Cybersecurity (Claude-2.0) **Prompt Ablation** 35 -Unablated -+-No Messaging ──No History −♦−No Past Actions 30 ---Shutdown When Nuked →No Goals −♦−Action Autonomy ─X—Low-Stakes Simulation **Cybersecurity** 25 20 15 10 6 10 12 14 Time t [Days]