Speedup of GEMM on A100 (Weight Quantize)

Legend:
- cuBLAS-$W_{FP16}A_{FP16}$
- FasterTransformer-$W_{INT4}A_{FP16}$
- TensorRTLLM-$W_{INT4}A_{FP16}$
- vLLM-$W_{INT4}A_{FP16}$
- Marlin-$W_{INT4}A_{FP16}$
- BitBLAS-$W_{INT4}A_{FP16}$
- BitBLAS-$W_{FP4}A_{FP16}$
- BitBLAS-$W_{INT1}A_{INT8}$

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

X-axis: Shapes from LLM (M0, M1, M2, M3, M4, M5, M6)