

Effective Bits of Different Quantization Methods for Llama 3.1 8B 4 bits

