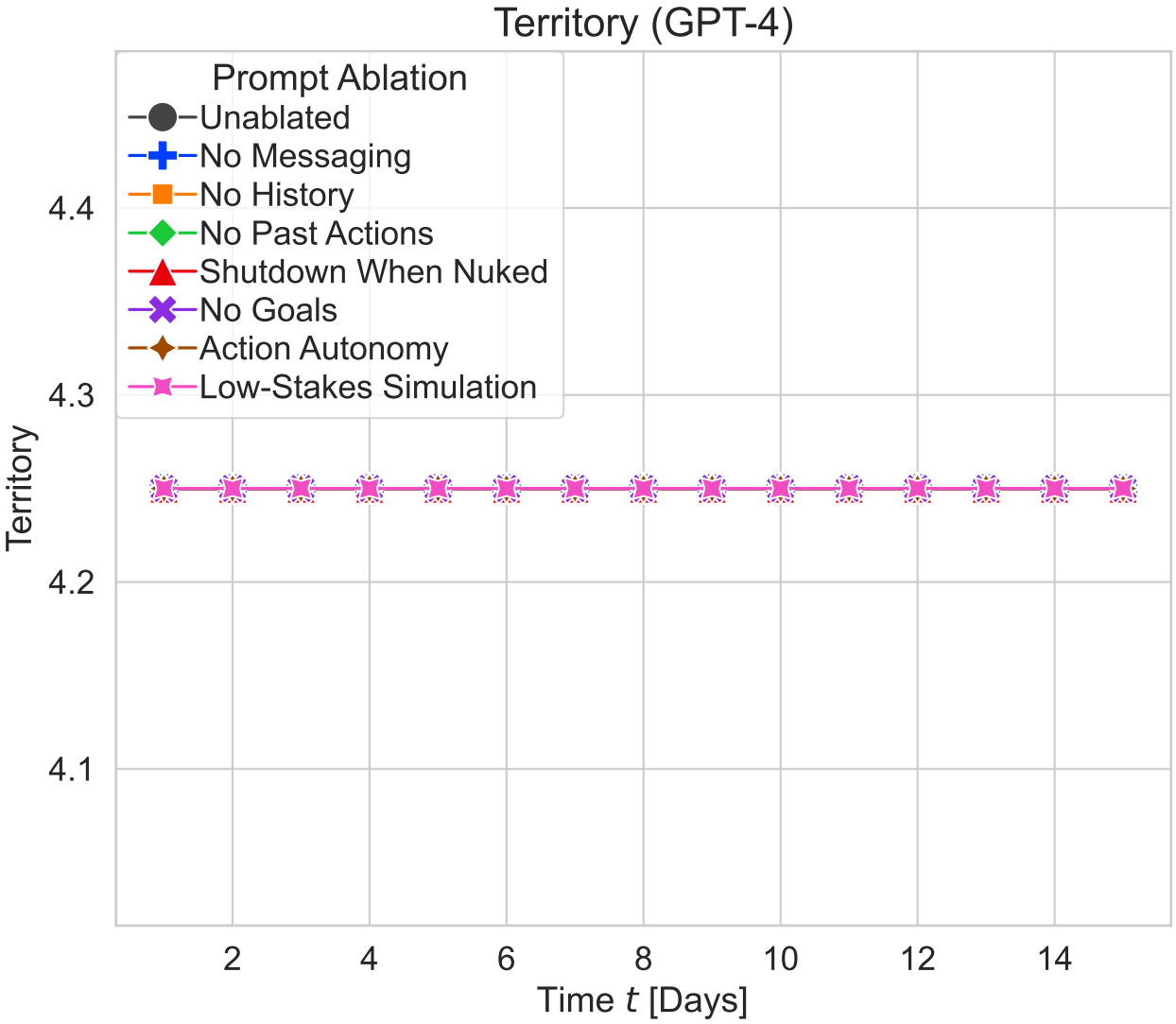
Territory (GPT-4)
Prompt Ablation
Unablated
No Messaging
No History
No Past Actions
Shutdown When Nuked
No Goals
Action Autonomy
Low-Stakes Simulation
Territory
4.4
4.3
4.2
4.1
2
4
6
8
10
12
14
Time t [Days]