Llama 3.2-1B Quantization: Key Results

2.44×

Memory Compression

+8.2%

NF4 vs FP16 Baseline

+15.2%

NF4 vs FP4 Advantage