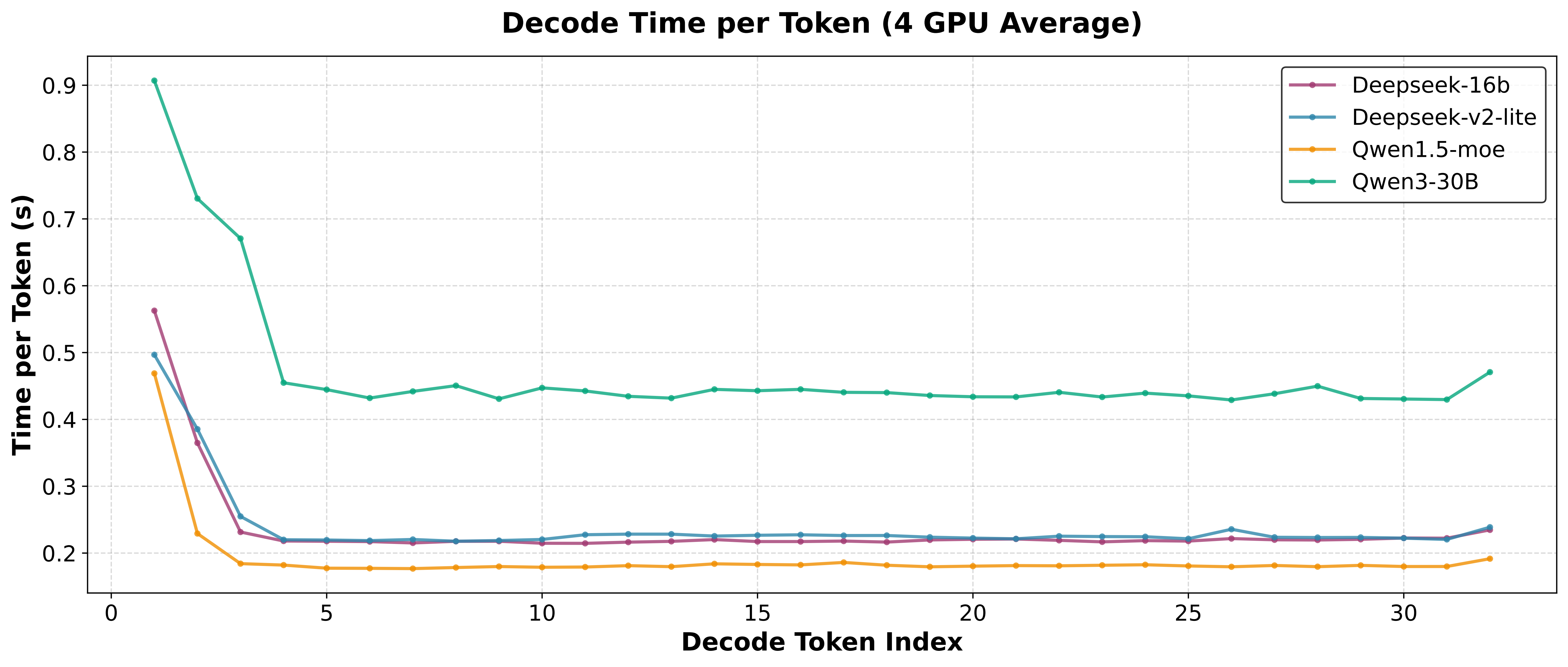
Decode Time per Token (4 GPU Average)
Time per Token (s)
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
Decode Token Index
0
5
10
15
20
25
30
Deepseek-16b
Deepseek-v2-lite
Qwen1.5-moe
Qwen3-30B