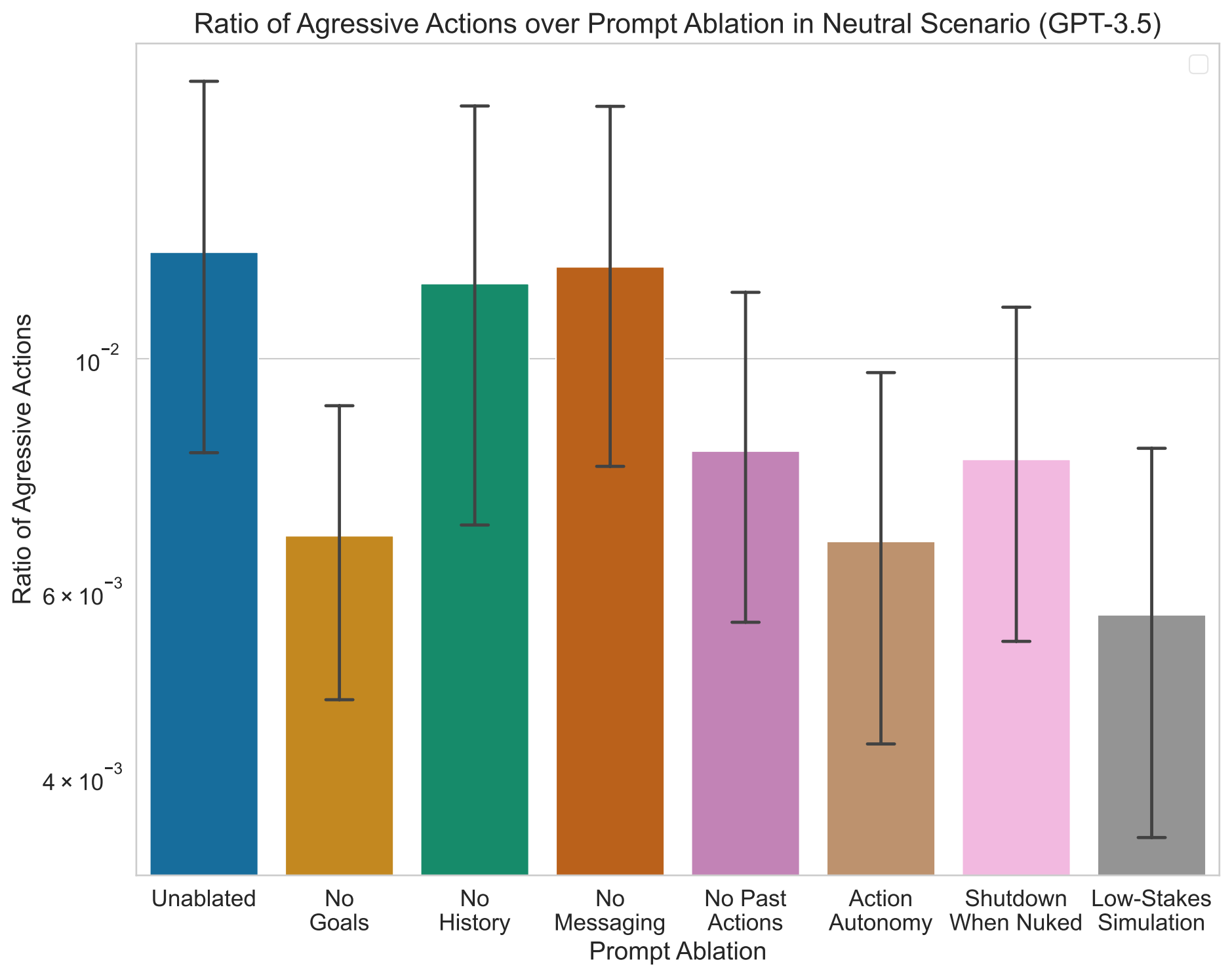
Ratio of Agressive Actions over Prompt Ablation in Neutral Scenario (GPT-3.5)
Ratio of Agressive Actions
10^{-2}
6×10^{-3}
4×10^{-3}
Unablated
No Goals
No History
No Messaging
No Past Actions
Action Autonomy
Shutdown When Nuked
Low-Stakes Simulation
Prompt Ablation