

Performance for Model: 8B, Seq Len: 8192 on RTX5090

