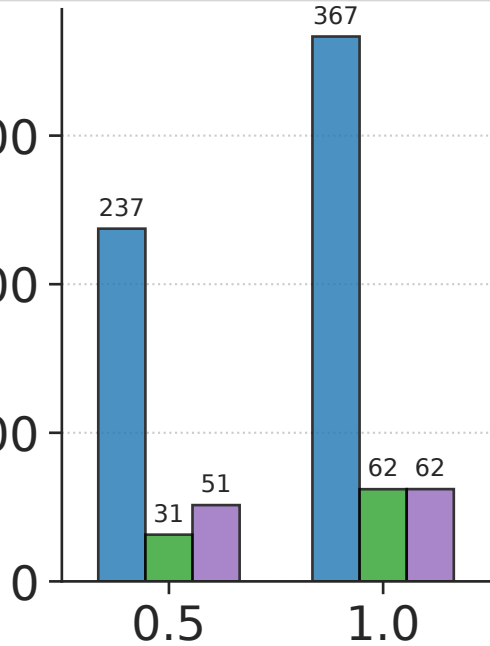


E2E Latency (s)

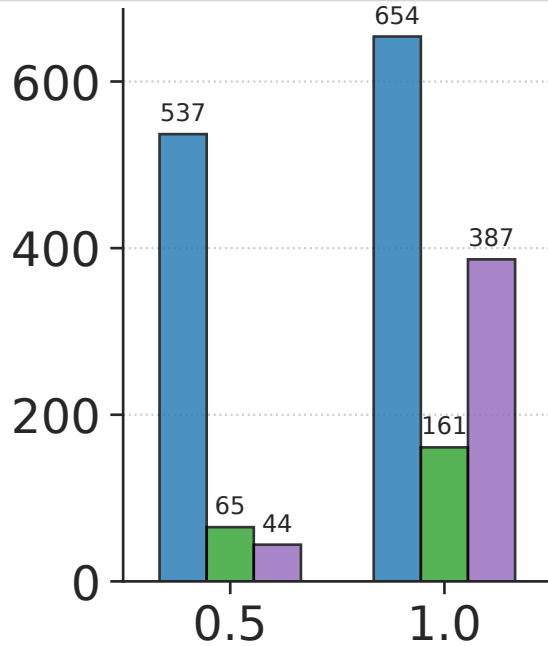
vLLM+SCB

+Delta (N=8)

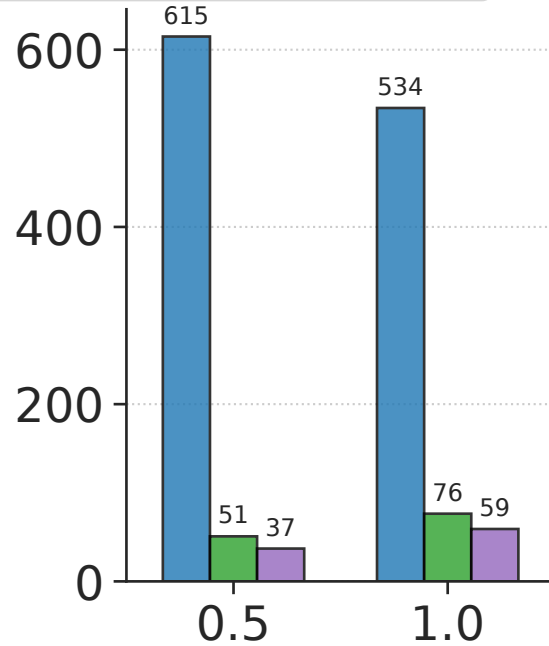
+Delta (N=12)



(a) Azure



(b) Uniform



(c) Zipf:1.5