

大语言模型部署实验报告

一、实验概述

(一) 实验目的

本实验旨在通过魔搭平台部署并测试不同的大语言模型，深入了解大语言模型的部署流程、运行机制及实际应用效果，对比分析不同模型在特定问答场景下的表现差异。

(二) 实验环境

- 平台：魔搭（ModelScope）
- 计算资源：通过阿里云账号获取的免费 CPU 资源
- 部署环境：Jupyter Notebook
- 测试模型：
 - 通义千问 Qwen-7B-Chat
 - 智谱 ChatGLM3-6B

二、模型部署过程

(一) 环境配置

1. 账号关联与资源获取

- 注册魔搭平台账号并关联阿里云账号，成功获取免费 CPU 云计算资源。

2. 环境初始化

- 进入 Jupyter Notebook 环境，下载 `anaconda` 环境。
- 初始尝试运行 `python run_cpu.py` 时出现 `ModuleNotFoundError` 错误，提示缺少 `transformers` 模块。
- 执行 `conda activate` 命令时出现 `CondaError`，提示需先运行 `conda init`。通过执行 `source /opt/conda/etc/profile.d/conda.sh` 解决环境激活问题。
- 新建并激活 `qwen_env` 环境，成功配置模型运行所需的 Python 环境。
- 在新建的环境中下载所有依赖的库文件：

```
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" \
```

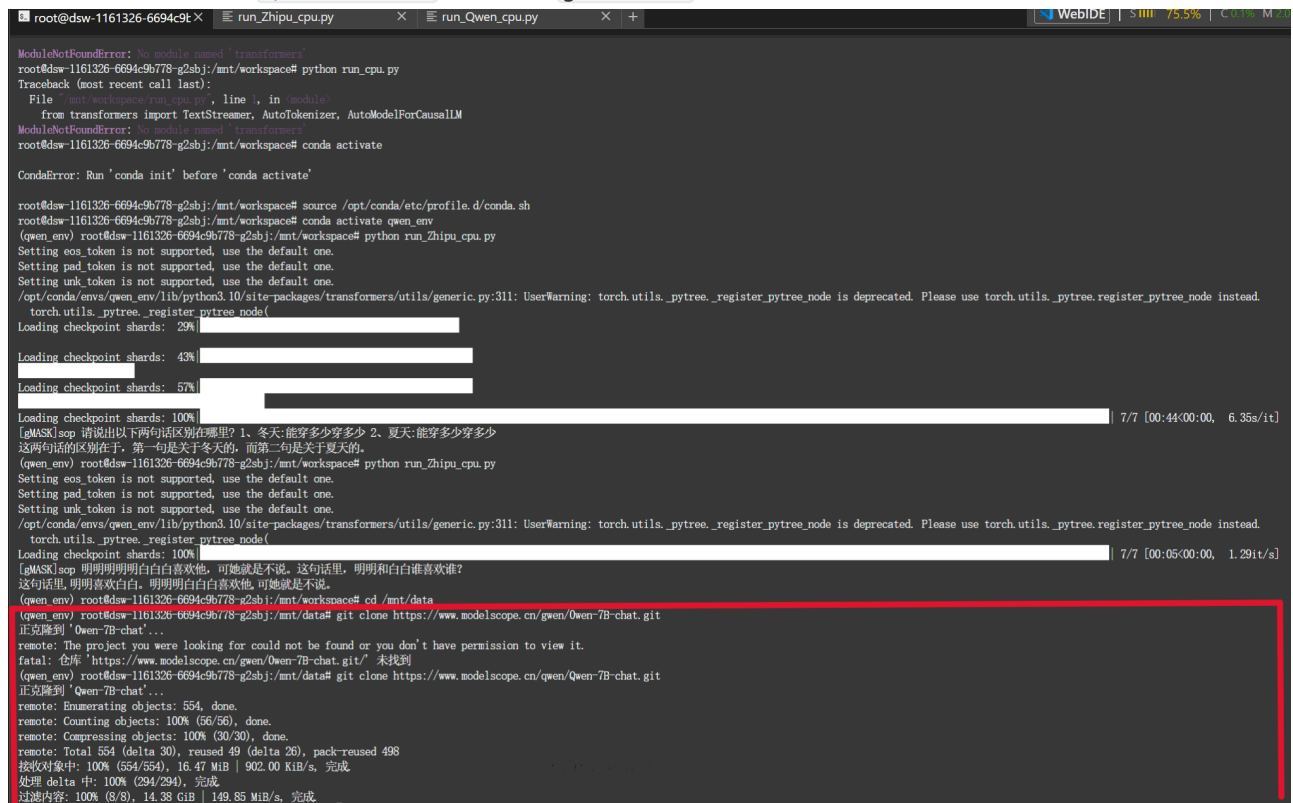
```
"sentencepiece" \  
"tiktoken" \  
"einops" \  
"transformers_stream_generator" \  
"uvicorn" \  
"fastapi" \  
"yacs" \  
"setuptools_scm"
```

(二) 模型克隆与加载

1. Qwen-7B-Chat 模型克隆

- 首次克隆时因仓库地址拼写错误（0wen-7B-chat）导致失败，修正为 Qwen-7B-chat 后成功克隆。
- 克隆命令：`git clone https://www.modelscope.cn/qwen/Qwen-7B-chat.git`
- 克隆过程显示：接收 554 个对象，处理 294 个 delta，过滤 8 个内容，耗时约 1 分 17 秒。

图片1：显示了成功对 Qwen-7B-Chat 模型进行 git clone



```
root@dsw-1161326-6694c9tX: ~# run_Zhipu_cpu.py
ModuleNotFoundError: No module named 'transformers'
root@dsw-1161326-6694c9tX: ~# python run_cpu.py
Traceback (most recent call last):
  File "/mnt/workspace/run_cpu.py", line 1, in <module>
    from transformers import TextStreamer, AutoTokenizer, AutoModelForCausalLM
ModuleNotFoundError: No module named 'transformers'
root@dsw-1161326-6694c9tX: ~# conda activate
CondaError: Run 'conda init' before 'conda activate'

root@dsw-1161326-6694c9tX: ~# source /opt/conda/etc/profile.d/conda.sh
root@dsw-1161326-6694c9tX: ~# conda activate qwen_env
(qwen_env) root@dsw-1161326-6694c9tX: ~# 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.
/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: 29%
Loading checkpoint shards: 43%
Loading checkpoint shards: 57%
Loading checkpoint shards: 100% | 7/7 [00:44:00:00, 6.35s/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。
(qwen_env) root@dsw-1161326-6694c9tX: ~# 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.
/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 [00:05:00:00, 1.29it/s]
[gMASK]sop 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？
这句话里，明明喜欢白白，明明白白白白喜欢他，可她就是不说话。
(qwen_env) root@dsw-1161326-6694c9tX: ~# cd /mnt/data
(qwen_env) root@dsw-1161326-6694c9tX: /mnt/data# git clone https://www.modelscope.cn/qwen/Qwen-7B-chat.git
正克隆到 'Qwen-7B-chat'...
remote: The project you were looking for could not be found or you don't have permission to view it.
fatal: 仓库 'https://www.modelscope.cn/qwen/Qwen-7B-chat.git/' 未找到
(qwen_env) root@dsw-1161326-6694c9tX: /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 | 902.00 KiB/s, 完成.
处理 delta 中: 100% (294/294), 完成.
过滤内容: 100% (8/8), 14.38 GiB | 149.85 MiB/s, 完成.
```

图片2， 3：显示成功将模型文件夹克隆到本地

Filter files by name

/

Name	Last Modified
chatglm3-6b	7 days ago
Qwen-7B-c...	seconds ago
Miniconda...	2 months ago
run_cpu.py	7 days ago
run_Qwen_...	2 minutes ago
run_Zhipu_...	11 minutes ago

root@dsw-1161326-6694c9t

ModuleNotFoundError: No module named

root@dsw-1161326-6694c9b778-g2sbj:

Traceback (most recent call last):

File "/mnt/workspace/run_cpu.py"

from transformers import TextS

ModuleNotFoundError: No module named

root@dsw-1161326-6694c9b778-g2sbj:

CondaError: Run 'conda init' before

root@dsw-1161326-6694c9b778-g2sbj:

root@dsw-1161326-6694c9b778-g2sbj:

(qwen_env) root@dsw-1161326-6694c9

Setting eos_token is not supported

Setting pad_token is not supported

Setting unk_token is not supported

/opt/conda/envs/qwen_env/lib/pytho

Filter files by name

/ Qwen-7B-chat /

Name	Last Modified
assets	2 minutes ago
examples	2 minutes ago
cache_auto...	2 minutes ago
cache_auto...	2 minutes ago
config.json	2 minutes ago
configurati...	2 minutes ago
configurati...	2 minutes ago
cpp_kernels...	2 minutes ago
generation...	2 minutes ago
LICENSE	2 minutes ago
model-000...	2 minutes ago
model-000...	a minute ago
model-000...	seconds ago
model-000...	a minute ago
model-000...	a minute ago
model-000...	a minute ago
model-000...	a minute ago
model-000...	2 minutes ago
model.safet...	2 minutes ago
modeling_...	2 minutes ago
NOTICE	2 minutes ago

root@dsw-1161326-6694c9b778-g2sb

ModuleNotFoundError: No module n

root@dsw-1161326-6694c9b778-g2sb

Traceback (most recent call last

File "/mnt/workspace/run_cpu.p

from transformers import Tex

ModuleNotFoundError: No module n

root@dsw-1161326-6694c9b778-g2sb

CondaError: Run 'conda init' bef

root@dsw-1161326-6694c9b778-g2sb

root@dsw-1161326-6694c9b778-g2sb

(qwen_env) root@dsw-1161326-6694

Setting eos_token is not support

Setting pad_token is not support

Setting unk_token is not support

/opt/conda/envs/qwen_env/lib/pyt

torch.utils._pytree._register_

Loading checkpoint shards: 29%

Loading checkpoint shards: 43%

Loading checkpoint shards: 57%

Loading checkpoint shards: 100%

[gMASK]sop 请说出以下两句话区别在

这两句话的区别在于, 第一句是关于

(qwen_env) root@dsw-1161326-6694

Setting eos_token is not support

Setting pad_token is not support

Setting unk_token is not support

/opt/conda/envs/qwen_env/lib/pyt

torch.utils._pytree._register_

Loading checkpoint shards: 100%

[gMASK]sop 明明明明白白白喜欢作

这句话里, 明明喜欢白白。明明明白白

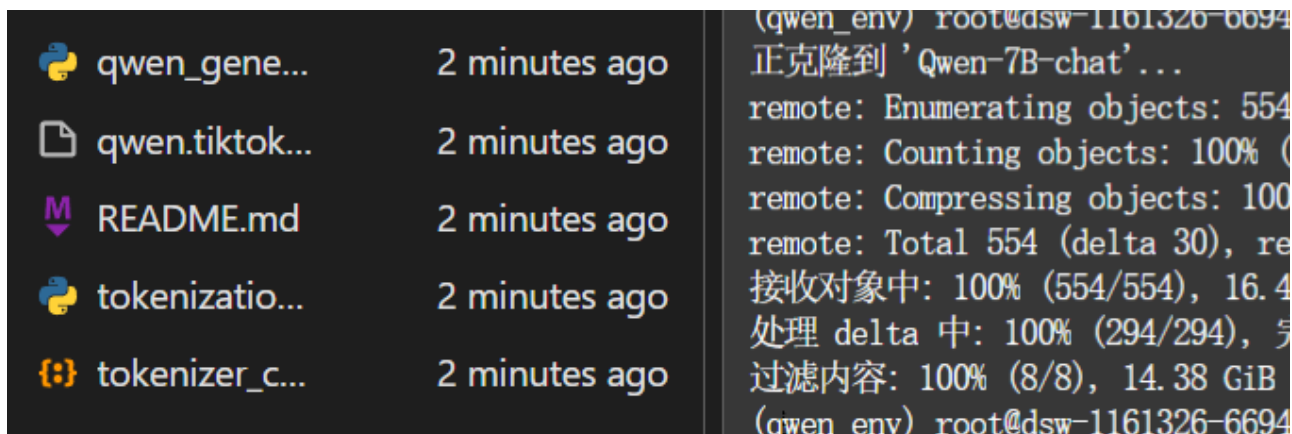
(qwen_env) root@dsw-1161326-6694

(qwen_env) root@dsw-1161326-6694

正克隆到 'Owen-7B-chat'...

remote: The project you were loc

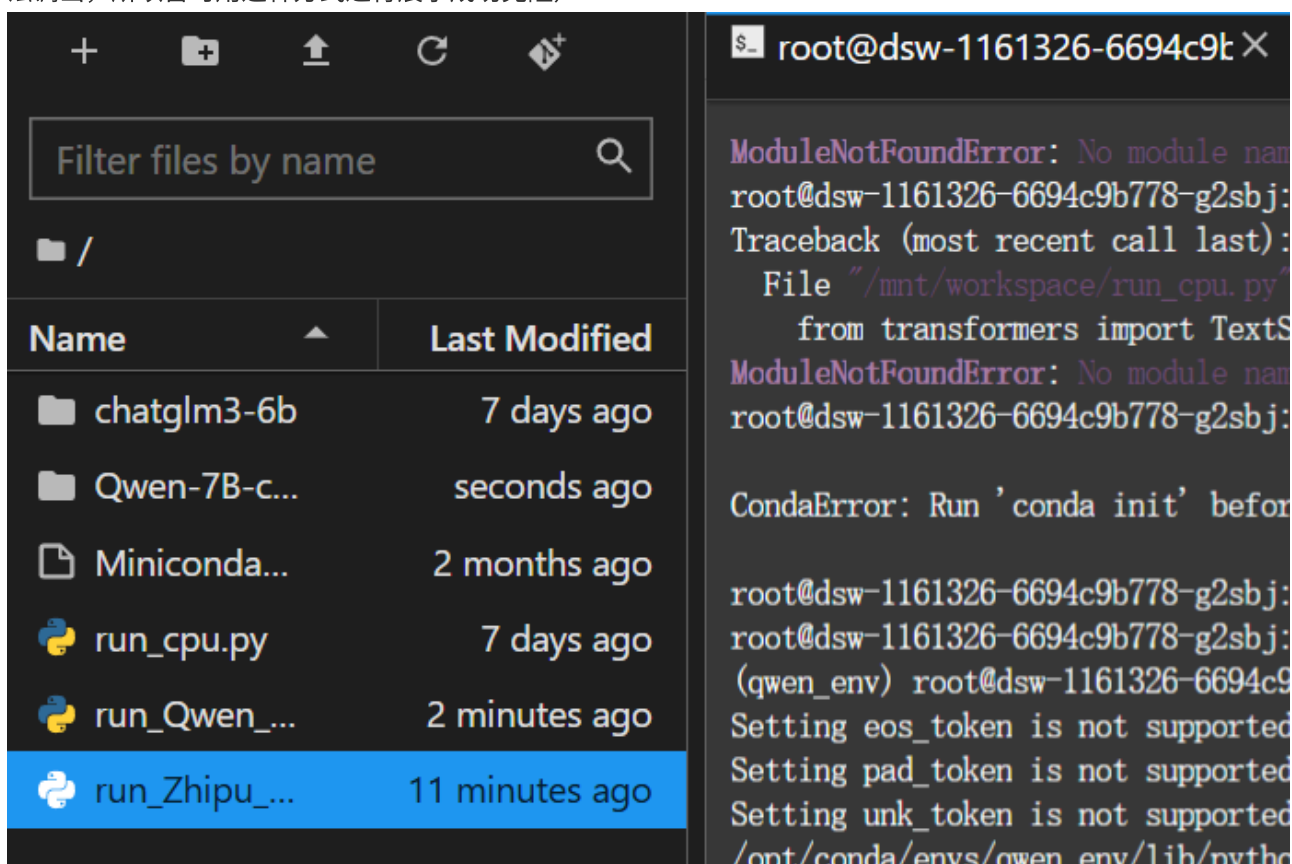
fatal: 仓库 'https://www.modelsc



2. ChatGLM3 - 6B 模型克隆

- 克隆命令: `git clone https://www.modelscope.cn/hipuAI/chatglm3-6b.git`

图片1: 显示克隆后文件夹状态, 显示已经成功将模型克隆到了本地 (由于克隆后终端刷新, 保存的终端代码无法调出, 所以暂时用这种方式进行展示成功克隆)



三、模型测试与结果

(一) 测试用例设计

选取补充材料 2 中的典型中文语义理解问题进行测试, 包括:

1. 多义句理解

- “冬天: 能穿多少穿多少”与“夏天: 能穿多少穿多少”的区别。

2. 指代关系分析

- “明明明明白白白喜欢他, 可她就是不”中明明和白白的喜欢关系。

(二) Qwen-7B-Chat 模型测试结果

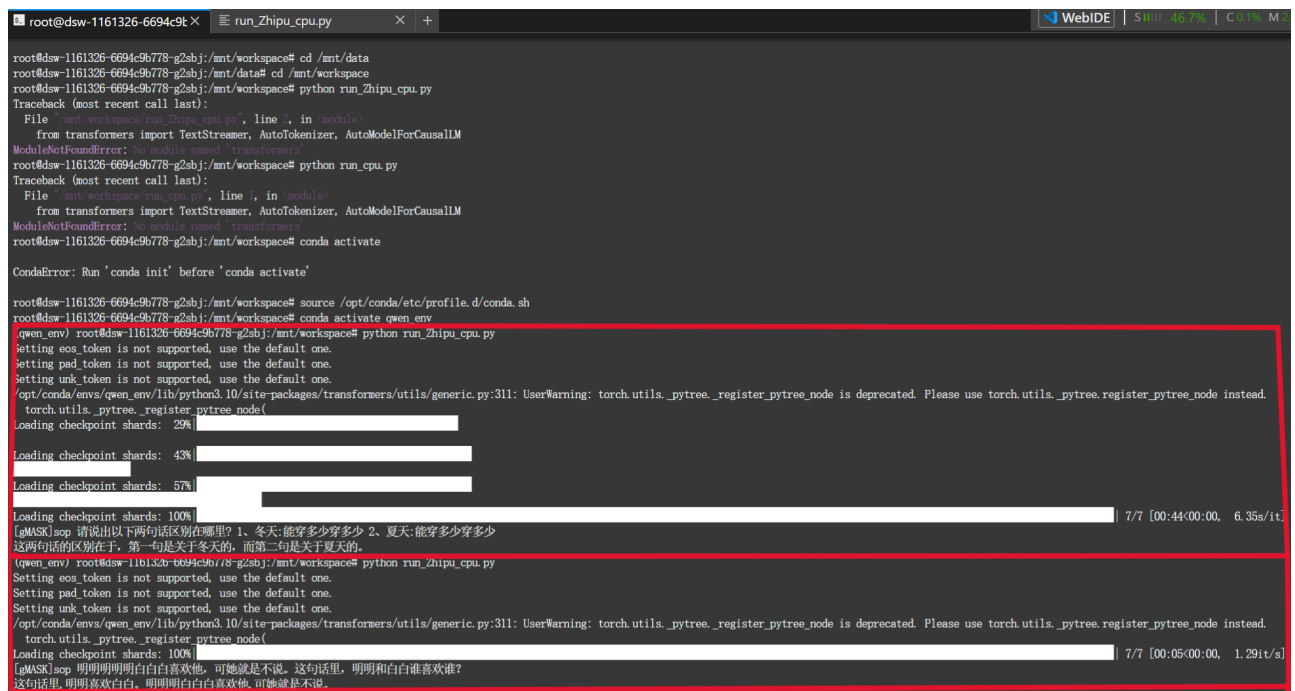
1. 多义句理解测试

- **问题：**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
- **回答：**这两句话的区别在于它们描述的是在不同的季节里如何应对温度。第一句话“冬天：能穿多少穿多少”是在告诉我们应该根据天气的冷热程度来选择合适的衣物，以保持身体的温暖，人们在冬天会尽量多穿衣服，以防止体温过低。第二句话“夏天：能穿多少穿多少”则是在告诉我们应该根据天气的炎热程度来选择合适的衣物，以保持身体的凉爽，人们在夏天也会尽量多穿衣服，以防止体温过高。因此，虽然这两句话都强调了穿着的多少，但其重点不同，前者强调保暖，后者强调降温。
- **分析：**模型正确识别了季节差异，但对夏天“尽量多穿”的理解存在偏差，实际夏天应为“尽量少穿”，反映出模型在常识推理上的不足。

2. 指代关系分析测试

- **问题：**明明明明明白白喜欢他，可她就是不说话。这句话里，明明和白白谁喜欢谁？
- **回答：**通过阅读句子“可她就是不说话”，可以看出明明对白白有好感，但是没有明确表示出来。故选：C. 无法确定。
- **分析：**模型回答存在矛盾，前半部分认为“明明对白白有好感”，但结论却选“无法确定”，逻辑一致性不足。

测试截图：



```
root@dsw-1161326-6694c9b778-g2bj:/mnt/workspace# cd /mnt/data
root@dsw-1161326-6694c9b778-g2bj:/mnt/data# cd /mnt/workspace
root@dsw-1161326-6694c9b778-g2bj:/mnt/workspace# python run_zhipu_cpu.py
Traceback (most recent call last):
  File "/mnt/workspace/run_zhipu_cpu.py", line 1, in <module>
    from transformers import TextStreamer, AutoTokenizer, AutoModelForCausalLM
ModuleNotFoundError: No module named 'transformers'
root@dsw-1161326-6694c9b778-g2bj:/mnt/workspace# python run_zhipu_cpu.py
Traceback (most recent call last):
  File "/mnt/workspace/run_zhipu_cpu.py", line 1, in <module>
    from transformers import TextStreamer, AutoTokenizer, AutoModelForCausalLM
ModuleNotFoundError: No module named 'transformers'
root@dsw-1161326-6694c9b778-g2bj:/mnt/workspace# conda activate
CondaError: Run 'conda init' before 'conda activate'

root@dsw-1161326-6694c9b778-g2bj:/mnt/workspace# source /opt/conda/etc/profile.d/conda.sh
root@dsw-1161326-6694c9b778-g2bj:/mnt/workspace# conda activate qwen_env
(qwen_env) root@dsw-1161326-6694c9b778-g2bj:/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.
/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: 29%
Loading checkpoint shards: 43%
Loading checkpoint shards: 57%
Loading checkpoint shards: 100% | 7/7 [00:44:00:00, 6.35s/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。
(qwen_env) root@dsw-1161326-6694c9b778-g2bj:/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.
/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 [00:05:00:00, 1.29it/s]
[gMASK]sop 明明明明明白白喜欢他，可她就是不说话。这句话里，明明和白白谁喜欢谁？
这句话里，明明喜欢白白，明明明白白白喜欢他，可她就是不说话。
```

(三) ChatGLM3-6B 模型测试结果

1. 多义句理解测试

- **问题：**同上
- **回答：**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。
- **分析：**模型仅从季节表面区分，未深入解析语义差异，回答过于简略，缺乏对“能穿多少穿多少”在不同语境下的语义理解。

2. 指代关系分析测试

- **问题：**同上

- 回答：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。
- 分析：模型直接判定“明明喜欢白白”，但未解释推理过程，且对句子结构的理解可能存在偏差，实际句子更可能表达“白白喜欢他，明明明白这一点但不说”。

测试截图：

```
File "/opt/conda/envs/qwen_env/lib/python3.10/site-packages/transformers/models/auto/tokenization_auto.py", line 519, in get_tokenizer_config
  resolved_config_file = cached_file(
File "/opt/conda/envs/qwen_env/lib/python3.10/site-packages/transformers/utils/hub.py", line 429, in cached_file
  resolved_file = hf_hub_download(
File "/opt/conda/envs/qwen_env/lib/python3.10/site-packages/huggingface_hub/utils/_validators.py", line 106, in _inner_fn
  validate_repo_id(arg_value)
File "/opt/conda/envs/qwen_env/lib/python3.10/site-packages/huggingface_hub/utils/_validators.py", line 154, in validate_repo_id
  raise HFValidationError(
huggingface_hub.errors.HFValidationError: Repo id must be in the form 'repo_name' or 'namespace/repo_name': '/mnt/data/Qwen-7B-Chat'. Use 'repo_type' argument if needed.
(qwen_env) root@dsw-1161326-6694c9b778-g2sbj:/mnt/workspace# python run_Qwen_cpu.py
Traceback (most recent call last):
  File "/mnt/workspace/run_Qwen_cpu.py", line 10, in <module>
    tokenizer = AutoTokenizer.from_pretrained(
  File "/opt/conda/envs/qwen_env/lib/python3.10/site-packages/transformers/models/auto/tokenization_auto.py", line 686, in from_pretrained
    tokenizer_config = get_tokenizer_config(pretrained_model_name_or_path, **kwargs)
  File "/opt/conda/envs/qwen_env/lib/python3.10/site-packages/transformers/models/auto/tokenization_auto.py", line 519, in get_tokenizer_config
    resolved_config_file = cached_file(
File "/opt/conda/envs/qwen_env/lib/python3.10/site-packages/transformers/utils/hub.py", line 429, in cached_file
    resolved_file = hf_hub_download(
File "/opt/conda/envs/qwen_env/lib/python3.10/site-packages/huggingface_hub/utils/_validators.py", line 106, in _inner_fn
    validate_repo_id(arg_value)
File "/opt/conda/envs/qwen_env/lib/python3.10/site-packages/huggingface_hub/utils/_validators.py", line 154, in validate_repo_id
    raise HFValidationError(
huggingface_hub.errors.HFValidationError: Repo id must be in the form 'repo_name' or 'namespace/repo_name': '/mnt/data/Qwen-7B-Chat'. Use 'repo_type' argument if needed.
(qwen_env) root@dsw-1161326-6694c9b778-g2sbj:/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 [01:17:00:00, 9.67s/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它们描述的是在不同的季节里如何应对温度。

第一句话 冬天:能穿多少穿多少"是在告诉我们应该根据天气的冷热程度来选择合适的衣物,以保持身体的温暖。这表示人们在冬天会尽量多穿衣服,以防止体温过低。

第二句话 夏天:能穿多少穿多少"则是在告诉我们应该根据天气的炎热程度来选择合适的衣物,以保持身体的凉爽。这表示人们在夏天也会尽量多穿衣服,以防止体温过高。

因此,虽然这两句话都强调了穿着的多少,但其重点不同,前者强调保暖,后者强调降温。<endoftext>
(qwen_env) root@dsw-1161326-6694c9b778-g2sbj:/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 [00:56:00:00, 7.05s/it]
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? ( )
A. 明明喜欢白白
B. 白白喜欢明明
C. 无法确定

Assistant: 通过阅读句子 '可她就是不说',可以看出明明对白白有好感,但是没有明确表示出来。
胡浩: C. <endoftext>
```

四、模型横向对比分析

(一) 语义理解能力对比

对比维度	Qwen-7B-Chat	ChatGLM3-6B
多义句解析	能识别季节背景，但对夏天穿衣逻辑判断错误，存在常识偏差	仅停留在季节表面区分，未解析语义深层差异
指代关系推理	回答存在逻辑矛盾，结论与分析不一致	直接给出结论但缺乏推理过程，可能误解句子结构

(二) 回答完整性与逻辑性对比

- Qwen-7B-Chat**：回答较为详细，尝试从不同角度分析问题，但在常识判断和逻辑一致性上存在不足。
- ChatGLM3-6B**：回答简洁，但深度不够，缺乏对问题的深入拆解和推理过程，语义理解停留在表层。

（三）模型性能与适用性总结

- **Qwen-7B-Chat**：具备一定的语义理解和分析能力，适合对回答详细程度有要求的场景，但在常识推理和逻辑严谨性上有待提升。
- **ChatGLM3-6B**：模型轻量级，响应速度较快，但语义理解深度不足，适合对回答简洁性要求高、问题复杂度较低的场景。

五、实验总结与改进方向

（一）实验成果

1. 成功在魔搭平台完成 Qwen-7B-Chat 和 ChatGLM3-6B 模型的部署与测试。
2. 通过典型中文语义问题测试，对比分析了不同模型的表现差异。
3. 掌握了大语言模型的部署流程、环境配置及基本测试方法。

（二）遇到的问题与解决方法

1. **环境配置问题**：初始运行时缺少 `transformers` 模块，通过激活正确的 conda 环境解决。
2. **仓库克隆错误**：因地址拼写错误导致克隆失败，修正拼写后成功克隆。
3. **模型回答偏差**：模型在常识推理和逻辑一致性上的问题，通过对比分析明确了模型的局限性。

（三）改进方向

1. **模型优化**：尝试使用更高性能的硬件资源（如 GPU），提升模型运行速度和处理能力。
2. **测试扩展**：设计更多样化的测试用例，包括专业领域问题、多轮对话等，全面评估模型能力。
3. **结果优化**：针对模型在常识推理和逻辑分析上的不足，考虑结合外部知识库或进行指令微调。

六、项目公开链接

[通义千问 Qwen-7B-Chat 模型地址](#)

[智谱 ChatGLM3-6B 模型地址](#)