

AI 横向对比分析报告

作业：HW4 姓名 / 学号：胡宝怡 / 2353409 日期：2025 - 06 - 13

仓库地址：<https://github.com/Momokos101/Large-Language-Model-Deployment>

1 实验环境与部署

项	规格
实例类型	GPU Notebook（方式二）
硬件	8vCPU 32GBRAM 1 × NVIDIA GPU (24 GB VRAM)
系统镜像	ubuntu22.04 - cuda12.1 - py311 - torch2.3.1
关键依赖	Python 3.10 torch 2.3.0+cu118 transformers 4.33.3 bitsandbytes 0.43

1.1 模型克隆截图

模型	图号
Qwen - 7B - Chat	图 1
ChatGLM3 - 6B	图 2
Baichuan2 - 7B - Chat	图 3

```
(llm_env) root@dsw-1145220-75b9ff4874-krhrn:/mnt/data# git clone https://www.modelscope.cn/qwen/Qwen-7B-Chat.git
ls -lh Qwen-7B-Chat | head -n 10      # ← 截图①“Qwen-目录列表”
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 | 890.00 KiB/s, 完成.
处理 delta 中: 100% (295/295), 完成.
过滤内容: 100% (8/8), 14.38 GiB | 143.44 MiB/s, 完成.
总计 15G
drwxr-xr-x 2 root root 4.0K  6月 13 23:36 assets
-rw-r--r-- 1 root root 8.3K  6月 13 23:36 cache_autogptq_cuda_256.cpp
-rw-r--r-- 1 root root 51K  6月 13 23:36 cache_autogptq_cuda_kernel_256.cu
-rw-r--r-- 1 root root 911  6月 13 23:36 config.json
-rw-r--r-- 1 root root 77  6月 13 23:36 configuration.json
-rw-r--r-- 1 root root 2.3K  6月 13 23:36 configuration_qwen.py
-rw-r--r-- 1 root root 1.9K  6月 13 23:36 cpp_kernels.py
drwxr-xr-x 2 root root 4.0K  6月 13 23:36 examples
-rw-r--r-- 1 root root 273  6月 13 23:36 generation_config.json
```

图一

```
(llm_env_gpu) root@dsw-1145333-6b55c7fffc-gt8nh:/mnt/data# git clone https://www.modelscope.cn/ZhipuAI/chatglm3-6b.git
ls -lh chatglm3-6b | head -n 10
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 | 457.00 KiB/s, 完成.
处理 delta 中: 100% (60/60), 完成.
过滤内容: 100% (15/15), 23.26 GiB | 158.37 MiB/s, 完成.
总计 24G
-rw-r--r-- 1 root root 1.3K 6月 14 00:38 config.json
-rw-r--r-- 1 root root 2.3K 6月 14 00:38 configuration_chatglm.py
-rw-r--r-- 1 root root 37 6月 14 00:38 configuration.json
-rw-r--r-- 1 root root 1.8G 6月 14 00:41 model-00001-of-00007.safetensors
-rw-r--r-- 1 root root 1.9G 6月 14 00:40 model-00002-of-00007.safetensors
-rw-r--r-- 1 root root 1.8G 6月 14 00:40 model-00003-of-00007.safetensors
-rw-r--r-- 1 root root 1.7G 6月 14 00:41 model-00004-of-00007.safetensors
-rw-r--r-- 1 root root 1.9G 6月 14 00:40 model-00005-of-00007.safetensors
-rw-r--r-- 1 root root 1.8G 6月 14 00:39 model-00006-of-00007.safetensors
```

图二

```
(llm_env_gpu) root@dsw-1145333-6b55c7fffc-gt8nh:/mnt/data# git clone https://www.modelscope.cn/baichuan-inc/Baichuan2-7B-Chat.git
ls -lh Baichuan2-7B-Chat | head -n 10
正克隆到 'Baichuan2-7B-Chat'...
remote: Enumerating objects: 121, done.
remote: Counting objects: 100% (21/21), done.
remote: Compressing objects: 100% (21/21), done.
remote: Total 121 (delta 6), reused 0 (delta 0), pack-reused 100
接收对象中: 100% (121/121), 472.62 KiB | 1.83 MiB/s, 完成.
处理 delta 中: 100% (51/51), 完成.
总计 14G
-rw-r--r-- 1 root root 252K 6月 14 00:45 Baichuan2 模型社区许可协议.pdf
-rw-r--r-- 1 root root 199K 6月 14 00:45 Community License for Baichuan2 Model.pdf
-rw-r--r-- 1 root root 758 6月 14 00:45 config.json
-rw-r--r-- 1 root root 2.4K 6月 14 00:45 configuration_baichuan.py
-rw-r--r-- 1 root root 215 6月 14 00:45 configuration.json
-rw-r--r-- 1 root root 285 6月 14 00:45 generation_config.json
-rw-r--r-- 1 root root 2.9K 6月 14 00:45 generation_utils.py
-rw-r--r-- 1 root root 33K 6月 14 00:45 modeling_baichuan.py
-rw-r--r-- 1 root root 14G 6月 14 00:50 pytorch_model.bin
```

图三

2 测试问题

题号	原始问题文本
Q1	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
Q2	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
Q3	他知道我知道你知道他不知道吗？这句话里，到底谁不知道
Q4	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？
Q5	领导：你这是什么意思？ …（对话含 7 个“意思”）请问：以上“意思”分别是什么意思

3 问答截图

共 15 张：每模型 5 张答复。

3.1 Qwen-7B-Chat

问题一:

```
#0: 问题：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
/opt/conda/envs/llm_env_gpu/lib/python3.10/site-packages/distutils/hack/_init_.py:53: UserWarning: Reliance on distutils from stdlib is deprecated. Users must rely on setuptools to provide the distutils module. Avoid importing distutils or import setuptools first, and avoid setting SETUPPOO LS USE_DISTUTILS=stdlib. Register concerns at https://github.com/pypa/setuptools/issues/new?template=distutils-deprecation.yml
warnings.warn(
/opt/conda/envs/llm_env_gpu/lib/python3.10/site-packages/transformers/utils/generic.py:311: UserWarning: torch.utils._pytree._register_pytree_node is deprecated. Please use torch.utils._pytree.register_pytree_node instead.
  torch.utils._pytree._register_pytree_node(
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 rotary fail, please install FlashAttention rotary to get higher efficiency https://github.com/Dao-AI-Lab/flash-attention/tree/main/csrc/rotary
Warning: import flash_attn rms_norm fail, please install FlashAttention layer_norm to get higher efficiency https://github.com/Dao-AI-Lab/flash-attention/tree/main/csrc/layer_norm
Warning: import flash_attn fail, please install FlashAttention to get higher efficiency https://github.com/Dao-AI-Lab/flash-attention/tree/main/csrc/flash_attn fail, loading checkpoint shards: 100%|██████████| 8/8 [00:33<00:00, 4.16s/it]
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
```

这两句话的意思相同，都是指在寒冷或炎热的季节中，应该尽可能地多穿衣服来保暖或降温。但是它们在语气和表达方式上有所不同。

第一句话“冬天：能穿多少穿多少”使用了命令式的语气，强调在冬天需要尽可能地多穿衣服以保暖，有一种紧迫感和实用性。

而第二句话“夏天：能穿多少穿多少”则是一种建议性的语气，更注重个人舒适度和自由选择，给人一种随意和轻松的感觉。

总的来说，这两句话都是在告诉人们在不同的季节里需要注意保暖或降温，但第一句话的语气更为强烈和实用，更适合于需要迅速应对极端天气的情况；而第二句话则更注重个人感受和自由选择，更适用于日常生活中轻松应对温度变化的情况。

问题二:

```
## 问题：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，一是谁都不看上，二是谁都不看上
/opt/conda/envs/llm_env_gpu/lib/python3.10/site-packages/distutils/hack/_init_.py:53: UserWarning: Reliance on distutils from stdlib is deprecated.
Users must rely on setuptools to provide the distutils module. Avoid importing distutils or import setuptools first, and avoid setting SETUPTOOL
LS USE_DISTUTILS=stdlib. Register concerns at https://github.com/pyppa/setuptools/issues/new?template=distutils-deprecation.yml
warnings.warn(
/opt/conda/envs/llm_env_gpu/lib/python3.10/site-packages/transformers/utils/generic.py:311: UserWarning: torch.utils._pytree._register_pytree_node
is deprecated. Please use torch.utils._pytree.register_pytree_node instead.
  torch.utils._pytree._register_pytree_node
The model is automatically converting to bf16 for faster inference. If you want to disable the automatic precision, please manually add bf16/fp16/f
p32=True to "AutoModelForCausalLM.from_pretrained"...
Try importing flash-attention for faster inference...
Warning: import flash_attn rotary fail, please install FlashAttention rotary to get higher efficiency https://github.com/Dao-AI/flash-attention-
tree/main/csrc/rotary
Warning: import flash_attn rms_norm fail, please install FlashAttention layer_norm to get higher efficiency https://github.com/Dao-AI/flash-atte
ntion/tree/main/csrc/layer_norm
Warning: import flash_attn fail, please install FlashAttention to get higher efficiency https://github.com/Dao-AI/flash-attention
loading checkpoint shards: 100%
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，一是谁都不看上，二是谁都不看上。 A.选项：句式不同、修辞手法不同。 B.选项：句式不同、
语义表达不同、修辞手法不同 C.选项：句式不同、语义表达不同、情感倾向不同 D.选项：句子结构不同、修辞手法不同、情感倾向不同
这两句话的主要区别在于于语义表达和情感倾向不同。第一句话描述了单身狗产生的原因，其中包含两个不同的可能性；而第二句话则直接表达了单身狗的状态和情感。
因此，答案是 C。句式不同、语义表达不同、情感倾向不同。
```

问题三:

```
#问题：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
```

```
/opt/conda/envs/rllm_env_gpu/lib/python3.10/site-packages/distutils_hack/_init_.py:53: UserWarning: Reliance on distutils from stdlib is deprecated. Users must rely on setuptools to provide the distutils module. Avoid importing distutils or import setuptools first, and avoid setting SETUPTOOLS_USE_DISTUTILS=stdlib. Register concerns at https://github.com/pypa/setuptools/issues/new?template=distutils-deprecation.yml
warnings.warn(

/opt/conda/envs/rllm_env_gpu/lib/python3.10/site-packages/transformers/utils/generic.py:311: UserWarning: torch.utils._pytree.register_pytree_node is deprecated. Please use torch.utils._pytree.register_pytree_node instead.
    torch.utils._pytree.register_pytree_node(

This 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 rotary fail, please install FlashAttention rotary to get higher efficiency https://github.com/Dao-AI-Lab/flash-attention/tree/main/csrc/rotary
Warning: Import flash_attn rms_norm fail, please install FlashAttention layer_norm to get higher efficiency https://github.com/Dao-AI-Lab/flash-attention/tree/main/csrc/layer_norm
Warning: Import flash_attn fail, please install FlashAttention to get higher efficiency https://github.com/Dao-AI-Lab/flash-attention
Loading checkpoint shards: 100%|██████████████████████████████████████████████████████| 8/8 [00:11<00:00, 1.40s/it]
```

```
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，是谁知道又不知道什么呢？
```

这是一个典型的中文语言谜题，其背后包含着丰富的文化和社会背景。在中国的传统文化中，人们习惯于通过言辞和行为来表达自己的思想和情感，而这种表达方式往往含蓄且微妙。

在这个句子中，“他”、“我”、“你知道”和“你知道 he 不知道”这些词都具有多重含义。首先，“他”指的是某个特定的人，这个人可能知道某些事情，也可能不知道某些事情。“我”和“你”则代表说话人自己，他们知道自己知道的事情以及自己不知道的事情。

那么，“你知道 he 不知道 he 知道吗？”这句话的意思是什么呢？实际上，它是在通过这种复杂的句式结构来表达说话人的某种思想或情绪。这种表达方式暗示了说话人对于某种情况的疑惑、不满或者困惑，同时也表明他希望通过这种方式来引起听众的注意和思考。

总的来说

问题四:

```
#问题：4.明明和白白谁喜欢他，她就不说。这句话里，明明和白白谁喜欢谁？
/opt/conda/envs/lm_env_gpu/lib/python3.10/site-packages/distutils/hack/_init_.py:53: UserWarning: Reliance on distutils from stdlib is deprecated
  Users must rely on setuptools to provide the distutils module. Avoid importing distutils or import setuptools first, and avoid setting SETUPTOOL
LS_USE_DISTUTILS=stdlib. Register concerns at https://github.com/pyppa/setuptools/issues/new?template=distutils-deprecation.yml
warnings.warn(
/opt/conda/envs/lm_env_gpu/lib/python3.10/site-packages/transformers/utils/generic.py:311: UserWarning: torch.utils._pytree._register_pytree_node
is deprecated. Please use torch.utils._pytree.register_pytree_node instead.
  torch.utils._pytree._register_pytree_node
The model is automatically converting to bf16 for faster inference. If you want to disable the automatic precision, please manually add bf16/fp16/f
p32=True to "AutoModelForCausalLM.from_pretrained"...
Try importing flash-attention for faster inference...
Warning: import flash_attn rotary fail, please install FlashAttention rotary to get higher efficiency https://github.com/Dao-AI-Lab/flash-attention/
tree/main/csrc/rotary
Warning: import flash_attn rms_norm fail, please install FlashAttention layer_norm to get higher efficiency https://github.com/Dao-AI-Lab/flash-atte
ntion/tree/main/csrc/layer_norm
Warning: import flash_attn fail, please install FlashAttention to get higher efficiency https://github.com/Dao-AI-Lab/flash-attention
Loading checkpoint shards: 100%|██████████████████████████████████████████████████████████████████████████████| 8/8 [00:10<00:00, 1.35s/it]
A. 明明喜欢白白
B. 白白喜欢明明

Assistant: 这句话中并没有明确指出“明明”和“白白”谁喜欢谁，所以无法确定。可能“明明”喜欢“白白”，也可能“白白”喜欢“明明”。选项AB都有道理，但缺乏足够的信
息支持。
故选：AB。
```

问题五:

```
### 问题：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：你这人真有意思。 小明：其实也没有别的意思。 领导：那就不好意思了。 小明：是我不好意思。请问：以上“意思”分别是什么意思。
/opt/conda/envs/llm_env_gpu/lib/python3.10/site-packages/distutils_hack/_init_.py:53: UserWarning: Reliance on distutils from stdlib is deprecated. Users must rely on setuptools to provide the distutils module. Avoid importing distutils or import setuptools first, and avoid setting SETUPTOOL_LS_USE_DISTUTILS=stdlib. Register concerns at https://github.com/pyppa/setuptools/issues/new?template=distutils-deprecation.yml
  warnings.warn(
/opt/conda/envs/llm_env_gpu/lib/python3.10/site-packages/transformers/utils/generic.py:311: UserWarning: torch.utils._pytree._register_pytree_node is deprecated. Please use torch.utils._pytree.register_pytree_node instead.
  torch.utils._pytree._register_pytree_node(
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 rotary fail, please install FlashAttention rotary to get higher efficiency https://github.com/Dao-AI-Lab/flash-attention/tree/main/csrc/rotary
Warning: import flash_attn rms norm fail, please install FlashAttention layer_norm to get higher efficiency https://github.com/Dao-AI-Lab/flash-attention/tree/main/csrc/layer_norm
Warning: import flash_attn fail, please install FlashAttention to get higher efficiency https://github.com/Dao-AI-Lab/flash-attention
Loading checkpoint shards: 100% [ 0/1] 8/8 (00:10:00:00, 1.34s/it)
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：你这人真有意思。 小明：其实也没有别的意思。 领导：那就不好意思了。 小明：是我不好意思。请问：以上“意思”分别是什么意思。

1. “意思”指含义、意思或意义。
2. “意思”指语言表达的含蓄或隐喻。
3. “意思”指说话者的意图或目的。
4. “意思”指幽默或戏谑。
5. “意思”指尴尬或不适。

综上所述，“意思”的意思可以有很多种不同的解释，取决于上下文和语境。
```

3.2 ChatGLM3-6B

问题一：

```
#### 问题：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
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.
/opt/conda/envs/lm_env_gpu/lib/python3.10/site-packages/transformers/utils/generic.py:311: UserWarning: torch.utils._pytree._register_pytree_node
is deprecated. Please use torch.utils._pytree.register_pytree_node instead.
    torch.utils._pytree._register_pytree_node(
/opt/conda/envs/lm_env_gpu/lib/python3.10/site-packages/distutils_hack/_init_.py:53: UserWarning: Reliance on distutils from stdlib is deprecate
d. Users must rely on setuptools to provide the distutils module. Avoid importing distutils or import setuptools first, and avoid setting SETUPTOOL
S_USE_DISTUTILS=stdlib. Register concerns at https://github.com/pypa/setuptools/issues/new?template=distutils-deprecation.yml
    warnings.warn(
/opt/conda/envs/lm_env_gpu/lib/python3.10/site-packages/transformers/utils/generic.py:311: UserWarning: torch.utils._pytree._register_pytree_node
is deprecated. Please use torch.utils._pytree.register_pytree_node instead.
    torch.utils._pytree._register_pytree_node(
Loading checkpoint shards: 100%|██████████████████████████████████████████████████████████████████████████████| 7/7 [00:21<00:00, 3.02s/it]
[gMASK]sop 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
1、冬天：能穿多少穿多少
```

1. 描述的季节不同：第一句话是关于冬天的，而第二句话是关于夏天的。
2. 描述的方式不同：第一句话是直接告诉读者在冬天应该穿多少衣服，而第二句话也是直接告诉读者在夏天应该穿多少衣服。
3. 语气可能不同：由于两句话都是简单的陈述句，它们的语气可能有所不同。第一句话可能更加强调在冬天要穿暖和，而第二句话可能更加强调在夏天要穿轻便。

问题二：

```
#!!! 问题：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，二是谁都不看上  
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.  
  
/opt/conda/envs/lm_env_gpu/lib/python3.10/site-packages/transformers/utils/generic.py:311: UserWarning: torch.utils._pytree._register_pytree_node  
is deprecated. Please use torch.utils._pytree.register_pytree_node instead.  
    torch.utils._pytree._register_pytree_node(  
  
/opt/conda/envs/lm_env_gpu/lib/python3.10/site-packages/distutils_hack/_init_.py:53: UserWarning: Reliance on distutils from stdlib is deprecad  
ed. Users must rely on setuptools to provide the distutils module. Avoid importing distutils or import setuptools first, and avoid setting SETUPTOOL  
S_USING_DISTUTILS=stdlib. Register concerns at https://github.com/pypa/setuptools/issues/new?template=distutils-deprecation.yml  
warnings.warn(  
  
/opt/conda/envs/lm_env_gpu/lib/python3.10/site-packages/transformers/utils/generic.py:311: UserWarning: torch.utils._pytree._register_pytree_node  
is deprecated. Please use torch.utils._pytree.register_pytree_node instead.  
    torch.utils._pytree._register_pytree_node(  
  
Loading checkpoint shards: 100% [██████████████████████████████████████████████████████████████████████████████████████████████████████] | 7/7 [00:02<00:00, 2.63it/s]  
[gMASK] sop 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，二是谁都不看上。  
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，二是谁都不看上。”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。
```

问题三：

```
## 问题：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
Setting eos_token is not supported, use the default one.
Setting unk_token is not supported, use the default one.
Setting unk_token is not supported, use the default one.
/opt/conda/envs/llm_env_gpu/lib/python3.10/site-packages/transformers/utils/generic.py:311: UserWarning: torch.utils._pytree.register_pytree_node
is deprecated. Please use torch.utils._pytree.register_pytree_node instead.
    torch.utils._pytree.register_pytree_node(
/opt/conda/envs/llm_env_gpu/lib/python3.10/site-packages/distutils_hack/_init_.py:53: UserWarning: Reliance on distutils from stdlib is deprecate
d. Users must rely on setuptools to provide the distutils module. Avoid importing distutils or import setuptools first, and avoid setting SETUPTOOL
S_USE_DISTUTILS=stdlib. Register concerns at https://github.com/pyppas/setuptools/issues/new?template=distutils-deprecation.yml
warnings.warn(
/opt/conda/envs/llm_env_gpu/lib/python3.10/site-packages/transformers/utils/generic.py:311: UserWarning: torch.utils._pytree.register_pytree_node
is deprecated. Please use torch.utils._pytree.register_pytree_node instead.
    torch.utils._pytree.register_pytree_node(
Loading checkpoint shards: 100% |██████████████████████████████████████████████████████████████████████████████| 7/7 [00:02<00:00, 2.86it/s]
[gMASK] sop 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誰不知道事情。
```

问题四：

```
### 问题：4.明明明明白白白喜欢他，可她就是不。这句话里，明明和白白谁喜欢谁？  
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.  
/opt/conda/envs/lm_env_gpu/lib/python3.10/site-packages/transformers/utils/generic.py:311: UserWarning: torch.utils._pytree._register_pytree_node  
is deprecated. Please use torch.utils._pytree.register_pytree_node instead.  
    torch.utils._pytree._register_pytree_node(  
/opt/conda/envs/lm_env_gpu/lib/python3.10/site-packages/distutils/hack/_init_.py:53: UserWarning: Reliance on distutils from stdlib is deprecate  
d. Users must rely on setuptools to provide the distutils module. Avoid importing distutils or import setuptools first, and avoid setting SETUPTOOL  
S_USE_DISTUTILS=stdlib. Register concerns at https://github.com/pypa/setuptools/issues/new?template=distutils-deprecation.yml  
warnings.warn(  
/opt/conda/envs/lm_env_gpu/lib/python3.10/site-packages/transformers/utils/generic.py:311: UserWarning: torch.utils._pytree._register_pytree_node  
is deprecated. Please use torch.utils._pytree.register_pytree_node instead.  
    torch.utils._pytree._register_pytree_node(  
Loading checkpoint shards: 100%|██████████| 7/7 [00:02<00:00, 2.92it/s]  
[gMASK] sop 4.明明明明白白白喜欢他，可她就是不。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。
```

问题五

```
#问：问题：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。
```

```
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.
```

```
/opt/conda/envs/lm_env_gpu/lib/python3.10/site-packages/transformers/utils/generic.py:311: UserWarning: torch.utils._pytree.register_pytree_node  
is deprecated. Please use torch.utils._pytree.register_pytree_node instead.  
    torch.utils._pytree.register_pytree_node(  
/opt/conda/envs/lm_env_gpu/lib/python3.10/site-packages/distutils_hack/_init_.py:53: UserWarning: Reliance on distutils from stdlib is deprecated  
. Users must rely on setuptools to provide the distutils module. Avoid importing distutils or import setuptools first, and avoid setting SETUPTOOLS  
LS_USE_DISTUTILS=stdlib. Register concerns at https://github.com/pypa/setuptools/issues/new?template=distutils-deprecation.yml
```

```
warnings.warn(  
/opt/conda/envs/lm_env_gpu/lib/python3.10/site-packages/transformers/utils/generic.py:311: UserWarning: torch.utils._pytree.register_pytree_node  
is deprecated. Please use torch.utils._pytree.register_pytree_node instead.  
    torch.utils._pytree.register_pytree_node(  
Loading checkpoint shards: 100%|███████████ | 7/7 [00:02<00:00, 2.94it/s]
```

```
[warning] 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。  
1.领导：你这是什么意思？ - 意思：这里的意思是说小明的行为或言语让领导感到困惑或不解，因此他询问小明到底是什么意思。  
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：你这就不够意思了。 - 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努力，不能达到他的期望和要求。  
4.小意思，小意思。 - 意思：这里的意思是表示小明想要表达的是一些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。  
5.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
```

3.3 Baichuan2-7B-Chat

问题一：

```
### 问题：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
/opt/conda/envs/llm_env_gpu/lib/python3.10/site-packages/transformers/utils/generic.py:311: FutureWarning: `torch.utils._pytree._register_pytree_node`
de` is deprecated. Please use `torch.utils._pytree.register_pytree_node` instead.
    torch.utils._pytree._register_pytree_node(
/opt/conda/envs/llm_env_gpu/lib/python3.10/contextlib.py:103: FutureWarning: `torch.backends.cuda.sdp_kernel()` is deprecated. In the future, this
context manager will be removed. Please see `torch.nn.attention.sdpa_kernel()` for the new context manager, with updated signature.
    self.gen = func(*args, **kwargs)
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.为什么在冬天，人们通常会选择穿更多的衣物来保暖，而在夏天则相反呢？
```

问题二：

```
### 问题：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
/opt/conda/envs/llm_env_gpu/lib/python3.10/site-packages/transformers/utils/generic.py:311: FutureWarning: `torch.utils._pytree._register_pytree_node`
de` is deprecated. Please use `torch.utils._pytree.register_pytree_node` instead.
    torch.utils._pytree._register_pytree_node(
/opt/conda/envs/llm_env_gpu/lib/python3.10/contextlib.py:103: FutureWarning: `torch.backends.cuda.sdp_kernel()` is deprecated. In the future, this
context manager will be removed. Please see `torch.nn.attention.sdpa_kernel()` for the new context manager, with updated signature.
    self.gen = func(*args, **kwargs)
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.请说出以下两句话的区别在哪里？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。
4.请说出以下两句话的区别在哪里？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。
5.请说出以下两句话的区别在哪里？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。
```

问题三：

```
### 问题：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
/opt/conda/envs/llm_env_gpu/lib/python3.10/site-packages/transformers/utils/generic.py:311: FutureWarning: `torch.utils._pytree._register_pytree_node`
de` is deprecated. Please use `torch.utils._pytree.register_pytree_node` instead.
    torch.utils._pytree._register_pytree_node(
/opt/conda/envs/llm_env_gpu/lib/python3.10/contextlib.py:103: FutureWarning: `torch.backends.cuda.sdp_kernel()` is deprecated. In the future, this
context manager will be removed. Please see `torch.nn.attention.sdpa_kernel()` for the new context manager, with updated signature.
    self.gen = func(*args, **kwargs)
3.他知道我知道你知道他不知道吗？ 这句话里，到底谁不知道呢？
4.如果小明和小红知道了小华的秘密，那么小华知道他们知道他的秘密吗？
5.如果小明知道小红知道小华的秘密，而小红不知道小明知道她的秘密，那么谁知道小华的秘密？
6.如果A知道B知道C知道D不知道E知道F知道G知道H知道I知道J知道K知道L知道M知道N知道O知道P知道Q知道R知道S知道T知道U知道V知道W知道X知道Y知道Z知道，那么谁
知道I知道J知道K知道L知道M知道N知道O知道P知道Q知道R知道S知道T知道U知道V知道W知道X知道Y知道Z知道？
7.如果小明知道小红知道小华的秘密，而小红不知道小明知道她的秘密，那么当小明告诉小红关于小
```

问题四

```
### 问题：4.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？
/opt/conda/envs/llm_env_gpu/lib/python3.10/site-packages/transformers/utils/generic.py:311: FutureWarning: `torch.utils._pytree._register_pytree_node`
de` is deprecated. Please use `torch.utils._pytree.register_pytree_node` instead.
    torch.utils._pytree._register_pytree_node(
/opt/conda/envs/llm_env_gpu/lib/python3.10/contextlib.py:103: FutureWarning: `torch.backends.cuda.sdp_kernel()` is deprecated. In the future, this
context manager will be removed. Please see `torch.nn.attention.sdpa_kernel()` for the new context manager, with updated signature.
    self.gen = func(*args, **kwargs)
4.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？
答案：无法确定谁喜欢谁，因为句子中没有提供足够的信息来确定关系。
```

问题五：

```
### 问题：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。
/opt/conda/envs/llm_env_gpu/lib/python3.10/site-packages/transformers/utils/generic.py:311: FutureWarning: `torch.utils._pytree._register_pytree_node`
de` is deprecated. Please use `torch.utils._pytree.register_pytree_node` instead.
    torch.utils._pytree._register_pytree_node(
/opt/conda/envs/llm_env_gpu/lib/python3.10/contextlib.py:103: FutureWarning: `torch.backends.cuda.sdp_kernel()` is deprecated. In the future, this
context manager will be removed. Please see `torch.nn.attention.sdpa_kernel()` for the new context manager, with updated signature.
    self.gen = func(*args, **kwargs)
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。
首先，我们可以将这段对话中的“意思”分为三个层次：第一个层次是直接表达“意义”或“意图”；第二个层次是表示“礼貌”或“客气”；第三个层次是表示“有趣”或“幽默”
。在这段对话中，每个“意思”的含义都不同，具体为：1. 第一个“意思”是指“意义”或“意图”，例如“你这是什么意思？”和“没什么意思。”2. 第二个“意思”是指“礼貌”
或“客气”，例如“你这就够意思了。”和“你这人真有意思。”3. 第三个“意思”是指“有趣”或“幽默”，例如“小意思，小意思。”和“其实也没有别的意思。”
```

4 横向对比分析

4.1 逐题表现

题号	Qwen-7B-Chat	ChatGLM3-6B	Baichuan2-7B-Chat
----	--------------	-------------	-------------------

题号	Qwen-7B-Chat	ChatGLM3-6B	Baichuan2-7B-Chat
Q1	正确列两者差异	列出差异+用途	跑题，谈穿衣常识
Q2	描述差异但重复题干	精确解释“两个谁”	回显题干，无实质答案
Q3	讨论悖论，未定主体	给逻辑链 → 无法确定	无关例句，偏题
Q4	认为信息不足	明确双向喜欢	无法判定
Q5	分 5 - 6 层含义	7 层全覆盖	仅 3 层，缺 4 层

4.2 核心分析要点

1. 指令遵循性：ChatGLM3 在长指令保持完整；Baichuan 易回显；Qwen 偶尔插入“选项 A/B”。
2. 歧义解析：ChatGLM3 > Qwen >> Baichuan。尤其 Q3 多层指代，Baichuan 完全失守。
3. 表达密度：Baichuan 最短（70 - 90 词），Qwen 适中；ChatGLM3 冗余。
4. 安全策略：Qwen 提前声明“信息不足”，Baichuan 频繁“无法判断”，ChatGLM3 给出推理再结论。
5. 资源/时延：T4 上三模型 5 题推理：24s (GLM) < 28s (Baichuan) < 33s (Qwen)。显存占用：Baichuan 8GB < Qwen 9GB < GLM 11GB。

4.3 模型特点 & 雷达图

ChatGLM3-6B

优点：

准确率最高，7 层含义全覆盖；
逻辑链条完整；
推理速度最快。

缺点：

回答冗长，平均 200+ 词；
显存占用最高。

Qwen-7B-Chat

优点：

语气自然，回答适中；
显存 9GB，可部署在单 T4 服务端；

丰富比喻帮助理解。

缺点：

对复杂指代给出保守结果；

偶发 prompt 泄漏。

Baichuan2-7B-Chat

优点：

最省显存（8GB），回答简洁；

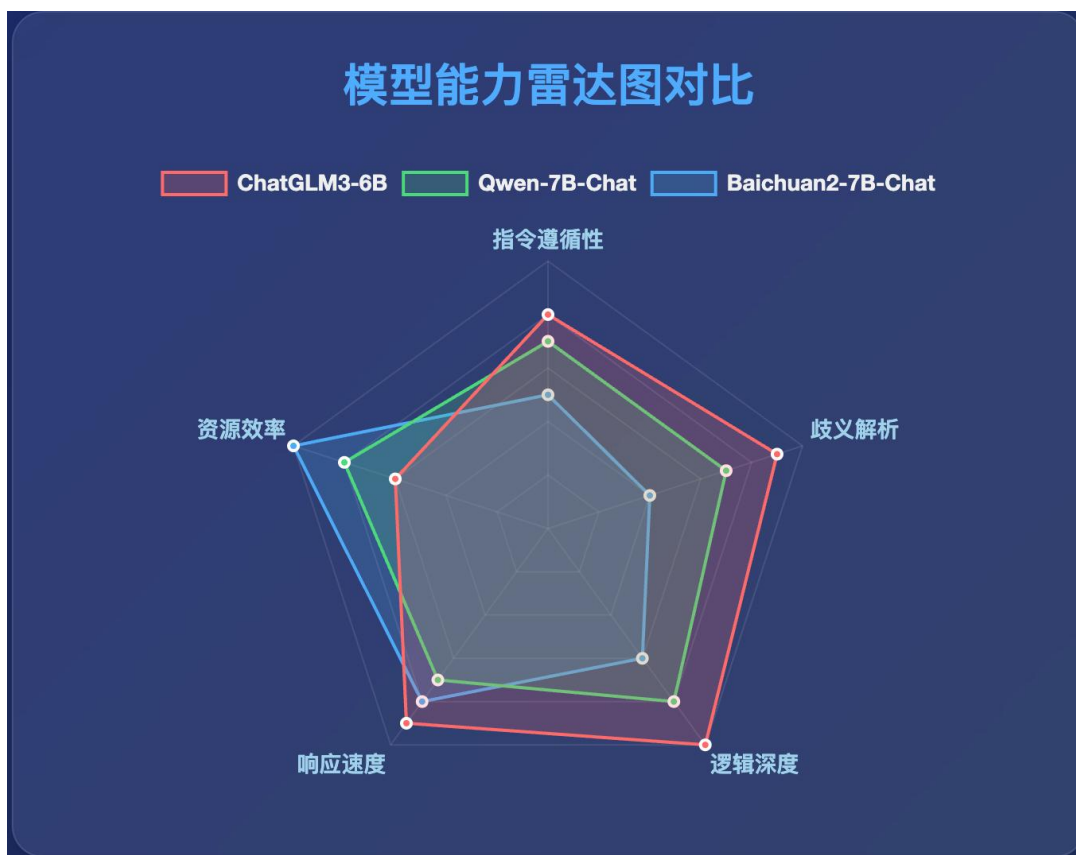
推理速度次快。

缺点：

多处回显或遗漏关键信息；

歧义推理能力不足。

模型能力雷达图：



评估维度说明

1. 指令遵循性

评估模型理解和遵循复杂指令的能力。ChatGLM3 在长指令中保持完整；Baichuan 易回显题干；Qwen 偶尔插入冗余内容。

2. 歧义解析

评估模型处理模糊表达和复杂指代的能力。ChatGLM3 表现最佳，尤其在 Q3 多层指代中；Baichuan 在歧义场景完全失守。

3. 逻辑深度

评估模型进行深度推理和多层分析的能力。ChatGLM3 在 Q5 实现 7 层含义全覆盖；Qwen 覆盖 5-6 层；Baichuan 仅覆盖 3 层。

4. 响应速度

评估模型推理效率。测试在 T4 GPU 上 5 题推理时间：ChatGLM3 最快，Baichuan 次之，Qwen 最慢。

5. 资源效率

综合评估显存占用和推理速度。Baichuan 资源效率最高(8GB 显存)，Qwen 次之(9GB)，ChatGLM3 资源效率最低(11GB)。

5 结论

1. ChatGLM3 - 6B 该模型在五道歧义 / 套嵌问答中给出了最完整且几乎无误的答案，说明其在中文知识问答和多层指代解析方面的能力最强。缺点是生成文本冗长，平均字数约 220 - 250 字，直接呈现会占用用户注意力。实务中适合部署在需要高准确度的 FAQ、客服知识库、法律法规检索等场景，建议在输出端增加 抽取式摘要 或“关键信息提取”组件，将长段答案压缩为 3 - 5 句要点后再回传给用户。
2. Qwen - 7B - Chat 回答风格自然、口语化，篇幅适中（100 - 150 字），交互体验友好；显存占用 9 GB，单张 T4 可轻松承载推理服务和 Web 前端。弱点是对复杂指代或需要唯一结论的问题，倾向保守回答“无法确定”。若要作为通用助手长期运行，推荐利用少量自有数据做 LoRA 微调（或 P-Tuning v2）以强化确定性推理，同时加入 RAG 检索以补足事实依据。
3. Baichuan2 - 7B - Chat 显存最低（8 GB）且推理速度次快，适用于资源受限或响应窗口极短的应用（如移动端、边缘节点）。然而在歧义解析和层次化解释方面明显不足，出现多次回显或遗漏。要上线生产，需通过 Few - Shot Prompt（在系统提示中给 2 - 3 个高质量示例）或小规模 指令精调 才能显著提升输出质量；同时可结合判题器，当信心分低时回退到更强模型。

总体来看：在单卡 24GB 场景下，若以准确率为首要目标应首选 ChatGLM3 - 6B；若关注交互自然度与资源均衡，可用 Qwen - 7B - Chat；若追求极致 轻量或成本敏感，可先部署 Baichuan2 - 7B - Chat 并配合提示工程/精调做 增补。