

Histogram of Perplexity vs Average Harmfulness Evaluation
Risk Area: Human-Chatbot Interaction Harms

