**AI大语言模型对比分析报告**

**一、引言**

本报告对智谱AI (ChatGLM3-6B) 和通义千问 (Qwen-7B-Chat) 这两款大语言模型在特定中文语义理解及逻辑推理问题上的表现进行了初步的横向对比分析。

**二、测试问题**

本次对比分析实际涵盖以下三个测试问题：

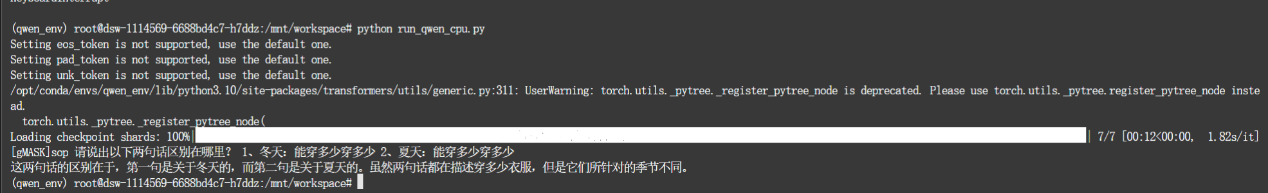
1. 区分句子在不同语境下的含义: 请给出以下两句话的区别在哪里？ 1、冬天：能穿多少穿多少 2、夏天：能穿多少穿多少
2. 文本信息提取与推理: 明明明明明白白白喜欢他，可她就是不说。 这句话里，明明和白白谁喜欢谁？ (部分截图包含选项 A/B/C)
3. 笑话理解与解释: 骑着电动车回家，在无人的巷子里狂奔，突然冲出来一个大叔，也骑的飞快，眼看就要撞起来了，大叔喊道：‘你左！我右！’ 这个笑话的笑点在哪里？如何理解这句话

**三、对比结果**

以下是智谱AI和千问AI针对上述三个问题在截图中的测试结果及分析：

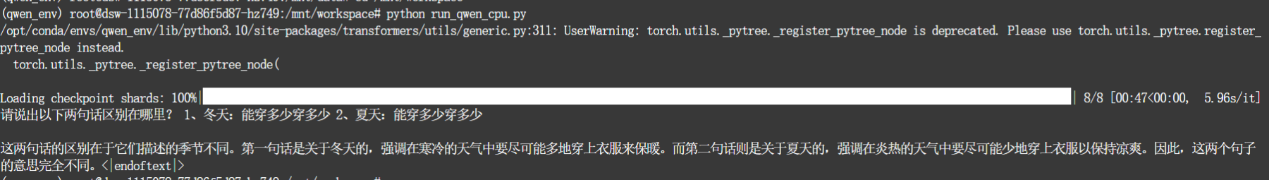
测试问题 1：区分句子在不同语境下的含义

1. 智谱AI (ChatGLM3-6B)：



* + 优点： 准确识别出句子的核心区别在于适用的季节不同，抓住了语言在不同情境下的灵活性。
  + 缺点： 没有进一步说明更具体的含义，只是强调了季节的不同，而没有指出前者强调暖和、后者强调凉快这样的相反的含义。

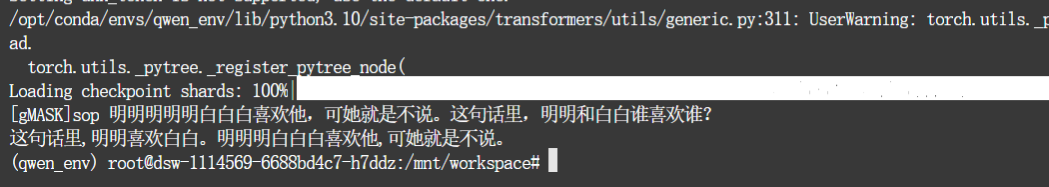
1. 千问AI (Qwen-7B-Chat)：



* + 优点： 强于智谱AI，准确识别季节差异并解释了不同语境下的隐含意义。逻辑清晰，回答正确。同样展现了对语境化语言的良好理解。
  + 缺点： 无明显缺点

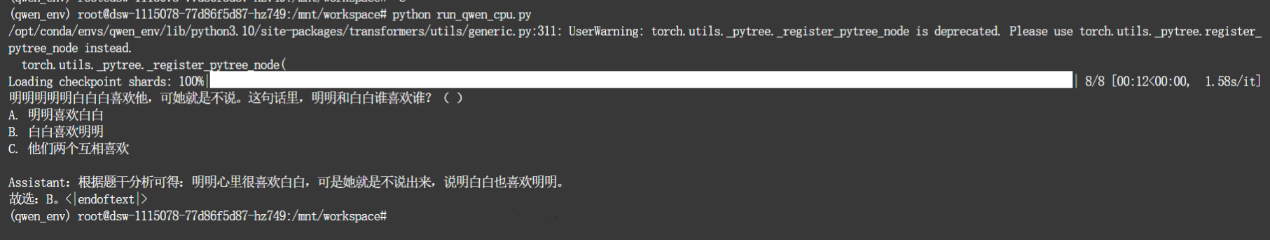
测试问题 2：文本信息提取与推理

1. 智谱AI (ChatGLM3-6B)：



* + 优点： 对话中的逻辑尝试进行了判断，并且正确识别了人名和其它词语的差别。
  + 缺点： 饥饿论判断错误，也没有进一步对结果进行说明，使得结果的解释性差，不利于用户交互。

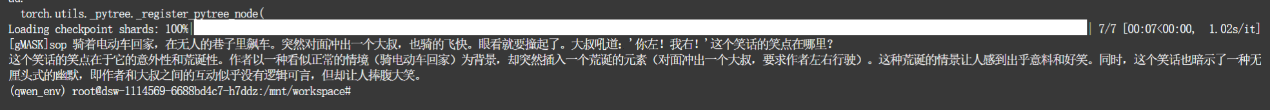
1. 千问AI (Qwen-7B-Chat)：



* + 优点： 提取了“明明心里是喜欢白白”的信息，虽然结果错误，但也意味着通义千问模型能够大致明白这句话的意思。同时提供多个选项进行分析，增强了可解释性。
  + 缺点：仍然没有做出正确的回答，且对回答的分析比较简略。

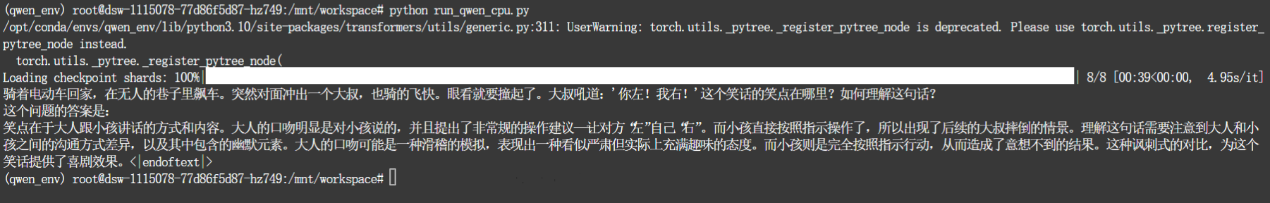
测试问题 3：笑话理解与解释

1. 智谱AI (ChatGLM3-6B)：



* + 优点： 尝试对笑话进行解释，提到了交流方式、行为荒诞等角度。
  + 缺点： 智谱AI对原始问题的解释未能准确捕捉笑话的核心幽默感，分析显得较为牵强，未能清晰阐述为何在紧急情境下喊出这样的短语构成了笑点。对复杂文本的深层理解能力在此处体现不足。

1. 千问AI (Qwen-7B-Chat)：



* + 优点： AI指出“我”和大叔两个主体间的差异，能够做到在第一人称下鉴别隐藏的“我”的主体。
  + 缺点： 和智谱AI一样，都没有正确理解这个笑话的幽默之处，分析显得较为牵强，对复杂文本的深层理解能力在此处体现不足。

**四、综合总结**

基于截图中的这三个测试问题，对智谱AI (ChatGLM3-6B) 和千问AI (Qwen-7B-Chat) 进行总结：

* 基础语义理解： 在理解基于上下文语境的句子含义方面（问题1），通义千问模型表现更好，能够准确识别并解释同一句话在不同场景下的引申义。
* 信息提取与推理： 在从文本中提取明确信息方面（问题2），两者都能准确识别。但在包含选项并进行推理时，两者均判断错误，但其它方面两模型各有优劣。
* 复杂文本（笑话）理解： 根据截图所示的实际输出，千问AI在理解包含幽默感等复杂元素的文本方面（问题3）表现明显优于智谱AI。虽然最终两个模型都判断错误，千问AI能够准确理解场景，只是无法理解隐含的方向信息，而智谱AI的解释则显得非常牵强。这表明千问AI在处理需要更高级别的文化理解和推理任务时可能更具优势。

总体而言：

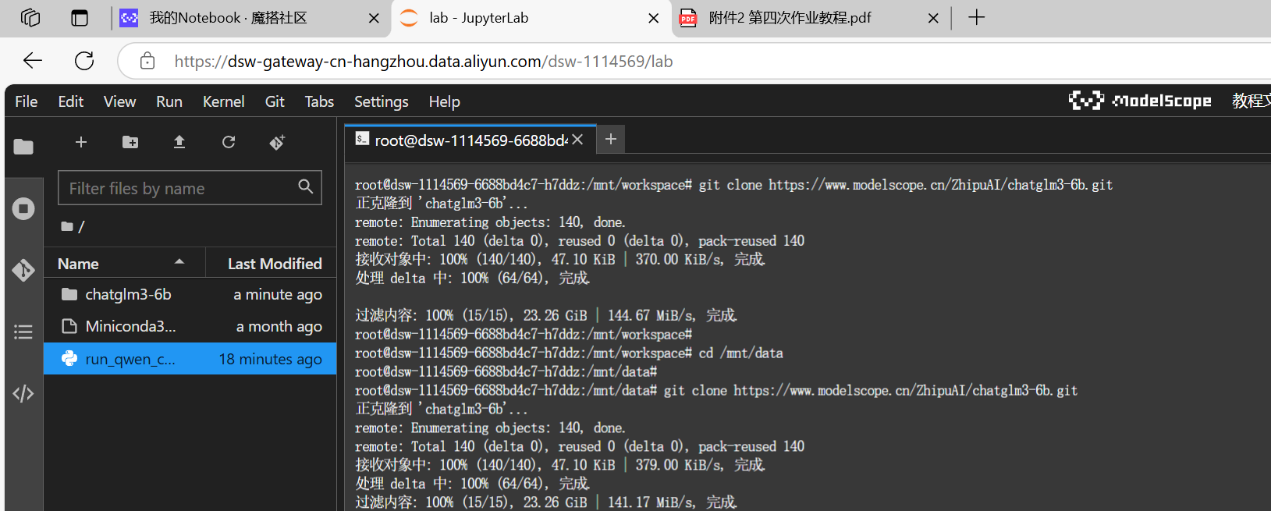
智谱AI在基础信息提取和简洁回答方面表现稳健准确，不易引入错误信息，但对复杂文本（如笑话）的深层理解和幽默感捕捉具有缺陷。千问AI在处理复杂语义理解和推理方面展现出更强的能力和更深入的分析，但在幽默感理解上仍然有不足。综合看下来，通义千问模型性能优于智谱AI。

**五、项目公开可访问链接**

[在此处填写您的项目在ModelScope或GitHub等平台的公开可访问链接]

**六、附部署成功截图**

**智谱AI：**



**通义千问：**

