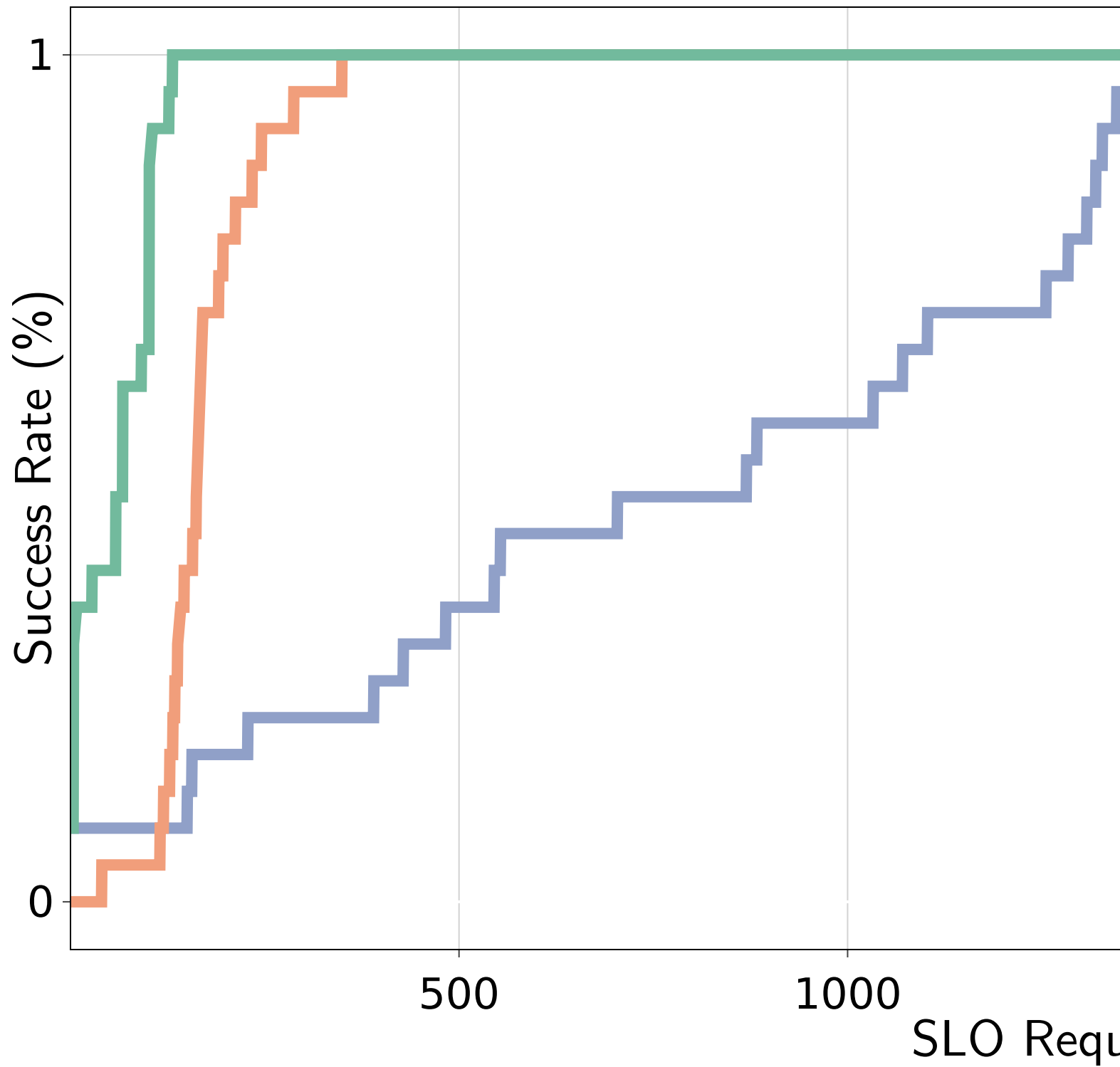


Time to First Token (TTFT)



End-to-End Latency

