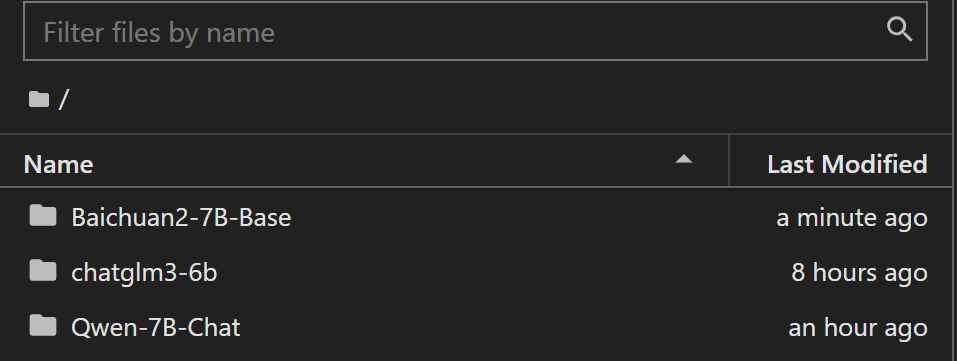
**人工智能导论第四次作业**

1. **模型部署**

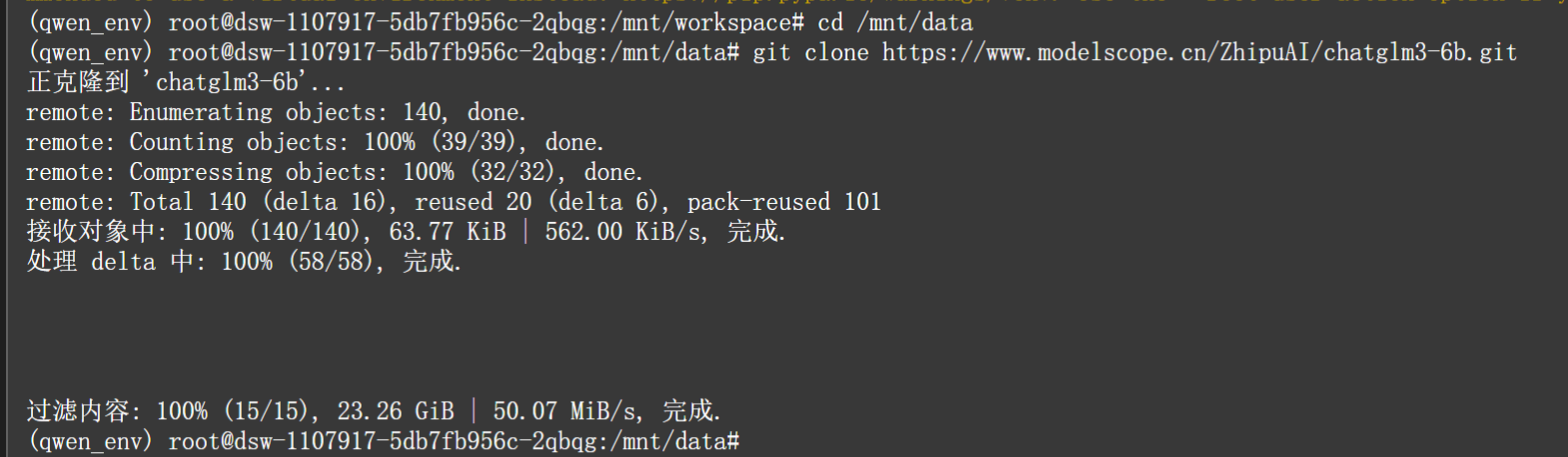
**1.1 Anaconda的下载：**



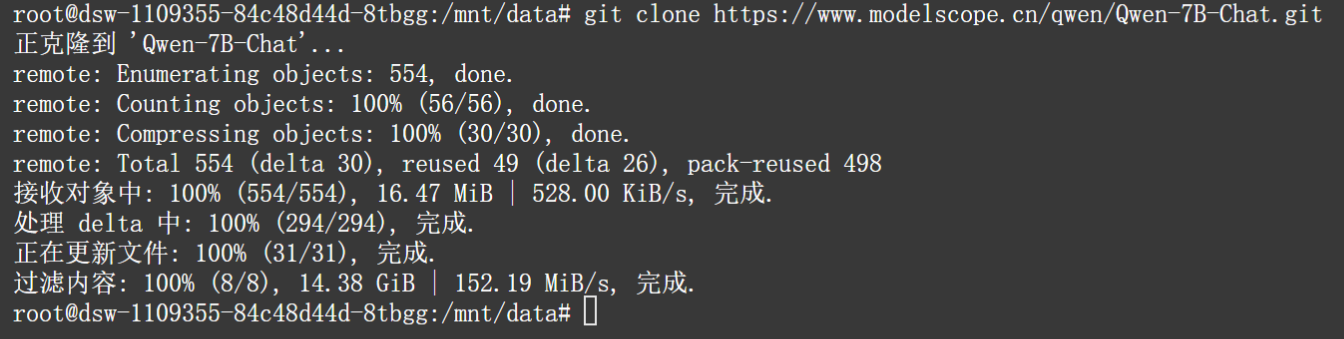
**1.2 3个模型的部署：**



**• chatglm3-6b 的部署结果：**

****

**• Qwen-7B-Chat 的部署结果：**



**• Baichuan2-7B-Base 的部署结果：**



1. **问答测试结果**

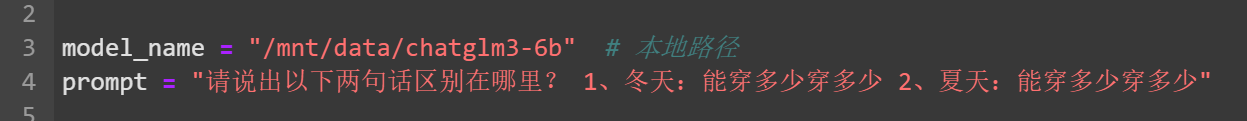
**对三个大模型分别询问相同的三个问题，得到的回复如下：**

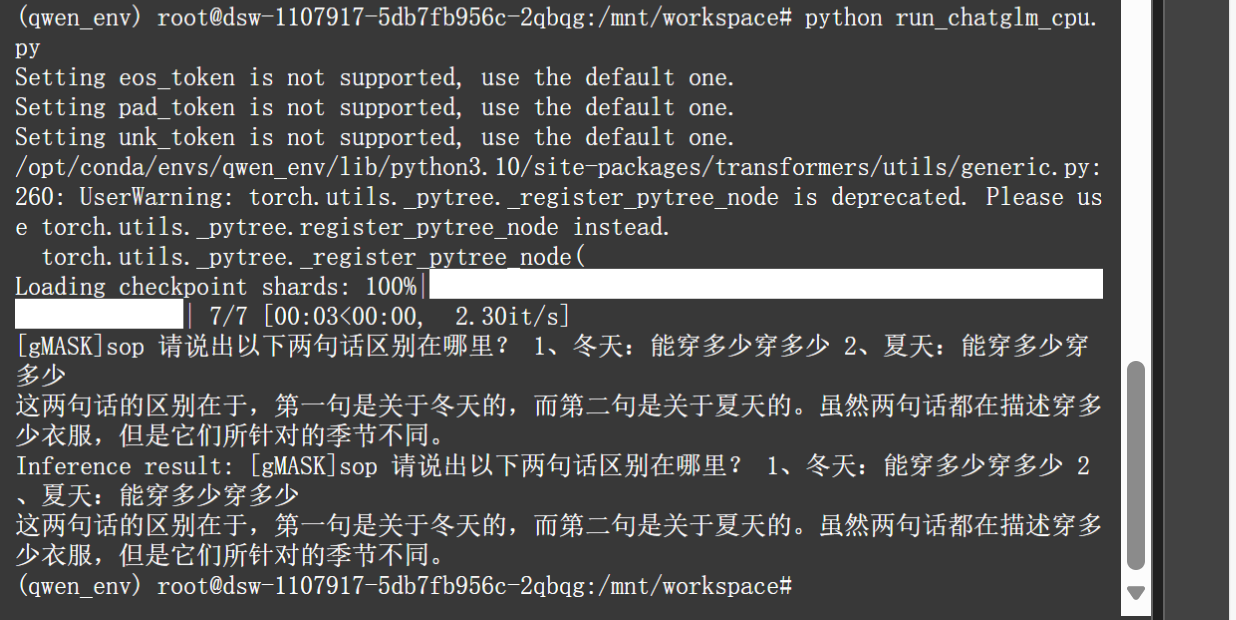
**• chatglm3-6b 的回复：**

**① 请说出以下两句话区别在哪里？**

**1、冬天：能穿多少穿多少**

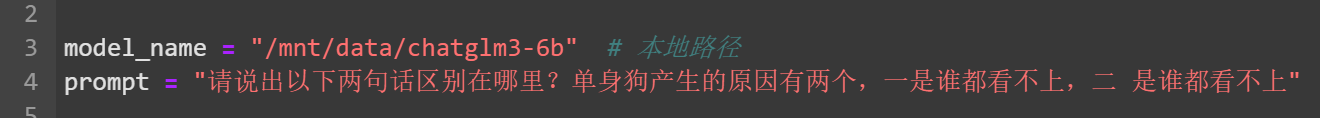
**2、夏天：能穿多少穿多少**

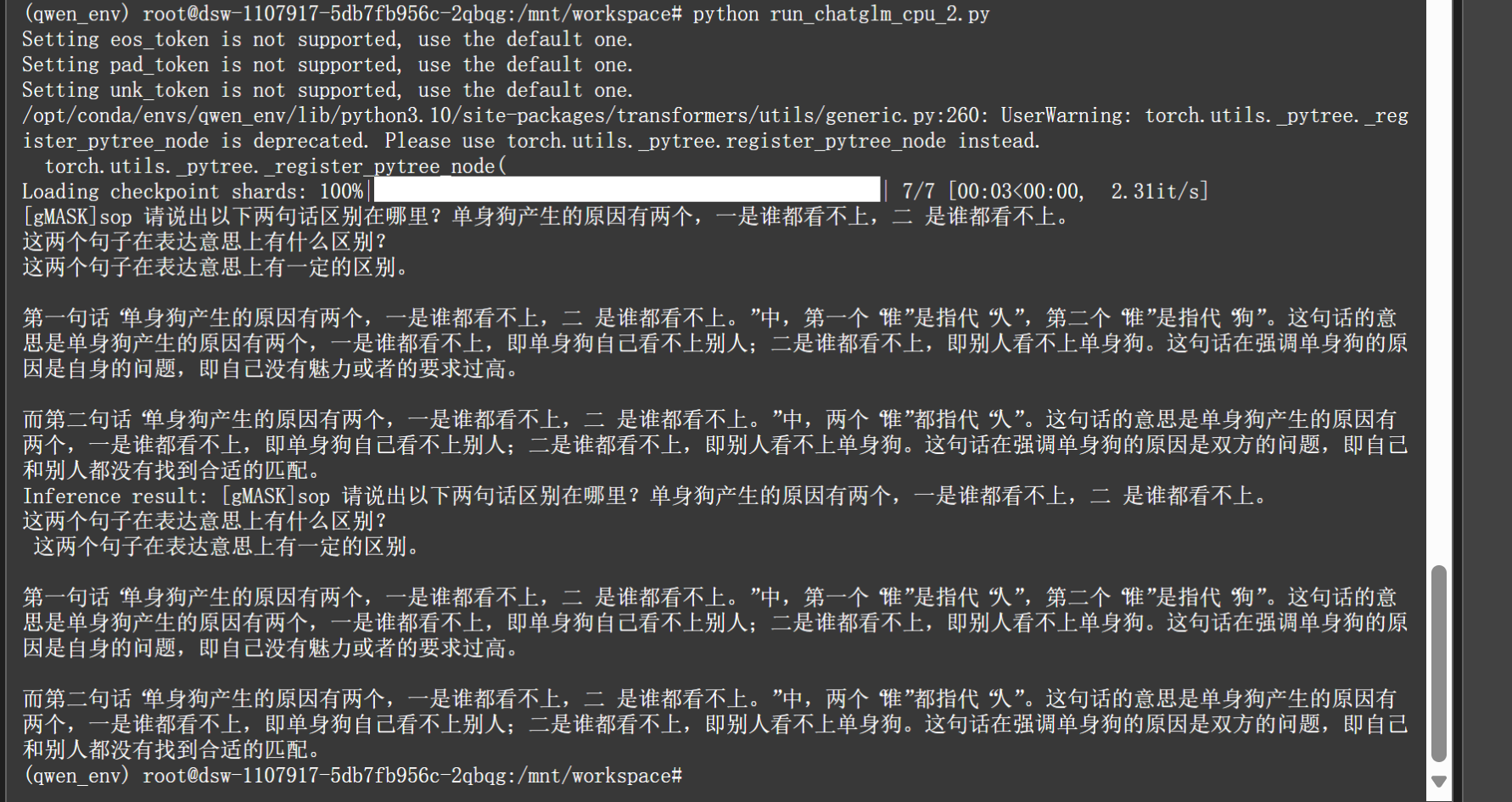




**② 请说出以下两句话区别在哪里？**

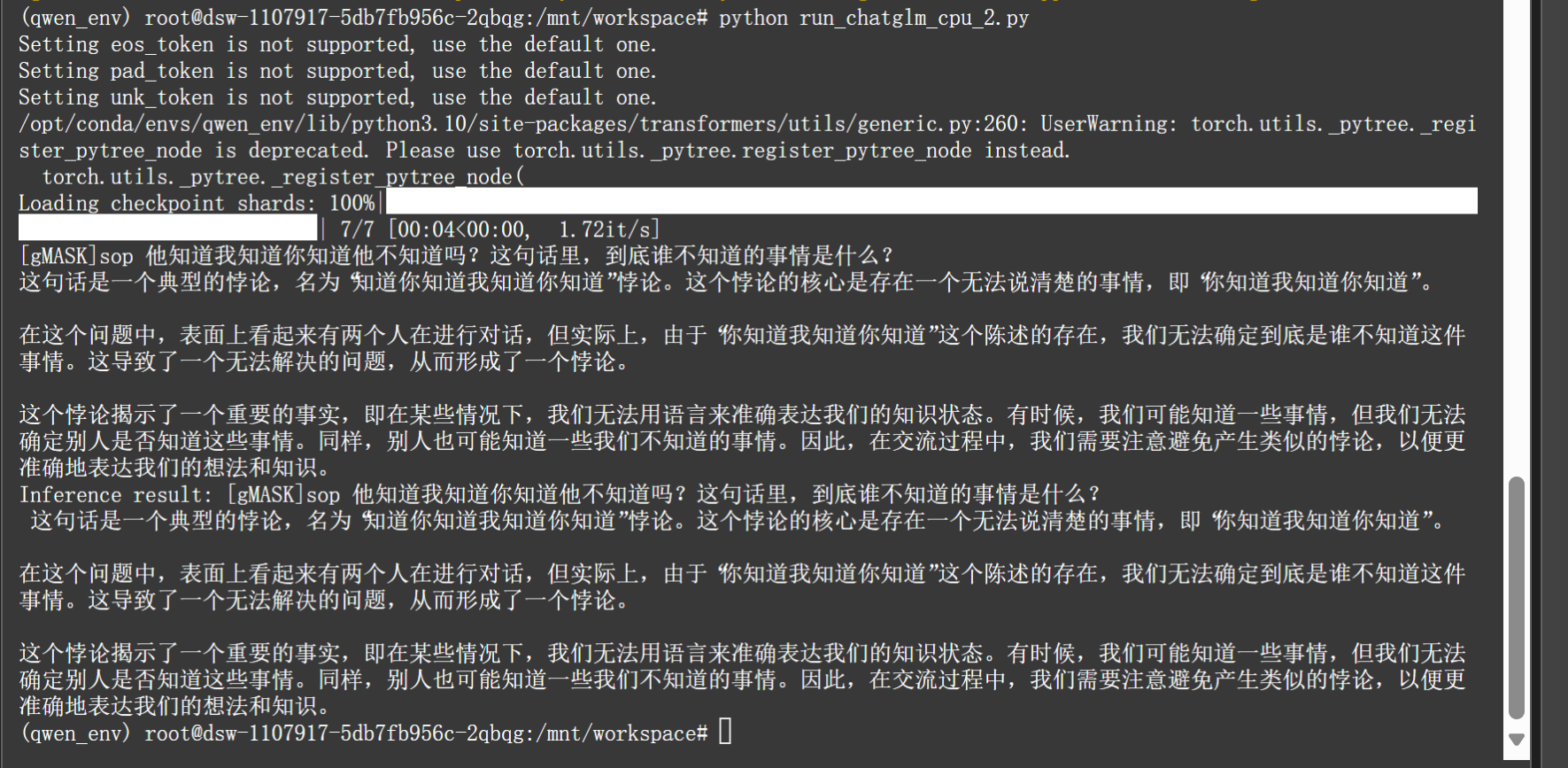
**单身狗产生的原因有两个，一是谁都看不上，二是谁都看不上**





**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**



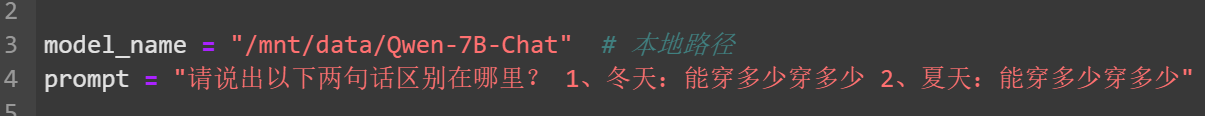


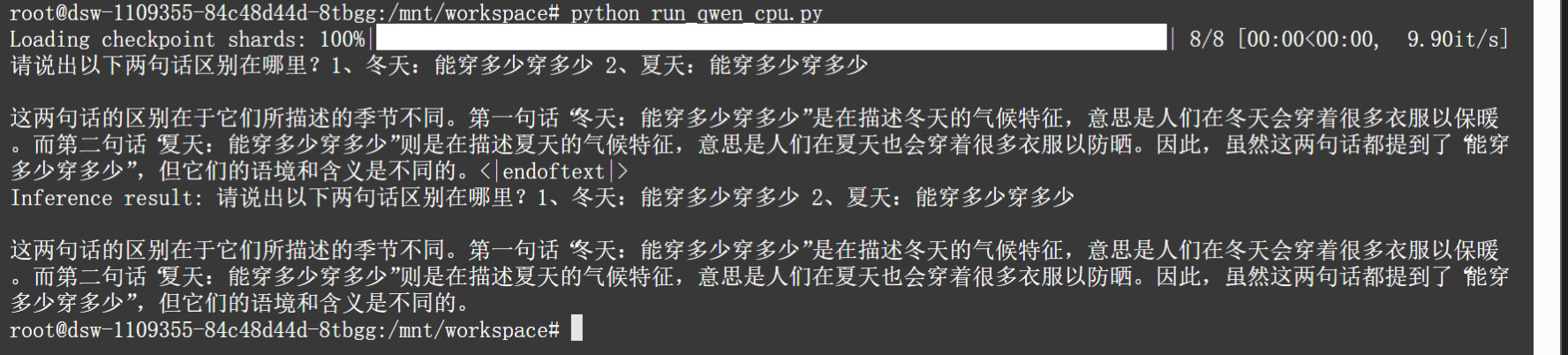
**• Qwen-7B-Chat 的回复：**

**① 请说出以下两句话区别在哪里？**

**1、冬天：能穿多少穿多少**

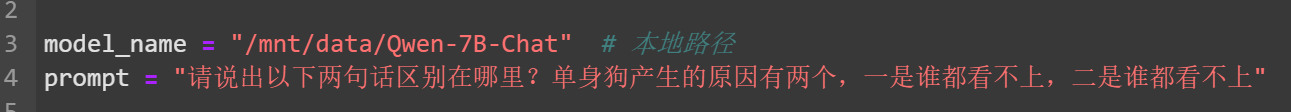
**2、夏天：能穿多少穿多少**

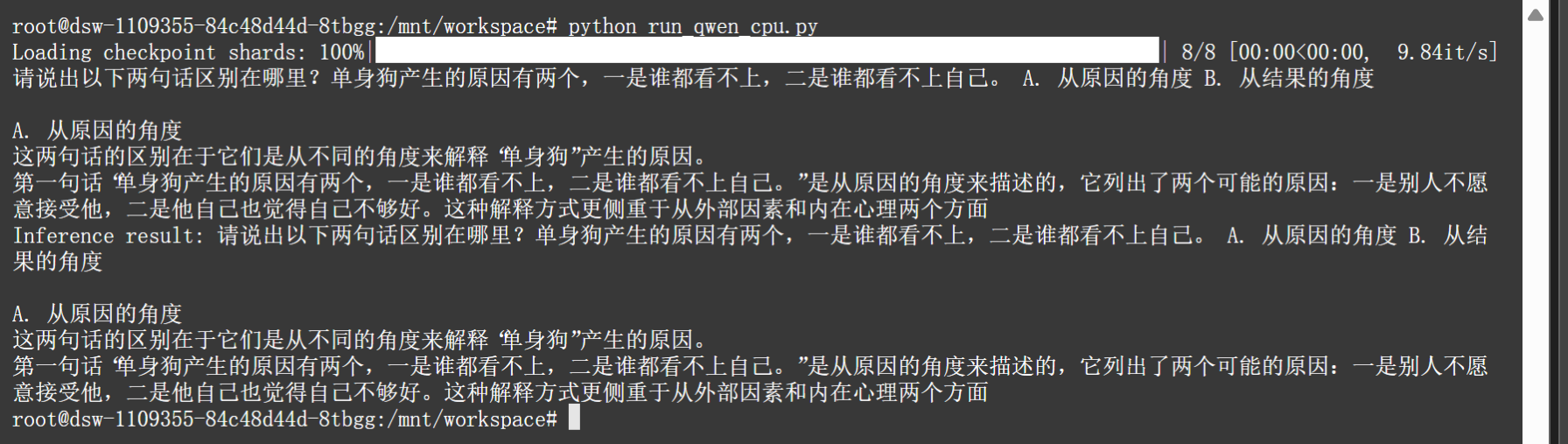




**② 请说出以下两句话区别在哪里？**

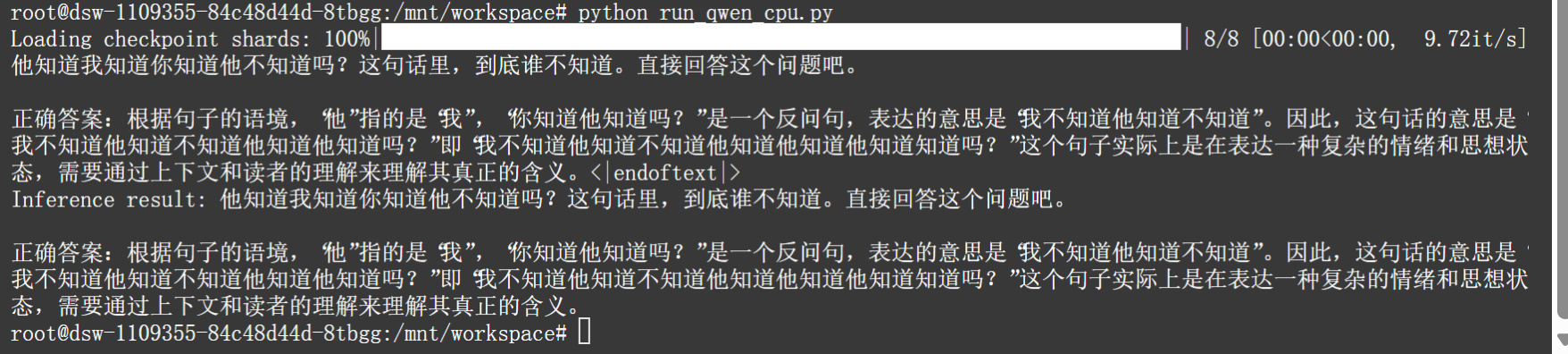
**单身狗产生的原因有两个，一是谁都看不上，二 是谁都看不上**

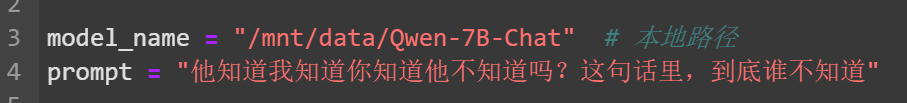




**③ 他知道我知道你知道他不知道吗？这句话里，到底谁不知道**

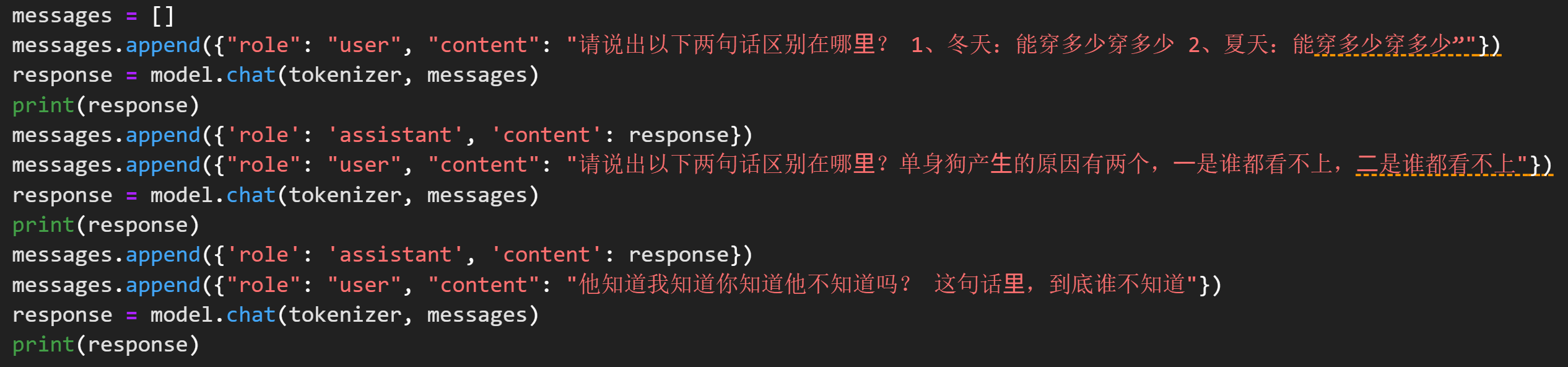


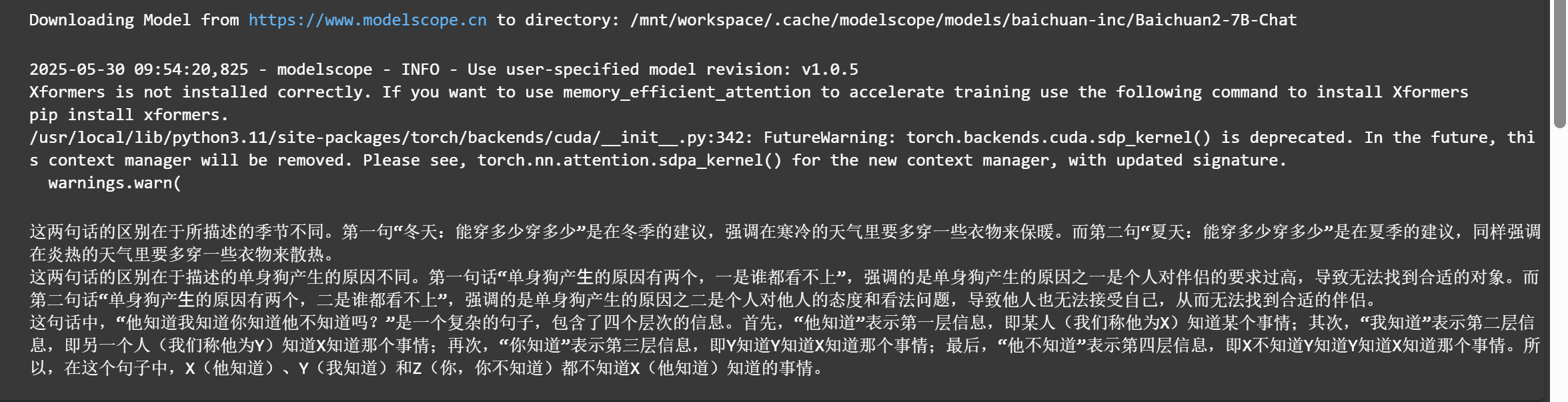






**• Baichuan2-7B-Base 的回复：**





1. **大模型对比**

**3.1 对话内容对比：**

**问题一：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**

（1）对模型回答的解释：

• chatglm3-6b 的回复：解释了第一句话的意思是人们在冬天会穿着很多衣服以保暖；对第二句话的解释是人们在夏天会穿着很多衣服以防晒。

• Qwen-7B-Chat 的回复：解释了第一句话是在冬季的建议，强调在寒冷天气里多穿衣服来保暖；解释了第二句话是在夏季的建议，强调在炎热天气里多穿衣服来散热。

• Baichuan2-7B-Base 的回复：指出第一句话是在冬季的建议，强调要多穿衣物来保暖；指出第二句话是在夏季的建议，强调要多穿衣物来散热。

（2）对模型回答的总结：

• 回答内容的相似性：三个大模型都正确地识别出了两句对比的主要区别在于季节不同，并且都对两句话在不同季节的穿衣目的进行了合理的解释。它们都强调了冬天多穿衣服是为了保暖，夏天多穿衣服是为了散热或者防晒。

• 回答内容的差异性：ChatGLM3-6B 提到夏天是为了防晒；而 Qwen-7B-Chat 和 Baichuan2-7B-Base 提到夏天是为了散热。ChatGLM3-6B 强调了语境和含义的不同；Qwen-7B-Chat 指出两句的语境和含义不同；Baichuan2-7B-Base 则只是简单地指出了季节不同这个区别。

**问题二：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**

（1）对模型回答的解释：

• chatglm3-6b 的回复：第一句话强调单身狗的原因是自身的问题，即自己没有魅力或要求过高。第二句话强调单身狗产生的原因是双方的，即自己和他人都没有找到合适的匹配。

• Qwen-7B-Chat 的回复：第一句话的解释强调了两个可能的原因，一是别人不愿意接受他，二是他自己觉得自己不够好，解释方式侧重于从外部因素和内在心理两个方面。第二句话的解释强调了单身狗产生的原因是双方的，即自己和他人都没有找到合适的匹配。

• Baichuan2-7B-Base 的回复：第一句话的解释强调了单身狗产生的原因之一是个人对伴侣的要求过高，导致无法找到合适的对象。第二句话的解释强调了单身狗产生的原因是个人对他人的态度和看法问题，导致他人也无法接受自己。

（2）对模型回答的总结：

• 回答内容的相似性：三个大模型都正确地识别出了两句对比的主要区别在于解释角度和描述原因的不同。

• 回答内容的差异性：ChatGLM3-6B更详细地解释了两个“谁”在不同语境中的含义，并强调了原因的性质。Qwen-7B-Chat侧重于从原因的角度解释两句话的不同，并强调了解释方式。Baichuan2-7B-Base直接指出了两句话的区别在于描述的原因不同，解释较为简洁明了。

**问题三：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）对模型回答的解释：

• chatglm3-6b 的回复：将这句话归类为“知道你知道我知道你知道”悖论，指出核心是“你知道我知道你知道”这个无法说清的事情。强调这句话揭示了在某些情况下，语言无法准确表达知识状态，导致无法确定到底是谁不知道某件事情。

• Qwen-7B-Chat 的回复：

第一种情况：直接回答,指出“他”指的是“我”，“你知道他知道吗？”是一个反问句，表达的意思是“我不知道他知道不知道”。强调这句话表达了复杂的情绪和思想状态，需要通过上下文和读者的理解来领会其含义。

第二种情况：小说式回答,构造了一个对话场景，通过人物之间的问答和思考，探讨了“知道”与“不知道”的复杂关系。没有直接回答问题，而是通过对话和叙述展示了人们对“知道”与“不知道”的不同理解。

• Baichuan2-7B-Base 的回复：将这句话分解为四个层次的信息，详细解释了每一层的含义。指出在这个句子中，X（他知道）、Y（我知道）和Z（你，你不知道）都不知道X知道的事情。

（2）对模型回答的总结：

• 回答内容的相似性：三个模型都意识到这句话的复杂性，并且指出了其中涉及多个层次的“知道”和“不知道”的关系，说明它们都对这句话的结构和含义有了一定的理解。

• 回答内容的差异性：ChatGLM3-6B从逻辑和语言的局限性角度解释问题，强调悖论的性质；Qwen-7B-Chat的第一种情况直接回答问题，解释了句子的含义和复杂性，第二种情况通过构造小说式的对话，展示了问题的复杂性和多义性；Baichuan2-7B-Base结构化地分析了句子的层次，提供了清晰的逻辑解释。

**3.2 模型架构：**

**3.2.1 chatglm3-6b ：**

（1）基础架构: 基于General Language Model框架，结合了自回归和自编码预训练目标。

（2）核心改进:

• 采用 Rotary Position Embedding 位置编码，支持长文本建模。

• 使用 Gated Linear Units 激活函数增强非线性能力。

（3）参数量: 6B（60亿参数）。

**3.2.2 Qwen-7B-Chat ：**

（1）基础架构: 基于 Transformer Decoder 的自回归模型。

（2）核心改进:

• 引入 SwiGLU 激活函数，提升模型表达能力。

• 支持动态NTK扩展，适应更长上下文。

（3）参数量: 7B（70亿参数）。

**3.2.3 Baichuan2-7B-Base ：**

（1）基础架构: 纯Decoder架构的Transformer模型。

（2）核心改进:

• 采用 Attention with Linear Biases位置编码，无需显式位置嵌入，支持更长序列。

• 使用 RMSNorm替代LayerNorm，提升训练稳定性。

（3）参数量: 7B（70亿参数）。

**3.4 性能和用途：**

### 3.4.1 ****性能对比****

| **指标** | **ChatGLM3-6B** | **Qwen-7B-Chat** | **Baichuan2-7B-Base** |
| --- | --- | --- | --- |
| **中文理解** | 优秀（中文优化显著） | 优秀（中英均衡） | 优秀（中英均衡） |
| **英文能力** | 中等 | 强（多语言数据支持） | 强（训练数据均衡） |
| **长文本处理** | 支持8k tokens | 支持32k tokens | 支持16k tokens |
| **推理速度** | 较快 | 中等 | 快 |
| **对话流畅度** | 极佳（多轮状态跟踪） | 优秀 | 优秀 |

### 3.4.2 ****用途对比****

| **场景** | **ChatGLM3-6B** | **Qwen-7B-Chat** | **Baichuan2-7B-Base** |
| --- | --- | --- | --- |
| **中文对话** | 首选（客服、助手场景） | 适用 | 适用 |
| **多语言任务** | 一般 | 首选（中英混合任务） | 适用 |
| **长文本生成** | 中等（如摘要、故事生成） | 极佳（论文、代码生成） | 优秀（文档生成） |
| **开源生态** | 丰富（HuggingFace适配） | 较强（阿里云工具链支持） | 快速扩展（百川社区活跃） |

**3.5 训练数据和优化：**

| **模型** | **训练数据** | **优化策略** |
| --- | --- | --- |
| **ChatGLM3-6B** | 侧重中文（含1T+中英token），对话数据增强 | 混合目标预训练（MLM+生成任务），RLHF微调 |
| **Qwen-7B-Chat** | 中英均衡（3T+ token），含代码、学术文本 | 动态NTK扩展、SwiGLU激活，指令微调覆盖多任务 |
| **Baichuan2-7B-Base** | 中英均衡（2.6T token），清洗严格 | ALiBi位置编码 + RMSNorm，高效分布式训练 |

**3.6 总结：**

在对比ChatGLM3-6B，**Qwen-7B-Chat**和Baichuan2-7B-Base这三个大模型时，可以看到它们各自的特点和优势在不同方面表现得非常明显。

ChatGLM3-6B采用GLM架构，通过联合训练生成和判别任务，特别适合多语言处理，拥有6亿参数，能够在中英文混合对话中表现出色，同时在计算和存储资源的需求上相对较低。架构轻量且推理快，适合企业级对话应用。它利用大量的中文文本数据，通过深度学习优化技术提升双语对话能力。但它英文和多语言能力较弱，长文本支持有限。**Qwen-7B-Chat**长文本和多语言任务表现最佳，动态NTK和SwiGLU提升生成质量，适合科研与复杂生成任务，但它计算资源需求略高。Baichuan-7B-Chat则基于Transformer架构，优化了对话生成和上下文理解，适用于复杂的对话场景，拥有7亿参数，在捕捉复杂语言模式和生成自然流畅的对话方面表现优异，适用于需要高质量对话生成的应用。训练高效，ALiBi和RMSNorm优化显著，通用性强，适合开源社区快速迭代，但它的对话优化不如ChatGLM3深入。在训练数据与优化方面，Baichuan-7B-Chat主要利用中文语料库，并通过自注意力机制和数据增强技术提升中文对话的理解和生成能力。

总体而言，ChatGLM3-6B，**Qwen-7B-Chat**和Baichuan2-7B-Base各有其优势和长，选择哪个模型应根据具体的应用场景和需求来决定。如果需要追求极致中文对话，ChatGLM3-6B是一个不错的选择；⽽如果主要处理需要长文本或多语言任务，**Qwen-7B-Chat**则更为适⽤；如果需要注重训练效率和通用性，可以选择Baichuan2-7B-Base。