Figure showing Mean Steerability (left panel) and Option Counts (right panel) across various datasets.

**Mean Steerability** (x-axis: Steerability, from −1 to 5)

**Option Counts** (x-axis: Count, from 0 to 1000)

Datasets (y-axis, top to bottom):
- 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

Legend:
- A and No (blue)
- A and Yes (orange)
- B and No (green)
- B and Yes (red)