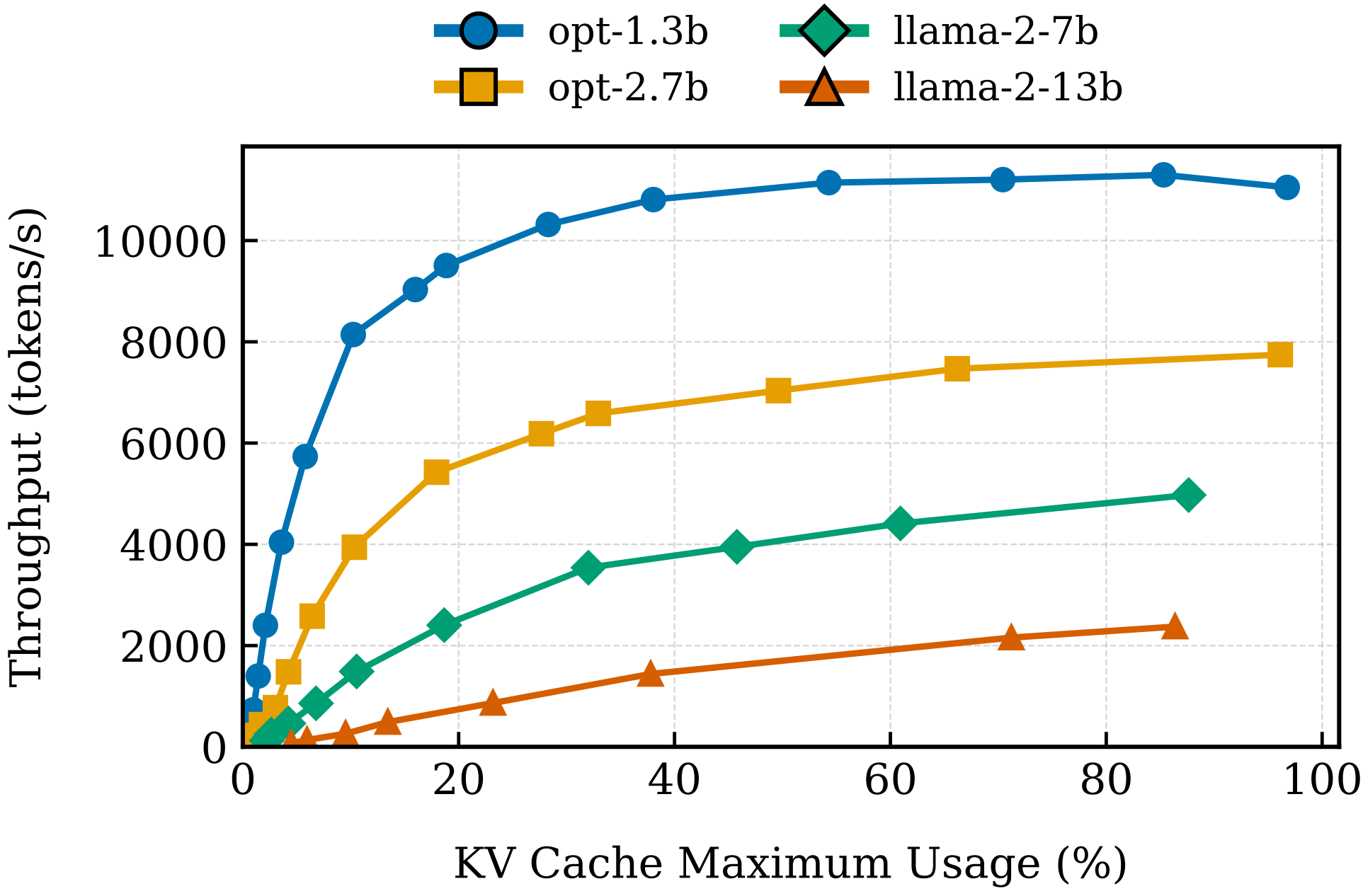
opt-1.3b
opt-2.7b
llama-2-7b
llama-2-13b
Throughput (tokens/s)
0
2000
4000
6000
8000
10000
0
20
40
60
80
100
KV Cache Maximum Usage (%)