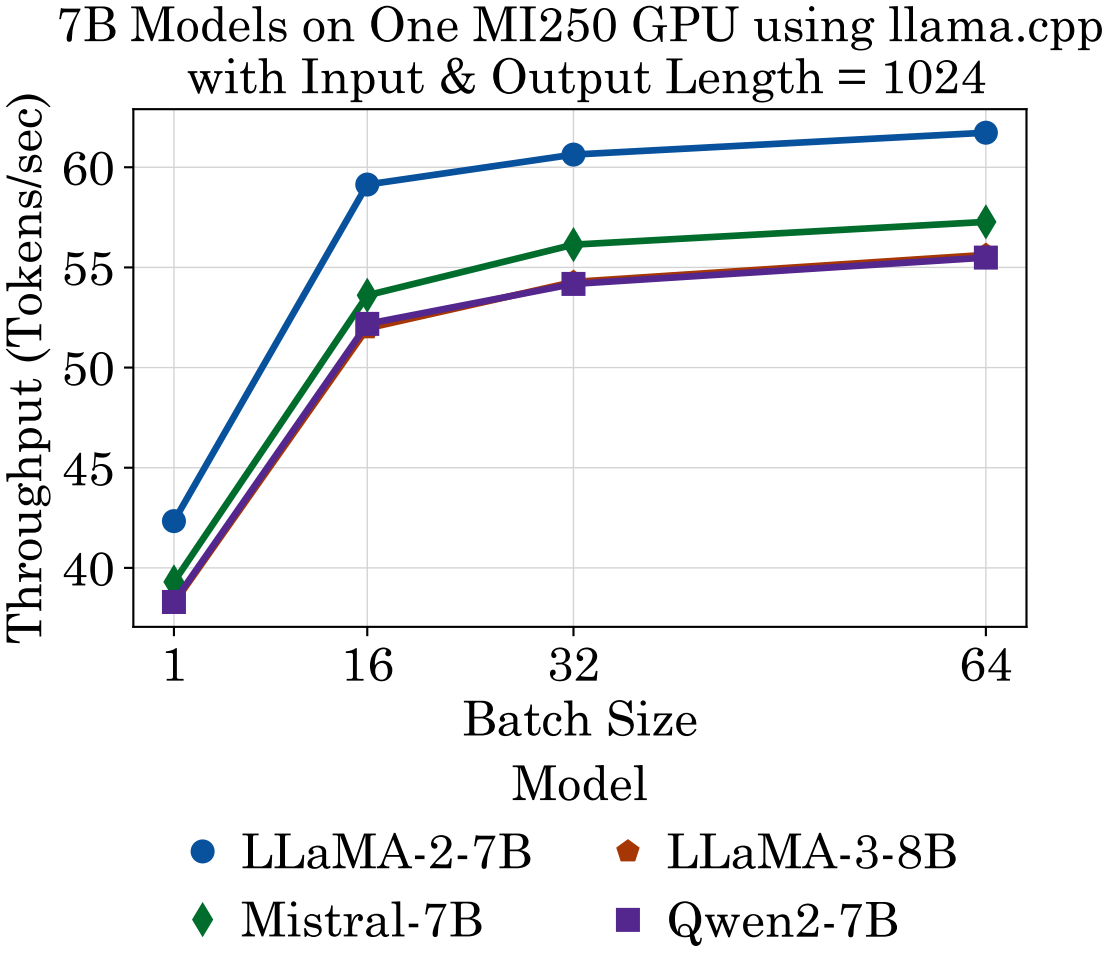
7B Models on One MI250 GPU using llama.cpp
with Input & Output Length = 1024
Throughput (Tokens/sec)
60
55
50
45
40
1
16
32
64
Batch Size
Model
LLaMA-2-7B
LLaMA-3-8B
Mistral-7B
Qwen2-7B