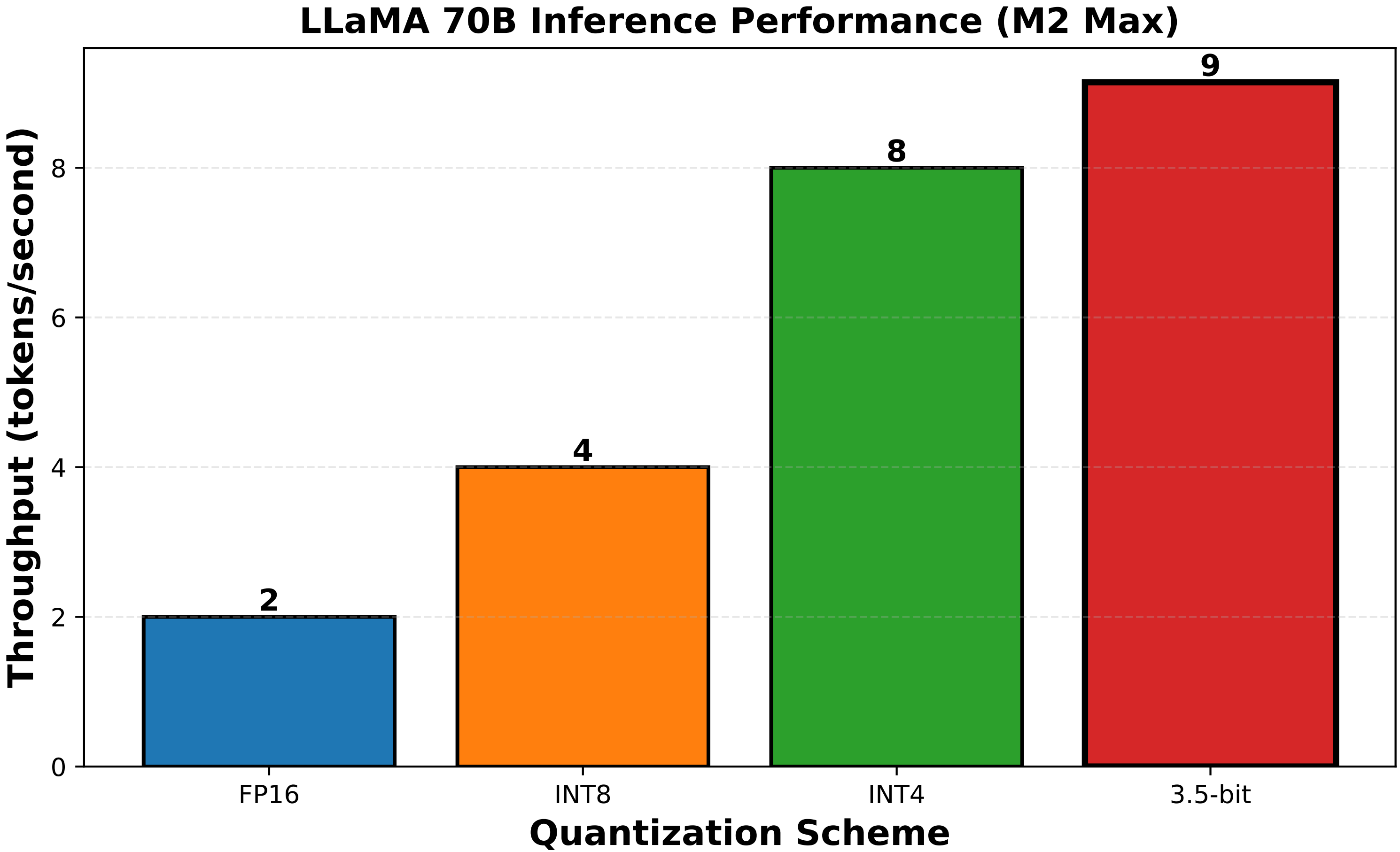
LLaMA 70B Inference Performance (M2 Max)
Throughput (tokens/second)
0
2
4
6
8
2
4
8
9
FP16
INT8
INT4
3.5-bit
Quantization Scheme