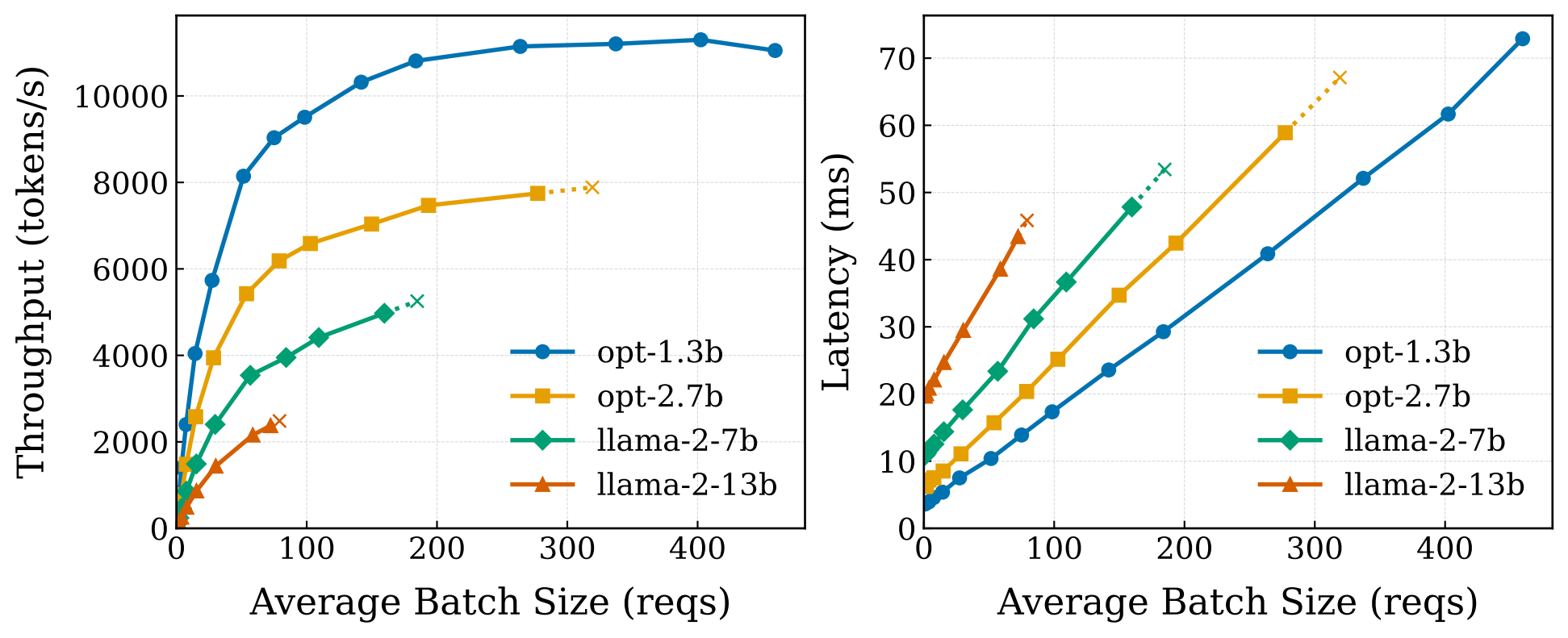
Throughput (tokens/s)
0
2000
4000
6000
8000
10000
0
100
200
300
400
Average Batch Size (reqs)
opt-1.3b
opt-2.7b
llama-2-7b
llama-2-13b
Latency (ms)
0
10
20
30
40
50
60
70
0
100
200
300
400
Average Batch Size (reqs)
opt-1.3b
opt-2.7b
llama-2-7b
llama-2-13b