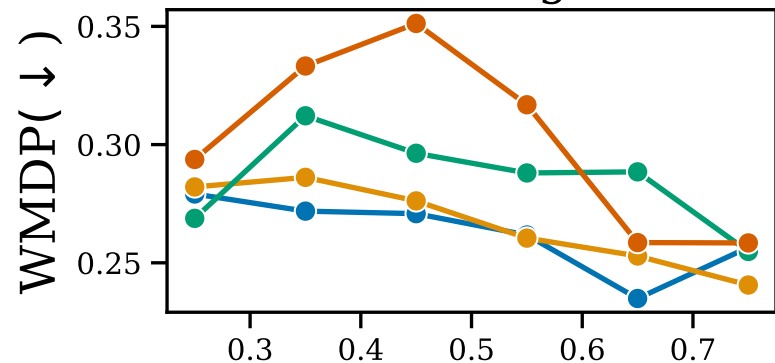
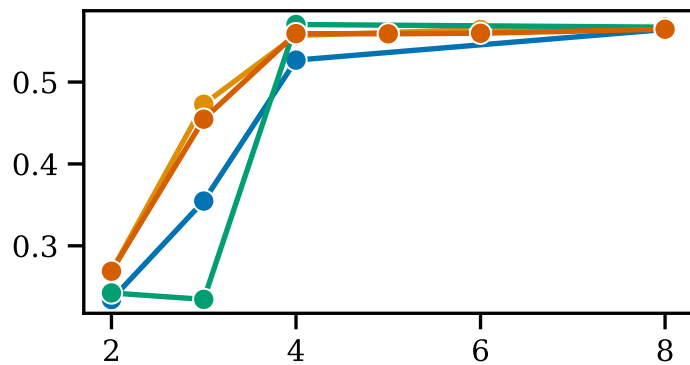
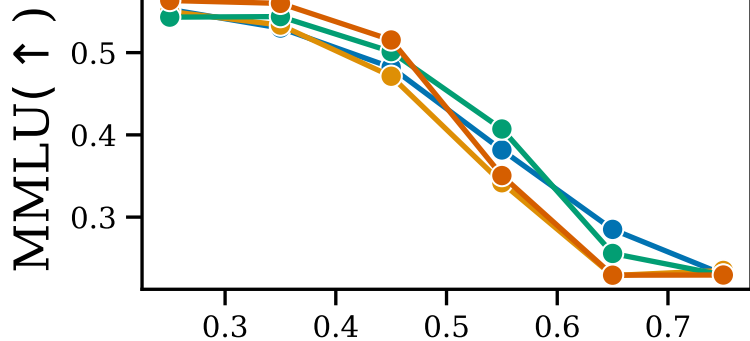
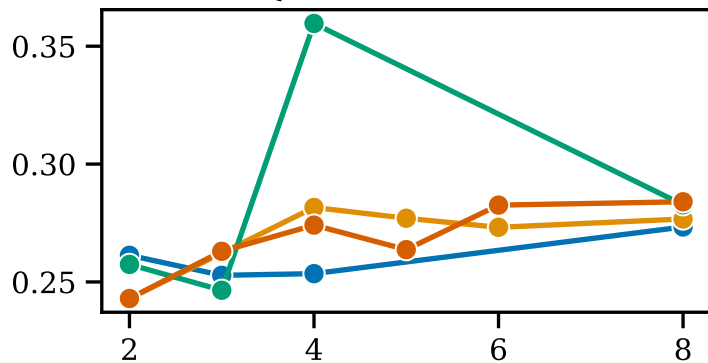


Pruning



Quantization



Sparsity

Bits

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

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