Mean Steerability — Option Counts

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	

Steerability — Count

Legend: A and No | A and Yes | B and No | B and Yes