大语言模型部署实验报告

作 者 姓 名： 虞澍

学 号： 2352979

指 导 教 师： 邓浩

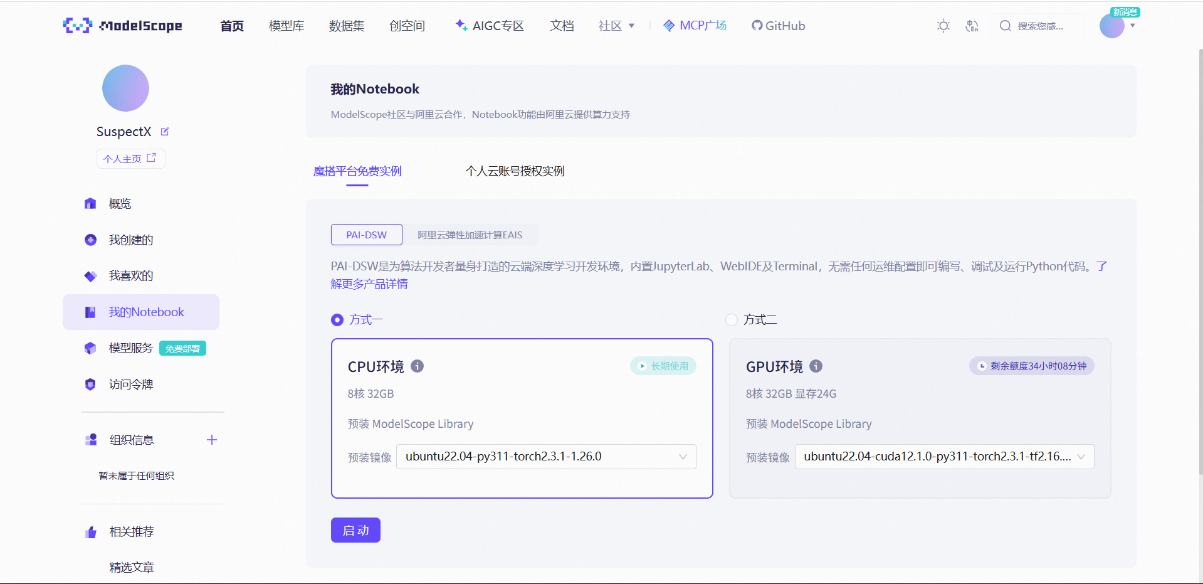
专 业： 软件工程

同济大学

Tongji University

**一、部署过程**

1. **登录并使用魔搭平台，获取CPU云计算资源；**

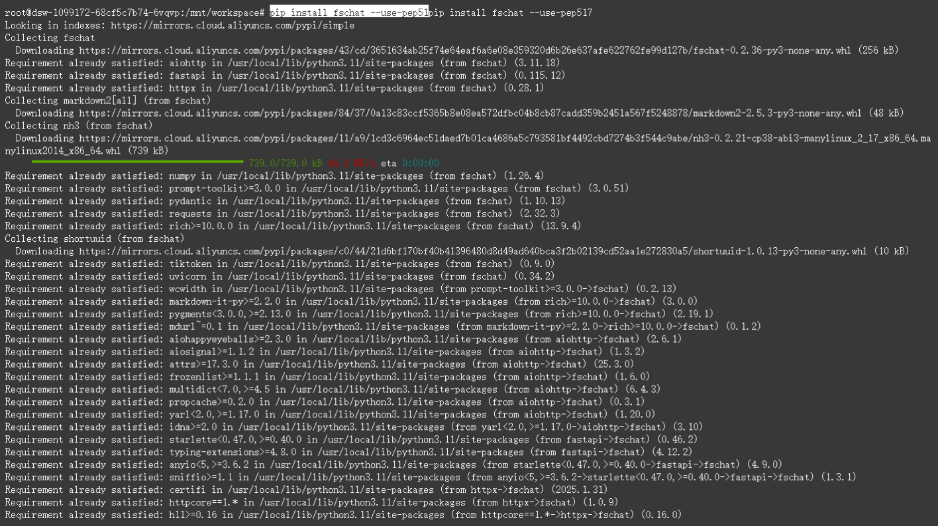


1. **通过Jupyter Notebook进行环境配置**

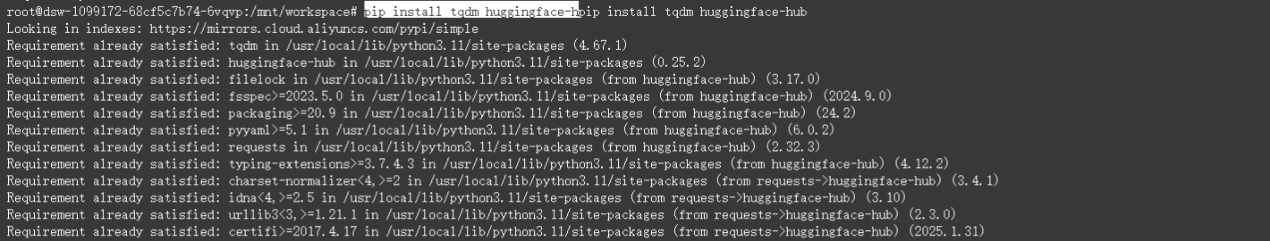
安装基础依赖



安装fschat

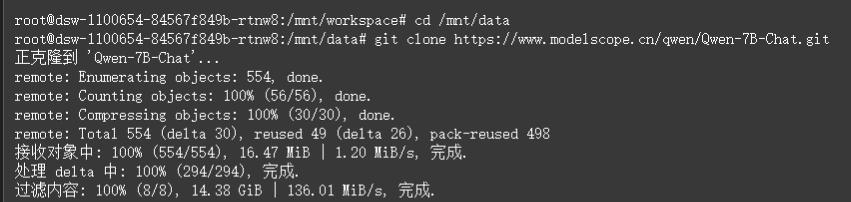


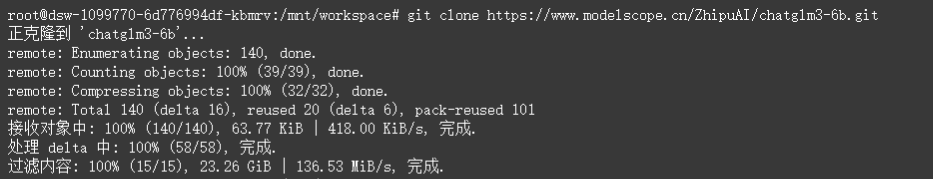
安装 tqdm、huggingface-hub

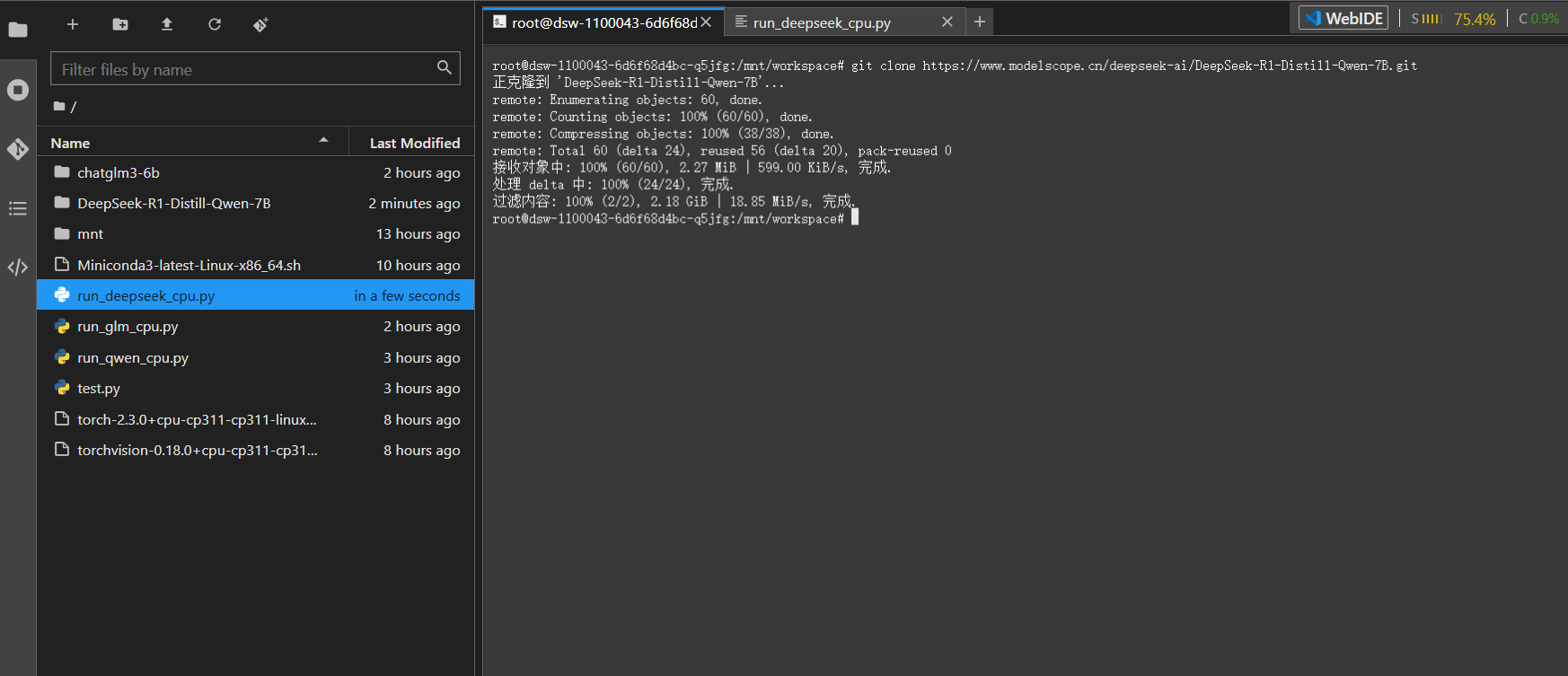


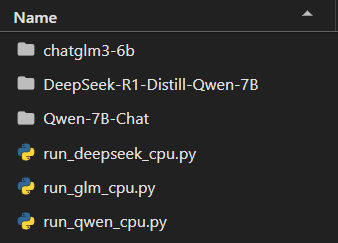
1. **LLM下载**

切换至工作目录，git clone下载中文大模型至本地



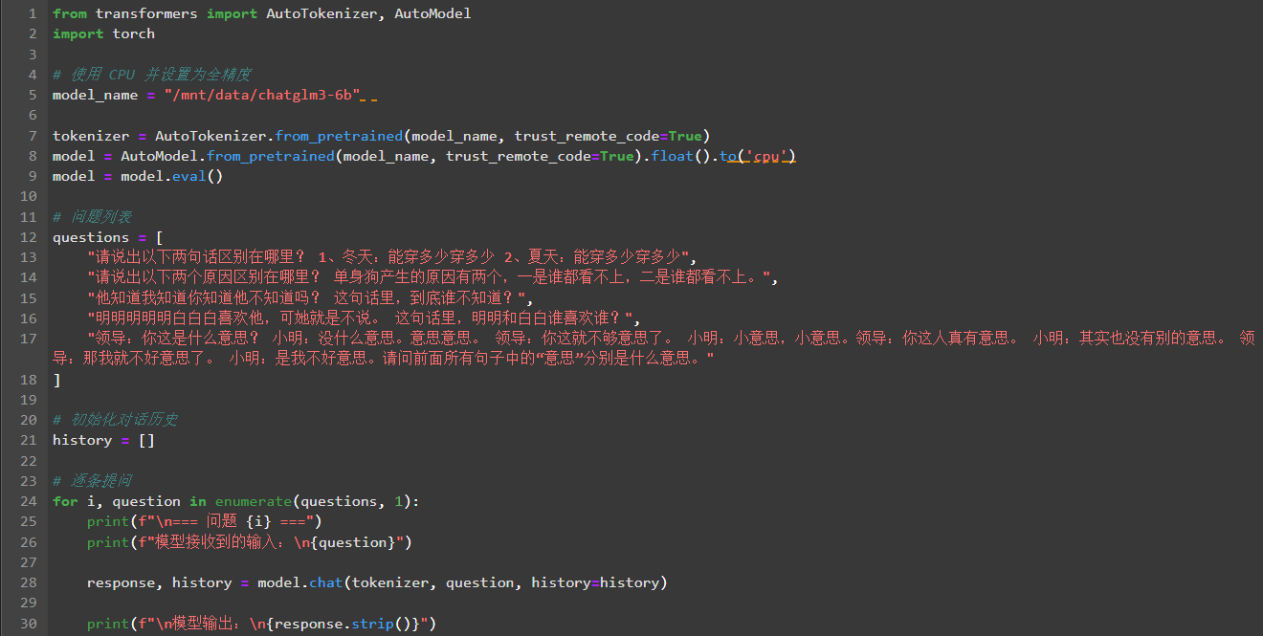






**二、编写Python测试程序并进行问答**

1. **基于chatglm3-6b模型**

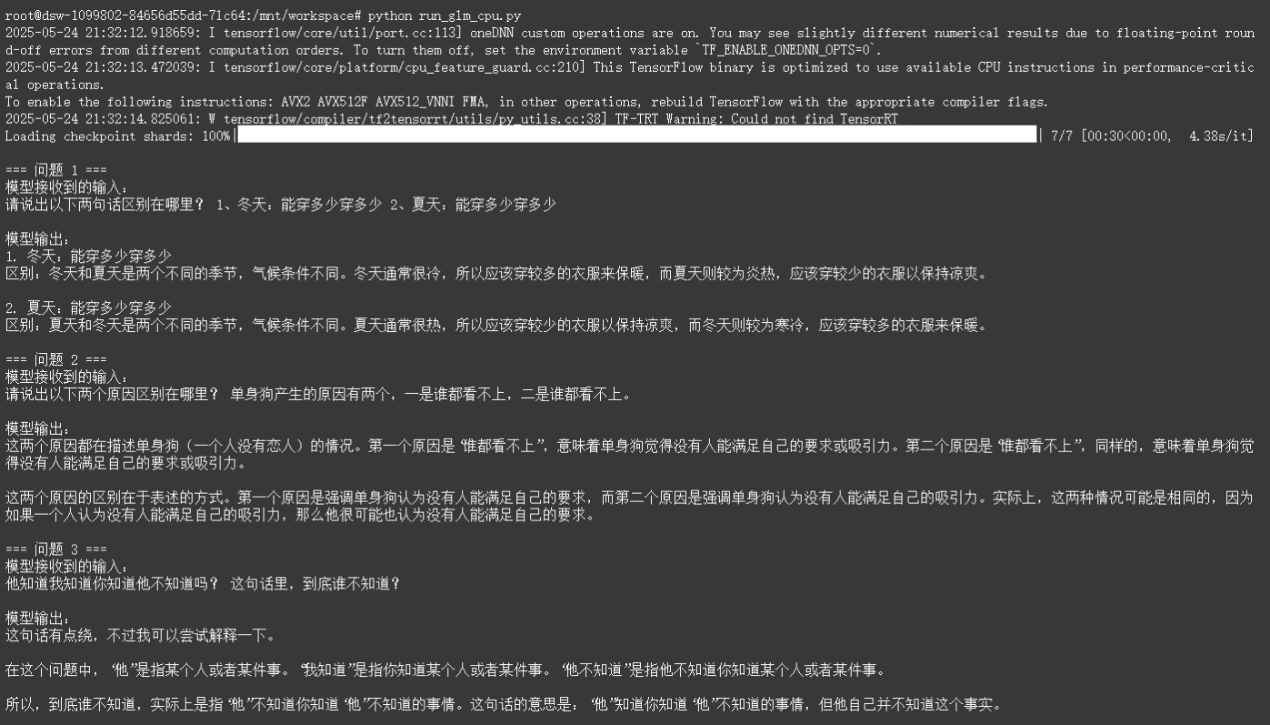


得到回答：

1）



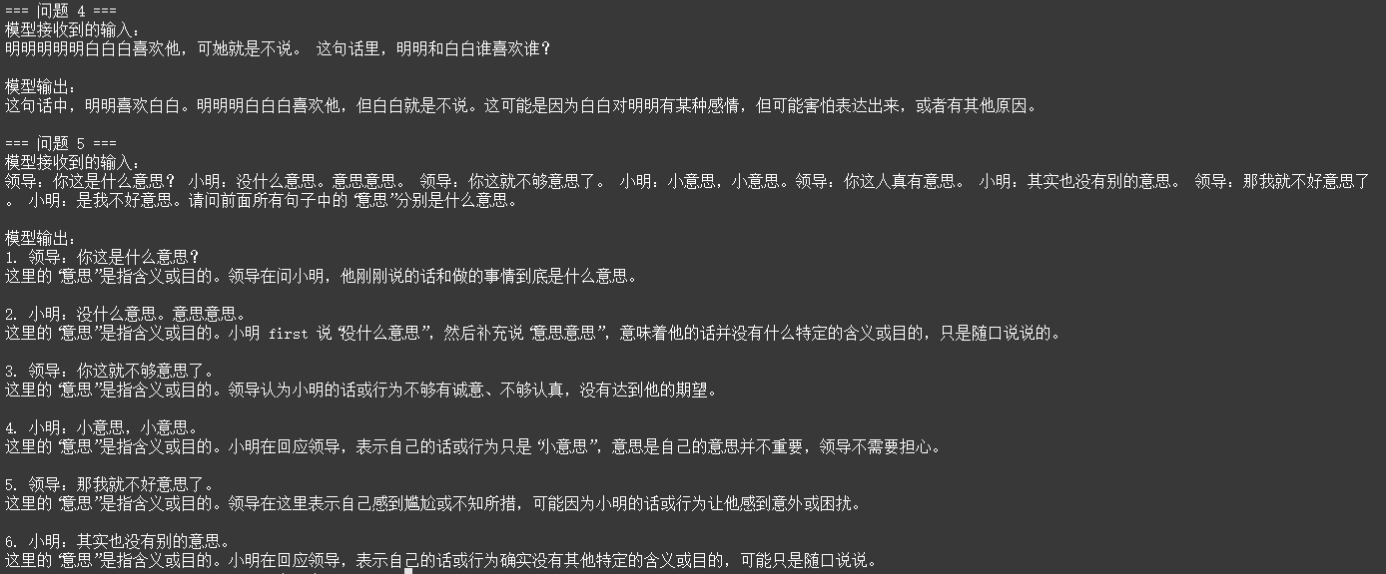
2）



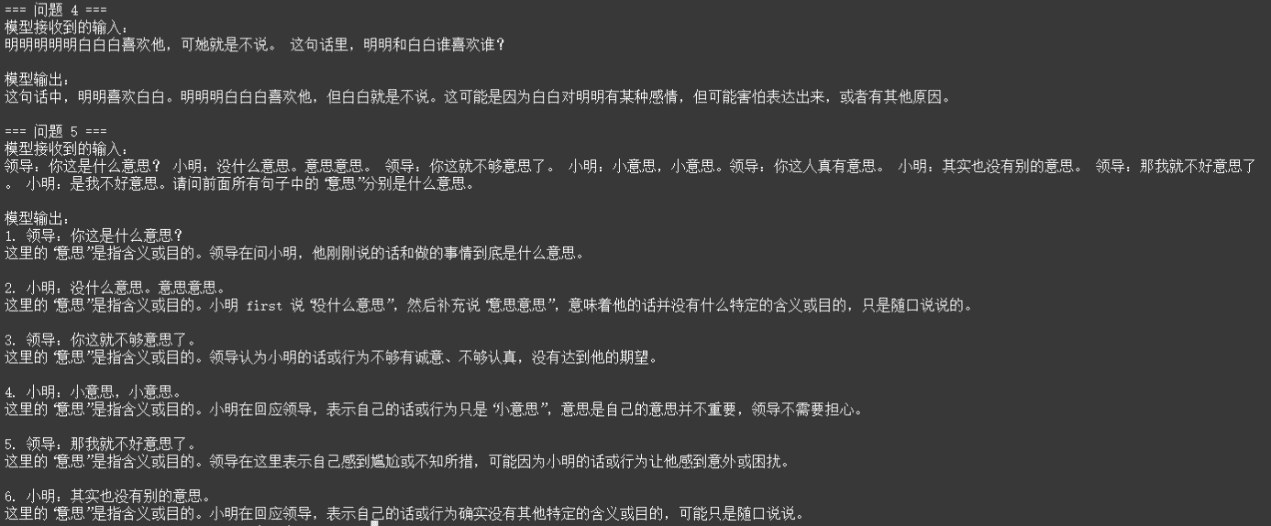
3）



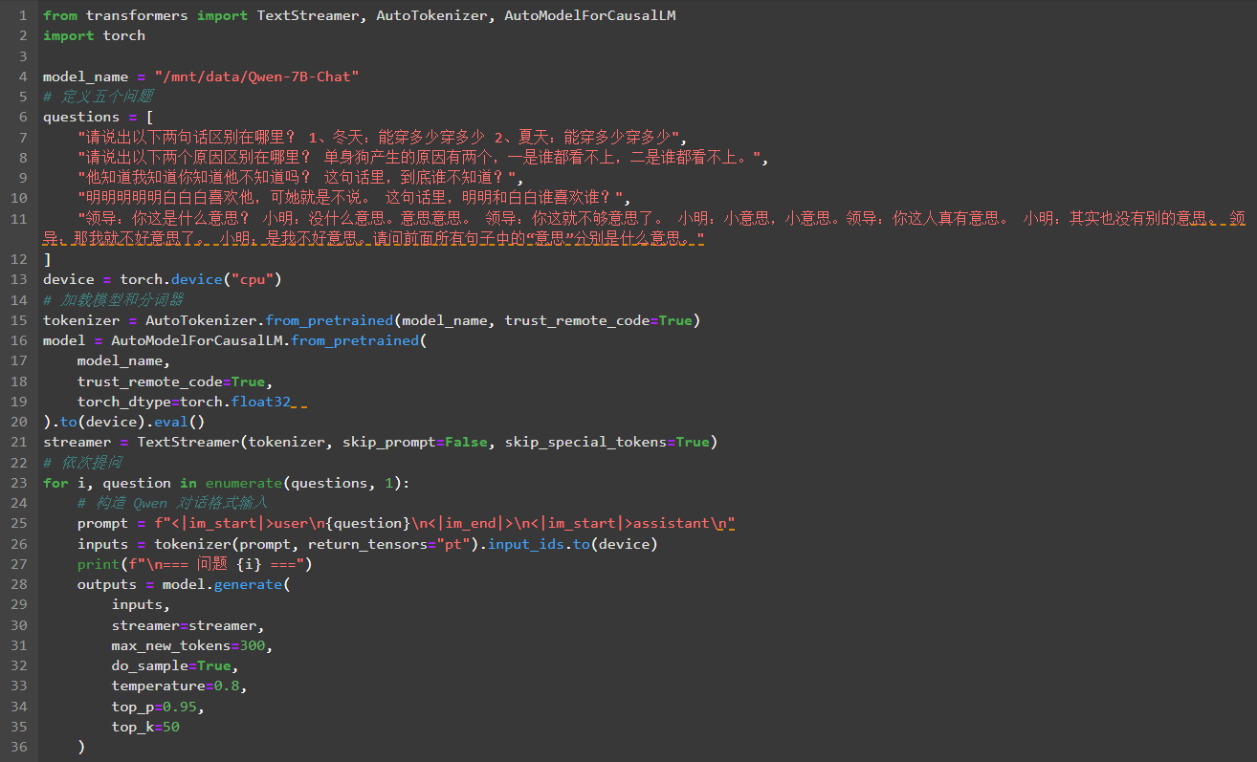
4）



5）

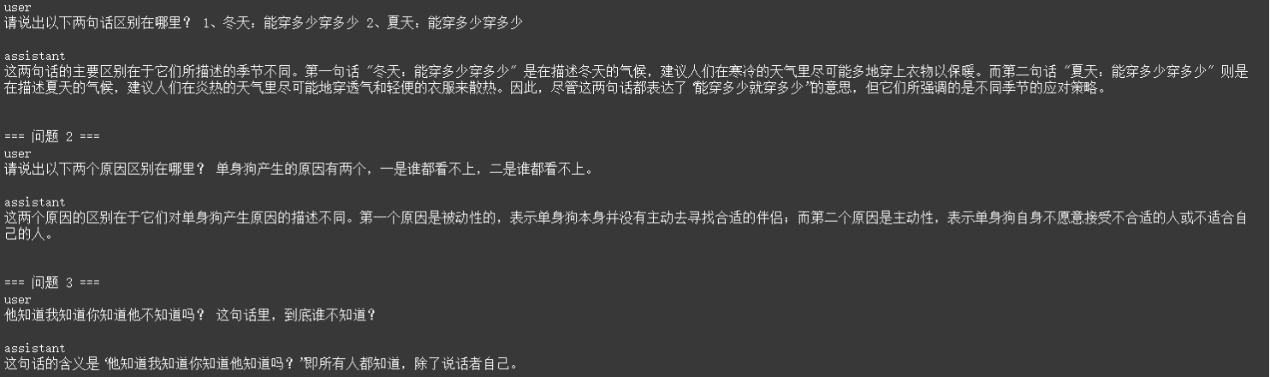


1. **基于Qwen-7B-Chat模型**

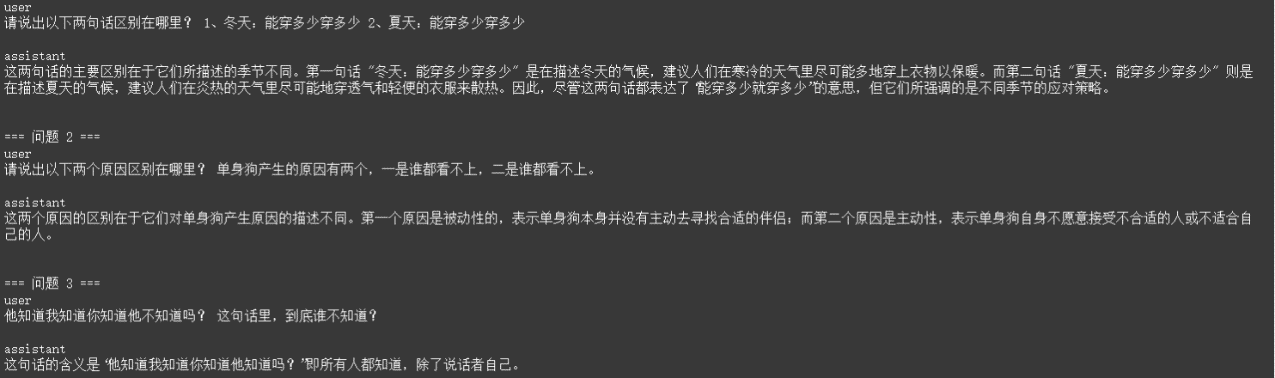


得到回答:

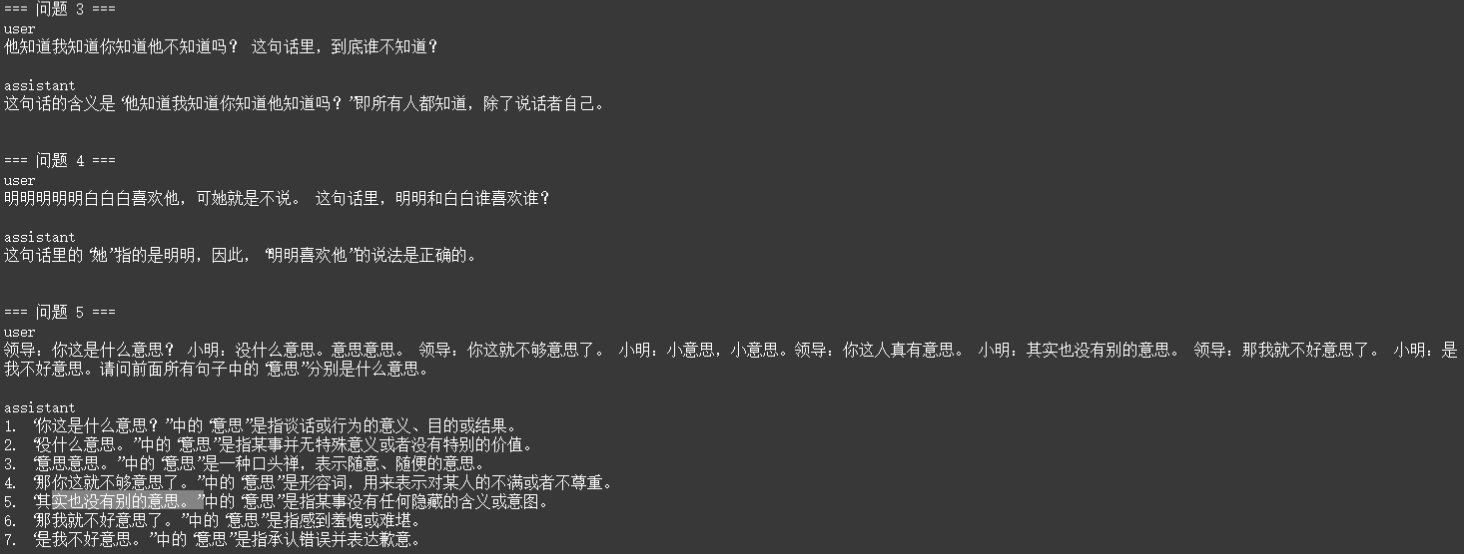
1）



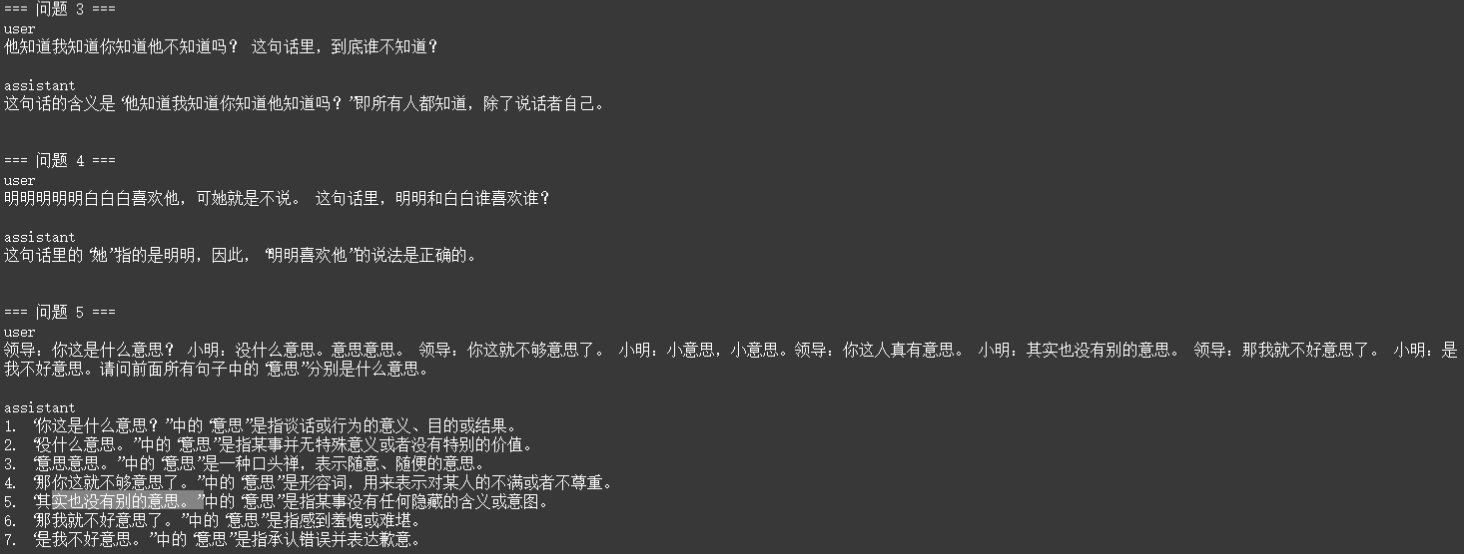
2）



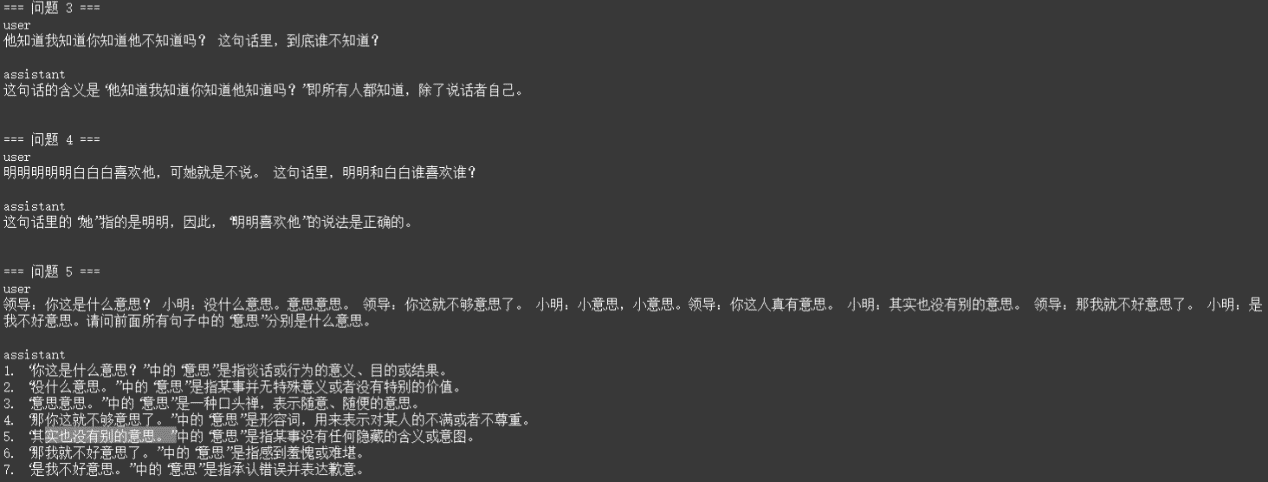
3）



4）



5）

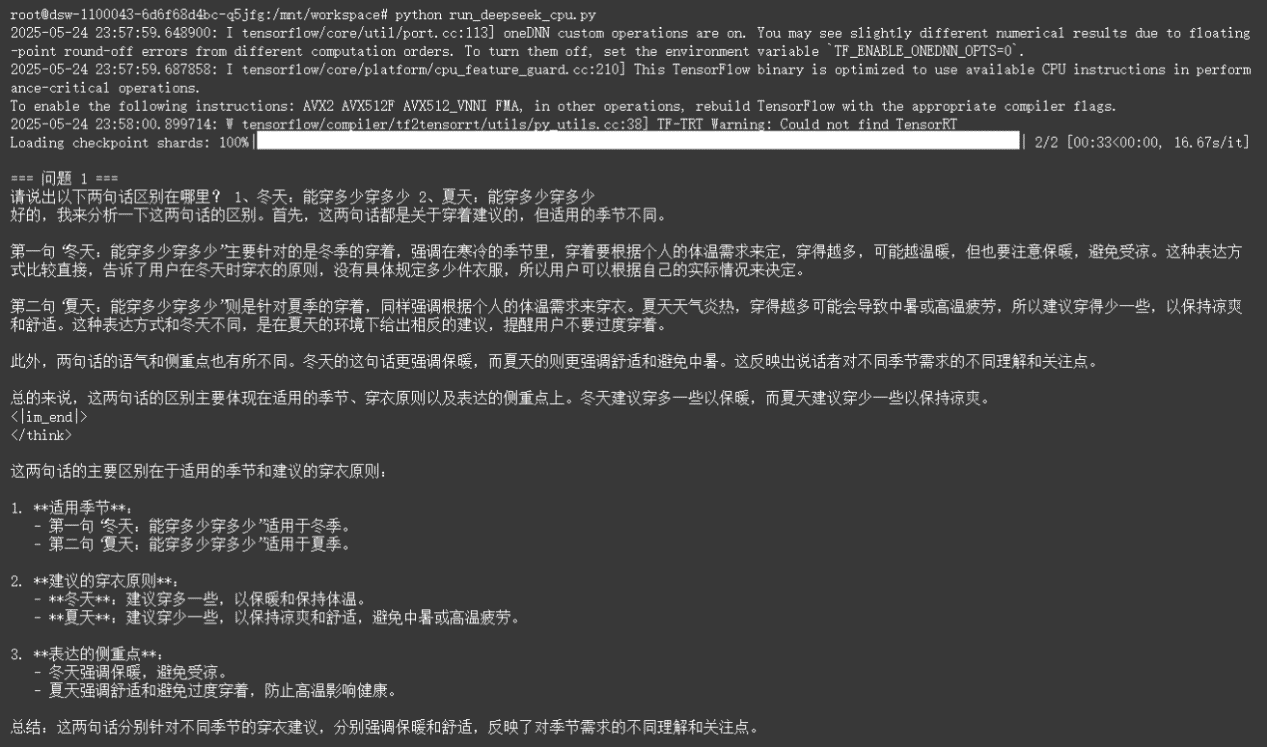


1. **基于DeepSeek-R1-Distill-Qwen-7B模型**

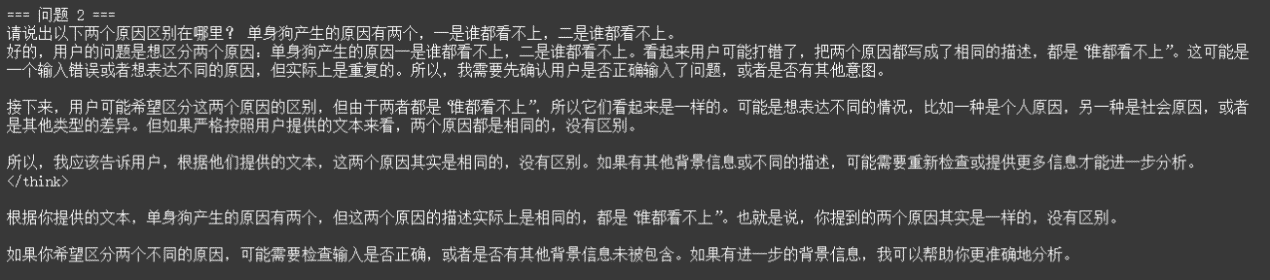


得到回答:

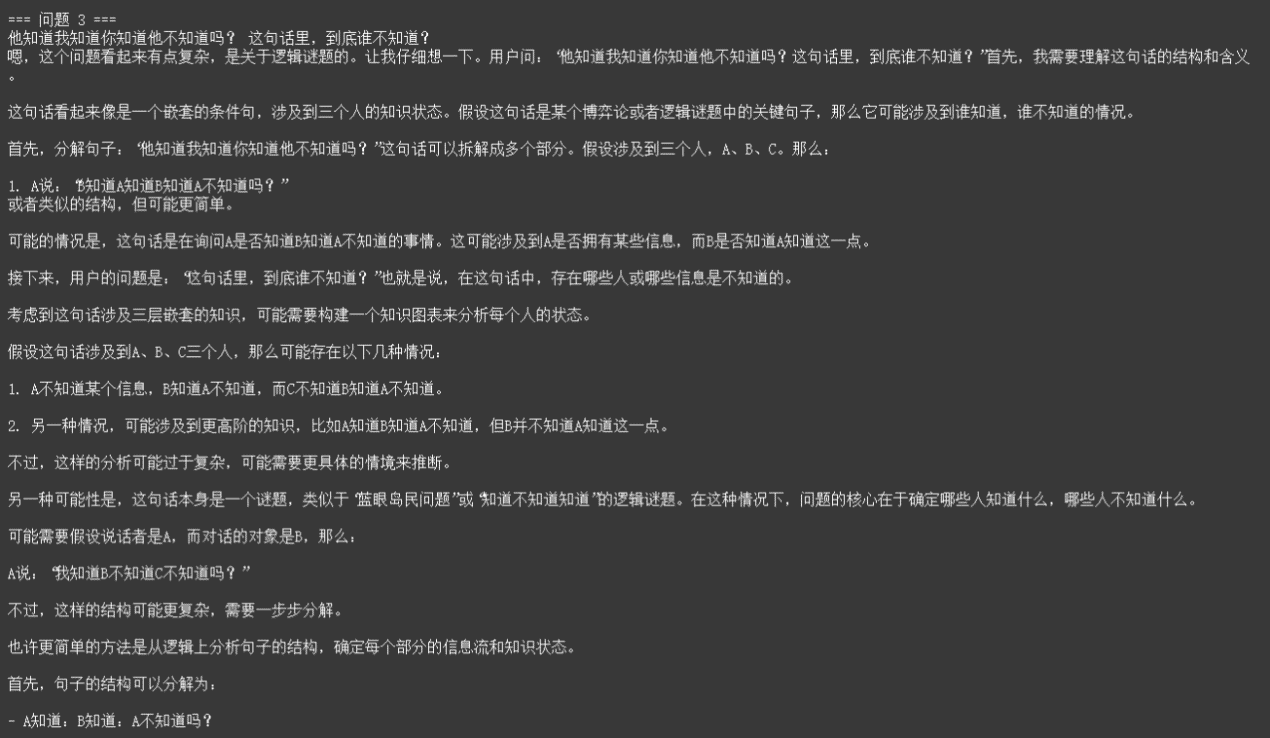
1）

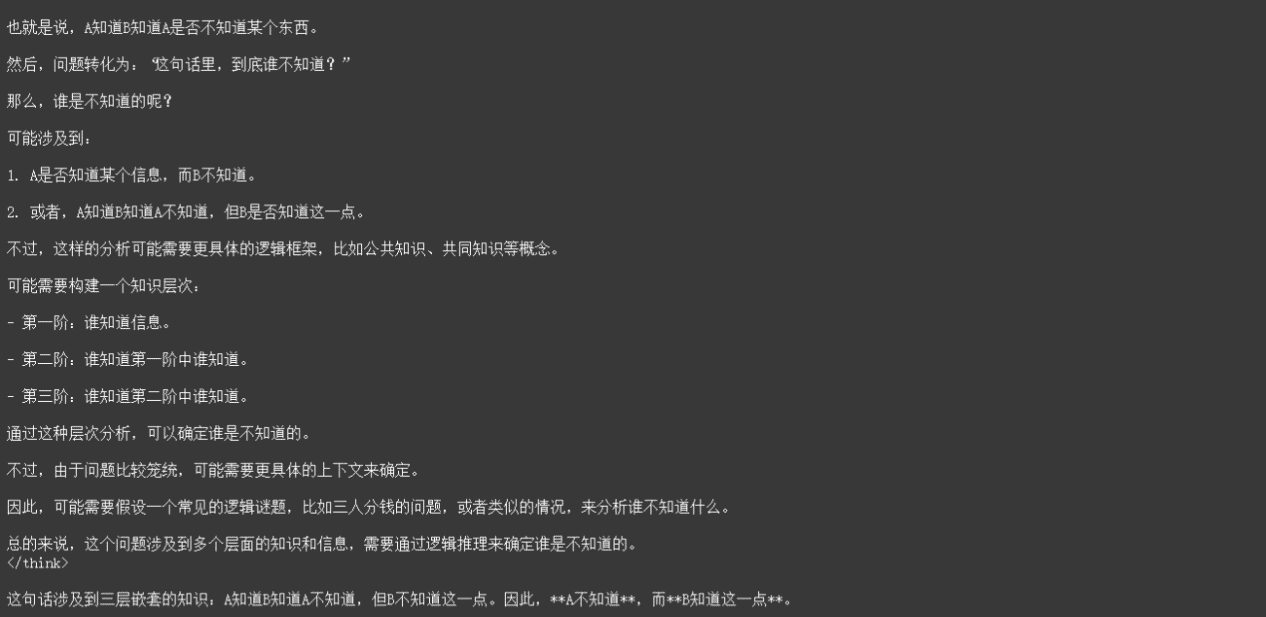


2）

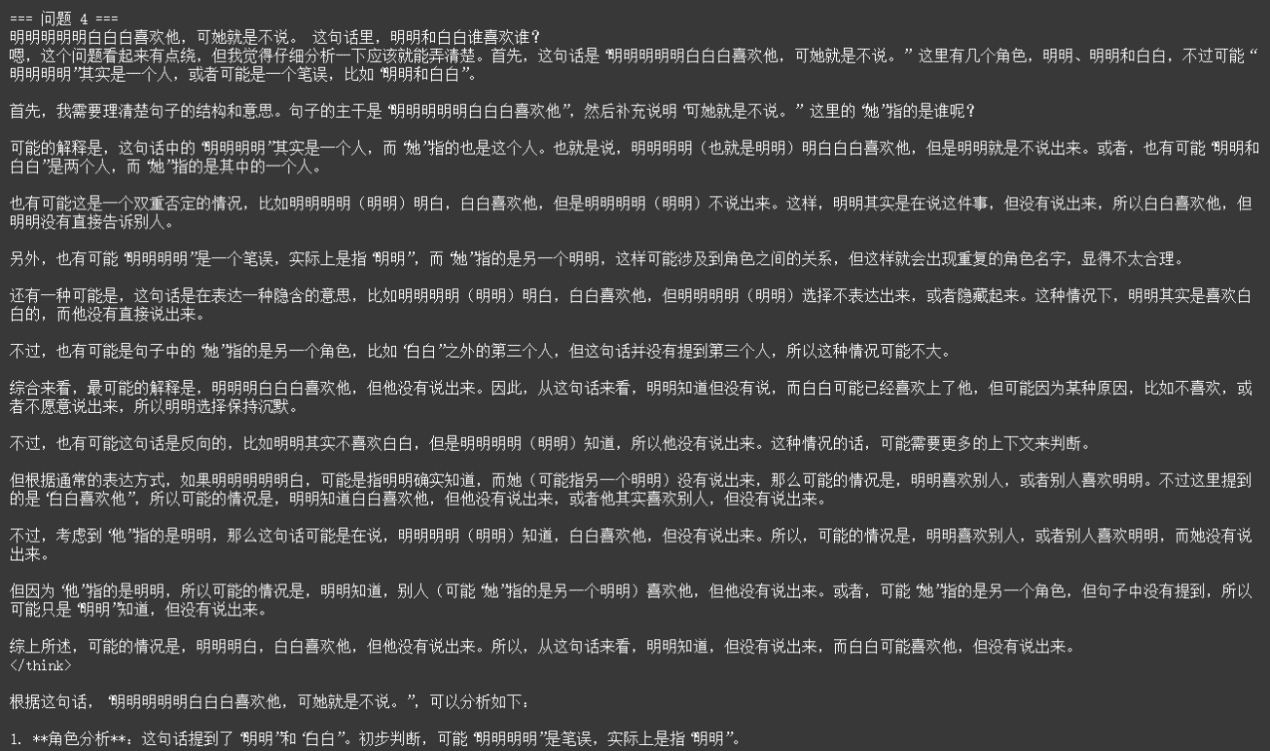


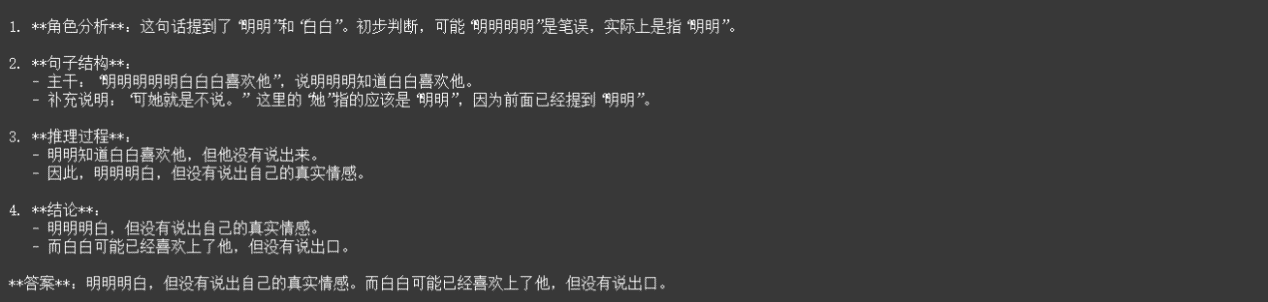
3）



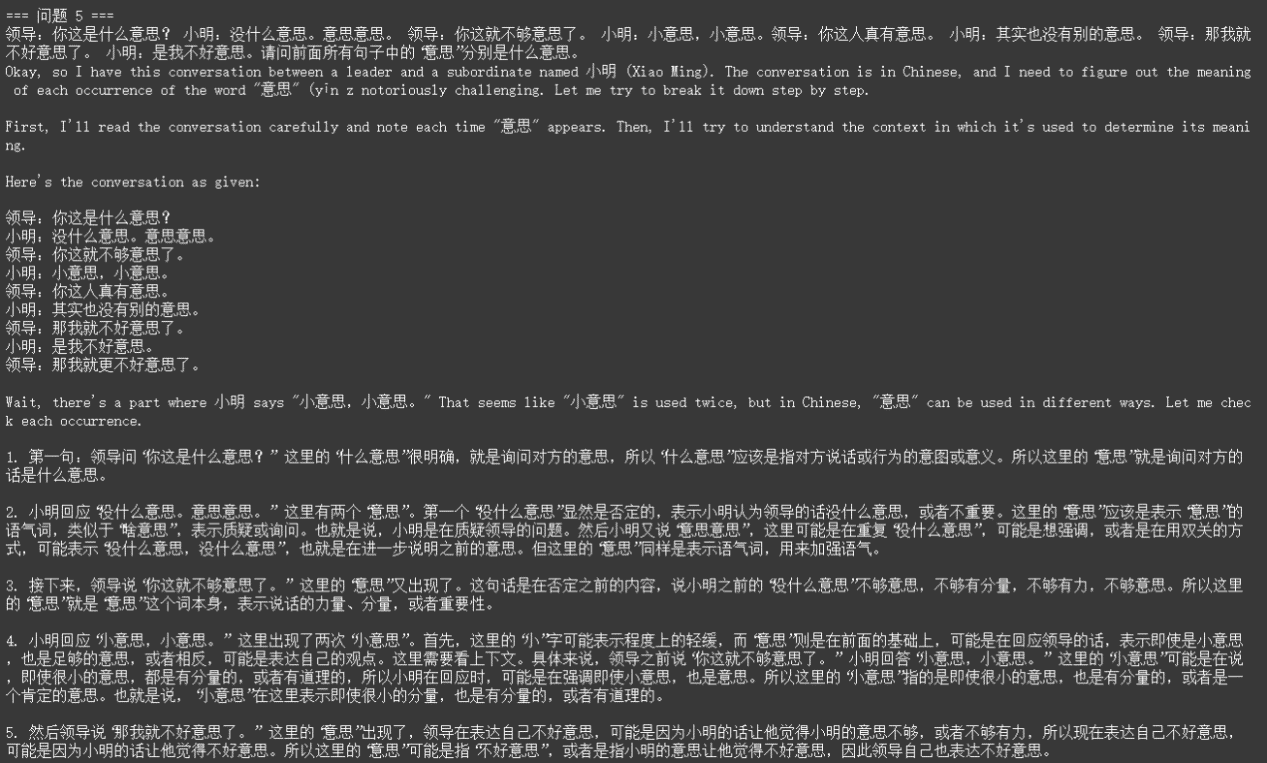
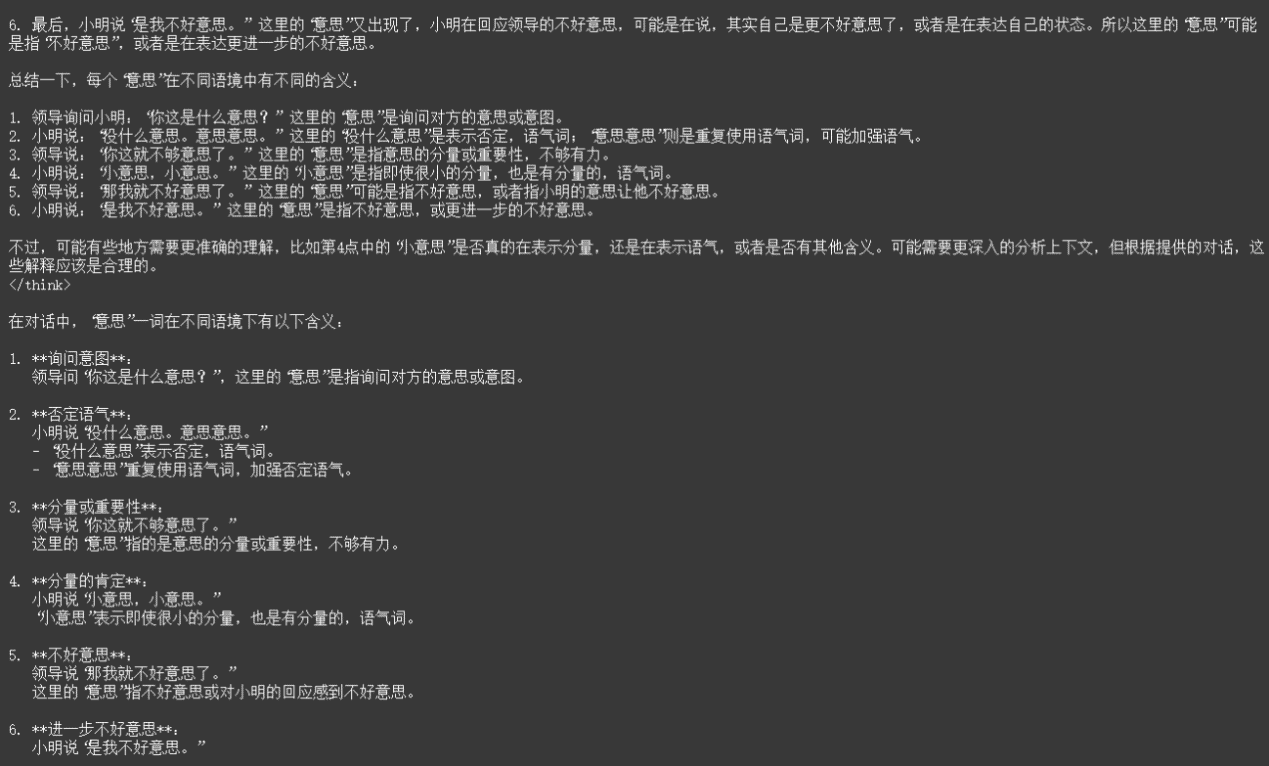
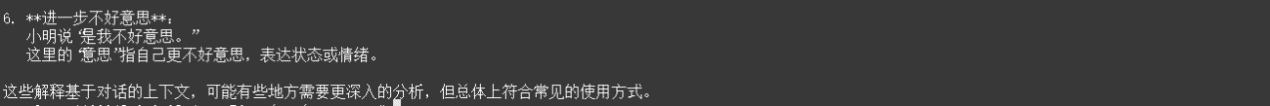


4）





5）

**三、横向对比分析：ChatGLM3-6B、Qwen-7B-Chat、Baichuan2-7B-Chat**

1. **模型架构对比**
2. ChatGLM3-6B 采用的是清华智谱开发的 GLM 架构，是一种融合了编码器与解码器的双向建模结构，但在生成时只用解码器部分。它使用了 RoPE（旋转位置编码）和 GELU 激活函数，参数量约为 60 亿。整体架构偏向中英文混合任务设计，支持多种任务类型。
3. Qwen-7B-Chat 是阿里达摩院推出的大模型，结构上类似 GPT，采用标准的 Transformer 解码器架构，但融合了一些自研优化（如更强的 tokenizer、数据清洗等），激活函数为 SwiGLU，参数量为 70 亿，专为中文对话优化，模型容量大、语义理解强。
4. DeepSeek-R1-Distill-Qwen-7B 是 DeepSeek 团队对其主力模型 R1 所进行的一次蒸馏压缩，架构整体保持 GPT 风格（与 Qwen-7B 接近），主要通过模型剪枝和知识蒸馏等方式减小参数规模和计算开销。虽然也是 70 亿参数，但由于蒸馏优化，其推理效率和负载明显更轻。
5. **训练数据与语言能力对比**
6. ChatGLM3-6B 的训练数据来自清华团队收集的大规模中英文语料，涵盖网页内容、百科、对话数据、代码、论文等，数据量级在万亿 tokens 以上。它的中文能力强，英文能力也不弱，是一个双语能力比较均衡的模型。
7. Qwen-7B-Chat 更专注中文，主要训练数据来源包括中文网页、问答社区、百科知识库和多轮对话语料，数量级也在万亿 tokens 以上。它的语言理解和生成能力在中文环境中表现尤为出色，英文能力相对弱一些，尤其在开放域问答中。
8. DeepSeek-R1-Distill-Qwen-7B 的训练方式是通过蒸馏从 DeepSeek-R1 主模型迁移而来，其语料并不公开完整细节，但可以推测主要包括 QA 类任务数据、数学推理数据、多轮对话和代码等。由于是蒸馏模型，训练数据的覆盖面不如前两者广，但由于继承了大模型的知识，其在多种任务上依然表现稳定。
9. **推理能力与对话风格**
10. ChatGLM3-6B 在对话连贯性、多轮理解和逻辑表达上有不错的表现，语气自然偏中性，适合用于结构化问答、对话机器人和英文写作辅助。它在多轮对话保持上下文一致性方面也有一定能力。
11. Qwen-7B-Chat 在中文对话领域非常强，特别擅长带有幽默感、人类风格的表达。例如写段子、解释网络梗或应对中文复杂句式时尤为灵活。它在“类人化”表达方面是这三者中最突出的一个。
12. DeepSeek-R1-Distill-Qwen-7B 更偏向“冷静理性型”，虽然不如 Qwen 在风格拟人上那么灵活，但逻辑推理、代码解读、事实问答的准确率很高。输出通常简洁有力，适合做技术问答、教育辅助、文案草稿等任务。
13. **推理资源与部署效率**
14. ChatGLM3-6B 在 float16 模式下需要大约 13GB 显存才能顺畅运行，加载速度中等，部署起来并不算轻量。但它支持 INT4、INT8 等量化方式，可在消费级显卡上做适当压缩部署。
15. Qwen-7B-Chat 的显存需求稍高，float16 模式下需 14GB 以上显存，且由于 tokenizer 的特殊性，对 tokenizer 的加载速度和显存也有一定要求。虽然可以量化，但在部署方面略重于 ChatGLM。
16. DeepSeek-R1-Distill-Qwen-7B 在这三者中是推理效率最高的，float16 模式下约 10GB 显存即可运行，加载快、响应快，非常适合用于边缘部署或开发场景下频繁加载测试。它的模型压缩得当，推理性能和质量保持了良好平衡。
17. **针对具体问答的对比分析**
18. 冬天：能穿多少穿多少 vs. 夏天：能穿多少穿多少

ChatGLM3-6B： 解释准确但有点重复，一句话正着说一遍，反着又说一遍，缺乏深入分析，没有解释清楚“多少”和“多‘少’”在两种语境下的区别。

Qwen-7B-Chat：描述了季节不同，冬天需要尽可能多穿以保暖，夏天需要尽可能少穿以保持凉爽。解释清晰，准确地指出了反向意义，且逻辑清晰、理由充分。

DeepSeek-R1-Distill-Qwen-7B：描述了季节不同，冬天需要尽可能多穿以保暖，夏天需要尽可能少穿以保持凉爽。解释清晰，准确地指出了反向意义，且逻辑清晰、理由充分，只是稍显啰嗦。

1. 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

ChatGLM3-6B：逻辑混乱，术语使用不准确；表述重复且啰嗦；没有有效传达新的观点或区分点。

Qwen-7B-Chat：没看出语义的设计，把“谁都看不上”和“谁都看不上”之间强行解读成“主动”和“被动”的差别，却没有解释出语法结构或主语变化，所以是瞎圆场+生搬硬套。

DeepSeek-R1-Distill-Qwen-7B：无法正确理解语义，将两句话认为是一样的内容，但给出了建议：如果希望区分，提供更多信息。

1. 他知道我知道你知道他不知道吗？

ChatGLM3-6B：似乎给出了答案：“他不知道”，但解释过程重复绕弯并没有说服力，效果不算特别好。

Qwen-7B-Chat：没有理解嵌套结构，给出了错误答案“我不知道”。

DeepSeek-R1-Distill-Qwen-7B：问题理解有误，逻辑前后矛盾，先是说可以分为3个人ABC，但是后面只出现了AB，明显是将句子中的人物混淆，无法处理复杂的嵌套逻辑，表现不佳。

1. 明明明明明白白白喜欢他，可她就是不说。

ChatGLM3-6B：给出回答“明明喜欢白白”是错误的。未能准确理解句子结构和逻辑。

Qwen-7B-Chat：理解了句子的意思，但是分析有误，未能正确解析句子结构，而且连问题本身都没有理解。

DeepSeek-R1-Distill-Qwen-7B：给出了正确答案“明明喜欢白白”，但是分析过程有漏洞，比如认为“明明明明”是笔误，认为没有说出口的是白白。

1. 领导和小明的对话中“意思”分别是什么意思

ChatGLM3-6B：能正确给出大部分意思的含义，但是顺序混乱，且统计个数不正确，有漏掉的情况。

Qwen-7B-Chat：解释了不同语境下“意思”的含义。解释详细且准确，涵盖了各语境下的不同意思，但依旧缺少了一些“意思”的解释。

DeepSeek-R1-Distill-Qwen-7B：解释了不同语境下“意思”的含义，但是有些不太准确，比如“意思意思”解释为加强否定，其次也没有给出所有解释，统计个数不正确。

1. **对比总结与评价**

在对 ChatGLM3-6B、Qwen-7B-Chat 和 DeepSeek-R1-Distill-Qwen-7B 三个模型的具体问答能力测试中，三者在语义理解、语言游戏处理、逻辑推理等方面展现出不同特点。

整体来看，Qwen-7B-Chat 在多数问题中的表现相对稳定，尤其擅长处理常规中文语义分析、对话语境理解、多义词解释等任务，输出风格自然流畅，口语化程度高，具备较强的实用性。在“能穿多少穿多少”这类具有反义结构的表达中，Qwen 能准确辨析语义反转，并用简洁有力的方式表达清楚。它在处理“意思”这种多义用法时也表现较好，覆盖面广，解释准确，逻辑清晰，是整体表现最均衡的一个。

DeepSeek-R1-Distill-Qwen-7B 在基本逻辑表达、语言组织和格式整洁度方面表现优秀，输出稳重、条理清晰，擅长处理结构化问答和有明确语境的语义分析。在“明明明明明白白白喜欢他”这种句式复杂的表达中，它成功得出了正确答案，尽管分析过程中有漏洞。此外，它在一些解释不清的问题上保持了较高的谨慎性，有时会建议提供更多上下文来完善理解。不过在面对嵌套逻辑复杂、涉及多主体推理的问题时，它的表现较差，例如在“他知道我知道你知道他不知道吗”中人物识别和逻辑顺序混乱，导致结果错误。

ChatGLM3-6B 在语义理解上处于中等水平，输出风格略显重复，解释往往缺乏深入挖掘，有时会出现“绕弯”表达而非直击重点。在面对语言游戏或幽默类问题时，常采取机械翻译式理解，缺乏灵活性和语感。例如在“单身狗”句中，它未能准确解析两个“谁都看不上”之间的语义差异，也未能提出替代解释；而在“他知道我知道你知道他不知道吗”中虽然得出结果，但过程表达重复、逻辑链条不清晰。在句子分析类任务中，如“明明明明明白白白喜欢他”，它未能准确划分语法结构，结论错误。

综合而言，如果任务涉及语言灵活性较高、需处理复杂语境、表达风格自然的中文问答，Qwen-7B-Chat 是首选。DeepSeek-R1-Distill-Qwen-7B 更适合逻辑清晰、结构明确的问题，尤其适合追求稳健输出的场景。ChatGLM3-6B 则作为基础模型在多数任务中能给出可用回答，但在处理语言双关、语法复杂或逻辑嵌套问题时仍有明显不足。

1. **优化建议**

Qwen-7B-Chat：可以通过优化信息量管理，避免信息过载，同时进一步提升在非常复杂语境下的解析能力，尤其是在人物关系混乱或语言嵌套较深的句式中增强推理能力。

ChatGLM3-6B：可以加强对复杂逻辑和嵌套语义的处理能力，提升其在深度分析问题上的表现，同时优化语言表达的简洁性与条理性，减少重复和绕弯的输出。

DeepSeek-R1-Distill-Qwen-7B：需要提升对复杂逻辑和语义结构的理解能力，特别是在多人物关系、嵌套逻辑以及语言幽默处理方面，同时保持其在输出条理性、语言稳定性和回答结构方面的优势。

通过进一步优化和增强这些方面，这三种模型都可以在各自的适用场景中提供更为优质的服务。