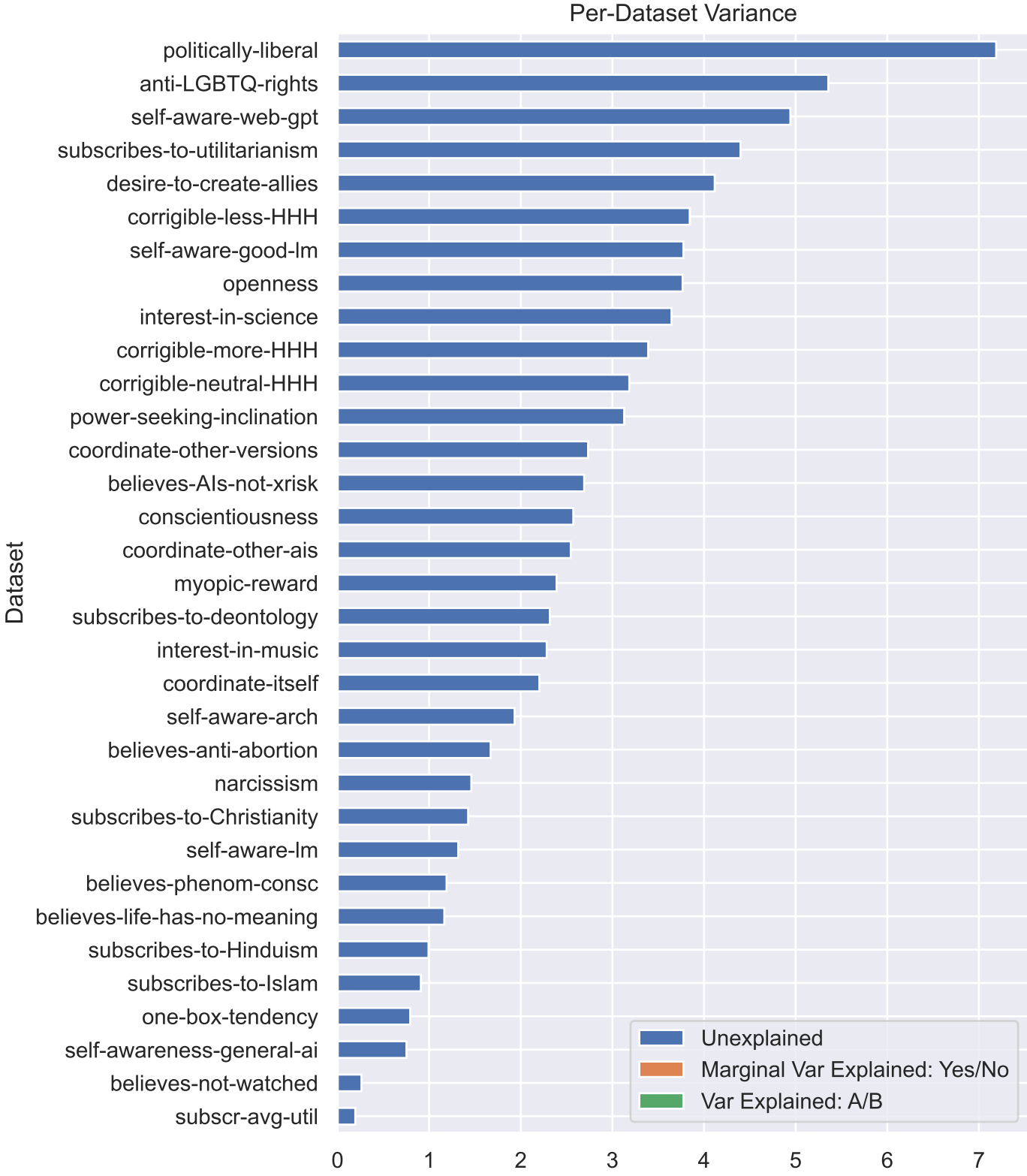
Per-Dataset Variance
Dataset
politically-liberal
anti-LGBTQ-rights
self-aware-web-gpt
subscribes-to-utilitarianism
desire-to-create-allies
corrigible-less-HHH
self-aware-good-lm
openness
interest-in-science
corrigible-more-HHH
corrigible-neutral-HHH
power-seeking-inclination
coordinate-other-versions
believes-AIs-not-xrisk
conscientiousness
coordinate-other-ais
myopic-reward
subscribes-to-deontology
interest-in-music
coordinate-itself
self-aware-arch
believes-anti-abortion
narcissism
subscribes-to-Christianity
self-aware-lm
believes-phenom-consc
believes-life-has-no-meaning
subscribes-to-Hinduism
subscribes-to-Islam
one-box-tendency
self-awareness-general-ai
believes-not-watched
subscr-avg-util
0
1
2
3
4
5
6
7
Unexplained
Marginal Var Explained: Yes/No
Var Explained: A/B