

Throughput (tokens/sec)

200  
150  
100  
50  
0

100

200

300

500

750

1000

1500

Generation Length (tokens)

