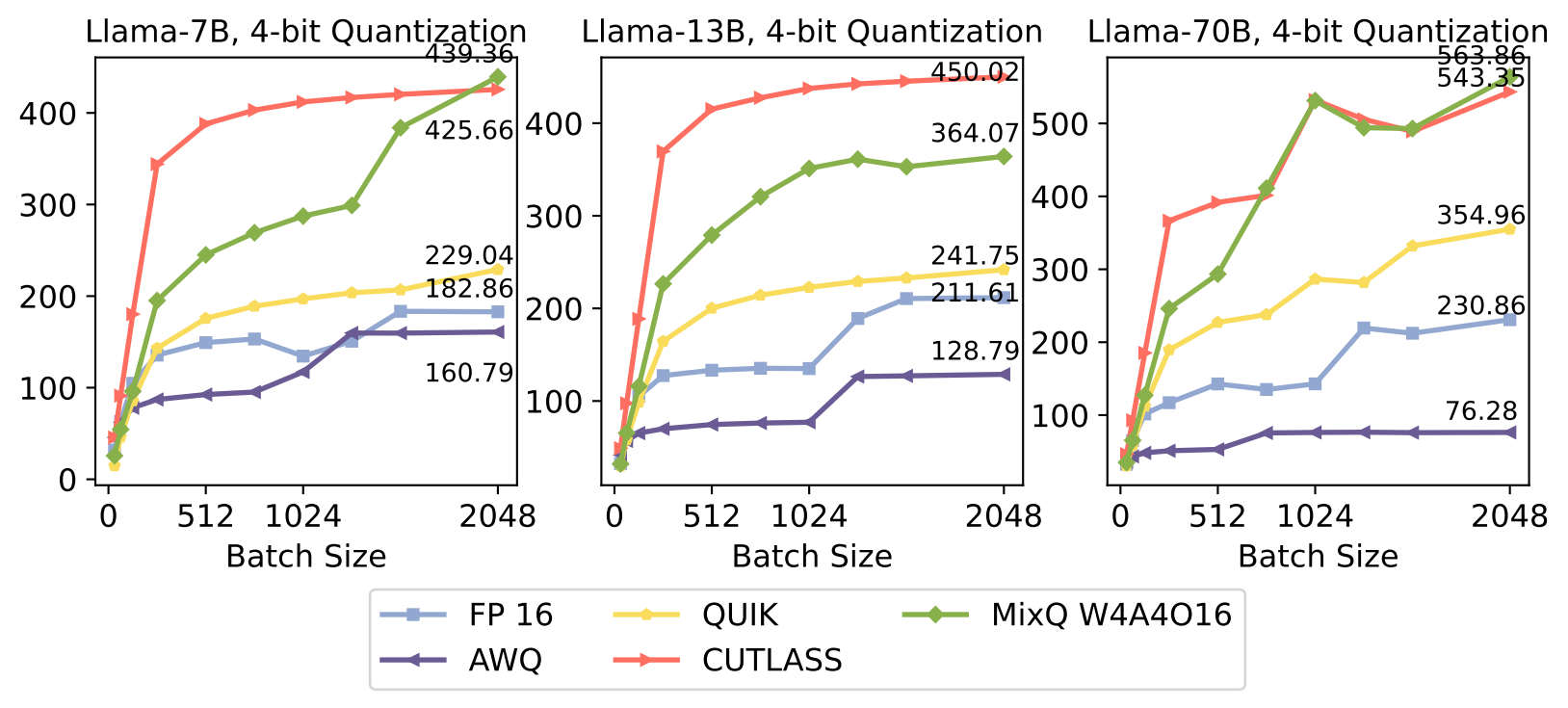
Llama-7B, 4-bit Quantization
Llama-13B, 4-bit Quantization
Llama-70B, 4-bit Quantization
439.36
425.66
229.04
182.86
160.79
450.02
364.07
241.75
211.61
128.79
563.86
543.35
354.96
230.86
76.28
0
100
200
300
400
500
0
512
1024
2048
Batch Size
FP 16
AWQ
QUIK
CUTLASS
MixQ W4A4O16