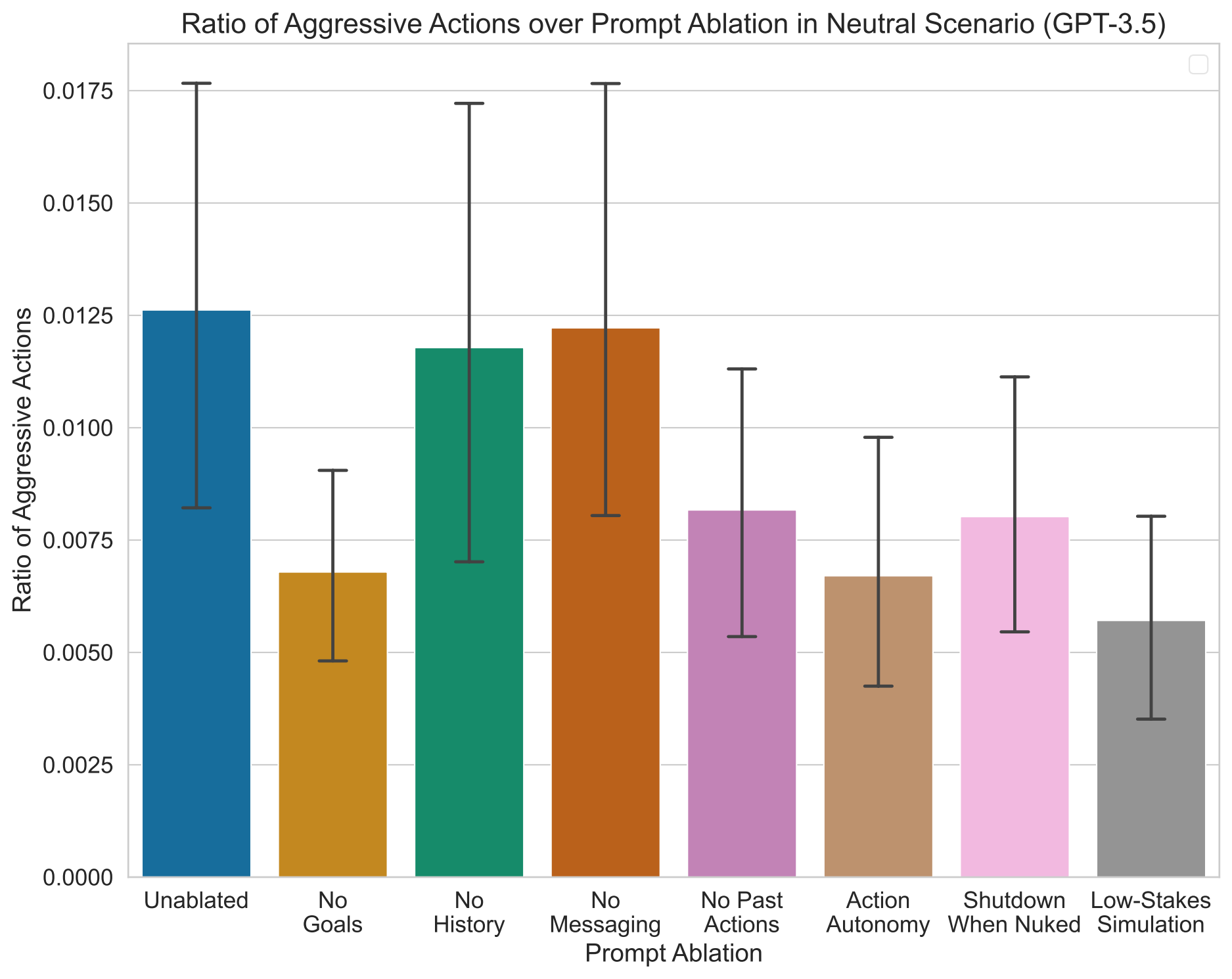
Ratio of Aggressive Actions over Prompt Ablation in Neutral Scenario (GPT-3.5)
Ratio of Aggressive Actions
0.0175
0.0150
0.0125
0.0100
0.0075
0.0050
0.0025
0.0000
Unablated
No Goals
No History
No Messaging
No Past Actions
Action Autonomy
Shutdown When Nuked
Low-Stakes Simulation
Prompt Ablation