

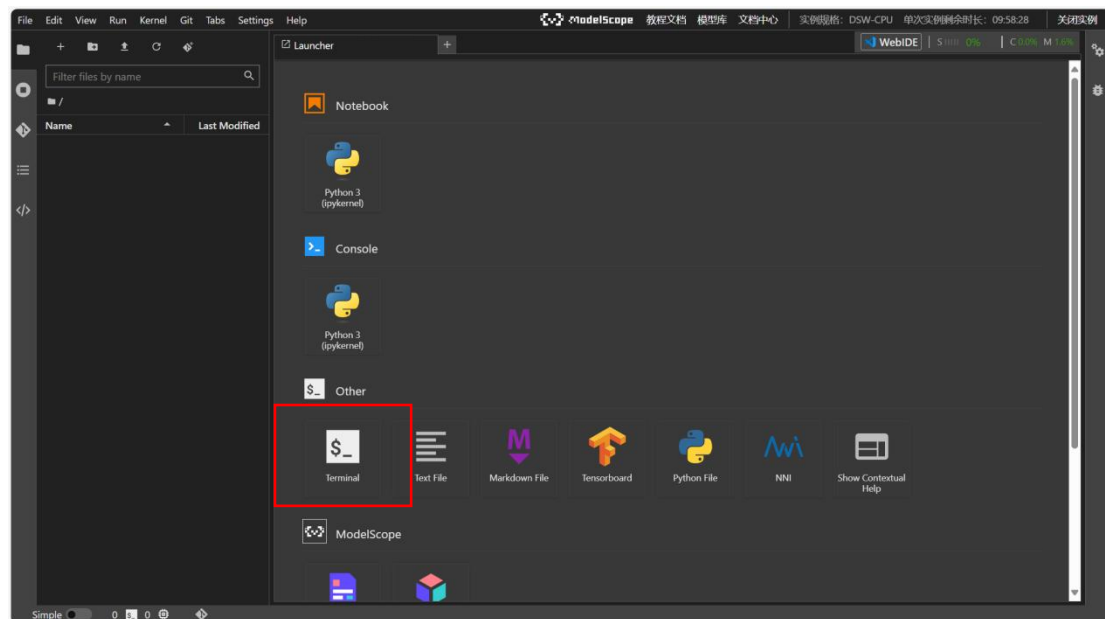
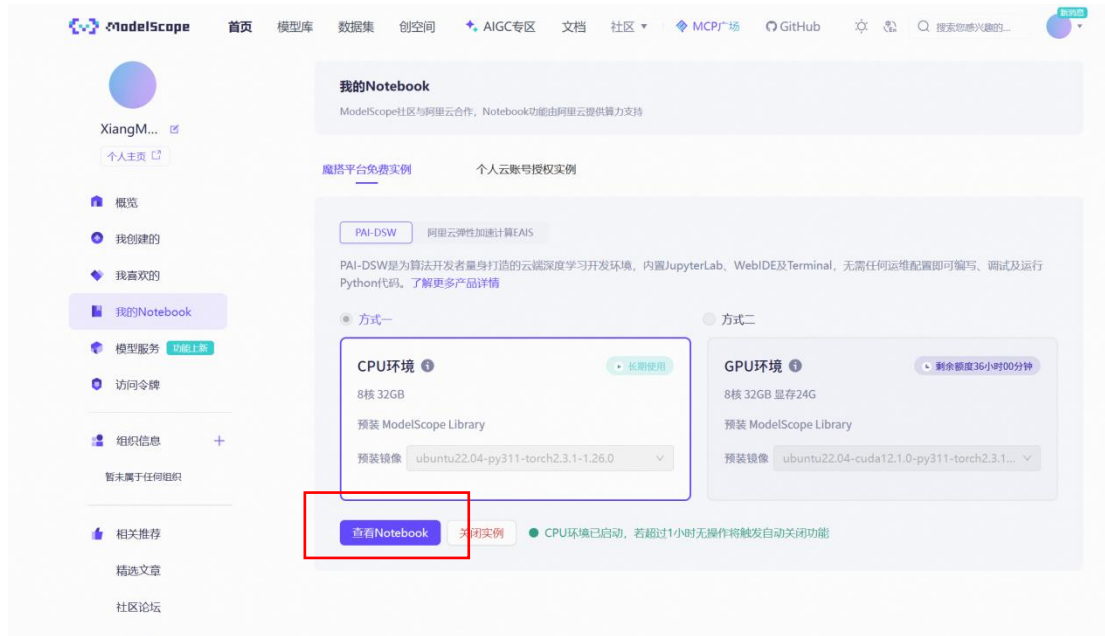
人工智能导论课程作业

—— 大语言模型部署

姓名：	吴瑞翔
学号：	2351716
学院：	计算机科学与技术学院
专业：	软件工程
指导教师：	曾进

1. 配置和部署

登录魔搭平台，关联阿里云账号后，打开 Notebook, 进入终端命令环境。



环境搭建，并根据需要将实验所需要的大模型 git clone 至本地。

(1) 通义千问 Qwen-7B-Chat

```
root@dsw-1133525-7c966c9f56-7hx49:/mnt/workspace# 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 | 547.00 KiB/s, 完成
处理 delta 中: 100% (294/294), 完成
过滤内容: 100% (8/8), 14.38 GiB | 128.51 MiB/s, 完成
```

Name	Last Modified
Qwen-7B-Chat	7 hours ago
• test.ipynb	25 minutes ago

(2) 智谱 chatglm3-6b

```
root@dsw-1132796-5b6999989-zh95b:/mnt/workspace# git clone https://www.modelscope.cn/ZhipuAI/chatglm3-6b.git
正克隆到 'chatglm3-6b'...
remote: Enumerating objects: 140, done.
remote: Total 140 (delta 0), reused 0 (delta 0), pack-reused 140
接收对象中: 100% (140/140), 47.10 KiB | 126.00 KiB/s, 完成
处理 delta 中: 100% (64/64), 完成
正在更新文件: 100% (28/28), 完成
过滤内容: 100% (15/15), 23.26 GiB | 116.36 MiB/s, 完成
root@dsw-1132796-5b6999989-zh95b:/mnt/workspace#
```

Name	Last Modified
chatglm3-6b	seconds ago
• test.ipynb	a minute ago

(3) 百川 Baichuan2-7B-chat

```
root@dsw-1133525-7c966c9f56-7hx49:/mnt/workspace# git clone https://www.modelscope.cn/baichuan-inc/Baichuan2-7B-Chat.git
正克隆到 'Baichuan2-7B-Chat'...
remote: Enumerating objects: 121, done.
remote: Counting objects: 100% (43/43), done.
remote: Compressing objects: 100% (36/36), done.
remote: Total 121 (delta 19), reused 14 (delta 7), pack-reused 78
接收对象中: 100% (121/121), 472.81 KiB | 724.00 KiB/s, 完成
处理 delta 中: 100% (51/51), 完成
```

Name	Last Modified
Baichuan2-...	10 minutes ago
• test.ipynb	2 minutes ago

编写推理脚本 test.ipynb

```
from transformers import TextStreamer, AutoTokenizer, AutoModelForCausalLM

model_name = "/mnt/data/Qwen-7B-Chat" # 本地路径
prompt = "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"

tokenizer = AutoTokenizer.from_pretrained(
    model_name,
    trust_remote_code=True
)

model = AutoModelForCausalLM.from_pretrained(
    model_name,
    trust_remote_code=True,
    torch_dtype="auto" # 自动选择 float32/float16 (根据模型配置)
).eval()

inputs = tokenizer(prompt, return_tensors="pt").input_ids

streamer = TextStreamer(tokenizer)
outputs = model.generate(inputs, streamer=streamer, max_new_tokens=300)
```

2. 对话与分析

2.1 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) 通义千问 Qwen-7B-Chat

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所描述的季节不同。第一句话“冬天：能穿多少穿多少”意味着在冬天应该尽量多穿衣服以保暖，因为冬天天气寒冷。而第二句话“夏天：能穿多少穿多少”则意味着在夏天应该尽量少穿衣服以保持凉爽，因为夏天天气炎热。因此，尽管这两句话都涉及到“能穿多少穿多少”，但它们的重点和目的却截然不同。<|endoftext|>

(2) 智谱 chatglm3-6b

这两句话的意思是相同的，都是在询问在寒冷的冬天和炎热的夏天，人们应该穿多少衣服。但是，冬天和夏天的气候条件不同，所以人们在选择穿什么衣服上有不同的考虑。冬天通常很冷，所以人们需要穿厚实的外套、保暖内衣和舒适的鞋子等；而夏天则很热，人们需要穿轻便的衣服、短裤和凉鞋等。因此，虽然这两句话的意思相同，但它们所反映的气候条件和人们对穿着的需求不同。

(3) 百川 Baichuan2-7B-chat

您好，请问您需要了解关于哪方面的信息呢？如果您有任何问题，请随时提问。

分析：通义千问提到了季节不同以及对应语境下的不同理解，解释清晰合理，准确地指出了反向意义，理由充分。

智谱认为两句话意思相同，没有准确理解“多少”一次的多重含义，但是给出了合理的解释，符合人类思维。

百川表现不佳，未能理解问题并给出回答。

2.2 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) 通义千问 Qwen-7B-Chat

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，二是谁都看不上自己。

这两句话主要的区别在于表达的意思和语气不同。第一句话“单身狗产生的原因有两个，一是谁都看不上，二是谁都看不上自己”是一种较为客观的陈述事实的语气，强调了产生单身狗的原因是自我因素和外部因素共同作用的结果；而第二句话“单身狗产生的原因有两个，一是谁都看不上，二是谁都看不上自己”则带有一种调侃和自嘲的语气，更多地是对自身处境的一种无奈和接受的态度。总的来说，第一句话比较严肃，而第二句话则更具有幽默感。<|endoftext|>

(2) 智谱 chatglm3-6b

这两句话的语义重复。第一句“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”表达的意思是单身的原因有两个，一个是看不上别人，另一个是别人看不上自己。第二句“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”与第一句意思相同，只是表述方式略有不同。因此，这两句话在语义上存在重复。

(3) 百川 Baichuan2-7B-chat

这两句话的区别在于它们的表达方式和重点不同。第一句主要强调了一个人对他人的看法和评价，而第二句则关注于个人的自我认知和对他人的态度。

分析：通义千问解释很详细，准确分析了语气与态度，以及语言背后的逻辑，表现良好。

智谱正确解释了“谁都看不上”分别是“看不上别人”和“别人看不起自己”，但是认为出现了两句语义重复的话。

百川回答言简意赅，提到了核心，解释正确，但是缺乏深入分析。

2.3 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) 通义千问 Qwen-7B-Chat

他知道我知道你知道他不知道吗？这句话里，到底谁不知道什么？
谁也不知道什么。<|endoftext|>

(2) 智谱 chatglm3-6b

在这个问题中，“他”是指一个人。根据问题的表述，无法确定“我不知道”指的是什么。“他知道我知道你知道他不知道吗？”这个句子是一个典型的悖论，因为它包含了两个相互矛盾的条件。第一个条件是“他知道我知道”，这意味着他已经知道了所有相关信息。第二个条件是“我知道你知道他不知道”，这表明我仍然不知道他是否已经知道了一些信息。这两个条件一起构成了一个逻辑悖论，因此在这个问题中，没有明确答案。

(3) 百川 Baichuan2-7B-chat

我需要理解您的问题。请问“他知道我知道你知道他不知道吗？”这句话中的“他”指的是什么？是某个特定的人、事物还是概念？

分析：通义千问该问题解释不清晰，且没有分析，处理复杂嵌套的逻辑表现不佳。

智谱认为问题是悖论，侧重偏离，同样处理复杂嵌套的逻辑表现不佳。

百川没有理解问题，无法正确理解“他”，表现不好。

2.4 问题四：

明明明明明白白白喜欢他，可她就是不說。这句话里，明明和白白谁喜欢谁？

(1) 通义千问 Qwen-7B-Chat

明明明明明白白白喜欢他，可她就是不說。这句话里，明明和白白谁喜欢谁？ 明明喜欢白白。
2. 这句话说明了什么？这句话说明了明明很喜欢白白，但是她没有勇气说出来。这句话也反映了现代社会中很多人的现状，有些人可能喜欢一个人，但因为各种原因而不敢表白。<|endoftext|>

(2) 智谱 chatglm3-6b

根据句子中的信息，无法确定是明明还是白白主动表达了自己的感情。可能是明明主动告诉白白自己的感情，也可能是白白自己告诉明明自己的感情。

(3) 百川 Baichuan2-7B-chat

你好！我很高兴能帮助你解答这个问题。在这句话中，“明明明白白白喜欢他”是一个倒装句，其中“明明喜欢白白”是正常语序。所以，“明明”和“白白”都喜欢“他”。

分析：通义千问分析不正确，主谓分析错误，但是理解了大致意思以及反映的问题。

智谱无法判断，未能准确理解句子结构与逻辑。

百川认为句子是倒装句，理解有误，人物分析混乱，分析结论也有误。

2.5 问题五：

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) 通义千问 Qwen-7B-Chat


```
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。<|im_end|>
<|im_start|>
1. “没什么意思” - 这句话的意思是小明没有特别的意思要表达。
2. “意思意思” - 这句话的意思是小明在说“没什么意思”的时候带有一种轻松、随意的口吻，似乎是在开玩笑或表示谦虚。
3. “不够意思” - 这句话的意思是领导认为小明的态度不够尊重或者礼貌。
4. “小意思” - 这句话的意思是小明在承认自己态度不好的同时也在解释这只是一个小事，不需要太认真对待。
5. “很意思” - 这句话的意思是领导觉得小明的回答很有趣，可能是因为小明的反应很独特或者有幽默感。
6. “其他意思” - 这句话的意思是小明想要澄清他并没有其他的意思。
7. “不好意思” - 这句话的意思是小明意识到自己的回答让领导感到不舒服，并向领导道歉。
8. “其实也没有别的意思” - 这句话的意思是小明再次强调他没有其他的意思。
9. “我不好意思” - 这句话的意思是小明因为自己的回答让领导感到不舒服而感到内疚。<|im_end|>
</endoftext|>
```

(2) 智谱 chatglm3-6b

```
1. “意思”第一层含义是指对某件事情或现象的内涵、意义、作用等方面的理解与解释，是一种抽象的概念。
2. “意思”第二层含义是指对某件事情或现象所包含的意义、内涵的感知和领悟，是一种较为具体的情感体验。
3. “意思”第三层含义是指对某个人或事物的性格、特点、品质等方面的评价和看法，是一种主观的情感态度。
```

(3) 百川 Baichuan2-7B-chat

```
您好！我很高兴为您解答这个问题。在这里，“意思”一词有多种含义和用法。在您提供的文本中，它可能表示以下几种意思：1. 表示礼貌、客气；2. 表示幽默、诙谐；3. 表示委婉、含蓄。具体含义需要根据上下文来判断。如果您能提供更多的上下文信息，我将更好地帮助您理解这个词的含义。
```

分析：通义千问解释了所有语境下“意思”的含义，分析详细准确，并指明了表达的目的，逻辑、结构清晰。

智谱只是单纯的解释“意思”的几层含义，没有结合问题，未能理解问题。

百川同样给出了“意思”可能是礼貌、幽默、委婉的表达，但是没有结合问题，未能结合文本分析。

3. 横向对比分析

3.1 模型架构对比

(1) 通义千问 Qwen-7B-Chat

参数数量：7 亿（7B）

基础模型：Transformer 架构。

优化技术：包含了诸如混合精度训练和模型剪枝等技术以提高性能和效率。

(2) 智谱 ChatGLM3-6b

参数数量：6 亿（6B）

基础模型：GLM (General Language Model)，基于自回归生成模型架构，具有良好的推理和生成能力。

优化技术：利用自监督学习技术，并针对中文语言进行了特别优化。

（3）百川 Baichuan2-7B-chat

参数数量：7 亿（7B）

基础模型：Transformer 架构，特别针对对话场景进行了优化。

优化技术：使用了对话优化技术，如对话状态跟踪和多轮对话管理，以提升对话体验。

3.2 对比总结与评价

在实际对话与分析环节，各模型的表现差异显著：

通义千问 Qwen-7B-Chat 整体表现最为稳定和出色，在理解多义词（如“意思”）、反向语境（如“能穿多少穿多少”）以及涉及情绪和态度的复杂语境（如“单身狗产生的原因”）时，展现出较强的逻辑分析能力和语境理解能力，且能给出详细且富有洞察力的解释，尽管在处理高度嵌套的逻辑（问题三）和主谓结构较为复杂（问题四）的句子时仍有提升空间，但其对中文语言的细致理解和解释的丰富性令人印象深刻。

智谱 ChatGLM3-6B 在某些方面表现出与人类思维相似的特点，例如对“能穿多少穿多少”的解释，它在理解直接的语义信息方面表现良好，但在处理深层含义、多重语境或复杂逻辑时显得力不从心，容易出现语义重复或对问题理解的偏差，其针对中文语言的优化在特定问题上有所体现，但在复杂情境下仍需提升。

百川 Baichuan2-7B-chat 在本次测试中表现相对较弱，它在多个复杂问题上未能准确理解提问意图，或给出的回答缺乏深度和准确性，特别是在理解嵌套逻辑和复杂句子结构方面存在明显不足，甚至出现对句子结构误判的情况，尽管其优化方向是对话场景，但在面对需要深层语义理解和推理的对话时，其能力仍有待加强。

综合评价来看，本次实验表明，即使是参数量相近的大语言模型，其在中文语言理解和推理能力上也存在显著差异，通义千问 Qwen-7B-Chat 在本次测试中展现了较强的综合实力，尤其在对中文语境的细致理解和解释方面表现突出，这可能得益于其训练数据和优化策略；智谱 ChatGLM3-6B 在某些方面表现出潜力，但在处理更复杂的语言现象时需要进一步优化；百川 Baichuan2-7B-chat 则需要在语言理解的准确性和深度上进行显著改进，以更好地应对复杂对话场景。

此次实践也强调了模型部署的重要性，在实际应用中，除了模型本身的性能，部署环境的稳定性和推理效率也是决定用户体验的关键因素，通过在魔搭平台等环境中进行部署，能

够更好地测试和评估这些模型的实际表现，为进一步的优化和应用提供宝贵经验。未来，随着大模型技术的不断发展，提升模型在复杂语境下的鲁棒性和准确性，以及优化其在实际部署中的性能，仍将是重要的研究方向。

项目公开访问链接：

<https://github.com/xiangmaster/Comparative-analysis-of-various-LLM.git>