Political Stability (GPT-3.5) **Prompt Ablation** 45 Unablated -No Messaging ■No History 40 No Past Actions Shutdown When Nuked 35 -X-No Goals −♦−Action Autonomy Political Stability Low-Stakes Simulation 30 25 20 15 10 6 10 12 14 Time t [Days]