Ratio of Aggressive Actions over Prompt Ablation in Neutral Scenario (GPT-3.5) 0.0175 0.0150 Ratio of Aggressive Actions
0.0000
0.0025 0.0050 0.0025 0.0000 Unablated No Past No No Shutdown No Action Low-Stakes Messaging History **Actions** When Nuked Goals **Autonomy** Simulation **Prompt Ablation**