Performance for Model: 1B, Seq Len: 2048 on RTX5090