

Performance for Model: 1B, Seq Len: 8192 on RTX5090

