Llama 3.2-1B Quantization: Key Results

2.44×

Memory Compression

+5.3%

Accuracy vs Baseline

9.5%

NF4 vs FP4 Advantage