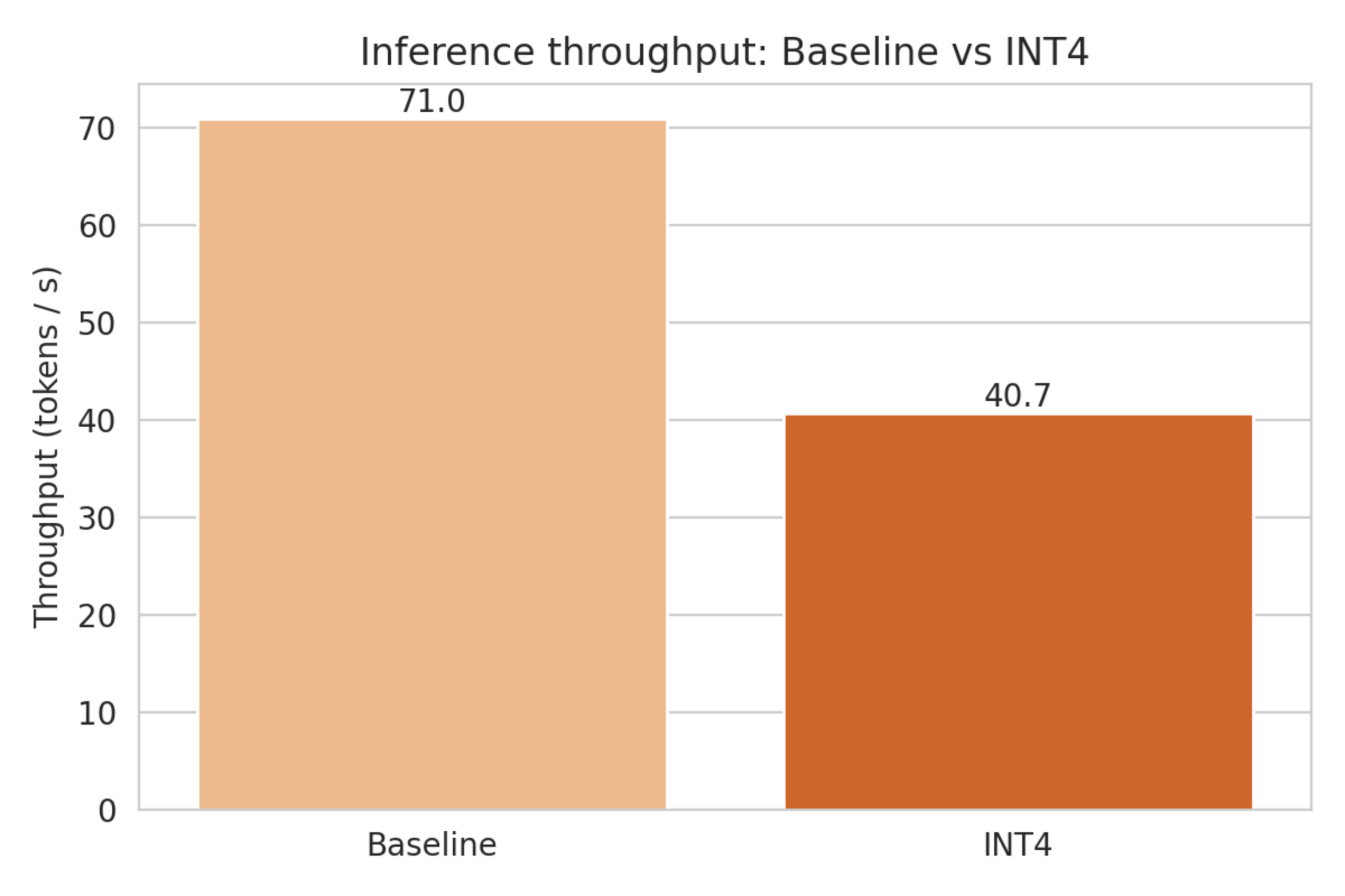
Inference throughput: Baseline vs INT4
Throughput (tokens / s)
0
10
20
30
40
50
60
70
71.0
40.7
Baseline
INT4