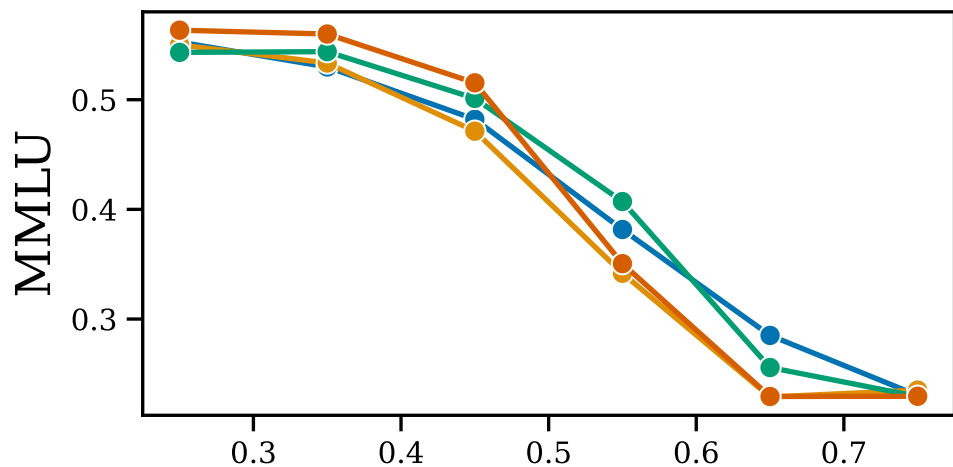
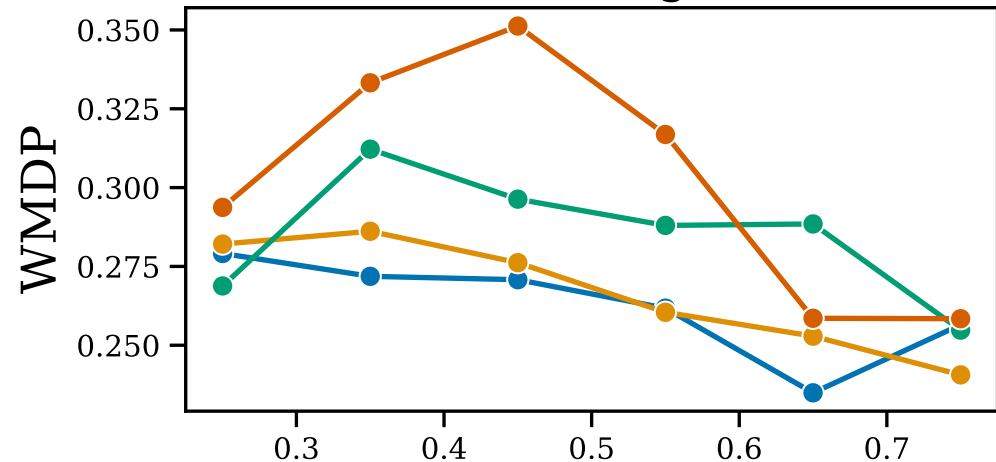
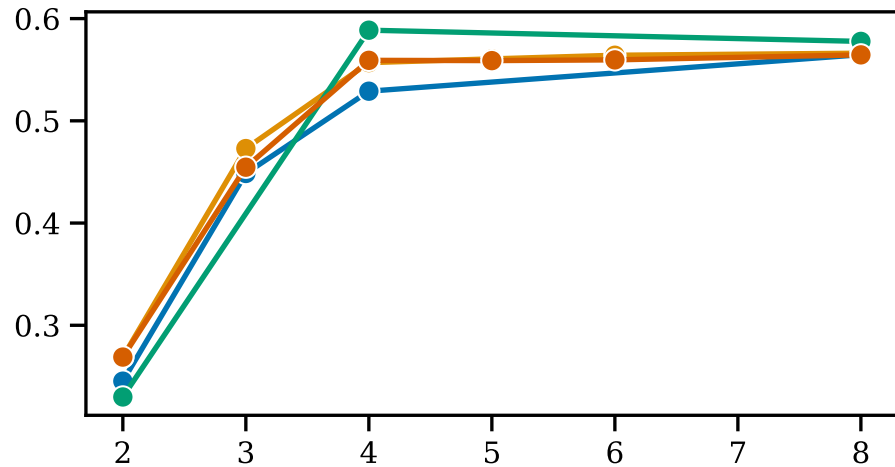
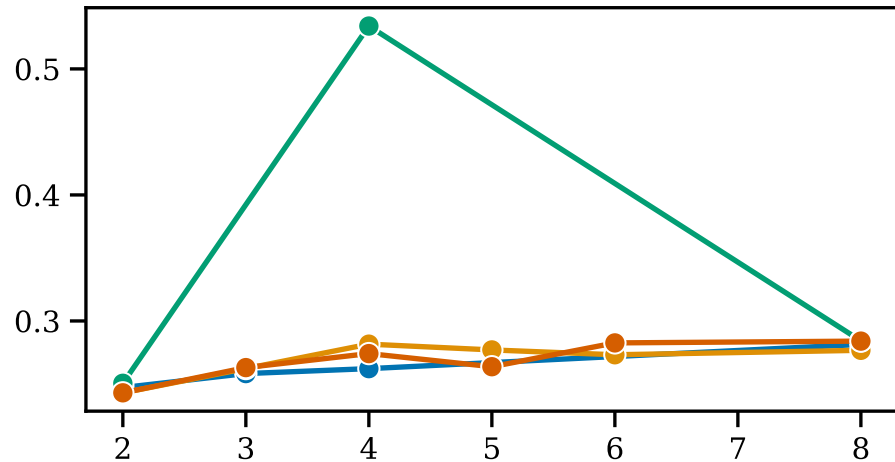


Pruning



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

Quantization



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