



Option Counts

Dataset

corrigible-neutral-HHH
self-aware-Im
self-aware-good-Im
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

