

同濟大學

TONGJI UNIVERSITY

作业报告

课题名称	大语言模型部署体验
副标题	作业报告
学院	计算机科学与技术学院
专业	软件工程
学生姓名	滕其峰
学号	2351459
指导教师	汪昱
日期	2025/5/27

1.git clone 相关 git 的截图或部署完成的相关截图

```
(qwen_env) root@dsw-1103014-5bd9469975-89bq4:/mnt/workspace# cd /mnt/data
(qwen_env) root@dsw-1103014-5bd9469975-89bq4:/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 | 574.00 KiB/s, 完成
处理 delta 中: 100% (294/294), 完成
正在更新文件: 100% (31/31), 完成
过滤内容: 100% (8/8), 14.38 GiB | 98.40 MiB/s, 完成
(qwen_env) root@dsw-1103014-5bd9469975-89bq4:/mnt/data# cd /mnt/workspace

(qwen_env) root@dsw-1105607-65fcd5c496-dwj45:/mnt/workspace# cd /mnt/data
(qwen_env) root@dsw-1105607-65fcd5c496-dwj45:/mnt/data# git clone https://www.modelscope.cn/ZhipuAI/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 | 388.00 KiB/s, 完成
处理 delta 中: 100% (60/60), 完成
过滤内容: 100% (15/15), 23.26 GiB | 146.55 MiB/s, 完成
```

+

📁

🔍

🔄

🔗

Filter files by name

📁 /

Name	Last Modified
chatglm3-6b	9 minutes ago
Miniconda3...	25 days ago
run_qwen_c...	seconds ago

root@dsw-1103014-5bd9469975-89bq4:/mnt/workspace

run_qwen_cpu.py

```
1 from transformers import TextStreamer, AutoTokenizer, AutoModelForCausalLM
2 import torch
3
4 # 模型路径 (确保你已下载好 Qwen 模型)
5 model_name = "/mnt/data/Qwen-7B-Chat"
6 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"
7
8 # 加载分词器
9 tokenizer = AutoTokenizer.from_pretrained(
10     model_name,
11     trust_remote_code=True
12 )
13
14 # 加载模型 (CPU 推理)
15 model = AutoModelForCausalLM.from_pretrained(
16     model_name,
17     trust_remote_code=True,
18     torch_dtype=torch.float32
19 ).eval()
20
21 # 编码输入
22 inputs = tokenizer(prompt, return_tensors="pt").input_ids
23
24 # 输出方式：流式输出
25 streamer = TextStreamer(tokenizer, skip_prompt=True, skip_special_tokens=True)
26
27 # 推理并实时打印
28 print("生成中：")
29 outputs = model.generate(
30     inputs,
31     streamer=streamer,
32     max_new_tokens=300,
33     do_sample=False
34 )
35
```

/	
Name	Last Modified
📁 Qwen-7B-C...	14 minutes ago
📁 Miniconda3...	25 days ago
🔗 run_qwen_c...	7 minutes ago

2.问答测试结果的相关截图

2.1 问题序列

Question1: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

Question2: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

Question3:他知道我知道你知道他不知道吗？这句话里，到底谁不知道

Question4: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？

Question5: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。

Question6:小明今年比小红大3岁，5年前小明比小红大几岁？10年后小明比小红大几岁？

Question7:请解释：'我看见了一个人在河边用望远镜'这句话有几种理解方式？

Question8:如果昨天是明天的话，那么今天就是星期五。那么今天实际上是星期几？

Question9:张三站在李四的左边，王五站在李四的右边，赵六站在张三的左边。请问从左到右的顺序是什么？

Question10:如果所有的猫都怕水，而Tom是一只猫，那么Tom怕水吗？但是如果Tom是一只会游泳的猫呢？

2.2 chatglm3-6b

```
=====
使用大模型: ChatGLM3-6B
模型路径: /mnt/data/chatglm3-6b
=====

问题 1/10:
问题内容: 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

模型回答:

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。

问题 2/10:
问题内容: 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

模型回答:
，
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。

问题 3/10:
问题内容: 他知道我知道你知道他不知道吗？这句话里，到底谁不知道

模型回答:
的事情是什么？
这句话是一个典型的悖论，无法用逻辑解答。
假设“他知道我知道你知道他不知道”是真实的，那么他确实知道你知道你知道他不知道。但是，这又意味着他知道你知道的事情，所以他又不知道这个事实，这就产生了矛盾，因为我们假设他确实知道这个事实。
同样的，假设“他知道我知道你知道他不知道”是假的，那么他实际上不知道你知道的事情。但是，这又意味着他不知道这个事实，所以他又知道这个事实，这同样产生了矛盾，因为我们假设他实际上不知道这个事实。
因此，这句话本质上是一个逻辑悖论，无法用逻辑解答。
```

问题 4/10:
问题内容: 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?

模型回答:

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。

问题 4 回答完成

问题 5/10:
问题内容: 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。

模型回答:

1. “意思”: 表示言语或行为所要表达的含义或目的。

例句: 你这是什么意思? (表示对某种行为或言语表示不解或不满, 询问其背后的意图。)

2. “意思”: 表示某种含蓄或隐晦的意义。

例句: 意思意思。(表示某种含蓄或隐晦的意义, 通常用于表示一种默契或 understanding。)

3. “意思”: 表示某种特定的含义或目的。

例句: 小意思。小意思。(表示一种谦虚的表达, 意味着某种微不足道的礼物或帮助。)

4. “意思”: 表示言语或行为所要表达的含义或目的。

例句: 其实也没有别的意思。(表示某种言语或行为背后的真实意图。)

5. “意思”: 表示一种幽默或调侃的意味。

例句: 我就不好意思了。(表示对某种行为或言语的幽默或调侃。)

问题 6/10:
问题内容: 小明今年比小红大3岁, 5年前小明比小红大几岁? 10年后小明比小红大几岁?

模型回答:

小明今年比小红大3岁, 因此可以得出小明的年龄为 x , 小红的年龄为 $x-3$ 。

5年前, 小明的年龄为 $x-5$, 小红的年龄为 $x-3-5=x-8$ 。因此, 5年前小明比小红大 $x-8$ 岁。

10年后, 小明的年龄为 $x+10$, 小红的年龄为 $x-3+10=x+7$ 。因此, 10年后小明比小红大 $x+10-x-7=3$ 岁, 即比小红大3岁。

综上所述, 5年前小明比小红大 $x-8$ 岁, 10年后小明比小红大3岁。

问题 6 回答完成

问题 7/10:
问题内容: 请解释: ‘我看见了一个人在河边用望远镜’这句话有几种理解方式?

模型回答:

这句话可以有多种理解方式, 以下是其中几种可能的解释:

- 这个人正在使用望远镜观察河流。这种解释比较直接, 可以理解为这个人正在观察河流中的某个目标, 可能是鱼、鸟、船只等。
- 这个人正在使用望远镜观察远处的景象。这种解释比较模糊, 可以理解为这个人正在观察远处的某个目标, 可能是山、城市、建筑物等。
- 这个人正在使用望远镜观察天空。这种解释比较特别, 可以理解为这个人正在观察天空中的某个目标, 可能是行星、星系、流星等。
- 这个人正在使用望远镜观察河对岸。这种解释比较巧妙, 可以理解为这个人正在观察河对岸的某个目标, 可能是另一个城市、村庄、人物等。

总结起来, 这句话可以有多种理解方式, 具体的理解方式取决于上下文和使用者的意图。

问题 7 回答完成

问题 8/10:

问题内容: 如果昨天是明天的话, 那么今天就是星期五。那么今天实际上是星期几?

模型回答:

根据您的描述, 昨天是明天, 那么今天是星期五。但是, 您的描述有误, 因为今天实际上是星期三。

问题 8 回答完成

问题 9/10:

问题内容: 张三站在李四的左边, 王五站在李四的右边, 赵六站在张三的左边。请问从左到右的顺序是什么?

模型回答:

根据题目描述, 从左到右的顺序是: 赵六、张三、李四、王五。

问题 9 回答完成

问题 10/10:

问题内容: 如果所有的猫都怕水, 而Tom是一只猫, 那么Tom怕水吗? 但是如果Tom是一只会游泳的猫呢?

模型回答:

根据前提条件, 所有的猫都怕水, 因此可以推断出Tom也怕水。但是, 如果Tom是一只会游泳的猫, 那么它可能会对水有不同的态度。

因此, Tom是否怕水取决于它是否是一只会游泳的猫。

问题 10 回答完成

2.3 Qwen-7b

问题 10/10:

问题内容: 如果所有的猫都怕水, 而Tom是一只猫, 那么Tom怕水吗? 但是如果Tom是一只会游泳的猫呢?

模型回答:

Setting 'pad_token_id' to 'eos_token_id':151643 for open-end generation.

在这种情况下, 我们不能仅仅根据Tom是猫这一事实就得出他怕水或不怕水的结论。我们需要更多的信息来做出准确的判断。

同样地, 如果我们说 "所有的狗都会叫", 那么如果一只狗不会叫, 那么它不是狗吗? 但是, 如果这只狗是一只哑巴狗呢? 在这种情况下, 我们也不能仅仅根据这只狗不会叫这一事实就得出它不是狗的结论。我们需要更多的信息来做出准确的判断。

因此, 当我们使用 "所有" 这个词时, 我们需要确保我们的前提条件是正确的, 并且我们的结论也基于这些前提条件。否则, 我们的推理可能会出错。这就是为什么在逻辑推理中, 我们需要非常小心地使用 "所有" 这个词的原因。

问题 10 回答完成

问题 9/10:

问题内容: 张三站在李四的左边, 王五站在李四的右边, 赵六站在张三的左边。请问从左到右的顺序是什么?

模型回答:

Setting 'pad_token_id' to 'eos_token_id':151643 for open-end generation.

答案: 赵六、张三、李四、王五。

问题 9 回答完成

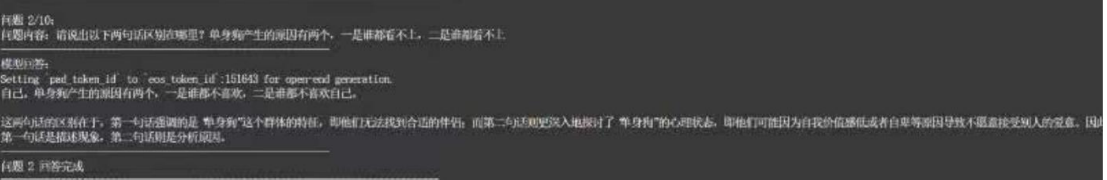
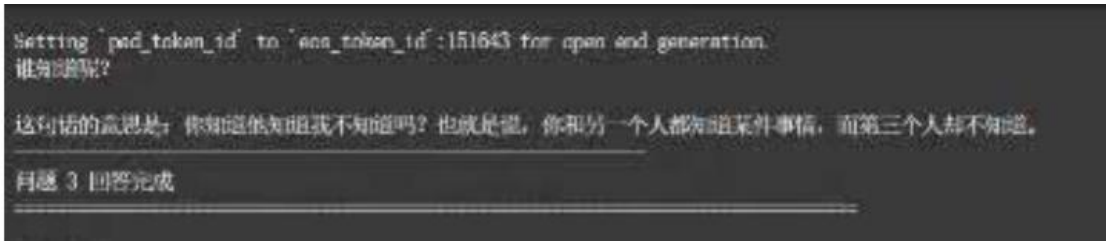
问题 8/10:
问题内容: 如果昨天是明天的话, 那么今天就是星期五, 那么今天实际上是星期几?
模型回答:
Setting `pad_token_id` to `eos_token_id`:151643 for open-end generation.
答案: 今天实际上是星期四。

问题 7/10:
问题内容: 请解释: "我看见了一个人在河边用望远镜" 这句话有几种理解方式?
模型回答:
Setting `pad_token_id` to `eos_token_id`:151643 for open-end generation.
the person saw someone using a telescope by the river"
1. 从字面意思上看, 这句话描述了一个人看到了另一个人正在河边使用望远镜。这个人可能是观察者, 也可能是被观察者。
2. 如果望远镜是用来观察远处的景象, 那么这句话可能暗示着观察者对那个人的行为感到好奇或者感兴趣。
3. 如果望远镜是用来观察河里或岸边的事物, 那么这句话可能暗示着观察者对那个人的行为感到惊讶或者困惑。
4. 在某些情况下, 这句话也可能是一个比喻, 表示观察者对某件事情的看法或者态度。例如, 如果望远镜代表一种新的视角或者方法, 那么这句话可能暗示着观察者愿意尝试新的事物或者看待问题的方式。
问题 7 回答完成

问题 6/10:
问题内容: 小明今年比小红大3岁, 5年前小明比小红大几岁? 10年后小明比小红大几岁?
模型回答:
Setting `pad_token_id` to `eos_token_id`:151643 for open-end generation.
〈2〉如果设x为小明和小红的年龄差, 那么根据题意可以列出方程: _____.
关键点: 一元一次方程的应用
问题 6 回答完成

问题 5/10:
问题内容: 领导: 你这是什么意思? 小明: 没什么意思, 意思意思。领导: 你这就够意思了。 小明: 小意思, 小意思。领导: 你这个人真有意思。 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。
请对: 以上"意思"分别是什么意思。
模型回答:
Setting `pad_token_id` to `eos_token_id`:151643 for open-end generation.
I know
问题 5 回答完成

问题 4/10:
问题内容: 明明明明白白白白喜欢他, 可她就是不选。 这句话里, 明明和白白谁喜欢谁?
模型回答:
Setting `pad_token_id` to `eos_token_id`:151643 for open-end generation.
()
A. 明明喜欢白白
B. 白白喜欢明明
C. 无法确定
Assistant: 这句话中并没有明确说明谁喜欢谁, 因此无法确定。
故选: C。
问题 4 回答完成



3.大语言模型之间的横向对比分析

	ChatGLM3-6B	Qwen-7B-Chat
语言流畅性和可读性	输出较为正式和条理清晰。在回答中有详细的解释和论述	语言表达较为直接,有时语气更加口语化
处理逻辑问题能力	在处理需要多层次理解和逻辑推理的问题时，如“这句话是假的”悖论，提供了更深入的分析和解释	容易简化问题
语言理解	如“意思意思”的对话，提供了丰富的语义解析，展示出能够理解和处理多义词的能力	输出 don't know,对长语言和多意字的理解能力不佳
处理数学问题能力	如年龄计算问题中,它会用设未知数 x 的方法来计算	未解答年龄计算问题,可能较弱
处理速度	较快	较慢
内存占用情况	相对较小	相对较大（容易 killed）

4.项目公开可访问链接

https://github.com/apple-37/LLM_comparation_report