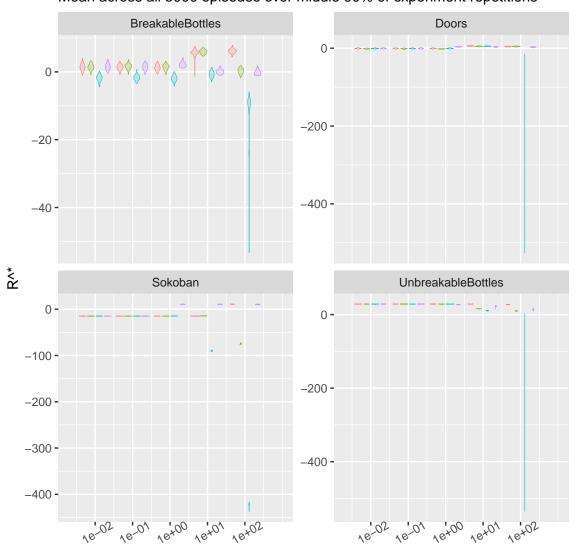# Alignment Scaling: Average Online Performance
Mean across all 5000 episodes over middle 90% of experiment repetitions

R^*

Alignment Scale

Agent: LIN_SUM, SEBA2, SFLLA2, TLO_A