# 大语言模型对比部署体验报告

2354177 陈雷诗语 指导老师：曾静

[1. 实验目的和内容概述 1](#_Toc200402637)

[2. 实验环境 2](#_Toc200402638)

[3. 模型部署过程 2](#_Toc200402639)

[3.1 环境部署 2](#_Toc200402640)

[3.2克隆模型 2](#_Toc200402641)

[3.3运行代码 3](#_Toc200402642)

[4. 应用场景测试与结果 4](#_Toc200402643)

[4.1问题设计 4](#_Toc200402644)

[4.2 Qwen回答结果 4](#_Toc200402645)

[4.3 ChatGLM回答结果 6](#_Toc200402646)

[4.4 Chatgpt-4o回答结果 8](#_Toc200402647)

[4.5 deepseek-R1模型 10](#_Toc200402648)

[5. 大模型回答对比分析 11](#_Toc200402649)

[5.1 几个模型特点 11](#_Toc200402650)

[5.2 模型对比 12](#_Toc200402651)

## 实验目的和内容概述

本次实验旨在通过魔搭ModelScope平台部署多个开源大语言模型，并对其在问答等应用场景中的表现进行横向对比分析。通过部署与测试不同模型，掌握其部署流程、运行环境配置和基本使用方法，并通过实际问题测试比较模型在准确性、流畅性和应用适配性等方面的性能表现。

具体实验内容包括：

1. 登录并注册使用魔搭平台，并完成与阿里云账号的绑定以获取免费云资源；
2. 选择并部署2至3个大语言模型（ChatGLM、Qwen、百川等）；
3. 使Jupyter Notebook或平台提供的镜像环境，在部署环境中运行模型；
4. 设计统一的问题用于模型测试，并对测试结果进行截图记录；
5. 对不同模型在多个维度下的问答表现进行对比分析；
6. 提交完整报告与截图，并提供项目公开可访问链接。

## 实验环境

使用平台：魔搭ModelScope

使用方式：Jupyter Notebook/镜像环境

云计算资源：免费CPU

## 模型部署过程

### 3.1 环境部署

(1)Qwen环境

在root中直接下载

配置基础环境

pipinstall\torch==2.3.0+cpu\torchvision==0.18.0+cpu\--index-url <https://download.pytorch.org/whl/cpu>

配置基础依赖

pip install -U pip setuptools wheel

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”

工作目录

cd /mnt/workspace

(2) ChatGLM环境

官方要求

pip install protobuf 'transformers>=4.30.2' cpm\_kernels 'torch>=2.0' gradio mdtex2html sentencepiece accelerate

实践发现，transformers等级太高会报错，采用4.31.0版本的transformers可以运行

pip install "transformers==4.31.0"

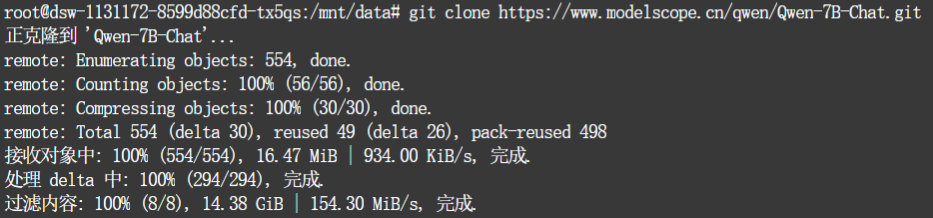
### 3.2克隆模型

在模型中选择了Qwen的Qwen-7B-Chat模型，ChatGLM的chatglm3-6b模型，以及Baichuan的Baichuan2-7B-Chat模型。在运行过程中发现Baichuan2-7B-Chat模型占用内存过大，在运行过程中总被杀死，于是只比较Qwen模型和ChatGLM模型的效果。

1. Qwen模型

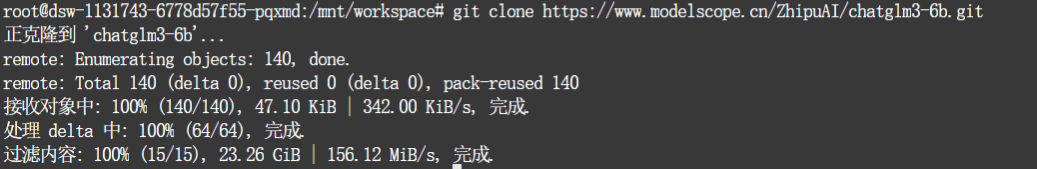
git clone <https://www.modelscope.cn/qwen/Qwen-7B-Chat.git>

git clone https://www.modelscope.cn/models/baichuan-inc/Baichuan2-7B-Chat-4bits



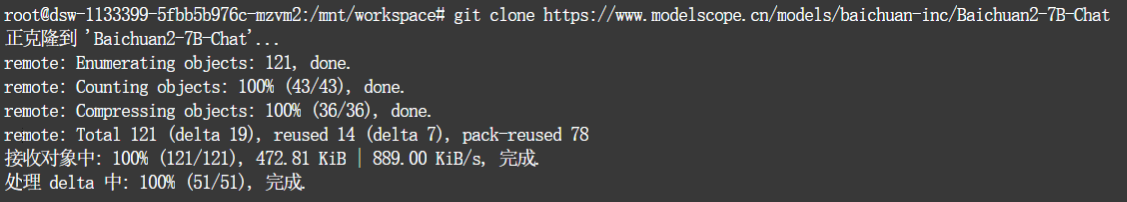
1. ChatGLM模型

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



1. Baichuan 2模型

git clone https://www.modelscope.cn/models/baichuan-inc/Baichuan2-7B-Chat



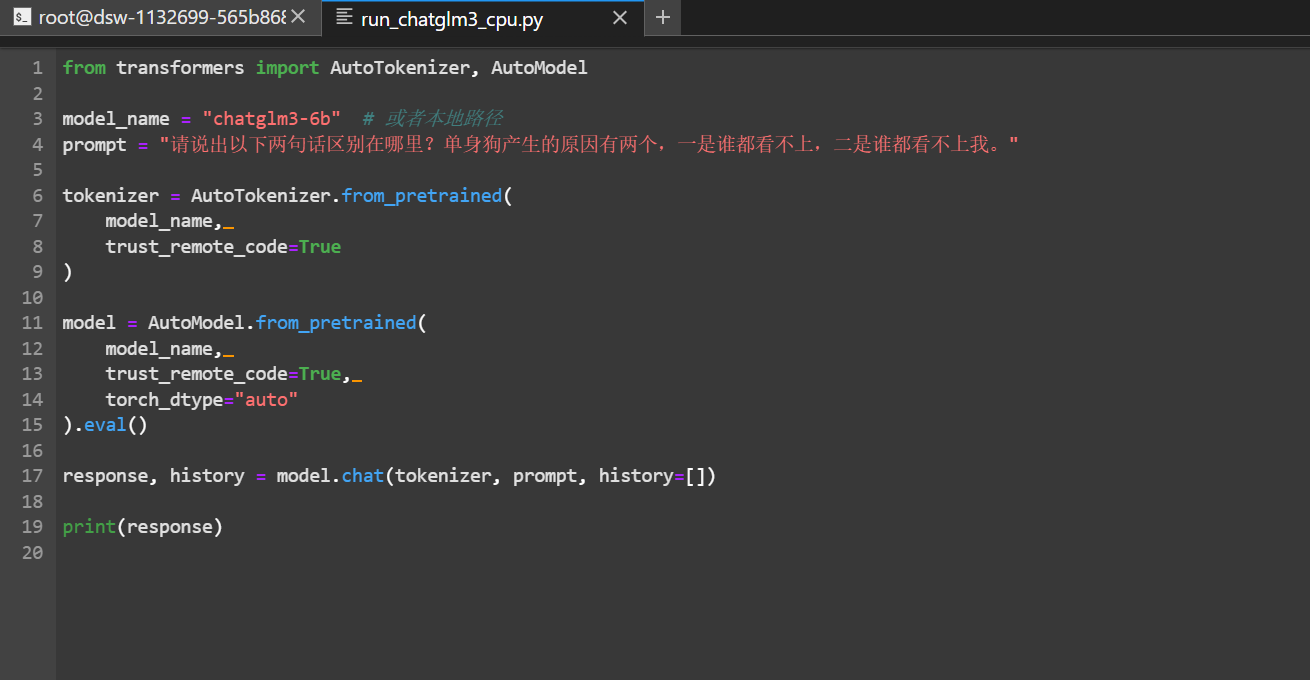
### **3.3运行代码**

在Cpu下运行代码，不同的程序接口不同。

(1)Qwen代码



(2) ChatGLM代码



## 应用场景测试与结果

为了广泛测试不同模型之间的回答差异，再原有魔搭平台部署大模型（共部署3个，1个因内存不足无法运行）的基础上加入了主流ai工具chatgpt4.1和deepseek-R1推理模型的回答。

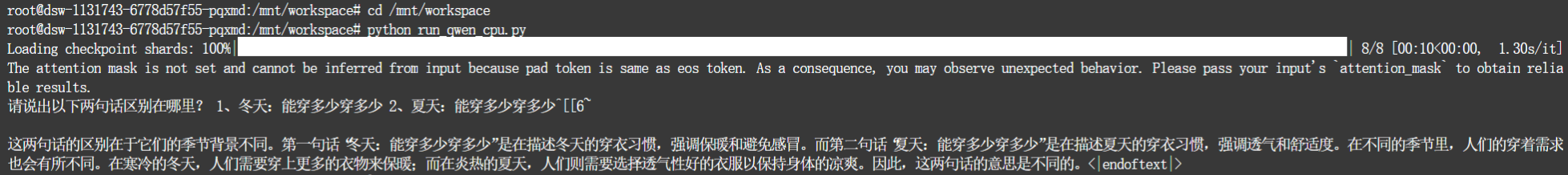
### 4.1问题设计

设计统一问题，每个问题都是一个语言游戏或歧义句，观察ai是否能分辨。

* 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。

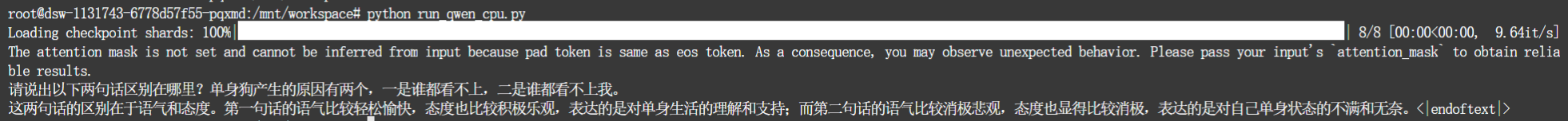
### 4.2 Qwen回答结果

(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



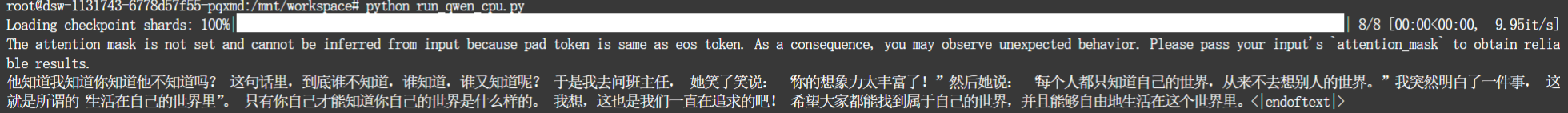
这两句话的区别在于它们的季节背景不同。第一句话“冬天：能穿多少穿多少”是在描述冬天的穿衣习惯，强调保暖和避免感冒。而第二句话“夏天：能穿多少穿多少”是在描述夏天的穿衣习惯，强调透气和舒适度。在不同的季节里，人们的穿着需求也会有所不同。在寒冷的冬天，人们需要穿上更多的衣物来保暖；而在炎热的夏天，人们则需要选择透气性好的衣服以保持身体的凉爽。因此，这两句话的意思是不同的。

(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，二是谁都看不上我。



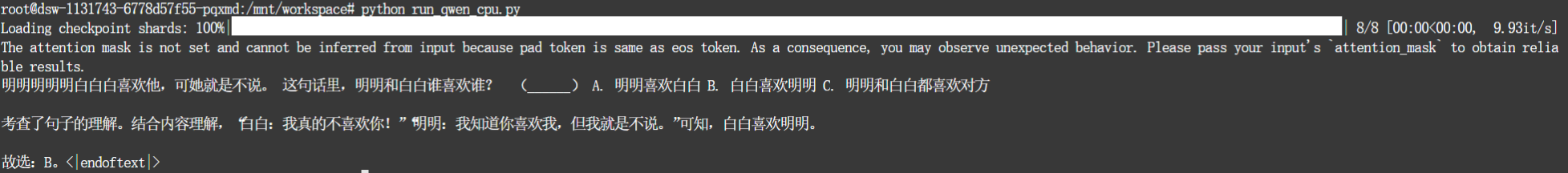
这两句话的区别在于语气和态度。第一句话的语气比较轻松愉快，态度也比较积极乐观，表达的是对单身生活的理解和支持；而第二句话的语气比较消极悲观，态度也显得比较消极，表达的是对自己单身状态的不满和无奈。

(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



这句话里，到底谁不知道，谁知道，谁又知道呢？ 于是我去问班主任， 她笑了笑说： “你的想象力太丰富了！” 然后她说： “每个人都只知道自己的世界，从来不去想别人的世界。” 我突然明白了一件事， 这就是所谓的“生活在自己的世界里”。 只有你自己才能知道你自己的世界是什么样的。 我想，这也是我们一直在追求的吧！ 希望大家都能找到属于自己的世界，并且能够自由地生活在这个世界里。

(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

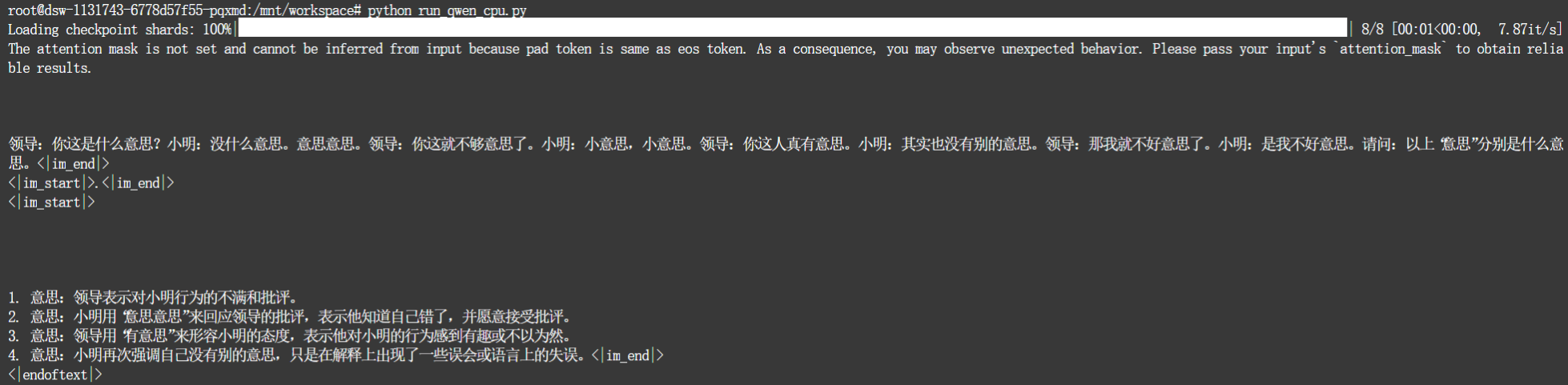


明明明明明白白白喜欢他，可她就是不说。 这句话里，明明和白白谁喜欢谁？ （\_\_\_\_\_\_） A. 明明喜欢白白 B. 白白喜欢明明 C. 明明和白白都喜欢对方

考查了句子的理解。结合内容理解，“白白：我真的不喜欢你！”“明明：我知道你喜欢我，但我就是不说。”可知，白白喜欢明明。

故选：B。

(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。



1. 意思：领导表示对小明行为的不满和批评。

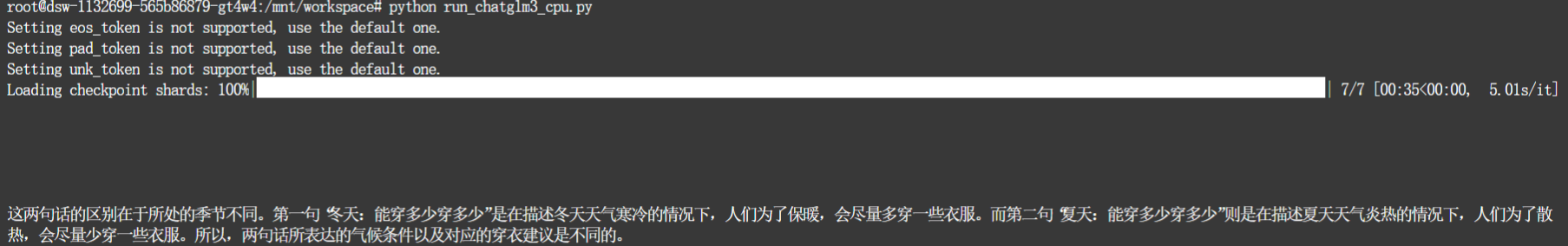
2. 意思：小明用“意思意思”来回应领导的批评，表示他知道自己错了，并愿意接受批评。

3. 意思：领导用“有意思”来形容小明的态度，表示他对小明的行为感到有趣或不以为然。

4. 意思：小明再次强调自己没有别的意思，只是在解释上出现了一些误会或语言上的失误。

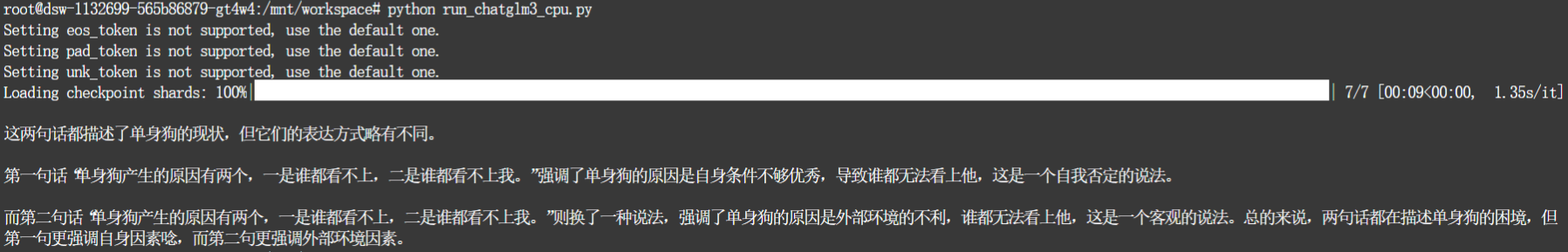
### 4.3 ChatGLM回答结果

(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



这两句话的区别在于所处的季节不同。第一句“冬天：能穿多少穿多少”是在描述冬天天气寒冷的情况下，人们为了保暖，会尽量多穿一些衣服。而第二句“夏天：能穿多少穿多少”则是在描述夏天天气炎热的情况下，人们为了散热，会尽量少穿一些衣服。所以，两句话所表达的气候条件以及对应的穿衣建议是不同的。

(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，二是谁都看不上我。

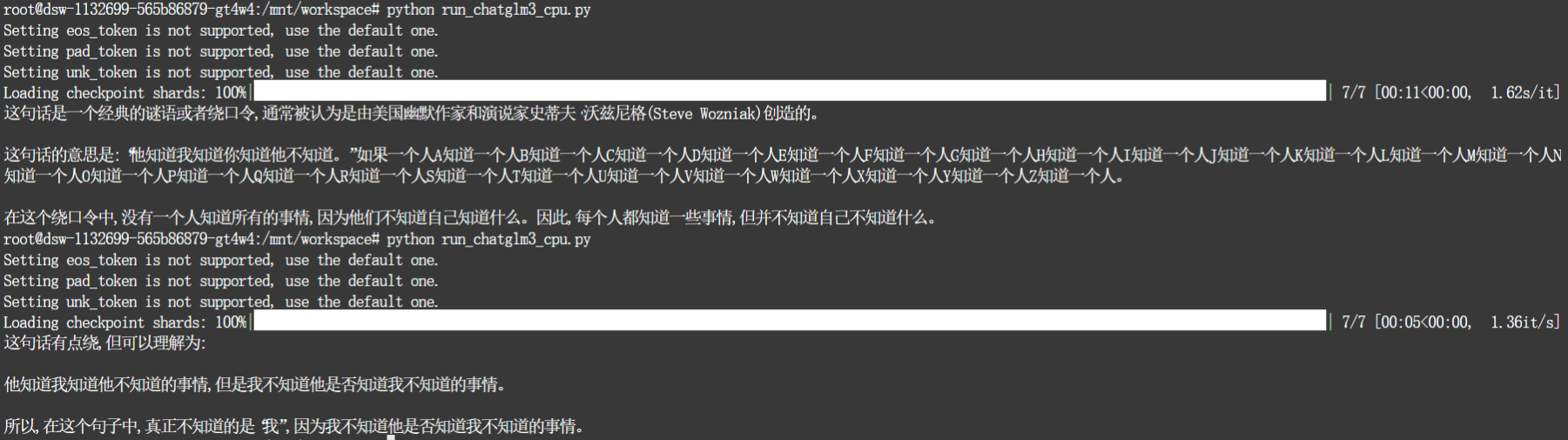


这两句话都描述了单身狗的现状，但它们的表达方式略有不同。

第一句话“单身狗产生的原因有两个，一是谁都看不上，二是谁都看不上我。”强调了单身狗的原因是自身条件不够优秀，导致谁都无法看上他，这是一个自我否定的说法。

而第二句话“单身狗产生的原因有两个，一是谁都看不上，二是谁都看不上我。”则换了一种说法，强调了单身狗的原因是外部环境的不利，谁都无法看上他，这是一个客观的说法。总的来说，两句话都在描述单身狗的困境，但第一句更强调自身因素唸，而第二句更强调外部环境因素。

(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



①第一次提问回答：

这句话是一个经典的谜语或者绕口令,通常被认为是由美国幽默作家和演说家史蒂夫·沃兹尼格(Steve Wozniak)创造的。

这句话的意思是:“他知道我知道你知道他不知道。”如果一个人A知道一个人B知道一个人C知道一个人D知道一个人E知道一个人F知道一个人G知道一个人H知道一个人I知道一个人J知道一个人K知道一个人L知道一个人M知道一个人N知道一个人O知道一个人P知道一个人Q知道一个人R知道一个人S知道一个人T知道一个人U知道一个人V知道一个人W知道一个人X知道一个人Y知道一个人Z知道一个人。

在这个绕口令中,没有一个人知道所有的事情,因为他们不知道自己知道什么。因此,每个人都知道一些事情,但并不知道自己不知道什么。

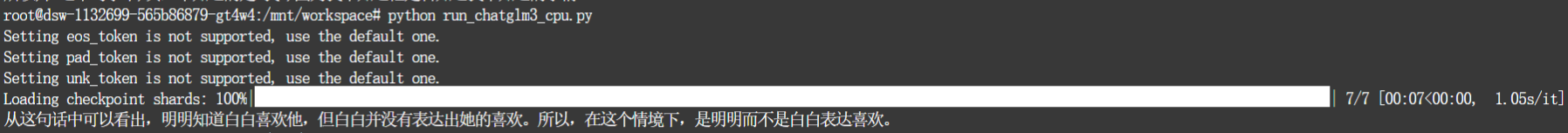
②第二次提问回答：

这句话有点绕,但可以理解为:

他知道我知道他不知道的事情,但是我不知道他是否知道我不知道的事情。

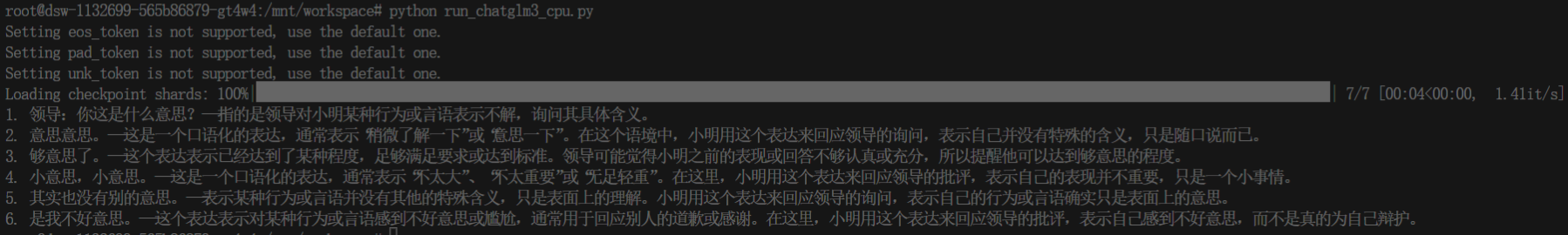
所以,在这个句子中,真正不知道的是“我”,因为我不知道他是否知道我不知道的事情。

(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



从这句话中可以看出，明明知道白白喜欢他，但白白并没有表达出她的喜欢。所以，在这个情境下，是明明而不是白白表达喜欢。

(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。



1.你这是什么意思？  
——指的是领导对小明某种行为或言语表示不解，询问其具体含义。

2.意思意思。  
——这是一个口语化的表达，通常表示“稍微了解一下”或“意思一下”。在这个语境中，小明用这个表达来回应领导的询问，表示自己并没有特殊的含义，只是随口说而已。

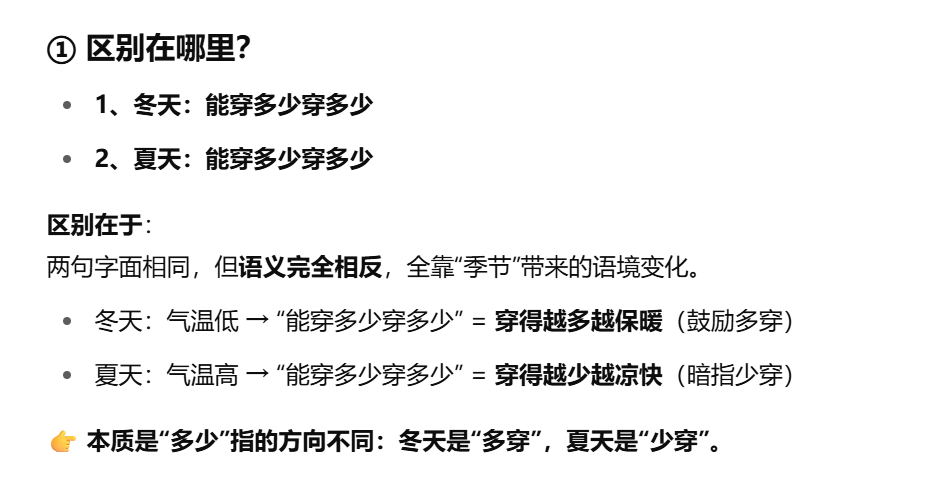
3.够意思了。  
——这个表达表示已经达到了某种程度，足够满足要求或达到标准。领导可能觉得小明之前的表现或回答不够认真或充分，所以提醒他可以达到“够意思”的程度。

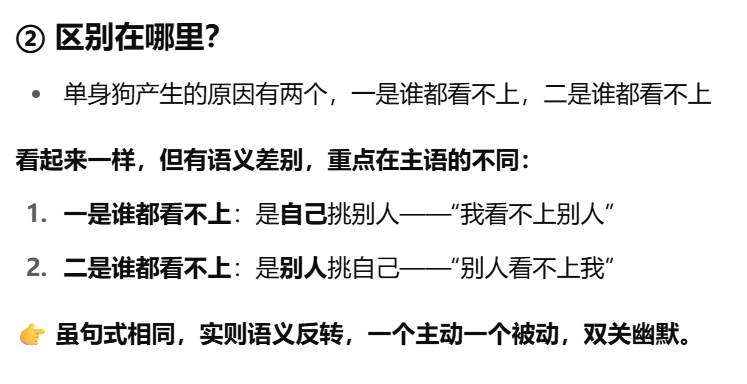
4.小意思，小意思。  
——这是一个口语化的表达，通常表示“无足轻重”。在这里，小明用这个表达来回应领导的批评，表示自己的表现并不重要，只是一个小事情。

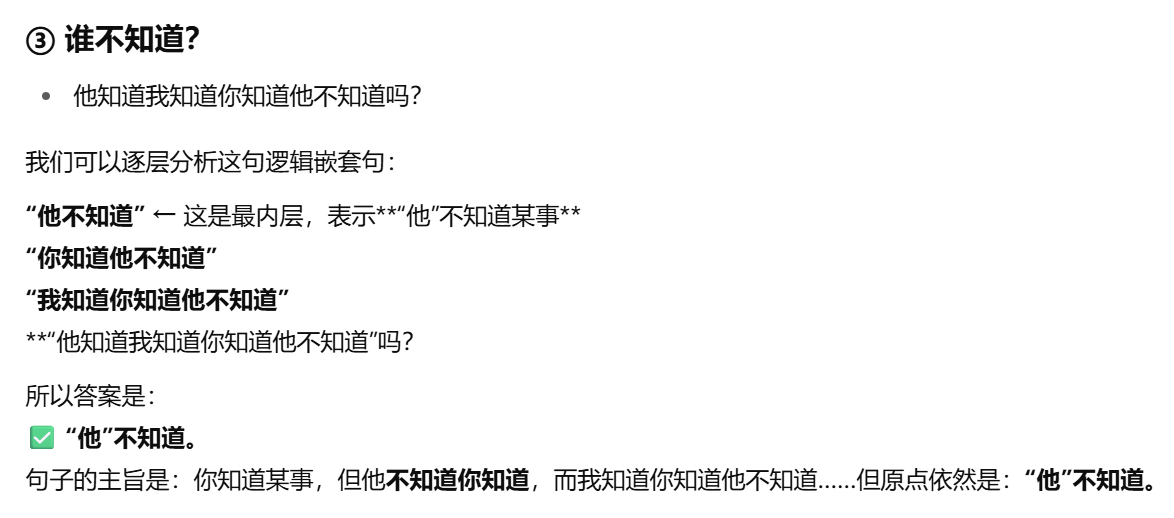
5.其实也没有别的意思。  
——表示某种行为或言语并没有其他的特殊含义，只是表面上的理解。小明用这个表达来回应领导的询问，表示自己的行为或言语确实只是表面上的意思。

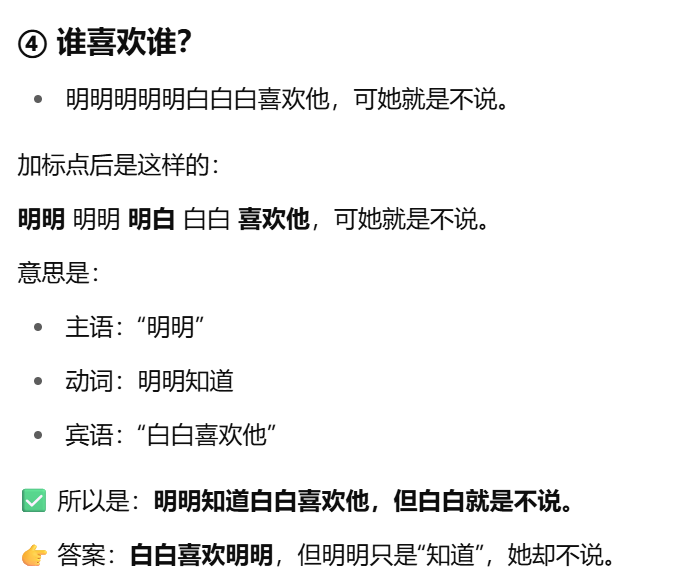
6.是我不好意思。  
——这个表达表示对某种行为或言语感到不好意思或惭愧，通常用于回应别人的道歉或感谢。在这里，小明用这个表达来回应领导的批评，表示自己感到不好意思，而不是真的为自己辩护。

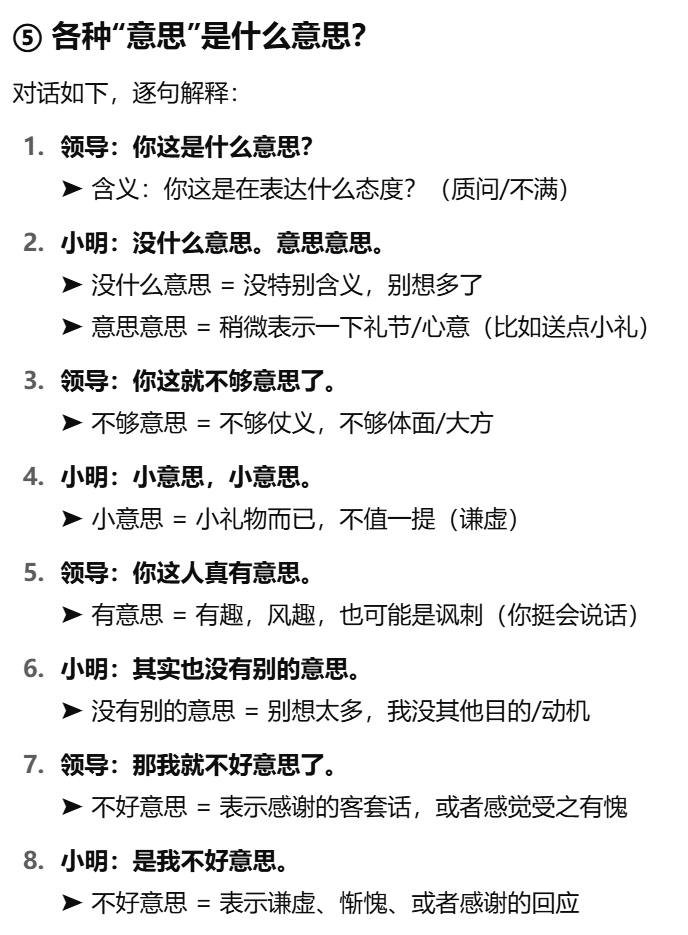
### 4.4 Chatgpt-4o回答结果



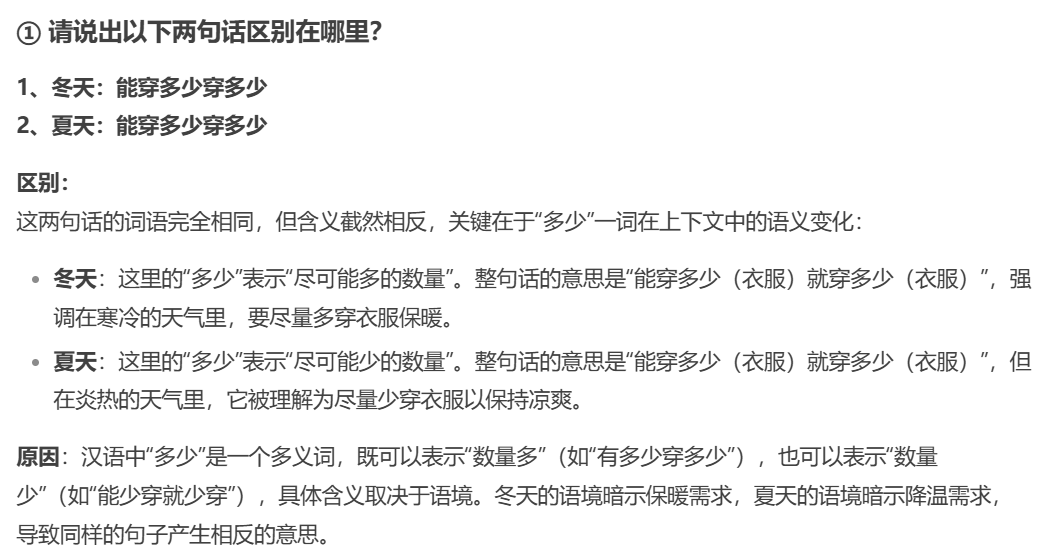


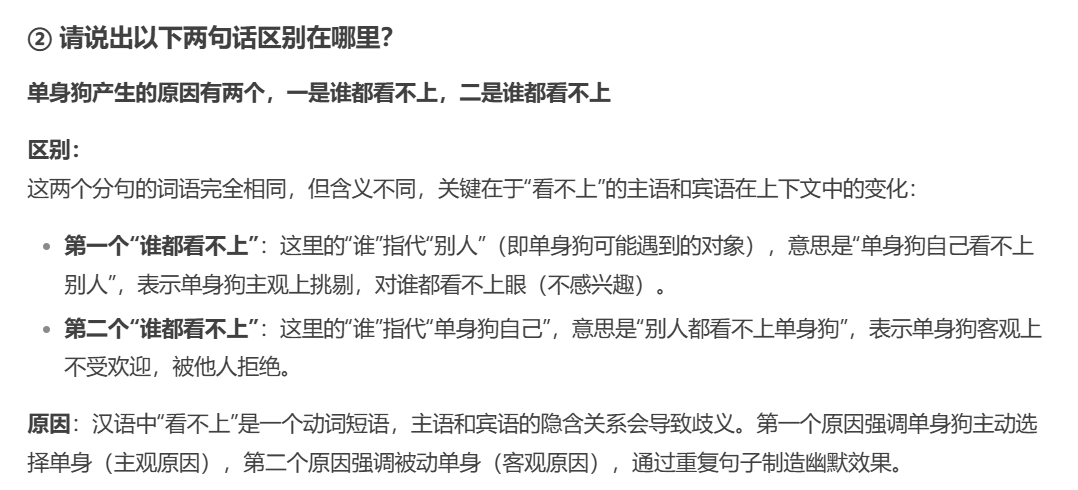


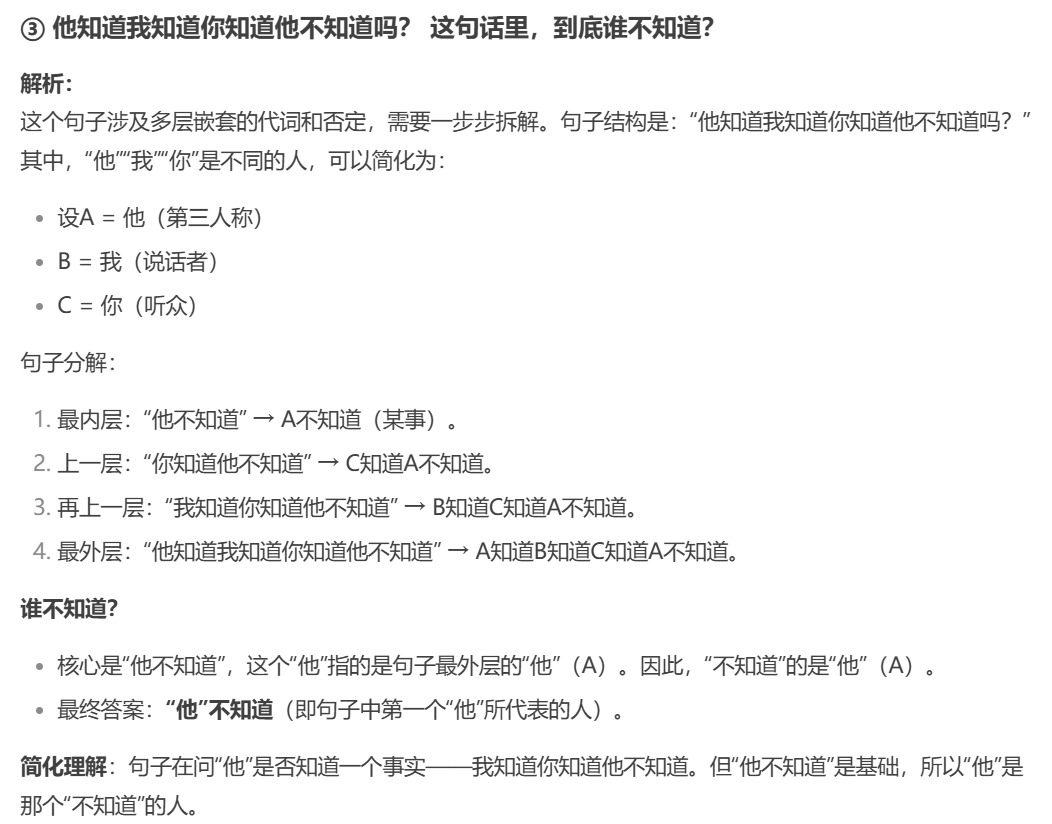


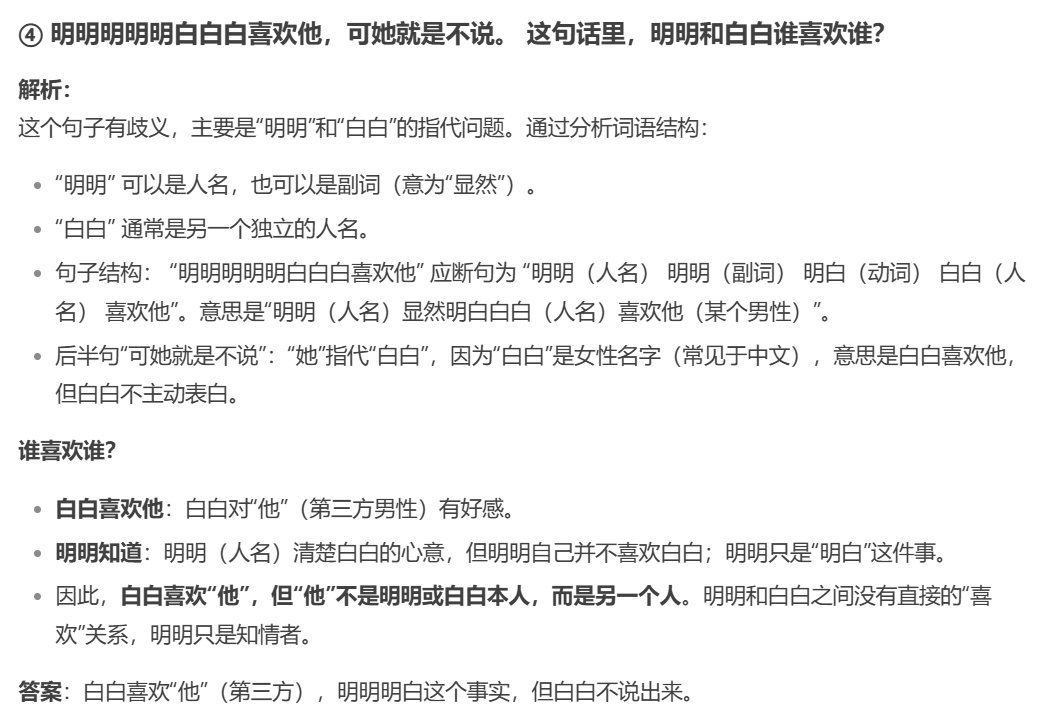


### 4.5 deepseek-R1模型











## 大模型回答对比分析

### 5.1 几个模型特点

Qwen模型的语言比较规范，逻辑也比较清楚，回答的时候喜欢先说句式或者语境上的差别，然后再解释这背后的意思。总体来看，它表达得比较有条理，适合用在教学或者分析类的场合，回答偏向标准答案和理性分析。

ChatGLM模型更灵活、接地气一点，语言更口语化，表现出比较强的“推理和互动”能力。在回答问题时，它会比较关注情境和心理反应的区别，分析角度也比较丰富。不过它的回答结构没那么紧凑，不像Qwen那么条理分明。

ChatGPT-4o模型的语言特别流畅，逻辑也很连贯，能很精准地抓住问题重点，给出的回答全面又细致。在面对本次测试中复杂的语义或者有歧义的句子时，它的表现很好，展现了强大的推理和理解能力，使用了自然又贴近人说话的方式。

DeepSeek-R1模型回答时准确且简洁，能快速抓住关键点，回答直接明了。它的语言风格比较实用，逻辑也比较清晰。面对歧义句时，能分清意思差别，但深度和扩展性还有所提高。

### 5.2 模型对比

从测试结果来看，不同模型在回答语言游戏和歧义句时表现出显著差异：

|  |  |  |
| --- | --- | --- |
| 准确性 | 流畅性 | 应用适配性 |
| * ChatGPT-4o表现最佳，能够精准解析歧义句的核心矛盾，并给出清晰的解释。 * Qwen和DeepSeek-R1次之，能够分辨语义差异，但部分回答略显刻板。 * ChatGLM在部分问题上存在逻辑跳跃，准确性稍逊。 | * ChatGPT-4o和ChatGLM的语言流畅性最高，回答自然且比较贴近日常表达。 * Qwen和DeepSeek-R1的回答较为规范，但是略显僵硬。 | * ChatGPT-4o和ChatGLM更适合需要互动性和创造力的场景，如对话、写作辅助等。 * Qwen和DeepSeek-R1更适合需要逻辑分析和快速响应的场景，如问答系统、技术支持等。 |

综上，不同模型各有优劣。总的来说，ChatGPT-4o和DeepSeek-R1的表现较为出色。因为他们是已经迭代的比较成熟、架构比较复杂的大模型，这种表现出色是较为正常的。而ChatGLM和Qwen是部署在平台上的轻量级模型，他们的表现也是可圈可点的，在面对本次测试中复杂的语义和有歧义的句子时，它们展现出了一定的分辨能力，在表达上也较为流畅。