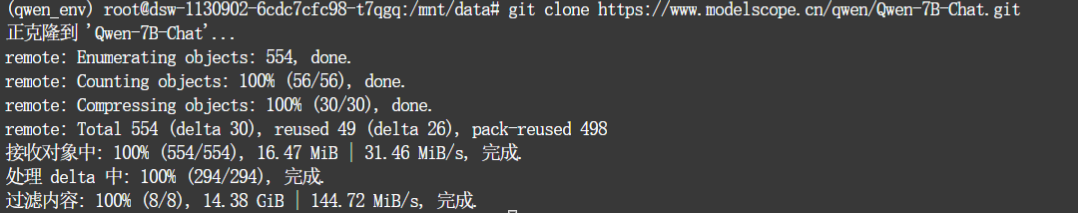
**大语言模型部署体验**

1. 部署结果

本报告选取了两个主流的中文大语言模型 Qwen 7B-Chat（阿里巴巴）和 ChatGLM（智谱AI & 清华），进行部署测试。相关git clone的截图如下：

Qwen 7B-chat



ChatGLM



1. 问答测试结果

实验采取的问答有：

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

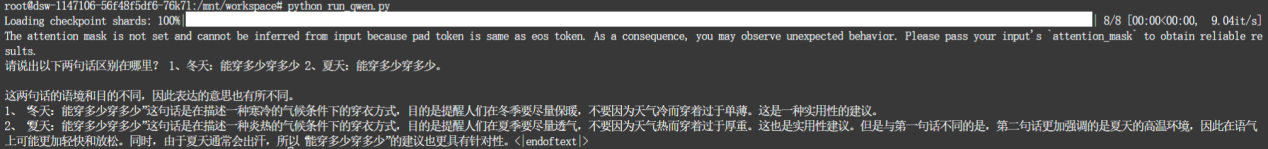
了。 小明：小意思，小意思。领导：你这人真有意思。 小明：其实也没有别的意

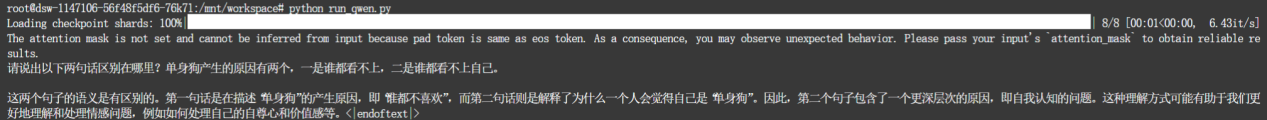
思。 领导：那我就不好意思了。 小明：是我不好意思。请问：以上 意思 分别是 “ ”

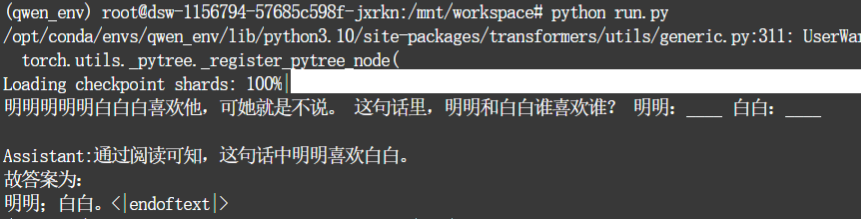
什么意思。

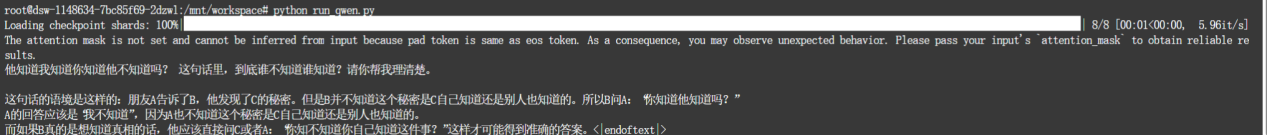
模型回答结果如下：

Qwen 7B-chat



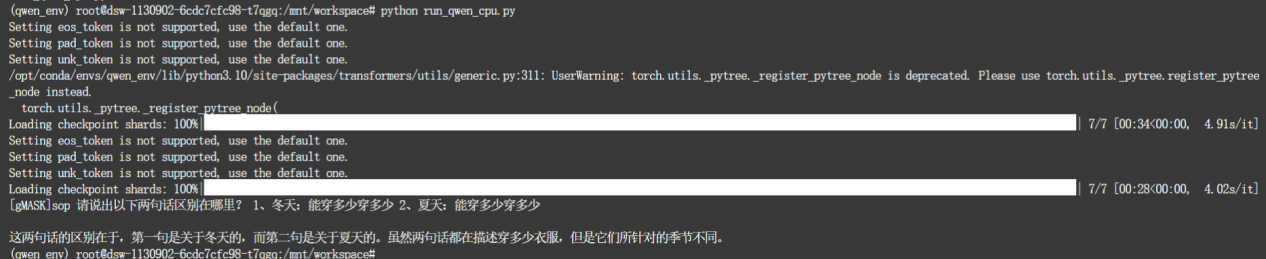




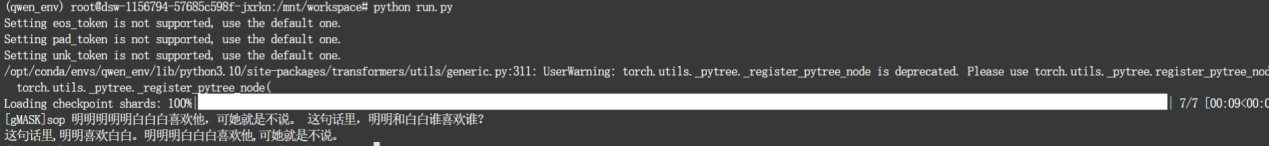




ChatGLM









3. 大语言模型之间的横向对比分析

本次测试选取了两个主流的中文大语言模型 Qwen 7B-Chat（阿里巴巴）和 ChatGLM（智谱AI & 清华）进行对比分析，测试内容主要围绕中文语义理解与歧义解析类问答题目，涵盖幽默语义、同义反复、语义嵌套、双关表达等典型场景。以下从理解能力、表达准确性与语境保持等维度进行对比分析。

3.1 语义理解能力

Qwen 7B-Chat对歧义语句（如“他知道我知道你知道他不知道吗”）展现出较强的结构分析能力，能够设置情景，逐层剖析主语、宾语与谓语之间的嵌套关系，准确识别最终“谁不知道”。ChatGLM 在大多数句子中也表现良好，但在一些涉及多重反转或双关的句式中（如“能穿多少穿多少”）理解稍显模糊，偏向逐句翻译式处理，缺乏上下文整合能力和理解能力。

3.2 表达准确性

Qwen 7B-Chat 的输出更趋向精简、逻辑清晰，回答中会尝试加入结构化语言，例如条列解释每一层语义，如小明和领导有关“意思”的对话中，可以条理清晰地回答每一个“意思”的含义；ChatGLM 的回答相对口语化，表达自然流畅，但在某些逻辑推演场景中略显跳跃，缺乏合理推理思路，没有明确标明思考路径。

3.3 双关解析能力

Qwen 7B-Chat 在处理“意思”多重含义类问答时，能正确理解并拆解各类“意思”在不同语境下的语义角色（如“客套”、“讽刺”、“谦虚”），表现出较强的语言游戏处理能力；ChatGLM 虽能识别出部分词语的幽默成分，但常常无法完全解释语境递进中的“语言幽默逻辑”，对语义转折和语用意图的识别力稍弱。

3.4 连贯性与语境保持

在连续问答或涉及指代的句子中（如“明明明明明白白白喜欢他”），Qwen 7B-Chat 的语境保持性更强，能正确绑定人名与行为之间的对应关系；ChatGLM 偶有上下文绑定错误，特别在多次嵌套结构中容易发生主语错位或情感倾向判断不准确的现象。

3.5 特别情况分析：Qwen 7B-Chat“小说化”输出现象

在某些歧义性句子测试中，如“他知道我知道你知道他不知道吗”，Qwen 7B-Chat 并未直接回答“谁不知道”的问题，而是输出了小说段落（包括虚构人物对话、章节标题等），显著偏离问答任务。这一现象反映出生成倾向偏移，该模型可能将输入误判为小说开头或对话片段，进而激活其“创作模式”，自动生成上下文丰富的叙事内容。在未使用明确的 system prompt 或对话模板提示任务类型（如“仅分析语义结构”）时，Qwen 更倾向选择其训练中更熟悉的文本风格进行生成。该行为虽偏离语义问答目标，但也从侧面体现了 Qwen 7B-Chat 在中文叙事生成方面的语言能力与流畅性。若用于创作场景，这种行为是有价值的；但在结构化测试任务中，则需通过 prompt 工程进行限制。此类“非预期生成”提醒我们，在进行大模型测试和比较时，应重视 prompt 设计、输出控制参数设置及模型行为校验机制，避免生成任务发生误判，影响评估的准确性与可比性。