
Performance for Model: 1B, Seq Len: 4096 on RTX5090
Host Memory (GB)
176
160
144
128
112
96
80
64
48
32
16
153.27 160.52 160.40 160.63 160.41
152.99 160.55 160.42 160.75 160.10
152.34 160.40 160.31 160.21 160.26
152.74 160.23 160.40 160.42 160.27
151.56 160.40 160.51 160.16 160.64
153.02 160.08 160.52 160.38 160.22
152.43 160.35 160.51 160.15 159.92
152.57 160.14 160.10 159.75 159.92
152.17 159.98 160.04 160.00 159.93
151.32 159.91 160.16 160.27 160.05
151.62 160.34 159.94 159.75 160.17
16
20
24
28
30
Device Memory (GB)
TFLOPS/s
180
160
140
120
100
80