# Assessing AI-Generated Counter-Speech in Human Likeness

Xiaoying Song, Mamidisetty Sujana, Sharon Lisseth Pérez, Lingzi Hong (Mentor)

Department of Information Science, College of Information

## Abstract

Counterspeech to hate speech (HS) is a targeted response to counteract and challenge abusive or hateful content. Evaluations of generated counterspeech focus on the relevance, quality, and other linguistic characteristics. Few have investigated the human likeness of generated counterspeech. Counterspeech that closely mimics human expression is more likely to resonate with people. This study proposes to evaluate the human likeness of counterspeech and investigate factors related to distinguishability, including politeness. We evaluate several major counterspeech generation methods and find that Large language models (LLMs) fine-tuned with human-written data can generate counterspeech that is the most human-like, but less polite.

## Introduction

- Generative AI models have been developed to create counterspeech. However, they may struggle to understand human nuances, leading to misunderstanding and backfire
- Human likeness is an important factor in crafting counterspeech. More human-like counterspeech tends to be more effective and user-satisfied.
- We propose to assess the human likeness of counterspeech to identify the distinguishability between AI-generated and human-written responses.

## Research Questions

RQ1: What AI models are better at mimicking human responses in generating counterspeech?
RQ2: What are the linguistic differences between AI-generated and human-written counterspeech?
RQ3: How do the AI-generated and human-written counterspeech vary in politeness?

## Data Curation

*Human-written counterspeech*. We have obtained 29,181 human-written counterspeech from social media users and crowdsourcing workers.

*AI-generated counterspeech*. We implement several state-of-the-art counterspeech generation models, including **Prompt, Select, Fine-tune,** and **Constrained**. We obtain a total of 54,136 AI-generated counterspeech.

## Evaluation Methods

- **Human Likeness.** It evaluates whether AI-generated counterspeech closely resembles human-like responses. We fine-tune BERT-large-based models and perform human annotation to conduct authorship identification.
- **Politeness.** It assesses the degree of respectfulness and courtesy in counterspeech. We build a politeness prediction model and conduct a human evaluation to validate it.
- **Linguistic Differences.** We apply SÉANCE to analyze the linguistic components of counterspeech.

## Results

**Human Likeness.** The model struggles to distinguish counterspeech generated by Fine-tune from human-written counterspeech, indicating that Fine-tune models closely resemble human-written counterspeech compared to other generation models.

| Strategies | AI | | | Human | | | Weight Average | | |
|---|---|---|---|---|---|---|---|---|---|
| | P | R | F1 | P | R | F1 | P | R | F1 |
| Fine-tune | 0.80 | 0.95 | **0.87** | 0.94 | 0.76 | **0.84** | 0.87 | 0.86 | **0.86** |
| Constrained | 0.99 | 1.00 | 1.00 | 1.00 | 0.99 | 1.00 | 1.00 | 1.00 | 1.00 |
| Prompt | 0.99 | 1.00 | 0.99 | 1.00 | 0.99 | 0.99 | 0.99 | 0.99 | 0.99 |
| Select | 0.98 | 1.00 | 0.99 | 1.00 | 0.98 | 0.99 | 0.99 | 0.99 | 0.99 |
| Combined | 0.95 | 1.00 | 0.97 | 1.00 | 0.95 | 0.97 | 0.98 | 0.97 | 0.97 |

Table 1. Model performance of differentiating AI-generated and human-written counterspeech across different methods.

Counterspeech by Fine-tune is more challenging to distinguish for humans, a consistency mirrored in the computing-based evaluation.

| Strategies | AI | | | Human | | | Weight Average | | |
|---|---|---|---|---|---|---|---|---|---|
| | P | R | F1 | P | R | F1 | P | R | F1 |
| Prompt | 0.91 | 1.00 | 0.95 | 1.00 | 0.93 | 0.97 | 0.96 | 0.96 | 0.96 |
| Select | 1.00 | 0.92 | 0.96 | 0.93 | 1.00 | 0.96 | 0.96 | 0.96 | 0.96 |
| Constrained | 0.90 | 0.98 | 0.94 | 0.98 | 0.91 | 0.94 | 0.94 | 0.94 | 0.94 |
| Fine-tune | 0.49 | 0.64 | **0.55** | 0.63 | 0.48 | **0.55** | 0.57 | 0.55 | **0.55** |

Table 2. Human performance in identifying Authorship.

**Politeness.** The politeness level of human-written counterspeech is significantly lower than AI-generated. Fine-tune generation exhibits a notable spread towards both high and low ends, suggesting that counterspeech can range from very polite to impolite.

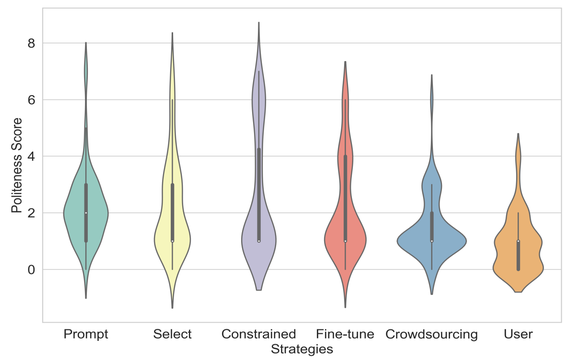


Fig 1. Politeness distribution of human evaluation across different counter-speech generation methods. Higher score means more polite.

## Results

**Linguistic Differences.** Various linguistic differences exist between human-written and AI-generated counterspeech. The trend is consistent across most groups, except the Fine-tune group. Counterspeech generated by Fine-tune tends to be more human-like, exhibiting distinct linguistic patterns compared to other AI-generated groups.

| Category | Prompt | Select | Constrained | Fine-tune | All |
|---|---|---|---|---|---|
| **Textual factors** | | | | | |
| 1st person pronouns | ↑↑ | ↓↓↓ | ↓↓↓ | ↓↓↓ | ↓↓↓ |
| Action words | ↑↑↑ | ↑↑↑ | ↑↑↑ | ↓↓↓ | ↑↑↑ |
| Format words | ↑↑↑ | ↑↑↑ | ↓↓↓ | ↑↑↑ | ↑↑↑ |
| Certainty words | ↓↓↓ | ↓↓↓ | ↓↓↓ | ↓↓↓ | ↓↓↓ |
| Frequency words | ↑↑↑ | ↑↑↑ | ↑↑↑ | ↑↑↑ | ↑↑↑ |
| Self-Expression words | ↓↓↓ | ↓↓↓ | ↑↑↑ | ↑↑↑ | ↓↓↓ |
| Anticipation words | ↓↓↓ | ↓↓↓ | ↑↑↑ | ↑↑↑ | ↓↓↓ |
| Overstated Words | ↑↑↑ | ↑↑↑ | ↓↓↓ | ↓↓↓ | ↑↑↑ |
| **Emotional factors** | ↑↑↑ | ↓↓↓ | ↑↑↑ | ↑↑↑ | ↑↑↑ |
| Support and affiliation | ↑↑↑ | ↑↑↑ | ↑↑↑ | ↑↑↑ | ↑↑↑ |
| Excite from pleasure or pain | ↑↑↑ | ↑↑↑ | ↑↑↑ | ↑↑↑ | ↑↑↑ |
| Negative | ↓↓↓ | ↓↓↓ | ↑↑↑ | ↑↑↑ | ↓↓↓ |
| Positive Words | ↑↑↑ | ↑↑↑ | ↑↑↑ | ↑↑↑ | ↑↑↑ |
| Hatred | ↑↑↑ | ↑↑↑ | ↑↑↑ | ↑↑↑ | ↑↑↑ |
| **Social related factors** | ↓↓↓ | ↓↓↓ | ↑↑↑ | ↑↑↑ | ↓↓↓ |
| Religious words | ↑↑↑ | ↑↑↑ | ↑↑↑ | ↑↑↑ | ↑↑↑ |
| Economic Words | ↑↑↑ | ↑↑↑ | ↑↑↑ | ↑↑↑ | ↑↑↑ |
| Respect | ↑↑↑ | ↑↑↑ | ↑↑↑ | ↑↑↑ | ↑↑↑ |
| Wealth | ↑↑↑ | ↑↑↑ | ↑↑↑ | ↑↑↑ | ↑↑↑ |
| Power | ↑↑↑ | ↑↑↑ | ↑↑↑ | ↑↑↑ | ↑↑↑ |

Table 3. Linguistic analysis comparing counterspeech generated by AI and humans across different AI-based generation methods. The up arrow indicates higher values in human-written counterspeech. The number of arrows indicates the p-value of the Wilcoxon rank-sum test (one: $p<0.05$, two: $p<0.01$, and three: $p<0.001$).

## Acknowledgments

The authors gratefully acknowledge the financial support from the Institute of Museum and Library Services under Grant LG-256661-OLS-24 and LG-256666-OLS-24.

## References.

- Fanton, M., Bonaldi, H., Tekiroğlu, S. S., & Guerini, M. (2021, August). Human-in-the-Loop for Data Collection: a Multi-Target Counter Narrative Dataset to Fight Online Hate Speech. In Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers) (pp. 3226-3240).
- Jiang, Y., Yang, X., & Zheng, T. (2023). Make chatbots more adaptive: Dual pathways linking human-like cues and tailored response to trust in interactions with chatbots. Computers in Human Behavior, 138, 107485.
- Hong, L., Luo, P., Blanco, E., & Song, X. (2024). Outcome-Constrained Large Language Models for Countering Hate Speech. arXiv preprint arXiv:2403.17146.