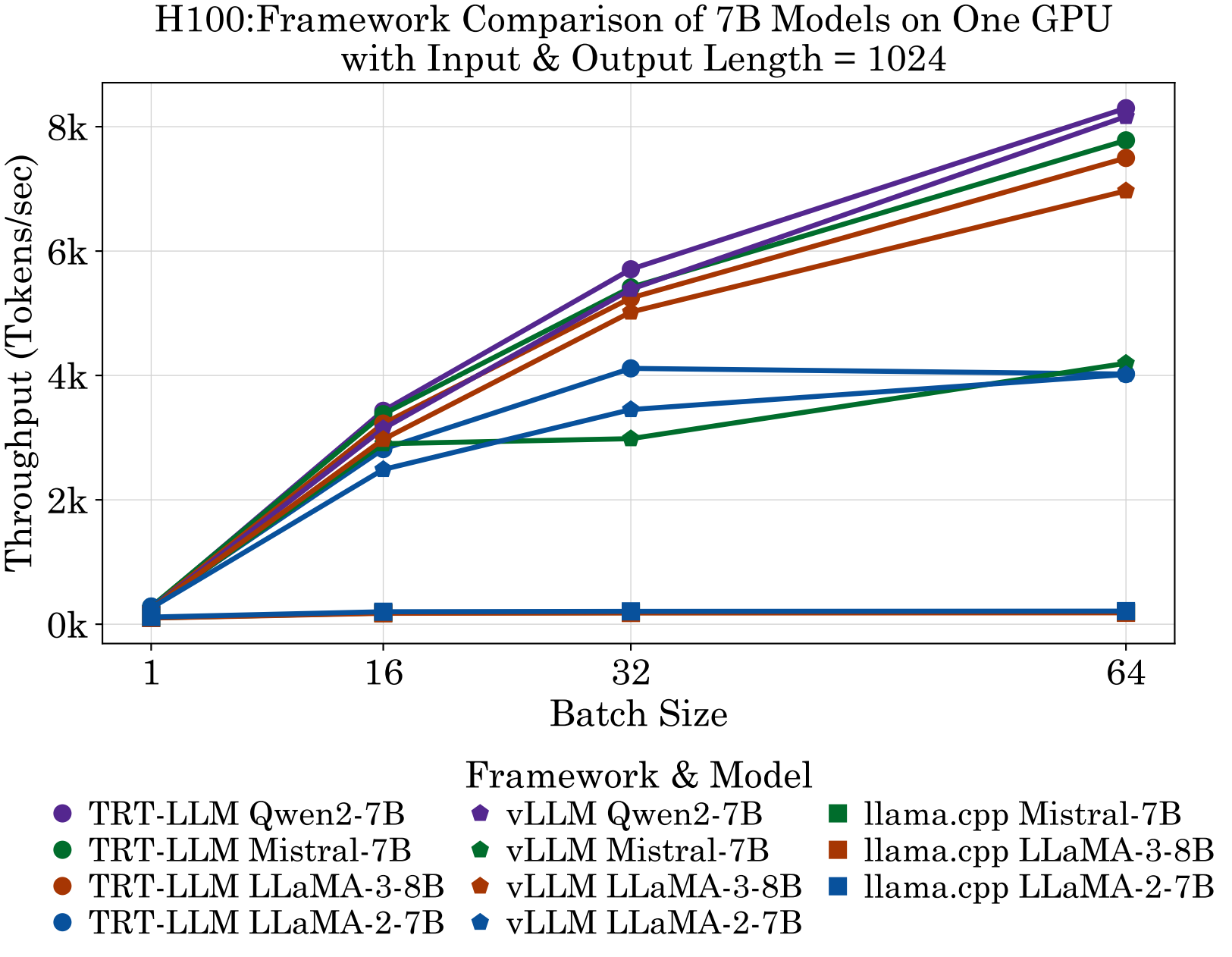
H100:Framework Comparison of 7B Models on One GPU
with Input & Output Length = 1024
Throughput (Tokens/sec)
8k
6k
4k
2k
0k
1
16
32
64
Batch Size
Framework & Model
TRT-LLM Qwen2-7B
TRT-LLM Mistral-7B
TRT-LLM LLaMA-3-8B
TRT-LLM LLaMA-2-7B
vLLM Qwen2-7B
vLLM Mistral-7B
vLLM LLaMA-3-8B
vLLM LLaMA-2-7B
llama.cpp Mistral-7B
llama.cpp LLaMA-3-8B
llama.cpp LLaMA-2-7B