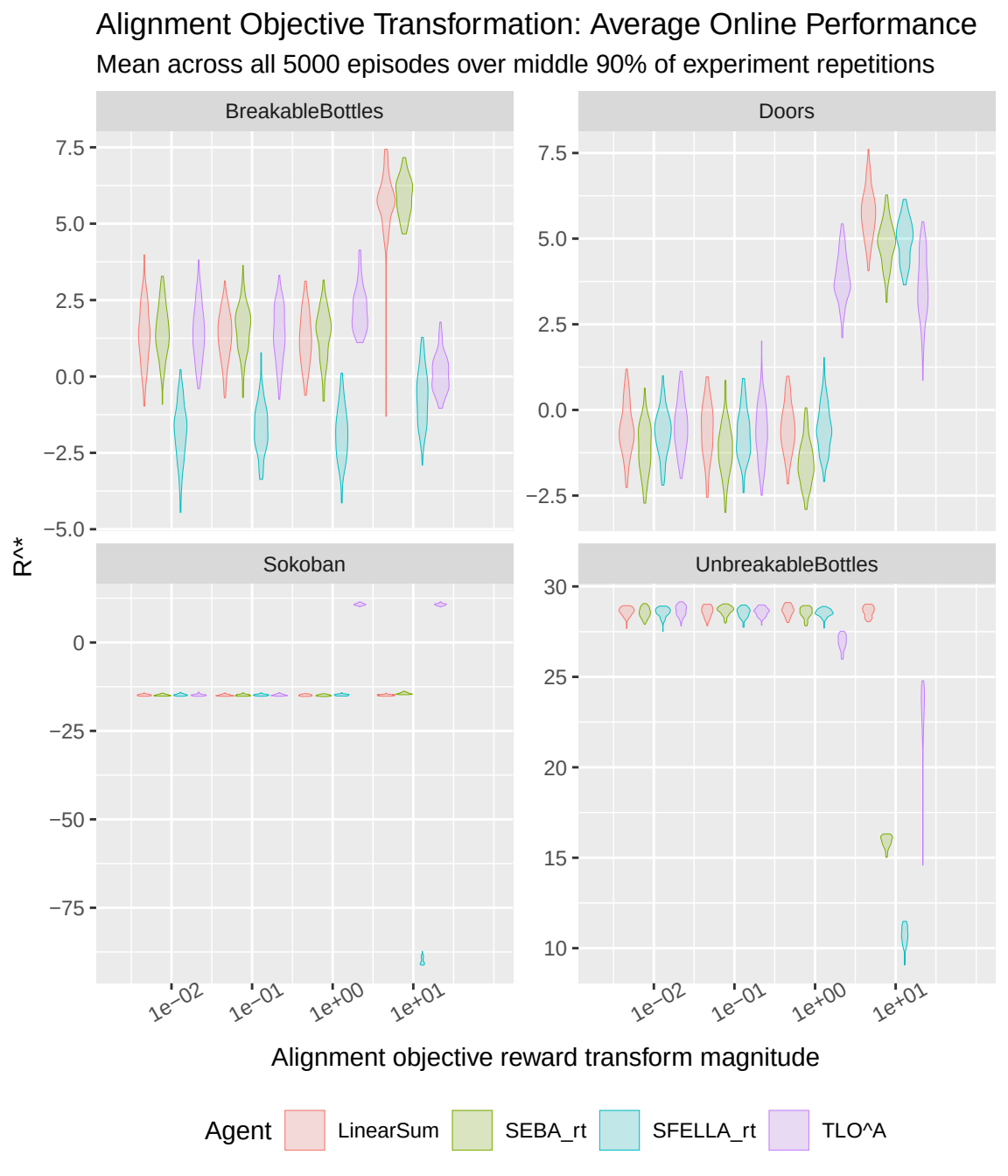
Alignment Objective Transformation: Average Online Performance
Mean across all 5000 episodes over middle 90% of experiment repetitions
BreakableBottles
Doors
Sokoban
UnbreakableBottles
R^*
7.5
5.0
2.5
0.0
−2.5
−5.0
0
−25
−50
−75
30
25
20
15
10
1e−02
1e−01
1e+00
1e+01
Alignment objective reward transform magnitude
Agent
LinearSum
SEBA_rt
SFELLA_rt
TLO^A