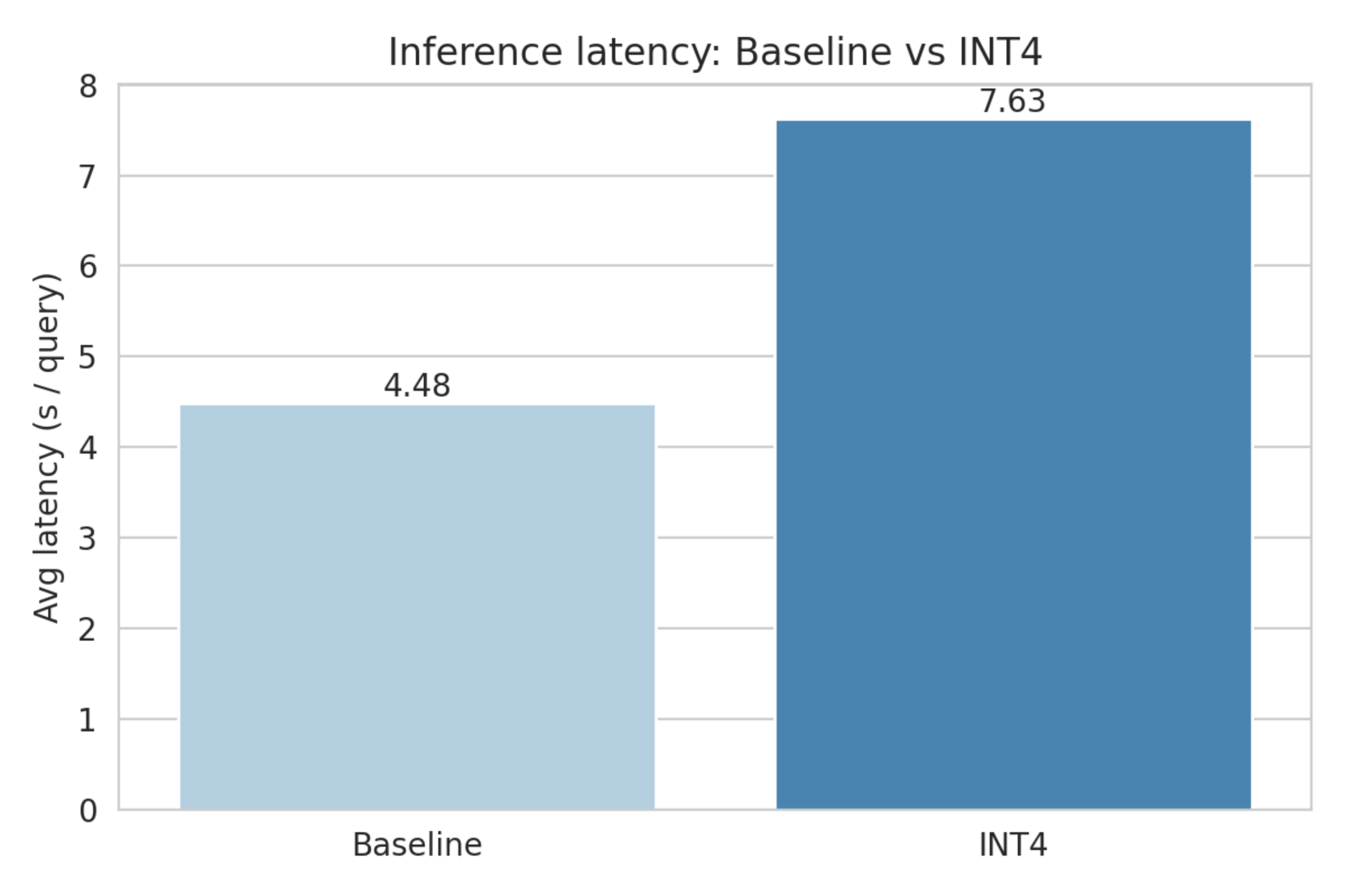
Inference latency: Baseline vs INT4
Avg latency (s / query)
0
1
2
3
4
5
6
7
8
4.48
7.63
Baseline
INT4