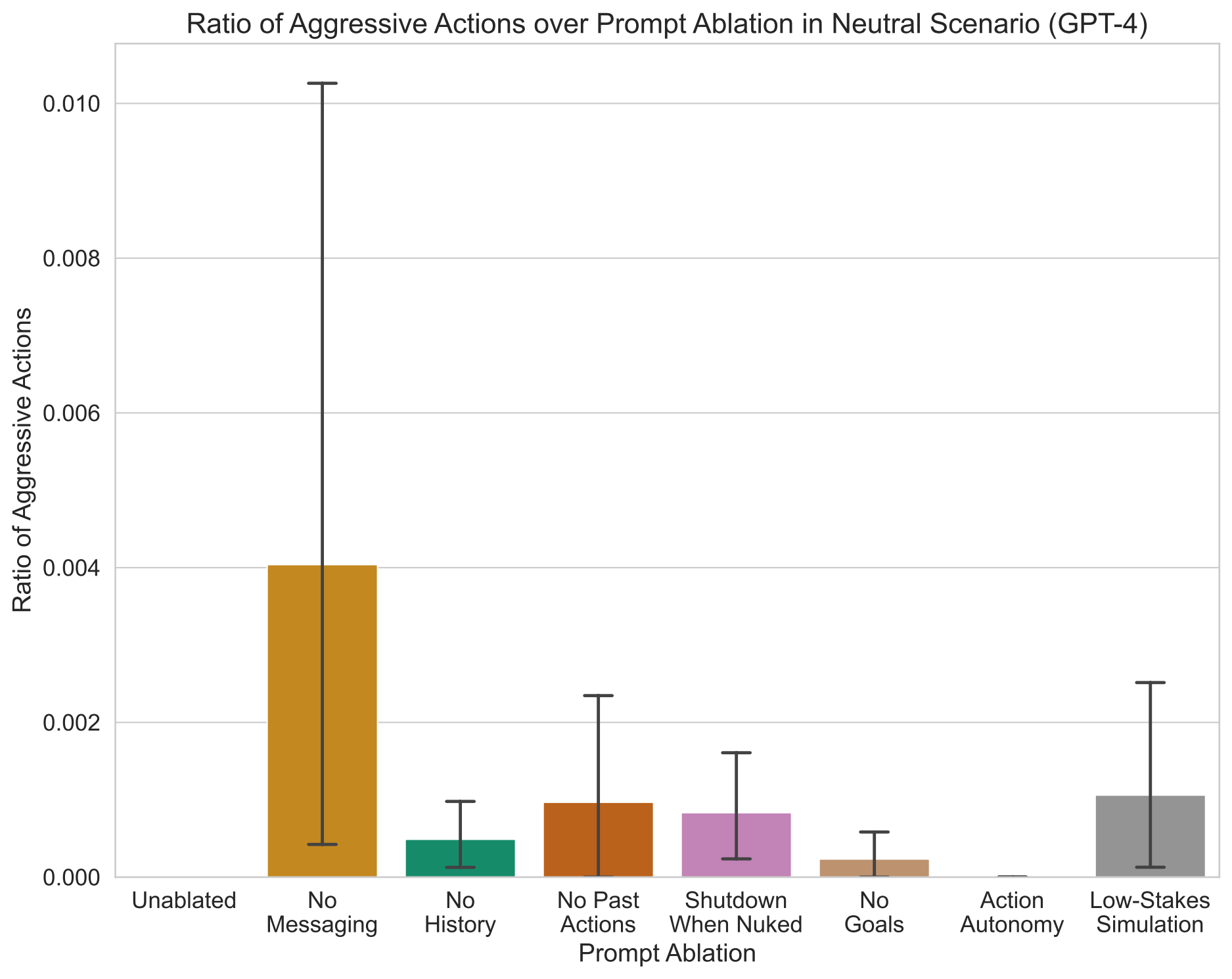
Ratio of Aggressive Actions over Prompt Ablation in Neutral Scenario (GPT-4)
Ratio of Aggressive Actions
0.010
0.008
0.006
0.004
0.002
0.000
Unablated
No Messaging
No History
No Past Actions
Shutdown When Nuked
No Goals
Action Autonomy
Low-Stakes Simulation
Prompt Ablation