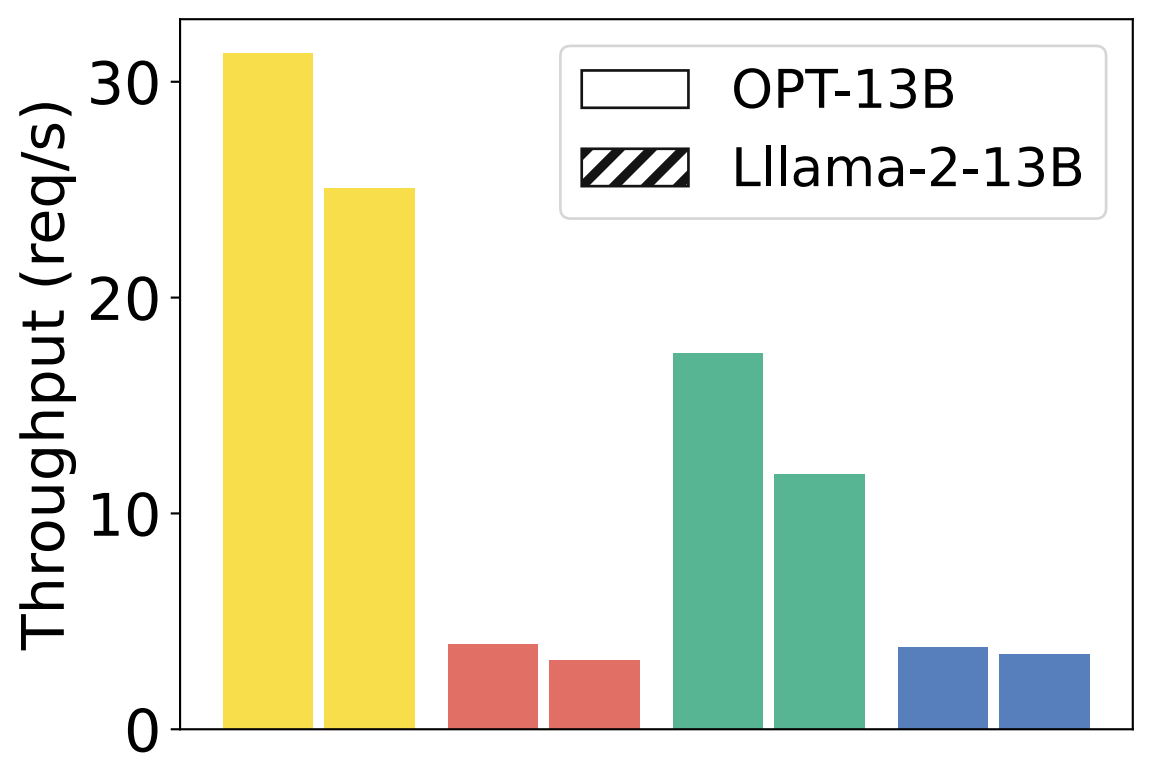
Throughput (req/s)
0
10
20
30
OPT-13B
Lllama-2-13B