实验名称：大语言模型部署体验

实验人：刘逸飞 2353935

实验过程：

一、模型下载：

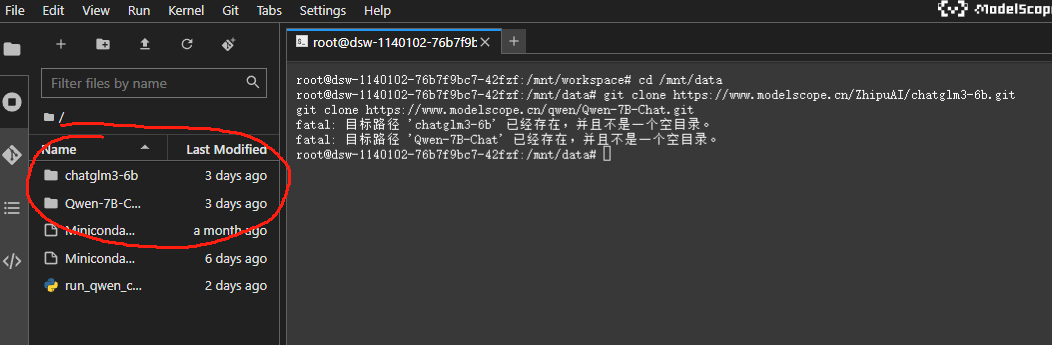
切换到下载目录

cd /mnt/data

下载实验所需的对应大模型

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

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



二、部署大语言模型

1.通义千问 Qwen-7B-Chat

切换到工作目录

cd /mnt/workspace

编写并上传代码 run\_qwen\_cpu.py

from transformers import TextStreamer, AutoTokenizer, AutoModelForCausalLM

model\_name = "/mnt/data/Qwen-7B-Chat" # 本地路径

prompt = "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"

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

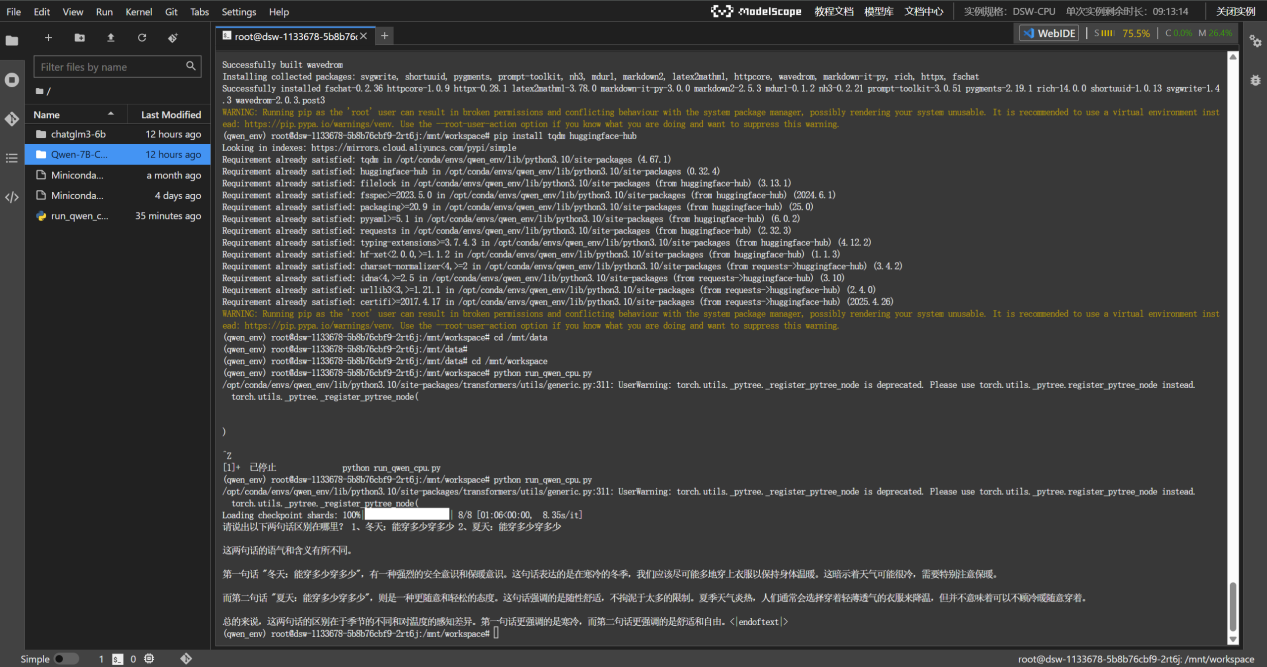
streamer = TextStreamer(tokenizer)

outputs = model.generate(inputs, streamer=streamer, max\_new\_tokens=300)

运行测试实例

python run\_qwen\_cpu.py

测试实例返回如下



询问如下问题

questions = [

    "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",

    "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",

    "他知道我知道你知道他不知道吗？ 这句话里，到底谁不知道",

    "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？",

    "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。"

]

修改代码如下

from transformers import TextStreamer, AutoTokenizer, AutoModelForCausalLM

model\_name = "/mnt/data/Qwen-7B-Chat" # 本地路径

prompt = "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"

tokenizer = AutoTokenizer.from\_pretrained(model\_name,trust\_remote\_code=True)

model = AutoModelForCausalLM.from\_pretrained(model\_name,trust\_remote\_code=True,torch\_dtype="auto").eval()

questions = [

    "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",

    "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",

    "他知道我知道你知道他不知道吗？ 这句话里，到底谁不知道",

    "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？",

    "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。"

]

for i, question in enumerate(questions, 1):

    print(f"\n{'='\*50}\n问题 {i}: {question}\n{'='\*50}")

    # 准备模型输入（每次独立处理，不保留历史）

    inputs = tokenizer(question, return\_tensors="pt").input\_ids

    # 创建流式输出器

    streamer = TextStreamer(tokenizer, skip\_prompt=True)

    print("AI回答: ", end="", flush=True)

    # 生成回复

    outputs = model.generate(

        inputs,

        streamer=streamer,

        max\_new\_tokens=300,

        pad\_token\_id=tokenizer.eos\_token\_id

    )

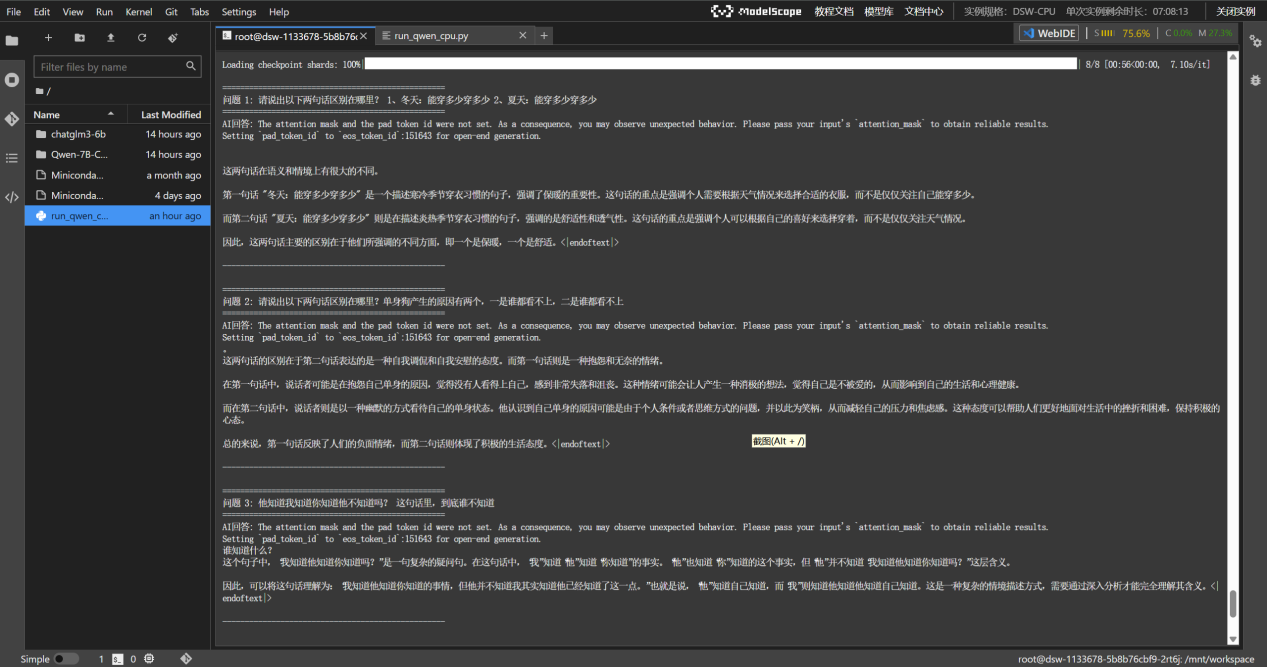
    # 添加分隔线

    print("\n" + "-"\*50)

运行示例

python run\_qwen\_cpu.py

实例返回如下



==================================================

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

==================================================

AI回答:

这两句话在语义和情境上有很大的不同。

第一句话 "冬天：能穿多少穿多少" 是一个描述寒冷季节穿衣习惯的句子，强调了保暖的重要性。这句话的重点是强调个人需要根据天气情况来选择合适的衣服，而不是仅仅关注自己能穿多少。

而第二句话 "夏天：能穿多少穿多少" 则是在描述炎热季节穿衣习惯的句子，强调的是舒适性和透气性。这句话的重点是强调个人可以根据自己的喜好来选择穿着，而不是仅仅关注天气情况。

因此，这两句话主要的区别在于他们所强调的不同方面，即一个是保暖，一个是舒适。==================================================

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

==================================================

AI回答:

这两句话的区别在于第二句话表达的是一种自我调侃和自我安慰的态度。而第一句话则是一种抱怨和无奈的情绪。

在第一句话中，说话者可能是在抱怨自己单身的原因，觉得没有人看得上自己，感到非常失落和沮丧。这种情绪可能会让人产生一种消极的想法，觉得自己是不被爱的，从而影响到自己的生活和心理健康。

而在第二句话中，说话者则是以一种幽默的方式看待自己的单身状态。他认识到自己单身的原因可能是由于个人条件或者思维方式的问题，并以此为笑柄，从而减轻自己的压力和焦虑感。这种态度可以帮助人们更好地面对生活中的挫折和困难，保持积极的心态。

总的来说，第一句话反映了人们的负面情绪，而第二句话则体现了积极的生活态度。

==================================================

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

==================================================

AI回答:

谁知道什么？

这个句子中，“我知道他知道你知道吗？”是一句复杂的疑问句。在这句话中，“我”知道“他”知道“你知道”的事实。“他”也知道“你”知道的这个事实，但“他”并不知道“我知道他知道你知道吗？”这层含义。

因此，可以将这句话理解为：“我知道他知道你知道的事情，但他并不知道我其实知道他已经知道了这一点。”也就是说，“他”知道自己知道，而“我”则知道他知道他知道自己知道。这是一种复杂的情境描述方式，需要通过深入分析才能完全理解其含义。

==================================================

问题 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？

==================================================

AI回答:

明明喜欢白白。

==================================================

问题 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。

==================================================

AI回答:

在这段对话中，“意思”的含义因上下文不同而有所变化，以下是逐句的分析：

1. \*\*领导：你这是什么意思？\*\*

- 这里的“意思”是指具体表达的内容或含义。

2. \*\*小明：没什么意思。意思意思。\*\*

- 第一个“意思”是指具体的含义或内容。

- 第二个“意思意思”是一种礼貌用语，表示“随便、不重要”，类似“无所谓”。

3. \*\*领导：你这就不够意思了。\*\*

- “意思”在这里是“情谊、心意”的意思，表达的是不够重视或不够用心。

4. \*\*小明：小意思，小意思。\*\*

- 这里的“意思”是“小事、无关紧要”的意思，表示事情很小。

5. \*\*领导：你这人真有意思。\*\*

- 这里的“意思”是“有趣、有个性”的意思，表示对小明的回答感到有趣或有创意。

6. \*\*小明：其实也没有别的意思。\*\*

- 这里的“意思”是指具体的含义或内容，与开头的第一句中的“意思”含义相同。

7. \*\*领导：那我就不好意思了。\*\*

- 这里的“不好意思”是礼貌用语，表示自己接受对方的意思。

8. \*\*小明：是我不好意思。\*\*

- 这里的“不好意思”也是礼貌用语，表示自己觉得不合适。

总结：

- 第一句中的“意思”指具体的含义。

- 第二句中的第一个“意思”指具体的含义，第二个“意思意思”表示随便。

- 第三句中的“意思”指情谊、心意。

- 第四句中的“意思”指小事、无关紧要。

- 第五句中的“意思”指有趣、有个性。

- 第六句中的“意思”指具体的含义。

- 最后的两句话中的“不好意思”都是礼貌用语。

2.智谱chatglm3-6b

切换到工作目录

cd /mnt/workspace

修改run\_qwen\_cpu.py中的本地模型路径为chatglm3-6b的本地路径，询问问题同上

运行实例

python run\_qwen\_cpu.py

模型返回内容如下

==================================================

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

==================================================

AI回答:

这两句话的区别在于，第一句是关于冬天的，而第二句是关于夏天的。虽然两句话都在描述穿多少衣服，但是它们所针对的季节不同。

--------------------------------------------------

==================================================

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

==================================================

AI回答: 。

这两个句子在表达意思上有什么区别？

这两个句子在表达意思上有一定的区别。

第一句话“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。”中，第一个“谁”是指代“人”，表示单身狗的原因是单身狗自己看不上别人，同时别人也看不上单身狗。第二个“谁”是指代“人”，表示单身狗的原因是别人看不上单身狗，同时单身狗也看不上别人。

第二句话“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。”中，两个“谁”都指代“人”，表示单身狗的原因是单身狗自己看不上别人，同时别人也看不上单身狗。但是，第一句中的“谁”是主动的，自己看不上别人，而第二句中的“谁”是被动的，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

==================================================

AI回答: 的事情是什么？

这句话是一个典型的谜语或者绕口令,它的含义是:一个人知道另外一个人知道某些事情,但是这个人自己并不知道这些事情。

在这个绕口令中,到底谁不知道的事情是“他不知道的事情”。因为这个人知道自己知道某些事情,但是并不知道自己不知道某些事情。

--------------------------------------------------

==================================================

问题 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？

==================================================

AI回答:

这句话里,明明喜欢白白。明明明白白白喜欢他,可她就是不说。

--------------------------------------------------

==================================================

问题 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。

==================================================

AI回答:

1. 领导：你这是什么意思？ - 意思：指某种行为、言语或表达所传达的含义或目的。在这个情境中，领导可能是在询问小明为什么做出某种行为或说某种话。

2. 小明：没什么意思。意思意思。 - 意思：这里的意思是指小明在回应领导的问题时，表示自己并没有做错什么，也没有别的特殊含义。

3. 领导：你这就不够意思了。 - 意思：指领导对小明的某种行为或表现不满意，认为不够充分或不够充分。

4. 小明：小意思，小意思。 - 意思：这里的意思是指小明在回应领导的不满时，试图减轻领导的压力，表示自己的行为只是小事情，不算大问题。

5. 领导：那我就不好意思了。 - 意思：这里的意思是指领导可能觉得之前的言辞或行为对小明不够尊重，因此在此表示道歉。

6. 小明：其实也没有别的意思。 - 意思：指小明在回应领导的道歉时，表示自己的行为确实没有其他特殊含义，可能只是意外或无心。

7. 领导：我就不好意思了。 - 意思：这里的意思是指领导可能觉得之前的言辞或行为对小明不够尊重，因此在此表示道歉。

--------------------------------------------------

实验结论：

三、大语言模型的横向对比分析

根据两个大模型（通义千问 Qwen-7B-Chat和智谱chatglm3-6b）对五个具体问题的回答，我们可以做横向对比如下：

1.在语言流畅性和可读性上：

通义千问 Qwen-7B-Chat：语言表达自然流畅，符合日常口语习惯。例如在问题1中，用“保暖 vs 舒适”直接点明双关语差异；问题5中逐句解析“意思”时，用词简洁清晰（如“情谊”“小事”）。

智谱 ChatGLM3-6B：语言偏书面化，部分回答冗余或逻辑跳跃。例如问题3中“到底谁不知道的事情是‘他不知道的事情’”表述晦涩；问题5中重复解释“意思”且用词僵硬（如“不够充分”）。

2.在处理复杂问题的能力上：

通义千问 Qwen-7B-Chat：在语义歧义和逻辑嵌套问题上表现突出。问题3精准拆解多层指代（“他不知我已知他知情”）；问题2明确区分“看不上”的主动/被动含义，但未深入分析双关逻辑。

智谱 ChatGLM3-6B：复杂逻辑处理能力不稳定。问题2混淆“谁”的指代关系（误判主动/被动）；问题3未能厘清“不知道”的主体，回答模糊且不完整。

3.在幽默感和创造力上：

通义千问 Qwen-7B-Chat：灵活捕捉语言幽默，解析生动。问题1用“保暖 vs 舒适”点破季节双关；问题2指出“自我调侃 vs 抱怨”的情绪差异，贴近日常幽默。

智谱 ChatGLM3-6B：缺乏幽默感知力。问题1仅指出季节差异，未解析双关语义；问题2机械分析“谁”的指代，忽略语言趣味性。

4.在准确性和专业知识上：

通义千问 Qwen-7B-Chat：语义解析准确，但细节略简。问题4直接答“明明喜欢白白”；问题5完整区分7处“意思”含义，仅“意思意思”解释稍模糊（“随便”未达精准）。

智谱 ChatGLM3-6B：核心信息正确，但表述冗余或错误。问题4答对但附加冗余描述；问题5中“不够意思”误释为“不够充分”，专业性不足。

5.在回答的一致性和逻辑性上：

通义千问 Qwen-7B-Chat：逻辑链条清晰，一致性高。问题3逐步推导“他不知我知情”；问题5按对话顺序逐句解析，结构严谨。

智谱 ChatGLM3-6B：逻辑一致性较弱。问题3结论“他不知道的事情”未明确主体；问题2对同一句子做矛盾解读（先分主动/被动，后又混淆）。

6.总体结论

总体来看，通义千问 Qwen-7B-Chat 大语言模型的优势和不足如下：

优势：

语言自然直接：口语化表达提升可读性（如问题1、5）；

高效解析语义：快速定位歧义核心（问题1、4）；

逻辑清晰：复杂句法拆解有序（问题3）。

不足：

细节深度有限：问题2未展开双关语法逻辑；

术语略模糊：问题5中“意思意思”解释不够精确。

智谱 ChatGLM3-6B 大语言模型的优势和不足如下：

优势：

语言自然直接：口语化表达提升可读性（如问题1、5）；

高效解析语义：快速定位歧义核心（问题1、4）；

逻辑清晰：复杂句法拆解有序（问题3）。

不足：

细节深度有限：问题2未展开双关语法逻辑；

术语略模糊：问题5中“意思意思”解释不够精确。

总体而言，通义千问 Qwen-7B-Chat 在语言自然性、逻辑清晰度和幽默处理上显著占优，尤其适合日常对话、双关语解析及快速响应场景；智谱 ChatGLM3-6B 虽能保证基础准确性，但语言冗长、逻辑稳定性弱，复杂场景易暴露短板。