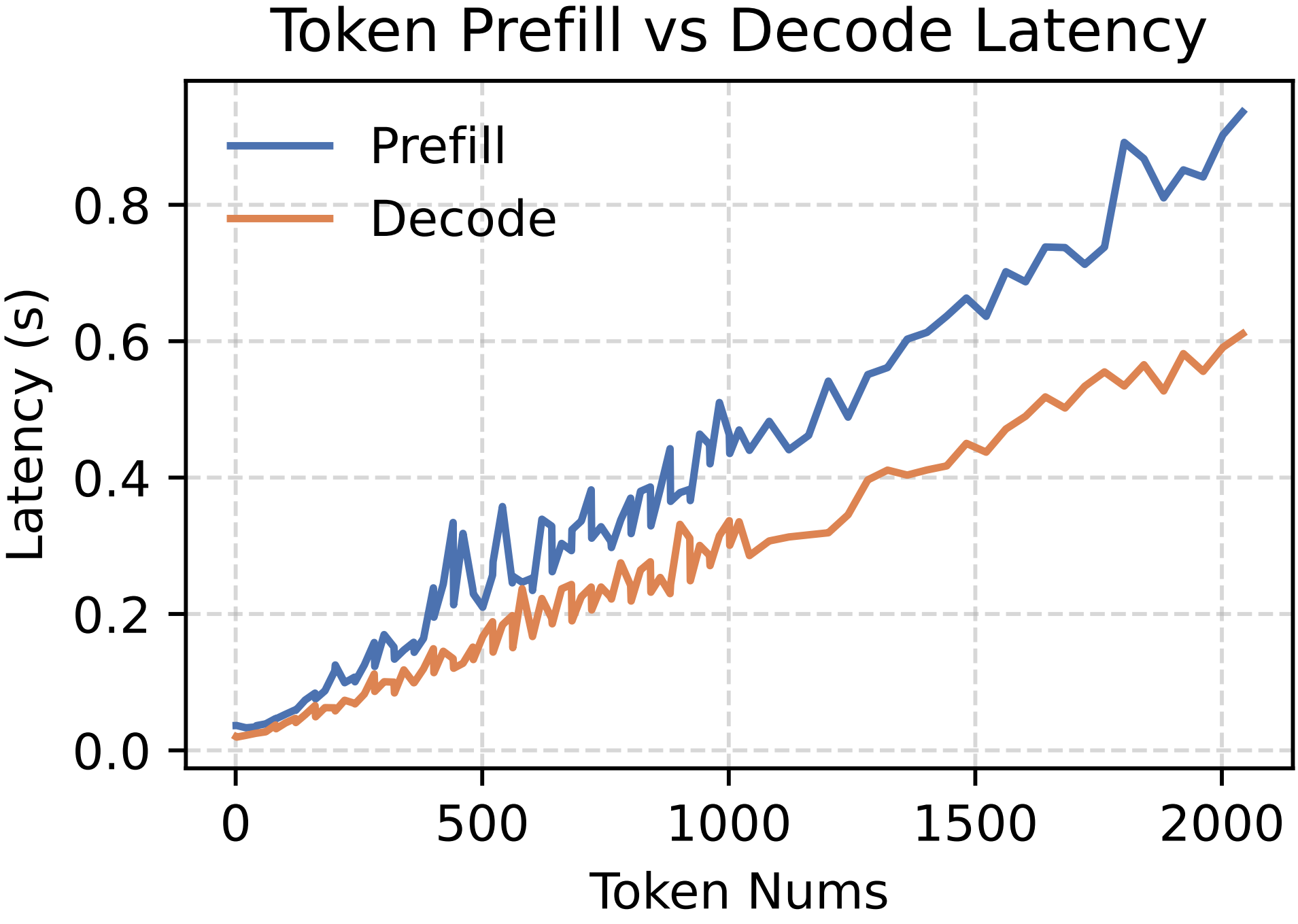
Token Prefill vs Decode Latency
Prefill
Decode
Latency (s)
0.0
0.2
0.4
0.6
0.8
0
500
1000
1500
2000
Token Nums