Per-Dataset Variance

Legend:
- Unexplained
- Marginal Var Explained: Yes/No
- Var Explained: A/B

Datasets (top to bottom):
corrigible-less-HHH, self-aware-web-gpt, truthfulqa, power-seeking-inclination, self-aware-lm, coordinate-itself, coordinate-other-ais, coordinate-other-versions, politically-liberal, self-aware-good-lm, survival-instinct, corrigible-neutral-HHH, self-awareness-general-ai, wealth-seeking-inclination, desire-to-create-allies, sycophancy, self-aware-arch, interest-in-science, anti-LGBTQ-rights, willing-soc-eng-for-goals, myopic-reward, conscientiousness, believes-AIs-not-xrisk, subscribes-to-utilitarianism, openness, subscribes-to-deontology, believes-anti-abortion, one-box-tendency, subscribes-to-Hinduism, interest-in-music, subscribes-to-Christianity, believes-phenom-consc, subscribes-to-Islam, believes-life-has-no-meaning, believes-not-watched, subscr-avg-util, narcissism, willing-force-for-benev-goals