LssWeight 1.0 quantize 5% 6% 6% 6% leverage 3% 3% 3% 3% sample pack 0.8 gemm 38% 44% 40% dequantize 45% LSQ 3% 3% 3% 2% 31% 30% 28% 26% 0.2 3% 3% 3% 3% 12% 12% 11% 11% (28612, 28612, 34816) (32168, 32168, 38912) 0.0 (25600, 25600, 32768) (286¹², 286¹², 348¹⁶)
(286¹², 286¹², 30¹²0, 30¹²0, 30¹²0)
(32¹
Matrix size (M,N,K)