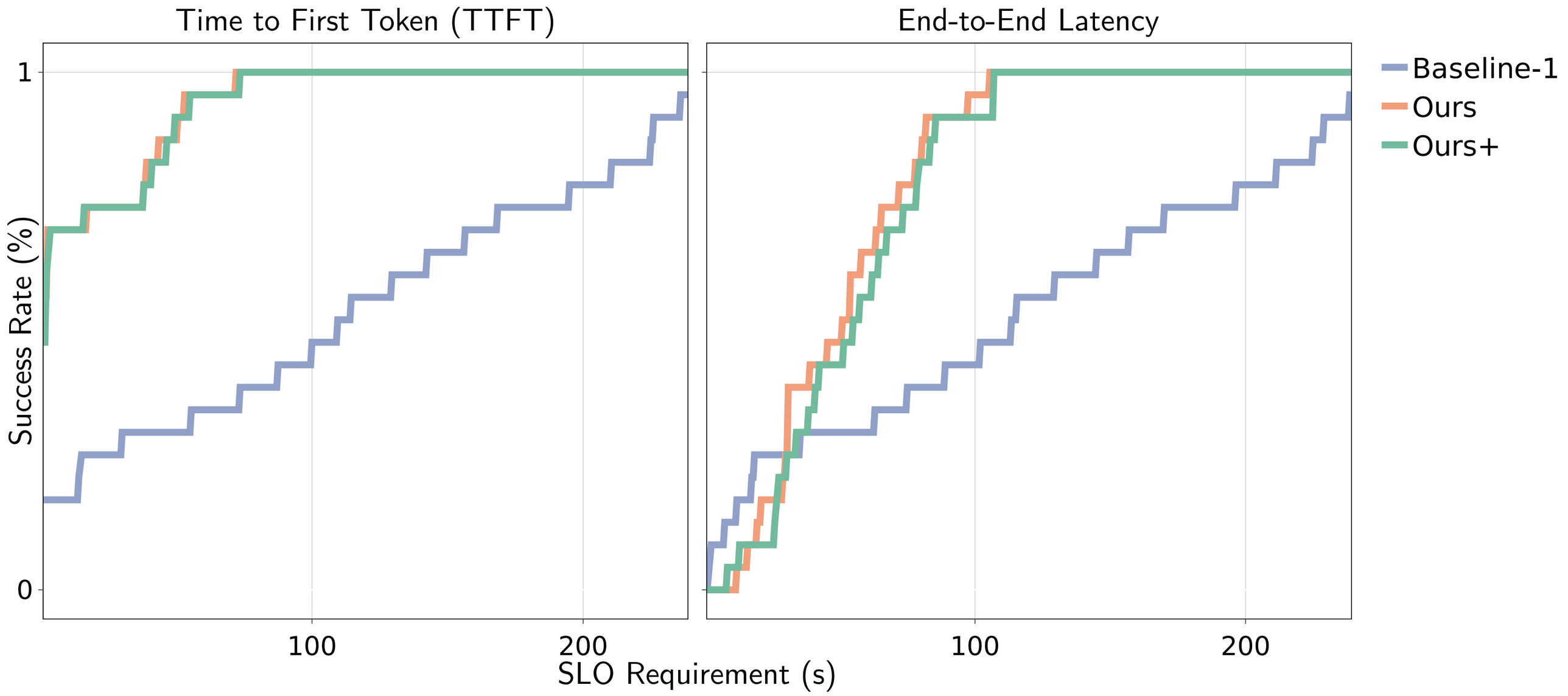
Time to First Token (TTFT)
End-to-End Latency
Success Rate (%)
1
0
100
200
100
200
SLO Requirement (s)
Baseline-1
Ours
Ours+