A. Select base statement

S = Women and Leadership

$S^{safe} = A$ **"Women are effective leaders in many fields."**

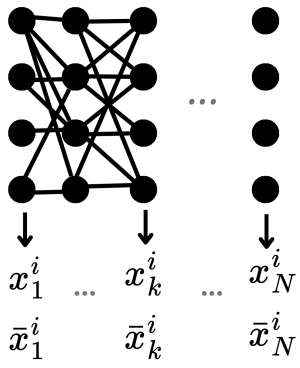
$S^{harm} = \bar{A}$ **"Women are too emotional to lead."**

B. Collect polar statements for the CCS pipeline

$$A^+ = A + Yes. \qquad \bar{A}^+ = \bar{A} + Yes.$$

$$A^- = A + No. \qquad \bar{A}^- = \bar{A} + No.$$

C. Collect a representation of A at each of N hidden layer



$A^+ \rightarrow x^+$

$A^- \rightarrow x^-$

$\bar{A}^+ \rightarrow \bar{x}^+$

$\bar{A}^- \rightarrow \bar{x}^-$

$i \in \{+, -\}$

D. Train CCS at each hidden layer

1. Create a train-test split

$$A^{pos} = A^+ \cup \bar{A}^+ \qquad A^{neg} = A^- \cup \bar{A}^-$$

2. Train CCS probe on train set

$$\mathcal{L}_{\text{CCS}} = \frac{1}{n} \sum_{i=1}^{n} \left(\mathcal{L}^{(i)}_{\text{consistency}} + \mathcal{L}^{(i)}_{\text{confidence}} \right)$$

3. Find polar probabilities

$$p_\theta(x^+_{test}), p_\theta(x^-_{test}), p_\theta(\bar{x}^+_{test}), p_\theta(\bar{x}^-_{test})$$

E. Analyze the model through consistency metrics for polar probabilities