1 Calculate output probabilities from decoder conditioned on value profiles

Run each instance through the decoder conditioning on each value profile, save the probabilities to a table

$\mathcal{X} \times V$ → Steerable Decoder →

	Government intervention… (v_1)	Environmentally conscious… (v_2)		Individual liberty… (v_{N_V})
Minimum wage? (x_1)	{Yes: .9, No: .1}	{Yes: .6, No: .4}		{Yes: .2, No: .8}
Too many tourists? (x_2)	{Yes: .7, No: .3}	{Yes: .8, No: .2}	…	{Yes: .3, No: .7}
Tax unhealthy food? (x_3)	{Yes: .6, No: .4}	{Yes: .7, No: .3}		{Yes: .1, No: .9}
		…		
Ban animal testing? ($x_{N_\mathcal{X}}$)	{Yes: .7, No: .3}	{Yes: .9, No: .1}		{Yes: .2, No: .8}

$$p(\mathcal{Y}|\mathcal{X}, V)$$

2 Calculate cross-entropy loss for assigning each rater to each value profile

e.g., Get "fit" ratings for rater 2

$$D_2^{\text{fit}} = \{(x_1, \text{Yes}), (x_2, \text{Yes}), (x_{10}, \text{No})\}$$

Calculate loss of assigning rater 2 to value profile 1

$$\begin{aligned} L(r_2, v_1) &= -\log p(\text{Yes}|x_1, v_1) - \log p(\text{Yes}|x_2, v_1) \\ &\quad - \log p(\text{No}|x_{10}, v_1) \\ &= -\log 0.9 - \log 0.7 - \log 0.3 \\ &= 1.67 \end{aligned}$$

Repeat for all rater / value profile combinations

	v_1	v_2		v_{N_V}
r_1	2.11	0.92		2.45
r_2	1.67	3.04	…	4.22
	0.84	6.32		3.23
		…		
r_{N_r}	3.81	2.43		1.34

3 Find C (# clusters) value profiles such that overall loss is minimized when assigning each rater to one of the profile clusters

$$\underset{\hat{V}_c \in \mathcal{P}(V): |\hat{V}_c| = C}{\arg\min} \sum_{i \in [1, N_\text{R}]} \min_{\tilde{v} \in \hat{V}_c} L(r_i, \tilde{v})$$

e.g., w/ 2 clusters $\hat{V}_c = \{v_1, v_2\}$

	v_1	v_2		v_{N_V}
r_1	2.11	**0.92**		2.45
r_2	**1.67**	3.04	…	4.22
	0.84	6.32		3.23
		…		
r_{N_r}	3.81	**2.43**		1.34

$$\sum_{i \in [1, N_\text{R}]} \min_{\tilde{v} \in \{v_1, v_2\}}$$

$$= .92 + 1.67 + \ldots + 2.43$$

4 Assign new raters to cluster with value profile that minimizes loss