

Performance for Model: 8B, Seq Len: 512 on RTX3090

