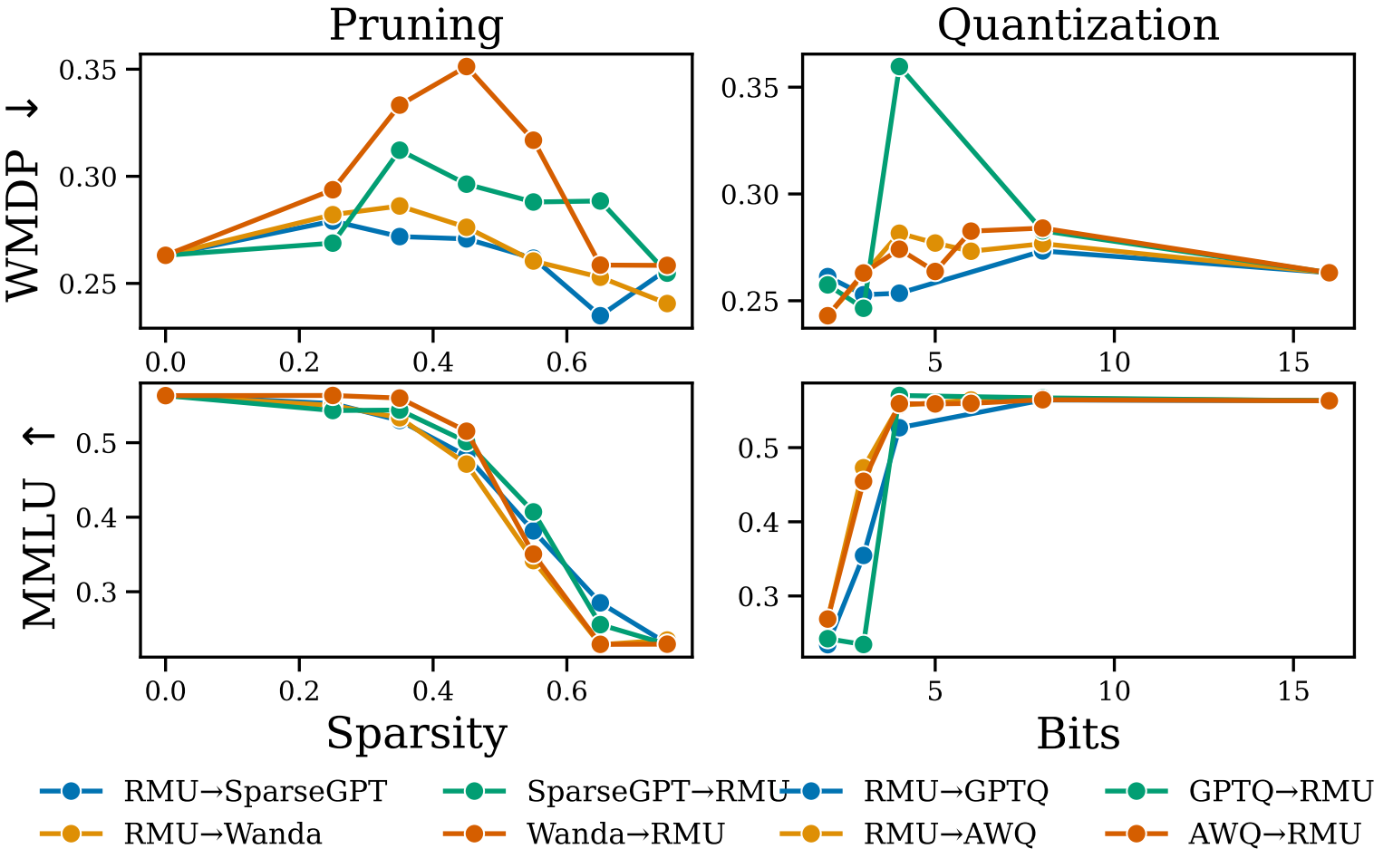
Pruning
Quantization
WMDP ↓
MMLU ↑
0.35
0.30
0.25
0.5
0.4
0.3
0.0
0.2
0.4
0.6
5
10
15
Sparsity
Bits
RMU→SparseGPT
SparseGPT→RMU
RMU→GPTQ
GPTQ→RMU
RMU→Wanda
Wanda→RMU
RMU→AWQ
AWQ→RMU