Speedup of GEMV on GTX3090 (AF4)

Legend: cuBLAS-$W_{FP16}A_{FP16}$, bitsandbytes-$W_{AF4}A_{FP16}$, BitBLAS-$W_{AF4}A_{FP16}$

y-axis: Speedup vs cuBLAS-$W_{FP16}A_{FP16}$

x-axis: Shapes from LLM (V0–V12)