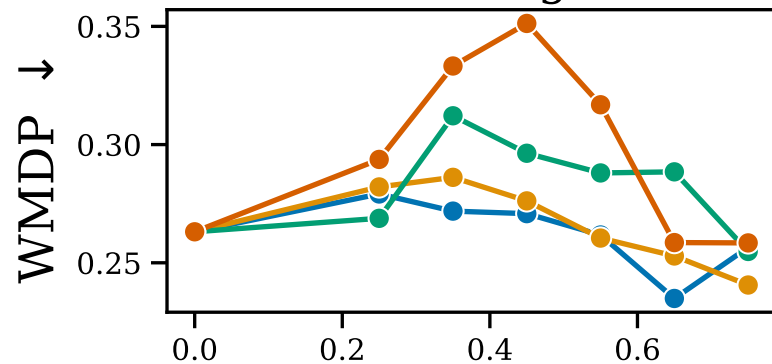
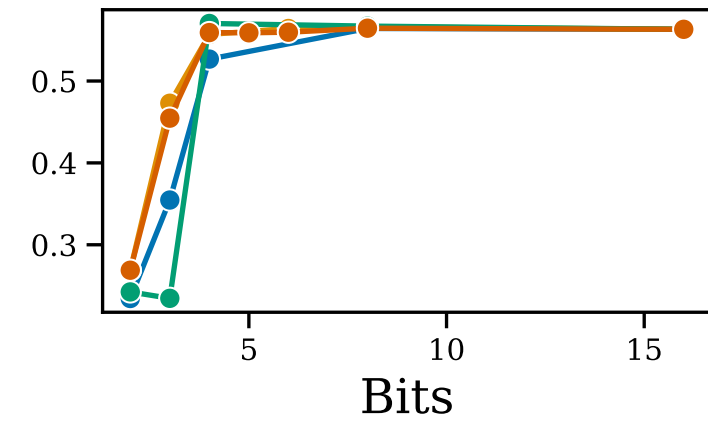
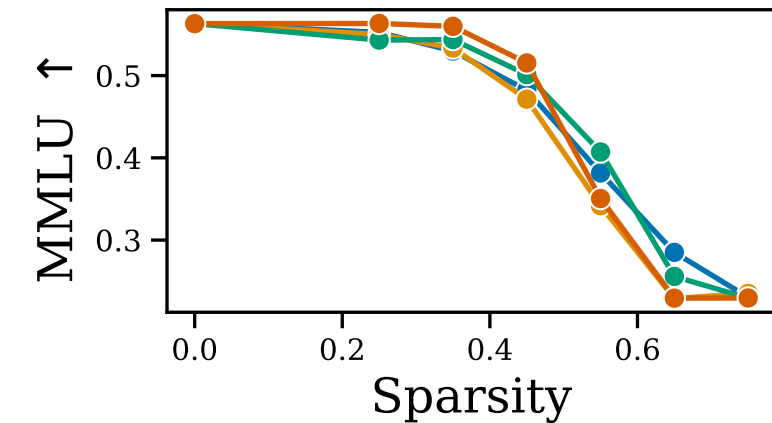
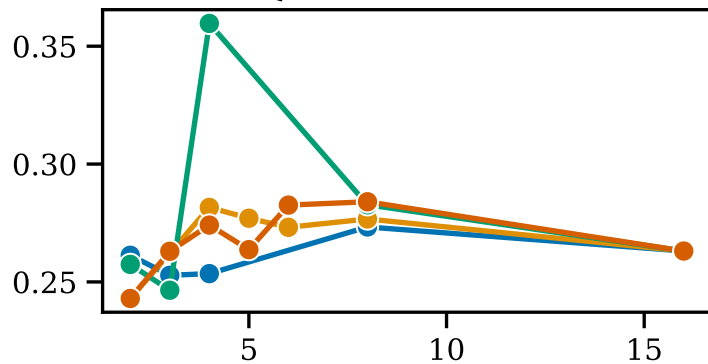


Pruning



Quantization



● RMU→SparseGPT ● SparseGPT→RMU ● RMU→GPTQ ● GPTQ→RMU
 ● RMU→Wanda ● Wanda→RMU ● RMU→AWQ ● AWQ→RMU