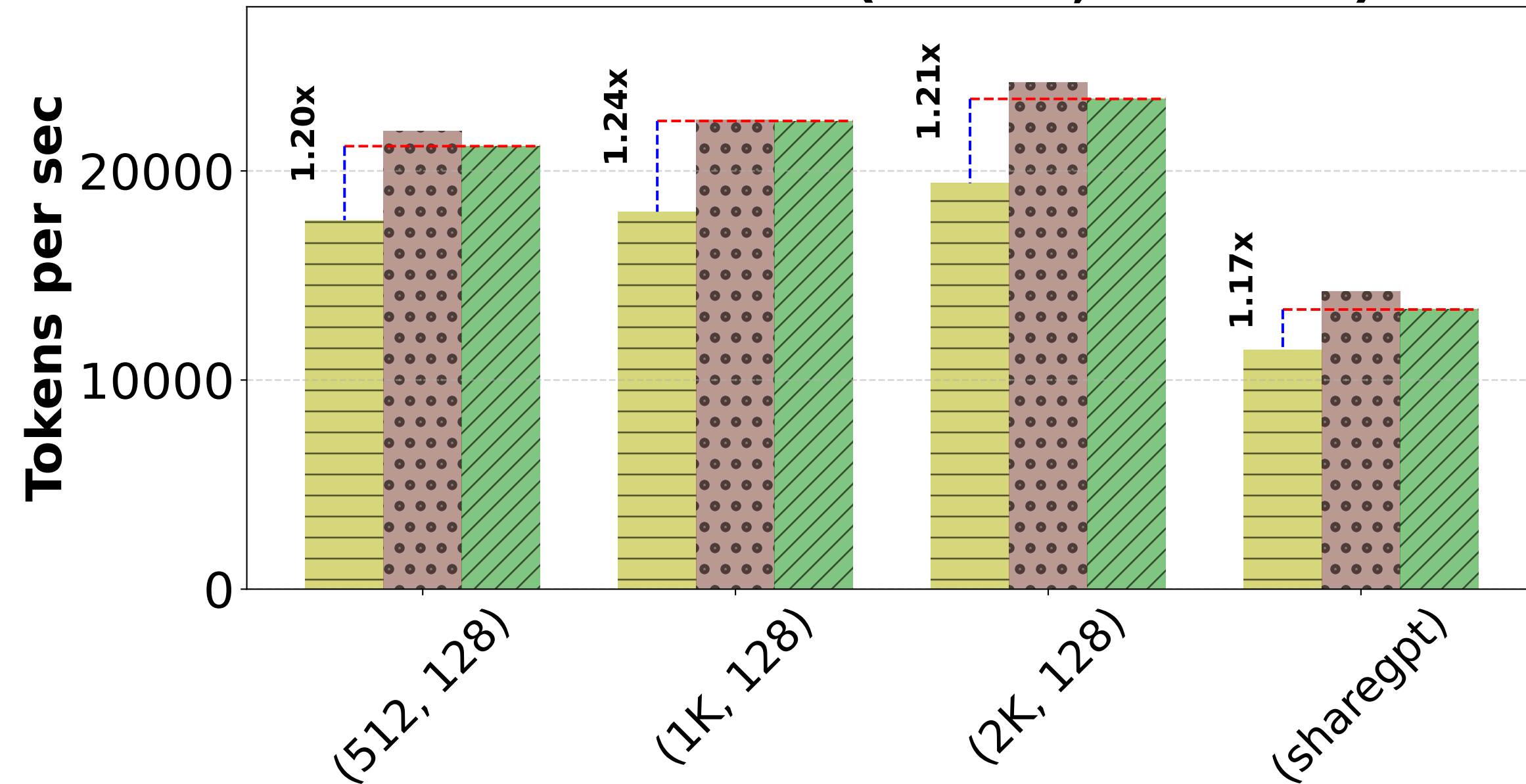
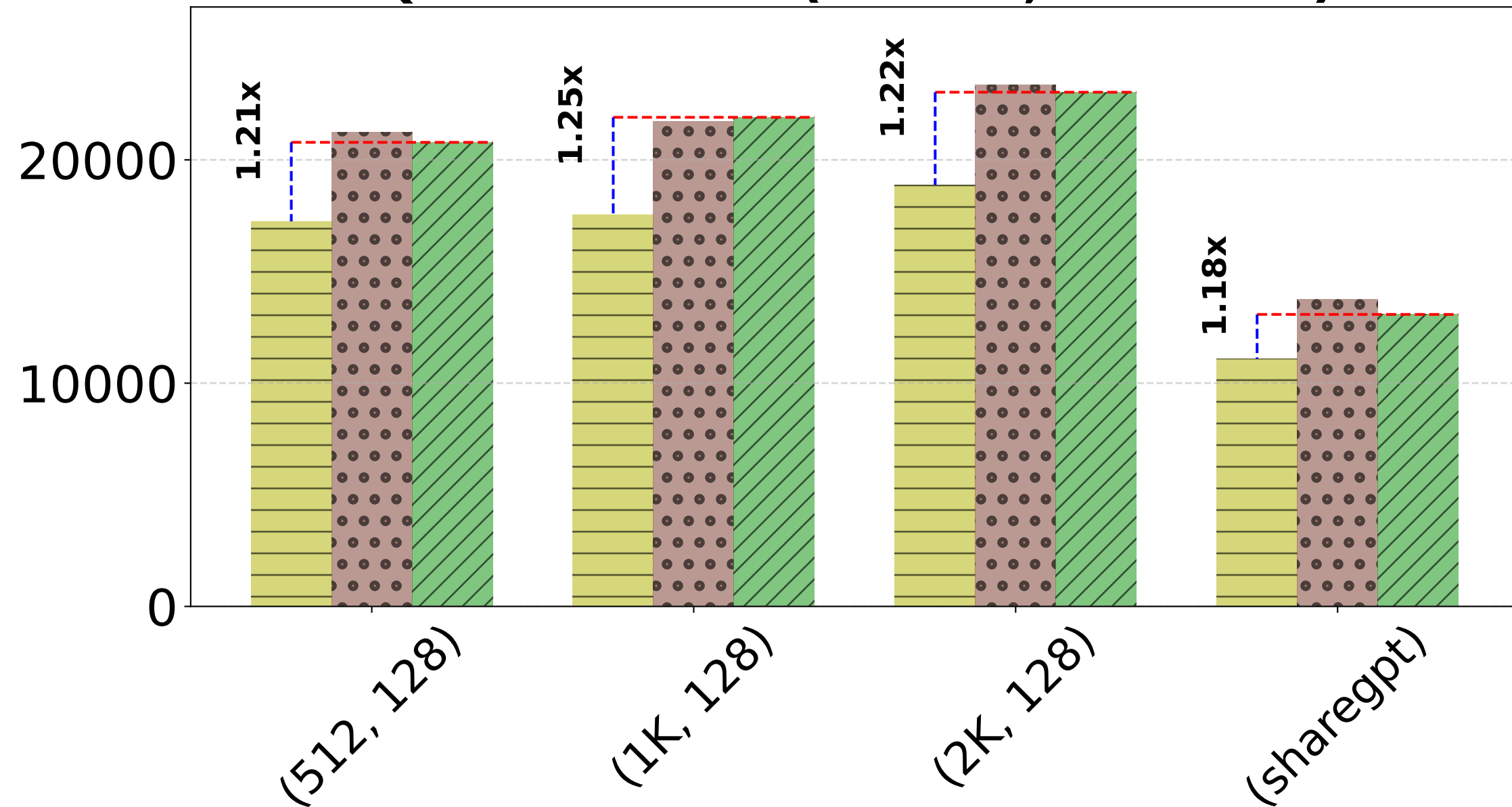


vllm-Multimem vllm-nocomm TokenWeave

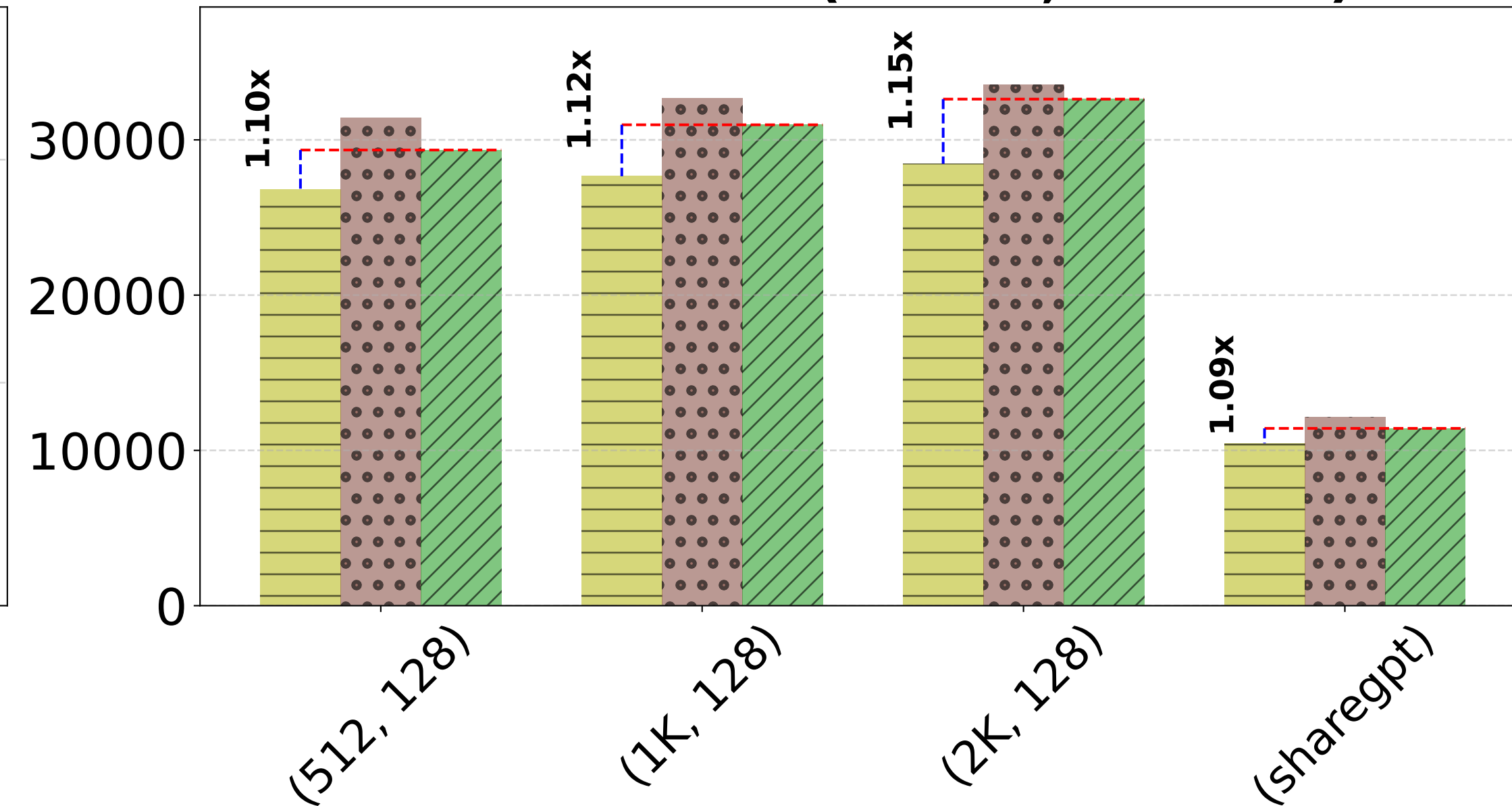
Llama-3.3-70B (CS: 2K, 8x H100)



Qwen2.5-72B (CS: 2K, 8x H100)



Mixtral-8x22B (CS: 4K, 8x H100)



Fixed(Input, Output) or ShareGPT