

Performance for Model: 1B, Seq Len: 32768 on RTX5090

