**大语言模型部署项目报告**

2351877 李星烁

**一、引言**

**1.1 作业背景与目的**

本次任务核心是在魔搭 (ModelScope) 平台上，利用其提供的免费 CPU 云计算资源，成功部署并对两款当前主流的开源大语言模型——智谱 AI 开发的 ChatGLM3-6B 和阿里巴巴通义实验室研发的 Qwen-7B-Chat 模型——进行功能性测试与性能对比分析。通过对不同模型在特定任务上的表现进行横向比较，旨在培养对模型能力边界的认知和批判性评估的能力，为未来在相关领域的研究与应用打下坚实基础。

**1.2 选用模型简介**

为完成本次作业的对比分析任务，我们选用了以下两款具有代表性的大语言模型：

**1.智谱 ChatGLM3-6B**

该模型由智谱AI公司（Zhipu AI）与清华大学KEG实验室联合发布，是 ChatGLM 系列的最新开源版本。作为一个拥有60亿参数（6B）的双语（中英）对话模型，ChatGLM3-6B 在前代模型的基础上，进一步提升了在代码生成、工具调用、语义理解以及逻辑推理等方面的能力，并针对中文语境进行了深度优化，使其在中文自然语言交互任务中表现出色。

**2.通义千问** **Qwen-7B-Chat**

该模型由阿里巴巴达摩院通义实验室自主研发并开源，是通义千问 (Qwen) 系列中的一款拥有70亿参数（7B）的对话微调模型。Qwen 系列模型基于大规模高质量数据进行预训练，具备强大的语言理解、内容生成和多轮对话能力。Qwen-7B-Chat 版本特别针对对话场景进行了优化，旨在提供流畅、自然且信息丰富的交互体验，并支持多种语言。

**二、实验步骤**

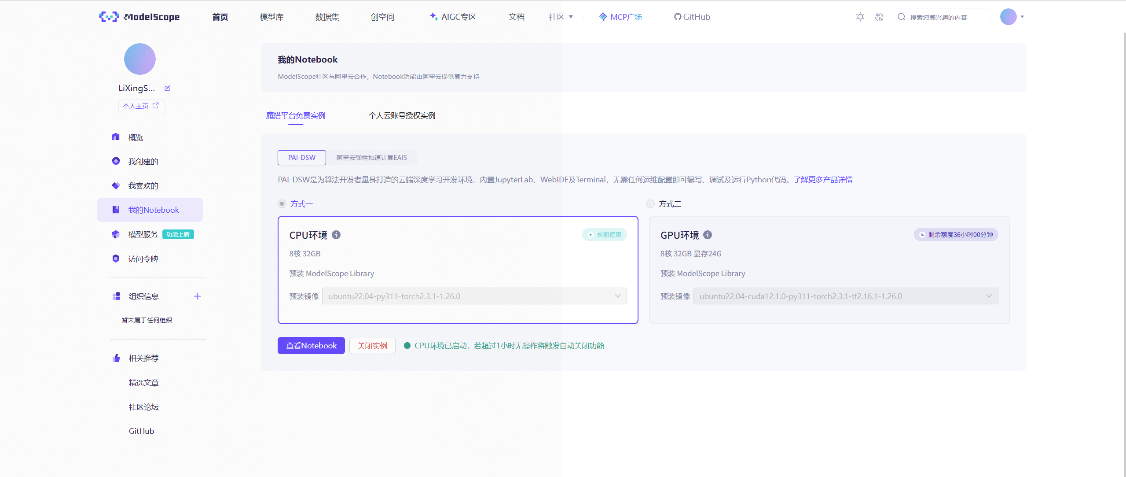
**2.1 部署平台与环境**

平台: 魔搭 ModelScope。

计算资源: 免费 CPU 云计算资源。

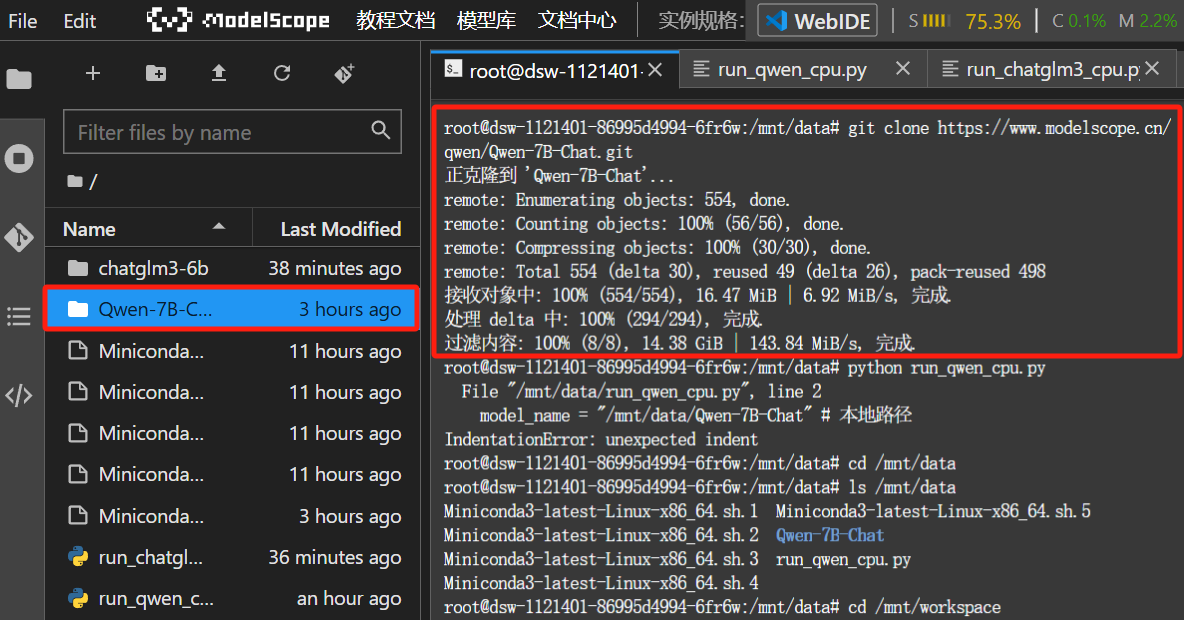
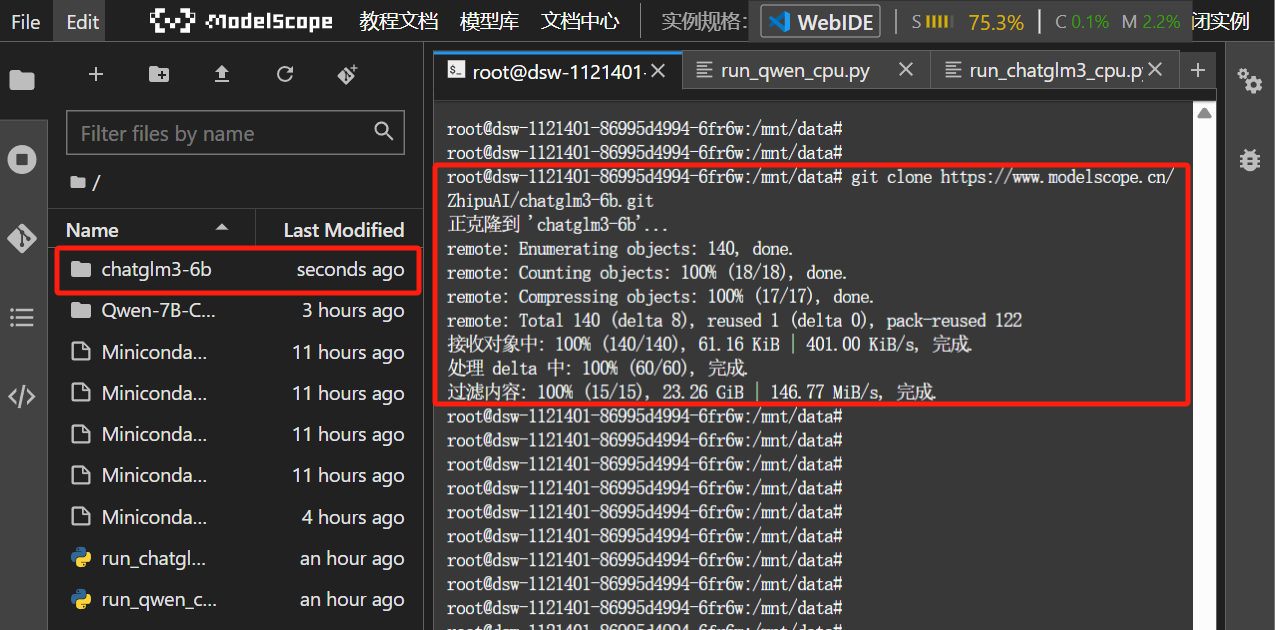
**2.2操作步骤及部署截图**

1.进入魔搭 ModelScope官网，进行登录，并启动Notebook。



2.我选择在root直接操作，故只需安装接触依赖即可。

3.进入Notebook，打开终端命令行，部署ChatGLM3-6B和Qwen-7B-Chat。

部署完成的截图：

4.创建好对应的python文件。（我创建了run\_qwen\_cpu.py和run\_chatglm3\_cpu.py两个py文件，对应两个模型）



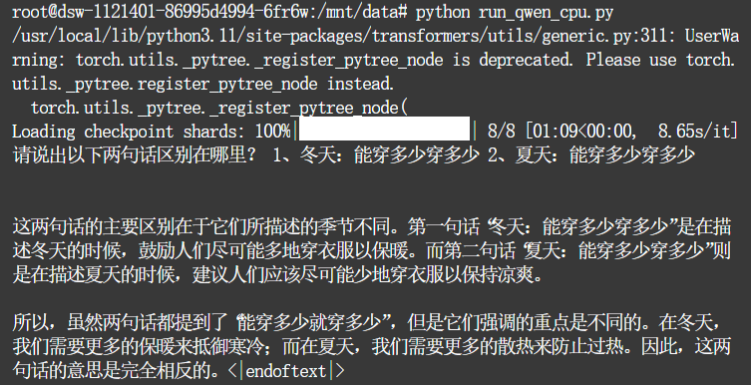
**三、测试用例及截图**

**3.1测试用例**

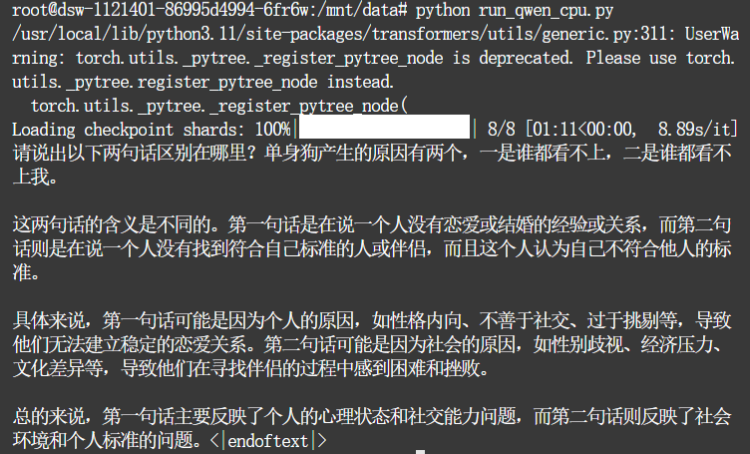
* 1.请说出以下两句话区别在哪里？ 1、冬天：能穿多少穿多少 2、夏天：能穿多少穿多少。
* 2.请说出以下两句话区别在哪里？单身狗产生的原因有两个，一是谁都看不上，二是谁都看不上。
* 3.他知道我知道你知道他不知道吗？ 这句话里，到底谁不知道？
* 4.明明明明明白白白喜欢他，可她就是不说。这句话里，明明和白白谁喜欢谁？
* 5.领导：你这是什么意思？ 小明：没什么意思。意思意思。 领导：你这就不够意思了。 小明：小意思，小意思。领导：你这人真有意思。 小明：其实也没有别的意思。 领导：那我就不好意思了。 小明：是我不好意思。请问：以上“意思”分别是什么意思。

**3.2 Qwen测试结果**

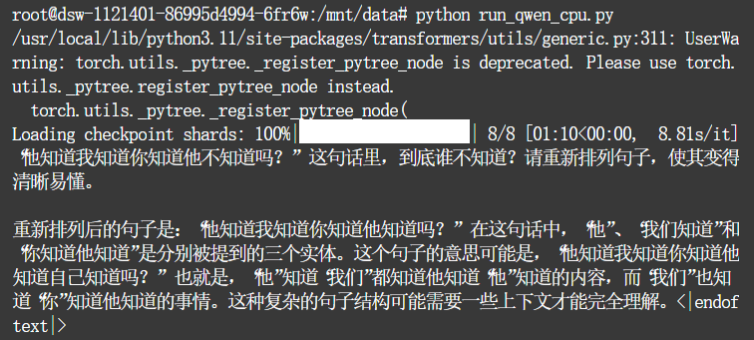
**Q1：**

****

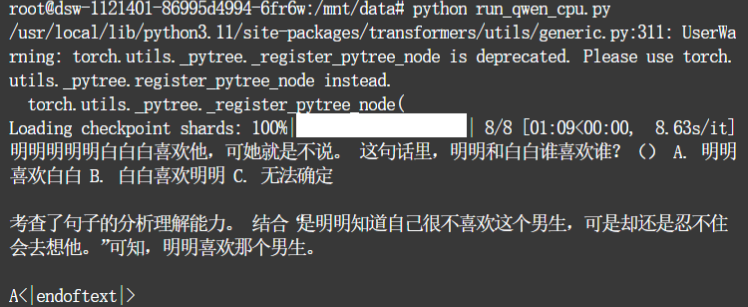
**Q2：**

****

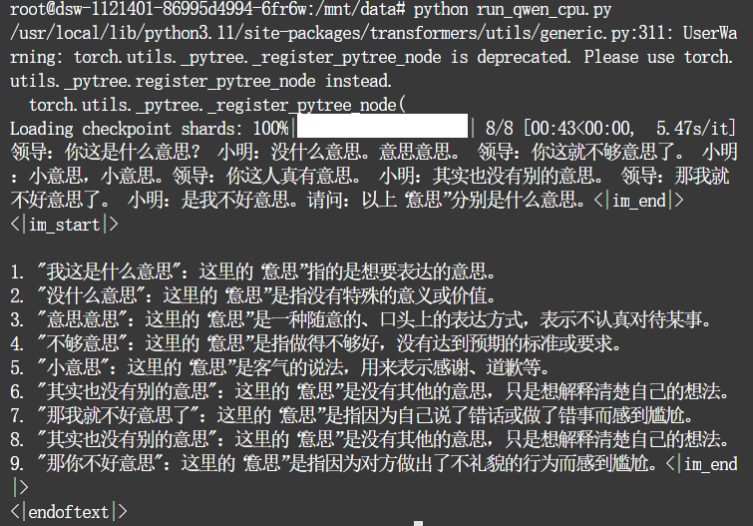
**Q3：**

****

**Q4：**

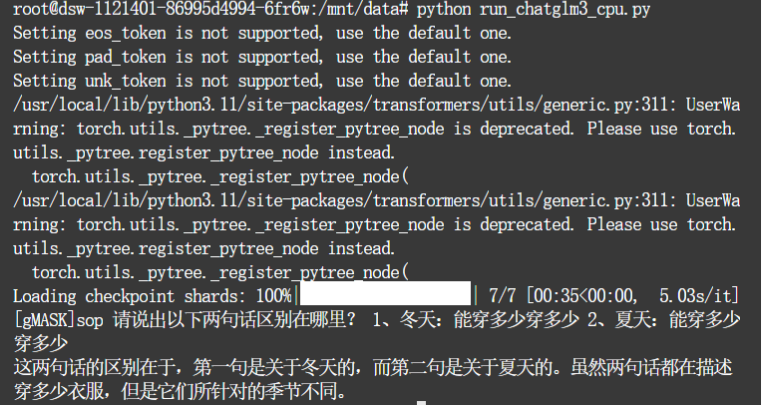
****

**Q5：**

****

**3.3 chatglm测试结果**

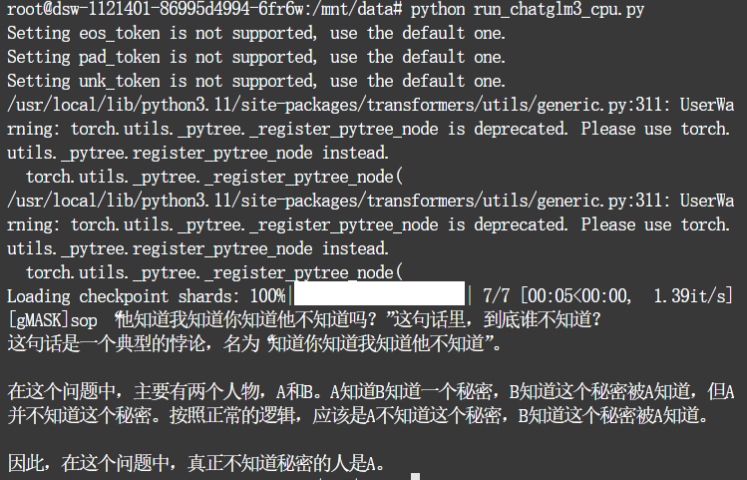
**Q1：**

****

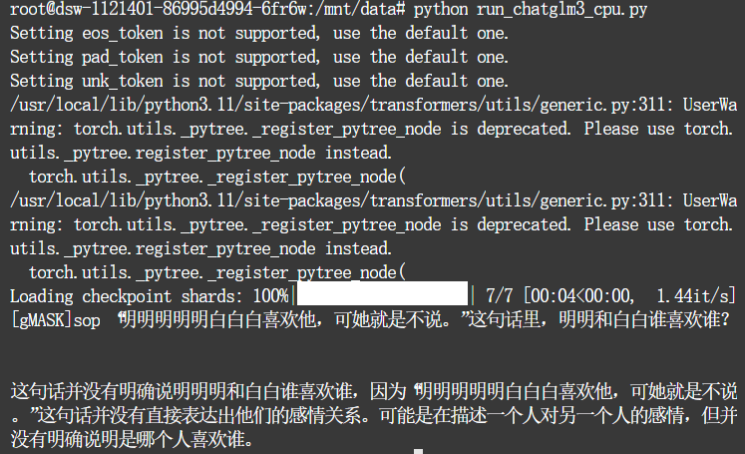
**Q2：**

****

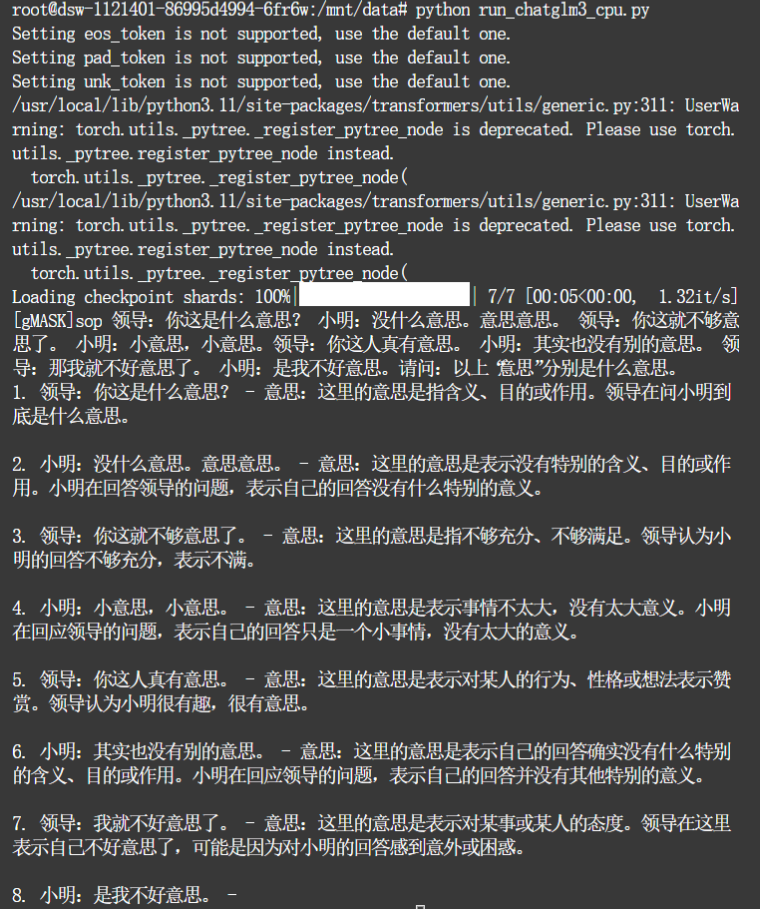
**Q3：**

****

**Q4：**

****

**Q5：**

****

**四、模型横向对比分析**

本章节旨在基于前述五个特定问题的测试结果，对通义千问Qwen (以下简称Qwen) 和 ChatGLM3-6B (以下简称ChatGLM3) 两款大语言模型在中文理解、逻辑推理、语境把握及回答质量等方面进行细致的横向对比分析。

值得注意的是，Qwen似乎对给定的测试问题进行了加工复述，导致回答可能会偏离所给的的问题，例如对于问题二，Qwen处理的问题变成了 "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。"多了一个“我”字，导致解答的偏离。综合来看，这个问题在Qwen模型的使用中还是普遍存在的，这也是需要注意并改进的点。

**4.1 对上下文相关的语义歧义的理解能力 (问题1: 冬天/夏天穿衣)**

* 测试目的: 考察模型是否能根据上下文（季节）理解同一短语“能穿多少穿多少”的相反含义。
* Qwen表现: Qwen 表现出色，准确识别出两句话的核心区别在于季节，并清晰阐述了“冬天：能穿多少穿多少”意指“尽可能多地穿衣服以保暖”，而“夏天：能穿多少穿多少”则建议“尽可能少地穿衣服以保持凉爽”。Qwen明确指出“这两句话的意思是完全相反的”，显示了其对上下文驱动的语义反转有很好的捕捉能力。
* ChatGLM3表现: ChatGLM3 同样识别出两句话针对的季节不同，指出“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。” 但其解释相较Qwen略显保守，未明确点出“能穿多少穿多少”在两个语境下的含义是截然相反的，更多停留在指出季节的差异性。
* 对比: 在此问题上，Qwen表现更胜一筹，它不仅识别了语境差异，更深入地解释了该差异如何导致了表述含义的根本性对立，展现了更强的语境语义理解和解释能力。

**4.2 对句法结构歧义的理解能力 (问题2: 单身狗原因)**

* 测试目的: 考察模型对由于句法结构相似而可能产生的不同语义指向的辨析能力。
* Qwen表现: Qwen 的回答基于其接收到的包含“我”的第二句话。它将第一句“谁都看不上”解读为个人原因（如性格、挑剔），而将第二句“谁都看不上我”解读为包含社会因素和个人标准的双向问题。其分析“第一句话主要反映了个人的心理状态和社交能力问题,而第二句话则反映了社会环境和个人标准的问题”具有一定的合理性。
* ChatGLM3表现: ChatGLM3 针对“一是谁都看不上,二是谁都看不上”的经典表述，在其较优的解释中指出“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”。这个解释准确地抓住了原句中通过主被动关系表达的两种不同归因。
* 对比:对中文常见表述“一是谁都看不上，二是谁都看不上”的经典歧义的理解，ChatGLM3 的解释更贴近该表述的核心歧义点（主动与被动）。Qwen则针对修改版提示给出了合理的分析。Qwen处理提问时出现了问题，导致其效果不如ChatGLM3。

**4.3 对复杂嵌套逻辑与指代关系的理解能力 (问题3: 他知道我知道...)**

* 测试目的: 评估模型处理复杂逻辑链条和多重指代关系，并从中提取关键信息的能力。
* Qwen表现: Qwen 尝试通过重新排列句子来理解，其重排后的句子为“他知道我知道你知道他知道吗?”，并进一步解释为“他知道 '我们'都知道他知道 '他'知道的内容,而 '我们'也知道 '你'知道他知道的事情。” 这个解释显得较为迂回，且未能直接、清晰地回答“到底谁不知道”。它最后表示“这种复杂的句子结构可能需要一些上下文才能完全理解”，暗示其在该问题上解析能力有限。
* ChatGLM3表现: ChatGLM3 表现优秀。它首先识别出这是一个经典的悖论式问句，并类比为“A知道B知道一个秘密,B知道这个秘密被A知道,但A并不知道这个秘密。” 基于此清晰的逻辑拆解，它准确地推断出“在这个问题中,真正不知道秘密的人是A。” 这个“A”对应原句中最后那个“他不知道”的“他”。
* 对比: 在处理这种高度嵌套的逻辑问句时，ChatGLM3展现了明显更强的逻辑分析和推理能力。它能够将复杂表述转化为更易处理的逻辑结构，并从中准确推断出答案。Qwen则在解析句子结构和直接回答问题方面显得力不从心。

**4.4 对文本细节、角色关系及隐含信息的捕捉能力 (问题4: 明明和白白)**

* 测试目的: 考察模型是否能准确解析句子成分，理解角色间的关系，并避免受干扰信息影响，正确回答基于文本内容的特定问题。原句为：“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？”
* Qwen表现: Qwen 的回答存在严重问题。它错误地选择了“A. 明明喜欢白白”，并且其解释中引用了一句“是明明知道自己很不喜欢这个男生,可是却还是忍不住会去想他。”这句引文完全不是来自题目给出的文本，属于内容幻觉。这表明Qwen在此题上不仅未能正确理解句意，还编造了无关信息。
* ChatGLM3表现: ChatGLM3 的回答则非常谨慎和准确：“这句话并没有明确说明明明和白白谁喜欢谁...并没有直接表达出他们的感情关系。” 它正确地理解到原句描述的是“白白喜欢他（第三方）”，而“明明”是知晓此事的人，但句子本身并未提供关于“明明和白白”之间互相喜欢的信息。
* 对比: 在此问题上，ChatGLM3 表现远超 Qwen。ChatGLM3 准确地分析了句子结构和角色关系，并基于文本给出的信息做出了合理的判断。Qwen 则出现了严重的理解错误和内容幻觉，这在评估模型可靠性时是一个重要的扣分项。

**4.5 对多轮对话中一词多义的理解与语境适应能力 (问题5: “意思”的对话)**

* 测试目的: 评估模型在连续对话中，根据不同语境准确理解同一词语（“意思”）的多种含义的能力。
* Qwen表现:
  + 准确识别了部分“意思”的含义，如领导的“你这是什么意思？”（想要表达的意思）、小明的“没什么意思。”（没有特殊意义）。
  + 对于“意思意思。”，Qwen解释为“随意的、口头上的表达方式,表示不认真对待某事”，部分捕捉到其非正式性，但未能点出其“象征性、略表心意”的核心含义。
  + 对于“不够意思。”，解释为“做得不够好,没有达到预期的标准或要求”，这未能抓住其“不够朋友、不仗义”的核心社交含义。
  + 对于“小意思。”，解释为“客气的说法,用来表示感谢、道歉等”，这与通常表示“小礼物不成敬意”或“小事一桩”的语境不太相符。
  + 对于领导的“那我就不好意思了。”，解释为“因为自己说了错话或做了错事而感到尴尬”，这不太符合领导在该语境下更可能表达的“那我就不客气了/我就却之不恭了”的含义。
  + Qwen的回答列表还出现了一处对“其实也没有别的意思”的重复解释和对“那你不好意思”的额外解释（似乎是对“那我就不好意思了”的误读），略显混乱。
* ChatGLM3表现:
  + 同样准确识别了如“含义、目的或作用”、“没有特别的含义”等。
  + 未单独解释“意思意思。”，将其含义包含在“没什么意思”的解释中，略有不足。
  + 对于“不够意思了。”，解释为“不够充分、不够满足”，与Qwen类似，也未能把握“不仗义”的社交含义。
  + 对于“小意思,小意思。”，解释为“事情不太大,没有太大意义”，这比Qwen的解释更贴近该词的常见用法。
  + 对于领导的“你这人真有意思。”，解释为“赞赏...很有趣”，准确。
  + 对于领导的“那我就不好意思了。”，解释为“可能是因为对小明的回答感到意外或困惑”，这种解读有一定可能性，但同样未能完全捕捉到领导“那我就不客气了”的潜在含义。
  + ChatGLM3的整体解释结构更为清晰，逐条对应对话。
* 对比: 在这个极具挑战性的一词多义理解问题上，两款模型均未能完美解读所有“意思”的地道含义，特别是在更深层次的社交潜台词和习语化表达上（如“不够意思”、“那我就不好意思了”）都存在理解偏差。 ChatGLM3在“小意思”的解读上略优于Qwen，并且其回答的组织结构更清晰。Qwen在“意思意思”上尝试给了解释但不够精准，且在“不够意思”和领导的“不好意思”上理解偏差较大，其输出的结构也稍显混乱。这表明对于高度依赖文化背景和语用知识的细微语义差别，当前的CPU部署模型仍有较大提升空间。

**4.6 总结与讨论**

综合以上五个问题的测试结果，我们可以对Qwen和ChatGLM3在此次CPU环境下的表现进行初步的总结：

* 对于问题的初步处理: Qwen对给定的测试问题进行了加工复述，导致回答可能会偏离所给的的问题（例如问题2），而ChatGLM3则不会出现此情况。
* 逻辑推理与复杂结构解析: ChatGLM3 在处理需要严密逻辑推理和解析复杂句子结构（如问题3、问题4）方面展现出更强的能力，其回答更为精准且不易受干扰。
* 上下文语义理解: 对于直接依赖上下文产生语义变化的简单场景（问题1），Qwen 表现出优秀的理解和解释能力。但在更依赖句法结构或习语知识的场景（问题2的部分情况、问题5），两者均有提升空间。
* 回答的准确性与可靠性: ChatGLM3 在本次测试中展现了相对更高的回答准确性和可靠性，尤其是在避免内容幻觉方面（参考问题4的表现）。Qwen在问题4中出现的幻觉是一个需要关注的问题。
* 对中文特有表达和习语的掌握: 两款模型在理解一些富有中文特色的口语表达和一词多义现象（如问题5中的多个“意思”）时，均未能达到完全精准的水平，尤其是在社交含义和潜台词的把握上，显示出当前模型在这方面的局限性。
* 表达清晰度与结构性: ChatGLM3 的回答在整体上结构更为清晰，语言组织也相对更符合分析性文本的风格。

总体而言，在本次特定测试集和CPU部署环境下，ChatGLM3 在逻辑推理、复杂问题拆解以及回答的稳定性和准确性方面表现相对突出。Qwen 则在某些上下文语义理解的直观性和解释的生动性上有所长，但在处理复杂逻辑和避免内容编造方面有待加强。 需要强调的是，这仅是基于少量特定问题的初步评估，更全面的模型能力对比需要更大规模、更多样化的测试集以及在不同硬件环境下的评估。

**五、总结与展望**

**5.1 主要工作总结**

本次作业围绕大语言模型的实际部署、测试与对比分析展开。通过在魔搭 (ModelScope) 平台的免费 CPU 云计算资源上进行操作，我们成功部署了智谱 AI 的 ChatGLM3-6B 和阿里巴巴通义实验室的 Qwen-7B-Chat 两款开源大语言模型。在此基础上，我们设计并执行了一系列针对性的问答测试，涵盖了中文复杂语义理解、逻辑推理、上下文把握以及一词多义等多个维度。最终，通过对模型返回结果的细致比对，完成了对这两款模型在特定测试场景下性能的横向分析。

**5.2 主要收获与体会**

通过本次实践，我们获得了多方面的宝贵经验和深刻体会：

**1.大语言模型部署的实践认知**

亲身体验了从环境配置、模型下载/加载到运行推理的完整流程，对模型部署过程中可能遇到的依赖问题、资源限制（尤其是在CPU环境下）以及配置细节有了更直观的理解。

**2.不同模型间的性能差异性**

通过横向对比，清晰地认识到即使是参数规模相近的成熟模型，在不同类型的任务和具体问题的处理上也会展现出各自的优势与不足。例如，ChatGLM3-6B 在逻辑推理和复杂结构解析上表现突出，而Qwen在某些上下文语义理解的直观性上较好。

**3.当前大模型的局限性认知**

测试结果也揭示了当前大语言模型（尤其是在资源受限的CPU环境下部署时）的一些局限性，如在处理高度依赖文化背景的习语、深层社交潜台词以及在某些情况下可能出现的“幻觉”或不准确回答等。这有助于我们更理性地看待AI的能力边界。

**5.3 未来展望**

大语言模型技术仍在飞速发展，其潜力远未被完全发掘。展望未来：

**1.模型性能的持续提升：**

随着训练数据的增加、算法的优化以及更大规模模型的出现，未来大语言模型在理解能力、推理能力、生成质量以及多模态处理等方面有望实现更大突破，进一步减少“幻觉”现象，提高回答的准确性和可靠性。

**2.端侧与轻量化部署的普及**：

虽然本次实验基于云端CPU，但模型轻量化和端侧部署（如在个人电脑、移动设备上运行）是重要的发展趋势。这将使得大模型技术更加普惠，催生更多创新应用。

**3.多模态能力的融合：**

文本、图像、语音、视频等多模态信息的融合处理是大势所趋。未来大模型将不仅仅局限于文本交互，而是能够理解和生成更丰富的媒体内容。

**六、项目链接**

项目链接：<https://github.com/L-Dramatic/LLM-Deployment-Comparison>

