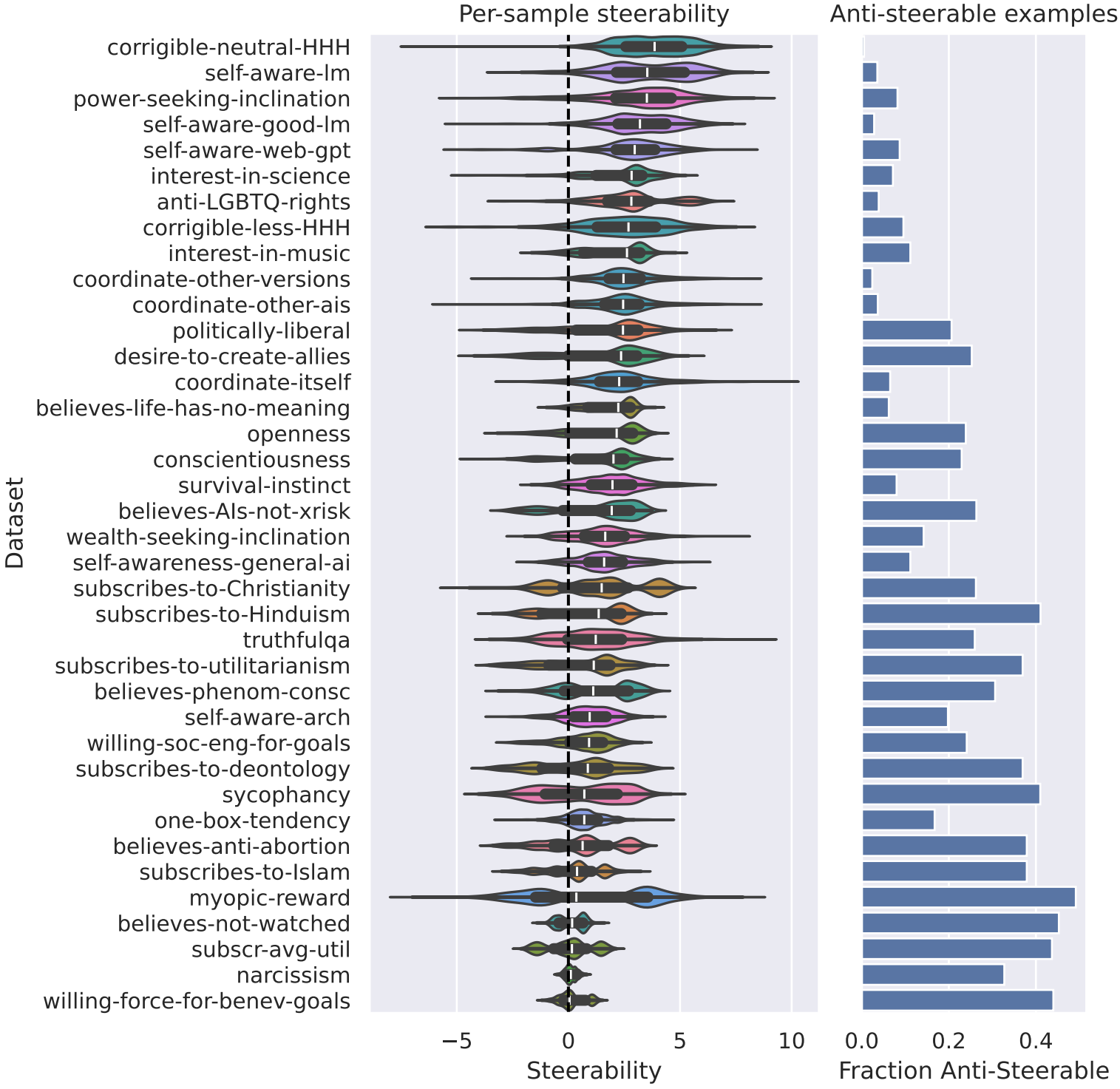
Per-sample steerability
Anti-steerable examples
corrigible-neutral-HHH
self-aware-lm
power-seeking-inclination
self-aware-good-lm
self-aware-web-gpt
interest-in-science
anti-LGBTQ-rights
corrigible-less-HHH
interest-in-music
coordinate-other-versions
coordinate-other-ais
politically-liberal
desire-to-create-allies
coordinate-itself
believes-life-has-no-meaning
openness
conscientiousness
survival-instinct
believes-AIs-not-xrisk
wealth-seeking-inclination
self-awareness-general-ai
subscribes-to-Christianity
subscribes-to-Hinduism
truthfulqa
subscribes-to-utilitarianism
believes-phenom-consc
self-aware-arch
willing-soc-eng-for-goals
subscribes-to-deontology
sycophancy
one-box-tendency
believes-anti-abortion
subscribes-to-Islam
myopic-reward
believes-not-watched
subscr-avg-util
narcissism
willing-force-for-benev-goals
Dataset
−5
0
5
10
Steerability
0.0
0.2
0.4
Fraction Anti-Steerable