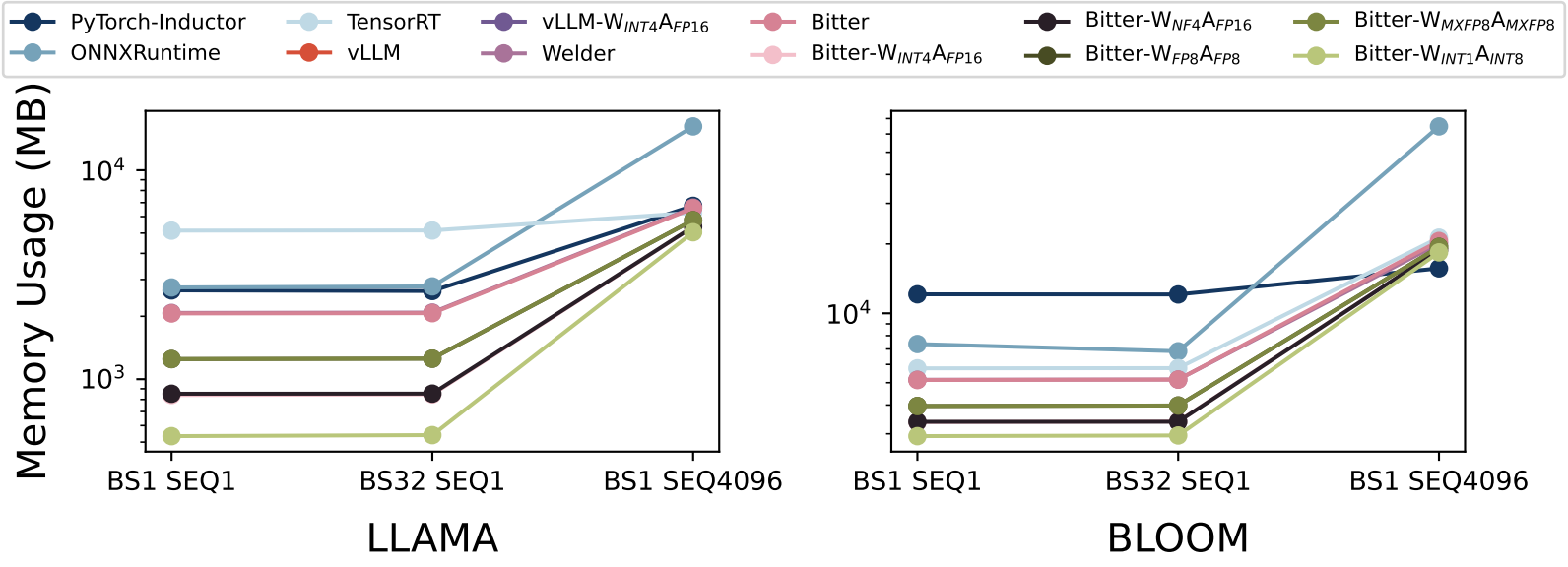
PyTorch-Inductor
ONNXRuntime
TensorRT
vLLM
vLLM-W$_{INT4}$A$_{FP16}$
Welder
Bitter
Bitter-W$_{INT4}$A$_{FP16}$
Bitter-W$_{NF4}$A$_{FP16}$
Bitter-W$_{FP8}$A$_{FP8}$
Bitter-W$_{MXFP8}$A$_{MXFP8}$
Bitter-W$_{INT1}$A$_{INT8}$
Memory Usage (MB)
10^4
10^3
BS1 SEQ1
BS32 SEQ1
BS1 SEQ4096
LLAMA
10^4
BS1 SEQ1
BS32 SEQ1
BS1 SEQ4096
BLOOM