# **Hw4:大语言模型部署测试报告**

# **2350283康凤轩**

## **部署流程**

### **1. 环境配置关键步骤**

* 安装基础依赖：

pip install -U pip setuptools wheel

# 安装基础依赖（兼容 transformers 4.33.3 和 neuralchat）

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"

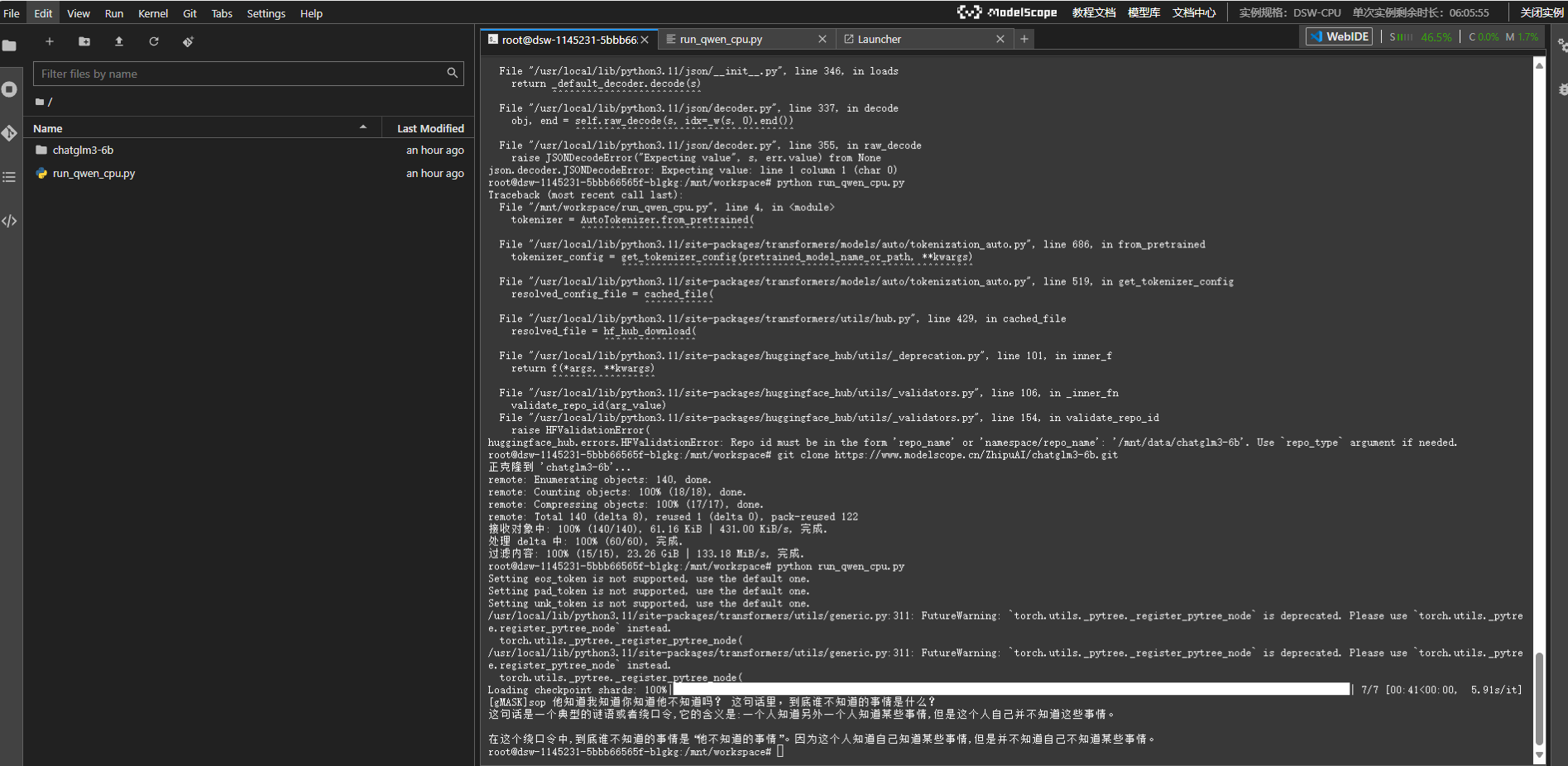
# 安装 fschat（需要启用 PEP517 构建）

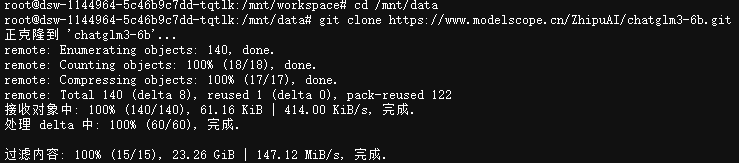
pip install fschat --use-pep517

* 模型推理脚本编写：编写run\_qwen\_cpu.py，使用AutoModelForCausalLM加载模型，并通过TextStreamer实现流式输出。

### **2. 模型搭建**

通过 ModelScope 平台完成以下模型的本地部署：

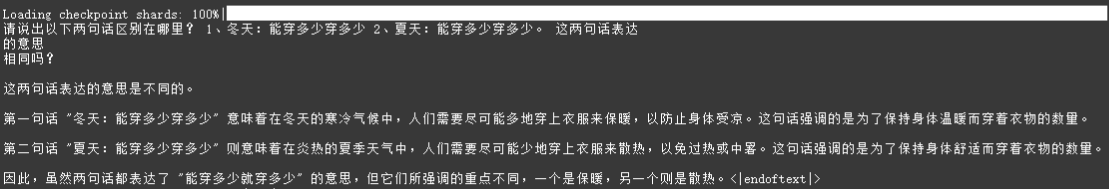


* ****ChatGLM3-6B****：执行git clone https://www.modelscope.cn/ZhipuAI/chatglm3-6b.git，成功获取模型文件。部署截图: 
* ****Qwen-7B-Chat****：通过git clone https://www.modelscope.cn/qwen/Qwen-7B-Chat.git,成功获取模型文件。部署截图: 

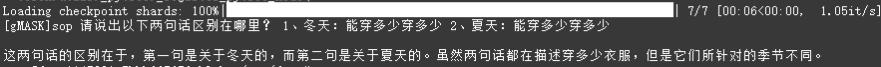
## **二、问答测试**

#### **问题 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**

* ****Qwen-7B-Chat 回答****：明确指出冬天 “尽可能多穿保暖”，夏天 “尽可能少穿散热”，强调季节需求差异。

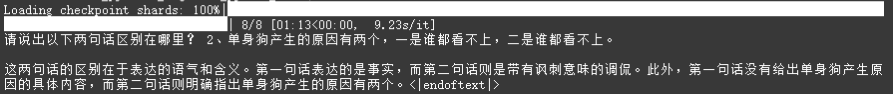


* ****ChatGLM3-6B 回答****：提到 “季节不同导致穿衣逻辑相反”，但未细化到保暖 / 散热场景。

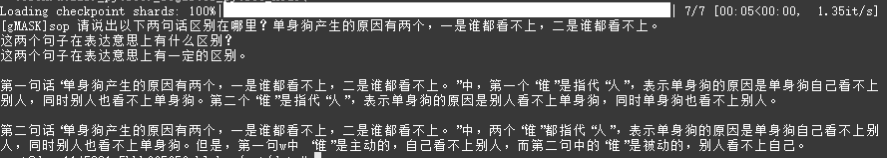


#### **问题 2：请说出以下两句话区别在哪里？单身狗产生的原因有两个，一是谁都看不上，二是谁都看不上**

* ****Qwen-7B-Chat 回答****：仅指出语气差异，第一句为事实陈述，第二句带讽刺意味。对句子真正含义未能理解。

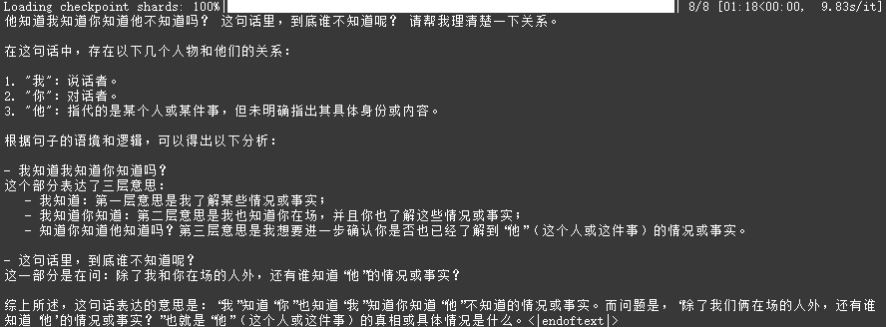


* ****ChatGLM3-6B 回答****：分析 “谁都看不上” 的双向性，即 “自己看不上别人” 与 “别人看不上自己”。

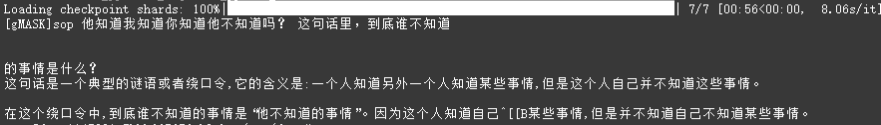


#### **问题3：他知道我知道你知道他不知道吗？ 这句话里，到底谁不知道**

* ****Qwen-7B-Chat 回答****：拆解为三层逻辑，明确 “他” 是不知道的主体，因 “我知道你知道他不知道”。



* ****ChatGLM3-6B 回答****：认为是绕口令，没能指出句子的实际含义，对未知主体“他”也缺乏解释。



## **三、模型横向对比分析**

| **维度** | **Qwen-7B-Chat** | **ChatGLM3-6B** |
| --- | --- | --- |
| ****语义理解**** | 精准捕捉歧义细节（问题1） | 能区分语义但细节稍欠（问题2） |
| ****逻辑推理**** | 逻辑步骤详细，但存在不够准确或回答冗长问题。 | 能拆解多层逻辑，但部分问题推理断层。 |
| ****响应速度**** | CPU 环境下生成速度约 1.35it/s | 约 1.05it/s |
| ****实用性**** | 对话交互与详细推理 | 对话交互与轻度推理 |

### **总结：**

* ****Qwen-7B-Chat****综合表现最优，语义理解与逻辑推理能力均衡，适合对准确性要求高的场景。
* ****ChatGLM3-6B****在对话流畅度上占优，但复杂问题处理不够理想。
* 两种模型都能满足轻度与入门使用，但由于二者在参数规模上仍不够大，都有可能出现答非所问或推理错误的情况，同时CPU环境下模型推理速度也不够迅速，可以考虑在GPU环境下部署规模更大的模型以获得更好体验。