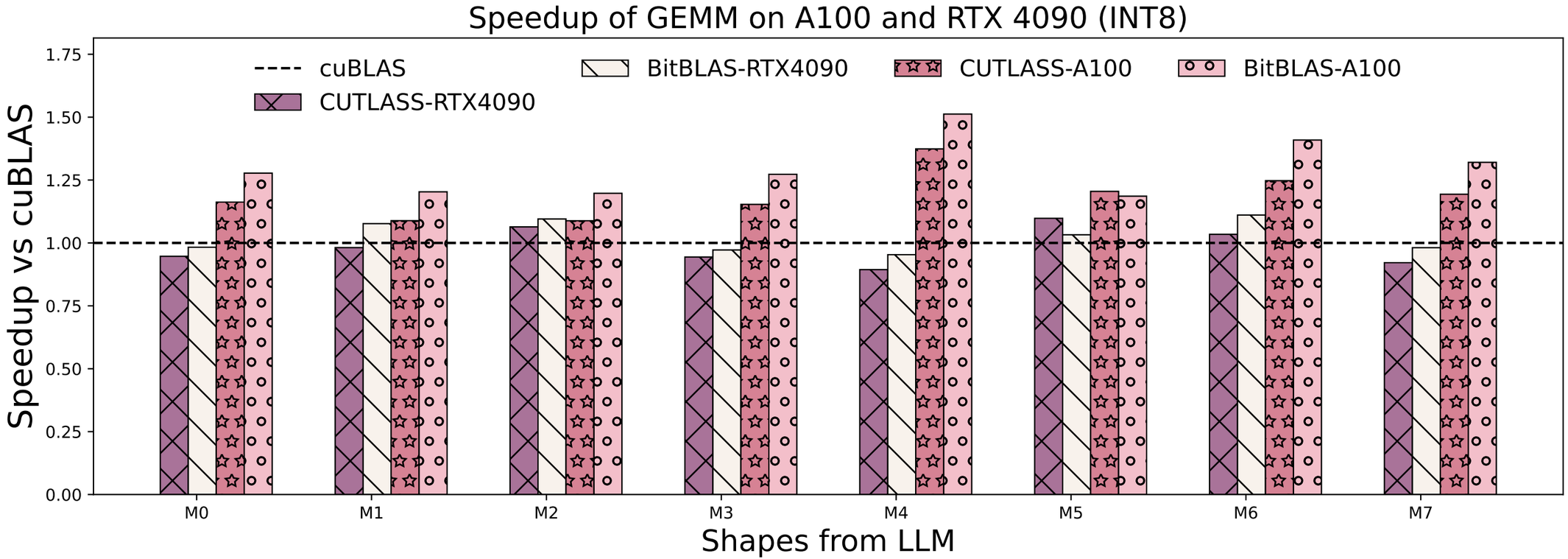
Speedup of GEMM on A100 and RTX 4090 (INT8)
cuBLAS
CUTLASS-RTX4090
BitBLAS-RTX4090
CUTLASS-A100
BitBLAS-A100
Speedup vs cuBLAS
1.75
1.50
1.25
1.00
0.75
0.50
0.25
0.00
M0
M1
M2
M3
M4
M5
M6
M7
Shapes from LLM