
Performance for Model: 1B, Seq Len: 2048 on RTX5090
Host Memory (GB)
176
160
144
128
112
96
80
64
48
32
16
139.05 167.11 166.79 167.01 166.86
140.82 166.89 166.72 166.68 166.75
139.02 167.04 166.59 166.86 166.44
139.31 166.70 166.57 166.61 166.85
139.00 166.82 166.48 166.51 166.57
139.63 166.65 166.59 166.61 166.60
139.35 166.87 166.48 166.78 166.72
138.54 167.06 166.55 166.70 166.48
139.33 166.58 166.11 166.46 166.46
138.48 166.43 166.22 166.13 166.40
138.57 166.23 166.07 165.88 166.38
16
20
24
28
30
Device Memory (GB)
TFLOPS
160
150
140
130
120