

Llama-7B, 8-bit Quantization

Llama-13B, 8-bit Quantization

Llama-70B, 8-bit Quantization

