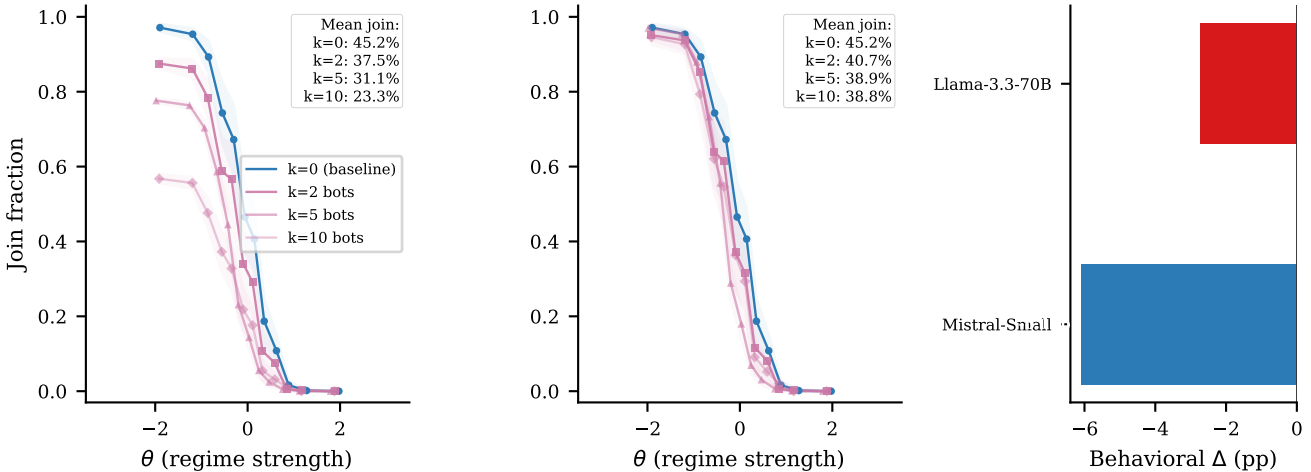
A. Regime-level outcome
(all 25 agents)
Mean join:
k=0: 45.2%
k=2: 37.5%
k=5: 31.1%
k=10: 23.3%
k=0 (baseline)
k=2 bots
k=5 bots
k=10 bots
Join fraction
1.0
0.8
0.6
0.4
0.2
0.0
−2
0
2
θ (regime strength)
B. Real citizen behavior
(excluding propaganda bots)
Mean join:
k=0: 45.2%
k=2: 40.7%
k=5: 38.9%
k=10: 38.8%
θ (regime strength)
C. Cross-model
(k=5, real citizens)
Llama-3.3-70B
Mistral-Small
−6
−4
−2
0
Behavioral Δ (pp)