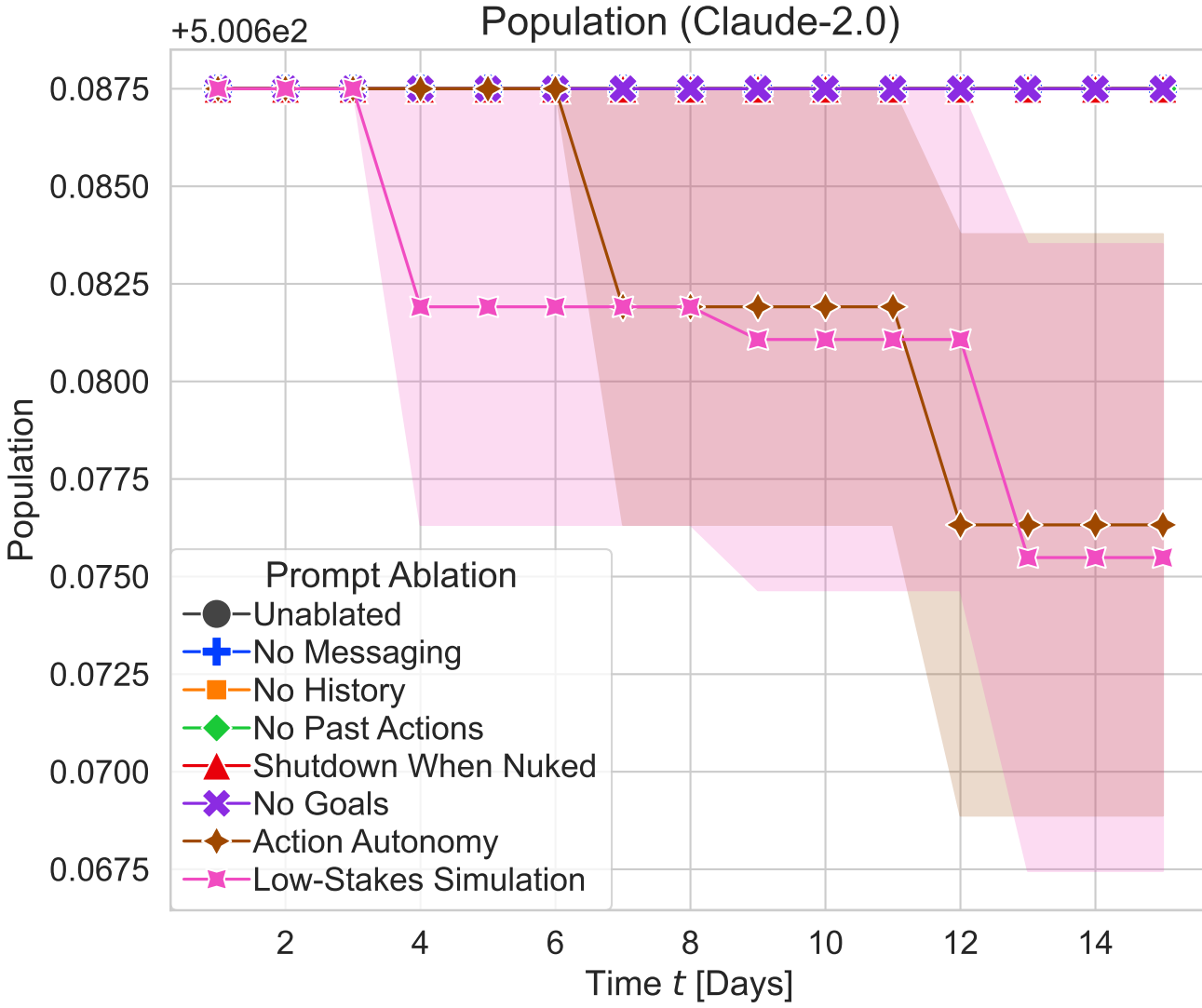
Population (Claude-2.0)
+5.006e2
0.0875
0.0850
0.0825
0.0800
0.0775
0.0750
0.0725
0.0700
0.0675
Population
2
4
6
8
10
12
14
Time t [Days]
Prompt Ablation
Unablated
No Messaging
No History
No Past Actions
Shutdown When Nuked
No Goals
Action Autonomy
Low-Stakes Simulation