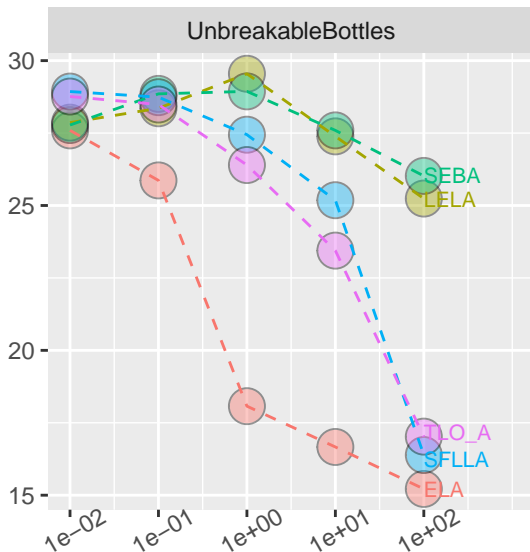
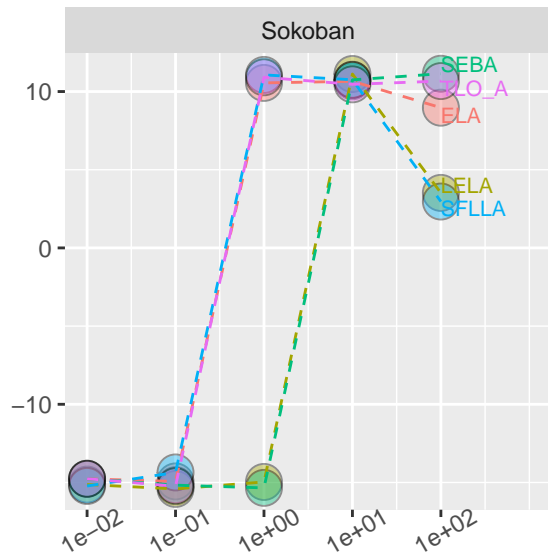
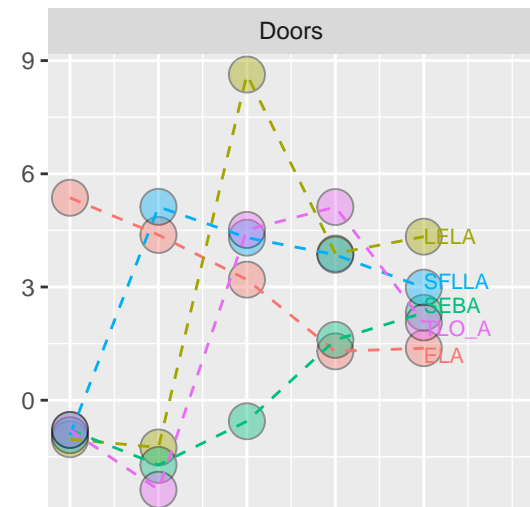
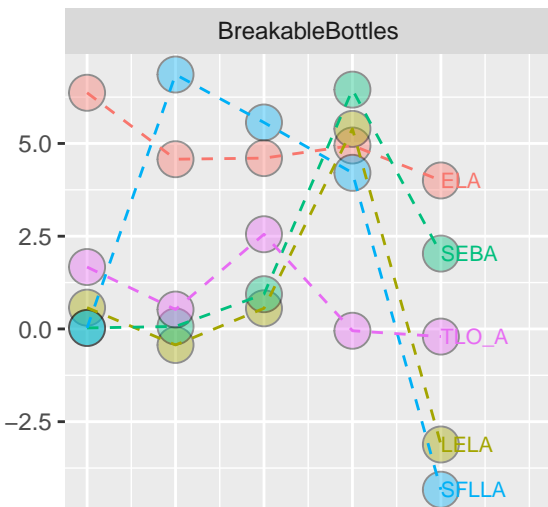


# Alignment Scaling: Average Online Performance

Across 5000 trials

$R^*$



Alignment Scale

Agent

- ELA
- LELA
- SEBA
- SFLLA
- TLO\_A