



同濟大學
TONGJI UNIVERSITY

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

二〇二五年六月

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
III 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：

```
root@dsw-1136888-58ddc6ff79-w5ttn:/mnt/workspace# cd /opt/conda/envs
bash: cd: /opt/conda/envs: 没有那个文件或目录
root@dsw-1136888-58ddc6ff79-w5ttn:/mnt/workspace# wget https://repo.anaconda.com/miniconda/Miniconda3-latest-Linux-x86_64.sh
--2025-06-10 16:31:18-- https://repo.anaconda.com/miniconda/Miniconda3-latest-Linux-x86_64.sh
root@dsw-1136888-58ddc6ff79-w5ttn:/mnt/workspace# bash Miniconda3-latest-Linux-x86_64.sh -b -p /opt/conda
PREFIX=/opt/conda
root@dsw-1136888-58ddc6ff79-w5ttn:/mnt/workspace# echo 'export PATH="/opt/conda/bin:$PATH"' >> ~/.bashrc
root@dsw-1136888-58ddc6ff79-w5ttn:/mnt/workspace# source ~/.bashrc
root@dsw-1136888-58ddc6ff79-w5ttn:/mnt/workspace# conda --version
conda 25.3.1
```

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" \
    "modelscope==1.9.5" \
    "pydantic==1.10.13" \
    "sentencepiece" \
    "tiktoken" \
    "einops" \
    "transformers_stream_generator" \
    "uvicorn" \
    "fastapi" \
    "yacs" \
    "setuptools_scm"
```

```
(qwen_env) root@dsw-1136888-58dde6ff79-w5ttn:/mnt/workspace# pip install fschat --use-pep517
```

```
(qwen_env) root@dsw-1136888-58ddc6ff79-w5ttn:/mnt/workspace# cd /mnt/data
(qwen_env) root@dsw-1136888-58ddc6ff79-w5ttn:/mnt/data#
```

```
(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/ZipuAI/chatglm3-6b.git
(qwen_env) root@dsw-1138778-7dc8fd48d-h7764:/mnt/data# git clone https://www.modelscope.cn/baichuan-inc/Baichuan2-7B-chat
```

```
(qwen_env) root@dsw-1136888-58dde6f79-w5ttt:/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。
```

3.2 智普 chatglm3-6b

具体代码见 `run_zhipu_cpu.py`。

3

3.3 百川| Baichuan2-7B-chat

具体代码见 `run baichaun cpu.py`。

```
(qwen_env) root@dsw-1138778-7dc8fd48d-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
ee_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 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 b
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散热
。</s>

(qwen_env) root@dsw-1138778-7dc8fd48d-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
ee_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 i
nstall Xformers
pip install xformers.
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/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.
  warnings.warn(

单身狗产生的原因有两个，一是谁都看不上，二是谁都看不上上。</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
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.
  warnings.warn(
呢？

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。</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
pip install xformers.
明明明明明白白白喜欢他，可她就是不。/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.
  warnings.warn(
这句话里，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。</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
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: 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.
  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”：表示自己的真诚和谦虚，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。</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。

d	能给出存在的可能性并给出正确结果。	分析后得出错误结论。	无法分清句子的主谓宾。
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：

	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