

Performance for Model: 1B, Seq Len: 16384 on RTX5090

