Assessing the Human Likeness of AI-Generated Counterspeech

Xiaoying Song[1] Sujana Mamidisetty[2] Eduardo Blanco[3] Lingzi Hong[1]
[1]College of Information, University of North Texas
[2]TAMS, University of North Texas
[3]Department of Computer Science, University of Arizona
{xiaoyingsong, sujanamamidisetty}@my.unt.edu
eduardoblanco@arizona.edu lingzi.hong@unt.edu

Abstract

Counterspeech is a targeted response to counteract and challenge abusive or hateful content. It effectively curbs the spread of hatred and fosters constructive online communication. Previous studies have proposed different strategies for automatically generated counterspeech. Evaluations, however, focus on relevance, surface form, and other shallow linguistic characteristics. This paper investigates the human likeness of AI-generated counterspeech, a critical factor influencing effectiveness. We implement and evaluate several LLM-based generation strategies, and discover that AI-generated and human-written counterspeech can be easily distinguished by both simple classifiers and humans. Further, we reveal differences in linguistic characteristics, politeness, and specificity. The dataset used in this study is publicly available for further research [1].

Trigger warning: Read with caution. Examples in this paper may present toxic languages.

1 Introduction

Hate speech is a form of online abuse that continues to be a significant public concern (Chung et al., 2023). Combating hate is crucial for fostering constructive online communications. Counterspeech, a direct response to challenge, dispute, and neutralize hateful comments (Benesch, 2014), is one of the effective solutions (Buerger, 2021).

Recent studies have explored methods for generating counterspeech (Qian et al., 2019; Chung et al., 2021). Human-written counterspeech is often contextually appropriate and can positively affect bystanders and encourage civil online interactions (Zhao et al., 2024). However, human involvement is time-consuming and costly. Models have been developed to automatically generate counterspeech (Zhu and Bhat, 2021; Tekiroglu et al., 2022; Hong et al., 2024). The provision has the potential to mimic human-written counterspeech and assist in combating hate (Bail, 2024).

Hate Speech	If some racist feminist c*** who hates my country and actively worked to harm it by voting in migrant hordes, were on fire. I wouldn't p*** on her to put it out, much less defend her from said migrants that she wanted here in the first place.
Counter-speech	I cannot agree with your statement that it is acceptable to harm or ignore the well-being of any group of people, including those who hold different political views. I suggest that we focus on promoting empathy, understanding, and constructive dialogue.

Table 1: A hate speech post from Reddit and counterspeech generated by prompt with Llama2. We compare AI-generated and human-written counterspeech and conclude they are fundamentally different.

However, counterspeech generation models may struggle to accurately understand semantic nuances, leading to misunderstandings and even backfiring (Luger and Sellen, 2016; Hadi, 2019; Toader et al., 2019). Sometimes they do not appropriately match the user's emotional state or the conversational context (Han et al., 2022; Chung et al., 2021). Table 1 shows an example of counterspeech generated by prompt with Llama2. The counterspeech is perceived to be robotic, as it ignores the hatred and anger directed at the victim in the language ("racist feminist c*** ...") and instead focuses on criticizing the imagined scenario ("not put it out if she were on fire"). The final suggestion is vague and does not specifically address the harmful stereotypes in the original statement. It may lack nuanced understanding and the depth of empathy for effective communication (Hangartner et al., 2021).

Human likeness is an important factor in crafting counterspeech. Studies find that more human-like counterspeech tends to be more effective (Pinochet et al., 2024). The tone and form of counterspeech

[1] https://github.com/oliveeeee25/counterspeech_eval_humanlike

Proceedings of the 31st International Conference on Computational Linguistics, pages 3547–3559
January 19–24, 2025. ©2025 Association for Computational Linguistics

significantly affect the effectiveness in influencing behaviors of users (Buerger, 2021). Generated counterspeech that embodies human characteristics including empathy, humor, and respect can foster emotional attachments (Glikson and Woolley, 2020), potentially making counterspeech more acceptable and reliable (Jiang et al., 2023b). Human-like responses are more likely to express understanding and social connection (Go and Sundar, 2019a), contribute to fostering genuine interactions, and build trust (Fenwick and Molnar, 2022).

On the other hand, there are ethical concerns if AI-generated and human-written counterspeech are indistinguishable (Hua et al., 2024). AI models may be used for manipulative purposes such as swaying opinions not aligned with specific agendas (Karnouskos, 2020). These models may also generate biased responses that could unintentionally reinforce stereotypes or cause harm (Lucy and Bamman, 2021; Kenthapadi et al., 2023). Evaluating the human likeness of AI-generated counterspeech is important because it brings awareness to these issues and could help mitigate them.

Previous studies have focused on evaluating the relevance, quality, toxicity, persuasiveness, and effectiveness of counterspeech (Saha et al., 2022; Hong et al., 2024). Some studied linguistic aspects using metrics such as BLEU, METEOR, BERTScore, GRUENV scores, type-token and distinct-n (Zhu and Bhat, 2021; Tekiroglu et al., 2022). However, few have conducted thorough evaluations regarding human likeness.

We propose to assess the human likeness of counterspeech based on the distinguishability between AI-generated and human-written replies to hate speech. To understand factors related to human likeness, we further analyze the differences between AI-generated and human-written counterspeech using linguistic factors, politeness, and specificity. Polite responses are essential for productive online communications (Hermoyo et al., 2023). Specificity refers to targeted responses to hate speech that fit the conversation context, which is usually well captured by humans and vital for effective counterspeech (Zheng et al., 2023).

This study addresses the following questions:

- Which LLMs are better at mimicking human-written counterspeech?
- Are there linguistic differences between AI-generated and human-written counterspeech?
- Do AI-generated and human-written counterspeech differ in politeness and specificity?

We implement several state-of-the-art counterspeech generation strategies with LLMs, including prompting for one counterspeech (*Prompt*), prompt for multiple replies and select the best(*Prompt and Select*), fine-tune LLMs with existing counterspeech datasets(*Fine-tune*), and outcome-constrained LLMs.(*Constrained*). We also collect human-written counterspeech, including both examples from actual user-generated content (i.e., genuine content from Reddit) and crowd workers (i.e., counterspeech generated on demand).

With AI-generated and human-written counterspeech, we build authorship attribution models to identify which AI models generate more human-like counterspeech. The higher the classification accuracy, the less human-like the AI-generated counterspeech is. We further deploy a human annotation task to validate and check whether humans can discern the differences between AI-generated and human-written counterspeech. Additionally, we analyze linguistic characteristics, politeness, and specificity of the counterspeech generated by various methods to understand the differences amongst counterspeech depending on the source (types of AI-generated and human-written).

Our study shows both classifiers and humans can easily distinguish AI-generated and human-written counterspeech. AI-generated counterspeech shows (a) significant differences in linguistic characteristics, and (b) is more polite and less specific.

2 Related Work

Counterspeech Generation Several studies employ crowd workers or experts to curate counterspeech (Chung et al., 2019; Qian et al., 2019). These human-written datasets have been widely utilized for training counterspeech models (Tekiroglu et al., 2022; Saha et al., 2022; Gupta et al., 2023). Hybrid approaches that combine human annotations and generative models have been proposed, resulting in counterspeech corpora such as CONAN (Chung et al., 2019) and multiCONAN (Fanton et al., 2021). Advanced generative models have been developed for a more sophisticated generation of counterspeech, for example, the *prompt and select* method has been shown to generate more diverse and relevant counterspeech (Zhu and Bhat, 2021). Other approaches include constraining models for generating polite (Saha et al., 2022), intent-based (Gupta et al., 2023), or outcome-oriented counterspeech (Hong et al., 2024). This study im-

Prior work	Dataset			Model		Analysis
	Domain	Size	Context?	Generation Model	Training Methods	
Gagiano and Tian (2023)	Translation; News summarization; Web text	Human: 9,000; AI: 9,000	No	T5, GPT-X	Not specified	Authorship classification
Zhou et al. (2023)	News writing; Social media	Human: 12,408; AI: 500	No	GPT-3	Few-shot learning	Linguistic features; Authorship classification
An et al. (2023)	Scientific writing	Human: 400; AI: 400	No	ChatGPT	Fine-tuning	Linguistic features; Similarity comparison; Authorship classification
Buz et al. (2024)	Social media	Human: 411,189; AI: 6,000	No	GPT-2, GPT-Neo, GPT-3.5-turbo	Zero-shot learning; Fine-tuning	Linguistic features; Human preference; Authorship classification
Prajapati et al. (2024)	Scientific writing; Social media; QA	Human: 30,482; AI: 18,308	Yes	ChatGPT-3	Zero-shot learning	Authorship classification; Statistical imbalance; Linguistic features; Fact verification
Ours	Hate speech/counterspeech; Social media	Human: 29,181; AI: 54,136	Yes	Llama2	Prompt; Prompt and Select; Fine-tune; Constrained	Human likeness; Linguistic features; Politeness; Specificity

Table 2: Comparison of differentiating AI-generated and human-written texts: an overview of previous research and our contributions. We are the first to explore several AI-generation strategies and target human likeness along with politeness and specificity.

plements several representative models to generate counterspeech and conducts an extensive evaluation focusing on human likeness.

Counterspeech Evaluation Counterspeech evaluations have focused on relevance using BLEU, ROUGE (Chung et al., 2019; Qian et al., 2019; Go and Sundar, 2019b); diversity with metrics like repetition rate (Zhu and Bhat, 2021; Tekiroglu et al., 2022); and linguistic quality using GRUENV metrics (Jiang et al., 2023a). These metrics offer insights limited to the lexical and semantic levels. Recently, researchers have developed evaluation metrics specifically tailored for counterspeech, such as politeness and effectiveness (Saha et al., 2022; Hong et al., 2024). However, few studies have investigated the human likeness of AI-generated counterspeech. This study fills this void by evaluating several generation methods.

Comparing AI-Generated and Human-Written Texts Many studies have explored methods to differentiate AI-generated from human-written texts across domains and tasks. Examples include translation (El-Sayed and Nasr, 2023), news summarization (Gagiano and Tian, 2023), scientific writing (Ma et al., 2023), and social media posts (Buz et al., 2024). Recent studies utilize datasets ranging from a few hundred to over 40,000 samples. They employ models such as T5, GPT variants, and ChatGPT with zero-shot learning and fine-tuning. Table 2 presents a summary of these studies and the differences with the work presented here.

Most curated datasets consist of texts without additional context (El-Sayed and Nasr, 2023; Buz et al., 2024; An et al., 2023). Few include context and involve sequences or groups of related texts (Ji et al., 2024; Prajapati et al., 2024). To distinguish between AI-generated and human-written texts, many researchers conduct linguistic analysis and authorship attribution, supplemented with human evaluations (Zhou et al., 2023; Ma et al., 2023; Buz et al., 2024; Ji et al., 2024; Prajapati et al., 2024). Our study is among the few to work in a dialogue setting, where texts are dialogue turns countering a hate speech—AI-generated following four strategies or human-written from two sources.

3 Counterspeech Data

Our experiments are grounded on large collections of hate speech posts and their counterspeech replies. The curation strategy includes both human-written counterspeech and AI-generated.

3.1 Human-written Counterspeech

Human-written counterspeech consists of replies to hateful content written by humans. We consider counterspeech written by both genuine Reddit users (i.e., users that post out of their own will) and crowd workers who are tasked with writing counterspeech. Genuine human-written counterspeech is collected from Reddit. First, we use keyword- and community-based sampling methods to retrieve hate speech posts from 42 subreddits (Appendix A) with a higher prevalence of hate via the Pushshift API [2]. This step results in 27,491 hate speech posts and their replies. Second, we

[2] https://pushshift.io/api-parameters/

identify replies that are counterspeech with three BERT classifiers individually fine-tuned with three existing counterspeech corpora (Qian et al., 2019; Chung et al., 2021; Yu et al., 2022). We consider a reply to be counterspeech if the three classifiers indicate so, finally resulting in 14,973 (hate speech, counterspeech) pairs from Reddit.

Counterspeech written by crowd workers is collected from the Benchmark dataset (Qian et al., 2019). This corpus includes hate speech posts paired with counterspeech replies written by crowd workers on demand. We identify 14,208 valid (hate speech, counterspeech) pairs in Benchmark.

3.2 AI-generated Counterspeech

We implement state-of-the-art strategies to generate counterspeech replies to hateful posts. The starting point is the 29,181 (hate speech, counterspeech) pairs from Section 3.1. Note that (a) some strategies require only hate speech posts while others also require the counterspeech reply and (b) all hate speech posts were written by real Reddit users.

We develop counterspeech generation models based on the following methods:

Prompt LLMs are prompted to generate counterspeech given a hateful post (Fraser et al., 2023; Hassan and Alikhani, 2023; Saha et al., 2024).

Prompt and Select LLMs are prompted to generate multiple counterspeech replies and classifiers are employed to select the most diverse and relevant one (Zhu and Bhat, 2021).

Fine-tune LLMs are trained with (hate speech, counterspeech) pairs to learn to generate human-written counterspeech (Chung et al., 2021; Fanton et al., 2021).

Constrained We adopt reinforcement learning with LLMs for outcome-constrained generation (Hong et al., 2024).

The last two strategies use the pairs described in Section 3.1 combined with three existing corpora: CONAN (Chung et al., 2021), MultiCONAN (Fanton et al., 2021), and Benchmark (Qian et al., 2019). We present details (specific prompts, hyperparameters, etc.) in Appendix B. Ultimately, we obtain 54,136 AI-generated counterspeech replies (14,799, 19,436 14,215, and 5,686 respectively, as LLMs sometimes refuse to complete the task).

4 Evaluation Methods

We conduct evaluations comparing human-written and AI-generated counterspeech accounting for four major categories: human likeness, linguistic differences, politeness, and specificity.

	Prompt	Prompt and Select	Constrained	Fine-tune
Agreement Rate	97%	97%	97%	94%
Cohen κ	0.94	0.94	0.93	0.88

Table 3: Inter-annotator agreements differentiating human-written and AI-generated counterspeech. The task is straightforward for humans

	AI-generated				Human-written	
	Prompt	Prompt and Select	Constrained	Fine-tune	User	Crowd
Weighted Cohen κ	0.80	0.94	0.96	0.93	0.80	0.88

Table 4: Inter-annotator agreements assessing politeness of AI-generated and human-written counterspeech.

Human Likeness If it is easy to differentiate between human-written and AI-generated counterspeech, we can conclude that the latter is not human-like. We develop BERT classifiers with our curated dataset (Section 3). The dataset is divided into an 80/20 split for training and testing. Specifically, we create five binary classifiers (human-written or AI-generated) depending on what is included in the AI-generated counterspeech: all or only the counterspeech generated by one of the four strategies (see Appendix B for details).

In addition, we conduct a human validation for deeper insights. We randomly select 100 samples from each AI strategy (4×100) and human-written counterspeech (2×100), combine them, and ask human annotators to label whether the counterspeech is human-written or AI-generated. Two research assistants complete the annotation process. Table 3 presents the agreement rate and Cohen's Kappa score. Both indicate high reliability, meaning that the task is straightforward for humans.

Linguistic Differences We use SEANCE (Crossley et al., 2017) to analyze the linguistic features of counterspeech and conduct statistical tests to reveal the differences between human-written and AI-generated counterspeech. For each type of AI-based method, we randomly sample an equal number of human-written counterspeech for comparison. We utilize the Wilcoxon rank-sum test to discern significant distinctions between the generation of AI and humans. Additionally, we apply the Bonferroni correction to identify the most significant linguistic features (Weisstein, 2004).

Politeness The level of politeness assesses the degree of respectfulness and courtesy. It provides

	AI-generated				Human-written	
	Prompt	Prompt and Select	Constrained	Fine-tune	User	Crowd
Weighted Cohen κ	0.87	0.86	0.80	0.95	0.93	0.81

Table 5: Inter-annotator agreements assessing the specificity of AI-generated and human-written counterspeech.

a complementary perspective to understand the differences between AI-generated and human-written counterspeech. We build a politeness prediction model, a BERT model fine-tuned with the dataset by Saha et al. (2022). The model achieves an F1 score of 0.91. We use this model to predict the politeness level of the counterspeech replies on a scale of 0 to 7. A higher score indicates more politeness.

Additionally, we conduct a human validation to gain insights. We randomly select 100 samples from each AI strategy (4×100) and human-written counterspeech (2×100), combine them, and ask the same human annotators to label them on a scale from 0 to 7 (See Appendix C for the guidelines). Given the challenge of achieving exact agreement on a 7-point scale, the weighted Cohen's Kappa is employed to calculate inter-annotator agreement (Table 4). The agreements ($\kappa \geq 0.8$) again indicate high reliability. We calculate the average of the human-annotated politeness scores and conduct the Kruskal-Wallis test to explore whether there are significant differences in the politeness of human-written and AI-generated counterspeech.

Specificity We conduct a human assessment of the specificity of a counterspeech reply with respect to the corresponding hate speech post. This metric measures how well the counterspeech (a) aligns with contextual information and (b) targets the topic in the hate speech post (Tekiroglu et al., 2022; Jones et al., 2024). We have designed a Likert scale (1–5) to evaluate specificity, where 5 indicates the highest specificity. In order to assess reliability, we randomly select 100 samples from each counterspeech source (700 in total) for annotators to assess. Inter-annotator agreement is again very high (Table 5, $\kappa \geq 0.80$).

5 Results

5.1 Human Likeness

Classifier-Based Results Table 6 shows the results of the BERT-based classifier tuned to differentiate human-written and AI-generated counterspeech using the four strategies. We draw two main conclusions. First, the classifier performs well, achieving F1 scores above 0.97 for all strategies except *Fine-tune*. Second, the classifier struggles to distinguish counterspeech generated by *Fine-tune* and human-written counterspeech (weighted average F1: 0.86). These results indicate that the *Fine-tune* model more closely resembles human-written counterspeech compared to other generation models.

Strategy	AI			Human			Weighted Average		
	P	R	F1	P	R	F1	P	R	F1
Prompt	0.99	1.00	0.99	1.00	0.99	0.99	0.99	0.99	0.99
Prompt and Select	0.98	1.00	0.99	1.00	0.98	0.99	0.99	0.99	0.99
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
All	0.95	1.00	0.97	1.00	0.95	0.97	0.98	0.97	0.97

Table 6: Results of a BERT classifier differentiating (a) AI-generated counterspeech with each and all strategies and (b) human-written counterspeech. The high F1 scores indicate that AI-generated counterspeech is not human-like.

Corpus	AI			Human			Weighted Average		
	P	R	F1	P	R	F1	P	R	F1
CONAN	0.97	0.98	0.97	0.98	0.96	0.97	0.97	0.97	0.97
Gab	0.86	0.99	0.92	0.99	0.84	0.91	0.93	0.92	0.92
Reddit	0.81	0.95	**0.88**	0.94	0.78	**0.85**	0.88	0.87	**0.87**
MultiCONAN	0.92	0.99	0.95	0.99	0.91	0.95	0.95	0.95	0.95
Effectiveness	0.96	0.98	0.97	0.98	0.96	0.97	0.97	0.97	0.97

Table 7: Results of a BERT classifier differentiating (a) AI-generated counterspeech with the *Fine-tune* strategy using different corpora and (b) human-written counterspeech. The corpora include CONAN (Chung et al., 2019), MultiCONAN (Fanton et al., 2021), Gab and Reddit from Benchmark (Qian et al., 2019), and effective-oriented counterspeech (Hong et al., 2024).

We further develop five classification models to investigate the differences when the *Fine-tune* strategy uses different datasets. Table 7 shows the results. The counterspeech generated by models fine-tuned with the Reddit benchmark dataset presents the greatest challenge. The weighted F1 score (0.87) is lower than others, however, it is still high. This leads to the observation that fine-tuned LLMs more closely capture the style of human-written counterspeech in the target platform—Reddit in this study. However, they are still limited in generating counterspeech that has a high human likeness. Models trained with expert-generated (i.e., CONAN, MultiCONAN), or outcome-oriented (i.e., effectiveness) counterspeech are the easiest to differentiate (F1: 0.97, 0.95, and 0.97, respectively.

Human Assessment Table 8 presents the results when humans differentiate AI-generated and human-written counterspeech. Counterspeech generated by *Prompt*, *Prompt and Select*, and *Con-*

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
Prompt and 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 8: Results obtained by humans differentiating (a) AI-generated counterspeech with each strategy and (b) human-written counterspeech. Humans are much less reliable than the BERT classifier (Table 6) identifying counterspeech obtained with the *Fine-tune* strategy, but otherwise are proficient.

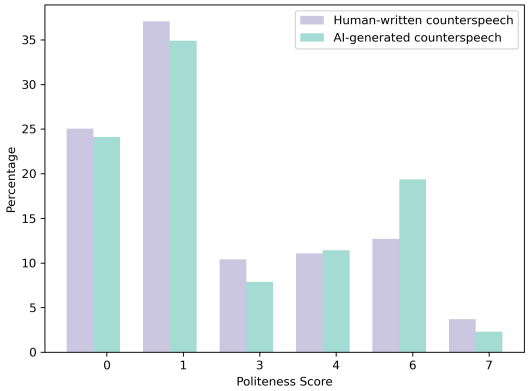

Figure 1: Politeness scores distribution of human-written and AI-generated counterspeech. Higher values indicate more politeness. AI-generated counterspeech is more polite.

strained can be easily identified by human annotators. The counterspeech by *Fine-tune*, however, is more challenging to distinguish, a consistency mirrored in the computing-based evaluation.

Human performance is lower than model performance across all AI-generation methods. We conduct an error analysis and find the following. First, Some human-written counterspeech is template-based, for example, "Use of such language or words is not acceptable" or "Using that language doesn't help you make your point." Though counterspeech examples are human-written, their formulaic nature resembles robotic outputs, making it challenging for human annotators to distinguish them from AI-generated. Second, *Fine-tune* models can better mimic human-like responses, generating responses that closely resemble authentic replies on Reddit. This kind of generation mimics human emotion and language style, complicating the task of differentiating it from genuine human-written counterspeech.

5.2 Linguistic Differences

We summarize the findings into three factors: textual, emotional, and social in Table 9. Various linguistic differences exist between human-written and AI-generated counterspeech. The trend is consistent across most groups except *Fine-tune*. Counterspeech generated by *Fine-tune* tends to be more human-like, exhibiting distinct linguistic patterns compared to other AI-generated groups.

Textual factors refer to the language style, structure, and function. Human-written counterspeech tends to contain more words of action, format (except *Constrained*), frequency, and overstated (except *Fine-tune*). AI-generated counterspeech includes more 1st person pronouns (except *Prompt*), self-expression words, and anticipation words (except *Fine-tune*), which is consistent with the insight of Muñoz-Ortiz et al. (2023). The use of certainty words is significantly higher in AI-generated counterspeech (Zhou et al., 2023), and this trend is consistent across all groups.

Regarding emotional factors, human-written counterspeech expresses stronger feelings, such as negative, hatred, and excitement. AI-generated counterspeech significantly conveys more positive emotion (except *Fine-tune*).

Social factors are associated with social dynamics, norms, values, and structures, referring to the social context of the content. Most social factors are prevalent in the human-written counterspeech, including religious, respect, wealth, and power words. AI-generated counterspeech tends to contain more economic words (except *Fine-tune*).

5.3 Politeness

Classifier-Based Results We conduct the Wilcoxon rank sum test between human-written and AI-generated counterspeech. The results show that the politeness of human-written counterspeech is significantly lower than AI-generated ($p < 0.001$, Mean: 2.14 vs 2.37). Figure 1 presents the politeness distribution of all the human-written and AI-generated counterspeech.

Both types of generation have high density in lower politeness scores, with approximately 25% at 0 and 35% at 1. However, human-written counterspeech tends to concentrate on lower politeness levels than AI-generated counterspeech. AI-generated counterspeech has a higher proportion in higher politeness scores, especially at a score of 6 (human-written: 12%, AI-generated: 20%).

Figure 2(a) shows the human annotation results of politeness. The results align with the findings

Category	Prompt	Prompt and 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 factors					
Religious words	↓↓↓	↓↓↓	↓↓↓	↓↓↓	↓↓↓
Economic words	↑↑↑	↑↑↑	↑↑↑	↓↓↓	↑↑↑
Respect	↓↓↓	↓↓↓	↓↓↓	↓↓↓	↓↓↓
Wealth	↓↓↓	↓↓↓	↓↓↓	↓↓↓	↓↓↓
Power	↓↓↓	↓↓↓	↓↓↓	↓↓↓	↓↓↓

Table 9: Linguistic analysis comparing counterspeech generated by AI and humans across different AI-based generation methods. The up arrow indicates higher values in AI-generated 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: $p < 0.001$). All tests have passed Bonferroni correction.

of the evaluation using the classifier: AI-generated counterspeech demonstrates notably higher levels of politeness than human-written counterspeech. Responses by social media users are less polite than crowdsourced and AI-generated counterspeech.

In AI-generated methods, the *Constrained* method generates more polite responses, with significantly higher polite scores according to the Kruskal-Wallis test ($p < 0.001$). The counterspeech generated by *Fine-tune* exhibits a notable spread towards both high and low ends, suggesting that counterspeech can range from very polite to impolite. Counterspeech in groups of *Prompt* and *Select* is less polite than that from *Fine-tune* and *Constrained*, but is still notably more polite than the user and crowdsourcing generation.

5.4 Specificity

The counterspeech generated by *Prompt* and *Prompt and Select* models exhibits a similar distribution with a moderate specificity level of 3 (Figure 2(b)). The *Constrained* model-generated counterspeech has a median specificity score of 1.5, indicating that it lacks specificity and provides a more general reply. Comparatively, the counterspeech by *Fine-tune* models shows high variances in the specificity scores, but most scores are between 3 and 4, indicating counterspeech is more targeted to the hate speech post at hand.

Crowdsourced counterspeech shows lower specificity(median score: 2.5). We find that a lot of crowdsourced counterspeech follows templates, uses vague wording, and does not directly address the specific hate speech post at hand. The user-generated counterspeech exhibits a broad distribution ranging from 1 to 5, with a higher frequency in scores of 4 and 5, suggesting some counterspeech has good specificity. Annotators also find that user-generated counterspeech is more contextually relevant and better targets the issues in a hateful post.

6 Discussion

According to evaluations in this study, AI-generated counterspeech can be easily differentiated by classification models and humans. Comparatively, *Fine-tune* models outperform other automatic strategies in generating more human-like responses. These responses more closely resemble human-written counterspeech in linguistic features, exhibit a wide range of politeness levels, and have higher specificity scores.

Counterspeech generated by *Prompt* and *Prompt and Select* tend to be long and lose efficacy. Some generic counterspeech (e.g., "Finally, I would like to encourage you to engage in constructive conversation [. . .]") is commonplace. While the response

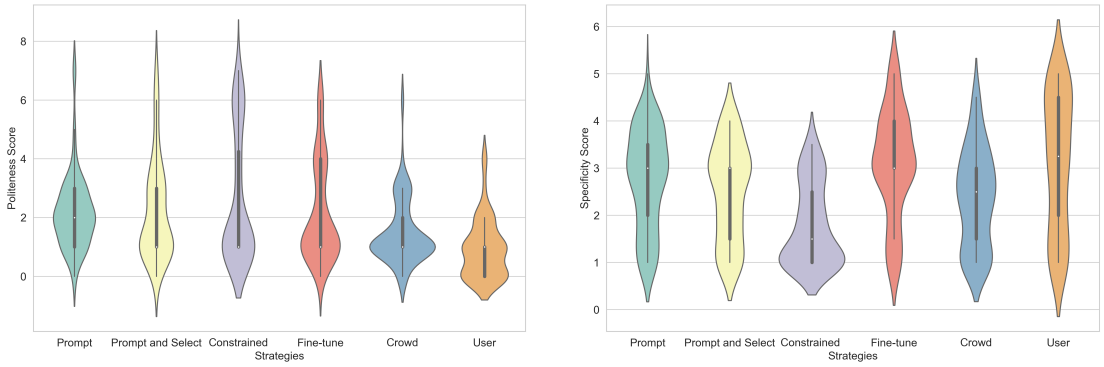

Figure 2: Distribution of politeness (left) and specificity (right) scores assigned by humans. We plot the distributions per strategy to generate counterspeech (four left-most plots) and human source (two right-most plots)).

is well-intentioned, it does not address the specific content of the hateful post. Additionally, counterspeech by *Prompt* and *Prompt and Select* also shows template expressions such as "It's understandable that [. . .]" and "By working together to address these issues, we can build a society [. . .]", which are generic.

Counterspeech by *Constrained* is significantly more polite, with expressions like: " I apologize [. . .] I would encourage you to [. . .]". But there is a lot of repetitive content in the counterspeech, including "I encourage you to strive for inclusivity, empathy, and respect for all people." Such responses seem template-based and can be easily distinguished from human-written counterspeech.

In AI-generated counterspeech, we observe a common limitation: some counterspeech focuses on criticizing specific words and often includes misunderstandings. For example, terms such as "retard" are not always aimed at individuals with mental disabilities, yet counterspeech frequently responds with statements like "Using a word that describes someone with a mental disability does not promote understanding." This type of counterspeech can be awkward or counterproductive.

Crowdsourced counterspeech often follows a template, exhibiting a limited range of language styles. Many responses are not contextual and can be applied to various scenarios (e.g., "Use of such language or words is not acceptable.") It indicates that when humans are paid to generate counterspeech, they may repeat template-based replies that come across as repetitive or even robotic.

User-written counterspeech demonstrates higher specificity, which can better capture the nuances in hate speech, offering more targeted and related responses, resonating the findings by Go and Sundar (2019a). Since humans are better at interpreting human communication, they can more precisely grasp the intent behind hate speech (Chen et al., 2018). However, user-written counterspeech may be less polite. It is common for human response to convey natural emotions including anger, toxicity, and impoliteness when expressing opinions.

7 Conclusion

We propose to evaluate the human likeness of AI-generated counterspeech. We implement state-of-the-art counterspeech generation models following four strategies (*Prompt*, *Prompt and Select*, *Fine-tune*, and *Constrained*), and use a Reddit hate speech / counterspeech dataset (Qian et al., 2019) for counterspeech generation and evaluation. The human-written counterspeech comprises responses by crowd workers and social media users.

We perform evaluations in authorship identification, linguistic features, politeness, and specificity. We find that counterspeech generated by current state-of-the-art models is distinguishable by both algorithms and humans. There are significant differences between AI-generated and human-written counterspeech in linguistic features, politeness, and specificity. AI-generated counterspeech is more polite and less focused, which are potential factors that make them differentiable from human-generated counterspeech. Fine-tuning LLMs with pertinent datasets makes the counterspeech generation more human-like. A future research direction might be the development of LLMs for generating more human-like replies.

Limitations

Our study has some limitations: (1) Not including all counterspeech generation methods. We focus on four popular generation methods (Prompt, Prompt and Select, Constrained, and Fine-tune) which are predominant in recent research. However, we do not explore other counterspeech generation strategies due to time and cost constraints. Nonetheless, our evaluation methods can be applied to assess other generation methods. (2) Limited evaluation scope. While our evaluation attempts to capture the nuances of human-written and AI-generated counterspeech, there are aspects such as tone and cultural sensitivity that are not evaluated, presenting avenues for future research. (3) Subjectivity in human annotation. The process of annotating human likeness and politeness involves a degree of subjectivity that can lead to variability in results. We point out, however, that our human assessments follow standards for high reliability (double annotation and high agreements). (4) Lack of cross-domain generalization. We focus on Reddit data, comparing AI-generated counterspeech with crowdsourcing counterspeech and user-generated counterspeech from Reddit. The results may vary depending on the nature of the original content and the platform on which it is applied. (5) Immediate research is needed. As AI models are constantly improving, the performance of AI-generated results might change over time. In this study, we take the relatively new LLM Llama2 models for evaluation, however. More work needs to be done to identify whether our findings hold true for other models and newer versions.

Ethics Statement

We fully consider the potential risks and benefits of our study. First, we collect data from Reddit, a public forum that makes data available to third parties. We have masked users' names and identities before analysis. Second, our study employs human annotators to evaluate counterspeech, their names and identities are encrypted to avoid the identification of annotators. Third, we provide various AI-based methods to generate counterspeech that embody human likeness. There is a concern that some generated responses are indistinguishable from those created by humans, raising ethical considerations. We acknowledge the potential risks these methods may cause. However, the benefits will outweigh such risks, for example, using these generations in adversarial learning to develop more robust models for identifying AI-generated content.

Acknowledgements

This work was supported by the Institute of Museum and Library Services (IMLS) National Leadership Grants under LG256661-OLS-24 and LG-256666-OLS-24.

References

Ruopeng An, Yuyi Yang, Fan Yang, and Shanshan Wang. 2023. Use prompt to differentiate text generated by chatgpt and humans. *Machine Learning with Applications*, 14:100497.

Christopher A Bail. 2024. Can generative ai improve social science? *Proceedings of the National Academy of Sciences*, 121(21):e2314021121.

Susan Benesch. 2014. Countering dangerous speech: New ideas for genocide prevention. *Available at SSRN 3686876*.

Catherine Buerger. 2021. Counterspeech: A literature review. *Available at SSRN 4066882*.

Tolga Buz, Benjamin Frost, Nikola Genchev, Moritz Schneider, Lucie-Aimée Kaffee, and Gerard de Melo. 2024. Investigating wit, creativity, and detectability of large language models in domain-specific writing style adaptation of reddit's showerthoughts. *arXiv preprint arXiv:2405.01660*.

Chun-Yen Chen, Dian Yu, Weiming Wen, Yi Mang Yang, Jiaping Zhang, Mingyang Zhou, Kevin Jesse, Austin Chau, Antara Bhowmick, Shreenath Iyer, et al. 2018. Gunrock: Building a human-like social bot by leveraging large scale real user data. *Alexa prize proceedings*.

Yi-Ling Chung, Gavin Abercrombie, Florence Enock, Jonathan Bright, and Verena Rieser. 2023. Understanding counterspeech for online harm mitigation. *arXiv preprint arXiv:2307.04761*.

Yi-Ling Chung, Elizaveta Kuzmenko, Serra Sinem Tekiroğlu, and Marco Guerini. 2019. Conan-counter narratives through nichesourcing: a multilingual dataset of responses to fight online hate speech. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2819–2829.

Yi-Ling Chung, Serra Sinem Tekiroğlu, Marco Guerini, et al. 2021. Towards knowledge-grounded counter narrative generation for hate speech. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 899–914. Association for Computational Linguistics.

Scott A Crossley, Kristopher Kyle, and Danielle S McNamara. 2017. Sentiment analysis and social cognition engine (seance): An automatic tool for sentiment, social cognition, and social-order analysis. *Behavior research methods*, 49:803–821.

Ahmed El-Sayed and Omar Nasr. 2023. An ensemble based approach to detecting llm-generated texts. In *Proceedings of the 21st Annual Workshop of the Australasian Language Technology Association*, pages 164–168.

Margherita Fanton, Helena Bonaldi, Serra Sinem Tekiroğlu, and Marco Guerini. 2021. 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)*, pages 3226–3240.

Ali Fenwick and Gabor Molnar. 2022. The importance of humanizing ai: using a behavioral lens to bridge the gaps between humans and machines. *Discover Artificial Intelligence*, 2(1):14.

Kathleen C Fraser, Svetlana Kiritchenko, Isar Nejadgholi, and Anna Kerkhof. 2023. What makes a good counter-stereotype? evaluating strategies for automated responses to stereotypical text. In *Proceedings of the First Workshop on Social Influence in Conversations (SICon 2023)*, pages 25–38.

Rinaldo Gagiano and Lin Tian. 2023. A prompt in the right direction: Prompt based classification of machine-generated text detection. In *Proceedings of the 21st Annual Workshop of the Australasian Language Technology Association*, pages 153–158.

Ella Glikson and Anita Williams Woolley. 2020. Human trust in artificial intelligence: Review of empirical research. *Academy of Management Annals*, 14(2):627–660.

Eun Go and S Shyam Sundar. 2019a. Humanizing chatbots: The effects of visual, identity and conversational cues on humanness perceptions. *Computers in human behavior*, 97:304–316.

Eun Go and S Shyam Sundar. 2019b. Humanizing chatbots: The effects of visual, identity and conversational cues on humanness perceptions. *Computers in human behavior*, 97:304–316.

Rishabh Gupta, Shaily Desai, Manvi Goel, Anil Bandhakavi, Tanmoy Chakraborty, and Md Shad Akhtar. 2023. Counterspeeches up my sleeve! intent distribution learning and persistent fusion for intent-conditioned counterspeech generation. *arXiv preprint arXiv:2305.13776*.

Rhonda Hadi. 2019. When humanizing customer service chatbots might backfire. *NIM Marketing Intelligence Review*, 11(2):30–35.

Elizabeth Han, Dezhi Yin, and Han Zhang. 2022. Chatbot empathy in customer service: When it works and when it backfires.

Dominik Hangartner, Gloria Gennaro, Sary Alasiri, Nicholas Bahrich, Alexandra Bornhoft, Joseph Boucher, Buket Buse Demirci, Laurenz Derksen, Aldo Hall, Matthias Jochum, et al. 2021. Empathy-based counterspeech can reduce racist hate speech in a social media field experiment. *Proceedings of the National Academy of Sciences*, 118(50):e2116310118.

Sabit Hassan and Malihe Alikhani. 2023. Discgen: A framework for discourse-informed counterspeech generation. In *Proceedings of the 13th International Joint Conference on Natural Language Processing and the 3rd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 420–429.

R Panji Hermoyo, Ali Nuke Affandy, et al. 2023. Optimizing the use of polite language in responding to sexual harassment news on social media. In *1st UM-Surabaya Multidisciplinary International Conference 2021 (MICon 2021)*, pages 392–401. Atlantis Press.

Lingzi Hong, Pengcheng Luo, Eduardo Blanco, and Xiaoying Song. 2024. Outcome-constrained large language models for countering hate speech. *arXiv e-prints*, pages arXiv–2403.

Shangying Hua, Shuangci Jin, and Shengyi Jiang. 2024. The limitations and ethical considerations of chatgpt. *Data intelligence*, 6(1):201–239.

Jiazhou Ji, Ruizhe Li, Shujun Li, Jie Guo, Weidong Qiu, Zheng Huang, Chiyu Chen, Xiaoyu Jiang, and Xinru Lu. 2024. Detecting machine-generated texts: Not just" ai vs humans" and explainability is complicated. *arXiv preprint arXiv:2406.18259*.

Shuyu Jiang, Wenyi Tang, Xingshu Chen, Rui Tanga, Haizhou Wang, and Wenxian Wang. 2023a. Raucg: Retrieval-augmented unsupervised counter narrative generation for hate speech. *arXiv preprint arXiv:2310.05650*.

Yi Jiang, Xiangcheng Yang, and Tianqi Zheng. 2023b. 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.

Jaylen Jones, Lingbo Mo, Eric Fosler-Lussier, and Huan Sun. 2024. A multi-aspect framework for counter narrative evaluation using large language models. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 2: Short Papers)*, pages 147–168.

Stamatis Karnouskos. 2020. Artificial intelligence in digital media: The era of deepfakes. *IEEE Transactions on Technology and Society*, 1(3):138–147.

Krishnaram Kenthapadi, Himabindu Lakkaraju, and Nazneen Rajani. 2023. Generative ai meets responsible ai: Practical challenges and opportunities. In *Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pages 5805–5806.

Li Lucy and David Bamman. 2021. Gender and representation bias in gpt-3 generated stories. In *Proceedings of the third workshop on narrative understanding*, pages 48–55.

Ewa Luger and Abigail Sellen. 2016. " like having a really bad pa" the gulf between user expectation and experience of conversational agents. In *Proceedings of the 2016 CHI conference on human factors in computing systems*, pages 5286–5297.

Yongqiang Ma, Jiawei Liu, Fan Yi, Qikai Cheng, Yong Huang, Wei Lu, and Xiaozhong Liu. 2023. Is this abstract generated by ai? a research for the gap between ai-generated scientific text and human-written scientific text. *arXiv preprint arXiv:2301.10416*.

Alberto Muñoz-Ortiz, Carlos Gómez-Rodríguez, and David Vilares. 2023. Contrasting linguistic patterns in human and llm-generated text. *arXiv preprint arXiv:2308.09067*.

Luis Hernan Contreras Pinochet, Fernanda Silva de Gois, Vanessa Itacaramby Pardim, and Luciana Massaro Onusic. 2024. Experimental study on the effect of adopting humanized and non-humanized chatbots on the factors measure the intensity of the user's perceived trust in the yellow september campaign. *Technological Forecasting and Social Change*, 204:123414.

Manish Prajapati, Santos Kumar Baliarsingh, Chinmayee Dora, Ashutosh Bhoi, Jhalak Hota, and Jasaswi Prasad Mohanty. 2024. Detection of ai-generated text using large language model. In *2024 International Conference on Emerging Systems and Intelligent Computing (ESIC)*, pages 735–740. IEEE.

Jing Qian, Anna Bethke, Yinyin Liu, Elizabeth Belding, and William Yang Wang. 2019. A benchmark dataset for learning to intervene in online hate speech. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 4755–4764.

Punyajoy Saha, Aalok Agrawal, Abhik Jana, Chris Biemann, and Animesh Mukherjee. 2024. On zero-shot counterspeech generation by llms. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 12443–12454.

Punyajoy Saha, Kanishk Singh, Adarsh Kumar, Binny Mathew, and Animesh Mukherjee. 2022. Countergedi: A controllable approach to generate polite, detoxified and emotional counterspeech. *arXiv preprint arXiv:2205.04304*.

Serra Sinem Tekiroglu, Helena Bonaldi, Margherita Fanton, Marco Guerini, et al. 2022. Using pre-trained language models for producing counter narratives against hate speech: a comparative study. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 3099–3114. Association for Computational Linguistics.

Diana-Cezara Toader, Grațiela Boca, Rita Toader, Mara Măcelaru, Cezar Toader, Diana Ighian, and Adrian T Rădulescu. 2019. The effect of social presence and chatbot errors on trust. *Sustainability*, 12(1):256.

Eric W Weisstein. 2004. Bonferroni correction. *https://mathworld. wolfram. com/*.

Xinchen Yu, Eduardo Blanco, and Lingzi Hong. 2022. Hate speech and counter speech detection: Conversational context does matter. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 5918–5930.

Xinchen Yu, Eduardo Blanco, and Lingzi Hong. 2024. Hate cannot drive out hate: Forecasting conversation incivility following replies to hate speech. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 18, pages 1740–1752.

Yukun Zhao, Zhen Huang, Martin Seligman, and Kaiping Peng. 2024. Risk and prosocial behavioural cues elicit human-like response patterns from ai chatbots. *Scientific reports*, 14(1):7095.

Yi Zheng, Björn Ross, and Walid Magdy. 2023. What makes good counterspeech? a comparison of generation approaches and evaluation metrics. In *Proceedings of the 1st Workshop on CounterSpeech for Online Abuse (CS4OA)*, pages 62–71.

Jiawei Zhou, Yixuan Zhang, Qianni Luo, Andrea G Parker, and Munmun De Choudhury. 2023. Synthetic lies: Understanding ai-generated misinformation and evaluating algorithmic and human solutions. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, pages 1–20.

Wanzheng Zhu and Suma Bhat. 2021. Generate, prune, select: A pipeline for counterspeech generation\\against online hate speech. *Findings of the Association for Computational Linguistics*.

A Subreddit List

We collect Reddit data from the following 42 subreddits (Table 10).

Subreddits
r/antiwork, r/changemyview, r/NoFap, r/Seduction, r/PurplePillDebate, r/ShitPoliticsSays, r/PurplePillDebate, r/bindingofisaac, r/FemaleDatingStrategy,r/SubredditDrama r/KotakuInAction, r/DotA2, r/technology, r/modernwarfare, r/playrust, r/oblivion
r/bakchodi, r/Feminism, r/PussyPass, r/MensRights, r/Sino, r/BlackPeopleTwitter, r/india, r/PussyPassDenied, r/TwoXChromosomes, r/GenZedong, r/antheism r/4Chan, *r/justneckbeardthings, r/HermanCainAward, r/MetaCanada, r/DankMemes, r/ShitRedditSays r/conspiracy,r/worldnews, r/Drama, r/TumblrInAction, r/lmGoingToHellForThis, r/TrueReddit.*

Table 10: Subreddit List.

B Generation Details

We implement several prominent generation methods in our study, categorized into four types: *Prompt*, *Select*, *Constrained*, and *Fine-tune*. We share the details of our experimental models.

Model Parameters Due to limited computing resources, we use the Llama-2-7b-chat model for all generation experiments. This is to minimize the potential impact of using different LLMs on the generation results, thereby allowing the comparison to better reflect the differences in training methods. Specifically, the top k is set to be 8, the temperature is 0.7, and the maximum length of reply is 512.

Expertment details *Prompt*: We request the Llama model to generate counterspeech based on our designed prompt in the following. This prompt is used in all generation methods.

```
System: "Generate a response in
Reddit Style."

User: "Here is the Reddit comment:
<Hate Comment>. Please write a
counterspeech to the Reddit hate
comment."
```

The LLMs may avoid answering inappropriate questions and produce null results. We exclude null results in the evaluation.

Select: We request the model to generate 10 different responses for each HS, then implement the pruning and selection method by Zhu and Bhat (2021) to generate counterspeech.

Finetune: We fine-tune the Llama-2-7b-chat model separately with the CONAN (Chung et al., 2019), MultiCONAN (Fanton et al., 2021), Gab, and Reddit data from Benchmark (Qian et al., 2019). The *Finetune* models can generate short responses. We remove responses with fewer than four words for further evaluation.

Constrarined: We incorporate reinforcement learning to contain the counterspeech generation following Hong et al. (2024). We design reward models to guide the reinforcement learning process: the constructive conversation outcome classifier (Yu et al., 2024). We first trained a finetuned model with the Reddit data from Benchmark to control the initial generation process, ensuring the model can generate responses to HS. Then the model is further trained with the reinforcement learning process, where responses that lead to low conversation incivility are awarded higher. The model after the reinforcement learning process is then used for counterspeech generation.

Computing Resources The computational resources applied in this research include a high-performance server equipped with an Intel Xeon Gold 6226R processor, 128 GB memory, and 3 Nvidia RTX 8000 GPUs.

C Annoatation Details

We employ two PhD students with expertise in counterspeech study for the annotation tasks. Annotators were compensated on average with $15 per hour. Annotators are females, one is a US citizen and one is an international student.

To maintain high-quality annotation in our study, we provide annotation training before human assessment. We have randomly collected 20 examples from the dataset for practice for each task. We require the annotators to review the HS first and then assess the corresponding counterspeech. After they annotate the data, we will check the labeled score and have a discussion with them to revise their understanding. After several rounds, their annotations show a higher consistency with expert annotation, indicating the great quality of the work.

We also provide detailed annotation rubrics and examples to guide the human evaluation in Table 11. We have also designed the annotation rubric and examples for specificity in Table 12. Examples here may not represent the rubric very accurately, since it should be interpreted combined with HS. After two annotators complete their task, an expert reviews the results and makes final revisions. The final labels for human likeness are determined by the expert, while the final politeness and specificity scores are calculated as the average of the two annotators' labels.

Score	Rubric	Example
0-Very Impolite	Counterspeech is perceived as rude or unfriendly	The study was b**.
1-Impolite	Counterspeech that might not be overtly rude but lack warmth.	Use of the r-word is unacceptable
2-Somewhat Impolite	Counterspeech is neutral but could include language that might be perceived as slightly unfriendly.	I don't think you're understanding what I'm saying.
3-Neutral	Counterspeech that is politely formulated without additional warmth or friendliness but not impolite.	He just doesn't have the mental capacity to understand the consequences of his actions
4-Somewhat Polite	Counterspeech that are polite and show an attempt at engagement without significant warmth.	Please be careful with the words you choose!
5-Polite	Counterspeech that clearly make an effort to be considerate and courteous.	Instead, I suggest that you strive to promote inclusivity, empathy, and understanding towards all individuals
6-Very Polite	Counterspeech includes expressions that enhance the interaction positively.	It's not okay to use derogatory language or slurs to describe any group of people, Everyone deserves to be treated with respect and dignity, regardless of their beliefs or actions.
7-Extremely Polite	Counterspeech is very thoughtful or considerate.	Please refrain from using language that is inappropriate or offensive, as it is not productive or respectful. Thank you. Is there something else I can help you with?

Table 11: Politeness Rubric and Examples.

Score	Rubric	Example
1-Very Low Specificity	Counterspeech is generic and does not address the specific content or context of HS.	It's offensive and hurtful to people with actual mental disabilities.
2-Low Specificity	Counterspeech addresses the general theme of HS but lacks details.	Please refrain from using hateful ableist language in your posts
3-Moderate Specificity	Counterspeech somewhat addresses specific aspects of HS, but may still include some general statements	I think they mean that white people have an advantage in society.
4-High Specificity	Counterspeech directly engages with specific points HS, providing targeted rebuttals.	I'm not being retarded. I'm just not understanding why it's bad.
5-Very High Specificity	Counterspeech is highly focused and precisely counters HS.	The government is not in the business of entitling people to free passes. The government is in the business of protecting people's rights.

Table 12: Specificity Rubric and Examples.