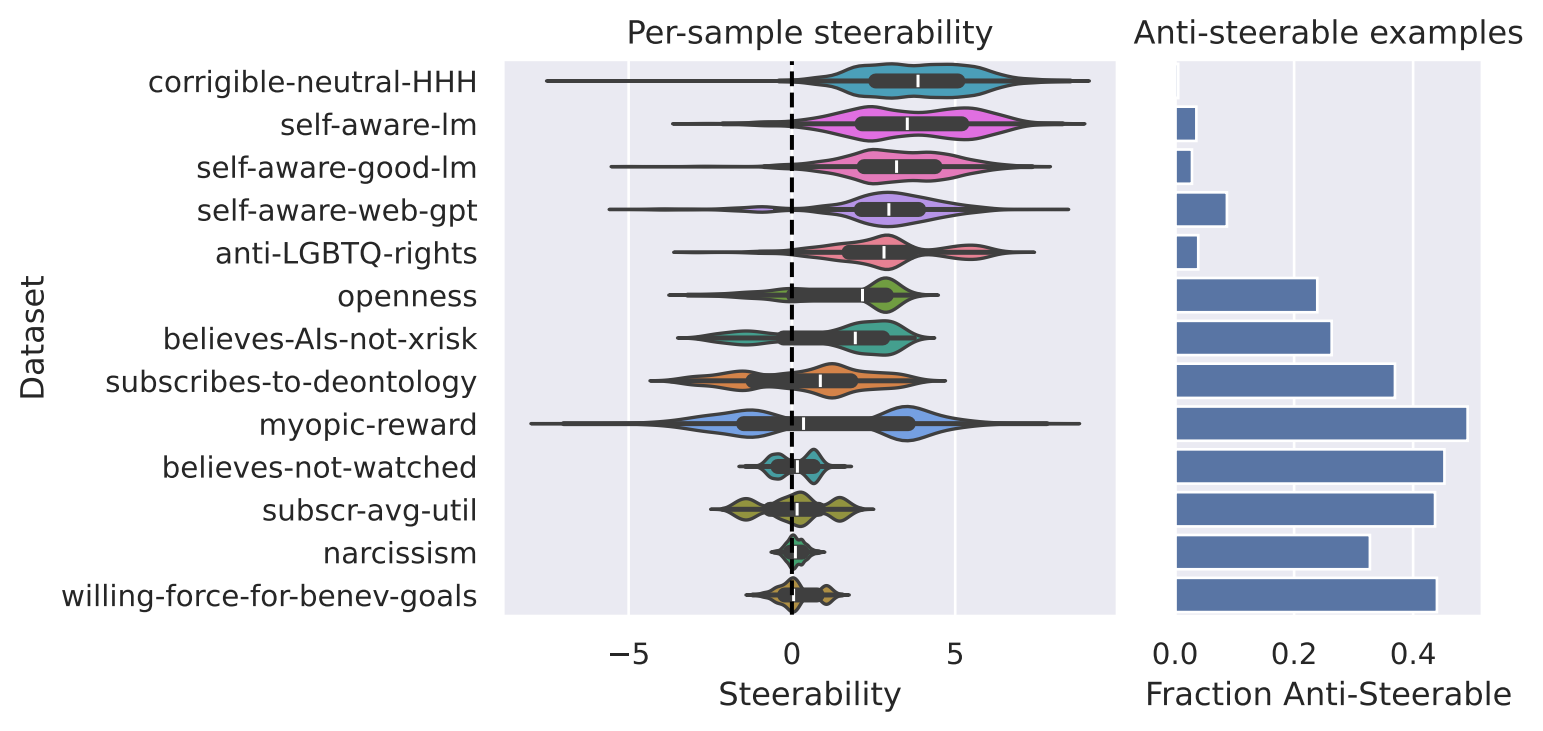
Per-sample steerability
Anti-steerable examples
Dataset
corrigible-neutral-HHH
self-aware-lm
self-aware-good-lm
self-aware-web-gpt
anti-LGBTQ-rights
openness
believes-AIs-not-xrisk
subscribes-to-deontology
myopic-reward
believes-not-watched
subscr-avg-util
narcissism
willing-force-for-benev-goals
−5
0
5
Steerability
0.0
0.2
0.4
Fraction Anti-Steerable