人工智能导论

项目报告

AI 中文语义理解能力横向对比分析

姓　　名：王炯昭
学　　号：2353819
学　　院：计算机科学与技术学院
专　　业：软件工程
指导教师：汪昱

二〇二五年六月

一、项目简介

本项目的公开访问链接为：

https://github.com/WJiongzha0/Benchmarking-Chinese-Semantic-Understanding-in-Large-Language-Models

（一）项目流程

本项目的流程如下：

1. 登录并使用魔搭平台，关联阿里云账号来获得免费的 CPU 云计算资源；
2. 通过 Jupyter Notebook 进入相应的项目部署环境，完成模型的部署；
3. 针对 2 个不同的模型进行一些应用场景的测试，并开展不同模型之间的横向对比；

（二）模型选取

本项目选取以下两个模型作为对比，进行分析：

1. 通义千问 Qwen-7B-Chat；
2. 智普 ChatGLM3-6B；

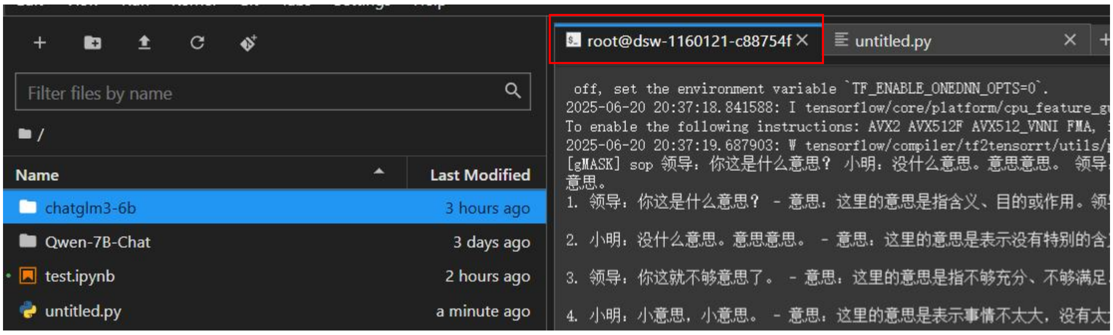
（三）测试问题

本项目共选取五个测试问题，分别如下：

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、配置流程

1. 在魔搭平台上启动 notebook 后，点击 Terminal，打开终端命令行环境；



```
 off, set the environment variable `TF_ENABLE_ONEDNN_OPTS=0`.
2025-06-20 20:37:18.841588: I tensorflow/core/platform/cpu_feature_g
To enable the following instructions: AVX2 AVX512F AVX512_VNNI FMA,
2025-06-20 20:37:19.687903: W tensorflow/compiler/tf2tensorrt/utils/
[gMASK] sop 领导：你这是什么意思？ 小明：没什么意思。意思意思。 领导
意思。
1. 领导：你这是什么意思？ - 意思：这里的意思是指含义、目的或作用。

2. 小明：没什么意思。意思意思。 - 意思：这里的意思是表示没有特别的

3. 领导：你这就不够意思了。 - 意思：这里的意思是指不够充分、不够满足

4. 小明：小意思，小意思。 - 意思：这里的意思是表示事情不太大，没有太
```

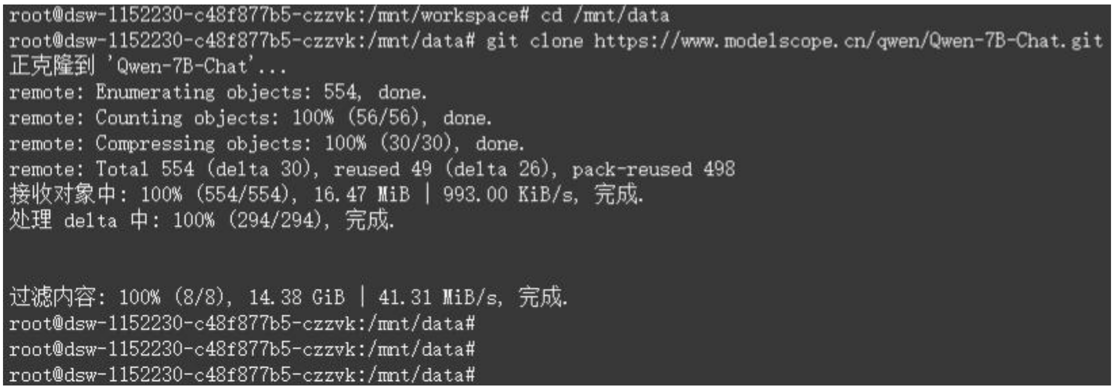
2. 这里采用 root 直接操作，输入以下代码安装基础环境和基础依赖：

```
pip install \
torch==2.3.0+cpu \
torchvision==0.18.0+cpu \
--index-url https://download.pytorch.org/whl/cpu
# 安装基础依赖（兼容 transformers 4.33.3 和 neuralchat）
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″
# 安装 fschat（需要启用 PEP517 构建）
pip install fschat --use-pep517
```

3.先切换到数据目录，之后分别输入以下代码，下载两个大模型到本地：

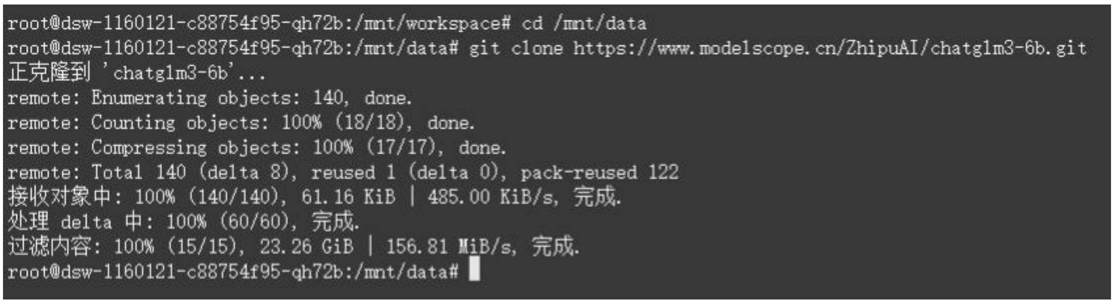
```
git clone https://www.modelscope.cn/ZhipuAI/chatglm3-6b.git
git clone https://www.modelscope.cn/qwen/Qwen-7B-Chat.git
```

下载模型 Qwen-7B-Chat:



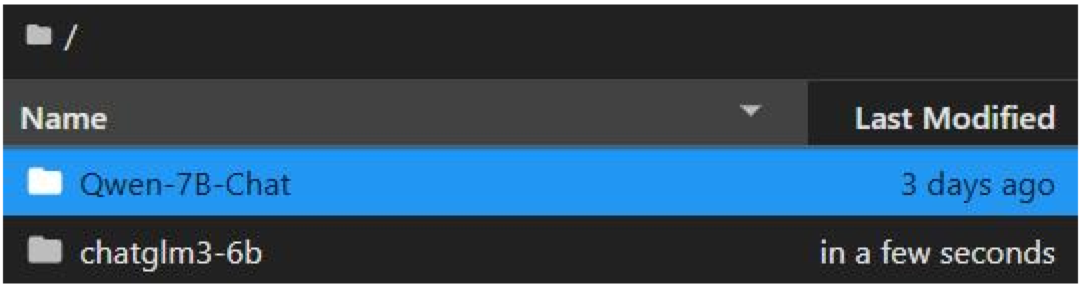
```
root@dsw-1152230-c48f877b5-czzvk:/mnt/workspace# cd /mnt/data
root@dsw-1152230-c48f877b5-czzvk:/mnt/data# git clone https://www.modelscope.cn/qwen/Qwen-7B-Chat.git
正克隆到 'Qwen-7B-Chat'...
remote: Enumerating objects: 554, done.
remote: Counting objects: 100% (56/56), done.
remote: Compressing objects: 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 | 993.00 KiB/s, 完成.
处理 delta 中: 100% (294/294), 完成.


过滤内容: 100% (8/8), 14.38 GiB | 41.31 MiB/s, 完成.
root@dsw-1152230-c48f877b5-czzvk:/mnt/data#
root@dsw-1152230-c48f877b5-czzvk:/mnt/data#
root@dsw-1152230-c48f877b5-czzvk:/mnt/data#
```

下载模型 chatglm3-6b:



```
root@dsw-1160121-c88754f95-qh72b:/mnt/workspace# cd /mnt/data
root@dsw-1160121-c88754f95-qh72b:/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 | 485.00 KiB/s, 完成.
处理 delta 中: 100% (60/60), 完成.
过滤内容: 100% (15/15), 23.26 GiB | 156.81 MiB/s, 完成.
root@dsw-1160121-c88754f95-qh72b:/mnt/data# █
```

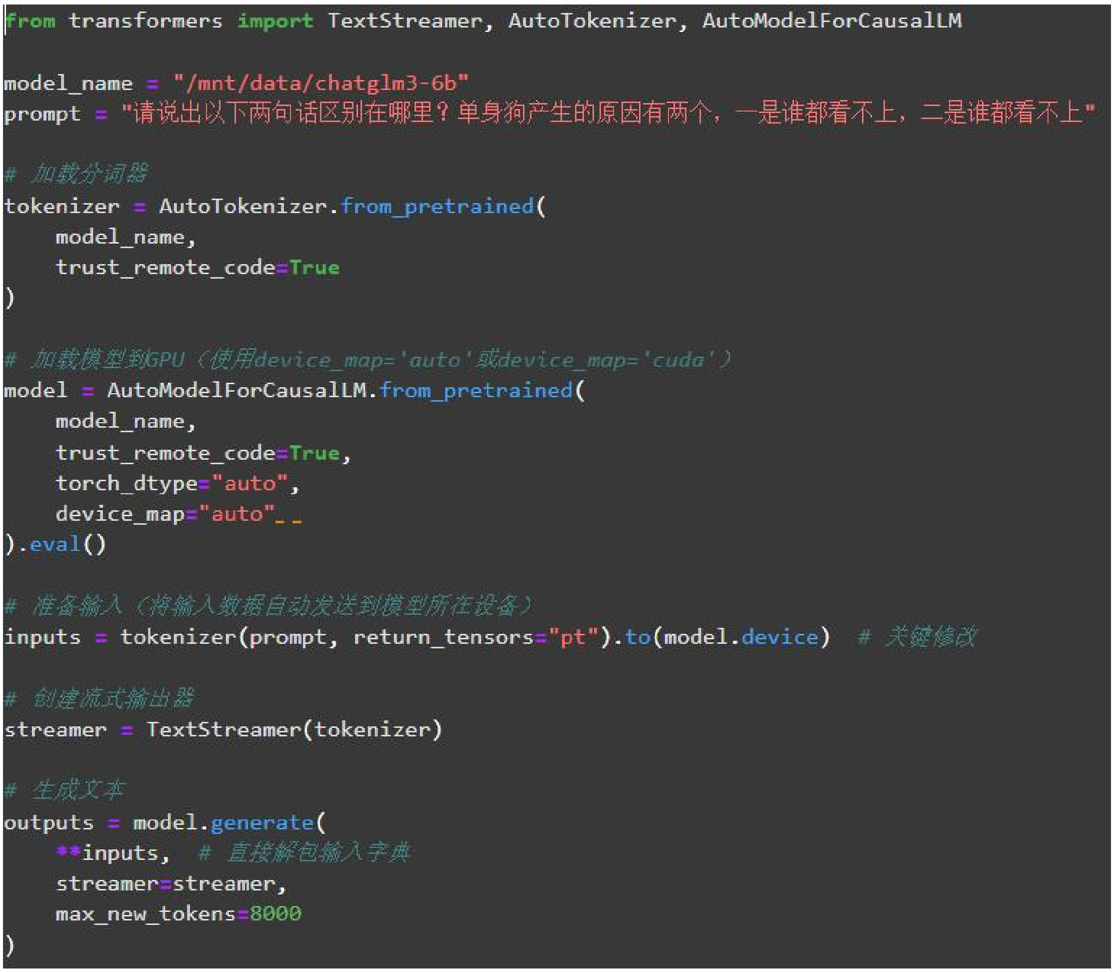
两个模型均下载完成：



/

Name	Last Modified
Qwen-7B-Chat	3 days ago
chatglm3-6b	in a few seconds

4. 编写 Python 脚本，准备运行：

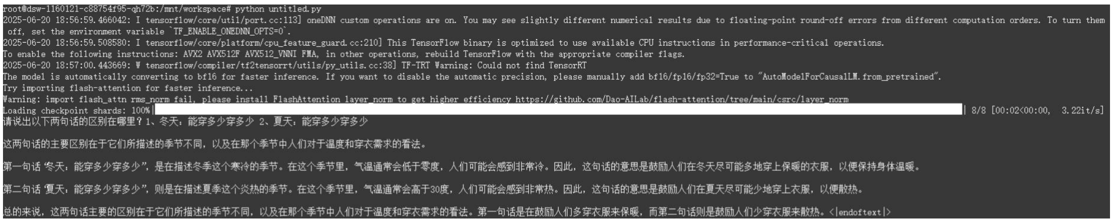


```
from transformers import TextStreamer, AutoTokenizer, AutoModelForCausalLM

model_name = "/mnt/data/chatglm3-6b"
prompt = "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"

# 加载分词器
tokenizer = AutoTokenizer.from_pretrained(
    model_name,
    trust_remote_code=True
)

# 加载模型到GPU（使用device_map='auto'或device_map='cuda'）
model = AutoModelForCausalLM.from_pretrained(
    model_name,
    trust_remote_code=True,
    torch_dtype="auto",
    device_map="auto"
).eval()

# 准备输入（将输入数据自动发送到模型所在设备）
inputs = tokenizer(prompt, return_tensors="pt").to(model.device)  # 关键修改

# 创建流式输出器
streamer = TextStreamer(tokenizer)

# 生成文本
outputs = model.generate(
    **inputs,  # 直接解包输入字典
    streamer=streamer,
    max_new_tokens=8000
)
```

三、结果分析

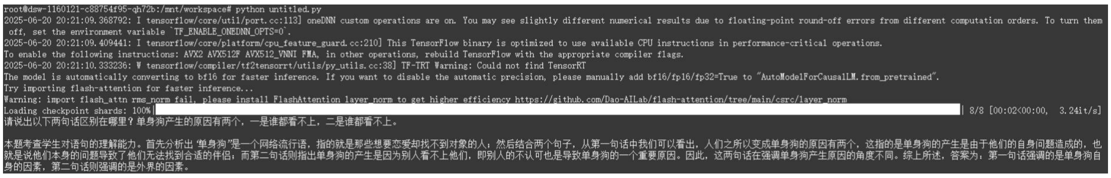
（一）Qwen-7B-Chat 回答

1. 问题 1：



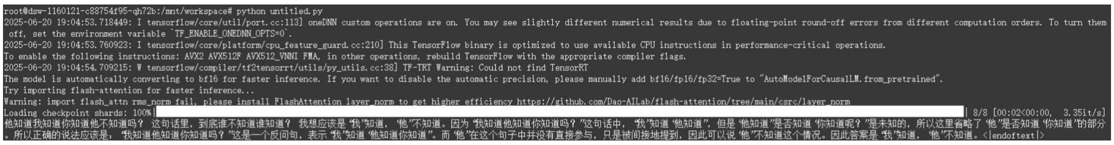
```
root@dsw-1160121-c88754f95-qh72b:/mnt/workspace# python untitled.py
2025-06-20 18:56:59.466042: I tensorflow/core/util/port.cc:113] oneDNN custom operations are on. You may see slightly different numerical results due to floating-point round-off errors from different computation orders. To turn them
 off, set the environment variable `TF_ENABLE_ONEDNN_OPTS=0`.
2025-06-20 18:56:59.508580: I tensorflow/core/platform/cpu_feature_guard.cc:210] This TensorFlow binary is optimized to use available CPU instructions in performance-critical operations.
To enable the following instructions: AVX2 AVX512F AVX512_VNNI FMA, in other operations, rebuild TensorFlow with the appropriate compiler flags.
2025-06-20 18:57:00.443669: W tensorflow/compiler/tf2tensorrt/utils/py_utils.cc:38] TF-TRT Warning: Could not find TensorRT
The model is automatically converting to bf16 for faster inference. If you want to disable the automatic precision, please manually add bf16/fp16/fp32=True to "AutoModelForCausalLM.from_pretrained".
Try importing flash-attention for faster inference...
Warning: import flash_attn rms_norm fail, please install FlashAttention layer_norm to get higher efficiency https://github.com/Dao-AILab/flash-attention/tree/main/csrc/layer_norm
Loading checkpoint shards: 100%|████████████████████████████████████████████████| 8/8 [00:02<00:00,  3.22it/s]
请说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

这两句话的主要区别在于它们所描述的季节不同，以及在那个季节中人们对于温度和穿衣需求的看法。

第一句话"冬天：能穿多少穿多少"，是在描述冬季这个寒冷的季节。在这个季节里，气温通常会低于零度，人们可能会感到非常冷。因此，这句话的意思是鼓励人们在冬天尽可能多地穿上保暖的衣服，以便保持身体温暖。

第二句话"夏天：能穿多少穿多少"，则是在描述夏季这个炎热的季节。在这个季节里，气温通常会高于30度，人们可能会感到非常热。因此，这句话的意思是鼓励人们在夏天尽可能少地穿上衣服，以便散热。

总的来说，这两句话主要的区别在于它们所描述的季节不同，以及在那个季节中人们对于温度和穿衣需求的看法。第一句话是在鼓励人们多穿衣服来保暖，而第二句话则是鼓励人们少穿衣服来散热。<|endoftext|>
```

2. 问题 2：



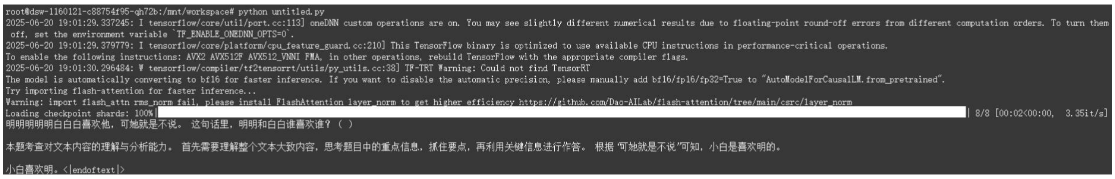
```
root@dsw-1160121-c88754f95-qh72b:/mnt/workspace# python untitled.py
2025-06-20 20:21:09.368792: I tensorflow/core/util/port.cc:113] oneDNN custom operations are on. You may see slightly different numerical results due to floating-point round-off errors from different computation orders. To turn them
 off, set the environment variable `TF_ENABLE_ONEDNN_OPTS=0`.
2025-06-20 20:21:09.409441: I tensorflow/core/platform/cpu_feature_guard.cc:210] This TensorFlow binary is optimized to use available CPU instructions in performance-critical operations.
To enable the following instructions: AVX2 AVX512F AVX512_VNNI FMA, in other operations, rebuild TensorFlow with the appropriate compiler flags.
2025-06-20 20:21:10.333236: W tensorflow/compiler/tf2tensorrt/utils/py_utils.cc:38] TF-TRT Warning: Could not find TensorRT
The model is automatically converting to bf16 for faster inference. If you want to disable the automatic precision, please manually add bf16/fp16/fp32=True to "AutoModelForCausalLM.from_pretrained".
Try importing flash-attention for faster inference...
Warning: import flash_attn rms_norm fail, please install FlashAttention layer_norm to get higher efficiency https://github.com/Dao-AILab/flash-attention/tree/main/csrc/layer_norm
Loading checkpoint shards: 100%|████████████████████████████████████████| 8/8 [00:02<00:00,  3.24it/s]
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。

本题考查学生对语句的理解能力。首先分析出“单身狗”是一个网络流行语，指的就是那些想要恋爱却找不到对象的人；然后结合两个句子，从第一句话中我们可以看出，人们之所以变成单身狗的原因有两个，这指的是单身狗的产生是由于他们的自身问题造成的，也
就是说他们本身的问题导致了他们无法找到合适的伴侣；而第二句话则指出单身狗的产生是因为别人看不上他们，即别人的不认可也是导致单身狗的一个重要原因。因此，这两句话在强调单身狗产生原因的角度不同。综上所述，答案为：第一句话强调的是单身狗自
身的因素，第二句话则强调的是外界的因素。
```

3. 问题 3：



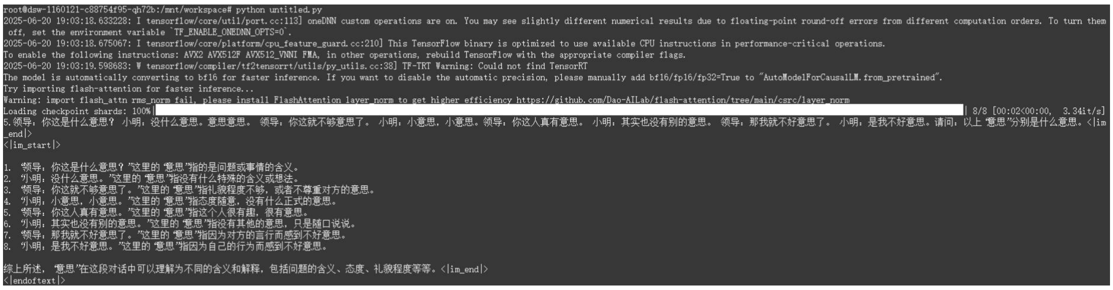
```
root@dsw-1160121-c88754f95-qh72b:/mnt/workspace# python untitled.py
2025-06-20 19:04:53.718449: I tensorflow/core/util/port.cc:113] oneDNN custom operations are on. You may see slightly different numerical results due to floating-point round-off errors from different computation orders. To turn them
 off, set the environment variable `TF_ENABLE_ONEDNN_OPTS=0`.
2025-06-20 19:04:53.760923: I tensorflow/core/platform/cpu_feature_guard.cc:210] This TensorFlow binary is optimized to use available CPU instructions in performance-critical operations.
To enable the following instructions: AVX2 AVX512F AVX512_VNNI FMA, in other operations, rebuild TensorFlow with the appropriate compiler flags.
2025-06-20 19:04:54.709215: W tensorflow/compiler/tf2tensorrt/utils/py_utils.cc:38] TF-TRT Warning: Could not find TensorRT
The model is automatically converting to bf16 for faster inference. If you want to disable the automatic precision, please manually add bf16/fp16/fp32=True to "AutoModelForCausalLM.from_pretrained".
Try importing flash-attention for faster inference...
Warning: import flash_attn rms_norm fail, please install FlashAttention layer_norm to get higher efficiency https://github.com/Dao-AILab/flash-attention/tree/main/csrc/layer_norm
Loading checkpoint shards: 100%|████████████████████████████████████████| 8/8 [00:02<00:00,  3.35it/s]
他知道我知道你知道他不知道吗？ 这句话里，到底谁不知道谁知道？ 我想应该是“我”知道，“他”不知道。因为“我知道他知道你知道吗？”这句话中，“我”知道“他知道”，但是“他知道”是否知道“你知道呢？”是未知的，所以这里省略了“他”是否知道“你知道”的部分
。所以正确的说法应该是：“我知道他知道你知道吗？”这是一个反问句，表示“我”知道“他知道你知道”。而“他”在这个句子中并没有直接参与，只是被间接地提到，因此可以说“他”不知道这个情况。因此答案是“我”知道，“他”不知道。<|endoftext|>
```

4. 问题 4：



```
root@dsw-1160121-c88754f95-qh72b:/mnt/workspace# python untitled.py
2025-06-20 19:01:29.337245: I tensorflow/core/util/port.cc:113] oneDNN custom operations are on. You may see slightly different numerical results due to floating-point round-off errors from different computation orders. To turn them
 off, set the environment variable `TF_ENABLE_ONEDNN_OPTS=0`.
2025-06-20 19:01:29.379779: I tensorflow/core/platform/cpu_feature_guard.cc:210] This TensorFlow binary is optimized to use available CPU instructions in performance-critical operations.
To enable the following instructions: AVX2 AVX512F AVX512_VNNI FMA, in other operations, rebuild TensorFlow with the appropriate compiler flags.
2025-06-20 19:01:30.296484: W tensorflow/compiler/tf2tensorrt/utils/py_utils.cc:38] TF-TRT Warning: Could not find TensorRT
The model is automatically converting to bf16 for faster inference. If you want to disable the automatic precision, please manually add bf16/fp16/fp32=True to "AutoModelForCausalLM.from_pretrained".
Try importing flash-attention for faster inference...
Warning: import flash_attn rms_norm fail, please install FlashAttention layer_norm to get higher efficiency https://github.com/Dao-AILab/flash-attention/tree/main/csrc/layer_norm
Loading checkpoint shards: 100%|████████████████████████████████████████| 8/8 [00:02<00:00,  3.35it/s]
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？（ ）

本题考查对文本内容的理解与分析能力。 首先需要理解整个文本大致内容，思考题目中的重点信息，抓住要点，再利用关键信息进行作答。 根据“可她就是不说”可知，小白是喜欢明的。

小白喜欢明。<|endoftext|>
```

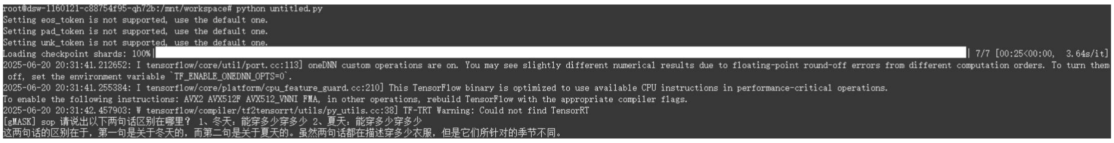
5. 问题 5：



```
root@dsw-1160121-c88754f95-qh72b:/mnt/workspace# python untitled.py
2025-06-20 19:03:18.633228: I tensorflow/core/util/port.cc:113] oneDNN custom operations are on. You may see slightly different numerical results due to floating-point round-off errors from different computation orders. To turn them
 off, set the environment variable `TF_ENABLE_ONEDNN_OPTS=0`.
2025-06-20 19:03:18.675067: I tensorflow/core/platform/cpu_feature_guard.cc:210] This TensorFlow binary is optimized to use available CPU instructions in performance-critical operations.
To enable the following instructions: AVX2 AVX512F AVX512_VNNI FMA, in other operations, rebuild TensorFlow with the appropriate compiler flags.
2025-06-20 19:03:19.598683: W tensorflow/compiler/tf2tensorrt/utils/py_utils.cc:38] TF-TRT Warning: Could not find TensorRT
The model is automatically converting to bf16 for faster inference. If you want to disable the automatic precision, please manually add bf16/fp16/fp32=True to "AutoModelForCausalLM.from_pretrained".
Try importing flash-attention for faster inference...
Warning: import flash_attn rms_norm fail, please install FlashAttention layer_norm to get higher efficiency https://github.com/Dao-AILab/flash-attention/tree/main/csrc/layer_norm
Loading checkpoint shards: 100%|████████████████████████████████████████| 8/8 [00:02<00:00,  3.34it/s]
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。<|im
_end|>
<|im_start|>

1.  “领导：你这是什么意思？”这里的“意思”指的是问题或事情的含义。
2.  “小明：没什么意思。”这里的“意思”指没有什么特殊的含义或想法。
3.  “领导：你这就不够意思了。”这里的“意思”指礼貌程度不够，或者不尊重对方的意思。
4.  “小明：小意思，小意思。”这里的“意思”指态度随意，没有什么正式的意思。
5.  “领导：你这人真有意思。”这里的“意思”指这个人很有趣，很有意思。
6.  “小明：其实也没有别的意思。”这里的“意思”指没有其他的意思，只是随口说说。
7.  “领导：那我就不好意思了。”这里的“意思”指因为对方的言行而感到不好意思。
8.  “小明：是我不好意思。”这里的“意思”指因为自己的行为而感到不好意思。

综上所述，“意思”在这段对话中可以理解为不同的含义和解释，包括问题的含义、态度、礼貌程度等等。<|im_end|>
<|endoftext|>
```

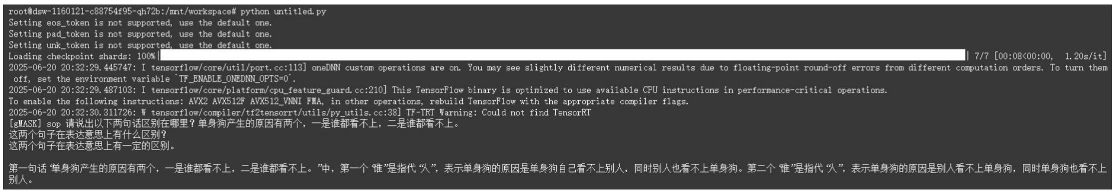
（二）Chatglm3-6B 回答

1. 问题 1：



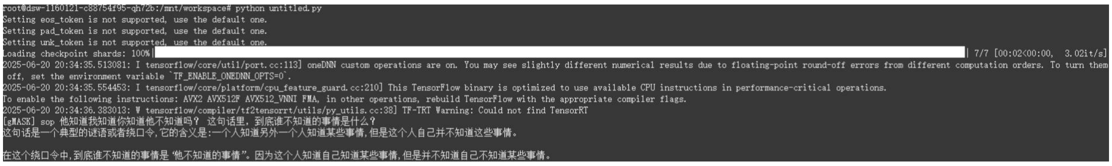
```
root@dsw-1160121-c88754f95-qh72b:/mnt/workspace# python untitled.py
Setting eos_token is not supported, use the default one.
Setting pad_token is not supported, use the default one.
Setting unk_token is not supported, use the default one.
Loading checkpoint shards: 100%|████████████████████████████████████████| 7/7 [00:25<00:00,  3.64s/it]
2025-06-20 20:31:41.212652: I tensorflow/core/util/port.cc:113] oneDNN custom operations are on. You may see slightly different numerical results due to floating-point round-off errors from different computation orders. To turn them
 off, set the environment variable `TF_ENABLE_ONEDNN_OPTS=0`.
2025-06-20 20:31:41.255384: I tensorflow/core/platform/cpu_feature_guard.cc:210] This TensorFlow binary is optimized to use available CPU instructions in performance-critical operations.
To enable the following instructions: AVX2 AVX512F AVX512_VNNI FMA, in other operations, rebuild TensorFlow with the appropriate compiler flags.
2025-06-20 20:31:42.457903: W tensorflow/compiler/tf2tensorrt/utils/py_utils.cc:38] TF-TRT Warning: Could not find TensorRT
[gMASK] sop 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，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。
```

2. 问题 2：



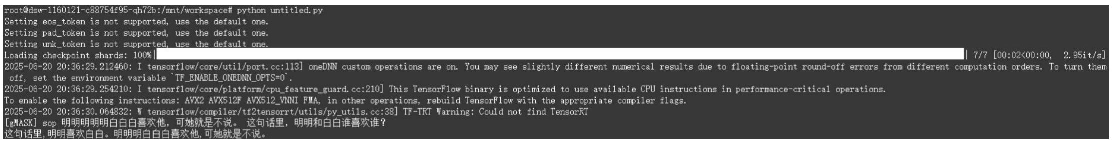
```
root@dsw-1160121-c88754f95-qh72b:/mnt/workspace# python untitled.py
Setting eos_token is not supported, use the default one.
Setting pad_token is not supported, use the default one.
Setting unk_token is not supported, use the default one.
Loading checkpoint shards: 100%|████████████████████████████████████████| 7/7 [00:08<00:00,  1.20s/it]
2025-06-20 20:32:29.445747: I tensorflow/core/util/port.cc:113] oneDNN custom operations are on. You may see slightly different numerical results due to floating-point round-off errors from different computation orders. To turn them
 off, set the environment variable `TF_ENABLE_ONEDNN_OPTS=0`.
2025-06-20 20:32:29.487103: I tensorflow/core/platform/cpu_feature_guard.cc:210] This TensorFlow binary is optimized to use available CPU instructions in performance-critical operations.
To enable the following instructions: AVX2 AVX512F AVX512_VNNI FMA, in other operations, rebuild TensorFlow with the appropriate compiler flags.
2025-06-20 20:32:30.311726: W tensorflow/compiler/tf2tensorrt/utils/py_utils.cc:38] TF-TRT Warning: Could not find TensorRT
[gMASK] sop 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。
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。
```

3. 问题 3：



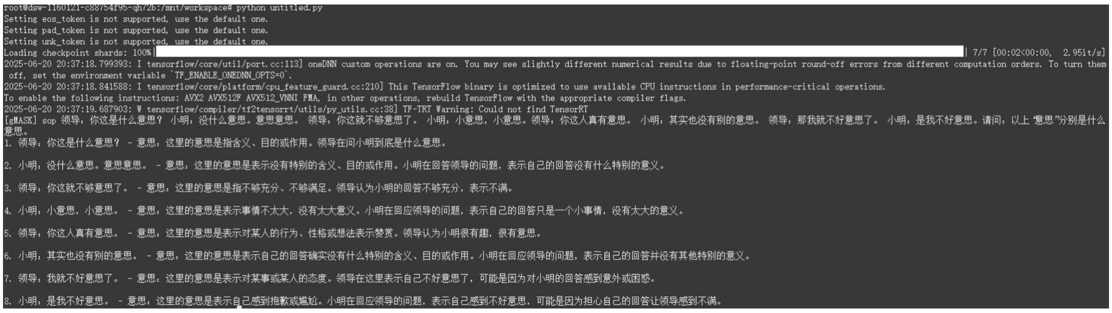
```
root@dsw-1160121-c88754f95-qh72b:/mnt/workspace# python untitled.py
Setting eos_token is not supported, use the default one.
Setting pad_token is not supported, use the default one.
Setting unk_token is not supported, use the default one.
Loading checkpoint shards: 100%|████████████████████████████████████████| 7/7 [00:02<00:00,  3.02it/s]
2025-06-20 20:34:35.513081: I tensorflow/core/util/port.cc:113] oneDNN custom operations are on. You may see slightly different numerical results due to floating-point round-off errors from different computation orders. To turn them
 off, set the environment variable `TF_ENABLE_ONEDNN_OPTS=0`.
2025-06-20 20:34:35.554453: I tensorflow/core/platform/cpu_feature_guard.cc:210] This TensorFlow binary is optimized to use available CPU instructions in performance-critical operations.
To enable the following instructions: AVX2 AVX512F AVX512_VNNI FMA, in other operations, rebuild TensorFlow with the appropriate compiler flags.
2025-06-20 20:34:36.383013: W tensorflow/compiler/tf2tensorrt/utils/py_utils.cc:38] TF-TRT Warning: Could not find TensorRT
[gMASK] sop 他知道我知道你知道他不知道吗？ 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谜语或者绕口令,它的含义是:一个人知道另外一个人知道某些事情,但是这个人自己并不知道这些事情。

在这个绕口令中,到底谁不知道的事情是“他不知道的事情”。因为这个人知道自己知道某些事情,但是并不知道自己不知道某些事情。
```

4. 问题 4：



```
root@dsw-1160121-c88754f95-qh72b:/mnt/workspace# python untitled.py
Setting eos_token is not supported, use the default one.
Setting pad_token is not supported, use the default one.
Setting unk_token is not supported, use the default one.
Loading checkpoint shards: 100%|████████████████████████████████████████████████████████████| 7/7 [00:02<00:00,  2.95it/s]
2025-06-20 20:36:29.212460: I tensorflow/core/util/port.cc:113] oneDNN custom operations are on. You may see slightly different numerical results due to floating-point round-off errors from different computation orders. To turn them
 off, set the environment variable `TF_ENABLE_ONEDNN_OPTS=0`.
2025-06-20 20:36:29.254210: I tensorflow/core/platform/cpu_feature_guard.cc:210] This TensorFlow binary is optimized to use available CPU instructions in performance-critical operations.
To enable the following instructions: AVX2 AVX512F AVX512_VNNI FMA, in other operations, rebuild TensorFlow with the appropriate compiler flags.
2025-06-20 20:36:30.064832: W tensorflow/compiler/tf2tensorrt/utils/py_utils.cc:38] TF-TRT Warning: Could not find TensorRT
[gMASK] sop 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？
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。
```

5. 问题 5：



```
root@dsw-1160121-c88754f95-qh72b:/mnt/workspace# python untitled.py
Setting eos_token is not supported, use the default one.
Setting pad_token is not supported, use the default one.
Setting unk_token is not supported, use the default one.
Loading checkpoint shards: 100%|████████████████████████████████████████████████████████████| 7/7 [00:02<00:00,  2.95it/s]
2025-06-20 20:37:18.799393: I tensorflow/core/util/port.cc:113] oneDNN custom operations are on. You may see slightly different numerical results due to floating-point round-off errors from different computation orders. To turn them
 off, set the environment variable `TF_ENABLE_ONEDNN_OPTS=0`.
2025-06-20 20:37:18.841588: I tensorflow/core/platform/cpu_feature_guard.cc:210] This TensorFlow binary is optimized to use available CPU instructions in performance-critical operations.
To enable the following instructions: AVX2 AVX512F AVX512_VNNI FMA, in other operations, rebuild TensorFlow with the appropriate compiler flags.
2025-06-20 20:37:19.687903: W tensorflow/compiler/tf2tensorrt/utils/py_utils.cc:38] TF-TRT Warning: Could not find TensorRT
[gMASK] sop 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。
1. 领导：你这是什么意思？ - 意思：这里的意思是指含义、目的或作用。领导在问小明到底是什么意思。

2. 小明：没什么意思。意思意思。 - 意思：这里的意思是表示没有特别的含义、目的或作用。小明在回答领导的问题，表示自己的回答没有什么特别的意义。

3. 领导：你这就不够意思了。 - 意思：这里的意思是指不够充分、不够满足。领导认为小明的回答不够充分，表示不满。

4. 小明：小意思，小意思。 - 意思：这里的意思是表示事情不太大，没有太大意义。小明在回应领导的问题，表示自己的回答只是一个小事情，没有太大的意义。

5. 领导：你这人真有意思。 - 意思：这里的意思是表示对某人的行为、性格或想法表示赞赏。领导认为小明很有趣，很有意思。

6. 小明：其实也没有别的意思。 - 意思：这里的意思是表示自己的回答确实没有什么特别的含义、目的或作用。小明在回应领导的问题，表示自己的回答并没有其他特别的意义。

7. 领导：我就不好意思了。 - 意思：这里的意思是表示对某事或某人的态度。领导在这里表示自己不好意思了，可能是因为对小明的回答感到意外或困惑。

8. 小明：是我不好意思。 - 意思：这里的意思是表示自己感到抱歉或尴尬。小明在回应领导的问题，表示自己感到不好意思，可能是因为担心自己的回答让领导感到不满。
```

（三）横向对比分析

对于问题 1，qwen 可以完全理解题目的意思并给出正确的回答，但 glm 给出的回答中关于区别集中在季节而非穿衣，且比较简短，似乎没有理解题目的意思；

对于问题 2，qwen 可以完全理解题意并给出正确的回答，指出题目中“单身狗的原因”是其本身和其他人两方面的原因，而从 glm 的回答中不难看出，其理解的题目的意思，但在解答过程中出现了问题，其给出两方面的原因相同，且均包含了正确的两方面原因，推测是 glm 在回答时组织语言出现了问题；

对于问题 3，qwen 很好的理解了题意，给出了正确的回答，并且详细的给出了推理分析的过程，而 glm 没有体会到题目的要求，并且分析过程出现了逻辑混乱的现象，推理分析能力不足；

对于问题 4，qwen 和 glm 都理解了题意。在答案上，qwen 给出了正确的答案但是人名出现了小错误，glm 给出的答案错误且分析过程体现出混乱的逻辑；

对于问题 5，qwen 和 glm 都理解了题意，且都选择正确的分条作答的形式。在答案上，qwen 和 glm 分析的正确率都达到了 100%，均体现出良好的一词多义在具体语境中的强大理解能力。

综合来讲，qwen 模型对于中文语义的理解能力较强，能够良好的理解题目要求，并进行合理正确的推理分析，给出完整正确的答案。对于一词在具体语境中的多义，一句在不同语气下的多义，连续多个重复词不同词性的分割与理解，复杂语境等现象，qwen 模型均表现良好。

而对于 glm 模型，其中文语义的理解能力一般，甚至无法正确理解题目要求，在推理分析的过程中无法把握合理方向，且表现出逻辑混乱，语言组织能力弱等现象。对于一词在具体语境中的多义，glm 模型表现良好；而对于一句在不同语气下的多义，连续多个重复词不同词性的分割与理解，复杂语境等现象，glm 模型均表现一般甚至较差，无法理解题目要求，也无法通过正确的推理分析过程得到正确的答案。