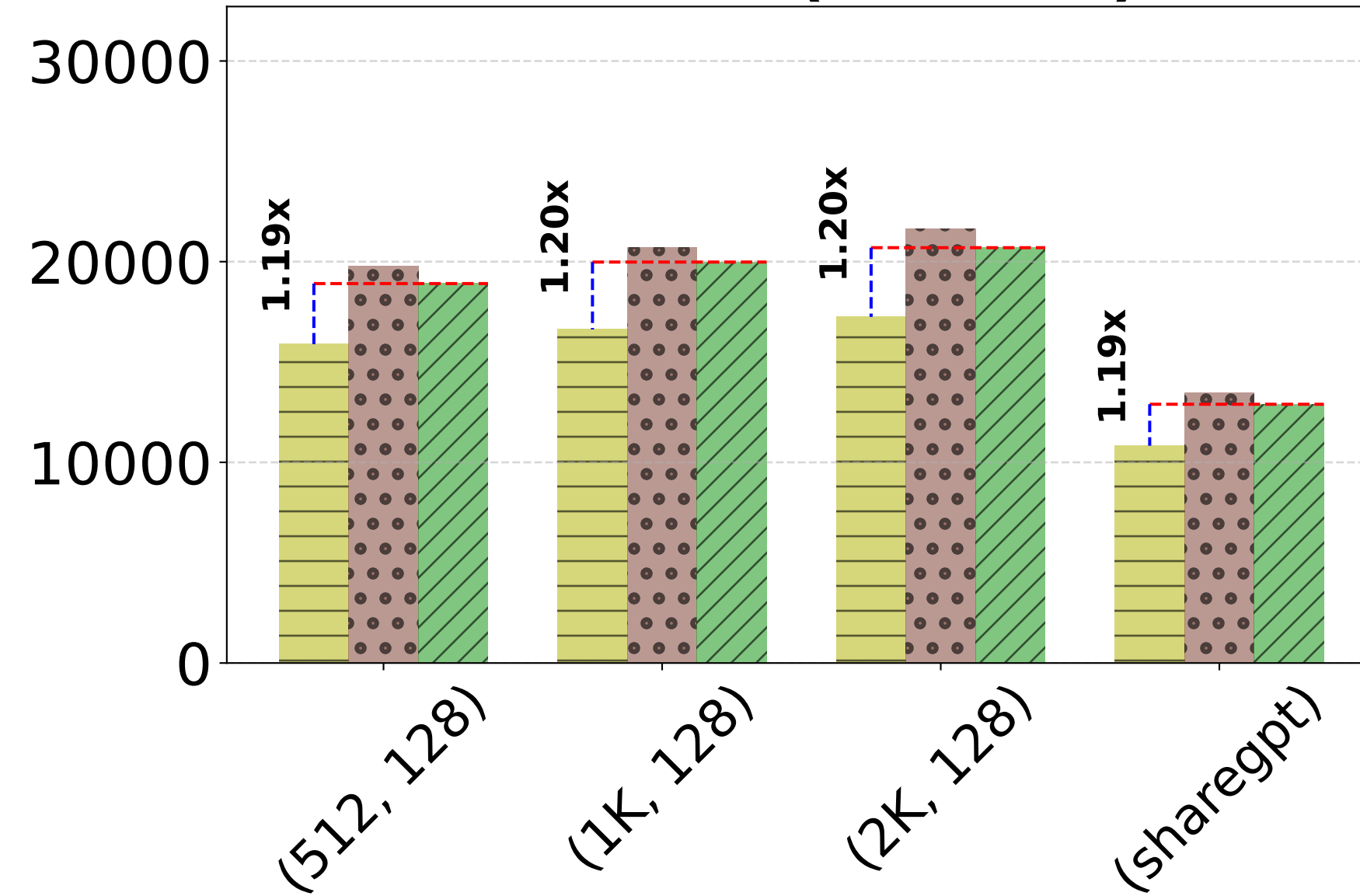
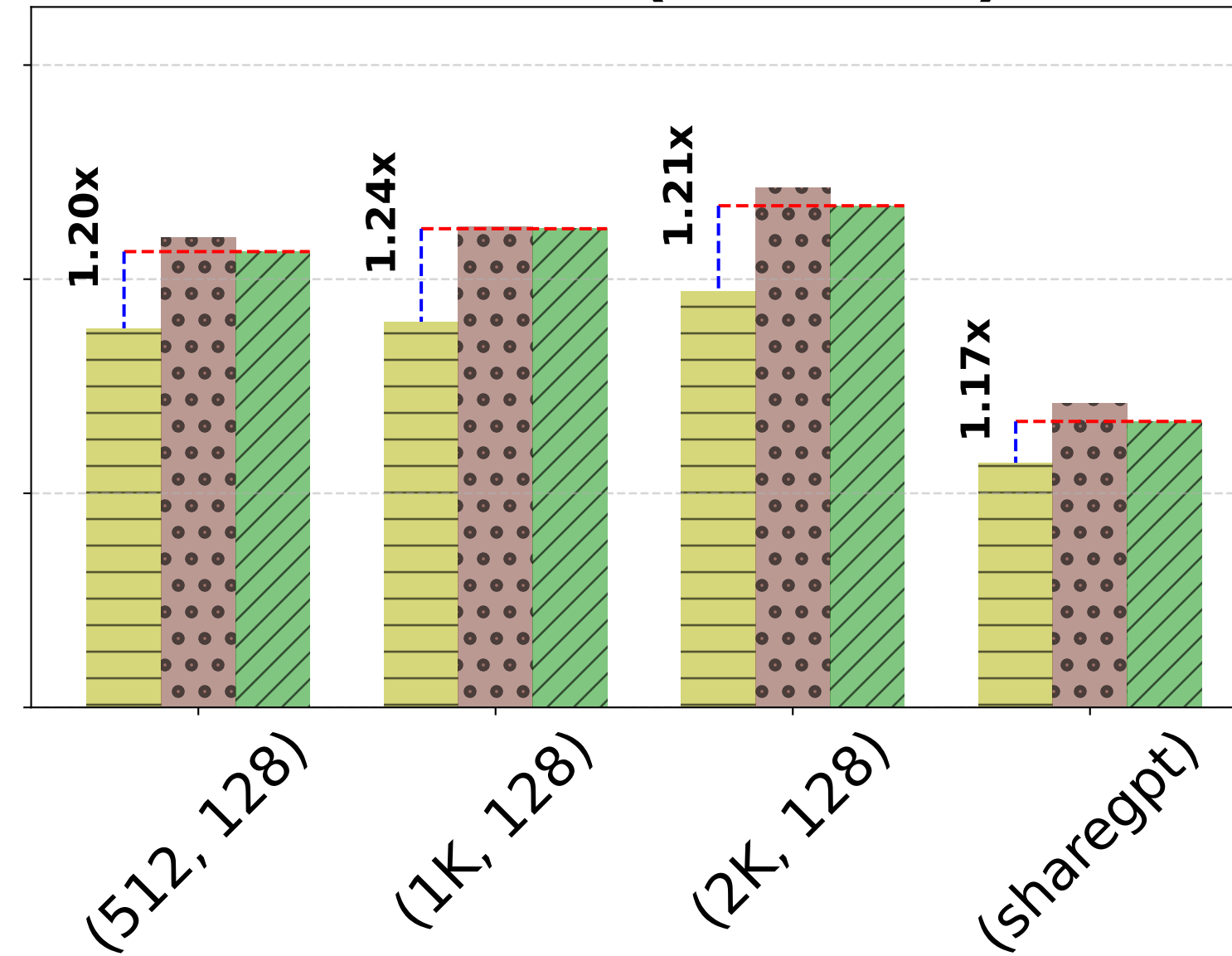


vllm-Multimem      vllm-nocomm      TokenWeave

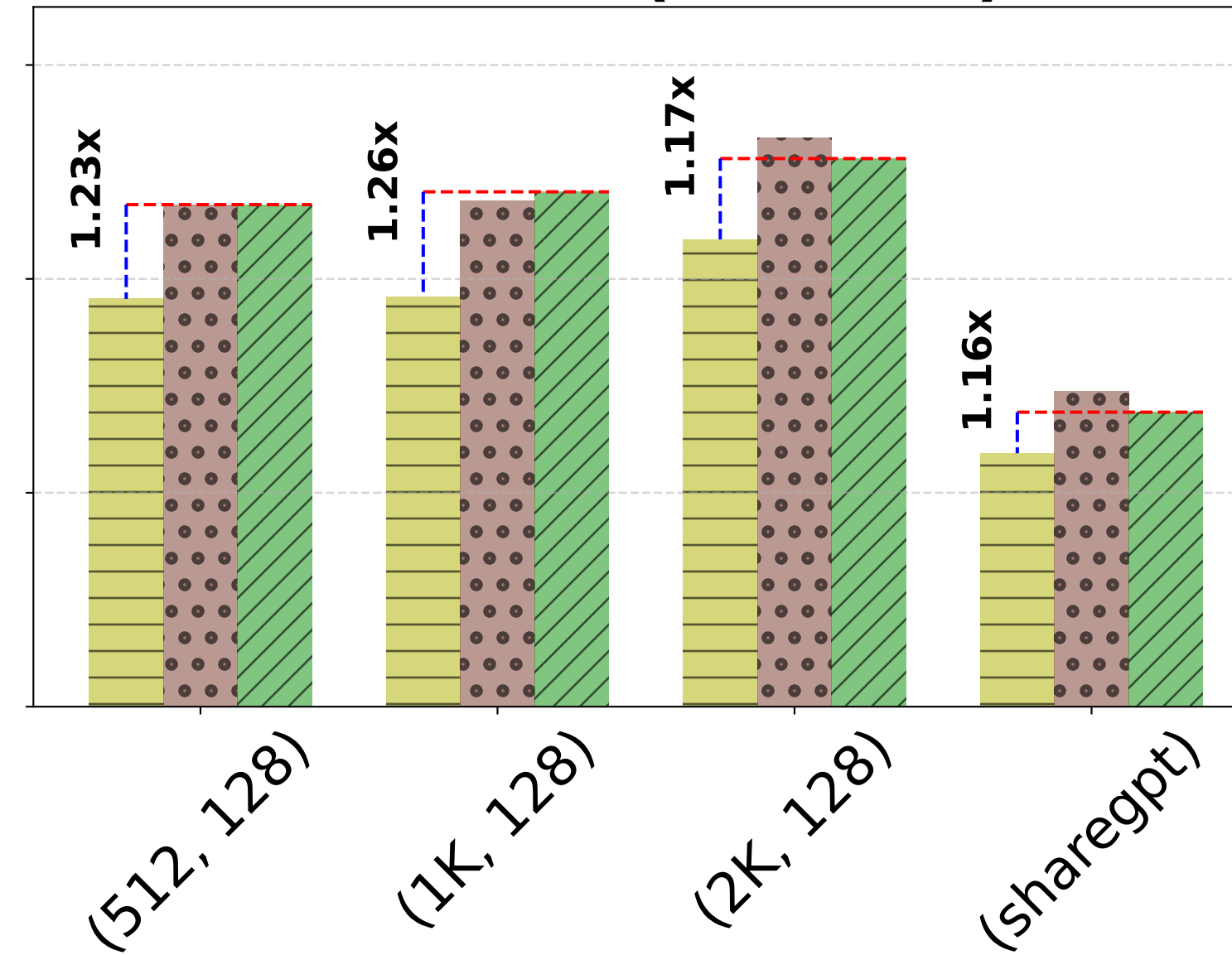
**CS: 1024 (8x H100)**



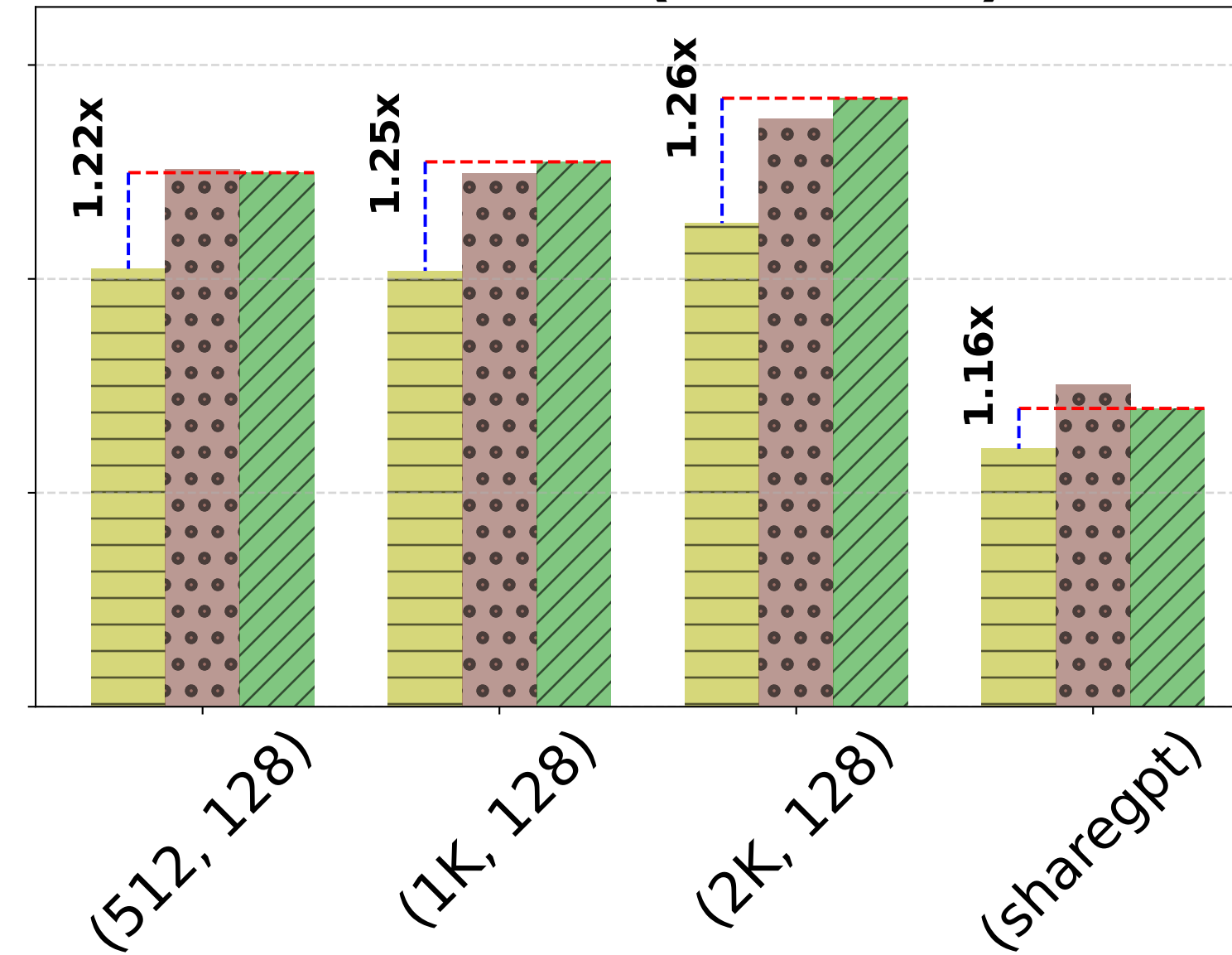
**CS: 2048 (8x H100)**



**CS: 4096 (8x H100)**



**CS: 8192 (8x H100)**



**Fixed(Input, Output) or ShareGPT**