What Makes Good Counterspeech?
A Comparison of Generation Approaches and Evaluation Metrics

Yi Zheng, Björn Ross and **Walid Magdy**
Institute for Language, Cognition & Computation
School of Informatics
University of Edinburgh
Y.Zheng-77@sms.ed.ac.uk, b.ross@ed.ac.uk, wmagdy@inf.ed.ac.uk

Abstract

Counterspeech has been proposed as a solution to the proliferation of online hate. Research has shown that natural language processing (NLP) approaches could generate such counterspeech automatically, but there are competing ideas for how NLP models might be used for this task and a variety of evaluation metrics whose relationship to one another is unclear. We test three different approaches and collect ratings of the generated counterspeech for 1,740 tweet-participant pairs to systematically compare the counterspeech on three aspects: quality, effectiveness and user preferences. We examine which model performs best at which metric and which aspects of counterspeech predict user preferences. A free-form text generation approach using ChatGPT performs the most consistently well, though its generations are occasionally unspecific and repetitive. In our experiment, participants' preferences for counterspeech are predicted by the quality of the counterspeech, not its perceived effectiveness. The results can help future research approach counterspeech evaluation more systematically.

1 Introduction

Social platforms are known to create an echo chamber for extreme opinions, and the associated online hate can motivate violent actions in real life (Müller and Schwarz, 2021). Many researchers see counterspeech as an effective way to intervene with online hate and extremist content (Garland et al., 2022; Schieb and Preuss, 2016; Cypris et al.). Counterspeech is a non-negative response to hateful content that aims to refute stereotypes and misinformation with arguments, while upholding the principle of free speech (Kiritchenko et al., 2021).

It is hoped that such counterspeech can be automatically generated, using Natural Language Processing (NLP) techniques. Recent research has explored a wide range of state-of-the-art neural models for the task of generating counterspeech for a given example of hate speech, in particular pretrained transformer models such as BERT, BART, GPT (Qian et al., 2019; Pranesh et al., 2021; Zhu and Bhat, 2021; Chung et al., 2021; Ashida and Komachi, 2022; Gupta et al., 2023). To evaluate the generated counterspeech, most research uses classical NLP metrics (BLEU, ROUGE, BERTScore) to measure language quality, as well as human evaluation of counterspeech effectiveness and suitability (Zhu and Bhat, 2021; Chung et al., 2021; Ashida and Komachi, 2022; Tekiroğlu et al., 2022). Some researchers have examined correlations between characteristics of the counterspeech and its effectiveness (Munger, 2021; Hangartner et al., 2021; Kim et al., 2022; Obermaier et al.). However, existing research occasionally conflates the different aspects of counterspeech in evaluation, or only one of the aspects is measured, so we separate them into different evaluation metrics and study how they are related.

We compare three approaches to counterspeech generation. All are based on LLMs but they differ fundamentally: a template-based approach using GPT-2 fine-tuned on training data, a template-based approach using GPT-3 without fine-tuning (i.e. zero-shot) and zero-shot free-form text generation (ChatGPT). We conduct a survey-based experiment to address the following research questions: (1) How do different types of approaches to counterspeech generation differ in the quality of the counterspeech generated? (2) How do different types of approaches to counterspeech generation differ in the perceived effectiveness of the counterspeech generated? (3) What are the aspects that people prefer in counterspeech?

2 Background

2.1 Approaches to Generating Counterspeech

Over the last few years, different types of approaches have been used to generate counterspeech.

Proceedings of the 1st Workshop on CounterSpeech for Online Abuse (CS4OA), pages 62–71
September 11, 2023 ©2023 Association for Computational Linguistics

Approximately since the release of BERT and GPT in 2018, the dominant approach for many NLP tasks has been to fine-tune a pre-trained transformer model on a task-specific dataset. For example, Pranesh et al. (2021) fine-tuned BERT, DialoGPT and BART, and Chung et al. (2021) fine-tuned GPT models.

This was followed by a generation of models including GPT-3 (released in 2020) that have performed reasonably well on many NLP tasks with only minimal fine-tuning, or none at all (also known as “few-shot learning” and “zero-shot transfer learning”, respectively, though this terminology is not without ambiguity). Ashida and Komachi (2022) have explored the potential of LLMs for the generation of counterspeech. They use zero-shot and few-shot prompts on GPT-2, GPT-Neo (Black et al., 2022) and GPT-3 to counter both hate and microaggressions. Their results show that GPT-2 is unable to generate high quality counterspeech with zero-shot and one-shot prompts, but that GPT-3 can produce meaningful output. Their study demonstrates the potential of using automated counterspeech on online social platforms.

The latest generation of LLMs such as ChatGPT (released in 2023) has demonstrated an improved ability to generate free-form text, though given how recently ChatGPT was released, it is still unclear how much it outperforms previous models in terms of counterspeech generation.

2.2 User Preferences and Evaluation of Generated Counterspeech

Various approaches have been used to evaluate the quality of the generated counterspeech. Besides automatic metrics such as BLEU and ROUGE, many use human annotators to evaluate different facets such as appropriateness and effectiveness. Chung et al. (2021); Zhu and Bhat (2021) asked annotators to rate “suitableness” using a five-point Likert scale, while Ashida and Komachi (2022) segmented counterspeech suitableness into two labels called text “offensiveness” and “stance”. Fraser et al. (2023) also used human annotators (4 participated) to test ChatGPT’s ability to generate high quality counterspeech. One of their questions asked annotators to judge the effectiveness of counterspeech in countering stereotypes from a bystander’s perspective (whether the generated counters-stereotype is “one of the best possible responses to this stereotype”, “an OK response, but it could be better”, or “not a good response to this stereotype”).

Some approaches try to use objective criteria, such as changes in user behaviours to measure effectiveness. Nuti et al. (2022) created a simulated conversational testbed to evaluate counterspeech mediation on text sentiment and the number of hate speeches during discourse flow. Studies such as Hangartner et al. (2021) and Kim et al. (2022) tested whether different counterspeech characteristics could influence effectiveness. A field experiment in Hangartner et al. (2021) tested counterspeech with different strategies to reduce exclusionary behaviour of hate speakers on Twitter. The results show that empathy-based counterspeech is the most effective one. Kim et al. (2022) examines the correlation between several counterspeech characterstics (gender of the counter speaker and the number of upvotes) and the willingness of bystanders to report hate speech.

3 Methodology

We generate counterspeech using three models that represent different types of approaches: the first is to use GPT-2 fine-tuned on a task-specific dataset to augment templates with information specific to the countered hate speech, the second approach uses GPT-3 without any fine-tuning (i.e. zero-shot) to augment the templates, and the third approach uses ChatGPT to generate free-form text.

3.1 GPT-2 Pipeline

The GPT-2 pipeline works in two steps. First, the GPT-2 model performs a task called implied statement generation (proposed by ElSherief et al., 2021). The implied statements are then transformed into counterspeech responses using counterspeech templates.

Online hate speech is not always easy to detect, as users sometimes use coded hate symbols and idiomatic speech to evade censorship. Generating hate speech-specific counterspeech therefore requires recognition of these implicit statements. We fine-tune a pre-trained GPT-2 model (HuggingFace’s Pytorch version with 124M parameters) on the Latent Hatred dataset (ElSherief et al., 2021) using the method and training settings described in the same paper. The dataset contains about 6346 pairs of implicit hate speech and implied statements. After fine-tuning, GPT-2 should be better at detecting hate targets and explaining metaphorical language. Some examples of targets and implied

Hate Speech	Targets	Implied Statement
Immigration, if done the right way, can help a nation. Mass immigration - no matter how you slice it is an invasion meant to destroy a nation	Immigrants	Immigrants are invaders.
It's just Jewish power flexing their muscles because they hate white people for existing.	Jews	Jews hate white people.
We the British people must try and get rid of the Muslims. To live in a civilised way.	Muslims	Muslims are uncivilised.

Table 1: Examples of targets and implied statements generated by the GPT-2 pipeline.

statements generated by GPT-2 are shown in Table. 1.

We create a total of 20 counterspeech templates from Qian et al. (2019)'s expert-written dataset. Parts of the counter speaker's statements are replaced by placeholders "[T]" and "[IS]", to be filled with generated targets and implied statements. The templates are neutral in tone and are designed not to affect people's understanding of the whole counterspeech. An example template for countering anti-immigrant hate speech:

> *Please don't say [IS]. The [T] are human, just like us. We're all doing our best in our jobs and with the lives dealt to us.*

3.2 GPT-3 Pipeline

GPT-3 is a more recent model than GPT-2 and it has been shown that for many tasks, it requires less training data to achieve equal or better performance (Brown et al., 2020). We therefore use GPT-3 to generate targets and implied statements (the same task as for GPT-2) by zero-shot prompting. We use the GPT-3 model Text-DaVinci-003, which has far more parameters than its predecessor GPT-2. The expanded training data helps GPT-3 handle niche topics and complex instructions. We used the following prompt, using the OpenAI Playground:

> *What is the target and implied statement of the following hate speech:*

3.3 ChatGPT

Finally, we use ChatGPT (3 May 2023 version) which is built upon GPT-3.5. This model has been widely praised for its performance to generate free-form text (Rathore, 2023). We deliberately keep the prompt simple to investigate how well ChatGPT generates counterspeech without detailed instructions or "prompt engineering". The following prompt was used:

> *Give me a short counterspeech of fewer than 40 words to this hate speech:*

4 Research Design

4.1 Hate Speech Collection

Before conducting the counterspeech experiment, we launched a pilot study to collect hate speech from Twitter. We ask the participants to rate the intensity of hatred in each Tweet (called "hatefulness"). This is to ensure that the Tweets from which we generate counterspeech are actually perceived as hateful.

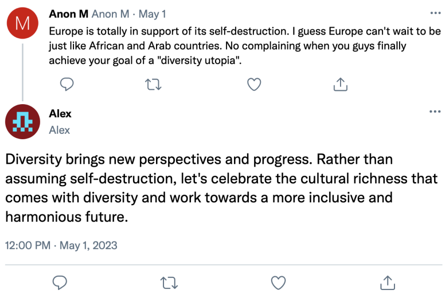
The data we collect focus on four topics of hate: racism, anti-immigration, Islamophobia and anti-Semitism. First, we manually review and collect a large number of potentially hateful Tweets (all posted in April 2023). We search using keywords and phrases, for example, "mass immigration". We then recruit 15 participants from Prolific to rate the hatefulness of each Tweet. Because some Tweets may contain implicit hatred, we also present them with the conversational context to help them understand. We ask the question:

- How hateful is the tweet meant to be towards the target group, please rate the speaker's intention.

The question uses a five-point Likert scale (1 = Not at all hateful, 2 = Slightly hateful, 3 = Hateful, 4 = Very hateful, 5 = Extremely hateful). Each tweet's hatefulness score is the average of all participants' ratings. We select 60 tweets with the highest hatefulness scores (15 tweets for each hatefulness topic). All selected tweets have an average rating higher than 3.

4.2 Experimental Design

Our counterspeech experiment approximates a real-life scenario in which participants encounter hate speech and counterspeech replies on Twitter. The



Anon M Anon M · May 1

Europe is totally in support of its self-destruction. I guess Europe can't wait to be just like African and Arab countries. No complaining when you guys finally achieve your goal of a "diversity utopia".

Alex

Alex

Diversity brings new perspectives and progress. Rather than assuming self-destruction, let's celebrate the cultural richness that comes with diversity and work towards a more inclusive and harmonious future.

12:00 PM · May 1, 2023

Figure 1: Example of a counterspeech screenshot. Counterspeech generated by ChatGPT.

overall goal of the experiment is to measure counterspeech quality and effectiveness and to understand human preferences for counterspeech. The experiment follows a within-subjects design, where each of the 29 participants recruited from Prolific reads hate speech and counterspeech generated by the three models.

We use an online tool[1] to create fake screenshots for the 60 Twitter conversations. Then, for each hateful Tweet, we create three screenshots of the Tweet followed by a counterspeech response. Fig. 1 is an example of the counterspeech screenshots. The names and profile pictures of the hate speakers and counter-speakers are randomly assigned in each screenshot. Participants read the 60 Twitter conversations in random order. Within each conversation, the order of the counterspeech screenshots shown to participants is also random. For each counterspeech screenshot, we ask four questions:

1. Does the counterspeech understand the hate speaker correctly? (Choices are no, partially, yes)

2. Does the counterspeech mention the correct target? (Choices are no, yes)

3. Do you think this response can make the targets feel better?

4. Do you think this response can help other users empathise with the target?

Q1 and Q2 measure the quality of generation. Q3 and Q4 measure counterspeech effectiveness using a 5-point Likert scale (1 = Strongly disagree, 3 = Neutral, 5 = Strongly agree). After answering questions for all the models, participants are asked to give their own preference for the three counterspeech responses to that particular hate speech (including a "none of the above" option). We ask the question:

- Which response do you prefer?

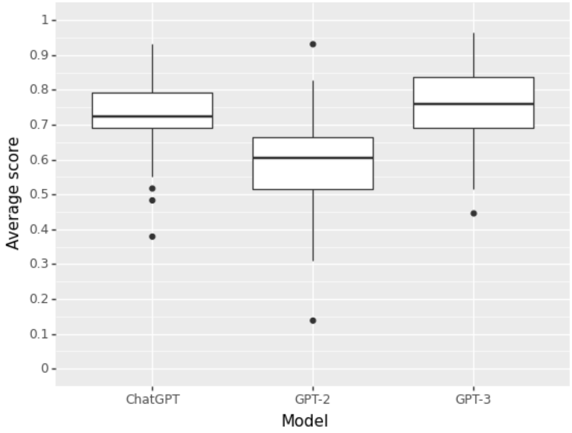


Figure 2: Results for Q1, analysis of counterspeech fully interpreted hate speech correctly. Boxplot of proportion of participants who agreed with the question (observations are tweets). Averaged over all tweets, the percentages of participants agreeing with ChatGPT, GPT-2 and GPT-3 are 72.8%, 59.4% and 76.0%.

5 Results

5.1 Counterspeech Quality

5.1.1 Interpreting Hate Speech

To analyse the results for Q1, we first calculate, for each of the 60 tweets, the proportion of participants who agreed that the counterspeech understood the hate speaker correctly. The box plot (Fig. 2) shows the distribution of this percentage over the 60 tweets. A two-tailed paired-samples t test shows that the mean percentage of people who agree with Q1 is significantly lower for GPT-2 than for GPT-3 at the 5% level ($p = 3\text{e}{-11}$). This result is expected because the data that was used to fine-tune GPT-2 consists mostly of short but broad implied statements, usually less than 10 words (ElSherief et al., 2021). The implied statement may not mention specific claims from the hate speech. The difference between GPT-3 and ChatGPT in the mean percentage of people who agreed with Q1 is not statistically significant ($p = 0.06$). However, Fig. 2 shows that data for ChatGPT is less scattered, indicating stable and consistent performance.

Fig. 3 shows the results after combining "Yes" and "Partially". On average, 88% of the partici-

[1] https://www.tweetgen.com/

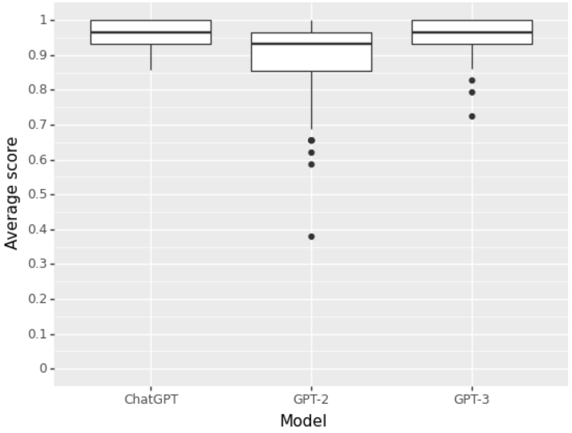


Figure 3: Results for Q1 after combining “Yes” and “Partially” together. Averaged across all hate speech, the percentages of participants agreeing with ChatGPT, GPT-2 and GPT-3 are 96.3%, 88.0% and 95.3%.

pants think that the counterspeeches generated by GPT-2 are at least partially correct. Without mentioning hate speakers’ claims, the counterspeech generated by GPT-2 is still somewhat relevant. However, GPT-2’s score is still significantly lower than that of GPT-3 ($\alpha = 0.05$, $p = 2\text{e}{-5}$) and ChatGPT ($\alpha = 0.05$, $p = 6\text{e}{-6}$).

5.1.2 Finding the Correct Targets

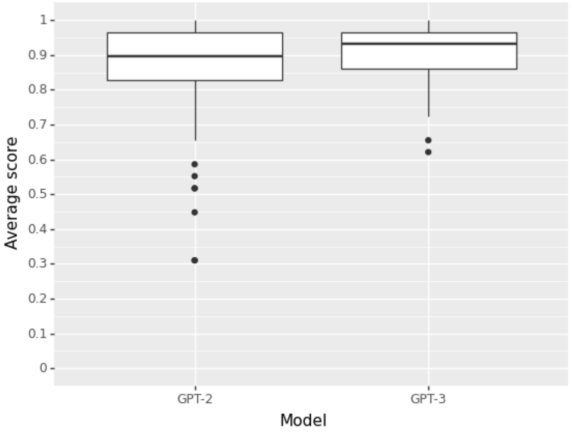


Figure 4: Results for Q2, analysis of counterspeech with the correct target. Boxplot of proportion of participants who agreed with the question (observations are tweets). Averaged over all tweets, the percentage of cases where participants agreed that the system had included the correct target in the counterspeech was 62.8% for ChatGPT (not included in the figure), 84.1% for GPT-2 and 90.7% for GPT-3.

Both GPT-2 and GPT-3 achieve good performance when generating the target of hate speech, as shown in Fig. 4, where the y axis represents the percentage of participants who agree a counterspeech found the correct target. There are cases for both GPT-2 and GPT-3 where all participants agree with the model. However, among the three models, GPT-3 has the fewest outliers in Fig. 4, and the data distribution is also the least dispersed. These observations show that GPT-3 has the most stable and consistent performance when generating targets.

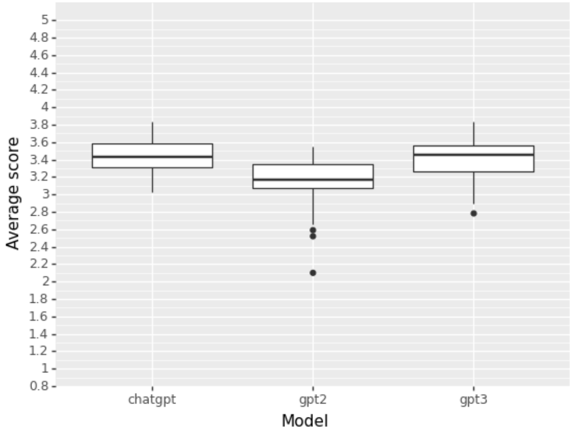


Figure 5: Result for Q3. Analysis of counterspeech making the targets feel better. Boxplot of mean ratings, averaged over all participants, on the question (observations are tweets). Averaged over all tweets, the mean ratings for ChatGPT, GPT-2 and GPT-3 are 3.45, 3.15 and 3.40.

ChatGPT is less likely to include the correct targets in the counterspeech. Only in about 62.8% of cases did participants respond that ChatGPT had included the correct target. However, because the prompts used for ChatGPT did not ask the model explicitly to mention the targets, a lower number here does not necessarily indicate worse performance. The difference in means between GPT-3 and GPT-2, as well as between GPT-3 and ChatGPT, are both statistically significant at the 5% level according to a two-tailed paired-samples t test ($p = 1\text{e}{-4}$ and $p = 6\text{e}{-23}$).

5.2 Counterspeech Effectiveness

We define counterspeech as effective if it can make the targets feel better, and help other bystanders empathise with the targets. The distributions of the mean ratings of all participants agreeing with each model on the two questions are shown in Fig. 5 and Fig. 6. The two graphs show that the mean ratings for all models are slightly better than 3 for both questions. There is only a slight tendency for participants to agree that counterspeech is effec-

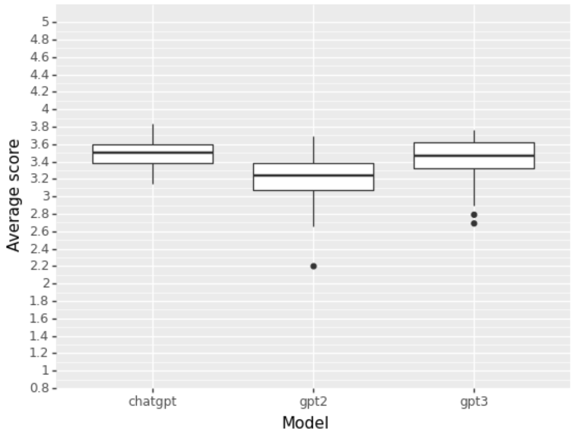


Figure 6: Result for Q4. Analysis of counterspeech helping bystanders empathise with the targets. Boxplot of mean ratings, averaged over all participants, on the question (observations are tweets). Averaged over all tweets, the ratings for ChatGPT, GPT-2 and GPT-3 are 3.50, 3.21 and 3.43.

tive. However, there's a stronger tendency when they read counterspeech generated by GPT-3 and ChatGPT.

For both questions, the differences between the means of GPT-3 and ChatGPT are not statistically significant at the 5% level according to a two-tailed paired-samples t test ($p = 0.16$ and $p = 0.08$ in each question). However, ChatGPT's data is always the least scattered of the three models with no outliers, indicating consistent performance.

5.3 Participants' Preferences in Generated Counterspeech

Table 2 shows the total number and percentage of cases where the counterspeech generated by each model is preferred by a participant. It shows that participants prefer the counterspeech of GPT-3 and ChatGPT in the majority of cases (35.1%, 41%). The participants' preference is consistent with our observations mentioned earlier, as ChatGPT usually has the most stable and consistent performance.

	ChatGPT	**GPT-2**	**GPT-3**	**None**
Count	711	292	610	127
%	40.9	16.8	35.1	7.3

Table 2: Number and percentage of times that each model's response was preferred by a participant.

We then use Multinominal Logistic Regression to investigate which aspects of the counterspeech predict participants' preferences. We use participants' ratings of their preferred model (the four questions in the survey) as the independent variables (predictors), and the model preference itself as the dependent variable (outcome). After removing cases where participants chose "None" as their preference, the outcome is a three-way multinomial variable with GPT-2 as the reference category. The results are shown in Table 3 (LLR $p = 2e{-39}$). When comparing the model preference for GPT-3 and GPT-2, only feature Q2 has $p < 0.05$ (0.001). The other predictors are not statistically significant. When comparing the model preference for ChatGPT and GPT-2, both Q1 and Q2 have $p < 0.05$. Predictors Q3 and Q4 are not statistically significant. This shows that under our test settings, the two measures of quality are the most important in determining participants' preference for counterspeech, more so than the measures of perceived effectiveness.

GPT-3	**coef**	**P>|Z|**
const	-1.342	0.002
Q1	0.308	0.123
Q2	1.083	0.001
Q3	0.169	0.279
Q4	0.003	0.984
ChatGPT	**coef**	**P>|Z|**
const	0.061	0.872
Q1	0.613	0.002
Q2	-1.645	0.000
Q3	0.238	0.128
Q4	0.142	0.383

Table 3: Multinominal Logistic Regression Results, with coefficient values and p-values for each feature. (LLR P-value = $2e{-39}$)

In the survey, we also asked participants directly to give reasons for their preferred counterspeech. Although this is an optional question, 15 out of 29 participants gave their reasons. Here we summarise some of the most common opinions:

1. Choice: None [of the models]
   - Don't seem to address the hate speaker's language directly.
   - This post is extremely hateful. All the counterspeech is too soft.
   - This post is extremely hateful. The hate speaker should just be banned.
   - All are broad statements. They don't

point to the correct group.

2. Choice: ChatGPT
   - Good wording. Text is fluent.
   - It addresses all parts of the hate speech.
   - It suggests an alternative route.
   - It is a clear and sharp comment denouncing hate speech.
3. Choice: GPT-3
   - It explains more about why it is wrong to stereotype.
4. Choice: GPT-2
   - It refers to the original text, and gives a positive message of not using those words again.

5.4 Analysis of ChatGPT's Counterspeech

When analysing the survey results, we observe some repetition in the counterspeech generated by ChatGPT. To test this observation, we compute pairwise binary bag-of-words cosine similarity scores between examples of generated counterspeech (the model output), and as a baseline for comparison, also between examples of hate speech (the model input). For each item in the set of anti-immigration generated counterspeech, we average the cosine similarity scores between that item and the rest of the items. We then compare the anti-immigration counterspeech to counterspeech from the other three topics. The same process is repeated for hate speech. The results of the within-topic and across-topic similarity are shown in Fig. 7. The counterspeech set always has higher similarity scores (around 0.20) compared to the hate speech test (around 0.75), and their scores remain unchanged even when comparing counterspeech across topics. This confirms our observation that there is at least a certain amount of repetition in the wording and phrasing of ChatGPT's counterspeech. This may be one of the explanations for ChatGPT's stable and consistent behaviour.

6 Discussions and Conclusion

We compare three LLM-based approaches to counterspeech generation on quality, perceived effectiveness and user preferences. The results show that all three approaches can generate customised counterspeech relevant to hate speech. Participants tend to agree that counterspeech by all three approaches is effective at helping targets of hate and bystanders. Among the three approaches, GPT-3 and ChatGPT have the most stable and consistent performance. Meanwhile, despite not imposing any structure or template on ChatGPT's output, we noticed some repetition in the counterspeech generated. Both GPT-2 and GPT-3 use the same template-based approach. However, we observe that GPT-2's fine-tuning approach doesn't fully "understand" hate speech, even though it requires the most manual effort. A larger LLM with more pre-training data using zero-shot outperformed a smaller model that had been fine-tuned on task-specific data. This is in line with results obtained by others on other tasks (Brown et al., 2020; Wu et al., 2023).

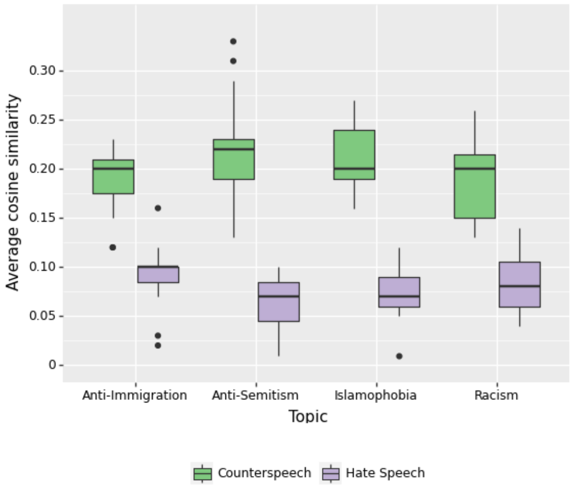


Figure 7: Means of cosine similarity scores when comparing anti-immigration counterspeech or hate speech within and across topics. For comparison, green is for counterspeech, purple is hate speech.

We investigate user preferences for counterspeech aspects. The results of the regression analysis show that both effectiveness questions do not predict model preference. Instead, people tend to prefer counterspeech with a better understanding of hate speech. ChatGPT is most often preferred by users, followed by GPT-3. Table 4 is an example where most participants prefer GPT-2's output (GPT-2: 13; GPT-3: 7; ChatGPT: 6; None: 3). In this example, none of the approaches are able to generate satisfactory counterspeech from hate speech:

> *The only value of the #EDL was to stand for English interests. They called out Muslims but didn't say shit about blacks despite blacks being just as bad as Mus-*

Model	Counterspeech	Q1	Q2	Q3	Q4
GPT-2	You're suggesting "Muslims are violent". This language is racist, offensive and constitutes hate speech directed at Muslims. It's derogatory and wrong.	55%	65%	3.41	3.52
GPT-3	You're suggesting "the EDL was a fake nationalist movement that was only interested in attacking Muslims, while ignoring other groups such as blacks, who may have been just as bad when it comes to crime". This language is racist, offensive and constitutes hate speech directed at EDL. It's derogatory and wrong.	65%	62%	2.79	2.69
ChatGPT	Prompting equality means addressing all forms of discrimination. Let's reject hate speech, work towards understanding, and build inclusive societies that value and respect individuals regardless of their race or religion.	37%	44%	3.21	3.17

Table 4: Counterspeech generated by each model and averages of participants' ratings to each question.

> *lims when it comes to crime etc. They were a red herring and fake nationalist movement. #altleft*

This example shows that GPT-3 can be very good at interpreting hate speech. However, without fine-tuning, an implied statement doesn't necessarily equate to implied hate speech (at least in this case). The counterspeech generated is actually offensive. Meanwhile, ChatGPT generates a platitude that participants tend to agree with, but the counterspeech itself is not specific or relevant to the hate speech and most participants prefer others.

When we invited participants to provide optional additional feedback on the generated counterspeech, many participants made clear that they prefer counterspeech that is direct and to the point, rather than platitudes. Good counterspeech should point out the correct targets of hate and give detailed explanations. Fluency and wording also influence people's opinions of counterspeech. Meanwhile, for extremely hateful posts, some participants mentioned that they would prefer a strong punishment for the hate speaker rather than friendly counterspeech, such as a ban or a more aggressive response. Mathew et al. (2019) also notes that different online communities have their own preferences when it comes to countering hate. This suggests the need for intent-specific or attribute-specific counterspeech generation models (Saha et al., 2022; Gupta et al., 2023), and models that tailor counterspeech to the characteristics of hate speech.

7 Limitations

Our approach is not without limitations. In the design of our evaluation measures, we cover three aspects of counterspeech systematically but there are other aspects of counterspeech that we do not evaluate. For example, we measure perceived effectiveness at making targets feel better and at helping others empathise with the targets but we do not measure effectiveness at convincing the hate speaker to reduce hate, which would require an entirely different research design. In our choice of models, we select one model from the GPT family as representative of each paradigm (e.g. GPT-2 for fine-tuning a smaller model and ChatGPT for zero-shot, free-form text generation) but there are many other language models available.

Still, our results point to key differences in the abilities of different generations of language models, differences between evaluation metrics designed to measure different aspects of counterspeech and differences in their usefulness in predicting user preferences. We hope that our results will be useful for future researchers when choosing an approach to generate counterspeech, and when designing counterspeech evaluation systematically.

8 Ethics Statement

The present study, in which we asked human participants to rate automatically generated counterspeech, was approved by the research ethics committee of the University of Edinburgh School of Informatics with reference number 340484.

Looking to the future, our analysis shows that the automatic generation of counterspeech remains a challenging task, even for current large language models. The prospect of using automatically generated counterspeech to counter hate speech on social media raises important ethical questions. For example, if users who are spreading hateful content are presented with counterspeech, do they need to be informed that it was automatically generated? Sometimes, models generate inappropriate counterspeech. How do we deal with situations where users are presented with such inappropriate counterspeech responses, or with false positives, where users who were not actually spreading hateful language are accidentally presented with counterspeech? While we explore the automatic generation of counterspeech and obtain some promising results, any proposals to actually deploy our models would require careful thoughth and further testing.

References

Mana Ashida and Mamoru Komachi. 2022. Towards automatic generation of messages countering online hate speech and microaggressions. In *Proceedings of the Sixth Workshop on Online Abuse and Harms (WOAH)*, pages 11–23.

Sid Black, Leo Gao, Phil Wang, Connor Leahy, and Stella Biderman. 2022. Gpt-neo: Large scale autoregressive language modeling with mesh-tensorflow, 2021. *URL: https://doi. org/10.5281/zenodo*, 5297715.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.

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

Niklas Felix Cypris, Severin Engelmann, Julia Sasse, Jens Grossklags, and Anna Baumert. Intervening against online hate speech: A case for automated counterspeech.

Mai ElSherief, Caleb Ziems, David Muchlinski, Vaishnavi Anupindi, Jordyn Seybolt, Munmun De Choudhury, and Diyi Yang. 2021. Latent hatred: A benchmark for understanding implicit hate speech. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 345–363.

Kathleen C. Fraser, Svetlana Kiritchenko, Nejadgholi Isar, 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)*.

Joshua Garland, Keyan Ghazi-Zahedi, Jean-Gabriel Young, Laurent Hébert-Dufresne, and Mirta Galesic. 2022. Impact and dynamics of hate and counter speech online. *EPJ data science*, 11(1):3.

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*.

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.

Jae Yeon Kim, Jaeung Sim, and Daegon Cho. 2022. Identity and status: When counterspeech increases hate speech reporting and why. *Information Systems Frontiers*, pages 1–12.

Svetlana Kiritchenko, Isar Nejadgholi, and Kathleen C Fraser. 2021. Confronting abusive language online: A survey from the ethical and human rights perspective. *Journal of Artificial Intelligence Research*, 71:431–478.

Binny Mathew, Punyajoy Saha, Hardik Tharad, Subham Rajgaria, Prajwal Singhania, Suman Kalyan Maity, Pawan Goyal, and Animesh Mukherjee. 2019. Thou shalt not hate: Countering online hate speech. In *Proceedings of the international AAAI conference on web and social media*, volume 13, pages 369–380.

Karsten Müller and Carlo Schwarz. 2021. Fanning the flames of hate: Social media and hate crime. *Journal of the European Economic Association*, 19(4):2131–2167.

Kevin Munger. 2021. Don't@ me: Experimentally reducing partisan incivility on twitter. *Journal of Experimental Political Science*, 8(2):102–116.

Gaurav Nuti, Louis Penafiel, Janelle Ward, Abel Salinas, Fred Morstatter, Nathan Schurr, and R McCormack. 2022. Productive online discourse for emergency response. In *Workshop Proceedings of the 16th International AAAI Conference on Web and Social Media. Retrieved from https://doi. org/10.36190*.

Magdalena Obermaier, Desirée Schmuck, and Muniba Saleem. I'll be there for you? effects of islamophobic online hate speech and counter speech on muslim in-group bystanders' intention to intervene. *new media & society*, page 14614448211017527.

Raj Ratn Pranesh, Ambesh Shekhar, and Anish Kumar. 2021. Towards automatic online hate speech intervention generation using pretrained language model.

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.

Bharati Rathore. 2023. Future of ai & generation alpha: Chatgpt beyond boundaries. *Eduzone: International Peer Reviewed/Refereed Multidisciplinary Journal*, 12(1):63–68.

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*.

Carla Schieb and Mike Preuss. 2016. Governing hate speech by means of counterspeech on facebook. In *66th ica annual conference, at fukuoka, japan*, pages 1–23.

Serra Sinem Tekiroğlu, Helena Bonaldi, Margherita Fanton, and Marco Guerini. 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.

Zihao Wu, Lu Zhang, Chao Cao, Xiaowei Yu, Haixing Dai, Chong Ma, Zhengliang Liu, Lin Zhao, Gang Li, Wei Liu, et al. 2023. Exploring the trade-offs: Unified large language models vs local fine-tuned models for highly-specific radiology nli task. *arXiv preprint arXiv:2304.09138*.

Wanzheng Zhu and Suma Bhat. 2021. Generate, prune, select: A pipeline for counterspeech generation against online hate speech. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 134–149.