2024.3.31 书生浦语趣味Demo

1. 部署 InternLM2-Chat-1.8B

2. 部署 优秀作品 八戒-Chat-1.8B

3. 运行 Lagent 智能体 Demo

4. 运行 灵笔 InternLM-XComposer2 → 浅尝

玩大模型 SIG

1. 角色扮演兴趣小组
2. RAG
3. 多模态
4. 多样的开源项目
5. 兴趣小组圆桌会议
6. Deploy并行与量化
7. Agent
8. 模型评测

以 XiYou 为例讲解

创建部署：

1. 创建开发机
2. 名称
3. 镜像 —— 11.7 cuda - conda
4. GPU —— 10% A10
5. 创建
6. 进入开发机 { Jupyterlab ✓
   Terminal
   vscode
7. 点击Terminal
8. 创建环境（copy 代码）
   跟据文档搭环境

看到User时，复制文档示例内容，回车 → 模型响应

输入 exit 退出

二、（八戒）

conda activate demo（上一次创建）

（根据文档操作）

powershell 中修改为开发机端口（ssh 连接）

密码为开发机密码
（成功连接时无响应）
回到 webIDE 界面
打开URL → 使用 Demo → 对话界面...

Ctrl + C 退出

三、30% A100（升级配置）— 需先关闭开发机

根据文档操作

模型加载需要时间
勾选 数据分析
对话界面键入内容...

四：灵笔 50% A100 配制
激活 demo
补充文件包（根据文档操作）
...
在 terminal 中输入指令，构造软链接快捷访问方式
...
点击 URL 之后
直接点击 submit

关闭 、 打开左下角的 terminal、Shutdown All

重新开启一个 terminal
激活 demo
复制文档代码（5.4）

(因前面改更改端口映射，
故此处不用再更改)

打开链接

点击文件上传进行上传图片文件(可任意一张)

对话界面输入：请分析一下图中内容，submit

结束

文档后面有下载其它模型以及
软链接清除方法