

Performance for Model: 8B, Seq Len: 32768 on RTX5090

