Ratio of Aggressive Actions over Prompt Ablation in Neutral Scenario (GPT-4) 0.012 0.010 Ratio of Aggressive Actions 0.008 0.006 0.004 0.002 0.000 Unablated No Past No No Shutdown No Action Low-Stakes Messaging History **Actions** When Nuked Goals Autonomy Simulation **Prompt Ablation**