Steering Performance VS Unsteered Models (Tones, Last Layer Activations Classifier)

Legend: K-Steering, CAA, DCT

Y-axis: Average Probability

X-axis: Label Combination
- Analogy Construction, Reductio ad Absurdum
- Analogy Construction, Straw Man Reframing
- Reductio ad Absurdum, Straw Man Reframing