

Performance for Model: 1B, Seq Len: 512 on RTX5090

