

Performance for Model: 8B, Seq Len: 2048 on RTX5090

