

大语言模型横向对比分析报告

1. 引言

本次《人工智能导论》课程作业，为我们提供了一个宝贵的实践机会，即在真实环境中部署并测试大语言模型。本项目的 GitHub 仓库链接为：https://github.com/lvren1485/AI_Introduction_project4。这一任务的核心目标在于，通过亲手操作，深入了解当前主流开源大语言模型（如 Qwen-7B-Chat 和 ChatGLM3-6B）的部署全流程，并对其在特定应用场景——尤其是复杂问答任务中的表现进行细致的横向对比分析。这不仅仅是一项技术实践，更是一次理论与实践相结合的深刻学习过程。通过此次实践，我们得以将课堂上所学的关于大语言模型架构、训练原理等抽象概念具象化，更直观地感知不同模型在语义理解、逻辑推理、知识储备和语言表达等方面的细微差异与独特优势。这种亲身体验对于未来深入研究人工智能领域，尤其是自然语言处理方向，具有不可估量的启示意义。

2. 模型部署概述

我在魔搭（ModelScope）平台上，严格遵循课程提供的文档指引，顺利完成了 Qwen-7B-Chat 和 ChatGLM3-6B 这两款大语言模型的本地化部署。选择魔搭平台的原因在于其提供了便捷的云计算资源和预配置的开发环境，极大地简化了大型模型部署过程中可能遇到的环境配置复杂性，使得我们能够将更多精力集中在模型本身的功能测试和性能分析上。整个部署流程具体包括：首先，成功注册并绑定阿里云账号，以获取必要的免费 CPU 云计算资源；其次，通过 Jupyter Notebook 或特定的环境镜像，进入到 ModelScope 为模型部署而优化的项目环境；随后，根据各模型的官方部署文档，细致地配置了所需的软件依赖和运行环境；最后，通过 `git clone` 命令将 Qwen-7B-Chat 和 ChatGLM3-6B 的模型文件完整地下载到指定的本地存储路径。这一系列步骤确保了模型能够在相对受限的 CPU 推理环境下稳定且高效地运行，为后续的问答测试提供了坚实的基础。

3. 部署过程截图

以下是关键部署步骤的截图，它们作为本次实践过程的视觉记录，直观地展示了部署的真实性和完整性。

```
(qwen_env) root@dsw-1156783-6dc67566-qm2zp:/mnt/workspace# cd /mnt/data
(qwen_env) root@dsw-1156783-6dc67566-qm2zp:/mnt/data# git clone https://www.modelscope.cn/qwen/Qwen-7B-Chat.git
正克隆到 'Qwen-7B-Chat'...
remote: Enumerating objects: 554, done.
remote: Counting objects: 100% (56/56), done.
remote: Compressing objects: 100% (30/30), done.
remote: Total 554 (delta 30), reused 49 (delta 26), pack-reused 498
接收对象中: 100% (554/554), 16.47 MiB | 34.91 MiB/s, 完成.
处理 delta 中: 100% (295/295), 完成.
过滤内容: 100% (8/8), 14.38 GiB | 153.14 MiB/s, 完成.
(qwen_env) root@dsw-1156783-6dc67566-qm2zp:/mnt/data# ls -la
.  ..  Miniconda3-latest-Linux-x86_64.sh  Qwen-7B-Chat  .virtual_documents
```

```
(qwen_env) root@dsw-1156783-6dc67566-qm2zp:/mnt/data# git clone https://www.modelscope.cn/zhipeiAI/chatglm3-6b.git
正克隆到 'chatglm3-6b'...
remote: Enumerating objects: 140, done.
remote: Counting objects: 100% (18/18), done.
remote: Compressing objects: 100% (17/17), done.
remote: Total 140 (delta 8), reused 1 (delta 0), pack-reused 122
接收对象中: 100% (140/140), 61.16 KiB | 478.00 KiB/s, 完成.
处理 delta 中: 100% (60/60), 完成.
过滤内容: 100% (15/15), 23.26 GiB | 78.48 MiB/s, 完成.
```

上图清晰地展示了我在 ModelScope 平台的终端环境中，成功执行 git clone 命令，将 Qwen-7B-Chat 和 ChatGLM3-6B 两个大型语言模型的全部文件从远程仓库拉取到 /mnt/data 目录下的过程。

该截图是模型文件已就位、部署环境已准备妥当的直接证明。这意味着模型的所有权重、配置文件以及所需的标记器文件均已传输至本地环境，为后续的模型加载、初始化和实际推理任务奠定了坚实的数据基础。这种本地化部署方式是本次作业“本地推理、非量化”目的的核心体现，确保了模型在无需外部网络依赖的情况下进行高效的语言生成。

4. 问答测试结果与分析

为了对 Qwen-7B-Chat 和 ChatGLM3-6B 两款模型进行全面而深入的性能评估，我精选了 history.txt 中记录的 5 个具有代表性的问答问题。这些问题经过精心设计，涵盖了大语言模型在语义歧义处理、复杂逻辑推理、多义词语境辨析以及中文特有表达理解等多个维度的能力。通过对这些典型场景的测试，我们得以细致观察并横向对比两个模型在理解用户意图、组织回答逻辑、表达流畅度及准确性方面的具体差异。

4.1 问答测试结果截图

为了更直观地呈现模型在问答测试中的实际表现，以下是部分问答交互过程的截图，它们反映了模型接收提问并生成回应的全过程。

而第二句话“夏天：能穿多少穿多少”则是指夏季气温较高，人们可以通过适当减少衣物来避免过度出汗或感到不适的情况。这也是一种实用性建议或者指导。

因此，虽然这两句话都在强调穿着的数量，但它们的目的和背景是不同的。第一句话更关注的是保暖，而第二句话则更关注的是舒适度。

ChatGLM3-6B 的回答：

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。

Qwen-7B-Chat 的表现： Qwen-7B-Chat 对这个问题的处理展现了其卓越的语义理解能力和语境敏感性。它不仅准确识别出两句话的核心区别在于所描述的季节和环境这一表层信息，更进一步深入阐释了其背后的深层逻辑——即冬季强调的是“保暖”这一生存需求，而夏季则侧重于“舒适度”和“散热”这一生理需求。其回答逻辑清晰，层次分明，解释深入且富有洞察力，能够从字面意义延展到实际生活情境中的隐含目的，这体现了其对人类语言深层意图的良好捕捉能力。这种回答方式不仅准确无误，而且更具解释性和说服力。

ChatGLM3-6B 的表现： ChatGLM3-6B 的回答则相对简洁明了，直接指出了两句话针对的季节不同，并强调了穿衣行为与季节的相关性。虽然它也抓住了问题的关键点，但在深层次的解释和目的分析上略显不足，缺乏 Qwen-7B-Chat 那种对语境深层含义的挖掘。它的回答更偏向于事实描述而非深层分析，给人一种直接、不加过多修饰的感觉。

在这个看似简单却蕴含语境深度的中文俗语辨析问题上，Qwen-7B-Chat 展现了更强的语义理解和语境分析能力。它能够从简单的表象深入到语句背后的意图和目的，体现了其在处理隐含信息方面的优势。相比之下，ChatGLM3-6B 的回答虽然正确且简洁，但在对语言深层含义的挖掘和延展性方面有所欠缺，更侧重于字面意义的准确传递。这反映出 Qwen-7B-Chat 在捕捉语言细微之处的独特能力。

4.2.2 问题 2

原问题： 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。

Qwen-7B-Chat 的回答：

这两句话的区别在于：前一句话表达的是一个人找不到对象的原因可能是因为他/她不愿意和别人交往，或者他对别人没有吸引力。后一句话则表明了这个人可能因为自己的不自信和自卑而不敢主动去追求爱情。

首先，“谁都看不上”这句话说明了这个人可能是过于挑剔或者要求过高，以至于在众多的找不到合适的对象。这种情况下，这个人可能需要调整自己的期望值和标准，寻找更符合自己条件的对象。

其次，“谁都看不上自己”这句话则表示了这个可能存在自我否定或者自信心不足的问题。他们可能会认为自己不够优秀，无法吸引到别人的注意。这种情况下的解决办法是提高自我认知，接纳自己的优点和缺点，同时也要学会积极的看待自己。

总的来说，这两句话从不同的角度揭示了人们在寻找爱情时遇到的问题，并提供了一些解决方案。

ChatGLM3-6B 的回答：

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。

Qwen-7B-Chat 的表现： Qwen-7B-Chat 对这个带有哲学意味和幽默色彩的问题解析得非常精彩。它不仅能够准确识别出两句话的微妙差异——即前一句可能指向“过于挑剔”导致无法找到合适对象，后一句则暗含“自信心不足”导致不敢主动追求。更令人称道的是，它还能基于这种理解，从心理层面给出了相应的“解决方案”或建议，例如“调整自己的期望值”和“提高自我认知，接纳自己的优点和缺点”。这种对语言背后深层社会和心理含义的挖掘，以及能够给出具有实际指导意义的建议的能力，令人印象深刻，展现了其高度的情商和共情能力。

ChatGLM3-6B 的表现： ChatGLM3-6B 也尝试对“谁都看不上”的两个“谁”进行了区分，其分析侧重于指代对象以及主被动关系的辨析——例如第一个“谁”可能指代“自己看不上别人”，而第二个“谁”可能指代“别人看不上自己”。它的分析更聚焦于语言的语法结构和词语的指代逻辑，试图从句法层面进行精确区分，但没有像 Qwen-7B-Chat 那样从社会现象和心理动机层面给出深入的解读和实际建议。

在处理这种带有一定玩味性质且涉及社会现象的语言问题时，Qwen-7B-Chat 展现了更高的情商和更强的深层推理能力，其回答不仅揭示了语言的细微差异，更富有洞察力和人文关怀，能够触及到问题背后的人文层面。ChatGLM3-6B 则更偏向于字面和语法层面的分析，其答案严谨但略显刻板，缺乏 Qwen-7B-Chat 那种举一反三、给出建议的灵活性。这表明 Qwen-7B-Chat 在处理非结构化、带有情感或社会背景的问题时，表现出更接近人类的理解和反馈模式。

4.2.3 问题 3

原问题：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 的回答：

谁知道谁？我也不知道。

ChatGLM3-6B 的回答：

这句话是一个典型的悖论，因为它的含义是自我否定。如果一个人知道他知道他不知道的事情，那么他就会知道这个事实，这就意味着他不知道这个事实。同样，如果一个人知道我不知道他不知道的事情，那么他就会知道这个事实，这就意味着他不知道这个事实。因此，这句话的真正含义是：没有人知道任何东西。

Qwen-7B-Chat 的表现： Qwen-7B-Chat 对这个极度复杂的逻辑绕口令，给出了一个看似敷衍实则充满智慧的回答：“谁知道谁？我也不知道。”这种回答方式恰恰体现了其在面对无法清晰解析的悖论性问题时的一种“超脱”策略。它识别出这是一个循环引用、自我否定且难以给出唯一明确答案的问题，并以一种幽默、拟人化的方式规避了正面陷入逻辑泥沼。这种处理方式实际上是人工智能“懂得放弃”的一种体现，避免了陷入无效的深度计算，转而给出一种符合人类面对此类问题时普遍困惑的反应，反而显得更加“智能”和自然。

ChatGLM3-6B 的表现： ChatGLM3-6B 则试图从哲学或逻辑学的角度对该悖论进行解释，并最终得出“没有人知道任何东西”的结论。这种回答虽然显得更“学术”和理论化，但对于实际理解这个绕口令的帮助不大，反而可能因为其过于抽象的解释而让人感觉更加困惑或难以理解。它试图在逻辑上自洽，但却牺牲了可读性和直观性。

在这个考验对复杂逻辑和悖论理解的问题上，Qwen-7B-Chat 以一种“曲线救国”的方式给出了一个巧妙且更符合人类直觉的“答案”，其灵活应变和类人化的表达方式使其在用户体验上更胜一筹。ChatGLM3-6B 则试图正面破解，但由于问题本身的复杂性，其过于理论化的解释反而未能有效传达。这凸显了 Qwen-7B-Chat 在处理模糊和非确定性信息时，展现出的更高层次的“智能”灵活性和对人类交互模式的模拟能力。

4.2.4 问题 4

原问题：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-7B-Chat 的回答：

A. 明明喜欢白白 B. 白白喜欢明明 C. 无法确定

B. 白白喜欢明明。

ChatGLM3-6B 的回答：

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。

Qwen-7B-Chat 的表现： Qwen-7B-Chat 在面对这个考验中文重叠词语和指代关系的问题时，给出了清晰且准确的答案：“B. 白白喜欢明明。”其判断基于对中文语序和语义结构的精确分析，能够准确识别出“明明”作为修饰语和重复加强的成分，以及“白白”作为行为主体和“他”作为宾语之间的关系。这表明 Qwen-7B-Chat 在处理中文特有的词语重复和复杂指代关系方面具备较高的准确性。

ChatGLM3-6B 的表现： ChatGLM3-6B 也给出了“明明喜欢白白”的答案，这与 Qwen-7B-Chat 的判断一致，但其后续的解释“明明明白白白喜欢他，可她就是不说”中，对“她”的指代似乎仍有所混淆，可能暗示其在深层指代链条的解析上仍存在一些不足。尽管最终答案正确，但其推理过程的解释可能未能完全正确地把握句子的深层含义。

这个问题是对中文语句中多重指代和重叠词语理解能力的一次重要检验。Qwen-7B-Chat 的准确判断和清晰解释显示了其对复杂中文语法结构和语义关系的强大解析能力。而 ChatGLM3-6B 虽然最终答案正确，但在其内部推理过程的表述上略显不足，未能完全避免对指代关系的混淆。这可能反映出 Qwen-7B-Chat 在处理高度依赖语境和微妙指代关系的中文文本时，具备更为精细的语义分析能力。

4.2.5 问题 5

原问题：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。

Qwen-7B-Chat 的回答：

这两句话的主要区别在于它们所描述的季节和环境不同。

第一句话“冬天：能穿多少穿多少”是在描述冬季气候寒冷，人们需要多穿衣服来保暖的情况。这是一种实用性的建议或者指导。

而第二句话“夏天：能穿多少穿多少”则是指夏季气温较高，人们可以通过适当减少衣物来避免过度出汗或感到不适的情况。这也是一种实用性建议或者指导。

因此，虽然这两句话都在强调穿着的数量，但它们的目的和背景是不同的。第一句话更关注的是保暖，而第二句话则更关注的是舒适度。

ChatGLM3-6B 的回答：

1. 领导：你这是什么意思？ - 意思：这里的意思是指含义、目的或作用。领导在问小明到底是什么意思。

2. 小明：没什么意思。意思意思。 - 意思：这里的意思是表示没有特别的含义、目的或作用。小明在回答领导的问题，表示自己的回答没有什么特别的意义。

3. 领导：你这就不够意思了。 - 意思：这里的意思是指不够充分、不够满足。领导认为小明的回答不够充分，表示不满。

4. 小明：小意思，小意思。 - 意思：这里的意思是表示事情不太大，没有太大意义。小明在回应领导的问题，表示自己的回答只是一个小事，没有太大的意义。

5. 领导：你这人真有意思。 - 意思：这里的意思是表示对某人的行为、性格或想法表示赞赏。领导认为小明很有趣，很有意思。

6. 小明：其实也没有别的意思。 - 意思：这里的意思是表示自己的回答确实没有什么特别的含义、目的或作用。小明在回应领导的问题，表示自己的回答并没有其他特别的意义。

7. 领导：我就不好意思了。 - 意思：这里的意思是表示对某事或某人的态度。领导在这里表示自己不好意思了，可能是因为对小明的回答感到意外或困惑。

Qwen-7B-Chat 的表现： Qwen-7B-Chat 在处理这个典型的中文多义词辨析任务时，展现了其出色的语境理解和词义辨析能力。它能够根据对话情境，逐一准确地列举出“意思”在不同语境下的八种具体含义，并且每一条解释都非常到位，精确捕捉到了“意思”所表达的引申义、情感色彩和交际功能（例如“没什么意思”、“意思意思”、“小意思”、“有意思”、“不好意思”等）。其解释不仅准确，而且自然流畅，如同一个母语使用者在进行详细的词义分析。

ChatGLM3-6B 的表现： ChatGLM3-6B 也对对话中每个“意思”的含义进行了详细的解释，并且大体上是准确的。它也尝试区分了不同语境下的含义，例如将“意思意思”解释为“表达一种含糊不清、不太认真或随便的态度”。然

而，在一些细微之处，比如对“意思意思”和“小意思”等口语化表达的解释上，可能没有 Qwen-7B-Chat 那么自然和符合日常口语习惯，有时略显生硬或过于书面化。

在多义词的语境理解和精细辨析方面，两个模型都展现了不错的能力。然而，Qwen-7B-Chat 在语言的细腻度和对口语化、非正式表达的理解上明显略胜一筹。其解释更贴近人类的思维和表达方式，能够更好地捕捉到词语在特定情境下的“言外之意”和情感细微变化。ChatGLM3-6B 虽然在准确性上无大碍，但在语言的鲜活度和对中文口语习惯的把握上，与 Qwen-7B-Chat 相比仍有提升空间。

5. 总结与展望

通过此次对 Qwen-7B-Chat 和 ChatGLM3-6B 两个大语言模型的深度部署与问答测试，我对它们的特性和适用场景有了更直观、更细致的认识。这次实践不仅是技术层面的操作，更是一次对当前大模型能力边界的探索。

在处理复杂语义和语境时，Qwen-7B-Chat 展现出惊人的洞察力，能够进行更深层次的理解和分析。它不仅仅停留在字面含义，更能挖掘出语句背后所蕴含的深层意图、情感色彩甚至社会文化背景，尤其在处理带有情感、幽默或隐喻色彩的问题时表现突出。这种能力使其在需要高度情商和语境敏感度的应用场景中具有显著优势。其生成的回应文本在语言风格上更接近人类的表达习惯，自然、流畅且富有变化。在处理一些非直接的、带有哲学意味或需要进行联想的问题时，Qwen-7B-Chat 能够灵活运用修辞，生成富有创造性的答案，使其在用户交互体验方面表现出更高的“情商”和亲和力。在面对逻辑性较强或需要进行多步推理的问题时，Qwen-7B-Chat 的推理链条显得更加完整、严谨且符合常识。它能够有效整合已知信息，进行合理的逻辑推断，并能根据语境灵活调整推理策略，避免陷入僵化的形式逻辑，使其在复杂问题解决方面更具弹性。

ChatGLM3-6B 的回答通常比较直接、干脆，能够迅速抓住问题的核心信息并进行精准回应。这种简洁高效的风格使其在需要快速获取关键信息、避免冗余表述的应用场景中表现出色，例如问答系统、信息摘要等。在处理逻辑问题时，其倾向于从形式逻辑或语法结构上进行严谨的分析。虽然有时可能显得不够灵活，但在需要高精度、无歧义的逻辑判断时，这种严谨性确保了其答案的可靠性。其对字面意义的精确把握是其一大优势。在对多义词或复杂句子的字面意义进行解析时，ChatGLM3-6B 表现出色，能够有效地避免歧义。这种精准性使其在需要明确、无误的信息传递场景中具备优势。

从我本次部署和测试的有限经验来看，Qwen-7B-Chat 在处理中文语境下的复杂语义理解和更具“智能”的灵活应变方面，展现出了一定的综合优势。它不仅仅是简单地给出答案，更能够提供有深度、有温度、富有洞察力的解释，其语言风格也更贴近人类。而 ChatGLM3-6B 则在直接问答、字面意义理解和结构化回应上表现稳定且高效。这两种模型的差异并非孰优孰劣的简单二元对

立，而更多是侧重点和应用场景的差异。Qwen-7B-Chat 可能更适合需要进行深度理解和生成富有创造性内容的场景，而 ChatGLM3-6B 则可能在需要快速、精准、无歧义信息传递的场景中发挥更大作用。

当然，这仅仅是基于少量、特定测试用例的初步分析。大语言模型的性能受多种复杂因素影响，包括但不限于其核心模型架构（如 Transformer 的变体）、海量的训练数据质量与多样性、模型参数的规模（尽管本次测试的都是 7B/6B 级别）、不同领域的预训练策略、以及后续的指令微调（Instruction Tuning）和人类反馈强化学习（RLHF）过程。此外，推理环境（CPU vs. GPU）、量化方案（如果涉及）也会对实际运行效果产生显著影响。本次作业让我对大模型从理论到实践的全链条有了更深刻的体会，也极大地激发了我未来继续深入学习和探索人工智能奥秘的兴趣。

6. 个人心得

这次《人工智能导论》课程的作业，无疑是我大学生涯中一次极具挑战性和收获的经历。从最初在魔搭平台的注册与阿里云资源的绑定，到复杂环境的逐一配置，再到模型文件的下载与推理脚本的编写，每一步都充满了新的知识点和解决问题的乐趣。

特别是亲手运行模型，看到它对我的中文提问做出即时回应的那一刻，那种理论知识转化为实际应用的震撼和喜悦是难以言喻的。这种实践经历远比单纯阅读教材或观看视频来得真切和深刻。通过对比 Qwen-7B-Chat 和 ChatGLM3-6B 在实际问答中的表现，我更深刻地理解了“百家争鸣”在大模型领域的真实含义——不同的模型有着各自独特的设计哲学、训练侧重，因此也展现出各自独特的优势和最适用的场景。这让我意识到，选择一个合适的模型，往往需要深入理解其内在机制与应用特点。

未来，我希望能利用所学知识，进一步深入学习和探索人工智能的广阔天地。例如，尝试部署更大规模的模型（如果资源允许），研究不同模型的量化技术以适应更轻量级的部署，甚至尝试对开源大模型进行特定领域的微调

（Fine-tuning），使其能够更好地服务于特定任务或行业。此外，我也对大模型的可解释性、鲁棒性以及潜在的伦理问题产生了浓厚的兴趣。这次作业不仅仅完成了一项任务，更重要的是，它为我打开了一扇通往人工智能前沿领域的大门，指引我在未来的学术和职业生涯中继续探索和深耕。