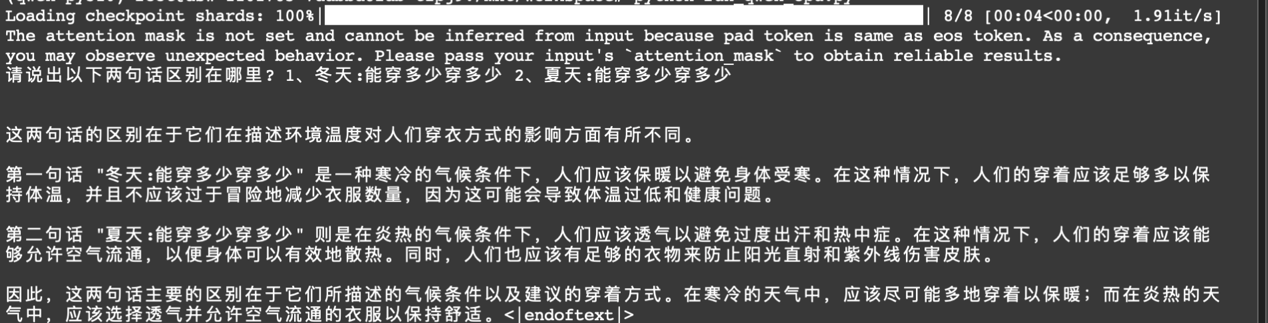
项目结构图

五、实验方法

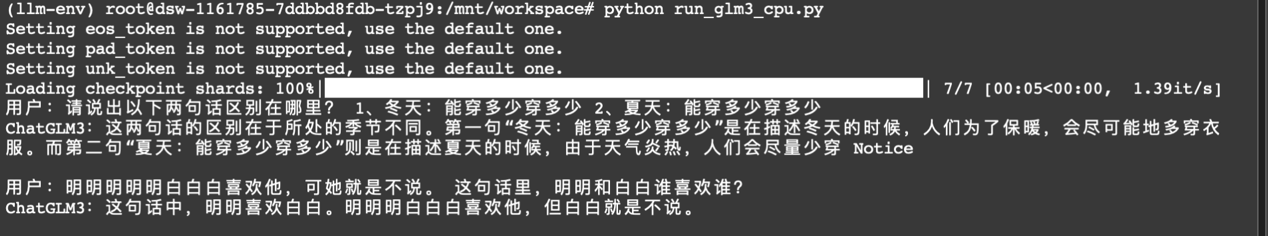
此次实验通过对两个模型实行问答测试来对比两个模型。

问题一：他知道我知道你知道他不知道吗？ 这句话里，到底谁不知道

问题二：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



千问大模型的回答



CHatGLM3大模型的回答

六、对比结果

由上方回答结果可以得到以下结论

1. 两个大模型回答准确性高，在语义理解上，对于同样的问题都没有出现错误。
2. 输出模式有差别：

千问大模型本地部署一次只能问一个问题，且输出答案为一个字符一个字符输出。

GLM3大模型本地部署可以同时问多个问题（类似于历史回答），并可以多轮提问回答。

1. 响应速度的差别：

在同样的环境和配置下，千问大模型回复速度响应速度较慢，GLM3大模型回复速度响应速度较快。

1. 回答问题详细程度：

在默认情况下，千问模型的回复比较简洁，GLM3大模型的回复比较详细。

1. 调用代码对比：

从调用代码的方向看，千问模型的调用代码较长，参数较多。

GLM3模型的调用代码较短，更方便快捷。

七、实验总结

本实验仅探究了千问模型和GLM3模型的部署体验和分析了两个模型的区别，得到了一些结论。这是初次在平台上进行操作，对于LMM大模型的认知得到了提升，学会了本地部署的方法。相信这次体验会为之后的学习实践道路提供灵感。

平台上传链接

https://github.com/AMROUTE/LMM\_comparison