Normalized Speedup vs cuBLAS-$W_{FP16}A_{FP16}$ on A100

Legend: cuBLAS-$W_{FP16}A_{FP16}$, CUTLASS-$W_{INT4}A_{FP16}$, Marlin-$W_{INT4}A_{FP16}$, BitBLAS-$W_{INT4}A_{FP16}$, BitBLAS-$W_{INT2}A_{INT8}$

Subplots: n16384_k16384, n8192_k8192, n8192_k28672, n28672_k8192

Y-axis: Normalized Speedup (vs cuBLAS)
X-axis: Batch Size