# 大语言模型横向对比分析报告

**一：项目简介**

本次项目使用魔搭平台，部署并测试多个大语言模型，针对不同的模型进行应用场景的测试，并进行不同模型之间的横向对比。

本次项目所测试的模型如下：

[通义千问Qwen-7B-Chat](https://www.modelscope.cn/models/qwen/Qwen-7B-Chat/summary" \t "https://canvas.tongji.edu.cn/courses/97328/assignments/_blank)

[智谱ChatGLM3-6B](https://www.modelscope.cn/models/ZhipuAI/chatglm3-6b/summary" \t "https://canvas.tongji.edu.cn/courses/97328/assignments/_blank)

[百川2-7B-对话模型](https://www.modelscope.cn/models/baichuan-inc/Baichuan2-7B-Chat/summary" \t "https://canvas.tongji.edu.cn/courses/97328/assignments/_blank)

问答测试如下：

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。

**二：模型搭建**

1. 安装基础依赖

pip install \

"intel-extension-for-transformers==1.4.2" \

"neural-compressor==2.5" \

"transformers==4.33.3" \

"modelscope==1.9.5" \

"pydantic==1.10.13" \"sentencepiece" \

"tiktoken" \

"einops" \

"transformers\_stream\_generator" \

"uvicorn" \

"fastapi" \

"yacs" \

"setuptools\_scm"

pip install fschat --use-pep517

2.下载大模型到本地

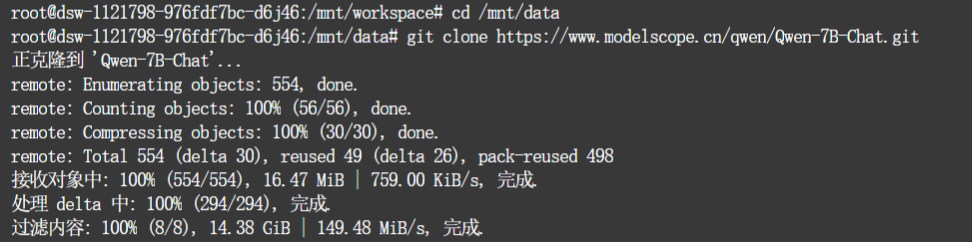
2.1 切换到数据目录

cd /mnt/data

2.2 下载对应大模型

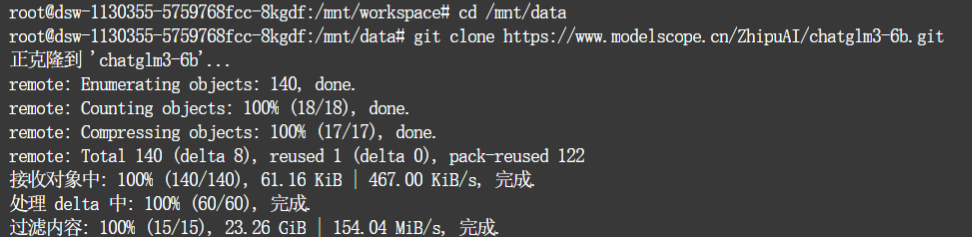
git clone <https://www.modelscope.cn/qwen/Qwen-7B-Chat.git>

出现以下信息代表模型下载成功

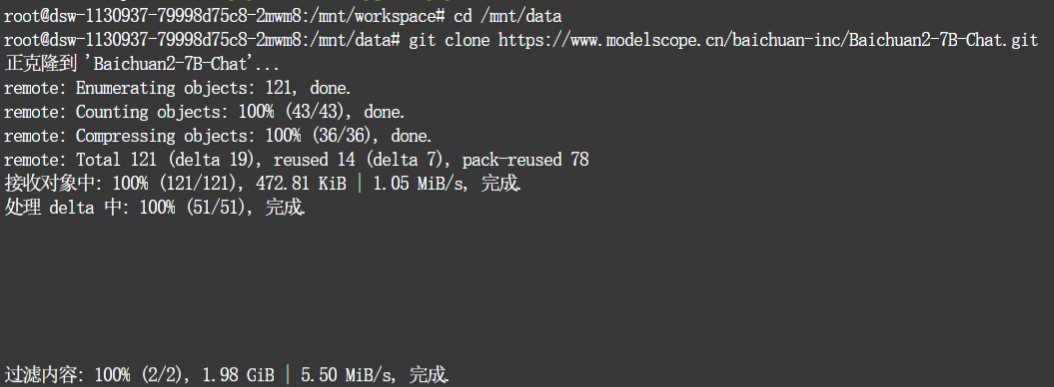


其余两模型类似

git clone <https://www.modelscope.cn/ZhipuAI/chatglm3-6b.git>



git clone https://www.modelscope.cn/baichuan-inc/Baichuan2-7B-Base.git



**三：问答横向对比**

问答截图附在报告结尾

**3.1 各应用场景具体对比**

**问题一**：冬夏穿衣提醒句子对比分析

各模型回答对比

| **模型** | **回答要点** | **准确性** | **完整性** | **逻辑性** |
| --- | --- | --- | --- | --- |
| 通义千问 | 1.季节不同（冬夏） 2.侧重点不同（保暖/防暑） | 准确指出季节差异与目的 | 覆盖季节、目的两方面，但表述较简略 | 逻辑清晰，直接对比 |
| 智谱AI | 1.季节差异（冬夏） 2.穿衣数量建议不同 | 正确区分季节与穿衣策略 | 仅说明现象，未提“保暖/防暑”目的 | 逻辑连贯，但深度不足 |
| 百川 | 1.季节穿衣策略差异 2.强调环境适应重要性 | 准确且补充了“环境适应”的深层逻辑 | 包含季节、策略、意义，最完整 | 逻辑层次分明，从现象到本质 |

总结:

百川模型表现最佳，不仅区分季节和策略，还升华到环境适应的重要性，回答最全面；

通义千问表现也不错，准确清晰说明侧重点；

智谱AI基础正确，但深度和完整性稍逊。

**问题二**：“单身狗原因”句子逻辑对比

各模型回答对比

| **模型** | **回答要点** | **准确性** | **完整性** | **逻辑性** |
| --- | --- | --- | --- | --- |
| 通义千问 | 1.逻辑顺序不同（先他人后自己/先自己后他人） 2.强调重点不同（自身因素/外界因素） | 准确分析顺序与重点差异，但对我的问题有所偏差，把两句话理解为后面两句的顺序问题（不过也可以） | 详细拆解逻辑链条，包含原因解释，但有些强加逻辑 | 逻辑严谨，分点清晰，但理解有一定偏差，有些强加逻辑 |
| 智谱AI | 1.“谁”的指代与主动/被动关系 2.双向原因表述 | 指代分析有些奇怪，并存在矛盾（两句表述相同却强行区分主动被动） | 试图从语法角度分析，但对句子实际含义理解 | 逻辑不够严谨，存在错误解读 |
| 百川 | “自身挑剔+自己缺乏自信” | 有些偏离原句内容，自行替换原因，第二点稍微准确，但第一点不太正确，不该是缺乏勇气而是别人看不上他，主语错误 | 回答比较简单且有一定问题 | 逻辑有些错误，原句结构理解一般 |

总结：

通义千问准确理解句子逻辑，分析深入且条理清晰，但强加前后逻辑关系。

智谱AI也有错误，将相同句子强行区分主动被动，逻辑混乱。

百川模型错误更多，两点里有一点表述错误有些明显

**问题三**：“他知道我知道你知道他不知道吗” 指代分析

各模型回答对比

| **模型** | **回答要点** | **准确性** | **完整性** | **逻辑性** |
| --- | --- | --- | --- | --- |
| 通义千问 | 1. 涉及三个角色（他、我、你） 2. 推断“他不知道”，但表述复杂绕口 | 部分正确（角色数量），但未明确“谁不知道” | 回答内容十分多，试图拆解对话关系，但结论模糊 | 逻辑有些混乱，未清晰定位“不知道”的主体 |
| 智谱AI | 1.谜语/绕口令性质 2.结论：“他不知道的事情”是未知对象 | 错误理解问题，未明确具体谁不知道 | 仅解释形式，没有拆解问题本质，未解决核心问题 | 逻辑不连贯，未回答关键疑问 |
| 百川 | 1.明确三个人物（A、B、C） 2.结论：C是不知道的人 | 准确划分角色，清晰定位“C不知道” | 回答简洁，完整拆解逻辑链条，从A到B到C | 逻辑严密，结论明确，解决核心问题 |

总结：

百川模型表现突出，通过角色定义和逻辑推导，明确指出“C不知道”，回答最准确，并且十分简洁。

通义千问和智谱AI均未清晰解决“谁不知道”的问题，智谱AI解答偏离更厉害。

**问题四**：“明明明明明白白白喜欢他” 指代分析

各模型回答对比

| **模型** | **回答要点** | **准确性** | **完整性** | **逻辑性** |
| --- | --- | --- | --- | --- |
| 通义千问 | 1. 结合生活常识选C（互相喜欢） 2. 强调“喜欢是双向的” | 错误解读，应该是白白喜欢明明 | 补充“生活常识”依据，但解释奇怪 | 有逻辑，未理解“明明 明明 明白 白白喜欢他”的含义 |
| 智谱AI | 直接结论：明明喜欢白白 | 错误解读，应该是白白喜欢明明 | 回答简单，如同复述，没有展开解释 | 逻辑片面，如同复述，未理解“明明 明明 明白 白白喜欢他”的含义 |
| 百川 | 结论：彼此喜欢 | 错误解读，应该是白白喜欢明明 | 表述简洁，但未展开解释 | 没有逻辑，答案错误 |

总结：

三个模型在这个问题上全军覆没，没有人回答正确是白白喜欢明明，通义千问有一定推理但结果错误，其余两模型直接给出答案但也同样错误。

**问题五**：“意思” 一词多义分析

各模型回答对比

| **模型** | **回答要点** | **准确性** | **完整性** | **逻辑性** |
| --- | --- | --- | --- | --- |
| 通义千问 | 1. 列举6种“意思”含义 2.涵盖目的、口语表达、歉意等 | 区分不同语境下的含义还算准确，理解有部分错误，比如对“那我就不好意思了”重复解释两次，两次解释不一样 | 覆盖场景不全面，回答中少了几个“意思”的理解，分析从词义到情感比较详细 | 按语境类型排列排序奇怪，部分分析有误 |
| 智谱AI | 1.结合领导与小明对话分析8处“意思” 2.包含目的、敷衍、感谢等含义 | 部分正确（如“含义/目的”），但存在错误（如将“我就不好意思”解释为“对小明的回答感到意外或困惑”） | 回答完整（可能由于我参数设置问题最后一句没解释） | 逻辑按对话顺序展开，但部分分析有误 |
| 百川 | 1.总结6种对话中的“意思”含义 2.涉及兴趣、敬意、不满等 | 部分正确，但存在不少错误（如将“没什么意思”解释为“不关心”） | 覆盖场景不全面，没有完全解释所有“意思”的含义 | 按顺序为逻辑分类，条理清晰，部分分析有误 |

总结：

三种模型均有对有错，准确率都差不多。分析都还算全面有逻辑，但都没有完整的回答所有的“意思”的含义并且对部分的解释有失偏颇。智谱AI回复在三者中比较突出，虽然有两句错误明显，但回答更加完整一些，正确的回答逻辑性很强。

**3.2 三个大模型整体综合对比**

| **维度** | **通义千问** | **智谱AI** | **百川** |
| --- | --- | --- | --- |
| **准确性** | 2题完全正确，1题错误，2题部分正确 | 3题部分正确，2题错误 | 2题完全正确，2题部分正确，1题错误 |
| **完整性** | 问题一、三回答完整，问题二分析深入；整体回答比较详细但有些繁琐 | 问题一、二、三回答完整性差，问题四像是复述，问题五完整性较好 | 问题二、三、四、五回答完整且深入，问题一补充环境适应逻辑 |
| **逻辑性** | 多数问题逻辑清晰，问题三稍显混乱，问题四分析错误 | 问题二、三逻辑断裂，问题四分析错误，整体分析多是重复原句 | 大部分问题逻辑严密，推导过程清晰，问题四分析错误，问题五错误较多 |
| **语言理解能力** | 能理解基础语义，但复杂逻辑（如问题三）处理稍弱 | 复杂语义理解能力较弱，易出现误读 | 复杂语义（如绕口令、多义词）理解能力最强 |
| **综合表现** | 中等偏上，适合基础语义分析 | 整体表现一般，除第五题处理不错，很多都是原句复述，表现较差 | 综合能力强，尤其在逻辑推理和多义分析中表现突出 |

**最终结论：**

百川2-7B-对话模型综合能力较好，在逻辑推理、复杂语义理解和回答完整性上表现不错，回答简单但大部分正确。

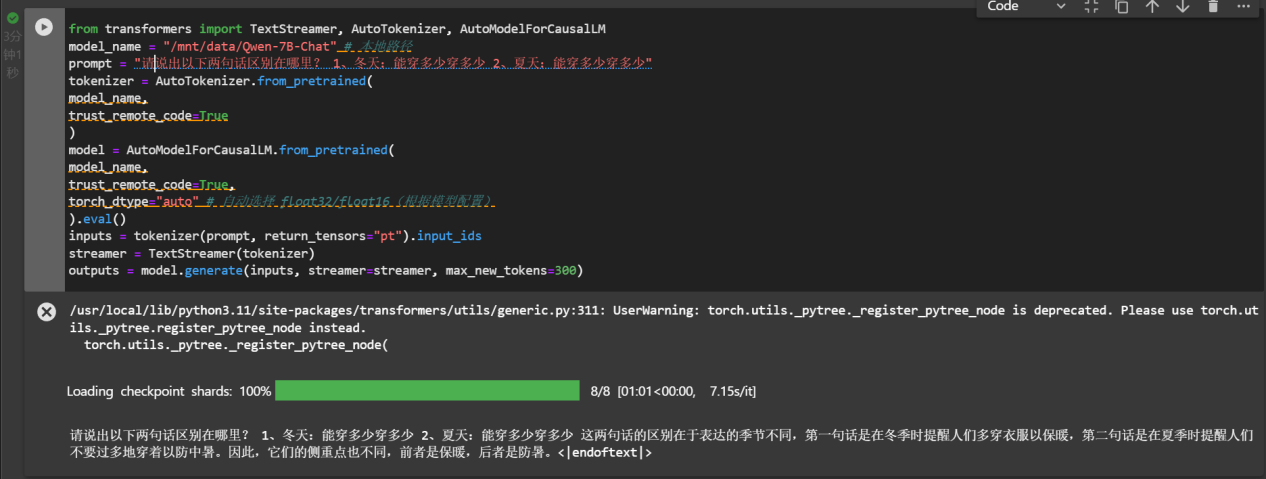
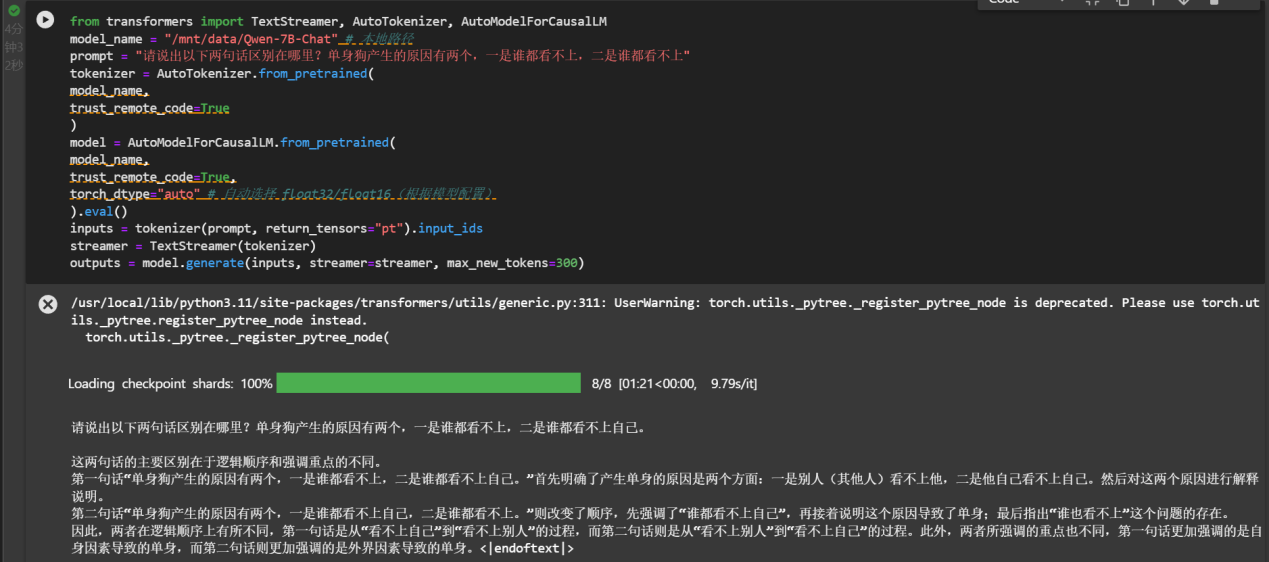
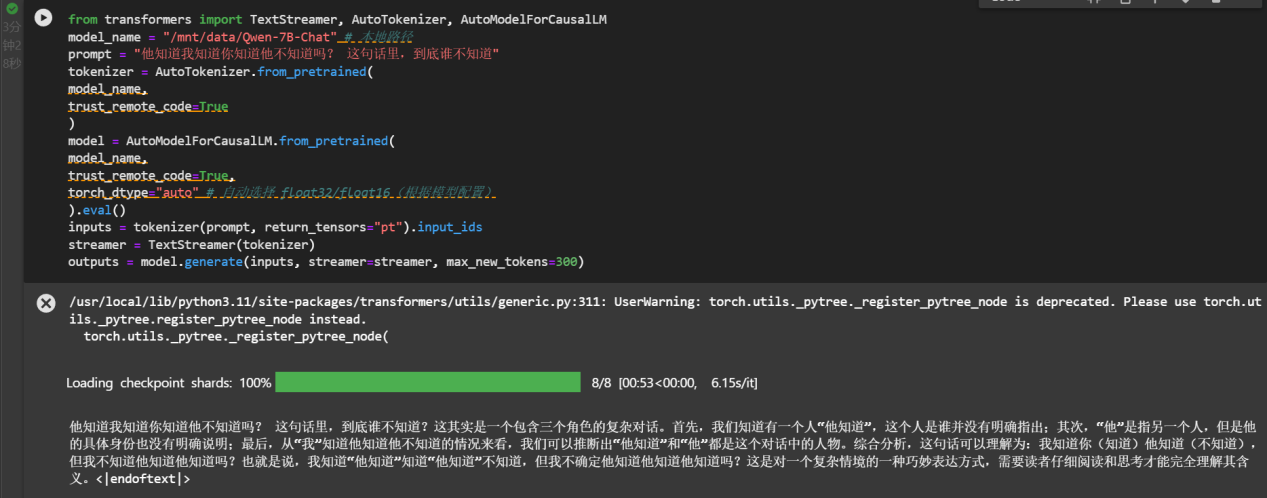
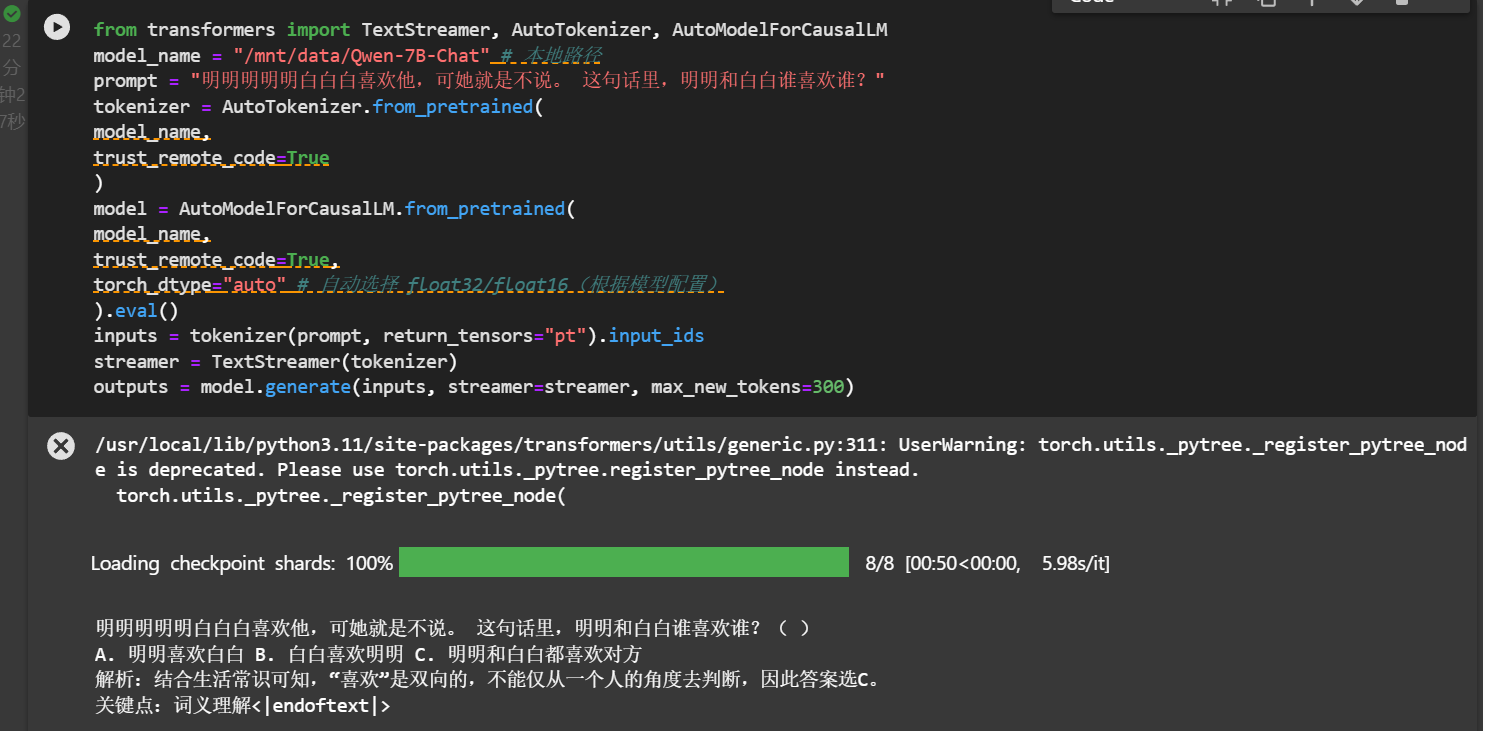
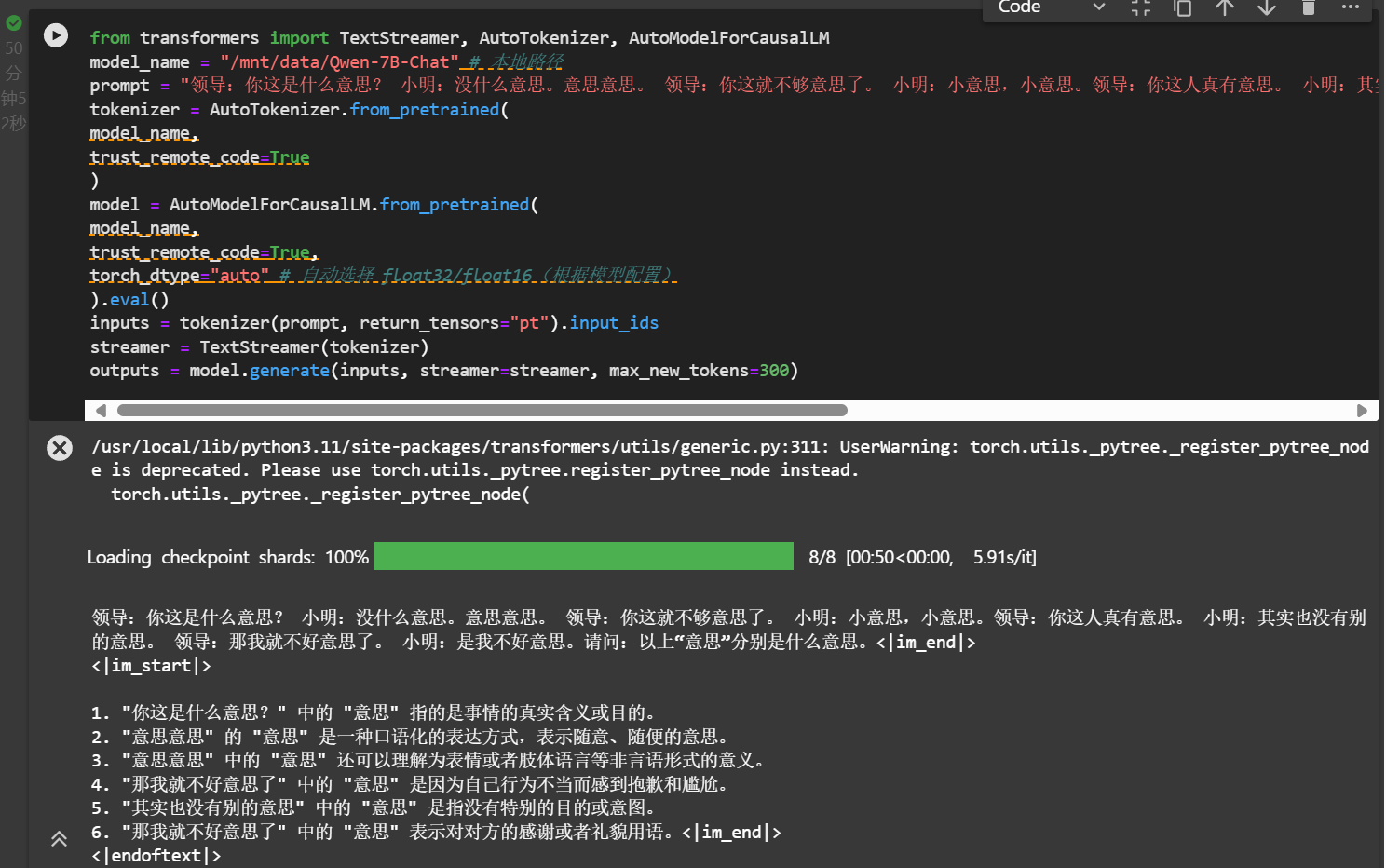
通义千问-7B-Chat基础扎实，多数问题回答准确，回答比较详细但繁琐，可能陷入大模型幻觉。

智谱ChatGLM3-6B表现不稳定，处理复杂语义存在明显误读和逻辑错误，但又在处理最长语句时表现突出，语言理解能力有待提升。

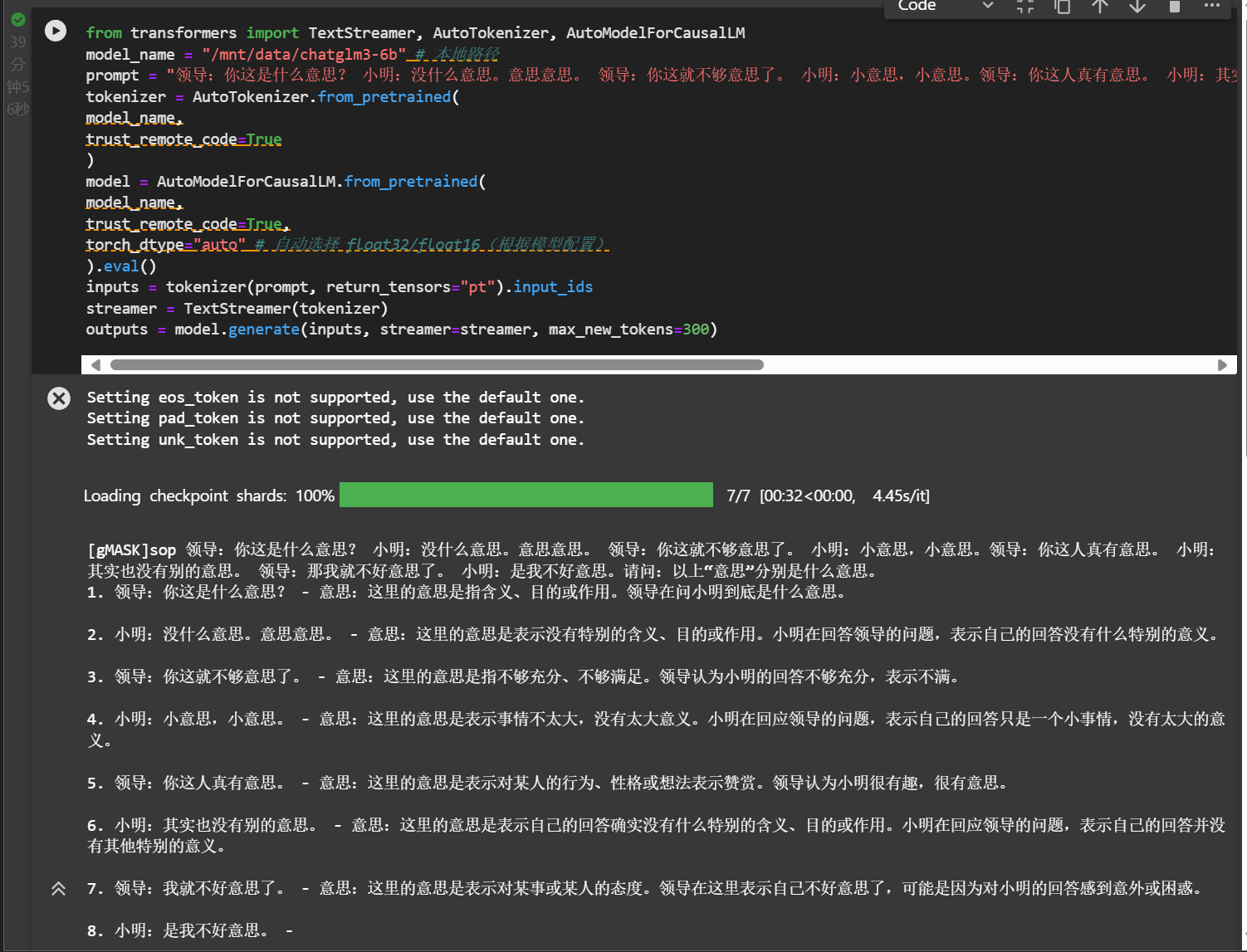
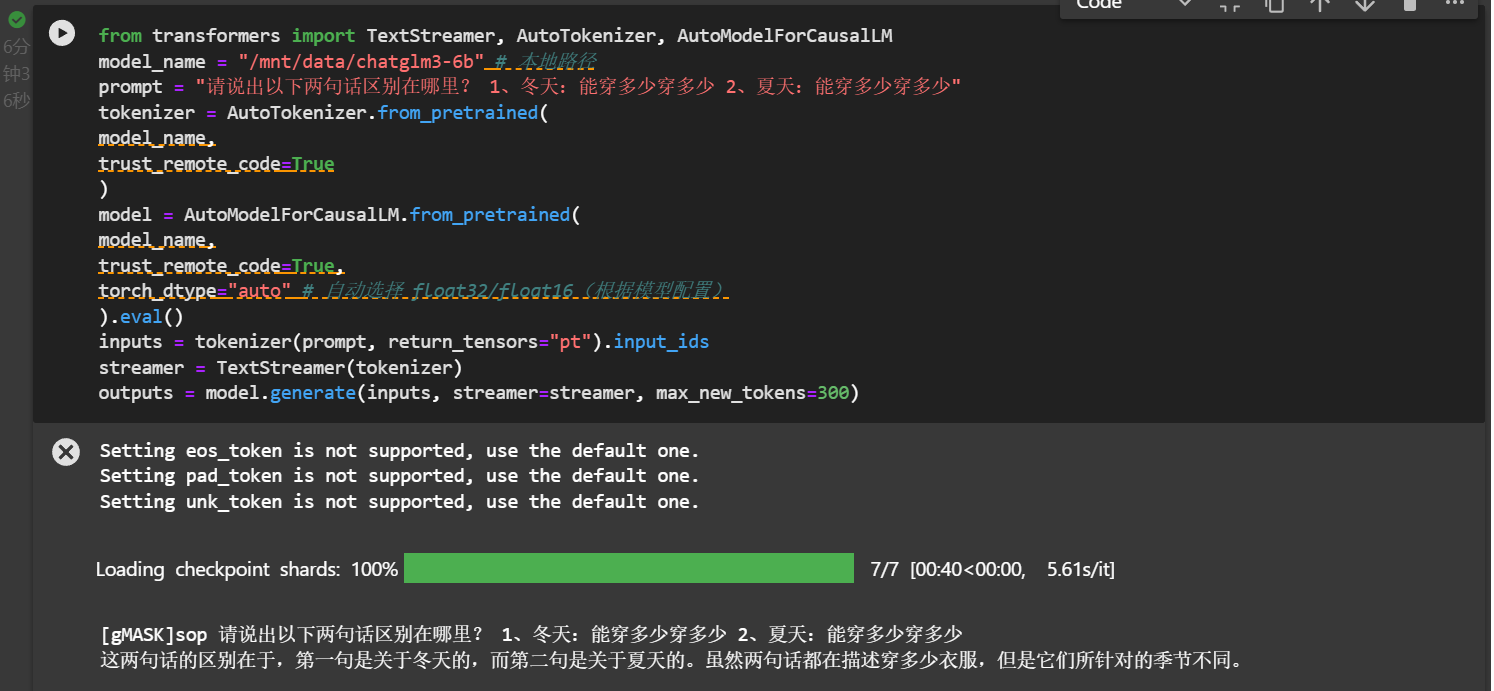
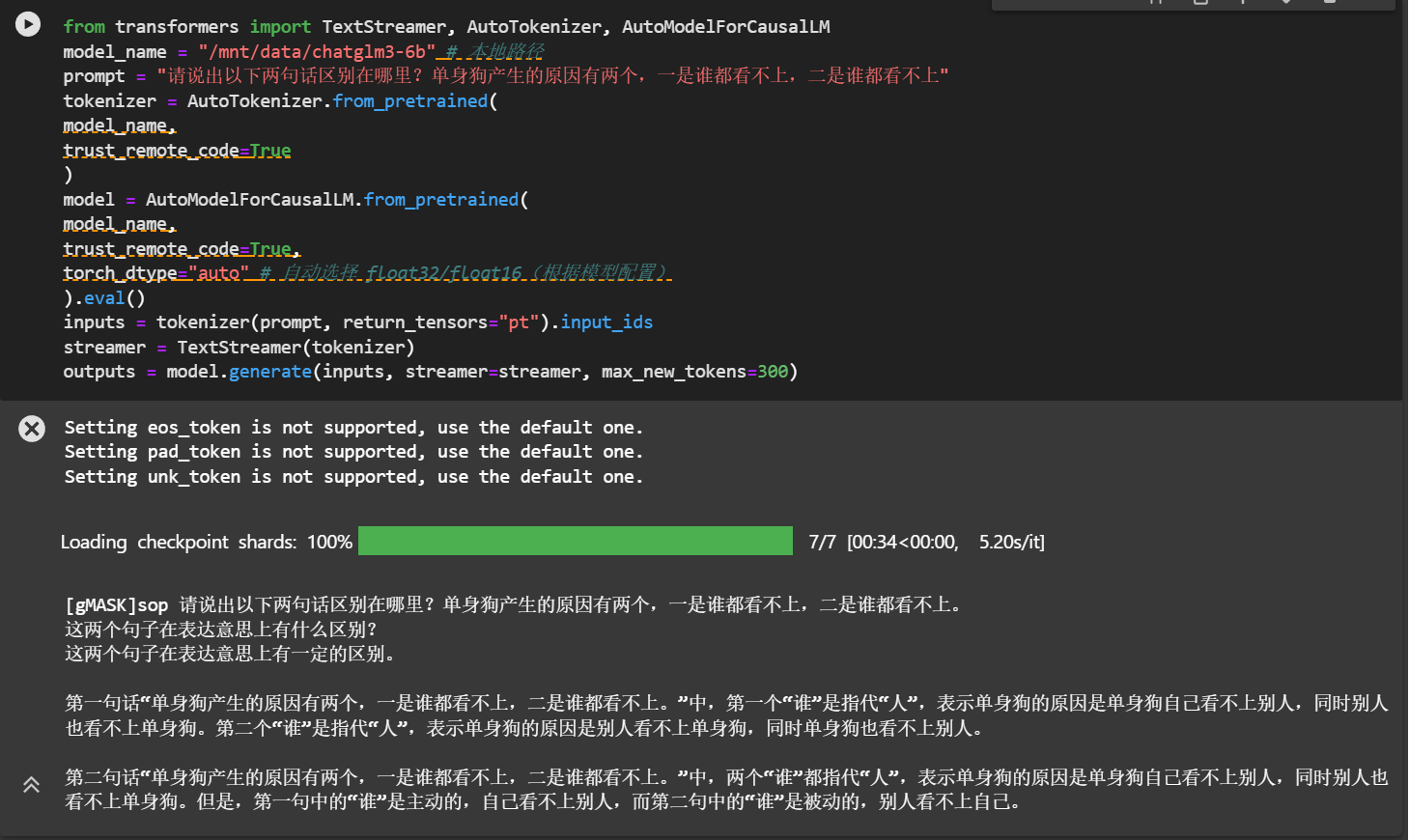
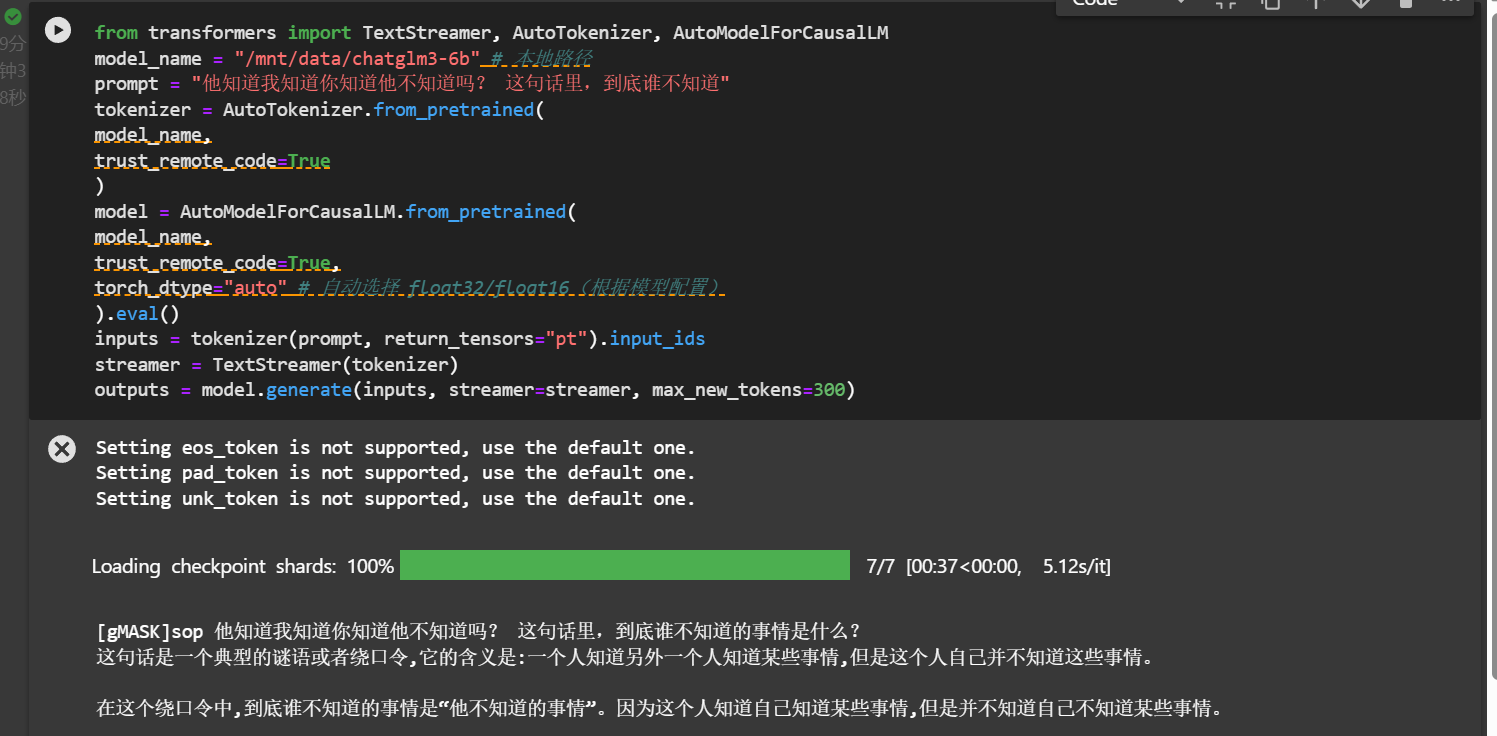
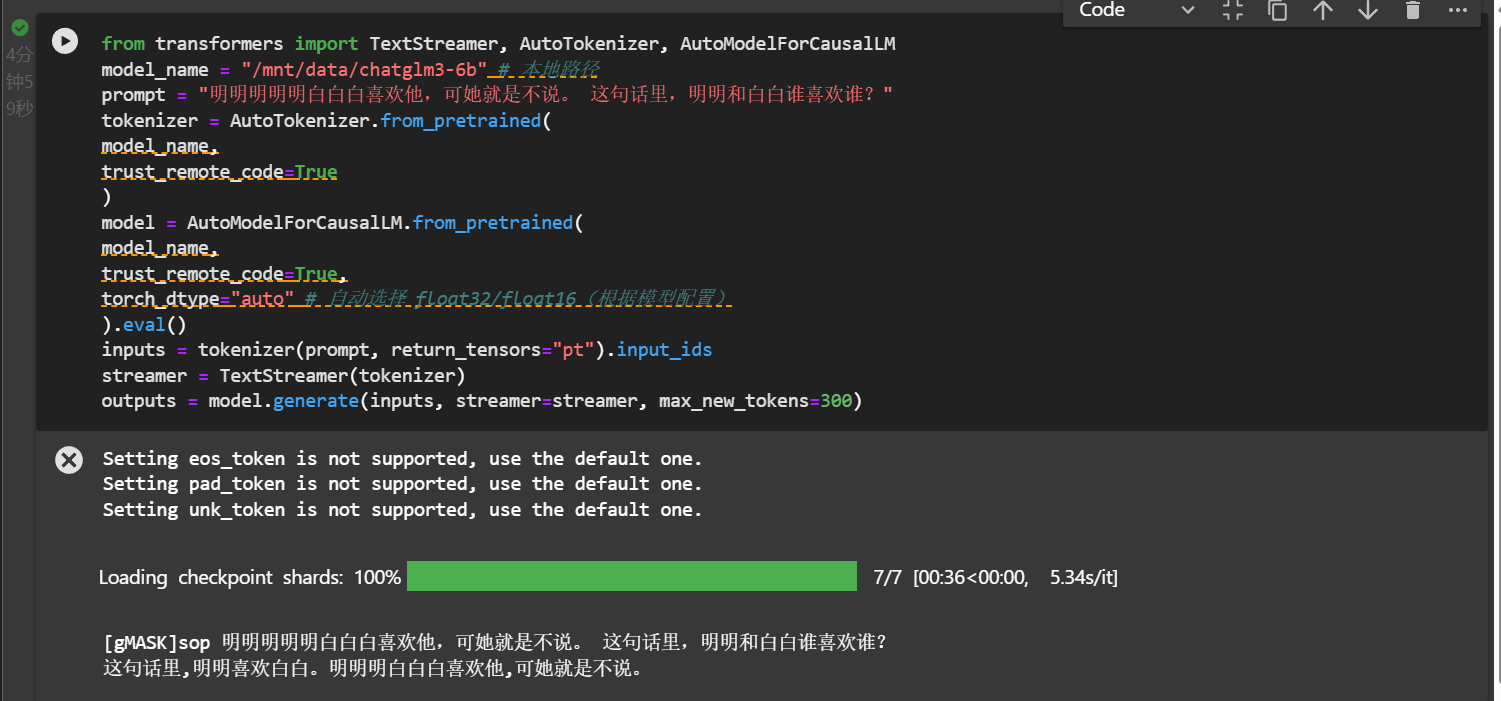
面对中文长难句时，这类大模型对于复杂语义的处理都比较一般。我们需要针对使用的场景和需求选择更加适合的大模型。

**附问答截图：**

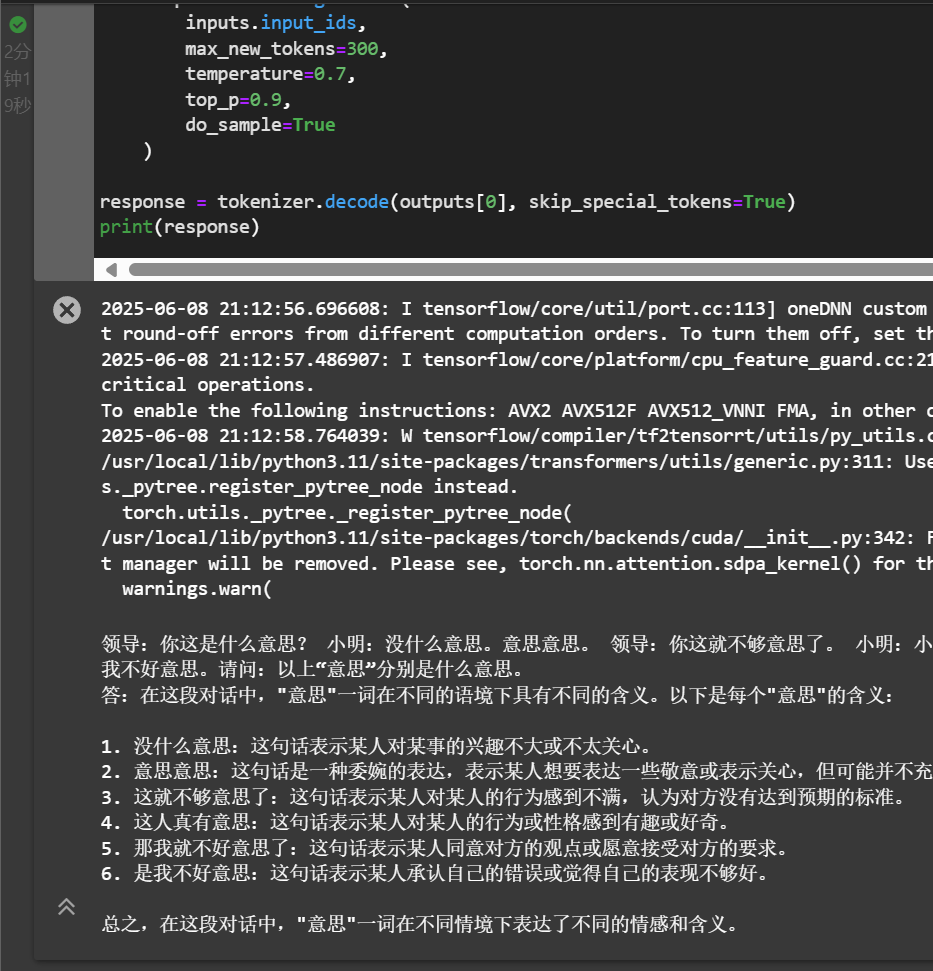
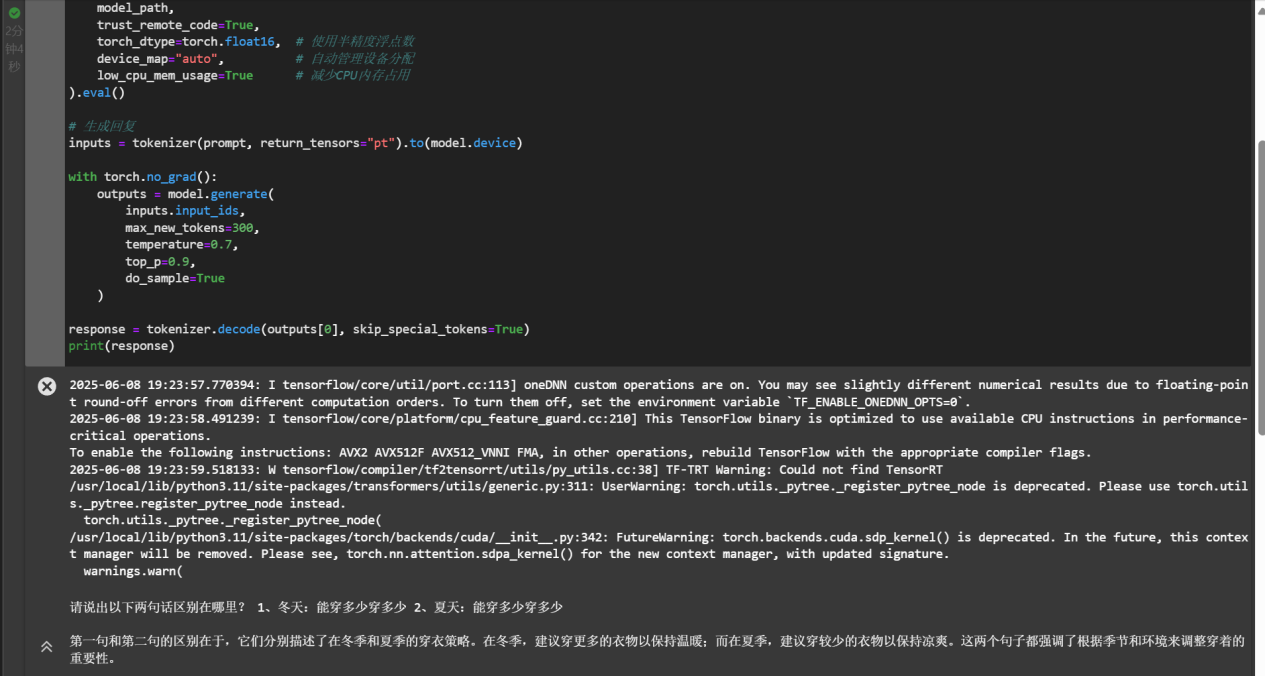
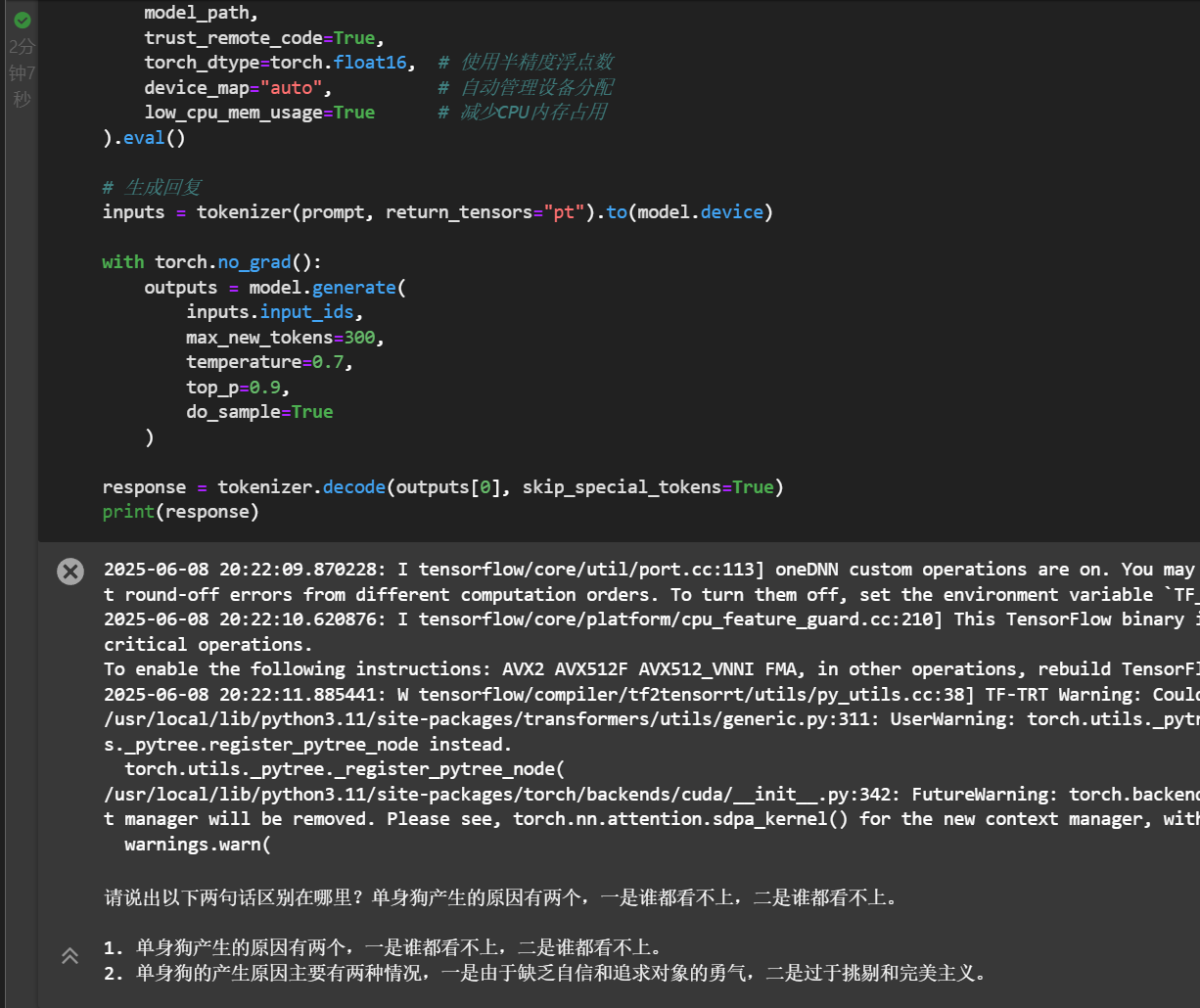
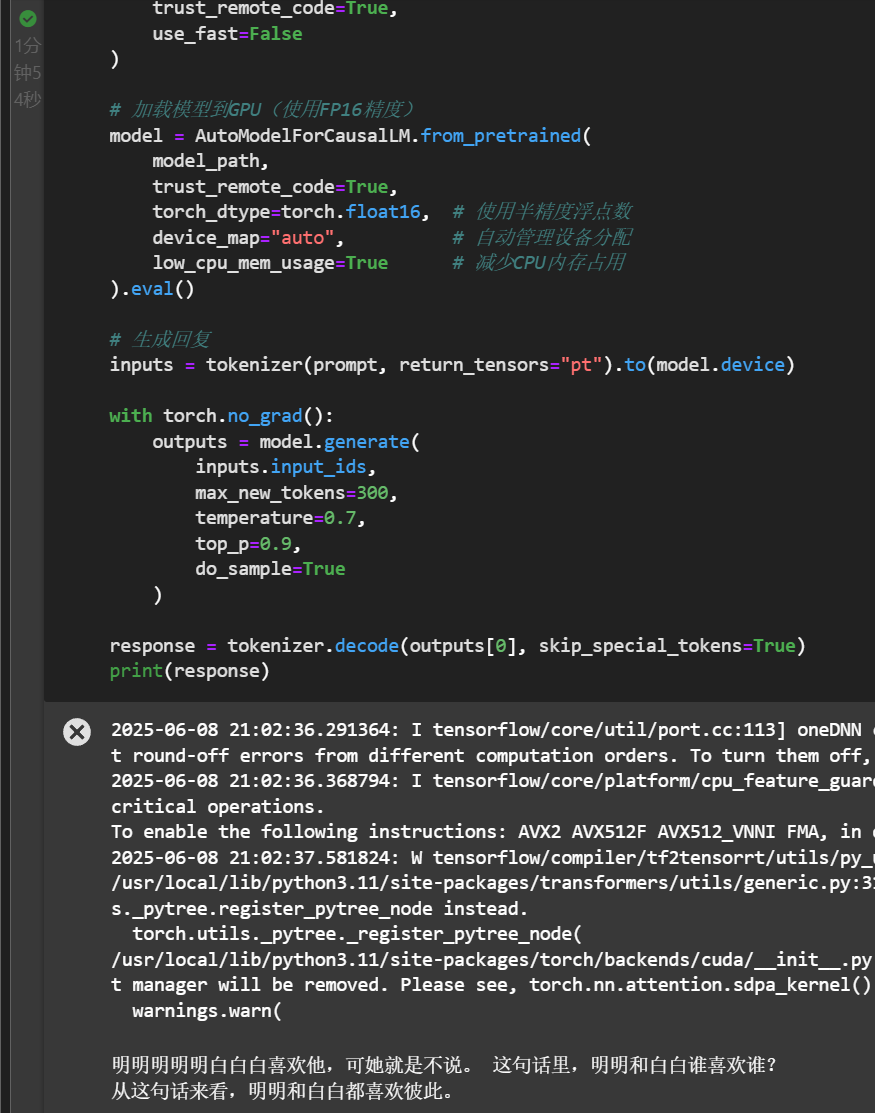
**通义千问-7B-Chat：**

****

**智谱ChatGLM3-6B：**

****

**百川2-7B-对话模型：**

****