# LLM-Inference-Bench Tool

# LLM-Inference-Bench  Tool

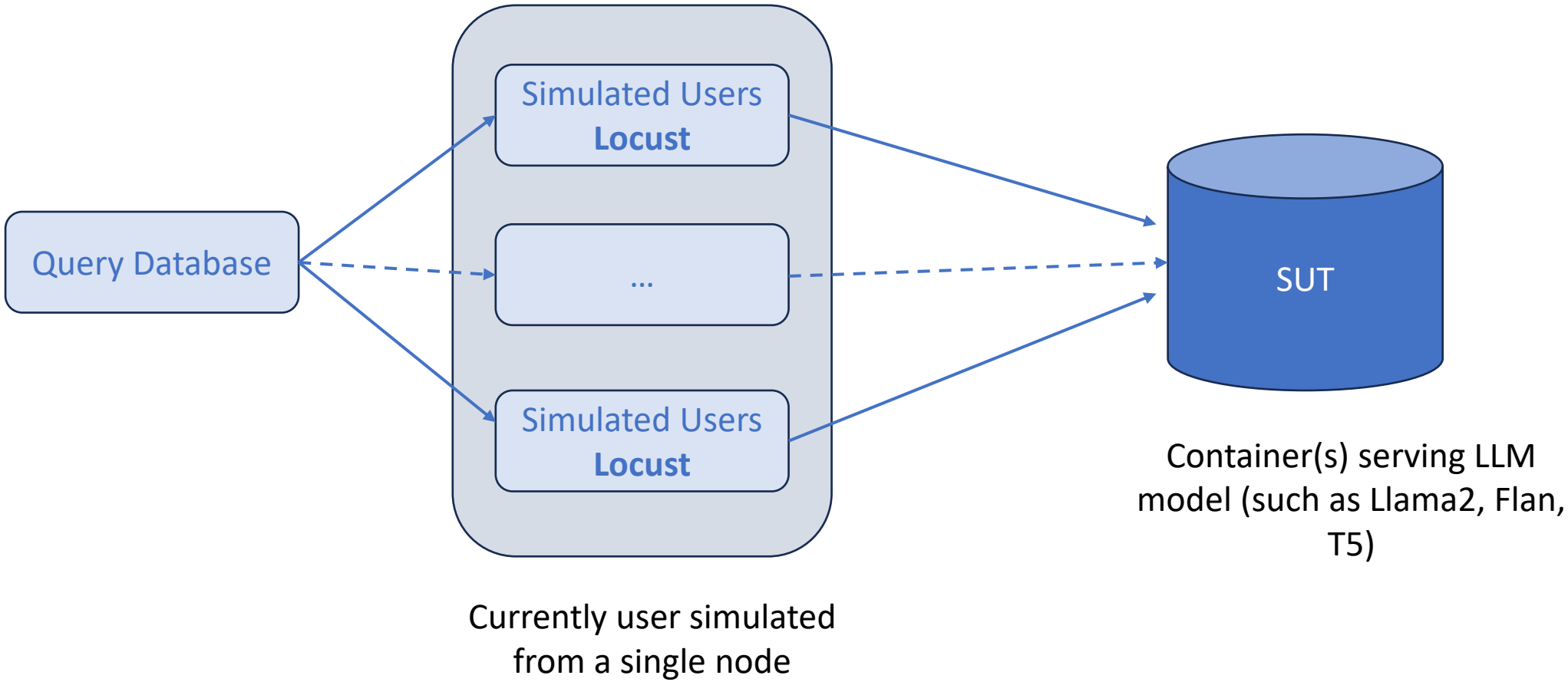
Goal of the Inferencing benchmark tool is to identify

- Latency of each request made and measured in millisecond/token
- Latency of the Time Taken for the First Token (TTFT) – higher the latency drastically affects user experience
- Throughput measured in number of tokens/second

This is measured with varying sized of

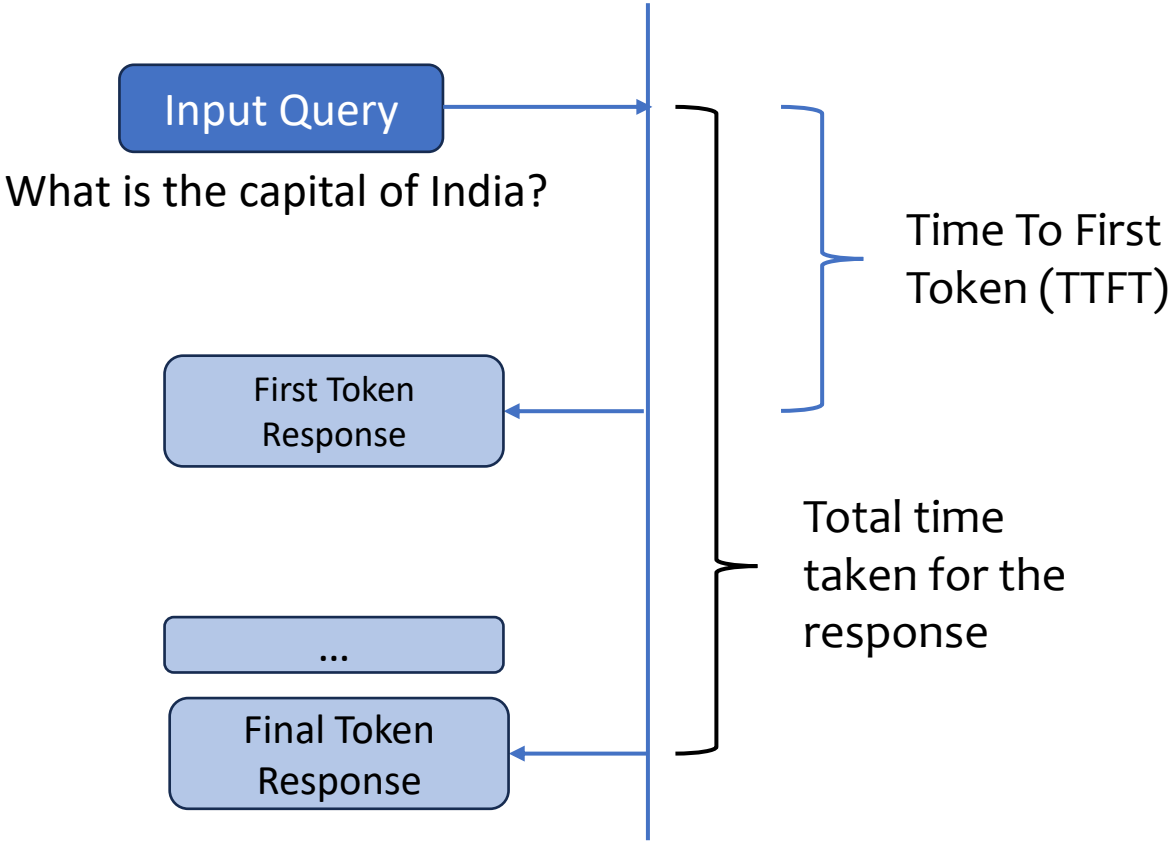
- Input Tokens (query length)
- Output Tokens (response length)
- Simulating parallel users

# LLM-Inference-Bench Tool



# Latency, Throughput & TTFT Calculations

- TTFT – Time To First Token
- Token Latency– Time taken for all tokens excluding the first
- Throughput - Tokens / second – Total number of tokens / total time



# Running the Benchmark

# Generating Input Dataset:

**Dataset** : HuggingFace Dataset “pvduy/sharegpt_alpaca_oa_vicuna_format”

Steps involved in dataset generation

- HuggingFace dataset consists of nearly 324k prompts
- Using LLaMA tokenizer prompts are tokenized to identify input query tokens for each of the prompts
- 7 files (.csv) of 1000 queries are built for input query sizes of 32, 64, 128, 256, 512, 1024(1k), 2048(2k)

# Benchmark Input Query

dataset_filtering.py

Process Hugging Face Query Dataset to generate

- 7 files with 1000 queries each for input token 32, 64, 128, 256, 512, 1k, 2k
- Input token size is not exact and is usually plus/minus 10 tokens the expected input token, example, file with 128 input tokens has queries with input token size 118 – 138

# LLM-Inference-Benchmark Script

llm_inference_benchmark.sh

- llm_inference_master.py: Run Inference benchmark on the llm
  a) for increasing parallel users ,
  b) run the benchmark for different combinations of Input tokens, Output tokens to obtain (latency, performance, TTFT) and
  c) write output in a different directory to csv

- llm_result_analysis.py: *Analyze all different csv output files stored in a directory for each combination to obtain*
  - Chart TTFT/Latency/Performance
  - Identify peak performance
  - Chart CPU/Memory for each user test
  - Chart TTFT/Latency/Performance