

Performance for Model: 8B, Seq Len: 16384 on RTX5090

