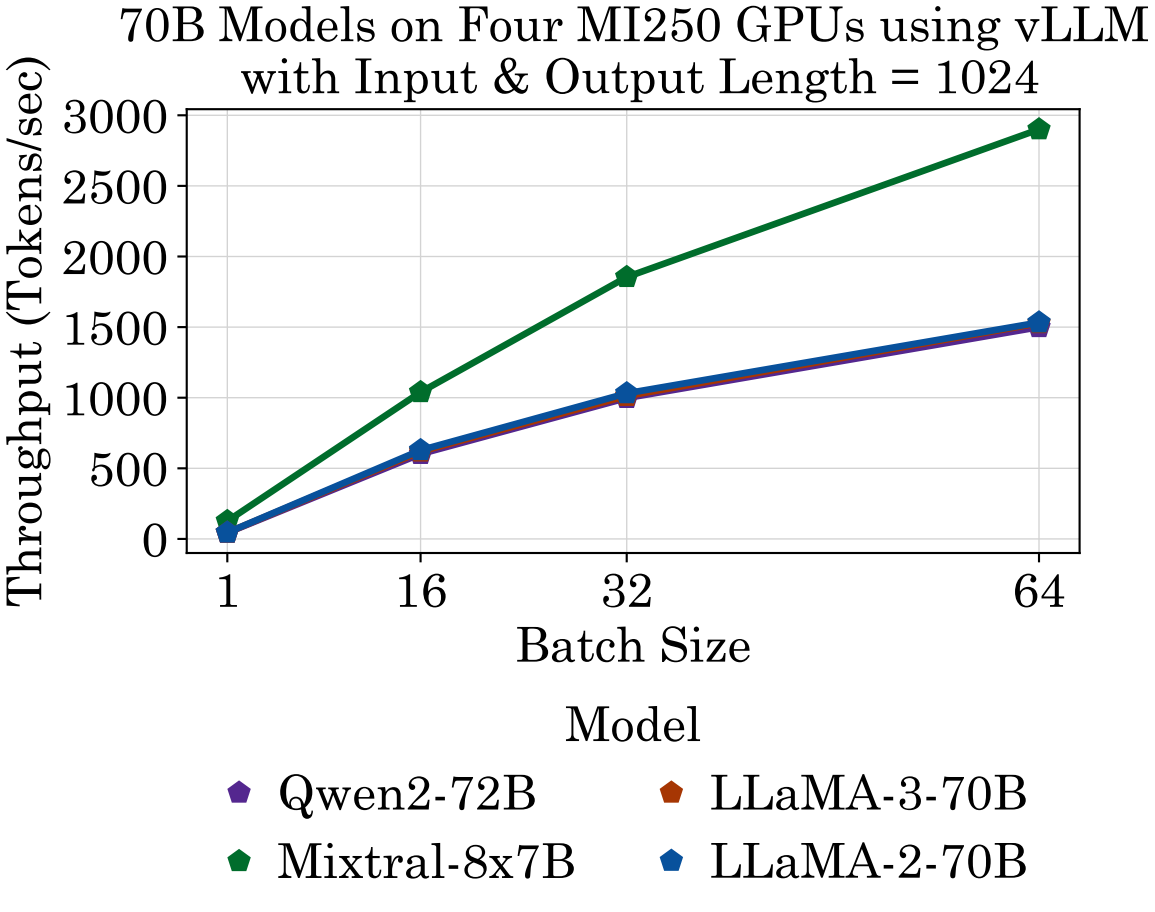
70B Models on Four MI250 GPUs using vLLM
with Input & Output Length = 1024
Throughput (Tokens/sec)
3000
2500
2000
1500
1000
500
0
1
16
32
64
Batch Size
Model
Qwen2-72B
LLaMA-3-70B
Mixtral-8x7B
LLaMA-2-70B