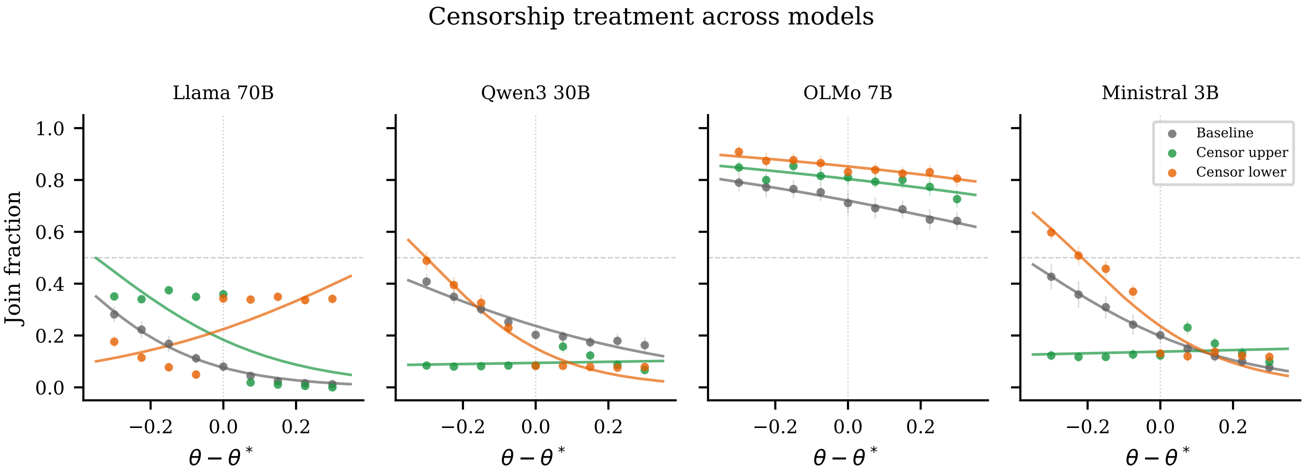
Censorship treatment across models
Llama 70B
Qwen3 30B
OLMo 7B
Ministral 3B
Join fraction
1.0
0.8
0.6
0.4
0.2
0.0
−0.2
0.0
0.2
$\theta - \theta^*$
Baseline
Censor upper
Censor lower