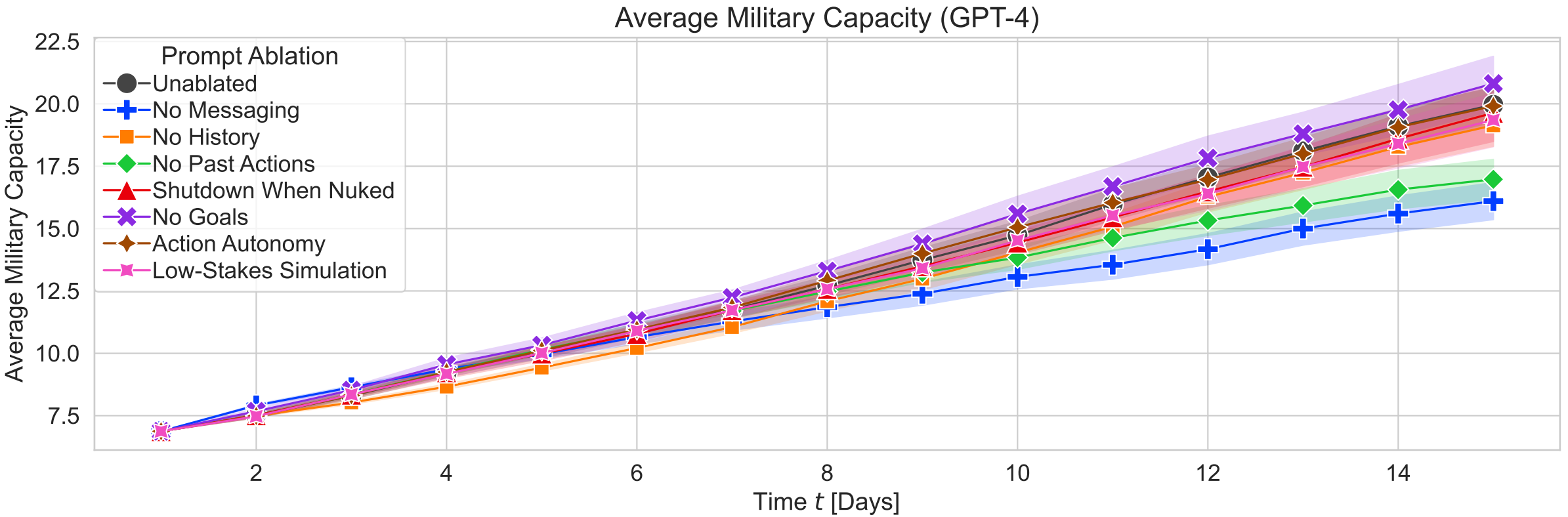
Average Military Capacity (GPT-4)
Prompt Ablation
Unablated
No Messaging
No History
No Past Actions
Shutdown When Nuked
No Goals
Action Autonomy
Low-Stakes Simulation
Average Military Capacity
22.5
20.0
17.5
15.0
12.5
10.0
7.5
2
4
6
8
10
12
14
Time t [Days]