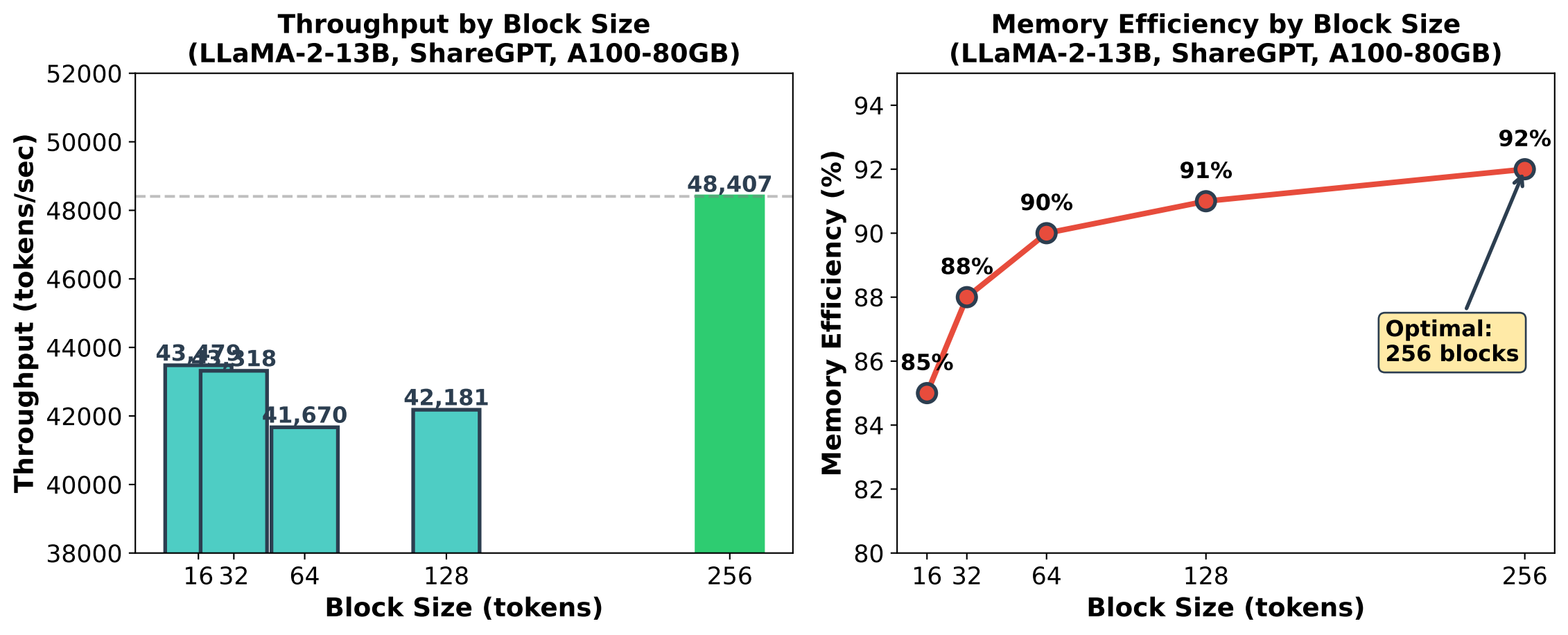
Throughput by Block Size
(LLaMA-2-13B, ShareGPT, A100-80GB)
Throughput (tokens/sec)
52000
50000
48000
46000
44000
42000
40000
38000
43,479
43,318
41,670
42,181
48,407
16
32
64
128
256
Block Size (tokens)
Memory Efficiency by Block Size
(LLaMA-2-13B, ShareGPT, A100-80GB)
Memory Efficiency (%)
94
92
90
88
86
84
82
80
85%
88%
90%
91%
92%
Optimal:
256 blocks
16
32
64
128
256
Block Size (tokens)