算能SC7 224T测试报告

1. 环境规格

硬件	组件	详情
服务器（10.110.181.137）	处理器	Intel(R) Xeon(R) Gold 5218 CPU @ 2.30GHz
	内存	125Gi
	型号	DELL R740
	CPU核数	64
	操作系统	Ubuntu 22.04.4 LTS
GPU	型号	SC7 224T
	显存	128GB LPDDR4x 1024bit
	显存带宽	470GB/s
	接口规格	PCIe Gen4 X16
	峰值算力	单卡8芯合起来的算力如下 FP32：14 TFLOPS FP16/BF16：112 TFLOPS INT8：224 TOPS
	TDP	300W
	最大操作温度	

2. 环境部署

2.1 驱动及软件包下载

从算能官网官网：https://developer.sophgo.com/site/index/material/all/all.html -> SDK-24.04.01

2.2 硬件检查

服务器插卡后，可以通过以下命令检查加速卡是否安装正确。

```
lspci | grep 168
```

此时会看到置于该计算机内的 BM1684X 设备，即板卡上 对应的芯片个数。（SC7 FP150 对应六颗芯片，SC7 224T 对应八颗芯片）

```
root@test-hpc-05:~# lspci | grep 168
3d:00.0 Processing accelerators: Device 1f1c:1686 (rev 01)
3e:00.0 Processing accelerators: Device 1f1c:1686 (rev 01)
3f:00.0 Processing accelerators: Device 1f1c:1686 (rev 01)
40:00.0 Processing accelerators: Device 1f1c:1686 (rev 01)
41:00.0 Processing accelerators: Device 1f1c:1686 (rev 01)
42:00.0 Processing accelerators: Device 1f1c:1686 (rev 01)
43:00.0 Processing accelerators: Device 1f1c:1686 (rev 01)
44:00.0 Processing accelerators: Device 1f1c:1686 (rev 01)
```

2.3 驱动+软件栈安装

```
# 安装依赖库，只需要执行一次：
sudo apt install dkms libncurses5
# 安装libsophon:
sudo dpkg -i sophon-driver_0.5.1_amd64.deb
sudo dpkg -i sophon-libsophon_0.5.1_amd64.deb
# 在终端执行如下命令，或者登出再登入当前用户后即可使用bm-smi等命令：
source /etc/profile
```

2.4 资源监控

安装完成后使用bm-smi命令进行检查驱动是否安装成功。

```
Wed Sep  4 10:21:26 2024
+-----------------------------------------------------------------------------------------+
| Lib Version:    0.5.1              Driver Version:  0.5.1                               |
+-----------------------------------+-----------------------------------------------------+
|card  Name      Mode        SN     |TPU  boardT  chipT   TPU_P  TPU_V  ECC  CorrectN  Tpu-Util|
|12V_ATX  MaxP boardP Minclk Maxclk  Fan|Bus-ID      Status   Currclk   TPU_C   Memory-Usage      |
|===================================+=====================================================|
| 0  SC7-224T    PCIE HQDZW59BDJEJD0078 | 0    33C     42C    4.9W   836mV  OFF    N/A          0% |
|8434mA  300W  101W   25M     875M   N/A|000:3d:00.0 Active    875M      5.9A    81MB/14787MB      |
|                                   |-----------------------------------------------------|
|                                   | 1    42C     44C    3.4W   843mV  OFF    N/A          0% |
|                                   |000:3e:00.0 Active    875M      4.0A    81MB/14787MB      |
|                                   |-----------------------------------------------------|
|                                   | 2    44C     46C    3.4W   843mV  OFF    N/A          0% |
|                                   |000:3f:00.0 Active    875M      4.0A    81MB/14787MB      |
|                                   |-----------------------------------------------------|
|                                   | 3    40C     43C    4.9W   836mV  OFF    N/A          0% |
|                                   |000:40:00.0 Active    875M      5.9A    81MB/14787MB      |
|                                   |-----------------------------------------------------|
|                                   | 4    42C     46C    3.4W   843mV  OFF    N/A          0% |
|                                   |000:41:00.0 Active    875M      4.1A    81MB/14787MB      |
|                                   |-----------------------------------------------------|
|                                   | 5    46C     46C    3.4W   843mV  OFF    N/A          0% |
|                                   |000:42:00.0 Active    875M      4.0A    81MB/14787MB      |
|                                   |-----------------------------------------------------|
|                                   | 6    43C     47C    5.1W   836mV  OFF    N/A          0% |
|                                   |000:43:00.0 Active    875M      6.1A    81MB/14787MB      |
|                                   |-----------------------------------------------------|
|                                   | 7    47C     46C    5.1W   835mV  OFF    N/A          0% |
|                                   |000:44:00.0 Active    875M      6.1A    81MB/14787MB      |
+===================================+=====================================================+

+-----------------------------------------------------------------------------------------+
| Processes:                                                                  TPU Memory  |
|  TPU-ID       PID   Process name                                            Usage       |
|=========================================================================================|
```

3. Qwen2-7b测试

3.1 环境安装

1. 创建虚拟环境

```
1 python3 -m venv sc7_venv
```

这里 sc7_venv 是你虚拟环境的名称，你可以根据需要更改。

2. 激活虚拟环境

```
1 source /home/lsc/SC7/sc7_venv/bin/activate
```

激活后，你会看到命令提示符前面有 `(sc7_venv)` ，表示你已进入虚拟环境。

3. **安装依赖**

在虚拟环境中，你可以使用 `pip` 安装所需的包。例如：

```
sudo apt-get update
pip3 install transformers_stream_generator einops tiktoken accelerate gradio transformers==4.41.2
pip3 install pybind11[global]
```

4. **退出虚拟环境（测试完再退出）**

当你完成工作后，可以使用以下命令退出虚拟环境：

```
# 可选，测试完再退出
deactivate
```

5. 下载bmodel

```
pip3 install dfss
python3 -m dfss --url=open@sophgo.com:/ext_model_information/LLM/LLM-TPU/qwen2-7b_int4_seq8192_1dev.bmodel
```

6. 编译chat.cpp

```
cd /home/lsc/SC7/LLM-TPU/models/Qwen2/python_demo/
mkdir build
cd build && cmake .. && make && cp *cpython* .. && cd ..
```

3.2 seq8192_1dev模型测试

7. 执行推理--CLI方式

```
python3 pipeline.py --model_path /home/lsc/SC7/bmodels/qwen2-7b_int4_seq8192_1dev.bmodel --tokenizer_path ../support/token_config/ --devid 0 --generation_mode greedy
```

```
(sc7_venv) root@test-hpc-05:/home/lsc/SC7/LLM-TPU/models/Qwen2/python_demo# python3 pipeline.py --model_path /home/lsc/SC7/bmodels/
qwen2-7b_int4_seq8192_1dev.bmodel --tokenizer_path ../support/token_config/ --devid 0 --generation_mode greedy
Load ../support/token_config/ ...
Special tokens have been added in the vocabulary, make sure the associated word embeddings are fine-tuned or trained.
Device [ 0 ] loading ....
open usercpu.so, init user_cpu_init
Model[/home/lsc/SC7/bmodels/qwen2-7b_int4_seq8192_1dev.bmodel] loading ....
Done!

Load Time: 8.042 s

==============================================================
1. If you want to quit, please enter one of [q, quit, exit]
2. To create a new chat session, please enter one of [clear, new]
==============================================================

Question: how are you

Answer: As an AI, I don don't have feelings, but I'm here to help you. How can I assist you today?
FTL: 15.430 s
TPS: 7.919 token/s

Question: who is the president of usa?

Answer: As of now, the President of the United States is Joe Biden. He took office on January 20, 2021.
FTL: 15.413 s
TPS: 7.907 token/s

Question:
```

4芯异常按FAE临时方案修复后加载时间延长，吞吐下降，详见问题4：

1.芯片跑挂之后，需要重置寄存器，运行allreduce_reg_init这个程序
2.跑4芯模型需要对驱动稍作修改：把libbmrt.so.1.0和libbmrt.a复制到/opt/sophon/libsophon-current/lib/

```
(sc7_venv) root@test-hpc-05:/home/lsc/SC7/LLM-TPU/models/Qwen2/python_demo# python3 pipeline.py --model_path /home/lsc/S
C7/bmodels/qwen2-7b_int4_seq8192_1dev.bmodel --tokenizer_path ../support/token_config/ --devid 0 --generation_mode greed
y
Load ../support/token_config/ ...
Device [ 0 ] loading ....
open usercpu.so, init user_cpu_init
Model[/home/lsc/SC7/bmodels/qwen2-7b_int4_seq8192_1dev.bmodel] loading ....
Done!

Load Time: 54.556 s

==============================================================
1. If you want to quit, please enter one of [q, quit, exit]
2. To create a new chat session, please enter one of [clear, new]
==============================================================

Question: how are you!

Answer: As an AI, I don't have feelings, but I'm here and ready to assist you. How can I help you today?
FTL: 15.461 s
TPS: 7.863 token/s

Question: who is the president of usa?

Answer: As of now, the President of the United States is Joe Biden. He took office on January 20, 2021.
FTL: 15.410 s
TPS: 7.874 token/s

Question: q
```

更新动态bmodel后：

```
python3 pipeline.py --model_path /home/lsc/SC7/bmodels/qwen2-7b_int4_seq8192_1dev_dyn.bmodel --tokenizer_path ../support/token_config/ --devid 0 --generation_mode greedy
```

```
(sc7_venv) root@test-hpc-05:/home/lsc/SC7/LLM-TPU/models/Qwen2/python_demo# python3 pipeline.py --model_path /home/lsc/SC7/bmode
ls/qwen2-7b_int4_seq8192_1dev_dyn.bmodel --tokenizer_path ../support/token_config/ --devid 0 --generation_mode greedy
Load ../support/token_config/ ...
Device [ 0 ] loading ....
open usercpu.so, init user_cpu_init
Model[/home/lsc/SC7/bmodels/qwen2-7b_int4_seq8192_1dev_dyn.bmodel] loading ....
Done!

Load Time: 3.896 s

================================================================
1. If you want to quit, please enter one of [q, quit, exit]
2. To create a new chat session, please enter one of [clear, new]
================================================================

Question: how are you!

Answer: As an AI, I don't have feelings, but I'm here to help you. How can I assist you today?
FTL: 0.778 s
TPS: 7.908 token/s

Question: who is the president of usa?

Answer: As of my last update in October 2021, the President of the United States is Joe Biden.
FTL: 0.785 s
TPS: 7.913 token/s

Question: q
```

3.3 seq8192_2dev模型测试

7. 执行推理--CLI方式

```
# 以下命令只需第一次执行，不需要重复执行
cd /home/lsc/SC7/LLM-TPU/
git submodule update --init
cd /home/lsc/SC7/LLM-TPU/models/Qwen2/python_demo_parallel/
mkdir build
cd build && cmake .. && make -j8 && cp *cpython* .. && cd ..

# 执行多芯推理
ulimit -HSn 65536
python3 pipeline.py --model_path /home/lsc/SC7/bmodels/qwen2-7b_int4_seq8192_2dev.bmodel --tokenizer_path ../support/token_config/ --devid 0,1
```

```
(sc7_venv) root@test-hpc-05:/home/lsc/SC7/LLM-TPU/models/Qwen2/python_demo_parallel# python3 pipeline.py --model_path /h
ome/lsc/SC7/bmodels/qwen2-7b_int4_seq8192_2dev.bmodel --tokenizer_path ../support/token_config/ --devid 0,1
Load ../support/token_config/ ...
Special tokens have been added in the vocabulary, make sure the associated word embeddings are fine-tuned or trained.
Device [ 0 1 ] loading ....
open usercpu.so, init user_cpu_init
Model[/home/lsc/SC7/bmodels/qwen2-7b_int4_seq8192_2dev.bmodel] loading ....
Done!

Load Time: 5.657 s

================================================================
1. If you want to quit, please enter one of [q, quit, exit]
2. To create a new chat session, please enter one of [clear, new]
================================================================

Question: how are you!

Answer: As an AI, I don't have feelings, but I'm here and ready to assist you. How can I help you today?
FTL: 10.515 s
TPS: 14.729 token/s

Question: who is the president of usa?

Answer: As of 2023, the President of the United States is Joe Biden. He took office on January 20, 2021.
FTL: 10.475 s
TPS: 14.669 token/s

Question:
```

4芯异常按FAE临时方案修复后加载时间延长，吞吐下降，详见问题4：

1.芯片跑挂之后，需要重置寄存器，运行allreduce_reg_init这个程序
2.跑4芯模型需要对驱动稍作修改：把libbmrt.so.1.0和libbmrt.a复制到/opt/sophon/libsophon-current/lib/

```
(sc7_venv) root@test-hpc-05:/home/lsc/SC7/LLM-TPU/models/Qwen2/python_demo_parallel# python3 pipeline.py --model_path /h
ome/lsc/SC7/bmodels/qwen2-7b_int4_seq8192_2dev.bmodel --tokenizer_path ../support/token_config/ --devid 0,1
Load ../support/token_config/ ...
Device [ 0 1 ] loading ....
open usercpu.so, init user_cpu_init
Model[/home/lsc/SC7/bmodels/qwen2-7b_int4_seq8192_2dev.bmodel] loading ....
Done!

Load Time: 38.140 s

================================================================
1. If you want to quit, please enter one of [q, quit, exit]
2. To create a new chat session, please enter one of [clear, new]
================================================================

Question: how are you!

Answer: As an AI, I don't have feelings, but I'm here and ready to assist you. How can I help you today?
FTL: 9.724 s
TPS: 13.757 token/s

Question: who is the president of usa?

Answer: As of 2023, the President of the United States is Joe Biden. He took office on January 20, 2021.
FTL: 9.726 s
TPS: 14.037 token/s

Question:
```

更新动态bmodel后：

```
python3 pipeline.py --model_path /home/lsc/SC7/bmodels/qwen2-7b_int4_seq8192_2dev_dyn.bmodel --tokenizer_path ../support/token_config/ --devid 0,1
```

```
(sc7_venv) root@test-hpc-05:/home/lsc/SC7/LLM-TPU/models/Qwen2/python_demo_parallel# python3 pipeline.py --model_path /home/lsc/
SC7/bmodels/qwen2-7b_int4_seq8192_2dev_dyn.bmodel --tokenizer_path ../support/token_config/ --devid 0,1
Load ../support/token_config/ ...
Device [ 0 1 ] loading ....
open usercpu.so, init user_cpu_init
Model[/home/lsc/SC7/bmodels/qwen2-7b_int4_seq8192_2dev_dyn.bmodel] loading ....
Done!

Load Time: 4.871 s

================================================================
1. If you want to quit, please enter one of [q, quit, exit]
2. To create a new chat session, please enter one of [clear, new]
================================================================

Question: how are you!

Answer: As an AI language model, I don't have feelings or emotions like humans do, so I don't have a personal experience of bein
g happy, sad, or anything else. However, I'm here to assist you with any questions or tasks you have. How can I help you today?
FTL: 1.894 s
TPS: 13.825 token/s

Question: who is the president of usa?

Answer: As of 2023, the President of the United States is Joe Biden. He took office on January 20, 2021.
FTL: 1.902 s
TPS: 13.735 token/s

Question: q
```

3.4 seq8192_4dev模型测试

7. 执行推理--CLI方式

```
# 以下命令只需第一次执行，不需要重复执行
cd /home/lsc/SC7/LLM-TPU/
git submodule update --init
cd /home/lsc/SC7/LLM-TPU/models/Qwen2/python_demo_parallel/
cd build && cmake .. && make -j8 && cp *cpython* .. && cd ..

# 执行多芯推理
ulimit -HSn 65536
python3 pipeline.py --model_path /home/lsc/SC7/bmodels/qwen2-
7b_int4_seq8192_4dev.bmodel --tokenizer_path ../support/token_config/ --devid
0,1,2,3
```

```
(sc7_venv) root@test-hpc-05:/home/lsc/SC7/LLM-TPU# python3 pipeline.py --model_path /home/lsc/SC7/bmodels/qwen2-7b_int4_
seq8192_4dev.bmodel --tokenizer_path ../support/token_config/ --devid 0,1,2,3
python3: can't open file '/home/lsc/SC7/LLM-TPU/pipeline.py': [Errno 2] No such file or directory
(sc7_venv) root@test-hpc-05:/home/lsc/SC7/LLM-TPU# cd /home/lsc/SC7/LLM-TPU/models/Qwen2/python_demo_parallel/
(sc7_venv) root@test-hpc-05:/home/lsc/SC7/LLM-TPU/models/Qwen2/python_demo_parallel# python3 pipeline.py --model_path /h
ome/lsc/SC7/bmodels/qwen2-7b_int4_seq8192_4dev.bmodel --tokenizer_path ../support/token_config/ --devid 0,1,2,3
Load ../support/token_config/ ...
Device [ 0 1 2 3 ] loading ....
open usercpu.so, init user_cpu_init
Model[/home/lsc/SC7/bmodels/qwen2-7b_int4_seq8192_4dev.bmodel] loading ....
Done!

Load Time: 52.313 s

================================================================
1. If you want to quit, please enter one of [q, quit, exit]
2. To create a new chat session, please enter one of [clear, new]
================================================================

Question: how are you!

Answer: As an AI, I don't have feelings, but I'm here and ready to assist you. How can I help you today?
FTL: 6.979 s
TPS: 20.938 token/s

Question: who is the president of usa?

Answer: As of 2023, the President of the United States is Joe Biden. He took office on January 20, 2021.
FTL: 6.985 s
TPS: 21.335 token/s

Question:
```

更新动态bmodel后：

```
python3 pipeline.py --model_path /home/lsc/SC7/bmodels/qwen2-7b_int4_seq8192_4dev_dyn.bmodel --tokenizer_path ../support/token_config/ --devid 0,1,2,3
```

```
(sc7_venv) root@test-hpc-05:/home/lsc/SC7/LLM-TPU/models/Qwen2/python_demo_parallel# python3 pipeline.py --model_path /home/lsc/
SC7/bmodels/qwen2-7b_int4_seq8192_4dev_dyn.bmodel --tokenizer_path ../support/token_config/ --devid 0,1,2,3
Load ../support/token_config/ ...
Device [ 0 1 2 3 ] loading ....
open usercpu.so, init user_cpu_init
Model[/home/lsc/SC7/bmodels/qwen2-7b_int4_seq8192_4dev_dyn.bmodel] loading ....
Done!

Load Time: 9.253 s

================================================================
1. If you want to quit, please enter one of [q, quit, exit]
2. To create a new chat session, please enter one of [clear, new]
================================================================

Question: how are you!

Answer: As an AI, I don't have feelings, but I'm here and ready to assist you. How can I help you today?
FTL: 2.683 s
TPS: 20.567 token/s

Question: who is the president of usa?

Answer: As of 2023, the President of the United States is Joe Biden. He took office on January 20, 2021.
FTL: 2.693 s
TPS: 22.381 token/s

Question: q
```

3.5 seq8192_8dev模型测试

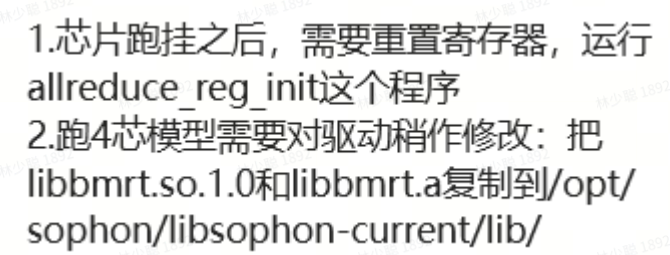
7. 执行推理--CLI方式

```
# 以下命令只需第一次执行，不需要重复执行
```

```
cd /home/lsc/SC7/LLM-TPU/
git submodule update --init
cd /home/lsc/SC7/LLM-TPU/models/Qwen2/python_demo_parallel/
cd build && cmake .. && make -j8 && cp *cpython* .. && cd ..
# 执行多芯推理
ulimit -HSn 65536 #可选
python3 pipeline.py --model_path /home/lsc/SC7/bmodels/qwen2-7b_int4_seq8192_8dev_static.bmodel --tokenizer_path ../support/token_config/ --devid 0,1,2,3,4,5,6,7
```

```
(sc7_venv) root@test-hpc-05:/home/lsc/SC7/LLM-TPU/models/Qwen2/python_demo_parallel# ulimit -HSn 65536
(sc7_venv) root@test-hpc-05:/home/lsc/SC7/LLM-TPU/models/Qwen2/python_demo_parallel# python3 pipeline.py --model_path /home/lsc/SC7
/bmodels/qwen2-7b_int4_seq8192_8dev_static.bmodel --tokenizer_path ../support/token_config/ --devid 0,1,2,3,4,5,6,7
Load ../support/token_config/ ...
Special tokens have been added in the vocabulary, make sure the associated word embeddings are fine-tuned or trained.
Device [ 0 1 2 3 4 5 6 7 ] loading ....
open usercpu.so, init user_cpu_init
Model[/home/lsc/SC7/bmodels/qwen2-7b_int4_seq8192_8dev_static.bmodel] loading ....
Done!

Load Time: 12.164 s

================================================================
1. If you want to quit, please enter one of [q, quit, exit]
2. To create a new chat session, please enter one of [clear, new]
================================================================

Question: how are you!

Answer: As an AI, I don't have feelings, but I'm here and ready to assist you. How can I help you today?
FTL: 8.007 s
TPS: 36.760 token/s

Question: who is the president of usa?

Answer: As of 2023, the President of the United States is Joe Biden. He took office on January 20, 2021.
FTL: 8.009 s
TPS: 37.019 token/s

Question:
```

4芯异常按FAE临时方案修复后加载时间延长，吞吐下降，详见问题4：



```
(sc7_venv) root@test-hpc-05:/home/lsc/SC7/LLM-TPU/models/Qwen2/python_demo_parallel# python3 pipeline.py --model_path /home/lsc/SC7/bmodels/qwen2-7b_int4_seq8192_8dev_static.bmodel --tokenizer_path ../support/token_config/ --devid 0,1,2,3,4,5,6,7
Load ../support/token_config/ ...
Device [ 0 1 2 3 4 5 6 7 ] loading ....
open usercpu.so, init user_cpu_init
Model[/home/lsc/SC7/bmodels/qwen2-7b_int4_seq8192_8dev_static.bmodel] loading ....
Done!

Load Time: 75.919 s

================================================================
1. If you want to quit, please enter one of [q, quit, exit]
2. To create a new chat session, please enter one of [clear, new]
================================================================

Question: how are you!

Answer: As an AI, I don't have feelings, but I'm here and ready to assist you. How can I help you today?
FTL: 5.793 s
TPS: 25.109 token/s

Question: who is the president of usa?

Answer: As of 2023, the President of the United States is Joe Biden. He took office on January 20, 2021.
FTL: 5.818 s
TPS: 25.955 token/s

Question: q
(sc7_venv) root@test-hpc-05:/home/lsc/SC7/LLM-TPU/models/Qwen2/python_demo_parallel# ulimit -HSn 65536
```

更新动态bmodel后：

```
python3 pipeline.py --model_path /home/lsc/SC7/bmodels/qwen2-7b_int4_seq8192_8dev_dyn.bmodel --tokenizer_path ../support/token_config/ --devid 0,1,2,3,4,5,6,7
```

3.6 性能测试数据汇总

	Output tokens_per_second (TPS)(token/s)	latency_per_token（TPOT）(ms)	first_token (TTFT)(ms)	Benchmark duration(s)
1. dev_num = 1 2. input_len = 8192 3. INT4	7.907	NA	15413	NA
1. dev_num = 2 2. input_len = 8192 3. INT4	14.669	NA	10475	NA
1. dev_num = 4	21.335	NA	6985	NA

2. input_len = 8192 3. INT4				
1. dev_num = 8 2. input_len = 8192 3. INT4	37.019	NA	8009	NA

4. Qwen2-14b VLLM测试（在算能工作站上测试）

4.1 基于vLLM在线推理功能测试

1. 进入容器，并启动一个基于 OpenAI 的 API 服务器。

```
# 进入容器
./docker_run.sh

# FP16
python3 -m vllm.entrypoints.openai.api_server \
        --model /workspace/qwen14b-bmodel/config \
        --device 'auto' \
        --host 0.0.0.0 \
        --trust-remote-code \
        --port 8080 \
        --enforce-eager
```

2. 打开另外一个终端，发送客户端请求。

注：下面IP需要替换为对应服务器IP

```
# FP16
curl http://172.18.97.235:8080/v1/completions \
-H "Content-Type: application/json" \
-d '{
"model": "/workspace/qwen14b-bmodel/config",
"prompt": "如何制作月饼",
"max_tokens": 256,
"temperature": 0.01
}'

```

```
sn@workstation:~/SC7/SC7$ curl http://172.18.97.235:8080/v1/completions \
-H "Content-Type: application/json" \
-d '{
"model": "/workspace/qwen14b-bmodel/config",
"prompt": "如何制作月饼",
"max_tokens": 256,
"temperature": 0.01
}'
{"id":"cmpl-496bdd64f23b4129a6521c67d17af165","object":"text_completion","created":90349,"model":"/workspace/qwen14b-bmodel/config","choices":[{"index":0,"text":"\n\n1. 准备好材料：转化糖浆100克、枧水2克、花生油30克、中筋面粉160克、咸蛋黄10个、红豆沙200克、莲蓉200克、月饼模具1个。\n\n2. 转化糖浆、枧水、花生油倒入盆中，搅拌均匀。\n\n3. 筛入中筋面粉，搅拌均匀。\n\n4. 揉成光滑的面团，盖上保鲜膜，静置1小时。\n\n5. 咸蛋黄喷上白酒，放入烤箱，180度烤5分钟。\n\n6. 红豆沙、莲蓉分成20份，每份约30克。\n\n7. 取一份红豆沙，包入一份莲蓉，包成团。\n\n8. 取一份面团，压扁，包入一份馅料，包成团。\n\n9. 月饼模具内撒上一层薄薄的面粉，放入月饼，压出花纹。\n\n10. 放入预热好的烤箱，180度烤5分钟，取出刷一层蛋黄液，再烤15分钟，取出晾","logprobs":null,"finish_reason":"length"}],"usage":{"prompt_tokens":3,"total_tokens":259,"completion_tokens":256}}sn@w
```

4.2 基于vLLM在线推理功能测试（自定义数据集）（!!! 测试未通过 !!!）

1. 用以下文件替换/home/sn/SC7/SC7/vllm/benchmarks/benchmark_serving.py文件

benchmark_serving.py

2. 继续沿用上面容器,启动服务端。

```
# 启动server端服务
cd inference_scripts/
./run_openai_api_server_gn.sh
```

附录：run_openai_api_server_gn.sh

run_openai_api_server_gn.sh

3. 打开另一个终端，并进入同个容器，启动客户端发送请求。

```
# 进入容器
./docker_run.sh

# 执行client请求
cd inference_scripts/
./run_openai_api_client_gn.sh
```

附录：run_openai_api_client_gn.sh

run_openai_api_client_gn.sh

input.json

4. 测试结果如下：

```
INFO:     127.0.0.1:57778 - "POST /v1/completions HTTP/1.1" 400 Bad Request
INFO:     127.0.0.1:57790 - "POST /v1/completions HTTP/1.1" 400 Bad Request
INFO:     127.0.0.1:57792 - "POST /v1/completions HTTP/1.1" 400 Bad Request
INFO:     127.0.0.1:57808 - "POST /v1/completions HTTP/1.1" 400 Bad Request
INFO:     127.0.0.1:57814 - "POST /v1/completions HTTP/1.1" 400 Bad Request
INFO:     127.0.0.1:57828 - "POST /v1/completions HTTP/1.1" 400 Bad Request
INFO:     127.0.0.1:57834 - "POST /v1/completions HTTP/1.1" 400 Bad Request
INFO:     127.0.0.1:57746 - "POST /v1/completions HTTP/1.1" 400 Bad Request
INFO:     127.0.0.1:57748 - "POST /v1/completions HTTP/1.1" 400 Bad Request
INFO:     127.0.0.1:57762 - "POST /v1/completions HTTP/1.1" 400 Bad Request
```

```
Namespace(backend='vllm', base_url=None, host='0.0.0.0', port=12345, endpoint='/v1/completions', dataset='/workspace/dataset/input.json', dataset_name='sharegpt', dataset_path=None, model='/workspace/qwen14b-bmodel/config', tokenizer='/workspace/qwen14b-bmodel/config', best_of=1, use_beam_search=False, num_prompts=10, sharegpt_output_len=4000, sonnet_input_len=550, sonnet_output_len=150, sonnet_prefix_len=200, request_rate=inf, seed=0, trust_remote_code=True, disable_tqdm=False, save_result=False, metadata=None, result_dir=None)
WARNING 09-13 09:17:26 tokenizer.py:64] Using a slow tokenizer. This might cause a significant slowdown. Consider using a fast tokenizer instead.
/workspace/vllm/benchmarks/benchmark_serving.py:644: UserWarning: The '--dataset' argument will be deprecated in the next release. Please use '--dataset-name' and '--dataset-path' in the future runs.
  main(args)
prompt_len:  3076
prompt_len:  4280
prompt_len:  4115
prompt_len:  4211
prompt_len:  2656
prompt_len:  3657
prompt_len:  3274
prompt_len:  4375
prompt_len:  2208
prompt_len:  3217
Traffic request rate: inf
100%|██████████| 10/10 [00:00<00:00, 63.68it/s]
============ Serving Benchmark Result ============
Successful requests:                     0
Benchmark duration (s):                  0.16
Total input tokens:                      0
Total generated tokens:                  0
Request throughput (req/s):              0.00
Input token throughput (tok/s):          0.00
Output token throughput (tok/s):         0.00
Output throughput is zero, cannot calculate latency per token.
---------------Time to First Token----------------
Mean TTFT (ms):                          0.00
Median TTFT (ms):                        0.00
P99 TTFT (ms):                           0.00
-----Time per Output Token (excl. 1st token)------
Mean TPOT (ms):                          0.00
Median TPOT (ms):                        0.00
P99 TPOT (ms):                           0.00
==================================================
root@workstation:/workspace/inference_scripts#
```

4.3 基于vLLM在线推理性能测试

1. 用以下文件替换/home/sn/SC7/SC7/vllm/benchmarks/benchmark_serving.py文件

benchmark_serving.py (备注：跟3.3节文件不一致)

2. 继续沿用上面容器,启动服务端。

```
# 启动server端服务
cd inference_scripts/
./run_openai_api_server_xn.sh
```

附录：run_openai_api_server_xn.sh

run_openai_api_server_xn.sh

3. 打开另一个终端，并进入同个容器，启动客户端发送请求。

```
# 进入容器
./docker_run.sh

# 执行client请求
cd inference_scripts
./run_openai_api_client_xn.sh
```

附录：run_openai_api_client_xn.sh

run_openai_api_client_xn.sh

4. 测试结果截图及记录：

a. batch size = 1；input_len = 2048；output_len = 256; FP16

```
Namespace(backend='vllm', base_url=None, host='localhost', port=12345, endpoint='/v1/completions', dataset='/workspace/dataset/ShareGPT_V3_unfiltered_cleaned_split.json', dataset_name='sharegpt', dataset_path=None, model='/workspace/qwen14b-bmodel/config', tokenizer='/workspace/qwen14b-bmodel/config', best_of=1, use_beam_search=False, num_prompts=1, sharegpt_input_len=2048, sharegpt_output_len=256, sonnet_input_len=550, sonnet_output_len=150, sonnet_prefix_len=200, request_rate=inf, seed=0, trust_remote_code=True, disable_tqdm=False, save_result=False, metadata=None, result_dir=None)
WARNING 09-13 10:44:01 tokenizer.py:64] Using a slow tokenizer. This might cause a significant slowdown. Consider using a fast tokenizer instead.
/workspace/vllm/benchmarks/benchmark_serving.py:634: UserWarning: The '--dataset' argument will be deprecated in the next release. Please use '--dataset-name' and '--dataset-path' in the future runs.
  main(args)
first 2217
Traffic request rate: inf
100%|          | 1/1 [01:17<00:00, 77.02s/it]
256
============ Serving Benchmark Result ============
Successful requests:                     1
Benchmark duration (s):                  77.02
Total input tokens:                      2048
Total generated tokens:                  256
Request throughput (req/s):              0.01
Input token throughput (tok/s):          26.59
Output token throughput (tok/s):         3.32
latency per token (ms):                  300.86
--------------Time to First Token----------------
Mean TTFT (ms):                          11034.38
Median TTFT (ms):                        11034.38
P99 TTFT (ms):                           11034.38
-----Time per Output Token (excl. 1st token)------
Mean TPOT (ms):                          258.76
Median TPOT (ms):                        258.76
P99 TPOT (ms):                           258.76
==================================================
```

b. batch size = 2；input_len = 2048；output_len = 256; FP16

```
Namespace(backend='vllm', base_url=None, host='localhost', port=12345, endpoint='/v1/completions', dataset='/workspace/dataset/ShareGPT_V3_unfiltered_cleaned_split.json', dataset_name='sharegpt', dataset_path=None, model='/workspace/qwen14b-bmodel/config', tokenizer='/workspace/qwen14b-bmodel/config', best_of=1, use_beam_search=False, num_prompts=2, sharegpt_input_len=2048, sharegpt_output_len=256, sonnet_input_len=550, sonnet_output_len=150, sonnet_prefix_len=200, request_rate=inf, seed=0, trust_remote_code=True, disable_tqdm=False, save_result=False, metadata=None, result_dir=None)
WARNING 09-13 10:47:34 tokenizer.py:64] Using a slow tokenizer. This might cause a significant slowdown. Consider using a fast tokenizer instead.
/workspace/vllm/benchmarks/benchmark_serving.py:634: UserWarning: The '--dataset' argument will be deprecated in the next release. Please use '--dataset-name' and '--dataset-path' in the future runs.
  main(args)
first 2217
first 2219
Traffic request rate: inf
100%|██████████| 2/2 [02:37<00:00, 78.84s/it]^[[C
256
256
============ Serving Benchmark Result ============
Successful requests:                     2
Benchmark duration (s):                  157.68
Total input tokens:                      4096
Total generated tokens:                  512
Request throughput (req/s):              0.01
Input token throughput (tok/s):          25.98
Output token throughput (tok/s):         3.25
latency per token (ms):                  615.94
---------------Time to First Token----------------
Mean TTFT (ms):                          51042.91
Median TTFT (ms):                        51042.91
P99 TTFT (ms):                           89446.74
-----Time per Output Token (excl. 1st token)------
Mean TPOT (ms):                          262.05
Median TPOT (ms):                        262.05
P99 TPOT (ms):                           264.45
==================================================
```

c. batch size = 4； input_len = 2048； output_len = 256; FP16

```
ataset/ShareGPT_V3_unfiltered_cleaned_split.json', dataset_name='sharegpt', dataset_path=None, model='/workspace/qwen14b-bmodel/config', tokenizer='/workspace/qwen14b-bmodel/config', best_of=1, use_beam_search=False, num_prompts=4, sharegpt_input_len=2048, sharegpt_output_len=256, sonnet_input_len=550, sonnet_output_len=150, sonnet_prefix_len=200, request_rate=inf, seed=0, trust_remote_code=True, disable_tqdm=False, save_result=False, metadata=None, result_dir=None)
WARNING 09-13 11:12:41 tokenizer.py:64] Using a slow tokenizer. This might cause a significant slowdown. Consider using a fast tokenizer instead.
/workspace/vllm/benchmarks/benchmark_serving.py:634: UserWarning: The '--dataset' argument will be deprecated in the next release. Please use '--dataset-name' and '--dataset-path' in the future runs.
  main(args)
Token indices sequence length is longer than the specified maximum sequence length for this model (20464 > 8192). Running this sequence through the model will result in indexing errors
first 2217
first 2219
first 2269
first 2181
Traffic request rate: inf
100%|██████████| 4/4 [05:32<00:00, 83.12s/it]
256
256
256
256
============ Serving Benchmark Result ============
Successful requests:                     4
Benchmark duration (s):                  332.50
Total input tokens:                      8192
Total generated tokens:                  1024
Request throughput (req/s):              0.01
Input token throughput (tok/s):          24.64
Output token throughput (tok/s):         3.08
latency per token (ms):                  1298.82
---------------Time to First Token----------------
Mean TTFT (ms):                          136481.77
Median TTFT (ms):                        136334.92
P99 TTFT (ms):                           258531.71
-----Time per Output Token (excl. 1st token)------
Mean TPOT (ms):                          276.85
Median TPOT (ms):                        276.44
P99 TPOT (ms):                           280.20
==================================================
```

d. batch size = 8； input_len = 2048； output_len = 256; FP16

```
Token indices sequence length is longer than the specified maximum sequence length for this model (20464 > 8192). Runnin
g this sequence through the model will result in indexing errors
first 2217
first 2219
first 2269
first 2181
first 2229
first 2247
first 2151
first 2175
Traffic request rate: inf
100%|          | 8/8 [11:45<00:00, 88.16s/it]
256
256
256
256
256
256
256
256
============ Serving Benchmark Result ============
Successful requests:                     8
Benchmark duration (s):                  705.25
Total input tokens:                      16384
Total generated tokens:                  2048
Request throughput (req/s):              0.01
Input token throughput (tok/s):          23.23
Output token throughput (tok/s):         2.90
latency per token (ms):                  2754.89
---------------Time to First Token----------------
Mean TTFT (ms):                          317657.95
Median TTFT (ms):                        316387.09
P99 TTFT (ms):                           619996.15
-----Time per Output Token (excl. 1st token)------
Mean TPOT (ms):                          295.85
Median TPOT (ms):                        295.28
P99 TPOT (ms):                           309.24
==================================================
```

4.4 基于VLLM性能测试数据汇总

	Output tokens_per_second (TPS)(tok/s)	latency_per_token（TPOT）(ms)	first_token (TTFT)(ms)	Benchmark duration(s)
1. batch size = 1 2. input_len = 2048 3. output_len = 256 4. FP16	3.32	258.76	11034.38	77.02
1. batch size = 2 2. input_len = 2048 3. output_len = 256 4. FP16	3.25	262.05	51042.91	157.68
1. batch size = 4 2. input_len = 2048	3.08	276.85	136481.77	332.50

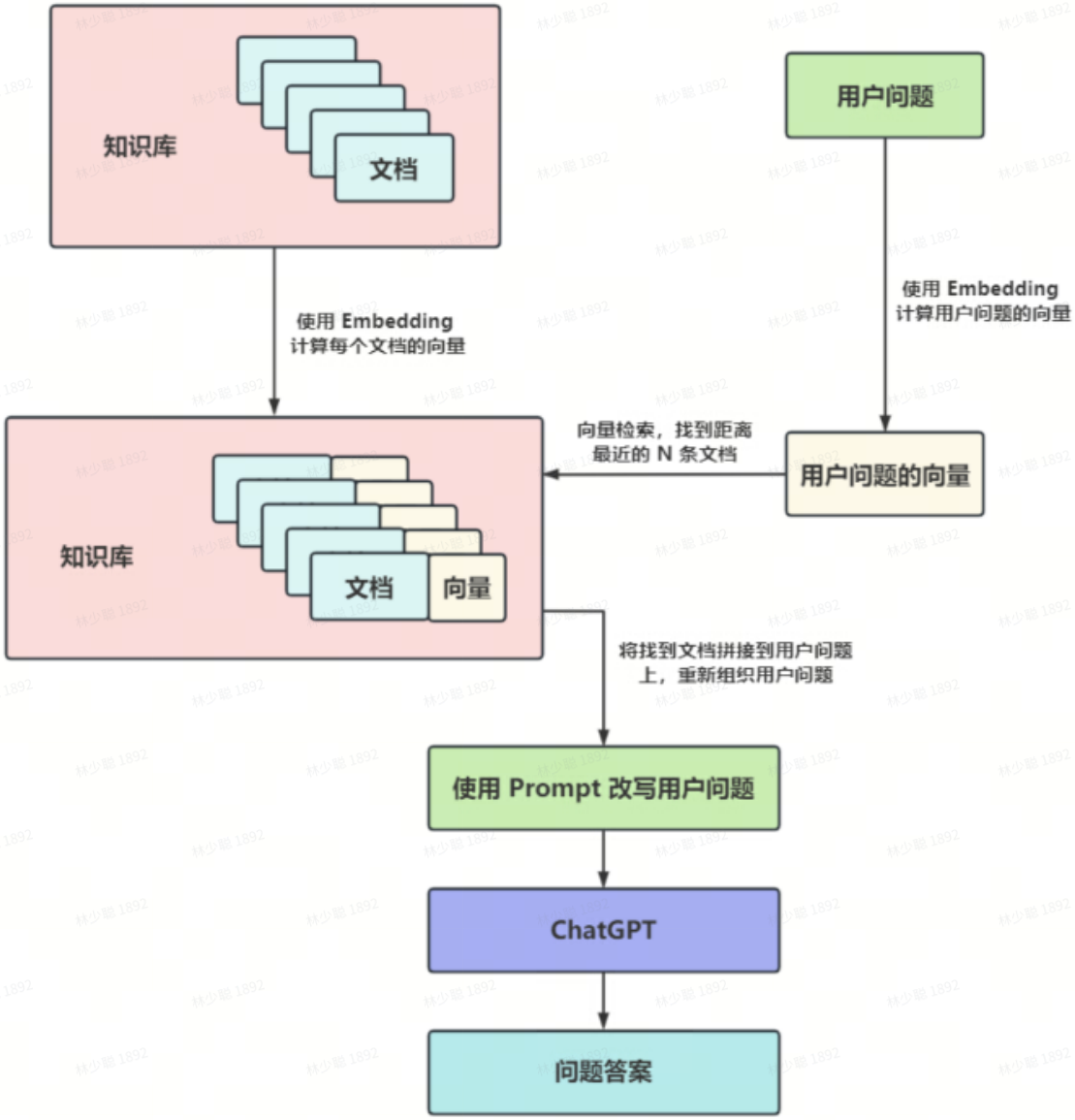
3. output_len = 256 4. FP16				
1. batch size = 8 2. input_len = 2048 3. output_len = 256 4. FP16	2.90	295.85	317657.95	705.25

5. 基于bge_large、bge_reranker部署的文档对话类RAG项目ChatDoc-TPU

5.1 介绍

该项目的主要目标是通过使用自然语言来简化与文档的交互，并提取有价值的信息。此项目使用LangChain、ChatGLM3-TPU或QWEN-TPU构建，以向用户提供流畅自然的对话体验。

以 ChatGPT 为例（可替换为其他LLM，本仓库已支持 Chatglm3-6B 和 Qwen-7B，需要保证接口一致），本地知识库问答流程如下：



5.2 特点

- 完全本地推理。
- 支持多种文档格式PDF, DOCX, TXT。
- 与文档内容进行聊天，提出问题根据文档获得相关答案。
- 用户友好的界面，确保流畅的交互。

5.3 环境安装

```
# 项目代码clone
git clone https://github.com/JackeyTakumi/ChatDoc-TPU.git

# 安装第三方库
cd ChatDoc-TPU
# 考虑到 langchain 和 sail 版本依赖，推荐在 python>=3.8 环境运行
sudo apt update
sudo apt install libgl1-mesa-glx libcairo2-dev
```

```
pip3 install -r requirements.txt -i https://pypi.tuna.tsinghua.edu.cn/simple
&& pip3 uninstall torchvision

# 安装sail
# 2.1节从算能官网下载的SDK包含sophon-sail_3.8.0.tar.gz安装包
# 下载SOPHON-SAIL源码,解压后进入其源码目录
cd /home/lsc/SC7/SDK/SDK-24.04.01/sophon-sail_20240606_085400/sophon-sail/
# 创建编译文件夹build,并进入build文件夹
mkdir build && cd build
# 执行编译命令
cmake -DONLY_RUNTIME=ON ..
make pysail
# 打包生成python wheel,生成的wheel包的路径为‘python/pcie/dist’,文件名为‘sophon-
3.8.0-py3-none-any.whl’
cd ../python/pcie
chmod +x sophon_pcie_whl.sh
./sophon_pcie_whl.sh
# 安装python wheel
pip3 install ./dist/sophon-3.8.0-py3-none-any.whl --force-reinstall
```

5.4 启动

```
# 回到ChatDoc-TPU主目录，启动程序，模型和配置文件自动下载，使用默认路径
./run.sh --dev_id 0
```

```
usage: ./run.sh  [--dev_id DEV_ID] [--server_address SERVER_ADDRESS] [--
server_port SERVER_PORT] [--chip CHIP]
--dev_id: 用于推理的 TPU 设备 ID。默认为 0。
--server_address: web server 地址。默认为 "0.0.0.0"。
--server_port: web sever 端口。如不设置，从 8501 起自动分配。
--chip: 需要下载模型对应的芯片类型。如不设置，默认为bm1684x。
```

config.ini

```ini
[bert_model]
bmodel_path = ./models/bert_model/bge_large_512_fp16.bmodel
token_path = ./models/bert_model/token_config

[reranker_model]
bmodel_path = ./models/reranker_model/bge_reranker_512_fp16.bmodel
token_path = ./models/reranker_model/token_config

[init_config]
base_url = http://127.0.0.1:18080/v1/
supported_model = qwen72b,qwen7b,chatglm3,chatglm3_BM1688,qwen1.5_BM1688
```

启动后您可以通过浏览器打开，URL: http://{host_ip}:8501，host_ip为启动ChatDoc的设备IP，或者您通过参数设置的server_address。

```
(sc7_venv) root@test-hpc-05:/home/lsc/SC7/ChatDoc-TPU# ./run.sh --dev_id 1
++ which unzip
+ res=/usr/bin/unzip
+ '[' 0 '!=' 0 ']'
+ pip3 install dfss -i https://pypi.tuna.tsinghua.edu.cn/simple --upgrade
Looking in indexes: https://pypi.tuna.tsinghua.edu.cn/simple
Requirement already satisfied: dfss in /home/lsc/SC7/sc7_venv/lib/python3.10/site-packages (1.7.11)
+ llm_model=qwen7b
+ dev_id=0
+ server_address=0.0.0.0
+ server_port=
+ chip=bm1684x
+ parse_args --dev_id 1
+ [[ 2 -gt 0 ]]
+ key=--dev_id
+ case $key in
+ dev_id=1
+ shift 2
+ [[ 0 -gt 0 ]]
+ '[' '!' -d /root/nltk_data ']'
+ echo '/root/nltk_dat already exist...'
/root/nltk_dat already exist...
+ '[' '!' -d ./models/bert_model ']'
+ echo 'bert_model already exist...'
bert_model already exist...
+ '[' '!' -d ./models/reranker_model ']'
+ echo 'reranker_model already exist...'
reranker_model already exist...
+ export LLM_MODEL=qwen7b
+ LLM_MODEL=qwen7b
+ export DEVICE_ID=1
+ DEVICE_ID=1
+ '[' '' == '' ']'
+ streamlit run web_demo_st.py --server.address 0.0.0.0

  You can now view your Streamlit app in your browser.

  URL: http://0.0.0.0:8501
```

5.5 启动chatglm3大模型接口服务

```bash
#下载chatglm3项目包openai_api_demo.zip并解压
cd /home/lsc/SC7/openai_api_demo/
# 下载模型和token_config
./scripts/download.sh --chip bm1684x
```

```
# 安装依赖
pip3 install -r requirements.txt
# 运行程序
python3 api_server.py --model chatglm3
```

```
(sc7_venv) root@test-hpc-05:/home/lsc/SC7/openai_api_demo# python3 api_server.py --model chatglm3
open usercpu.so, init user_cpu_init
INFO:     Started server process [873537]
INFO:     Waiting for application startup.
INFO:     Application startup complete.
INFO:     Uvicorn running on http://0.0.0.0:18080 (Press CTRL+C to quit)
INFO:     127.0.0.1:51856 - "POST /v1/chat/completions HTTP/1.1" 200 OK
INFO:     127.0.0.1:60778 - "POST /v1/chat/completions HTTP/1.1" 200 OK
INFO:     127.0.0.1:50270 - "POST /v1/chat/completions HTTP/1.1" 200 OK
INFO:     127.0.0.1:46158 - "POST /v1/chat/completions HTTP/1.1" 200 OK
INFO:     127.0.0.1:49650 - "POST /v1/chat/completions HTTP/1.1" 200 OK
INFO:     127.0.0.1:56820 - "POST /v1/chat/completions HTTP/1.1" 200 OK
INFO:     127.0.0.1:40198 - "POST /v1/chat/completions HTTP/1.1" 200 OK
INFO:     127.0.0.1:53918 - "POST /v1/chat/completions HTTP/1.1" 200 OK
INFO:     127.0.0.1:57364 - "POST /v1/chat/completions HTTP/1.1" 200 OK
INFO:     127.0.0.1:36416 - "POST /v1/chat/completions HTTP/1.1" 200 OK
INFO:     127.0.0.1:43480 - "POST /v1/chat/completions HTTP/1.1" 200 OK
^CINFO:     Shutting down
INFO:     Waiting for application shutdown.
INFO:     Application shutdown complete.
INFO:     Finished server process [873537]
(sc7_venv) root@test-hpc-05:/home/lsc/SC7/openai_api_demo# python3 api_server.py --model chatglm3
open usercpu.so, init user_cpu_init
INFO:     Started server process [930039]
INFO:     Waiting for application startup.
INFO:     Application startup complete.
INFO:     Uvicorn running on http://0.0.0.0:18080 (Press CTRL+C to quit)
INFO:     127.0.0.1:49914 - "POST /v1/chat/completions HTTP/1.1" 200 OK
```

5.6 操作说明

1. 点击Browse files选择要上传的文档，然后点击添加知识库。随后将embedding文档，完成后文档将被选中，接着就可开始对话。我们可重复上传文档，embedding成功的文档均会进入知识库。

6. 问题记录

6.1 问题1：安装libncurses5依赖库时报错。

1. 问题描述：

```
root@test-hpc-05:/home/lsc/SC7# sudo apt install dkms libncurses5
Reading package lists... Done
Building dependency tree... Done
Reading state information... Done
dkms is already the newest version (2.8.7-2ubuntu2.2).
You might want to run 'apt --fix-broken install' to correct these.
The following packages have unmet dependencies:
 libhwloc-plugins : Depends: libxnvctrl0 but it is not going to be installed
E: Unmet dependencies. Try 'apt --fix-broken install' with no packages (or specify a solution).
root@test-hpc-05:/home/lsc/SC7#
```

2. 问题解决

卸载NV相关的包，dpkg -l | grep nvidia和sudo apt remove [包名称]。

6.2 问题2：克隆 Git 子模块时遇到了连接问题：git submodule update --init报错。

1. 问题描述：

```
(sc7_venv) root@test-hpc-05:/home/lsc/SC7/LLM-TPU# git submodule update --init
Cloning into '/home/lsc/SC7/LLM-TPU/models/Qwen/demo_parallel/third_party/abseil-cpp'...
fatal: unable to access 'https://github.com/abseil/abseil-cpp.git/': Failed to connect to github.com port 443 after 130808 ms: Conn
ection timed out
fatal: clone of 'https://github.com/abseil/abseil-cpp.git' into submodule path '/home/lsc/SC7/LLM-TPU/models/Qwen/demo_parallel/thi
rd_party/abseil-cpp' failed
Failed to clone 'models/Qwen/demo_parallel/third_party/abseil-cpp'. Retry scheduled
Cloning into '/home/lsc/SC7/LLM-TPU/models/Qwen/demo_parallel/third_party/re2'...
Cloning into '/home/lsc/SC7/LLM-TPU/models/Qwen/demo_parallel/third_party/abseil-cpp'...
fatal: unable to access 'https://github.com/abseil/abseil-cpp.git/': Failed to connect to github.com port 443 after 129709 ms: Conn
ection timed out
fatal: clone of 'https://github.com/abseil/abseil-cpp.git' into submodule path '/home/lsc/SC7/LLM-TPU/models/Qwen/demo_parallel/thi
rd_party/abseil-cpp' failed
Failed to clone 'models/Qwen/demo_parallel/third_party/abseil-cpp' a second time, aborting
(sc7_venv) root@test-hpc-05:/home/lsc/SC7/LLM-TPU# git submodule update --init
Cloning into '/home/lsc/SC7/LLM-TPU/models/Qwen/demo_parallel/third_party/abseil-cpp'...
Submodule path 'models/Qwen/demo_parallel/third_party/abseil-cpp': checked out '08b21bd037990c18d44fda1691211e73835bf214'
Submodule path 'models/Qwen/demo_parallel/third_party/re2': checked out 'b84e3ff189980a33d4a0c6fa1201aa0b3b8bab4a'
```

2. 问题解决

重试几遍解决。

6.3 问题3：执行多芯推理报内存分配错误

1. 问题描述：

```
(sc7_venv) root@test-hpc-05:/home/lsc/SC7/LLM-TPU/models/Qwen2/python_demo_parallel# python3 pipeline.py --model_path /home/lsc/SC7
/bmodels/qwen2-7b_int4_seq8192_8dev_static.bmodel --tokenizer_path ../support/token_config/ --devid 0,1,2,3,4,5,6,7
Load ../support/token_config/ ...
Special tokens have been added in the vocabulary, make sure the associated word embeddings are fine-tuned or trained.
Device [ 0 1 2 3 4 5 6 7 ] loading ....
open usercpu.so, init user_cpu_init
Model[/home/lsc/SC7/bmodels/qwen2-7b_int4_seq8192_8dev_static.bmodel] loading ....
[bmlib_memory][error] bm_alloc_gmem failed, dev_id = 6, size = 0x80
[BM_CHECK][error] BM_CHECK_RET fail ../bmlib/src/bmlib_memory.cpp: bm_malloc_device_byte_heap_mask: 1142
[BMRT][must_alloc_device_mem:2897] FATAL:device mem alloc failed: size=32[0x20] type_len=4 status=5 desc=bd_cmd_mem
python3: /home/lsc/SC7/LLM-TPU/models/Qwen2/python_demo_parallel/chat.cpp:109: void Qwen::init(const std::vector<int>&, const strin
g&): Assertion `true == ret' failed.
Aborted (core dumped)
```

2. 问题解决

模型加载过程中无缘无故突然中断，可以运行ulimit -HSn 65536增加系统资源解决。

6.4 问题4：执行4芯推理，程序执行一半被强制终止

1. 问题描述：

```
(sc7_venv) root@test-hpc-05:/home/lsc/SC7/LLM-TPU/models/Qwen2/python_demo_parallel# python3 pipeline.py --model_path /h
ome/lsc/SC7/bmodels/qwen2-7b_int4_seq8192_4dev.bmodel --tokenizer_path ../support/token_config/ --devid 2,3,4,6
Load ../support/token_config/ ...
Special tokens have been added in the vocabulary, make sure the associated word embeddings are fine-tuned or trained.
Device [ 2 3 4 6 ] loading ....
open usercpu.so, init user_cpu_init
Model[/home/lsc/SC7/bmodels/qwen2-7b_int4_seq8192_4dev.bmodel] loading ....
Done!

Load Time: 7.479 s

================================================================
1. If you want to quit, please enter one of [q, quit, exit]
2. To create a new chat session, please enter one of [clear, new]
================================================================

Question: hello

Answer: python3: /home/lsc/SC7/LLM-TPU/models/Qwen2/python_demo_parallel/chat.cpp:81: void Qwen::net_launch(const string
&, std::vector<bm_tensor_s>&, std::vector<bm_tensor_s>&, int): Assertion `ret' failed.
Aborted (core dumped)
```

查看bm-smi,发现2,3,4,5芯片Fault异常。

```
Mon Sep  9 16:43:35 2024
+-----------------------------------------------------------------------------------------+
| Lib Version:    0.5.1              Driver Version:  0.5.1                               |
+-----------------------------------+-----------------------------------------------------+
|card  Name      Mode        SN     |TPU  boardT  chipT   TPU_P  TPU_V  ECC  CorrectN  Tpu-Util|
|12V_ATX  MaxP boardP Minclk Maxclk  Fan|Bus-ID      Status   Curclk   TPU_C   Memory-Usage      |
|===================================+=====================================================|
| 0  SC7-224T    PCIE HQDZW59BDJEJD0078 | 0    34C     43C    5.1W   835mV  OFF    N/A       100%  |
|8648mA  300W  103W   25M     875M   N/A|000:3d:00.0 Fault     875M      6.1A    81MB/14787MB      |
|                                   |-----------------------------------------------------|
|                                   | 1    44C     46C    3.5W   843mV  OFF    N/A       100%  |
|                                   |000:3e:00.0 Fault     875M      4.2A    81MB/14787MB      |
|                                   |-----------------------------------------------------|
|                                   | 2    46C     48C    3.5W   843mV  OFF    N/A       100%  |
|                                   |000:3f:00.0 Fault     875M      4.2A    81MB/14787MB      |
|                                   |-----------------------------------------------------|
|                                   | 3    41C     45C    5.1W   836mV  OFF    N/A       100%  |
|                                   |000:40:00.0 Fault     875M      6.1A    81MB/14787MB      |
|                                   |-----------------------------------------------------|
|                                   | 4    43C     48C    3.5W   843mV  OFF    N/A       100%  |
|                                   |000:41:00.0 Fault     875M      4.1A    81MB/14787MB      |
|                                   |-----------------------------------------------------|
|                                   | 5    49C     48C    3.5W   843mV  OFF    N/A       100%  |
|                                   |000:42:00.0 Fault     875M      4.2A    81MB/14787MB      |
|                                   |-----------------------------------------------------|
|                                   | 6    45C     49C    5.2W   835mV  OFF    N/A       100%  |
|                                   |000:43:00.0 Fault     875M      6.2A    81MB/14787MB      |
|                                   |-----------------------------------------------------|
|                                   | 7    50C     48C    5.3W   835mV  OFF    N/A       100%  |
|                                   |000:44:00.0 Fault     875M      6.3A    81MB/14787MB      |
+===================================+=====================================================+

+-----------------------------------------------------------------------------------------+
| Processes:                                                                  TPU Memory  |
|  TPU-ID       PID   Process name                                            Usage       |
|=========================================================================================|
```

2. 问题回复：

本地测试0,1,2,3也不行。按照最新回复解决,不过吞吐性能会下降，加载时间也多了很多（详见第3节测试记录）。

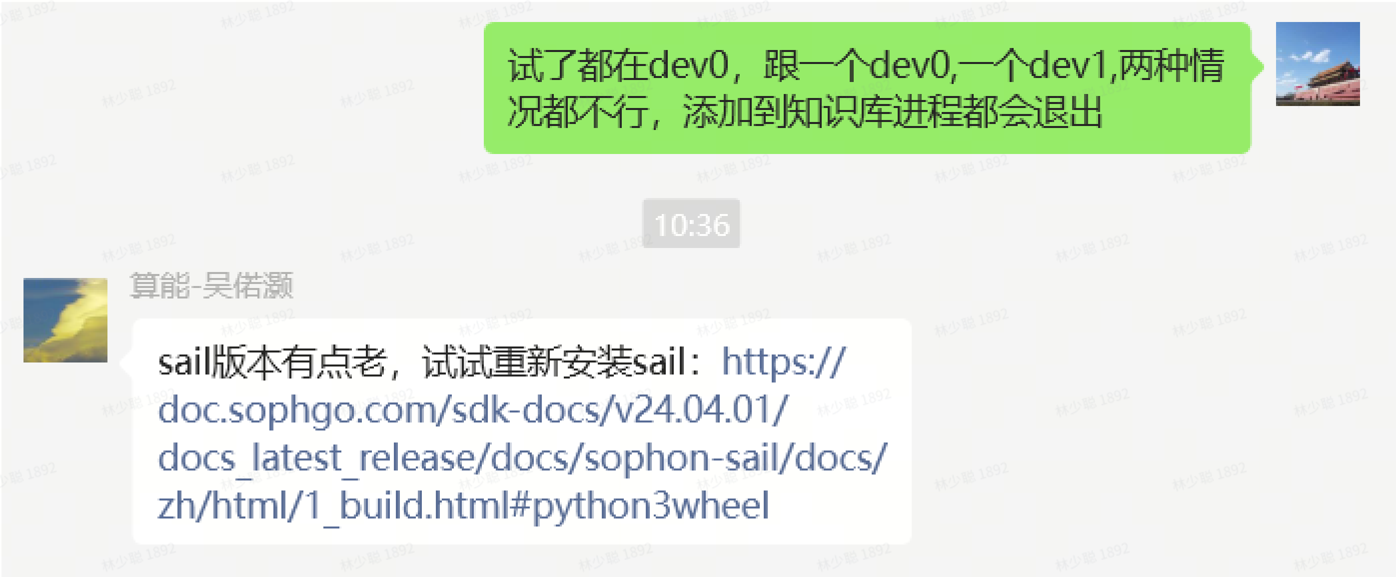
6.5 问题5：添加知识库时，报内存访问错误

1. 问题描述：

```
uploaded_file: [UploadedFile(id=1, name='现代化网络-技术洞察报告.docx', type='application/vnd.openxmlformats-officedocument.wordprocessingml.document', size=31000779)]
##############################
add_repository:  True
RapidOCRDocLoader  block index: 1292: 100%|████████████████████| 1293/1293 [02:15<00:00,  9.54it/s]
No section: 'qwen7b'
文档切分示例: page_content='现代化网络-技术洞察报告 1.' metadata={'source': './data/uploaded/现代化网络-技术洞察报告.docx'}
  0%|                                                            | 0/191 [00:00<?, ?it/s]
[2024-09-11 10:04:27.711] [error] [engine.cpp:1246] start alloc tensor!
./run.sh: line 119: 817891 Segmentation fault      (core dumped) streamlit run web_demo_st.py --server.address $server_address
```

2. 问题解决：

安装新版本sail解决。



6.6 问题6：自定义数据集进行推理，报400 bad Request

1. 问题描述：

```
INFO:     127.0.0.1:57778 - "POST /v1/completions HTTP/1.1" 400 Bad Request
INFO:     127.0.0.1:57790 - "POST /v1/completions HTTP/1.1" 400 Bad Request
INFO:     127.0.0.1:57792 - "POST /v1/completions HTTP/1.1" 400 Bad Request
INFO:     127.0.0.1:57808 - "POST /v1/completions HTTP/1.1" 400 Bad Request
INFO:     127.0.0.1:57814 - "POST /v1/completions HTTP/1.1" 400 Bad Request
INFO:     127.0.0.1:57828 - "POST /v1/completions HTTP/1.1" 400 Bad Request
INFO:     127.0.0.1:57834 - "POST /v1/completions HTTP/1.1" 400 Bad Request
INFO:     127.0.0.1:57746 - "POST /v1/completions HTTP/1.1" 400 Bad Request
INFO:     127.0.0.1:57748 - "POST /v1/completions HTTP/1.1" 400 Bad Request
INFO:     127.0.0.1:57762 - "POST /v1/completions HTTP/1.1" 400 Bad Request
```