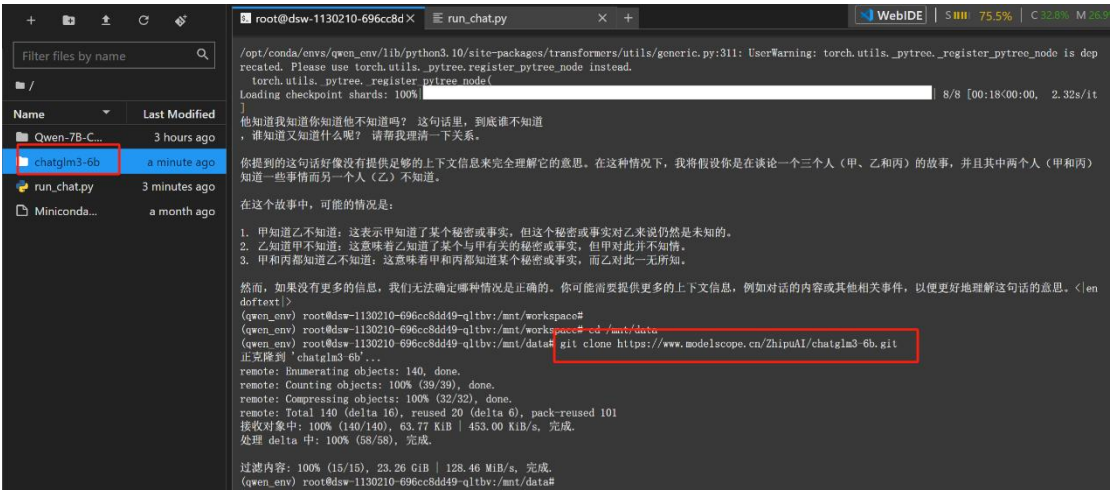
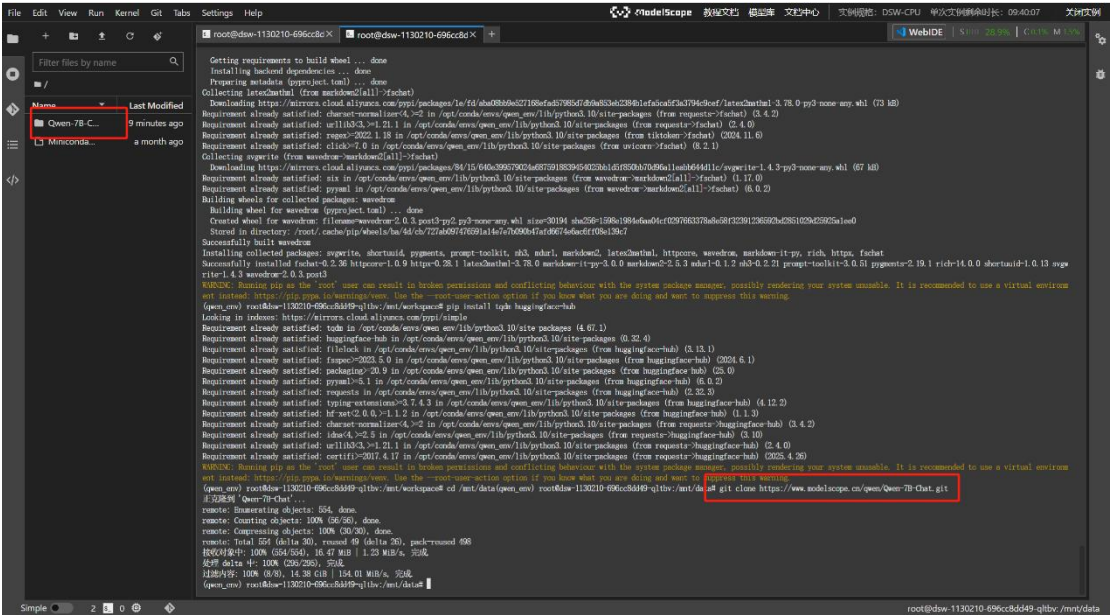


大语言模型部署体验报告

2352736 孙一宁

模型部署

根据指导手册在魔搭平台上完成环境的搭建和模型的部署，本次选用通义千问 Qwen-7B-Chat，智谱 ChatGLM3-6B 这两个开源的大语言模型；



模型测试

可以通过下图中的 python 脚本与大模型对话：

```

1 from transformers import TextStreamer, AutoTokenizer, AutoModelForCausalLM
2 model_name = "/mnt/data/Qwen-7B-Chat" # 本地路径
3 prompt = "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"
4 tokenizer = AutoTokenizer.from_pretrained(
5     model_name,
6     trust_remote_code=True
7 )
8 model = AutoModelForCausalLM.from_pretrained(
9     model_name,
10    trust_remote_code=True,
11    torch_dtype="auto" # 自动选择 float32/float16 (根据模型配置)
12 ).eval()
13 inputs = tokenizer(prompt, return_tensors="pt").input_ids
14 streamer = TextStreamer(tokenizer)
15 outputs = model.generate(inputs, streamer=streamer, max_new_tokens=300)

```

可以通过更改模型名称和问题内容，与不同的大模型进行对话，对比不同模型之间的差别

分别向通义千问 Qwen-7B-Chat，智谱 ChatGLM3-6B 提问三个相同的问题：

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
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

通义千问 Qwen-7B-Chat

```
(qwen_env) root@dsw-1129654-698774775c-66nfm:/mnt/workspace# python run_qwen_cpu.py
/opt/conda/envs/qwen_env/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% | 8/8 [00:45<00:00, 5.68s/it]
```

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语境和重点。

第一句话“冬天：能穿多少穿多少”是在谈论冬季保暖的建议，强调的是穿着暖和的重要性，以应对低温天气。

而第二句话“夏天：能穿多少穿多少”则是在谈论夏季防暑的建议，强调的是穿着透气和舒适的重要性，以应对高温天气。

因此，虽然两句都提到了“能穿多少就穿多少”的建议，但其侧重点不同，第一句关注的是保暖，第二句关注的是透气和舒适。<|endoftext|>

```
(qwen_env) root@dsw-1129654-698774775c-66nfm:/mnt/workspace#
```

```
(qwen_env) root@dsw-1130210-696cc8dd49-qltbv:/mnt/workspace#
python run_qwen_cpu.py
/opt/conda/envs/qwen_env/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% |
```

```
| 8/8 [00:27<00:00, 3.38s/it]
```

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，二是谁都看不上自己。

这两句话在表达上有所差异。第一句话将单身狗产生的原因归结为“谁都看不上”和“谁都看不上自己”，这是一种较为客观的表述方式，强调了单身狗在寻找伴侣的过程中遭遇困难和挫折的原因；而第二句话则更加主观，将单身狗的产生归因于自身的不足和缺陷，认为是因为“谁都看不上自己”，这更加强调个人的原因和心理状态，给人一种自我贬低的感觉。因此，这两句话在语气、情感色彩和表述方式上都存在差异。<|endoftext|>


```
(qwen_env) root@dsw-1130210-696cc8dd49-qltbv:/mnt/workspace#  
python run_qwen_cpu.py  
/opt/conda/envs/qwen_env/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%
```

```
| 8/8 [00:18<00:00, 2.32s/it]
```

他知道我知道你知道他不知道吗？这句话里，到底谁不知道，谁知道又知道什么呢？请帮我理清一下关系。

你提到的这句话好像没有提供足够的上下文信息来完全理解它的意思。在这种情况下，我将假设你是在谈论一个三个人（甲、乙和丙）的故事，并且其中两个人（甲和丙）知道一些事情而另一个人（乙）不知道。

在这个故事中，可能的情况是：

1. 甲知道乙不知道：这表示甲知道了某个秘密或事实，但这个秘密或事实对乙来说仍然是未知的。
2. 乙知道甲不知道：这意味着乙知道了某个与甲有关的秘密或事实，但甲对此并不知情。
3. 甲和丙都知道乙不知道：这意味着甲和丙都知道某个秘密或事实，而乙对此一无所知。

然而，如果没有更多的信息，我们无法确定哪种情况是正确的。你可能需要提供更多的上下文信息，例如对话的内容或其他相关事件，以便更好地理解这句话的意思。<|endoftext|>

智谱 ChatGLM3-6B

```
(qwen_env) root@dsw-1130210-696cc8dd49-qltbv:/mnt/data# python run_chat.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.  
/opt/conda/envs/qwen_env/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 [00:12<00:00, 1.81s/it]
```

[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。

```
(qwen_env) root@dsw-1130210-696cc8dd49-qltbv:/mnt/data#
```

```
(qwen_env) root@dsw-1130210-696cc8dd49-qltbv:/mnt/data# python run_chat.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.
/opt/conda/envs/qwen_env/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 [00:37<00:00, 5.42s/it]
[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。

第二句话“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，两个“谁”都指代“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。但是，第一句中的“谁”是主动的，自己看不上别人，而第二句中的“谁”是被动的，别人看不上自己。

```
(qwen_env) root@dsw-1130210-696cc8dd49-qltbv:/mnt/data# python run_chat.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.
/opt/conda/envs/qwen_env/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 [00:36<00:00, 5.28s/it]
[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，但是并不知道不知道自己不知道某些事情。

对于第一个问题，我们人类的正常理解是前半句的意思是冬天要穿多，夏天要穿少，两句的意思相反；

通义千问 Qwen-7B-Chat 的回答和人类的正常思维相符，能体现出两者的区别并指出区别的原因；智谱 ChatGLM3-6B 的回答说两句话意思相同，只是季节不同，让人感觉他并没有读懂这两句话区别的重点。综上，这个问题两个大语言模型都能找到季节的不同，通义千问 Qwen-7B-Chat 的回答要优于智谱 ChatGLM3-6B

对于第二个问题，我们人类的正常理解是人单身的原因，可能是自己看不上他人，也有可能是他人看不上自己。

通义千问 Qwen-7B-Chat 的回答让我有点没看明白，他的第一句话好像是看出来这两句话的区别了，但是后面的解释，第二句话是对的，第一句话不知道在说什么。智谱 ChatGLM3-6B 的回答感觉看似很有条理，实际在胡言乱语。综上，这一题通义千问 Qwen-7B-Chat 的回答可以算得分，但是智谱 ChatGLM3-6B 应该不可以得分

对于第三个问题，我们人类的正常理解应该是“他”不知道，这句话大致意思是，“你”知道“他不知道”这一事实，我又知道“你知道他不知道”这一事实，“我”想问“他”到底知不知道“我知道你知道他不知道”这一事实。

通义千问 Qwen-7B-Chat 转移了问题，把你我他变成甲乙丙，并且没有给出对应关系，这样你也不知道它最后回答的是“你我他”中的谁，既不能说他对，也不能完全说他错，那就算他没看懂问题吧；智谱 ChatGLM3-6B 定位到了这个知道一些东西，不知道一些东西的是“他”了，故可以说回答上来了这个问题，不过还有一些瑕疵，因为“他”到底知不知道‘我知道你知道他不知道’是未知的，也就是说可能他没有‘知道一些东西’。综上，这道题我认为智谱 ChatGLM3-6B 回答得更好，因为他找到了最关键的人物，而通义千问 Qwen-7B-Chat 没有看懂。

总结

通义千问-7B(Qwen-7B)是阿里云研发的通义千问大模型系列的 70 亿参数规模的模型。它是基于 Transformer 的大语言模型，在超大规模的预训练数据上进行训练

得到。ChatGLM3 是智谱 AI 和清华大学 KEG 实验室联合发布的对话预训练模型。两者都是当代优秀的大语言模型。

从这三个问题的回答可以看出，这两个大模型各有所长。通义千问 Qwen-7B-Chat 在语义逻辑差异识别、语境分析层面表现突出，对具象化场景（如季节穿衣逻辑）的理解更贴近人类思维，但在复杂逻辑推理题（如多主体认知嵌套问题）中存在逻辑转换偏差；智谱 ChatGLM3-6B 在主体定位与核心信息抓取上更具优势，尤其在多人物逻辑关系题中能快速锁定关键角色，但对语义对立关系的理解存在局限性，如因果差异，回答易偏离重点。

在优化方向上，通义千问 Qwen-7B-Chat 应加强复杂逻辑链的连续性（如多主体指代转换时需明确对应关系）；智谱 ChatGLM3-6B 应优化提升语义差异分析能力，避免在因果、主被动等基础对立关系题中漏答核心要点。