- `python3 -m sglang.bench\_serving --backend sglang --dataset-name random --num-prompts 100 --random-input 256 --random-output 256 --random-range-ratio 0.5`

- `python3 -m sglang.bench\_serving --backend sglang --dataset-name random --num-prompts 100 --random-input 1024 --random-output 512`



- `python3 -m sglang.bench\_serving --backend sglang --dataset-name random --num-prompts 100 --random-input 128 --random-output 128`



compared to the grafana visual metrics：





- `python3 -m sglang.bench\_serving --backend sglang --dataset-name random --num-prompts 100 --random-input 128 --random-output 128`



- `python3 -m sglang.bench\_serving --backend sglang --dataset-name random --num-prompts 100 --random-input 128 --random-output 128`



