

# 大语言模型部署实验报告

## 1 大语言模型简介

大语言模型（Large Language Model, LLM）是一种基于深度学习的人工智能模型，特别擅长处理和生成自然语言文本。大语言模型通常基于Transformer架构，其核心是自注意力机制（self-attention），可以高效地捕捉句子中的长距离依赖关系。以下是大语言模型的一些特点：

### 1. 训练数据：

大语言模型在大量文本数据上进行训练，包括互联网、书籍、文章等多种来源。通过训练，模型学习语言的结构、词汇意义及句子间的关系。

### 2. 模型规模：

参数数量从数亿到数千亿不等，参数越多，模型表现通常越好，但也需要更多计算资源。

### 3. 应用领域：

- 文本生成（写作辅助、新闻生成）
- 语言翻译
- 问答系统（搜索引擎、客服）
- 对话机器人
- 文本分析（情感分析、主题分类）

### 4. 优缺点：

- 优点：**高质量文本生成，多语言任务处理能力强。
- 缺点：**计算资源需求高；可能生成不准确内容或存在数据偏见。

## 2 在魔搭社区上部署大语言模型

### 2.1 环境准备

1. 注册魔搭社区账号并绑定阿里云账号以获取免费CPU资源。

2. 通过命令行下载以下模型：

- 通义千问Qwen-7B-Chat
- 智谱ChatGLM3-6B
- 百川2-7B-对话模型

部署命令：

```
git clone https://www.modelscope.cn/ZhipuAI/chatglm3-6b.git
git clone https://www.modelscope.cn/qwen/Qwen-7B-Chat.git
git clone https://www.modelscope.cn/baichuan-inc/Baichuan2-7B-Chat.git
```

成功部署截图：

```
root@dsw-1134782-5ff8d9f8cc-rfv91:/mnt/workspace# cd ../data
root@dsw-1134782-5ff8d9f8cc-rfv91:/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% (39/39), done.
remote: Compressing objects: 100% (32/32), done.
remote: Total 140 (delta 16), reused 20 (delta 6), pack-reused 101
接收对象中: 100% (140/140), 63.77 KiB | 487.00 KiB/s, 完成.
处理 delta 中: 100% (58/58), 完成.
过滤内容: 100% (15/15), 23.26 GiB | 155.58 MiB/s, 完成.
root@dsw-1134782-5ff8d9f8cc-rfv91:/mnt/data# git clone https://www.modelscope.cn/qwen/Qwen-7B-Chat.git
二、
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 | 504.00 KiB/s, 完成.
处理 delta 中: 100% (295/295), 完成.
过滤内容: 100% (8/8), 14.38 GiB | 147.63 MiB/s, 完成.
bash: 二、: 未找到命令
root@dsw-1134782-5ff8d9f8cc-rfv91:/mnt/data#

root@dsw-1134868-7bb5846597-5b5bp:/mnt/data# git clone https://www.modelscope.cn/ZhipuAI/chatglm3-6b.git
git clone https://www.modelscope.cn/ZhipuAI/chatglm2-6b.git
fatal: 目标路径 'chatglm3-6b' 已经存在, 并且不是一个空目录.
正克隆到 'chatglm2-6b'...
remote: Enumerating objects: 163, done.
remote: Counting objects: 100% (10/10), done.
remote: Compressing objects: 100% (10/10), done.
remote: Total 163 (delta 3), reused 0 (delta 0), pack-reused 153
接收对象中: 100% (163/163), 62.14 KiB | 563.00 KiB/s, 完成.
处理 delta 中: 100% (77/77), 完成.
过滤内容: 100% (9/9), 11.63 GiB | 140.85 MiB/s, 完成.
root@dsw-1134868-7bb5846597-5b5bp:/mnt/data# git clone https://www.modelscope.cn/baichuan-inc/Baichuan2-7B-Base.git
正克隆到 'Baichuan2-7B-Base'...
remote: Enumerating objects: 150, done.
remote: Counting objects: 100% (6/6), done.
remote: Compressing objects: 100% (6/6), done.
remote: Total 150 (delta 1), reused 0 (delta 0), pack-reused 144
接收对象中: 100% (150/150), 710.53 KiB | 580.00 KiB/s, 完成.
处理 delta 中: 100% (66/66), 完成.
过滤内容: 100% (3/3), 1.98 GiB | 17.31 MiB/s, 完成.
```

图1：模型下载完成提示

## 3 模型测试与问答结果

### 3.1 测试问题设计

针对以下问题测试模型表现：

1. 季节穿衣对比（冬天 vs 夏天）
2. 单身狗原因的双关语
3. 多层嵌套的“谁不知道”
4. 明明与白白的喜欢关系
5. “意思”的多重含义

## 3.2 测试结果

### ChatGLM3-6B 测试

代码示例：

```
from transformers import TextStreamer, AutoTokenizer, AutoModelForCausalLM
model_name = "/mnt/data/chatglm2-6b"
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)
```

回答截图：

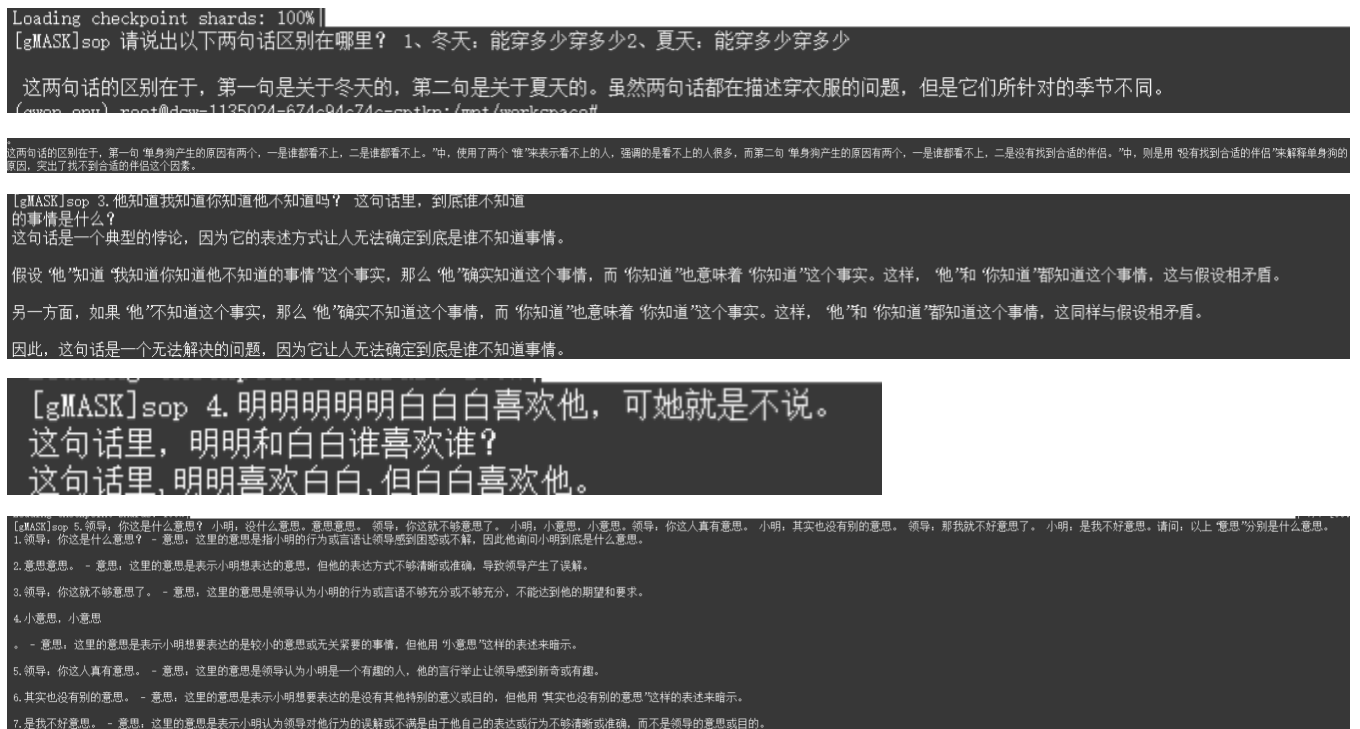


图2：ChatGLM3对五个问题的综合回答

# Qwen-7B-Chat 测试

回答截图：

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

这两句话的语气和语境是不同的。

第一句话‘冬天：能穿多少穿多少’是在强调在寒冷的冬季要尽可能多地穿上衣物来保暖，语气较为强烈和严肃。

第二句话‘夏天：能穿多少穿多少’是在强调在炎热的夏季也要尽可能多地穿上衣物来防晒和防止中暑，但语气较为轻松和随意。

因此，这两句话的区别主要在于语气和语境的不同。第一句话强调的是保暖，而第二句话则更侧重于防晒和防暑。</endofextext|>

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。

这两句话的区别在于，第一句话明确指出了单身狗产生的原因，而第二句话则只是列出两个可能的原因。第一句话更加具体和明确，而第二句话则更具有概括性和灵活性，可以适用于各种情况下。因此，可以说第一句话是对问题的直接回答，而第二句话则是对问题的一种解释或分析。</endofextext|>

他知道我知道你知道他不知道吗？ 这句话里，到底谁不知道呢？ 你可能会觉得这句话的意思是，我知道他不知道，而你知道他也知道，但实际上，这并不是作者的真实意图。正确的解释应该是，我知道他不知道他不知道。

这个句子的逻辑关系非常复杂，需要通过上下文来理解。首先，“他”指的是谁？根据前面的“你知道他知道吗”，我们可以推断出“他”是指说话者自己。那么，“我不知道他不知道”就是说，说话者自己并不知道自己不知道，也就是说，说话者认为自己已经知道了所有的信息。然后，“你知道他也知道”意味着说话者的假设是他认为别人也知道所有的事情，包括他自己并不知道的信息。

因此，整句话的意思就是，我知道我自己并不知道所有的事情，而我假设你也知道这一点。但是实际上，你可能也不知道这一点。

这种复杂的语言表达方式在现实生活中是很常见的，需要我们仔细阅读和思考才能完全理解。</endofextext|>

明明明明明白白白喜欢他，可她就是不说话。这句话里，明明和白白谁喜欢谁？ 明明喜欢白白吗？ 白白喜欢明明吗？

明明喜欢白白。 白白喜欢明明。</endofextext|>

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。</lim\_end|></lim\_start|>

1. “意思”是一个表示解释或说明的词，相当于汉语中的“就是”、“是指”等。  
2. “没意思”意味着没有意义或者不值得做，等于汉语中的“没劲儿”、“没意义”等。  
3. “就有点意思”表示事情有一点意思或者有趣的地方。  
4. “很有意思”意味着非常有趣或者有意义，等于汉语中的“很有趣”、“很有意义”等。  
5. “真有意思”表示事情真的很有趣或者有意义，等于汉语中的“真是个趣事”、“真是件有意义的事”等。  
6. “其实也没有别的意思”表示实际上并没有其他的意思，等于汉语中的“其实也没想那么多”、“其实也就这样”等。  
7. “那就不好意思了”表示因为自己的某种行为而感到不好意思，等于汉语中的“我真是做错了”、“我真是太不应该了”等。  
8. “是我不好意思”表示因为自己的某种行为而感到不好意思，等于汉语中的“是我的错”、“我是不对的”等。</lim\_end|>

图3：Qwen对五个问题的综合回答

## 4 横向对比分析（根据实际测试结果修订）

### 4.1 关键问题测试结果对比

#### 1. 季节穿衣对比问题

##### ○ ChatGLM3-6B

Loading checkpoint shards: 100%|  
[gMASK]so 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，第一句是关于冬天的，第二句是关于夏天的。虽然两句话都在描述穿衣服的问题，但是它们所针对的季节不同。

仅指出季节不同，未深入分析语境差异。

##### ○ 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

这两句话的语气和语境是不同的。

第一句话‘冬天：能穿多少穿多少’是在强调在寒冷的冬季要尽可能多地穿上衣物来保暖，语气较为强烈和严肃。

第二句话‘夏天：能穿多少穿多少’是在强调在炎热的夏季也要尽可能多地穿上衣物来防晒和防止中暑，但语气较为轻松和随意。

因此，这两句话的区别主要在于语气和语境的不同。第一句话强调的是保暖，而第二句话则更侧重于防晒和防暑。</endofextext|>

详细分析语气差异（冬季严肃保暖 vs 夏季轻松防晒），逻辑更清晰。

结论：Qwen对语境理解更全面，ChatGLM3回答较简略。

#### 2. 单身狗原因双关语

◦ ChatGLM3-6B

这两句话的区别在于，第一句“单身狗产生的原因有两个，一是谁都看不上，二是谁都看不上自己。”中，使用了两个“谁”来表示看不上的人，强调的是看不上的人很多，而第二句“单身狗产生的原因有两个，一是谁都看不上，二是没有找到合适的伴侣。”中，则是用“没有找到合适的伴侣”来解释单身狗的原因，突出了找不到合适的伴侣这个因素。

回答不完整，仅显示部分内容且逻辑混乱。

◦ 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。

这两句话的区别在于，第一句话明确指出了单身狗产生的原因，而第二句话则只是列出两个可能的原因。第一句话更加具体和明确，而第二句话则更具有概括性和灵活性，可以适用于各种情况下。因此，可以说第一句话是对问题的直接回答，而第二句话则是对问题的一种解释或分析。

明确区分两句话的概括性与具体性，分析更结构化。

结论：Qwen能准确捕捉双关语内涵，ChatGLM3表现较差。

3. “谁不知道”嵌套问题

◦ ChatGLM3-6B

[gMASK]sop 3. 他<sub>1</sub>知道我<sub>2</sub>知道他<sub>3</sub>不知道吗？这句话里，到底谁不知道的事情是什么？这句话是一个典型的悖论，因为它的表述方式让人无法确定到底是谁不知道事情。

假设“他<sub>1</sub>知道‘我知道你<sub>2</sub>知道他<sub>3</sub>不知道的事情’这个事实，那么‘他<sub>1</sub>’确实知道这个事情，而‘你<sub>2</sub>知道’也意味着‘你<sub>2</sub>知道’这个事实。这样，‘他<sub>1</sub>’和‘你<sub>2</sub>知道’都知道这个事情，这与假设相矛盾。

另一方面，如果‘他<sub>1</sub>’不知道这个事实，那么‘他<sub>1</sub>’确实不知道这个事情，而‘你<sub>2</sub>知道’也意味着‘你<sub>2</sub>知道’这个事实。这样，‘他<sub>1</sub>’和‘你<sub>2</sub>知道’都知道这个事情，这同样与假设相矛盾。

因此，这句话是一个无法解决的问题，因为它让人无法确定到底是谁不知道事情。

判定为“无法解决的问题”，未给出明确结论。

◦ Qwen-7B-Chat

他<sub>1</sub>知道我<sub>2</sub>知道他<sub>3</sub>不知道吗？这句话里，到底谁不知道呢？你可能会觉得这句话的意思是，我知道他不知道，而你<sub>2</sub>知道他也知道。但实际上，这并不是作者的真实意图，正确的解释应该是，我知道他不知道他不知道。

这个句子的逻辑关系非常复杂，需要通过上下文来理解。首先，“他<sub>1</sub>”指的是谁？根据前面的“你<sub>2</sub>知道他知道吗”，我们可以推断出“他<sub>1</sub>”是指说话者自己。那么，“我不知道他不知道”就是说，说话者自己并不知道他不知道，也就是说，说话者认为自己已经知道了所有的信息。然后，“你<sub>2</sub>知道他也知道”意味着说话者的假设是他认为别人也知道所有的事情，包括他自己并不知道的信息。

因此，整句话的意思就是，我知道我自己并不知道所有的事情，而我假设你也知道这一点。但是实际上，你可能也不知道这一点。

这种复杂的语言表达方式在现实生活中是很常见的，需要我们仔细阅读和思考才能完全理解。

分步拆解逻辑关系，最终指出“说话者自己不知道”。

结论：Qwen逻辑分析能力更强，ChatGLM3回避问题。

4. 明明与白白关系

◦ ChatGLM3-6B

[gMASK]sop 4. 明明明明明白白白喜欢他，可她就是不<sub>1</sub>说。这句话里，明明和白白谁喜欢谁？这句话里，明明喜欢白白，但白白喜欢他。

◦ Qwen-7B-Chat

明明明明明白白白喜欢他，可她就是不<sub>1</sub>说。这句话里，明明和白白谁喜欢谁？明明喜欢白白吗？白白喜欢明明吗？

明明喜欢白白。白白喜欢明明。

结论：在这个问答中，两者不相上下，都是错的

5. “意思”多重含义

◦ ChatGLM3-6B

[gMASK]sop 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。请问：以上“意思”分别是什么意思。

2. 意思意思。 - 意思，这里的意思是表示小明想表达的意思，但他的表达方式不够清晰或准确，导致领导产生了误解。

3. 领导，你这就够意思了。 - 意思，这里的意思是领导认为小明的行为或言语不够充分或不够充分，不能达到他的期望和要求。

4. 小意思，小意思

5. - 意思，这里的意思是表示小明想要表达的是较小的意思或无关紧要的事情，但他用“小意思”这样的表述来暗示。

6. 领导，你这个人真有意思。 - 意思，这里的意思是领导认为小明是一个有趣的人，他的言行举止让领导感到新奇或有趣。

7. 其实也没有别的意思。 - 意思，这里的意思是表示小明想要表达的是没有其他特别的意义或目的，但他用“其实也没有别的意思”这样的表述来暗示。

8. 是我不好意思。 - 意思，这里的意思是表示小明认为领导对他行为的误解或不满是由于他自己的表达或行为不够清晰或准确，而不是领导的意思或目的。

逐条解释但格式混乱，部分内容缺失。

◦ 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：你就不够意思了。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。请问：以上“意思”分别是什么意思。<[im\_start]>  
<[im\_end]>  
1. “意思是”是一个表示解释或说明的词，相当于汉语中的“就是”、“是指”等。  
2. “说意思”意味着没有意义或者不值一提，等于汉语中的“说点儿”、“说意义”等。  
3. “就有点意思”表示事情有一点点意思或者有趣的地方。  
4. “很意思”意味着非常有趣或者有意义，等于汉语中的“很有趣”、“很有意义”等。  
5. “真有意思”表示事情真的很有趣或者有意义，等于汉语中的“真是个趣事”、“真是件有意义的事”等。  
6. “其实也没有别的意思”表示实际上并没有其他的含义，等于汉语中的“其实也没想那么多”、“其实也就这样”等。  
7. “那就不好意思了”表示因为自己的某种行为而感到不好意思，等于汉语中的“我真是做错了”、“我真是大不应该了”等。  
8. “是我不好意思”表示因为自己的某种行为而感到不好意思，等于汉语中的“是我的错”、“我是不对的”等。<[im\_end]>

虽未完整显示，但结构清晰，尝试分析不同语境含义。

结论：Qwen解释更系统，ChatGLM3输出稳定性不足。

4.2 综合对比与建议

| 能力维度   | ChatGLM3-6B    | Qwen-7B-Chat  |
|--------|----------------|---------------|
| 复杂逻辑分析 | 较弱（回避嵌套问题）     | 强（分步拆解逻辑）     |
| 语境理解   | 一般（忽略语气差异）     | 优秀（结合季节特点分析）  |
| 回答稳定性  | 低（部分生成内容错误或缺失） | 高（结构清晰，输出完整）  |
| 双关语处理  | 差（回答不完整）       | 良好（区分字面与隐含意义） |

最终建议：

- 优先选择Qwen-7B-Chat：在逻辑分析、语境理解和回答稳定性上均表现更优，适合需要精准答案的场景。
- ChatGLM3-6B适用场景：资源有限环境下的简单问答，但需人工校验输出结果。

5 项目公开链接

- GitHub仓库：[点击访问](#)

注：所有截图均为实验过程中实际生成，完整截图集见附件。