Bert of different layers on V100 60 (records/s) (records/s) throughput of 10 exact 20 exact 28 exact 36 quantize 20 quantize 28 quantize 36 50 200 100 150 batch size