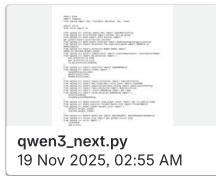


command

- docker: **rocm/ali-private:ubuntu22.04_rocm7.0.1_sglang_cda5676_20251030**
- model
 - [🇨🇳通义千问3-Next-80B-A3B-Instruct](#)
 - [🇨🇳通义千问3-Next-80B-A3B-Instruct-FP8](#)
- python/sglang/srt/models/qwen3_next.py文档修改



- server:
- FP16

```
1 export SGLANG_USE_AITER=1
2
3 TP=4
4 EP=1
5
6 echo "launching ${model}"
7 echo "TP=${TP}"
8 echo "EP=${EP}"
9
10 python3 -m sglang.launch_server \
11     --model-path /data/models/Qwen/Qwen3-Next-80B-A3B-Instruct \
12     --host localhost \
13     --port XXXX \
14     --tp-size ${TP} \
15     --ep-size ${EP} \
16     --trust-remote-code \
17     --chunked-prefill-size 32768 \
18     --mem-fraction-static 0.85 \
19     --disable-radix-cache \
20     --max-prefill-tokens 32768 \
21     --cuda-graph-max-bs 256 \
22     --page-size 64 \
23     --attention-backend triton \
24     --max-running-requests 128 \
```

- FP8

```
1 export SGLANG_USE_AITER=1
2
3
4 TP=4
5 EP=1
6
```

```

7 echo "launching ${model}"
8 echo "TP=${TP}"
9 echo "EP=${EP}"
10
11 python3 -m sglang.launch_server \
12     --model-path /data/models/Qwen/Qwen3-Next-80B-A3B-Instruct-FP8 \
13     --host localhost \
14     --port XXXX \
15     --tp-size ${TP} \
16     --ep-size ${EP} \
17     --trust-remote-code \
18     --chunked-prefill-size 32768 \
19     --mem-fraction-static 0.85 \
20     --disable-radix-cache \
21     --max-prefill-tokens 32768 \
22     --cuda-graph-max-bs 128 \
23     --max-running-requests 128 \
24     --page-size 64 \
25     --attention-backend triton \

```

- 安装kunlun-benchmark

- [kunlun-benchmark.tar.gz](#)
- [KUNLUN-BENCHMARK_使用教程.md](#)

- clinet

- ```

1 max_concurrency=300
2 num_prompts=$((10 * max_concurrency))
3
4
5
6 /mnt/md0/yixiongh/Qwen3_next/kunlun/kunlun-benchmark/kunlun-benchmark sglang server \
7 --port 8080 \
8 --work_mode manual \
9 --max_input_len 1000 \
10 --min_input_len 800 \
11 --max_output_len 500 \
12 --min_output_len 400 \
13 --concurrency ${max_concurrency} \
14 --query_num ${num_prompts} \
15 --result_dir /mnt/md0/yixiongh/Qwen3_next/client/result \
16 --model_path /data/models/Qwen/Qwen3-Next-80B-A3B-Instruct-FP8 \
17 --is_sla True \
18 --sla_decode 50 \

```