

## 大语言模型部署实验报告

姓 名: 雷思雨

学 号: 2354181

学 院: 计算机科学与技术学院

专业: 软件工程

学 科: 人工智能导论

指导教师: 邓浩

## 目录

I	大语言模型简介	1
II	魔搭社区大语言模型的部署	1
	2.1 环境配置	1
	2.1.1 conda 的配置	1
	2.1.2 基础依赖	1
	2.2 大模型的下载	2
Ш	大语言模型问答结果	2
	3.1 通义千问 Qwen-7B-Chat	2
	3.2 智普 chatglm3-6b	3
	3.3 百川 Baichuan2-7B-chat	4
IV	大语言模型横向对比	5

## I 大语言模型简介

大语言模型(Large Language Model,LLM)是一种基于深度学习技术的人工智能模型。其主要基于 Transformer 架构,通过对海量文本数据的训练,让模型学习到语言中词语、句子及段落之间的语义关联和语法规则,进而理解、生成和处理自然语言。

## II 魔搭社区大语言模型的部署

## 2.1 环境配置

详细的平台搭建和环境搭建步骤见"README.md"。

## 2.1.1 conda 的配置

#### conda 环境的下载:

#### 环境激活:

```
root@dsw-1136888-58ddc6ff79-w5ttn:/mnt/workspace# conda create -n qwen_env python=3.10 -y
root@dsw-1136888-58ddc6ff79-w5ttn:/mnt/workspace# source /opt/conda/etc/profile.d/conda.sh
root@dsw-1136888-58ddc6ff79-w5ttn:/mnt/workspace# conda activate qwen_env
```

#### 2.1.2 基础依赖

#### 基础环境:

```
(qwen_env) root@dsw-1136888-58ddc6ff79-w5ttn:/mnt/workspace# pip install \
torch==2.3.0+cpu \
torchvision==0.18.0+cpu \
--index-url https://download.pytorch.org/whl/cpu
```

#### 检查 pip 是否能正常联网:

(qwen\_env) root@dsw-1136888-58ddc6ff79-w5ttn:/mnt/workspace# pip install -U pip setuptools wheel

### 安装基础依赖:

```
(qwen_env) root@dsw-1136888-58ddc6ff79-w5ttn:/mnt/workspace# pip install \
    "intel-extension-for-transformers==1.4.2" \
    "neural-compressor==2.5" \
    "transformers==4.33.3" \
    "model scope==1.9.5" \
    "pydantic==1.10.13" \
    "sentencepiece" \
    "tiktoken" \
    "einops \
    "transformers_stream_generator" \
    "uvicorn" \
    "fastapi" \
    "yacs" \
    "setuptools_scm"
```

# 同僚大學

安装 fschat (需要启用 PEP517 构建):

(qwen\_env) root@dsw-1136888-58ddc6ff79-w5ttn:/mnt/workspace# pip install fschat --use-pep517

## 2.2 大模型的下载

切换到数据目录:

(qwen\_env) root@dsw-1136888-58ddc6ff79-w5ttn:/mmt/workspace# cd /mmt/data (qwen env) root@dsw-1136888-58ddc6ff79-w5ttn:/mmt/data#

下载大模型:

(qwen\_env) root@dsw-1136888-58ddc6ff79-w5ttn:/mnt/data# cd /mnt/workspace

(qwen\_env) root@dsw-1136888-58ddc6ff79-w5ttn:/mnt/data# git clone https://www.modelscope.cn/ZhipuAI/chatglm3-6b.git

(qwen\_env) root@dsw-1138778-7dc8fdf48d-h7764:/mnt/data# git clone https://www.modelscope.cn/baichuan-inc/Baichuan2-7B-chat

## III 大语言模型问答结果

本次实验采用的问题为:

- a) 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
- b) 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
- c) 他知道我知道你知道他不知道吗? 这句话里,到底谁不知道
- d) 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?
- e) 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。

### 3.1 诵义千问 Qwen-7B-Chat

具体代码见 run qwen cpu.py。

(qwen\_env) root@dsw-1136888-58ddc6ff79-w5ttn:/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 [00:50<00:00, 6.34s/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
这两句话的主要区别在于季节和语境。第一句话是在描述冬季的情况,强调的是保暖,因为冬季的天气通常比较冷,需要穿更多的衣服来保暖。而第二句话则是在描述夏季的情况,强调的是舒适度,因为夏季的天气通常比较热,需要穿少一点的衣服以保持凉爽。

此外,第一句话中的 能穿多少"表达了对衣物数量的限制或约束,可能是因为没有足够的衣物或为了节省开支。而第二句话中的 能穿多少"则表达了对衣物选择的自由度,可能是因为有足够的衣物或为了追求时尚或舒适。<|endoftext|>

人工智能导论

# 同僚大學

## 3.2 智普 chatglm3-6b

## 具体代码见 run\_zhipu\_cpu.py。

人工智能导论

# 同僚大學

```
(qwen_env) root@dsw-1136888-58ddc6ff79-w5ttn:/mnt/workspace# python run_zhipu_cpu.py
   Setting eos_token is not supported, use the default one.
Setting pad_token is not supported, use the default one.
  Setting pnd_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: 1005|

[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话或者绕口令,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。
   (qwen env) root@dsw-1136888-58ddc6ff79-w5ttn:/mnt/workspace# python run_zhipu_cpu.py
  (\text{Wer_env}\) rouseds\(^{\text{Piny local color}} = \text{Note: Note is not supported}, use the default one.} \)
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.
\(^{\text{Opt}/\text{conds-env}/\text{ib/python3.10/\text{site-packages/transformers/utils/generic.py:311: User\(^{\text{Warning: torch.utils.pytree._register_pytree_node}\) is deprecated. \(^{\text{Please use torch.utils.pytree.register_pytree_node}\) instead.
 torch. utils._pytree._register_pytree_node(
Loading_checkpoint_shards: 100%|
Loading_checkpoint_sh
                                                                                                                                                                                                                                                                                                                                                                  | 7/7 [00:03<00:00, 2.32it/s]
(gwen_env) root@dsw-ll36888-58ddc6ff79-w5ttn:/mnt/workspace# python run_zhipu_cpu.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.
Setting unk_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.
Setting unk_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.
Setting unk_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.
Setting unk_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.
Setting unk_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.
Setting unk_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.
Setting unk_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.
Setting unk_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.
Setting unk_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.
Setting unk_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.
Setting unk_token is not supported.
Setting unk_token is not supported.
Setting

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

 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
 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
           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。
```

## 3.3 百川 Baichuan2-7B-chat

## 具体代码见 run baichaun cpu.py。

```
(qwen_env) root@dsw-1138778-7dc8fdf48d-h7764:/mnt/workspace# python run_baichuan_cpu.py
/opt/conda/envs/qwen_env/lib/python3.10/site-packages/transformers/utils/generic.py:311: UserWarning: torch.utils._pytree._register_pytree_node instead.
torch.utils._pytree._register_pytree_node(
Xformers is not installed correctly. If you want to use memory_efficient_attention to accelerate training use the following command to i nstall Xformers
pip install xformers.
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/qwen_env/lib/python3.10/site-packages/torch/backends/cuda/_init__.py:342: FutureWarning: torch.backends.cuda.sdp_kernel() is deprecated. In the future, this context manager will be e removed. Please see, torch.nn.attention.sdpa_kernel() for the new context manager, with updated signature.

warnings.warn(

这两句话的区别在于,它们分别描述了在哪个季节应该采取什么样的穿衣策略。在冬天,建议穿更多的衣物来保暖;而在夏天,建议穿较少的衣物来散热热。〈/s〉
/opt/conda/envs/qwen_env/lib/python3.10/site-packages/transformers/utils/generic.py:311: UserWarning: torch.utils._pytree._register_pytree_node instead.
torch.utils._pytree._reg
```

(qwen\_env) root@dsw-1138778-7dc8fdf48d-h7764:/mmt/workspace# python run\_baichuan\_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(
Xformers is not installed correctly. If you want to use memory\_efficient\_attention to accelerate training use the following command to install Xformers

nstall Mormers
pip install xformers.
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/qwen\_env/lib/python3.10/site-packages/torch/backends/cuda/\_init\_.
py:342: 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.

在这个句子中,"他"不知道。因为句子的主题是"他",而"他"不知道是句子的主要信息息。〈/s〉

(qwen\_env) root@dsw-1138778-7dc8fdf48d-h7764:/mnt/workspace# python run\_baichuan\_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\_node(
Xformers is not installed correctly. If you want to use memory\_efficient\_attention to accelerate training use the following command to install Xformers

p明明明明白白白喜欢他,可她就是不说。/opt/conda/envs/qwen\_env/lib/python3.10/site-packages/torch/backends/cuda/\_\_init\_\_.py:342: Futur eWarning: torch.backends.cuda.sdp\_kernel() is deprecated. In the future, this context manager will be removed. Please see, torch.nn.atte ntion.sdpa\_kernel() for the new context manager, with updated signature.

warnings.warn() to the first manager, when posted a spinitule of warnings, warnings.warn() 这句话里,明明和白白谁喜欢谁? 这句话中,明明和白白谁喜欢谁? 从这句话中我们无法判断出明明和白白之间谁喜欢谁,因为"明明"和"白白"可能是指代两个人,也可能是指同一个人。为了更准确地理解这句话,我们需要更多的上下文信息息。〈/s〉

(qwen\_env) root@dsw-1138778-7dc8fdf48d-h7764:/mnt/workspace# python run\_baichuan\_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(

Xformers is not installed correctly. If you want to use memory\_efficient\_attention to accelerate training use the following command to install Xformers

nstall Atomers.
pip install xformers.
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。 /opt/conda/envs/qwen\_env/lib/python3.10/site-packages/torch/backends/cuda/\_\_init\_\_.py:342: FutureW arning: torch. backends.cuda.sdp\_kernel() is deprecated. In the future, this context manager will be removed. Please see, torch.nn. attent ion.sdpa\_kernel() for the new context manager, with updated signature.

warnings.warn( 小明:是我不好意思。请问:以上"意思"分别是什么意思。

答:在这段对话中,"意思"这个词在不同的语境下有不同的含义。以下是每个"意思"的含义:

- 1. 第一个"意思":表示某种行为或态度,例如领导对小明的不满意或者要求。 2. 第二个"意思":表示礼貌、谦逊,表示自己的行为还不够充分。 3. 第三个"意思":表示幽默、轻松、表示自己的行为并不重要。 4. 第四个"意思":表示自己的真如和谦虚,表示自己并没有其他目的。 5. 第五个"意思":表示自己对领导的歉意或者承认自己的错误。

总的来说,这段对话中的"意思"是一个多义词,在不同的语境下表示不同的含义义。</s>

## IV 大语言模型横向对比

## 三个大语言模型回答的对比:

问题	Qwen-7B-Chat	chatglm3-6b	Baichuan2-7B-chat
a	能识别两句话的不 同,但不能给出正确 原因。	仅能读懂句子的浅层 含义。	能识别两句话的不同,并给出解释。
b	能够理解两句话的正 确含义,并给出产生 原因。	能给出两句话的正确 涵义,但分析语序混 乱。	无法理解句子的不同。
c	能分析出多种可能 性,但不能得出正确 结论。	分析后得出正确结 论。	能给出正确答案,但 分析模糊。

同僚大學

人工智能导论

1	能给出存在的可能性	分析后得出错误结	无法分清句子的主谓
d	并给出正确结果。	论。	宾。
e	能较为准确地识别出 每个"意思"的含义, 但仍存在一定的语义 偏差。	能较识别出每个"意思"的不同含义,但部分语义偏差较大。	能较识别出"意思"的 不同含义,但指代不 清,且存在偏差。

综上所述,在一词多义或是一句多义的应用场景下,Qwen-7B-Chat 能够较好地识别关键词(或句)的不同含义,并给出相对可靠的分析; chatglm3-6b 次之,其能理解语境,但分析过程及结论存在一定偏差; Baichuan2-7B-chat 则是在理解上就存在偏差,分析和结论相对不可靠。

## 三个大语言模型架构及优缺点:

			1
	Qwen-7B-Chat	chatglm3-6b	Baichuan2-7B-chat
架构	Transformer	GLM	Transformer
优势	代码生成 多语言对话 轻量级部署	长文本处理 企业级复杂任务	多语言 医疗、法律领域领先
劣势	长文本处理稍弱	推理速度较慢	长文本处理稍弱 推理精度不足