Rater A

Demographics

Age: 40
Gender: Female
...
Income: £70k D_a

Fit dataset $\mathcal{D}^{\text{fit}}$ of n in-context ratings (E_n)

Is high immigration good for the UK? x_1

Strong yes y_{1a}

...

Should we ban cars in city centres? x_j

Yes y_{ja}

Is high immigration good for the UK? x_1

Neutral y_{1b}

...

Should we ban cars in city centres? x_k

No y_{kb}

Rater B

Demographics

Age: 40
Gender: Female
...
Income: £70k D_b

Value Encoder Q_ϕ

v_a

...

v_b

Generated Value Profile for Rater A v_a

social justice, equality; pro-immigration stance; pro-european...

Generated Value Profile for Rater B v_b

pro-car; fiscal conservatism; preference for existing systems ...

$\mathcal{D}^{\text{eval}}$ Held-out eval questions

Should we be in favour of universal basic income? x_i

v_a

Should we be in favour of universal basic income? x_i

v_b

Steerable Decoder P_θ

Predicted rating distribution given value profiles $P_\theta(y_{ia}|x_i, v_a)$

Strong No

Neutral

Strong Yes

0 10 20 30 40

$P_\theta(y_{ib}|x_i, v_b)$

Strong No

Neutral

Strong Yes

0 10 20 30 40

Encoder Q compresses rater representation as values

Decoder P uses value representation to predict ratings conditioned on values