

Quantized Model Accuracy vs Average Peak VRAM

