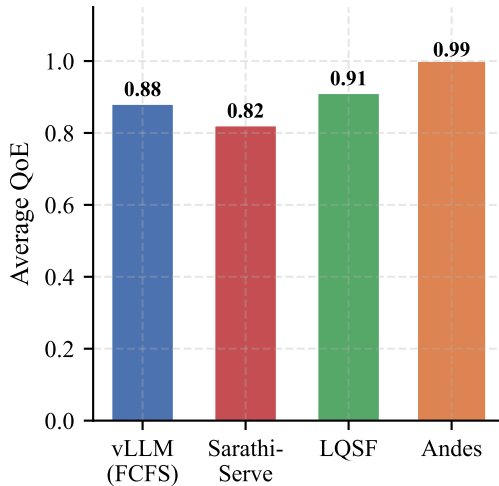
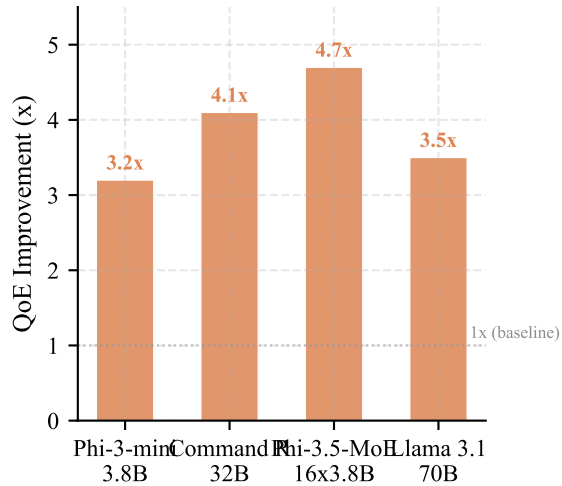


(a) QoE on BurstGPT Trace



(b) QoE Improvement vs vLLM



(c) Resource Efficiency

