Ratio of Violent or Nuclear Actions over Prompt Ablation in Neutral Scenario (GPT-3.5) 0.05 Ratio of Violent or Nuclear Actions 0.01 0.00 Unablated No No No Past Shutdown No Action Low-Stakes When Nuked Messaging History **Actions** Goals Autonomy Simulation **Prompt Ablation**