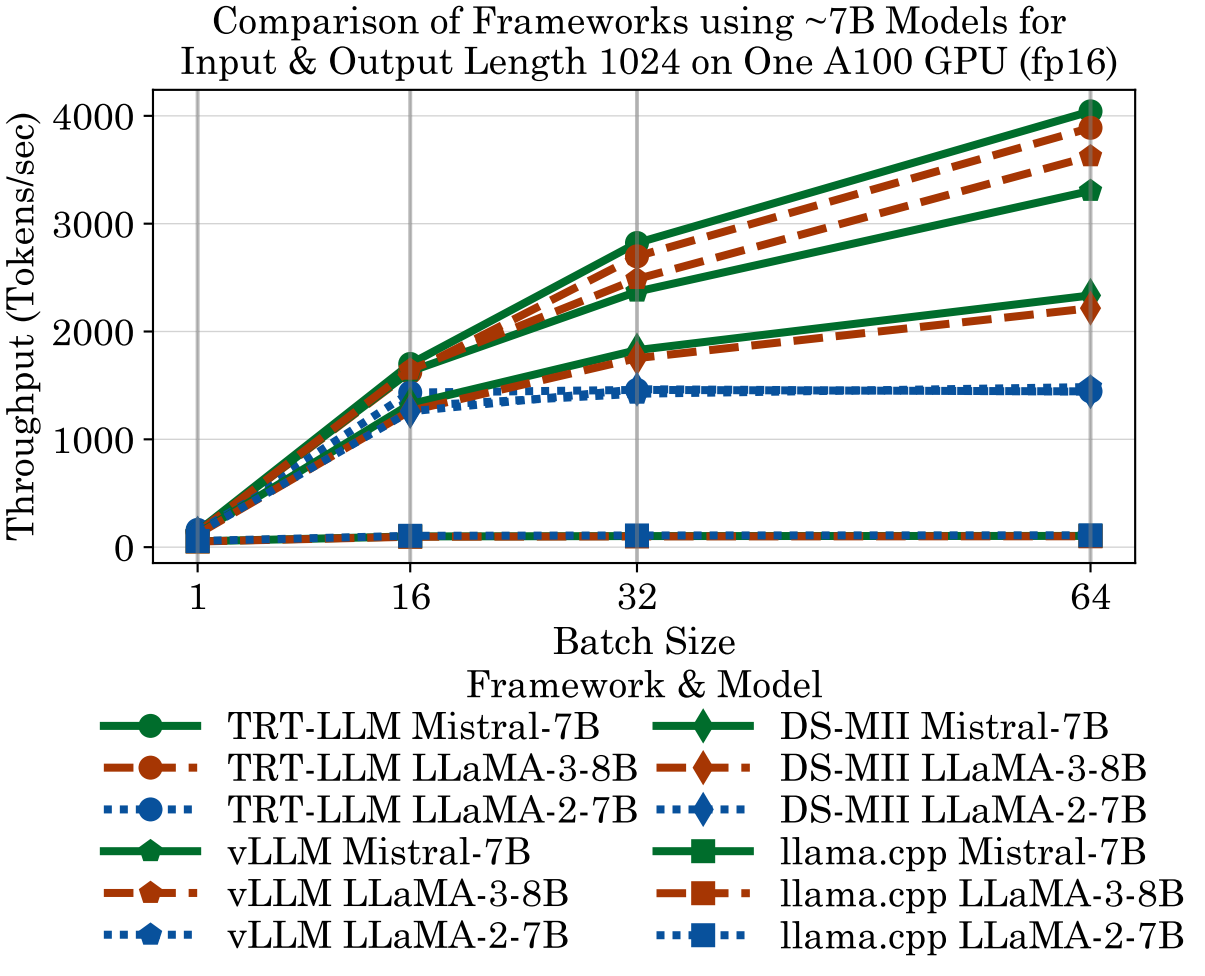
Comparison of Frameworks using ~7B Models for
Input & Output Length 1024 on One A100 GPU (fp16)
Throughput (Tokens/sec)
4000
3000
2000
1000
0
1
16
32
64
Batch Size
Framework & Model
TRT-LLM Mistral-7B
TRT-LLM LLaMA-3-8B
TRT-LLM LLaMA-2-7B
vLLM Mistral-7B
vLLM LLaMA-3-8B
vLLM LLaMA-2-7B
DS-MII Mistral-7B
DS-MII LLaMA-3-8B
DS-MII LLaMA-2-7B
llama.cpp Mistral-7B
llama.cpp LLaMA-3-8B
llama.cpp LLaMA-2-7B