Towards Automatic Generation of Messages Countering Online Hate Speech and Microaggressions

Mana Ashida* and **Mamoru Komachi**
Tokyo Metropolitan University
maashida@yahoo-corp.jp komachi@tmu.ac.jp

Abstract

Warning: *This paper discusses and contains content that may be deemed offensive or upsetting.*

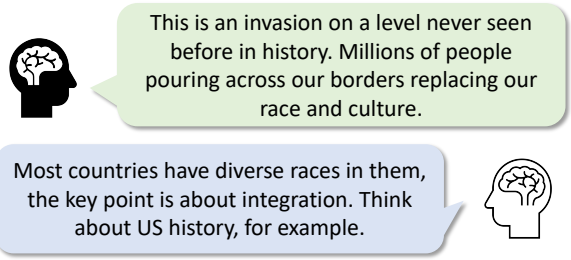
With the widespread use of social media, online hate is increasing, and microaggressions, unintentional offensive remarks in everyday life (Sue et al., 2007), are receiving attention. We explore the possibility of using pre-trained language models to automatically generate messages that combat the associated offensive texts. Specifically, we focus on using *prompting* to steer model generation as it requires less data and computation than *fine-tuning* and shows the potential for using *prompting* in the proposed generation task. We also propose a human evaluation perspective; offensiveness, stance, and informativeness. After obtaining 306 counter-speech and 42 micro intervention messages generated by GPT-2, textscGPT-Neo, and textscGPT-3, we conducted a human evaluation using Amazon Mechanical Turk and found that GPT-3 produces messages of the highest quality among three systems. Also, We discuss the pros and cons of using our evaluation perspectives. We release a corpus of **c**ountering **ha**te **s**peech and **m**icroaggressions (CHASM), annotated machine-generated counternarratives along with the annotation to promote further research on automatic counternarrative generation and its evaluation.

1 Introduction

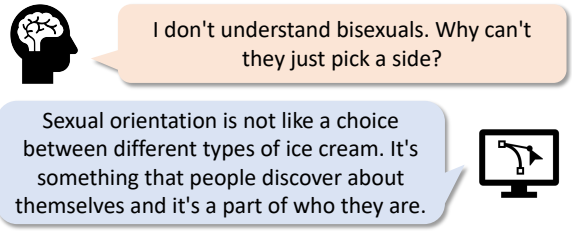
Concomitant with social media becoming a major means of communication, online abusive language is increasing. As abusive language can be harmful, countering it is an important way to reduce the level of danger on the Internet.

Hate speech is arguably the most well-studied form of abusive language across time and regions. It is defined by the United Nations Strategy and Plan of Action on Hate Speech[1] as "any kind of communication in speech, writing or behaviour, that attacks or uses pejorative or discriminatory language with reference to a person or a group on the basis of who they are, in other words, based on their religion, ethnicity, nationality, race, colour, descent, gender or other identity factor." Natural language processing (NLP) researchers have constructed several hate speech corpora, and some of them are publicly available (Madukwe et al., 2020).

Abusive language can be either explicitly offensive and harmful or implicitly offensive. Situations also exist where the offensiveness is executed in more subtly. One type of implicit offensive text is called "microaggression." **Microaggression** is a concept closely related to abusive language that



(a) A hate speech–counternarrative pair in Multitarget-CONAN (Fanton et al., 2021)



(b) An example of microaggressions in SELFMA (Breitfeller et al., 2019) and an intervention generated by GPT-3 davinci

Figure 1: Overview of the proposed message generation approach in action

*Currently working at Yahoo Japan Corporation, Tokyo, Japan.

[1] https://www.un.org/en/genocideprevention/hate-speech-strategy.shtml

Proceedings of the Sixth Workshop on Online Abuse and Harms (WOAH), pages 11 - 23

has been receiving increased attention recently. According to Sue et al. (2007), "microaggressions are brief and commonplace daily verbal, behavioral, or environmental indignities, whether intentional or unintentional, that communicate hostile, derogatory, or negative racial slights and insults." One characteristic of microaggression is invisibility; people exhibiting microaggressions are often unaware that they engage in such communications when they interact with targeted minorities. Despite its growing interests, research on microaggression in the field of NLP is quite limited.

Identifying hate speech or microaggressions, for example, "abusive language detection," is one technique for countering hate. However, detecting abusive language has some problems; simply flagging offensive content without providing a reason may result in the infringement of free speech. A better way to combat hate speech without infringing freedom of speech is to use a counternarrative.

Counterspeech or **counternarrative** is any message countering hate speech and offensive contents. The counternarrative has been studied as a means of confronting hate speech. Many researchers report that counternarratives are effective for reducing hate online (Hangartner et al., 2021). Several NGOs, such as Dangerous Speech Project[2], are working to promote counterspeech, and social networking platforms are also encouraging the use of counterspeech. Therefore, automatically generating counternarratives and thereby reducing the labor-intensiveness involved in countering online hate speech is an important application of NLP technology for social good.

Language generation by machines is becoming a viable option with the emergence of neural generative language models (LMs). The generation quality of the pretrained generative language models has increased to such an extent that humans cannot easily differentiate machine-generated text from texts written by a human (Clark et al., 2021). As such, we explore the automatic generation of counternarratives using LMs, namely Generative Pre-trained Transformers (GPTs). Conventionally, steering the generation process of a model relies on *fine-tuning*, which requires task-specific data that are not always easily obtainable. For this reason, recent studies have employed *prompting* as an alternative to *fine-tuning*. *Prompting* requires only a small number of examples of the task, and it does not require computation for optimizing the parameters of LMs. In this paper, we investigate the possibility of using a pretrained large-scale generative language model to generate counternarratives against hate speech and microaggressions using *prompting* instead of *fine-tuning*.

Whereas the traditional counternarrative generation task is primarily focused on countering hate speech, this study extends the target to microaggressions. Identifying microaggressions and understanding why they are offensive requires an understanding of the social context and the negative stereotypes that persist in the world. Consequently, countering microaggressions is more difficult than countering hate speech. To counter microaggressions, the concept of **microinterventions** has been proposed in recent years and is being studied from the perspective of psychology and sociology (Sue et al., 2019). However, we are the first to discuss the usefulness of NLP technology for this purpose.

In this study, we generated 696 counternarratives using LMs and evaluated their quality by conducting a human evaluation exercise on a crowdsourcing platform. In addition, for qualitative evaluation, we analyze some examples from the set of counternarratives and then discuss the issue related to the counternarrative generation task as well as its evaluation.

This study makes three main contributions:

1. We propose to include **microaggressions** as a target of counternarrative generation.

2. We design a *few-shot* prompt for generating counternarratives to assess the applicability of prompting for counternarrative generation using pretrained language models.

3. We propose an annotation scheme for machine-generated counternarratives evaluation and create a corpus of **c**ountering **ha**te **s**peech and **m**icroaggressions (CHASM), annotated machine-generated counternarratives along with the offensiveness score of the abusive language post.[3]

2 Related Work

Counterspeech Generation. Considering the positive effects of counternarratives, several NLP

[2] https://dangerousspeech.org/

[3] The corpus is accessible from https://github.com/tmu-nlp/CHASM.

studies have investigated the possibility of automatically generating counterspeech or using human-in-the-loop strategies to counteract hate and harmful speech online (Qian et al., 2019; Chung et al., 2019; Tekiroğlu et al., 2020; Fanton et al., 2021; Chung et al., 2021; Zhu and Bhat, 2021; Tekiroglu et al., 2022).

Qian et al. (2019) were the first to attempt automatic counternarrative generation. They created a resource of 10,243 counternarratives against 5,257 hate speech instances in 5,020 conversations containing 22,324 comments from Reddit and 31,487 counternarratives against 14,614 hate speech instances in 11,825 conversations containing 33,776 posts from Gab. They used crowdsourcing for obtaining counternarratives and used them to train neural models. Zhu and Bhat (2021) proposed a pipeline for generating counternarrative candidates using a recurrent neural network (RNN)-based generative model trained on this dataset, pruning only grammatical candidates, and selecting the most relevant candidate.

Chung et al. (2019) created a resource of counternarratives for Islamophobia—hate or fear against Islam and Muslims—written by expert operators from three NGOs. The CONAN dataset consists of 6,645 English hate speech–counternarrative pairs, including 2,781 translated pairs from French and Italian. Chung et al. (2021) used this dataset to fine-tune GPT-2 to automatically generate counternarratives. They also adopted the same methodology of data collection on hate speech targeting other religions, races, and gender to fine-tune GPT-2 for automated generation (Fanton et al., 2021). They reported data creation via the human-in-the-loop strategy of post-editing machine-generated counternarratives by expert operators from NGOs.

These strategies require substantial amounts of data as well as human resources. Although fine-tuning pretrained models rather than training neural models requires less data, a substantial amount of data is still necessary. Herein, we explore a method that requires only a few examples for generating counternarratives. This method is called *prompting*. *Prompting* has been receiving significant amounts of attention in recent years because of its effectiveness with only a few examples. Furthermore, it does not require the training of parameters for downstream tasks. This contrasts with fine-tuning of LMs, which requires the training of newly introduced parameters with different datasets, and thus more computation. *Prompting* has also reportedly achieved performance comparable with *fine-tuning*. Further details of prompting are presented in Sec. 4.2.

Microintervention Generation. Microaggression is a less well-known concept than hate speech, little research has been conducted regarding fighting against microaggressions. In the social sciences field, Sue et al. (2019) proposed the concept of "microinterventions" as a way to deal with everyday microaggressions. They state the following goals for microinterventions: (a) make the invisible visible, (b) disarm the microaggression, (c) educate the perpetrator, and (d) seek external reinforcement or support. Some of the core differences between the countering of hate speech and the countering of microaggression are (1) lack of recognition that a microaggression has occurred, and (2) harmful impact caused by good intent. However, no studies have been conducted in the NLP field on this subject.

Studies on the generation of microinterventions in NLP are rare. One of the closest is the work on anti-stereotype generation by Fraser et al. (2021). They investigated strategies to combat negative stereotypes using anti-stereotypes that help to deconstruct harmful beliefs, and proposed the anti-stereotype generation task. Further, they analyzed the kinds of stereotypes and showed that stereotypes are multidimensional and often ambivalent. Therefore, the anti-stereotypes can also be multidimensional, not just the antonym (e.g., the anti-stereotype for "caring nurse" is not "uncaring nurse" but "rude nurse"). They provided a few examples of anti-stereotypes that seem useful for countering stereotypes (e.g., "caring and mature mother" against "caring but childish mother") while mentioning the possibility that anti-stereotypes help us to look at others as individuals instead of stereotypical group representatives.

To the best of our knowledge, this is the first work that tackles the automatic generation of counternarratives against microaggressions and the evaluation of the machine-generated microintervention quality.

Counternarrative Evaluation. Evaluation of generated text is a bottleneck in the promotion of natural language generation tasks, especially for dialogue generation. The difficulty is that there are many acceptable responses when generating out-

put, and it is difficult to define what constitutes a good response. Therefore, the design of the evaluation scheme itself is also difficult. Whereas some relatively constrained generation tasks have established evaluation perspectives, such as "adequacy" and "fluency" for machine translation, evaluation perspectives for many other generation tasks have no common standard.

Similarly, evaluation methods for machine-generated counternarratives have not been established. Previous studies proposed various evaluation perspectives for human evaluation—such as suitableness, informativeness, intra-coherence (Chung et al., 2021), diversity, relevance, language quality (Zhu and Bhat, 2021), offensiveness, and stance (Baheti et al., 2021).

Diversity or **language quality** is designed to measure the generation ability of proposed models, and thus is not specifically designed for counternarrative generation. Because large pretrained models are known to generate fluent texts, we did not consider measuring general generation quality.

Alternatively, we adapt **offensiveness** and **stance** considering the characteristics of pretrained language models that previously found that they tend to agree with the previous comment during conversation (Baheti et al., 2021) and may generate abusive contents (Chung et al., 2021) in the counternarrative generation task. We also assume that **offensiveness** and **stance** can assess aspects that are measured by **relevance** or **suitableness** in previous studies (Chung et al., 2021; Zhu and Bhat, 2021).

Furthermore, we adapt **informativeness** from Chung et al. (2021) to reflect that counternarratives that are too generic are not considered effective. We can also presume that a system that often generates generic outputs cannot produce diverse contents. As such, we expected that **informativeness** could cover qualities that have been captured via **effectiveness** or **diversity** (Qian et al., 2019; Zhu and Bhat, 2021).

Most of the previous evaluation studies focused on comparing the generation quality of each model, and machine-generated counternarratives along with the evaluation have not been published. Baheti et al. (2021) provide the only available resource of human-written or machine-generated responses with annotations, but the original task does not involve assessing counternarrative quality but rather classifying the contextual toxicity of dialogue responses. In this study, we created a corpus of annotated machine-generated counternarratives along with the offensiveness score of the abusive language post. This allowed us to analyze counternarratives from multiple aspects. The details of the evaluation perspectives used in the experiments are presented in Sec. 4.4.

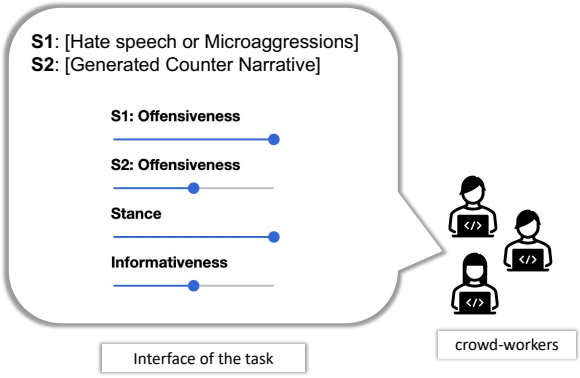


Figure 2: Interface of the annotation task

3 Counterspeech and Microintervention Generation

3.1 Task Formalization

Counternarrative generation can be viewed as a type of conditional or constrained text generation, in which the output is expected to oppose the input text. As the output is a response to the input, this task can also be considered dialogue generation with a single turn of conversation.

We formalize the counternarrative generation task following Zhu and Bhat (2021). Specifically, we assume access to a corpus of labeled pairs of conversations $D = (x_1, y_1), (x_2, y_2), ..., (x_n, y_n)$, where x_i is a hate speech or microaggression and y_i is the appropriate counternarrative as decided by experts or by crowdsourcing. The aim is to learn a model that takes as input a hate speech or a microaggression x and outputs a counternarrative y.

As output y, our goal is to produce a counternarrative that 1) is not offensive, 2) opposes the input hate speech, and 3) contains specific information on the corresponding offensive content. We propose the evaluation criteria along with the three features.

3.2 Evaluation

We conducted a human evaluation exercise to ascertain how effective and informative the counternar-

ratives are and to obtain a fine-grained quality assessment.

For the evaluation, we considered three dimensions: **offensiveness**, **stance**, and **informativeness**. These dimensions have been proposed in the literature regarding the counternarrative generation and dialogue generation, as explained in Sec. 2. Each perspective was measured on a five-point Likert scale for counternarratives. **Offensiveness** of input was also annotated to examine how humans' perception of the offensive input differs.

Offensiveness deals with whether the sentence is offensive to anyone, such as people of a certain race, including the individuals who wrote the offensive post. Certainly, counternarratives should not include text offending other people. Also, attacking the authors themselves rather than their behavior is undesirable. Attacking the person is called *ad-hominem* (Habernal et al., 2018; Sheng et al., 2021), and is a fallacy that often occurs during conversation on the Internet. Although attacking the author of the post can be considered a countermeasure of hate speech, it cannot be regarded as a good counternarrative. The labels are presented as 0 (not sure), 1 (not offensive), 2 (maybe safe), 3 (maybe offensive), and 4 (completely offensive).

Stance (of a post) is classified into three types: agreeing, neutral, and disagreeing. A counternarrative is required to oppose the original statement; therefore, we assume that outputs that are neutral or agree with the offensive statement are not good counternarratives. Prepared labels are as follows: 0 (irrelevant), 1 (clearly agreeing), 2 (weakly agreeing), 3 (fighting but partially agreeing), and 4 (clearly fighting).

Informativeness assesses how informative and specific the counternarrative is, while not being generic. This perspective was designed as a counternarrative evaluation perspective by Chung et al. (2021). Their annotation guideline presented examples against Islamophobic hate speech: "Do you really believe that they are a problem?" received the lowest score, and "Muslims should not be forced to assimilate, since it is not right and no one wants that. And polygamy is illegal and forbidden in UK and Muslims actually respect this ban." received the highest score. We set five labels, ranging from 0 to 4, with 0 (irrelevant), 1 (not informative), 2 (generic statement and little information), 3 (relatively specific but little information), and 4 (specific and informative).

It is important to note that **informativeness** does not ask if the information is true or not. We do not explicitly ask for consulting external sources to verify if the information generated by systems is true. We discuss the issue related to this setting in Section 5.3.

Further details about the experimental settings of human evaluation will be described in Sec. 4.4.

4 Experimental Settings

4.1 Models

As for the models, we used Generative Pre-trained Transformers (**GPT**), which is an auto-regressive language model. For a given corpus $U = \{u_1, u_2, \ldots, u_n\}$ of size n, GPT is trained to maximize objective (1) where k is the size of the context window.

$$L_1(U) = \sum_i \log P(u_i | u_{i-k}, \ldots, u_{i-1}; \Theta) \quad (1)$$

The conditional probability P that the token u_i appears in the context given the tokens $u_{i-k}, \ldots, u_{i-1}$ is modeled using a neural network with parameter θ.

In this study, we examined the GPT-Neo (Black et al., 2021) model and the GPT-2 model released from Huggingface (Radford et al., 2019) as well as the GPT-3 model released from OpenAI (Brown et al., 2020). We opted for using the biggest parameter size models for each GPT, for it is reported that the bigger parameters yield the better model performance as for neural models, which is called "scaling laws for language models (Kaplan et al., 2020)"; we used GPT-2 of 1.5 B parameters trained on WebText (Radford et al., 2019), GPT-Neo of 2.5 B parameters trained on the Pile (Gao et al., 2020), and GPT-3 text-davinci-001[4] trained on CommonCrawl[5].

4.2 Methods

Sampling Parameters. We tested several sampling parameters for GPT-Neo and GPT-2 using the parameters documented in Huggingface Transformers' generation function with a fixed seed. We applied either greedy search or nucleus sampling (Holtzman et al., 2020) (with a top-p in {0.5,

[4] https://beta.openai.com/docs/engines/gpt-3

[5] https://commoncrawl.org/

0.95, 1.0}). We compared the four outputs to 50 randomly sampled inputs of GPT-Neo and GPT-2, respectively in terms of overall suitableness as a counterspeech, and chose the best parameter setting as greedy search. The results showed that applying nucleus sampling increases fluency and output length, but the generations contain more hallucinations and often agree to the offensive post than when no sampling was applied. One of those examples is shown in Table 4 in the Appendix. The result of annotation is also included in our dataset.

See Appendix B for further details of parameter settings.

Prompt Design. Prompts are mostly designed according to the target downstream tasks, and the design of the prompts is largely divided into three methods: *zero-shot*, *one-shot*, and *few-shot*. In a *few-shot* learning setting, the number of examples is more than one. When using zero example (only description of the task) and one example for prompts, they are called *zero-shot* and *one-shot*, respectively.

We considered *one-shot* prompt in the form of a chat-bot prompt and *few-shot* using multiple examples. The *one-shot* chat-bot prompt was obtained from presets available in OpenAI.[6] The *few-shot* prompt was created using the counterspeech in the CONAN-KN dataset (Chung et al., 2021) because they are the latest counterspeech dataset generated by experts.

Among all the pairs, an offensive post–counterspeech pair was randomly selected from each of the following five categories: Anti-semitism, Homophobia, Islamophobia, Misogyny, and Racism. The actual prompt used in our experiment is shown in Table 5 in the Appendix.

As we observed that GPT-Neo and GPT-2 did not generate messages of high quality with *one-shot* prompt, we focused on using *few-shot*. However, note that GPT-3 produced some meaningful outputs, as shown in Table 6 in the Appendix; future work could analyze the differences between the use of two prompts.

4.3 Source Datasets

We used the CONAN (Chung et al., 2019), Multitarget-CONAN (Fanton et al., 2021), and Knowledge-grounded hate countering (Chung et al., 2021) datasets for hate speech inputs. For microaggression inputs, we used the Social Bias Inference Corpus (SBIC). The SBIC contains various degrees of offensive content collected from different websites. Because our interest is in microaggressions rather than directly offensive hate speech, we chose the category of "microaggression" from the dev set, which is based on the SELFMA dataset originally curated by Breitfeller et al. (2019). Further details of the chosen input texts used in the experiment are presented in Appendix A.

4.4 Evaluation

The evaluation was conducted via workers recruited through Amazon Mechanical Turk. All of the three perspectives (**offensiveness**, **stance**, **informativeness**) were evaluated using a five-point Likert scale. Each pair was evaluated by three crowd workers. We informed the workers about the risks of being exposed to offensive texts and asked for discretion. The instruction and examples presented to the workers are shown in Fig. 6.

Quality Control Recruitment was limited to those with a HIT approval rate of more than 98%, the number of approved HITs (the unit of task on Amazon Mechanical Turk) was more than 5,000. All workers were residents of the United States to ensure quality.[7] We also prepared our original qualification which can be easily answered by reading instructions. Only those who passed the additional qualification participated in our HITs.

Worker Payment We paid $2.7 per 25 sentence pair estimating 15 – 20 mins for completing. This adds up to an hourly wage of $8.4 – $11.2, which is above the federal minimum wage. Labels were obtained from three people for each pair. We collected data for 1020 sentence pairs for a total of about $400.

5 Results and Analysis

5.1 Annotation Statistics

Fig. 3 shows the annotation of offensiveness of **input** offensive text, categorized by dataset. Most of the CONAN texts are labeled as 4 (i.e., most offensive), whereas the SBIC texts have lower scores. This difference is possibly due to the characteristics of microaggressions described earlier; i.e., subtle and often unconscious discriminatory remarks.

[6] https://beta.openai.com/examples/default-chat

[7] However, lowering threshold is recommended considering unfair *qualification labour* to get qualified (Kummerfeld, 2021).

	CONAN			SBIC		
	off.	st.	inf.	off.	st.	inf.
GPT-2	.28	.38	.36	.25	.11	.39
GPT-Neo	.38	.32	.33	.38	.20	.31
GPT-3	.72	.76	.53	.77	.57	.42

Table 1: Inter-annotator agreement (Krippendorff's α). off., st., and inf. denote offensiveness, stance, informativeness, respectively.

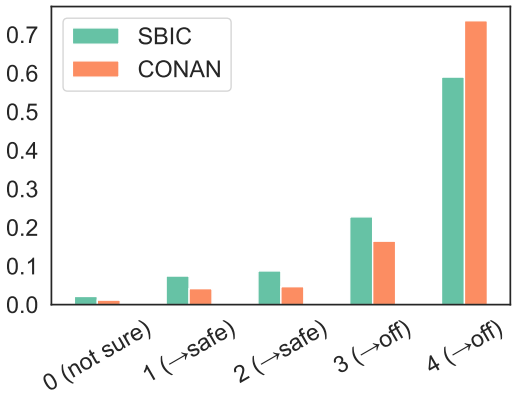


Figure 3: Offensive label distribution of input text per dataset.

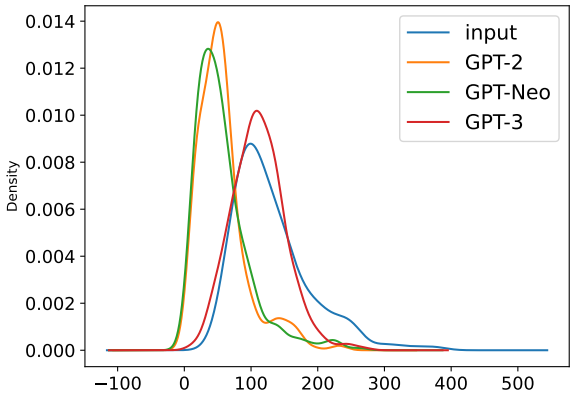


Figure 4: Kernel density estimation of the probability distribution of number of words per text for input and generated text by GPT-{2, Neo, 3}.

	DIST-1	DIST-2
GPT-2	.438	.737
GPT-Neo	.405	.681
GPT-3	**.495**	**.910**

Table 2: DIST-1 and DIST-2 of a set of generated texts by GPT-{2, Neo, 3}.

However, CONAN also receives low scores for some sentences. In these cases, the annotator's belief (Sap et al., 2021) may have influenced their judgment. For example, if one believes that migrants are a threat, it is likely for them to consider discriminatory texts about migrants as not offensive. Additionally, lack of context such as whom the author is addressing affects the certainty as to whether the texts are offensive.

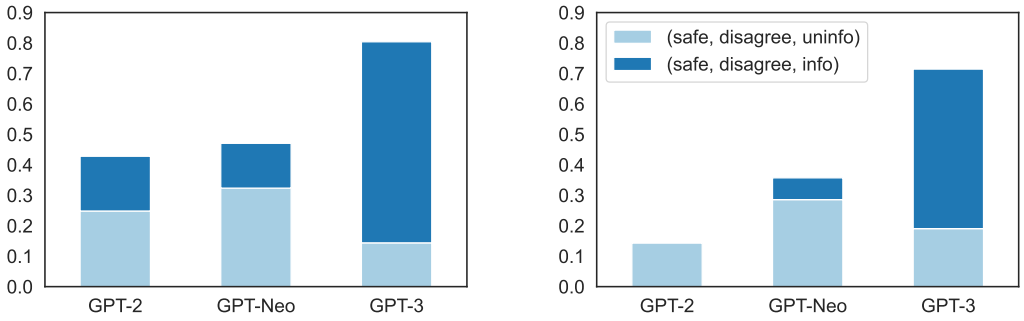
We report Krippendorff's α for each dataset per system in Table 1.[8] The values are comparable to previous studies dealing with relative subjectivity, such as $\alpha = 0.32$ for offensiveness and $\alpha = 0.18$ for stance of machine-generated responses reported in Baheti et al. (2021) and $\alpha = 0.51$ for offensiveness of human-written texts reported in Sap et al. (2020). Among the three systems, GPT-3 holds the higher scores for the offensive category. The higher agreement suggests that the quality of the output is more similar to the human-generated outputs, as it has been reported that machine-generated texts' agreement on offensiveness is lower than that of human-generated outputs (Baheti et al., 2021). The overall agreement reduction of SBIC compared to CONAN reflects that the generation quality is worse, and the task is more challenging. The label distribution of each perspective for CONAN is shown in Fig. 7 and for SBIC in Fig. 8 in the Appendix.

5.2 Quantitative Analysis

Generation Length. Fig. 4 shows the density distribution of the number of words for machine-generated texts. The distribution of GPT-3 corresponds to that of input texts written by a human, whereas that of GPT-2 and GPT-Neo shows that the output text lengths are much shorter. This suggests the performance of GPT-3 is the most similar to human among the three GPTs. Also, the GPT-3's output length is independent of the input length as the correlation between the number of words of each input and GPT-3 output is weak (Pearson's r of 0.29).

Generation Diversity. We report DIST (Li et al., 2016) over the outputs of three systems (Table 2). DIST calculates the percentage of different n-grams among the n-grams in all the raw sentences. DIST-1 and DIST-2 measure the proportion of different unigrams and different bigrams, respectively.

[8] To calculate α, we converted the labels of each perspective as follows: 1 and 2 of **offensiveness** into **safe**, and 3 and 4 into **offensive**; 1 and 2 of **stance** into **agree** and 3 and 4 into **disagree**; 1 and 2 of **informativeness** into **informative** and 3 and 4 into **uninformative**.



(a) CONAN (b) SBIC

Figure 5: Ratio of {safe, disagree, uninformative} to {safe, disagree, informative}

They are automatic measures of the diversity of the generated sentences. This results suggests that GPT-3's generation is the most diverse among the three systems.

Most outputs presents facts to counter hate-speech or microaggressions. According to (Benesch et al., 2016), the types of counternarratives are multiple, such as *warning of offline or online consequences* or using *humor*. Generating different types of counternarratives as well as how to evaluate the effectiveness of different types of counternarratives are left for future work.

Generation Quality. We hereafter proceed with the analysis based on the counternarratives annotated (safe, disagree, informative) and (safe, disagree, uninformative), as we consider the former to be valid counternarratives, and the latter to be acceptable counternarratives. The result will focus on how many (safe, disagree) counternarratives are generated by each system.

Fig. 5 shows the ratio of countering messages received (safe, disagree, uninfo) to (safe, disagree, info) against CONAN's hate speech and SBIC's microaggressions. For both cases, GPT-3 performs better, followed by GPT-Neo. This can be attributed to model size: GPT-3 is the largest, and GPT-Neo is the second-largest system among the three. In CONAN, more than 14.7% (for GPT-Neo) of the generated responses receive (safe, disagree, informative). This result suggests that all the systems hold the potential to generate valid counternarratives. In SBIC, the overall score falls compared to that of CONAN, and more than half of the responses of GPT-2 and GPT-Neo are invalid as counternarratives. This can be explained by the offensiveness annotation of the input sentences that vary by individual, as illustrated in Fig. 3. This result implies that subjective offensiveness is more difficult for language models to recognize and counter.

5.3 Qualitative Analysis

The examples in Table 3a show the generated messages that counter hate speech effectively. This confirms that it is possible to generate counternarratives via *prompting*.

However, the examples in Table 3b show that some of the generated counternarratives need to be post-edited even when judged as (safe, disagree, info.). As for GPT-3, the output uses the phrases "no evidence" and "in fact." Actually, the repeated use of these phrases is found in generated texts in the corpus. However, claiming "no evidence" and "In fact" can easily lead to misinformation or hallucination as it is almost impossible to guarantee that there is no evidence on some topics. Because it is difficult for non-expert annotators to distinguish fact from misinformation, and harmless hallucination from problematic hallucination, additional fact-checking procedures would be desirable to solve this problem. Moreover, both of the generated texts argue the legitimacy of education in the isolation of children with disabilities, instead of stating the importance of diversity and inclusion in society. This is also a point that needs to be improved.

Furthermore, the issue is not only limited to the quality of generation, but also the human evaluation of the generated text. For example, in Table 3c, the majority of annotators assigned "safe." In contrast, in SBIC, it is judged to be offensive, and the annotated harmful implications include "women are responsible for their assaults," and "women shouldn't go out after dark alone." The reason the annotators assigned "agree" may be because the first half of

Post	Muslim invasion is the new plague of London. They must be repatriated to their home countries!	(offensive)
Neo:	No, they are not. They are here to work and to live in peace.	(safe, disagree, info.)
GPT-3:	So you want people to feel like second class citizens, they're only invited here to work for us and must stay silent when discriminated.	(safe, disagree, info.)

(a) Example of generated counterspeech

Post	Children with Down Syndrome should be isolated, they cannot cope with education in normal schools.	(offensive)
GPT-2:	Down syndrome is a genetic condition, not a disability.	(safe, disagree, info.)
GPT-3:	There is no evidence that children with Down Syndrome cannot cope with education in normal schools. In fact, many of them excel in these settings.	(safe, disagree, info.)

(b) Example of generated counterspeech including hallucinated contents

Post	It's terrible! And it's not like she was doing something wrong, like walking around late at night.	(56% safe, 33% offensive), offensive in SBIC
GPT-3:	It's terrible that someone was attacked and it's not right that people are blaming the victim.	(safe, agree, informative)

(c) Example of generated microintervention

Table 3: Examples of generated messages against hate speech and microaggressions

the text is identical to the input, although the latter part of GPT-3's message shows the understanding of harmful implications and countering to it. This example highlights the difficulty of the task even for humans to reach a consensus and the need for additional quality control.

6 Conclusion

This paper explored the possibility of using pretrained language models on the counternarrative generation task against hate speech and harmful social implications. We used three LMs to generate counternarratives via prompting and conducted a human evaluation exercise to ascertain the quality of the generated counternarratives using "offensiveness," "stance," and "informativeness" as our evaluation criteria. Furthermore, we analyzed the models' performance based on how many generated counternarratives are safe, informative, and opposing to abusive language input.

The overall results show the potential of language models to generate controlled content using prompting, which requires only some examples of inputs and desired outputs, compared to fine-tuning, which is computation intensive. Among the three LMs we tested, GPT-3 performed the best in terms of generating safe, informative counternarratives that oppose abusive language input. However, some of the counternarratives considered informative contained misinformation or hallucinated contents. Applying a fact-checking process to the generated contents is a possible future task.

Ethical Considerations

Our study was conducted with the approval of the Internal Review Board. We informed workers about the risk of being exposed to the hate content through the HIT title visible to workers before accepting the HIT on Amazon Mechanical Turk. The paper's theme is important as online hate speech and microaggressions continue to increase; therefore, there is a need for combating hate automatically. We hope that our corpus encourages further studies on this topic. We acknowledge the limitations that the corpus is only in English and that the hate speech contents are not fully up-to-date, such as dealing with the increasing amounts of hate speech against Asians due to the COVID pandemic.

Acknowledgements

We gratefully acknowledge the support of LINE Corporation to conduct this research. This work was partially supported by JSPS KAKENHI Grant Number 22H03651. We would like to thank anonymous reviewers and those who gave feedback on earlier versions of this paper.

References

Ashutosh Baheti, Maarten Sap, Alan Ritter, and Mark Riedl. 2021. Just say no: Analyzing the stance of neural dialogue generation in offensive contexts. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 4846–4862, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Susan Benesch, Derek Ruths, Kelly P Dillon, Haji Mohammad Saleem, and Lucas Wright. 2016. Counterspeech on twitter: A field study. *A report for Public Safety Canada under the Kanishka Project*.

Sid Black, Gao Leo, Phil Wang, Connor Leahy, and Stella Biderman. 2021. GPT-Neo: Large Scale Autoregressive Language Modeling with Mesh-Tensorflow. Zenodo.

Luke Breitfeller, Emily Ahn, David Jurgens, and Yulia Tsvetkov. 2019. Finding microaggressions in the wild: A case for locating elusive phenomena in social media posts. 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 1664–1674, Hong Kong, China. Association for Computational Linguistics.

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. In *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901. Curran Associates, Inc.

Yi-Ling Chung, Elizaveta Kuzmenko, Serra Sinem Tekiroglu, 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, Florence, Italy. Association for Computational Linguistics.

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, Online. Association for Computational Linguistics.

Elizabeth Clark, Tal August, Sofia Serrano, Nikita Haduong, Suchin Gururangan, and Noah A. Smith. 2021. All that's 'human' is not gold: Evaluating human evaluation of generated text. 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 7282–7296, Online. Association for Computational Linguistics.

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, Online. Association for Computational Linguistics.

Kathleen C. Fraser, Isar Nejadgholi, and Svetlana Kiritchenko. 2021. Understanding and countering stereotypes: A computational approach to the stereotype content model. 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 600–616, Online. Association for Computational Linguistics.

Leo Gao, Stella Biderman, Sid Black, Laurence Golding, Travis Hoppe, Charles Foster, Jason Phang, Horace He, Anish Thite, Noa Nabeshima, Shawn Presser, and Connor Leahy. 2020. The Pile: An 800gb dataset of diverse text for language modeling. *arXiv preprint arXiv:2101.00027*.

Ivan Habernal, Henning Wachsmuth, Iryna Gurevych, and Benno Stein. 2018. Before name-calling: Dynamics and triggers of ad hominem fallacies in web argumentation. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 386–396, New Orleans, Louisiana. Association for Computational Linguistics.

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).

Ari Holtzman, Jan Buys, Li Du, Maxwell Forbes, and Yejin Choi. 2020. The curious case of neural text degeneration. In *International Conference on Learning Representations*.

Jared Kaplan, Sam McCandlish, T. J. Henighan, Tom B. Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeff Wu, and Dario Amodei. 2020. Scaling laws for neural language models. *ArXiv*, abs/2001.08361.

Jonathan K. Kummerfeld. 2021. Quantifying and avoiding unfair qualification labour in crowdsourcing. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 343–349, Online. Association for Computational Linguistics.

Jiwei Li, Michel Galley, Chris Brockett, Jianfeng Gao, and Bill Dolan. 2016. A diversity-promoting objective function for neural conversation models. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 110–119, San Diego, California. Association for Computational Linguistics.

Kosisochukwu Madukwe, Xiaoying Gao, and Bing Xue. 2020. In data we trust: A critical analysis of hate speech detection datasets. In *Proceedings of the Fourth Workshop on Online Abuse and Harms*, pages 150–161, Online. Association for Computational Linguistics.

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, Hong Kong, China. Association for Computational Linguistics.

Alec Radford, Jeff Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners.

Maarten Sap, Saadia Gabriel, Lianhui Qin, Dan Jurafsky, Noah A. Smith, and Yejin Choi. 2020. Social bias frames: Reasoning about social and power implications of language. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5477–5490, Online. Association for Computational Linguistics.

Maarten Sap, Swabha Swayamdipta, Laura Vianna, Xuhui Zhou, Yejin Choi, and Noah A. Smith. 2021. Annotators with attitudes: How annotator beliefs and identities bias toxic language detection. *ArXiv*, abs/2111.07997.

Emily Sheng, Kai-Wei Chang, Prem Natarajan, and Nanyun Peng. 2021. "nice try, kiddo": Investigating ad hominems in dialogue responses. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 750–767, Online. Association for Computational Linguistics.

Derald Wing Sue, Sarah Alsaidi, Michael N. Awad, Elizabeth Glaeser, Cassandra Z Calle, and Narolyn Mendez. 2019. Disarming racial microaggressions: Microintervention strategies for targets, white allies, and bystanders. *The American psychologist*, 74 1:128–142.

Derald Wing Sue, Christina M. Capodilupo, Gina C. Torino, Jennifer M Bucceri, Aisha M. B. Holder, Kevin L. Nadal, and Marta Elena Esquilin. 2007. Racial microaggressions in everyday life: implications for clinical practice. *The American psychologist*, 62 4:271–86.

Serra Sinem Tekiroglu, Helena Bonaldi, Margherita Fanton, and Marco Guerini. 2022. Using pre-trained language models for producing counter narratives against hate speech: a comparative study.

Serra Sinem Tekiroğlu, Yi-Ling Chung, and Marco Guerini. 2020. Generating counter narratives against online hate speech: Data and strategies. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1177–1190, Online. Association for Computational Linguistics.

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, Online. Association for Computational Linguistics.

A Data Preprocessing

Hatespeech. We gather all the English hate speech in CONAN, the seed dataset in Multitarget-CONAN, and the Knowledge-grounded hate countering dataset, then select those consist of more than 10 words and less than 100 words. After excluding overlapped sentences and the texts containing #, and , we obtain 306 hate speech.

Microaggressions. We select the post consisting of more than 10 words and less than 100 words and retained all the posts which have free-text implications. We exclude the texts containing #, and ::. In this way, we select a total of 42 microaggression statements.

B Reproducibility

For GPT-3 text-davinci-001 model, we use temperature of 0.7, max tokens of 50, top_p of 1, frequency penalty of 0, and presence penalty of 0. For GPT-{2, Neo}, we use temperature of 1.0, max tokens length of 1024, top_k of 50.

Task description

Thank you for your participation!
You will be presented with a pair of sentences; offensive post (S1) and the a counter narrative (counter-speech) (S2) to the post generated by a bot.
Your task: evaluate the quality of the counter narrative from the three persepectives; **Offensiveness**, **Stance** and **Informativeness**

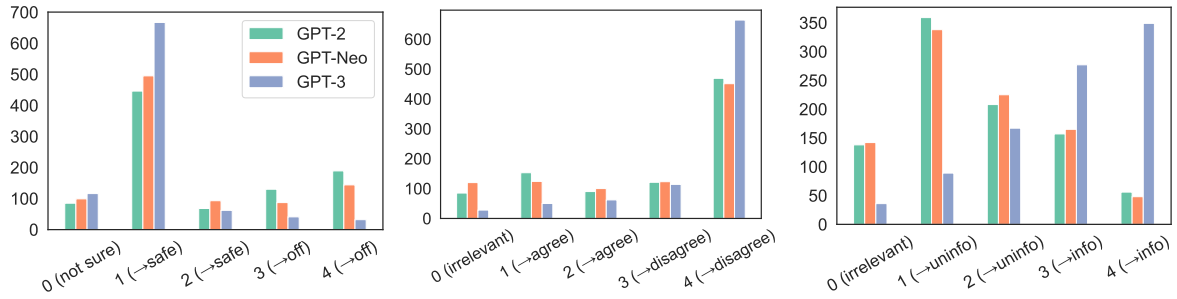
CAUTION:
The sentences presented in the task exibit overt Sexism, Racism, Xenophobia, Transphobia, Homophobia etc. Worker discretion is advised.

Offensiveness (applied to S1, S2)
Are the sentences 1 and 2 offensive to anyone, such as people of a certain race, gender or religion? "Anyone" also includes individuals such as the person who wrote S1. So if S2 looks like it is attacking the person who wrote the S1, rate them as offensive. When you don't understand the meaning of the sentence, rank it as 0 (not sure).

Stance (applied to S2 only)
Is the counter narrative successfully disagreeing against the given offensive post or, conforming to the post.
ATTENTION: some responses use sarcasm and rhetoric to express disagreement indirectly; in that case, you are supposed to label them as either 3 or 4.

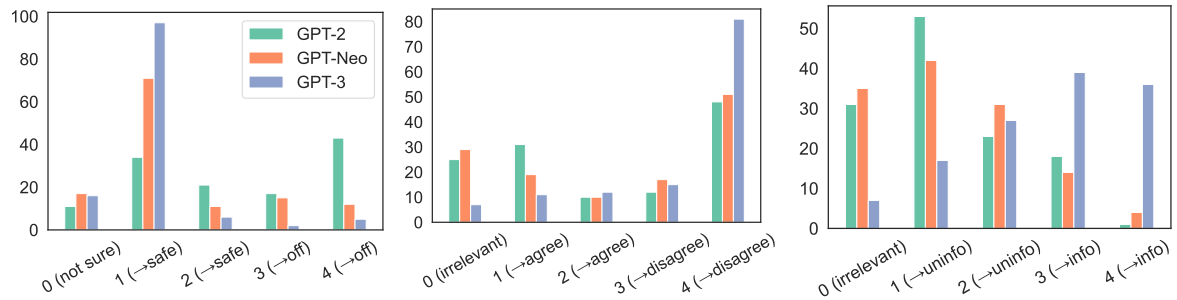
Informativeness (applied to S2 only)
How informative and specific is the counter narrative? If the counter narratives contain completely irrelevant information, mark them as 0.
If they are somewhat related, whether they agree or disagree to S1, rank them according to how much information they hold and how specific they are.

Figure 6: The instructions given to crowd-workers on Amazon Mechanical Turk



(a) offensiveness ↓ (b) stance ↑ (c) informativeness ↑

Figure 7: label distribution of CONAN



(a) offensiveness ↓ (b) stance ↑ (c) informativeness ↑

Figure 8: label distribution of SBIC

input	The overcrowd at migrant detention centers for illegals is also caused by the liberal left. Illegal immigrants have no right to be in your country and they should be taken back.
top-p=1.0	The law should be used, which has always been implemented by the American government to deport illegal immigrants who broke the law.
top-p=0.95	The liberals are also responsible for the illegal immigrants problem in the first place by not controlling the flow of illegal aliens.
top-p=0.50	Illegal immigrants are only here because of the liberal left. They should be deported and not allowed to live in your country.
greedy	You are a racist.

Table 4: Example of GPT-2's generation with different sampling parameters

Generate counterspeech to the given offensive post. Post: Jews are born greedy. Counterspeech: Like any ethno-racial or religious group, the population of Jews is internally diverse, so it is impossible to generalize. Post: Why gays do not try to heal? Counterspeech: Contrary to some insinuations, sexual orientation is not like a disease. Post: We need to Burn the Koran and all the Mosques down - this will solve all problems in this country. Counterspeech: NO, we need to close Radical mosques and to expel radical Muslims. That's all. Peaceful Muslims have the right to live in peace in our country. Post: I'm tired of woman taking part to political debate. Feminism is a divisive force in society. Counterspeech: How many women are represented in political conversations on television? Post: This is an invasion: we must stop immigrants because they are not accepted by our people. Counterspeech: So you want people to feel like second class citizens, they're only invited here to work for us and must stay silent when discriminated.

Table 5: *Few-shot* prompt used in the experiment

Prompt	The following is a conversation with an AI assistant. The assistant is helpful, creative, clever, and very friendly. Human: Hello, who are you? AI: I am an AI. How can I help you today? Human: Hey, I don't understand bisexuals. Why can't they just pick a side?
GPT-2	AI: ____
GPT-Neo	AI:????
GPT-3	AI: Bisexuality is a term used to describe someone who is attracted to people of both genders. Some people believe that bisexuals can't just pick a side, because they are attracted to both genders. Others believe that bisexuality is simply a phase, and that people will eventually settle down with one gender. Ultimately, it is up to the individual to decide what they believe.

Table 6: Example of chat-bot prompt and generated texts