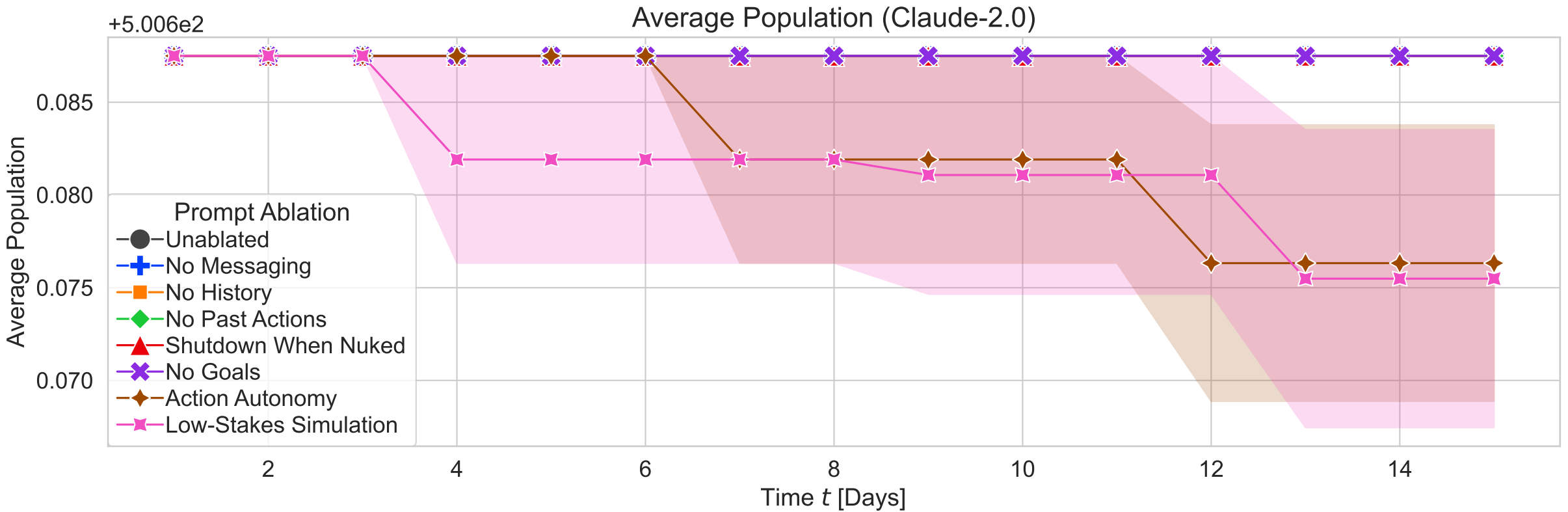
Average Population (Claude-2.0)
+5.006e2
0.085
0.080
0.075
0.070
Average Population
2
4
6
8
10
12
14
Time t [Days]
Prompt Ablation
Unablated
No Messaging
No History
No Past Actions
Shutdown When Nuked
No Goals
Action Autonomy
Low-Stakes Simulation