

A script to check speed(token/sec) of QWEN:

```
import time
from transformers import AutoModelForCausalLM, AutoTokenizer

model_name = "Qwen/Qwen2.5-0.5B-Instruct" # CPU-friendly small Qwen

tokenizer = AutoTokenizer.from_pretrained(model_name)
model = AutoModelForCausalLM.from_pretrained(model_name)

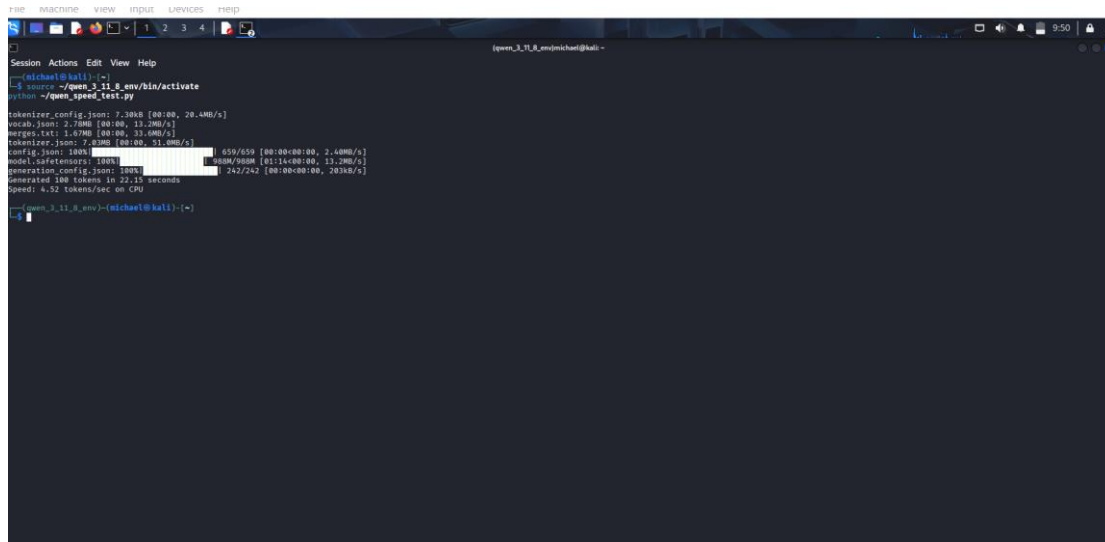
inputs = tokenizer(prompt, return_tensors="pt")

_ = model.generate(**inputs, max_new_tokens=10) # warm-up

start = time.time()
outputs = model.generate(**inputs, max_new_tokens=100)
end = time.time()

tokens_generated = outputs.shape[-1] - inputs["input_ids"].shape[-1]
time_taken = end - start
speed = tokens_generated / time_taken

print(f"Generated {tokens_generated} tokens in {time_taken:.2f} seconds")
print(f"Speed: {speed:.2f} tokens/sec on CPU")
```



```
Session Actions Edit View Help
[michael@kali:~]$ source ~/Qwen_3.11.8_env/bin/activate
[michael@kali:~]$ python ~/Qwen_speed_test.py

tokenizer_config.json: 7.38kB [00:00, 28.4MB/s]
vocab.json: 2.78MB [00:00, 13.2MB/s]
merges.txt: 1.47MB [00:00, 33.2MB/s]
tokenizer.json: 7.83MB [00:00, 51.0MB/s]
config.json: 1.04kB [00:00, 619.6KB/s]
model.safetensors: 100% [00:14<00:00, 2.40MB/s]
generation_config.json: 1.00kB [00:00, 988B/s]
Generated 100 tokens in 22.15 seconds
Speed: 4.52 tokens/sec on CPU

[michael@kali:~]$
```