

LLAMA-13B-INT4 Inference Speedup on A100 (FP16)

