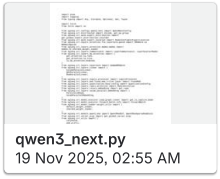
command

- 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

```
export SGLANG_USE_AITER=1

TP=4
EP=1

echo "launching ${model}"
echo "TP=${TP}"
echo "EP=${EP}"

python3 -m sglang.launch_server \
    --model-path /data/models/Qwen/Qwen3-Next-80B-A3B-Instruct \
    --host localhost \
    --port XXXX \
    --tp-size ${TP} \
    --ep-size ${EP} \
    --trust-remote-code \
    --chunked-prefill-size 32768 \
    --mem-fraction-static 0.85 \
    --disable-radix-cache \
    --max-prefill-tokens 32768 \
    --cuda-graph-max-bs 256 \
    --page-size 64 \
    --attention-backend triton \
    --max-running-requests 128 \
```

- FP8

```
export SGLANG_USE_AITER=1


TP=4
EP=1

```

```
echo "launching ${model}"
echo "TP=${TP}"
echo "EP=${EP}"

python3 -m sglang.launch_server \
    --model-path /data/models/Qwen/Qwen3-Next-80B-A3B-Instruct-FP8 \
    --host localhost \
    --port XXXX \
    --tp-size ${TP} \
    --ep-size ${EP} \
    --trust-remote-code \
    --chunked-prefill-size 32768 \
    --mem-fraction-static 0.85 \
    --disable-radix-cache \
    --max-prefill-tokens 32768 \
    --cuda-graph-max-bs 128 \
    --max-running-requests 128 \
    --page-size 64 \
    --attention-backend triton \
```

- 安装kunlun-benchmark
  - kunlun-benchmark.tar.gz
  - KUNLUN-BENCHMARK_使用教程.md
- clinet

```
 max_concurrency=300
num_prompts=$((10 * max_concurrency))



/mnt/md0/yixiongh/Qwen3_next/kunlun/kunlun-benchmark/kunlun-benchmark sglang server \
    --port 8080 \
    --work_mode manual \
    --max_input_len 1000 \
    --min_input_len 800 \
    --max_output_len 500 \
    --min_output_len 400  \
    --concurrency ${max_concurrency} \
    --query_num ${num_prompts} \
    --result_dir /mnt/md0/yixiongh/Qwen3_next/client/result \
    --model_path /data/models/Qwen/Qwen3-Next-80B-A3B-Instruct-FP8 \
    --is_sla True \
    --sla_decode 50 \
```