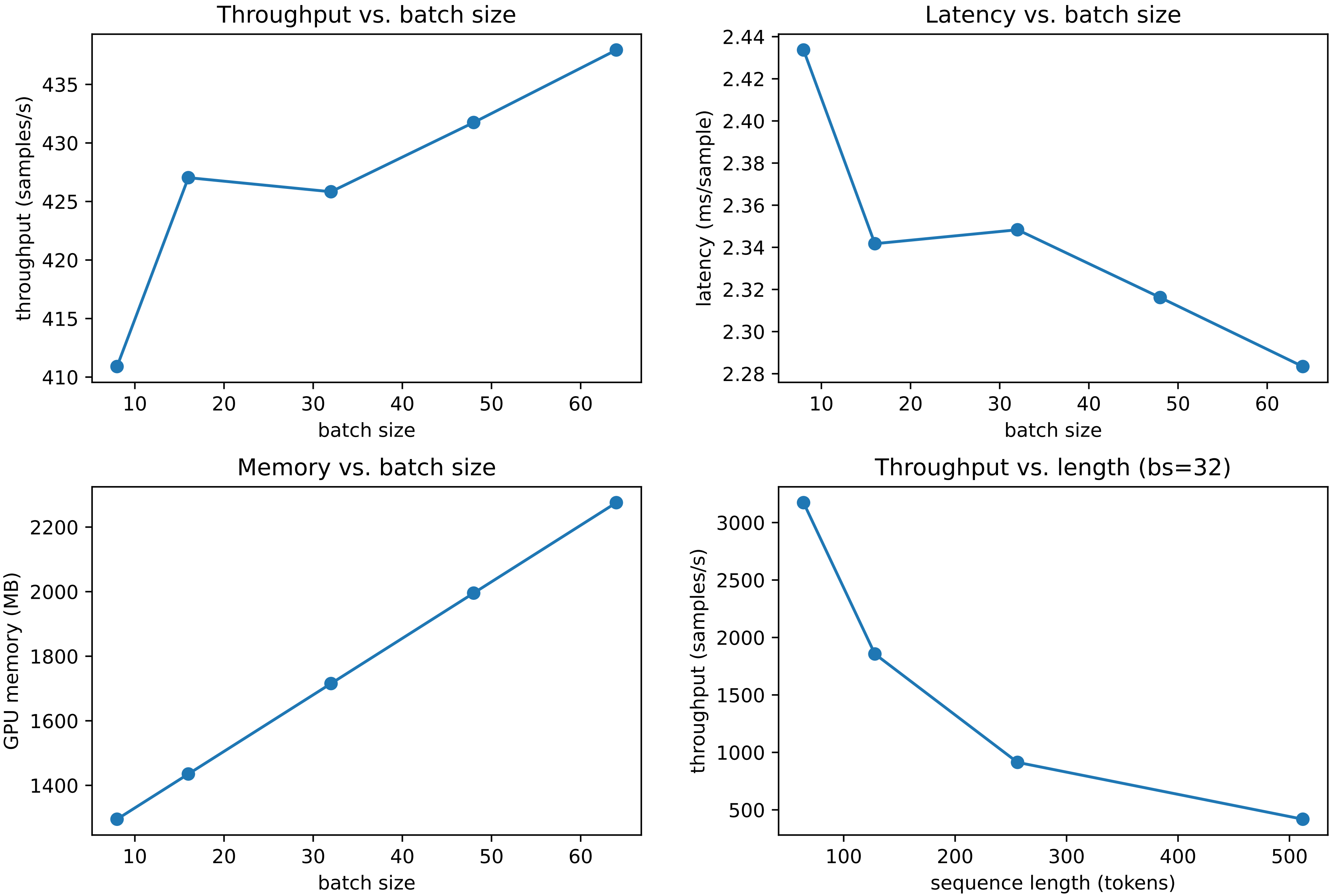
Efficiency & resource curves


Throughput vs. batch size
throughput (samples/s)
batch size
Latency vs. batch size
latency (ms/sample)
batch size
Memory vs. batch size
GPU memory (MB)
batch size
Throughput vs. length (bs=32)
throughput (samples/s)
sequence length (tokens)