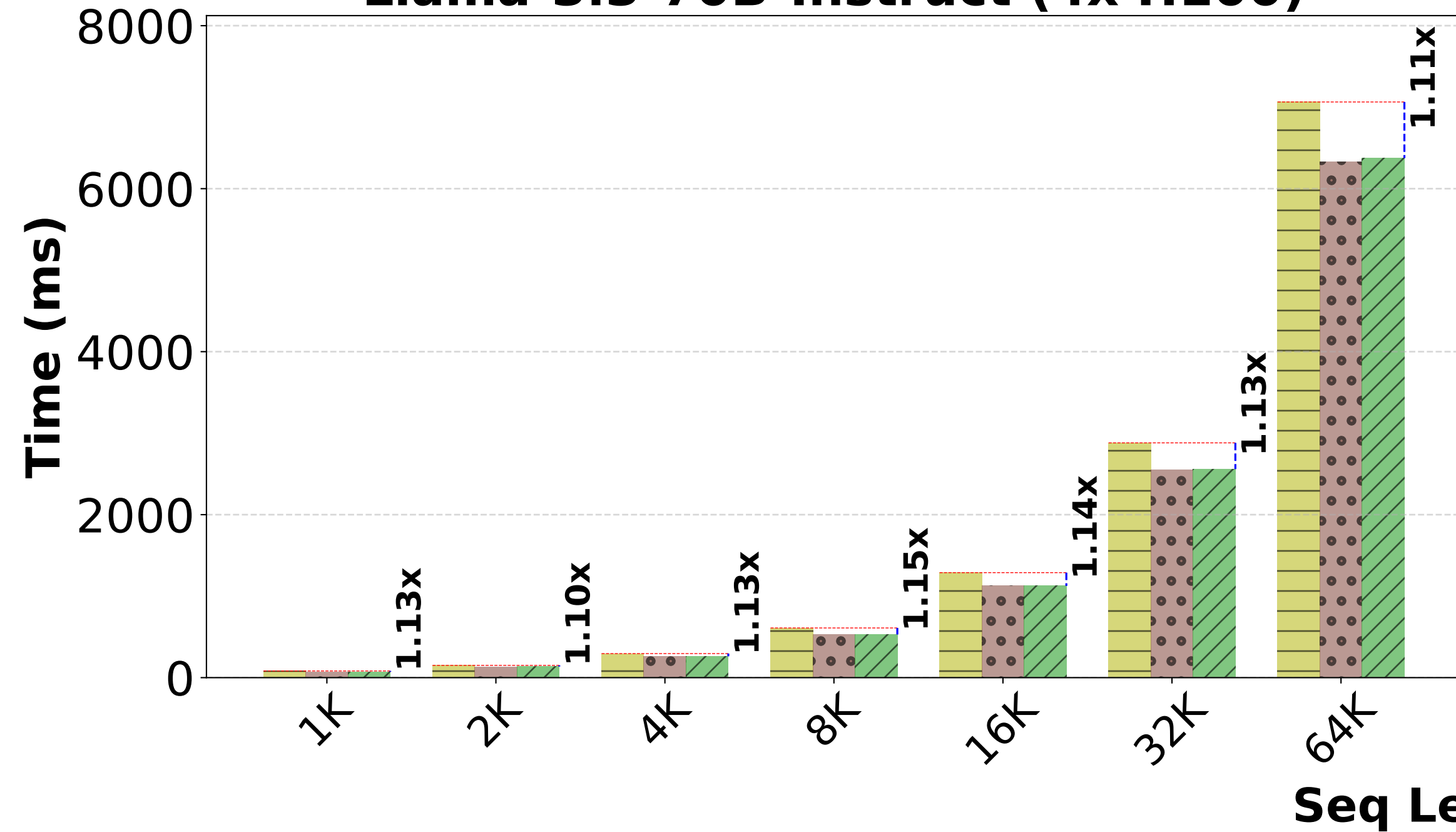


vllm-Multimem vllm-nocomm TokenWeave

Llama-3.3-70B-Instruct (4x H100)



Qwen2.5-72B-Instruct (4x H100)

