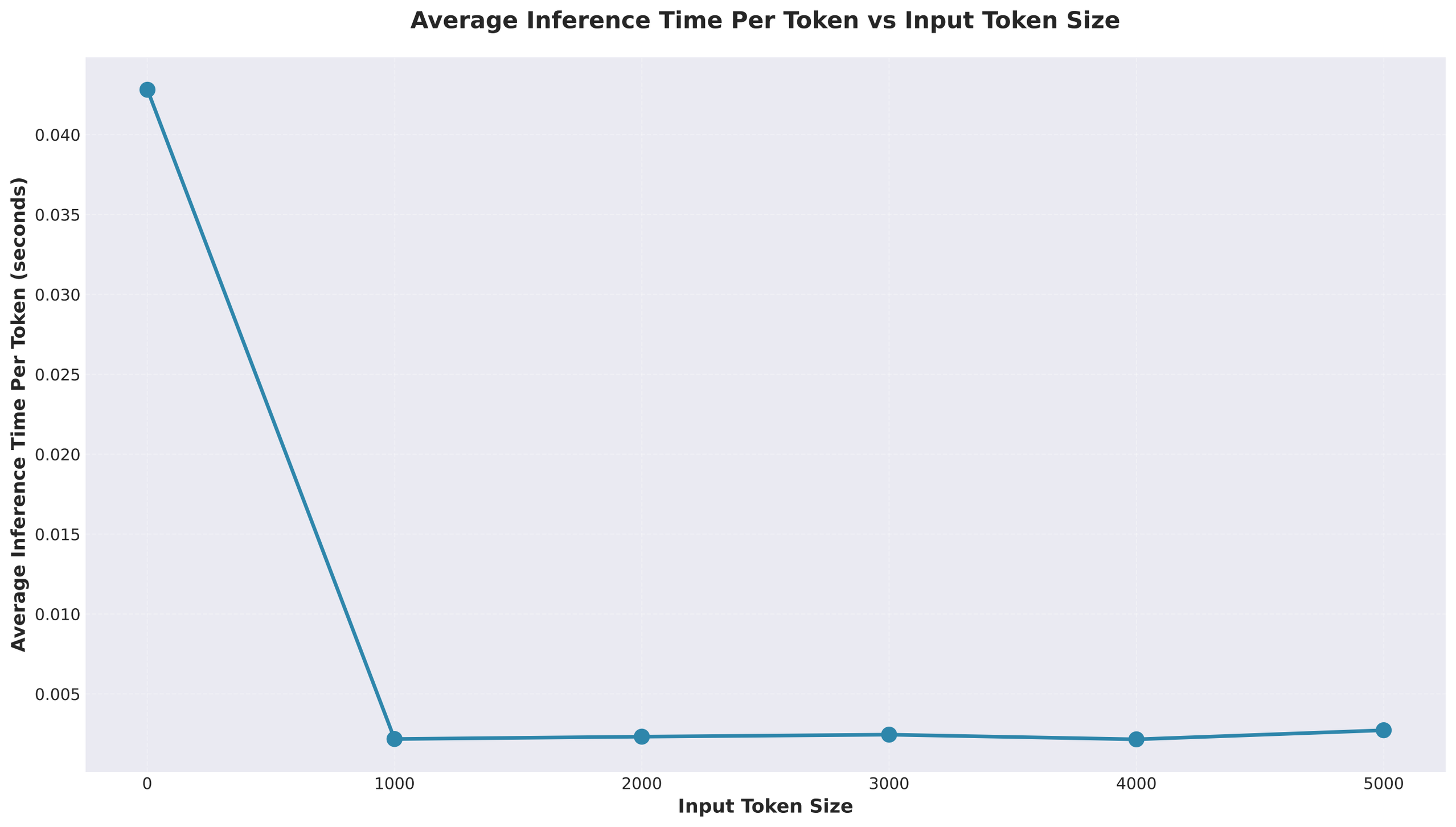
Average Inference Time Per Token vs Input Token Size
Average Inference Time Per Token (seconds)
0.040
0.035
0.030
0.025
0.020
0.015
0.010
0.005
0
1000
2000
3000
4000
5000
Input Token Size