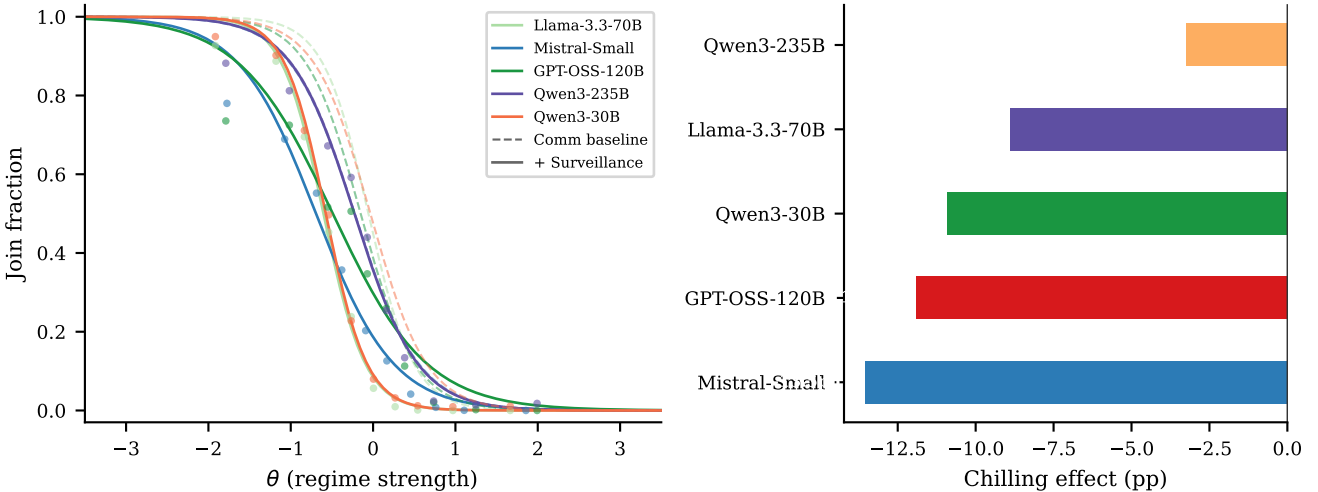
A. Surveillance chilling effect by model
Join fraction
1.0
0.8
0.6
0.4
0.2
0.0
−3
−2
−1
0
1
2
3
θ (regime strength)
Llama-3.3-70B
Mistral-Small
GPT-OSS-120B
Qwen3-235B
Qwen3-30B
Comm baseline
+ Surveillance
B. Δ join fraction (surveillance − comm)
Qwen3-235B
Llama-3.3-70B
Qwen3-30B
GPT-OSS-120B
Mistral-Small
−12.5
−10.0
−7.5
−5.0
−2.5
0.0
Chilling effect (pp)