

Performance for Model: 8B, Seq Len: 4096 on RTX5090

