

Model Run Time (ms)

140
120
100
80
60
40
20

0 1000 2000 3000 4000
Total Scheduled Tokens

