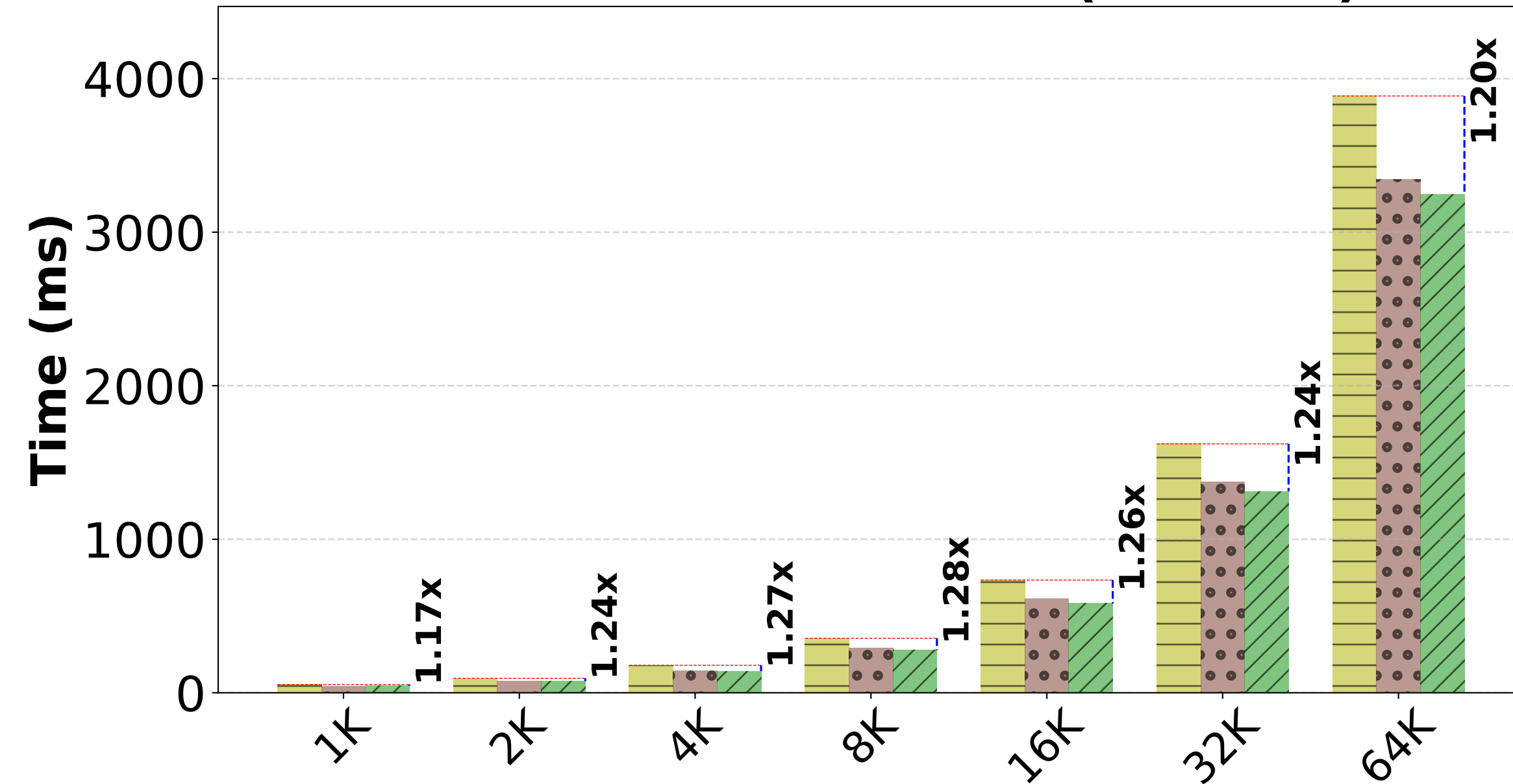
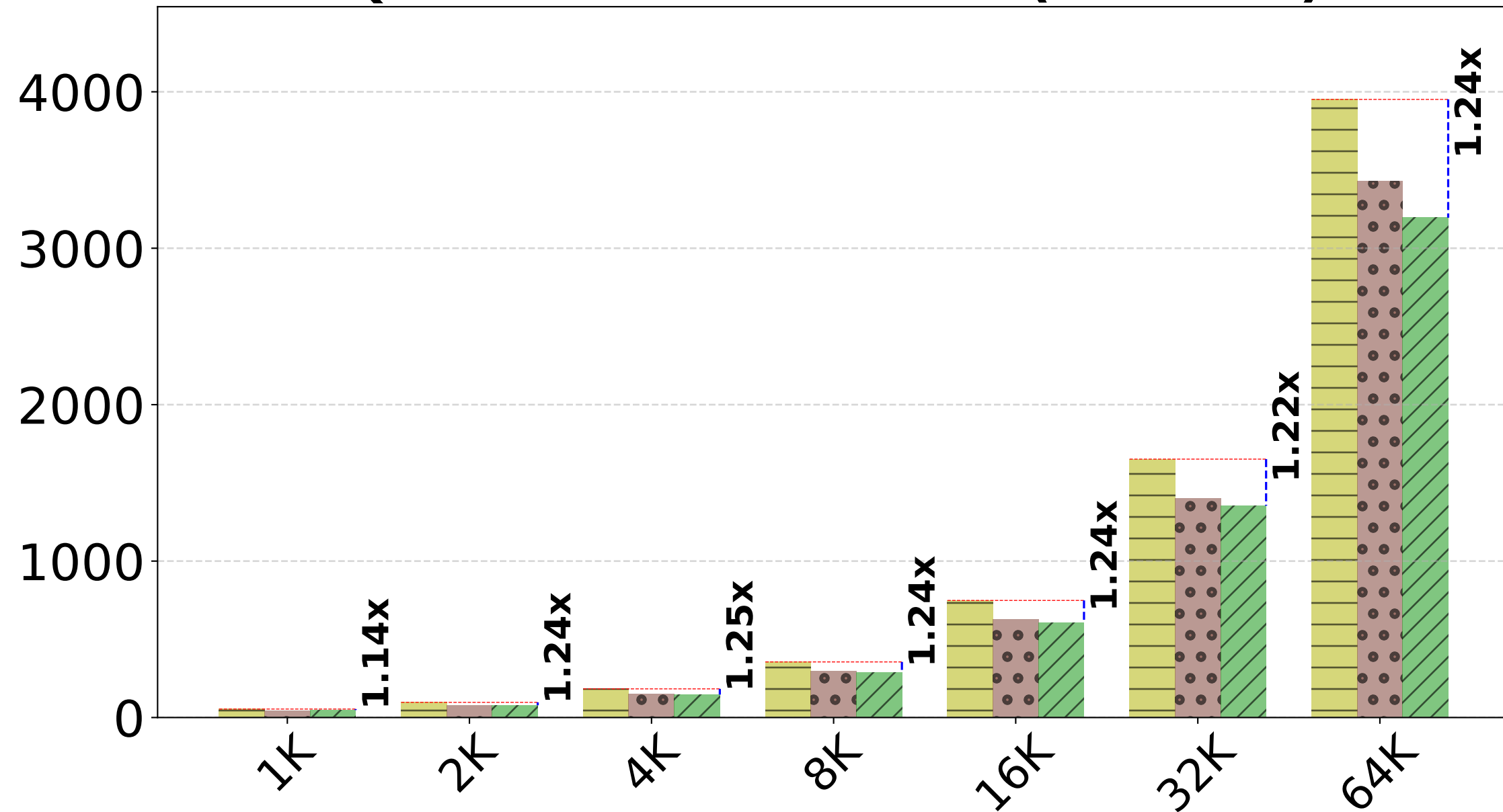


vllm-Multimem vllm-nocomm TokenWeave

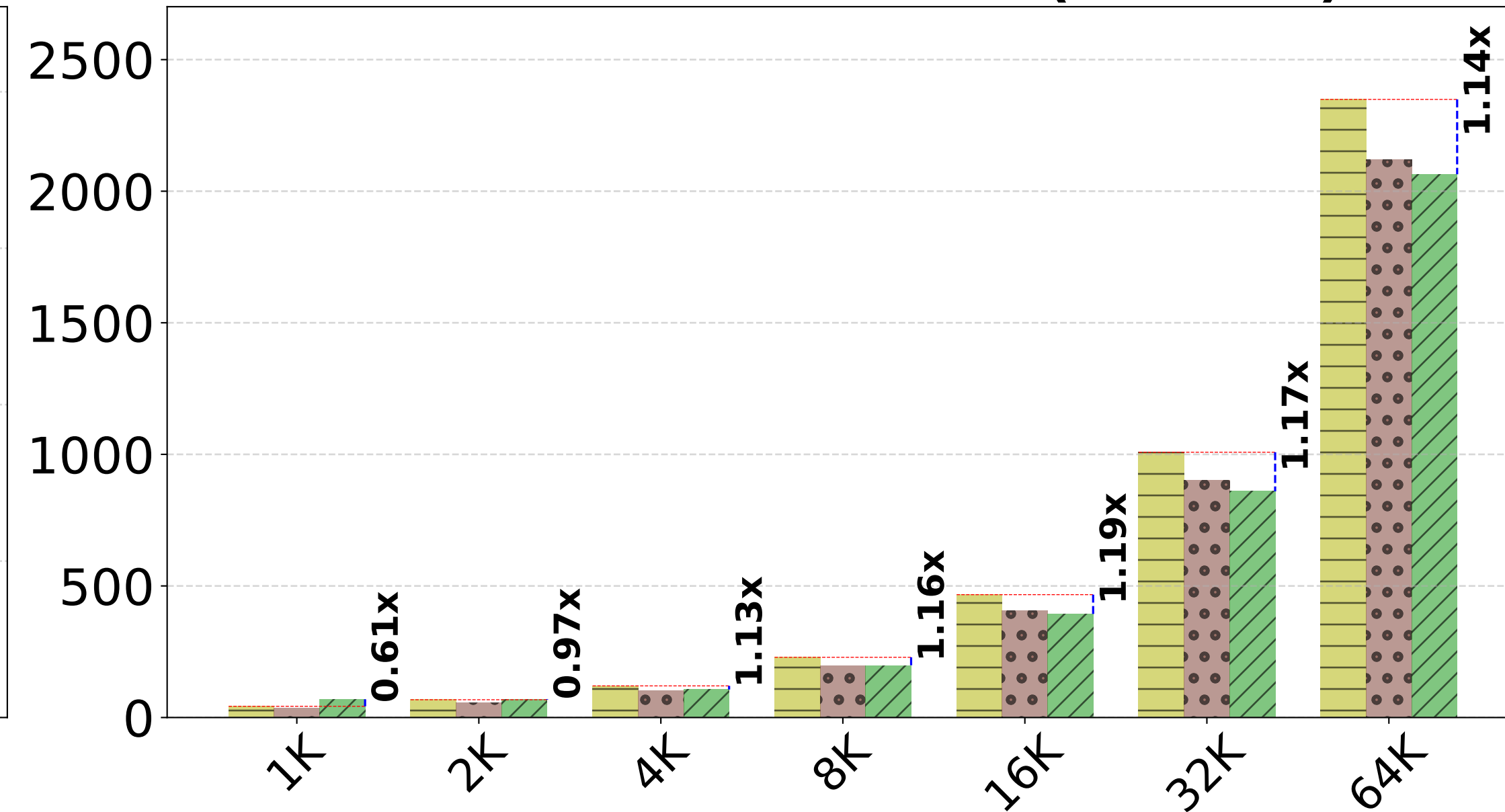
Llama-3.3-70B-Instruct (8x H100)



Qwen2.5-72B-Instruct (8x H100)



Mixtral-8x22B-Instruct (8x H100)



Seq Length