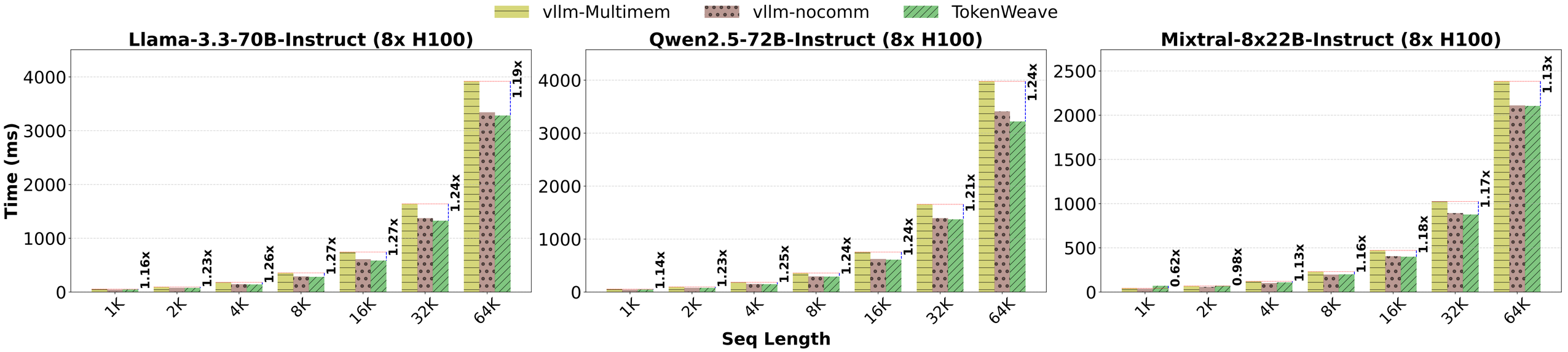
vllm-Multimem
vllm-nocomm
TokenWeave
Llama-3.3-70B-Instruct (8x H100)
Qwen2.5-72B-Instruct (8x H100)
Mixtral-8x22B-Instruct (8x H100)
Time (ms)
Seq Length
1K
2K
4K
8K
16K
32K
64K
1.16x
1.23x
1.26x
1.27x
1.27x
1.24x
1.19x
1.14x
1.23x
1.25x
1.24x
1.24x
1.21x
1.24x
0.62x
0.98x
1.13x
1.16x
1.18x
1.17x
1.13x