Performance for Model: 8B, Seq Len: 512 on RTX5090

Host Memory (GB)	18	20	22	24	26	28	30
176	156.55	164.32	169.49	172.54	180.52	185.50	188.11
160	157.73	164.03	169.38	173.53	181.14	185.52	188.19
144	157.08	164.87	168.52	173.41	181.44	186.08	188.44
128	157.93	164.68	168.89	173.65	181.29	186.47	188.70
112	157.62	165.06	168.51	173.88	180.60	185.93	188.52
96	157.76	164.76	169.92	173.40	181.24	186.63	188.86
80	157.17	163.80	168.73	173.13	181.27	186.37	188.27

Device Memory (GB)

TFLOPS/s