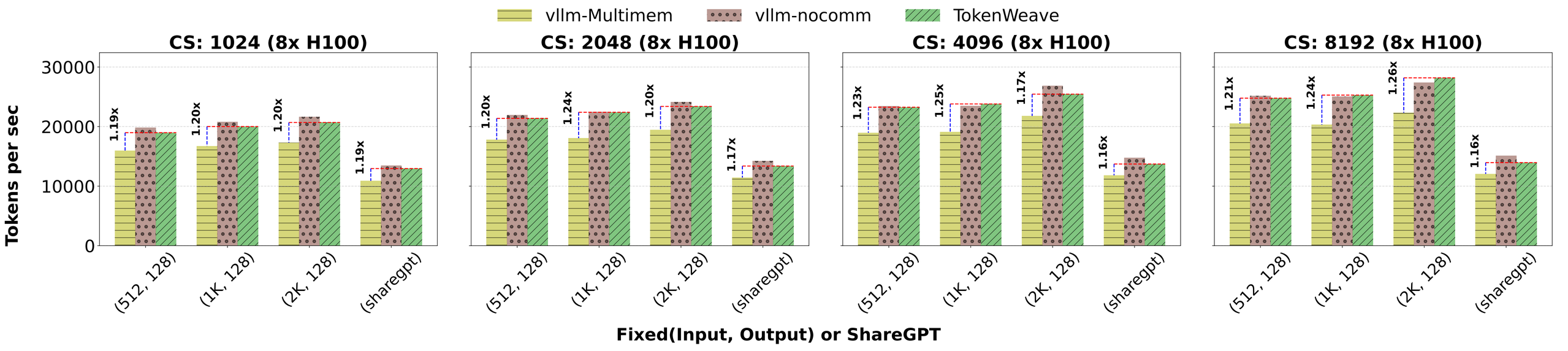
vllm-Multimem
vllm-nocomm
TokenWeave
CS: 1024 (8x H100)
CS: 2048 (8x H100)
CS: 4096 (8x H100)
CS: 8192 (8x H100)
Tokens per sec
0
10000
20000
30000
1.19x
1.20x
1.20x
1.19x
1.20x
1.24x
1.20x
1.17x
1.23x
1.25x
1.17x
1.16x
1.21x
1.24x
1.26x
1.16x
(512, 128)
(1K, 128)
(2K, 128)
(sharegpt)
Fixed(Input, Output) or ShareGPT