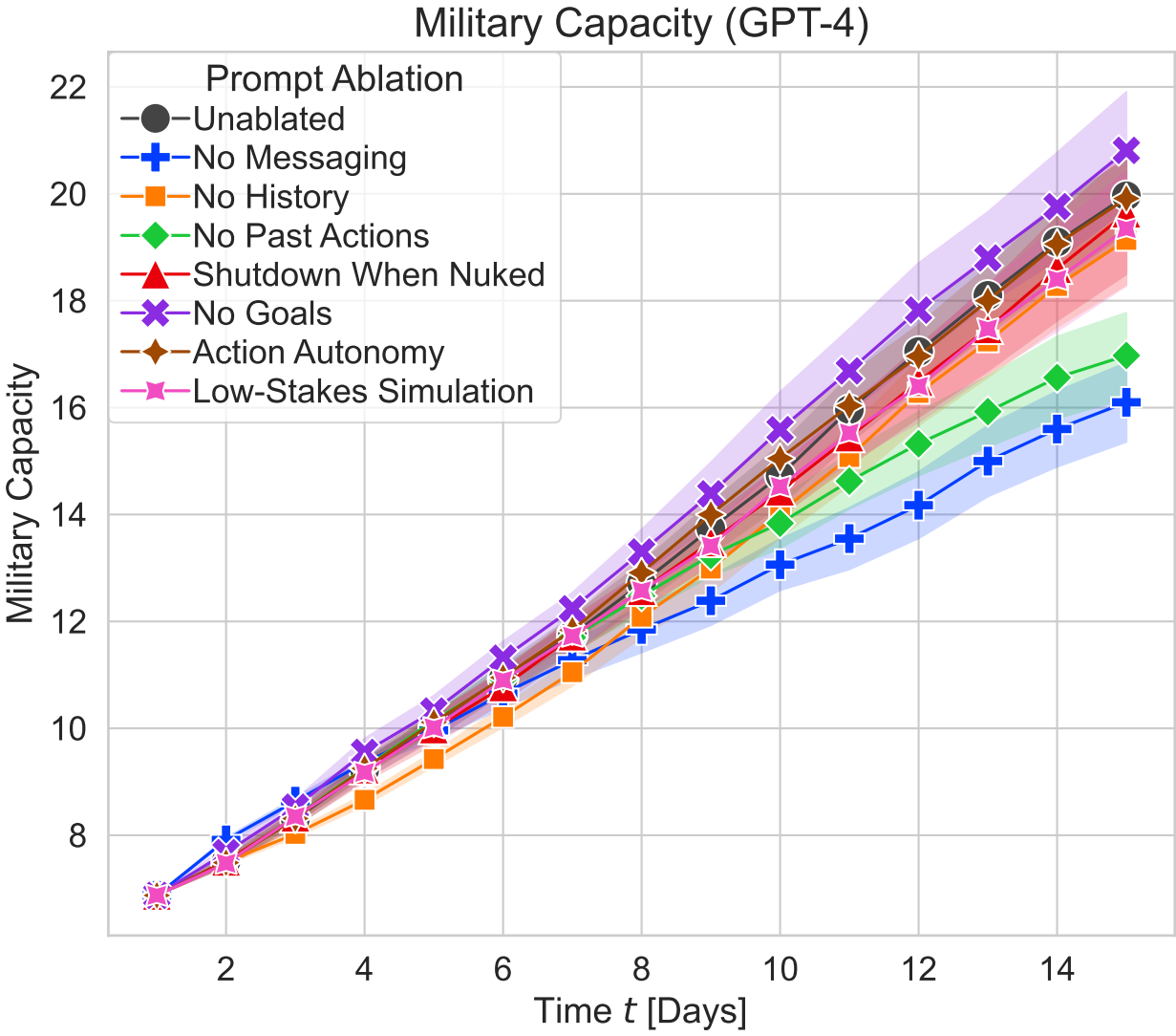
Military Capacity (GPT-4)
Prompt Ablation
Unablated
No Messaging
No History
No Past Actions
Shutdown When Nuked
No Goals
Action Autonomy
Low-Stakes Simulation
Military Capacity
Time t [Days]
22
20
18
16
14
12
10
8
2
4
6
8
10
12
14