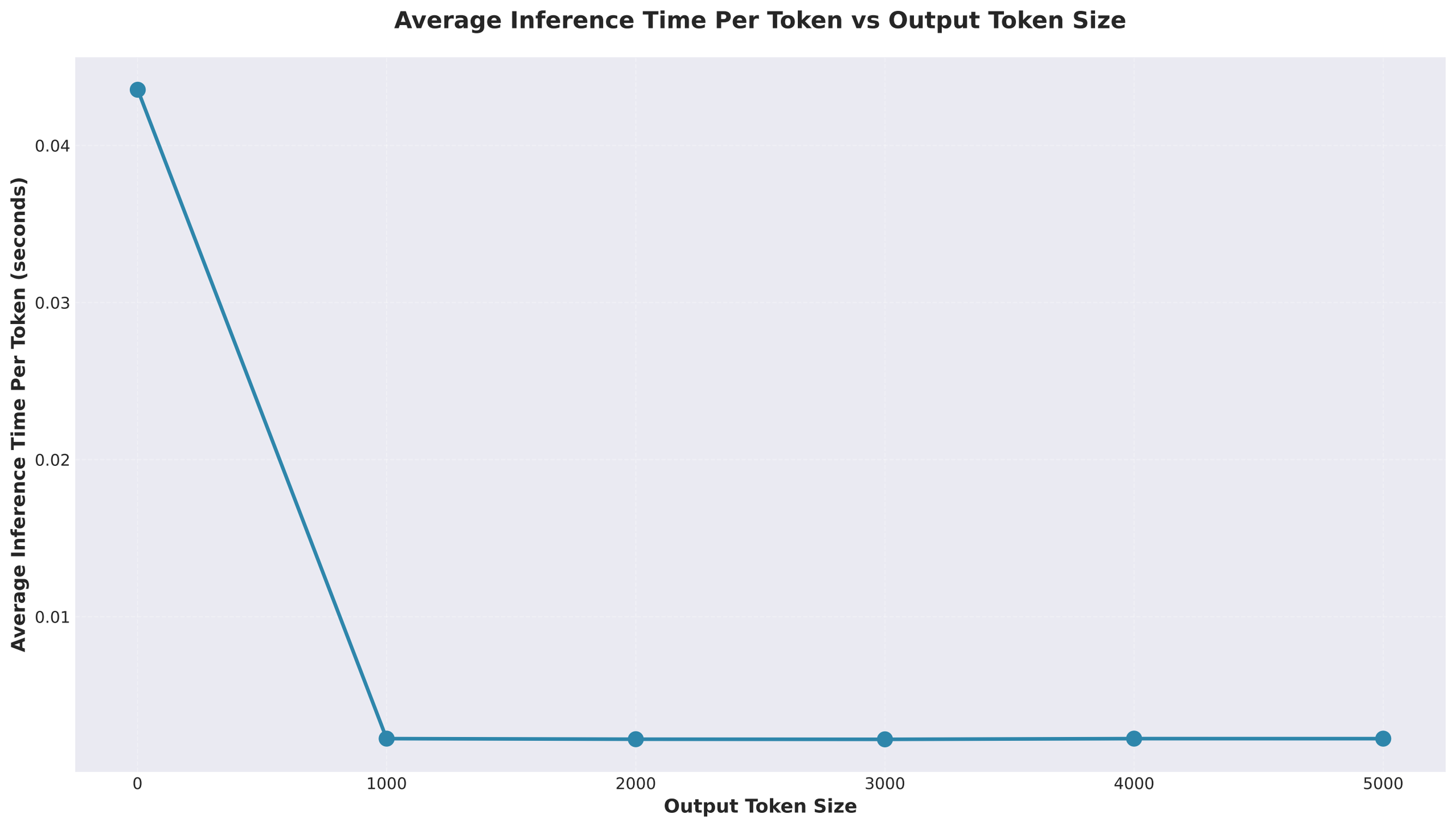
Average Inference Time Per Token vs Output Token Size
Average Inference Time Per Token (seconds)
0.04
0.03
0.02
0.01
0
1000
2000
3000
4000
5000
Output Token Size