Correlations between bias features and preference labels

Structure

Length

Jargon

Sycophancy

Vagueness

- - - Perfect alignment

r_{model}

r_{human}