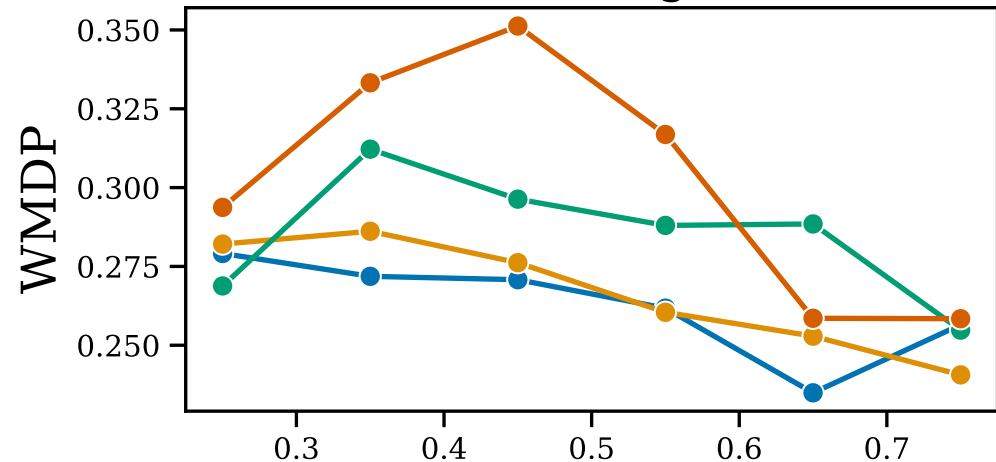
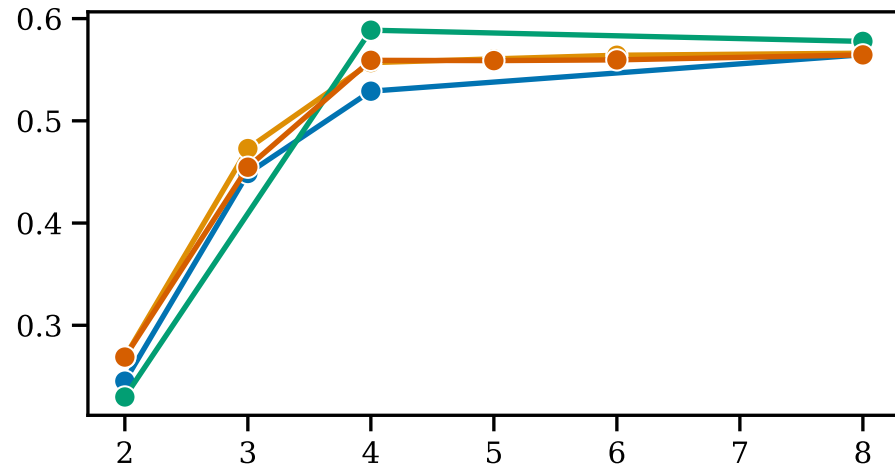
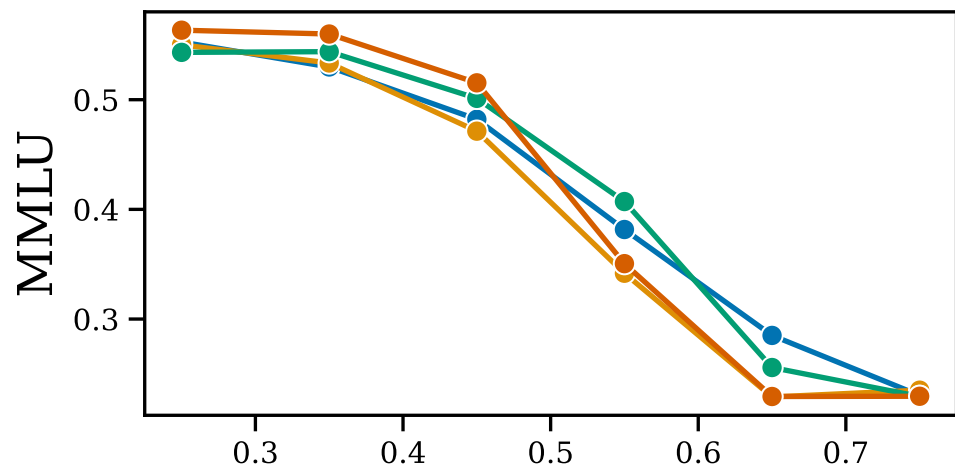
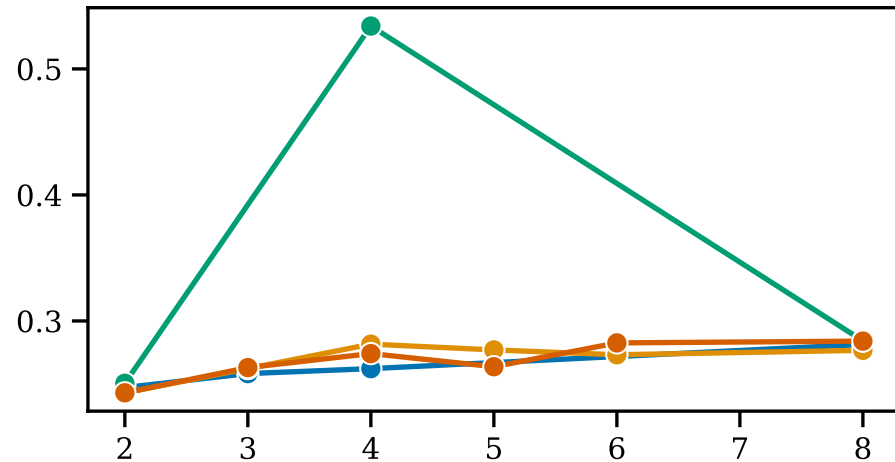


Pruning



Quantization



Sparsity

● RMU→SparseGPT    ● SparseGPT→RMU  
 ● RMU→Wanda    ● Wanda→RMU

Bits

● RMU→GPTQ    ● GPTQ→RMU  
 ● RMU→AWQ    ● AWQ→RMU