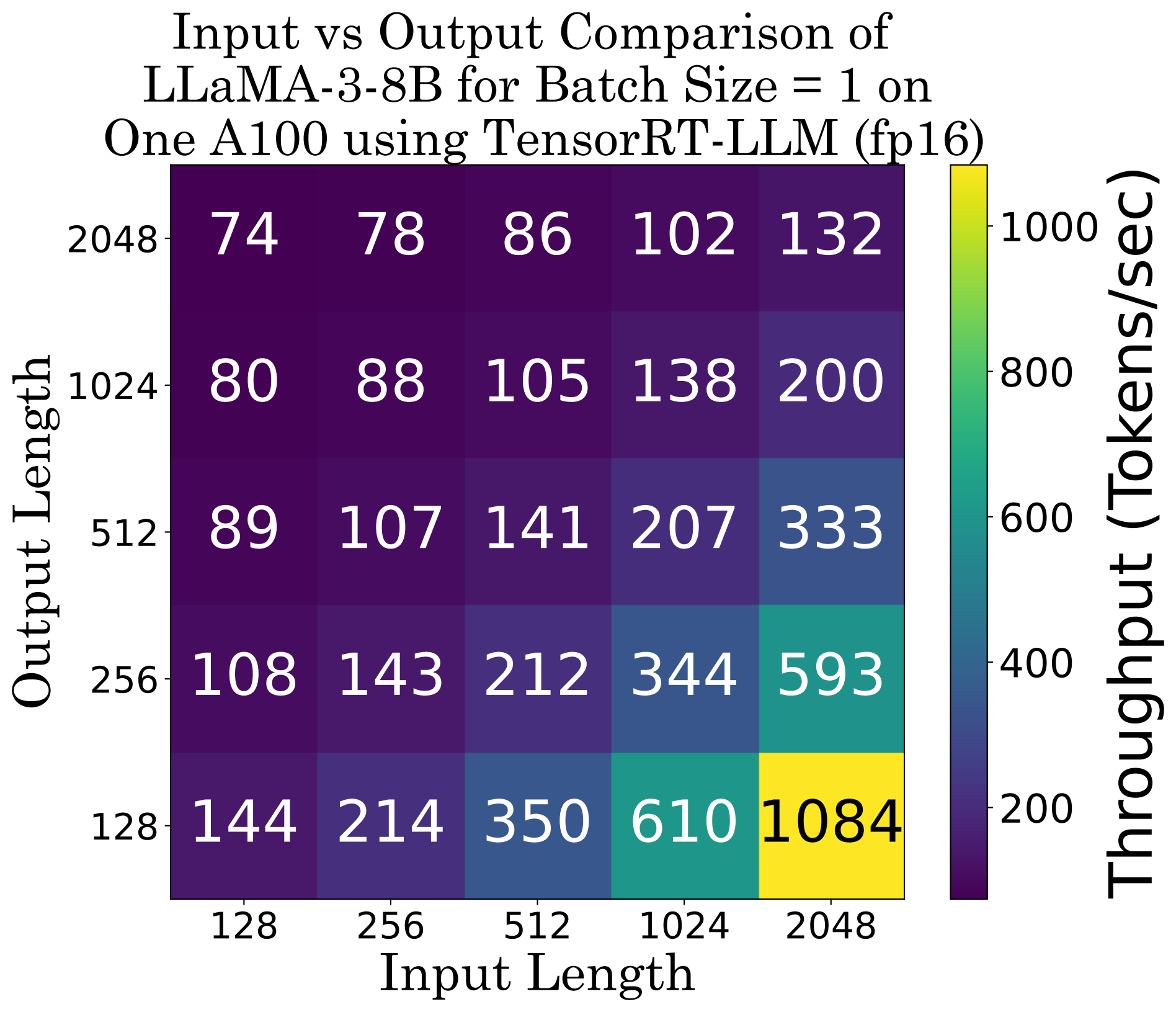
Input vs Output Comparison of
LLaMA-3-8B for Batch Size = 1 on
One A100 using TensorRT-LLM (fp16)
Output Length
2048
1024
512
256
128
74 78 86 102 132
80 88 105 138 200
89 107 141 207 333
108 143 212 344 593
144 214 350 610 1084
128 256 512 1024 2048
Input Length
1000
800
600
400
200
Throughput (Tokens/sec)