

LLAMA-70B-INT4 Inference Speedup on A100

