Outcome-Constrained Large Language Models for Countering Hate Speech

Lingzi Hong[1] Pengcheng Luo[2]
lingzi.hong@unt.edu luopc@pku.edu.cn

Eduardo Blanco[3] Xiaoying Song[1]
eduardoblanco@arizona.edu XiaoyingSong@my.unt.edu

[1]University of North Texas [2]Peking University [3]University of Arizona

February 8, 2026

Abstract

Automatic counterspeech generation methods have been developed to assist efforts in combating hate speech. Existing research focuses on generating counterspeech with linguistic attributes such as being polite, informative, and intent-driven. However, the real impact of counterspeech in online environments has not been considered. This study aims to develop methods for generating counterspeech constrained by conversation outcomes and evaluate their effectiveness. We experiment with large language models (LLMs) to incorporate into the text generation process two desired conversation outcomes: low conversation incivility and non-hateful hater reentry. Specifically, we experiment with instruction prompts, LLM finetuning, and LLM reinforcement learning (RL). Evaluation results show that our methods effectively steer the generation of counterspeech towards the desired outcomes. Our analyses, however, show that there are differences in the quality and style depending on the model.

1 Introduction

Hate speech has posed significant challenges to healthy and productive online communication. Counterspeech, which involves using constructive, positive, or factual responses to challenge or counteract hate speech, has shown to be effective in moderating online hostilities (**?**), promoting productive user engagement (**?**), and educating online users (**?**).

Automatic generation of counterspeech has been researched to support timely and effective efforts to fight hate speech. Synthetic counterspeech datasets have been developed using crowdsourcing (**?**) and human-in-the-loop strategies (**?**). These datasets have been used to develop counterspeech generation models. However, the impact of counterspeech in online environments has not been considered in the dataset creation. As a result, it is unknown whether generated counterspeech elicits civil or hateful follow-up conversations.

Recent counterspeech generation research focused on constrained generation with linguistic attributes (e.g., being polite, emotion-laden (**?**)), or embedded with knowledge (**?**). Questions about the impact of counterspeech with such attributes linger. Previous research also found one of the barriers counterspeakers face is their inability to determine the potential impact of counterspeech (**?**). However, there is a lack of research on generating outcome-oriented counterspeech, e.g., speech that leads to desired outcomes such as de-escalating user conflicts or encouraging constructive engagement in follow-up conversations.

Notably, previous studies indicate that language may influence the development of a conversation, including discourse popularity (**?**), reentry behaviors (**?**), and the rise of hate speech (**?**). This leads to our research questions:

- How can constraints on conversation outcomes be incorporated into developing LLMs for generating counterspeech?
- How effective are these methods in generating outcome-oriented counterspeech?

Unlike previous work that considers explicit linguistic attributes to guide language generation, we formulate counterspeech generation to achieve desired outcomes (e.g., constructive user engagement). Our study holds potential for broader applications. Anticipating the direction of a conversation is crucial in crafting effective

responses, allowing the conversation to meet the objectives (e.g., reducing hate speech, altering user behavior, and promoting positive discourse). This study makes the following contributions: (i) introducing conversation outcomes as a constraint to guide the generation of counterspeech, (ii) experimenting with LLMs for generating outcome-constrained counterspeech using instruction prompts, LLM finetuning, and LLM reinforcement learning (RL), and (iii) evaluating counterspeech generation models with various metrics to understand the strengths and weaknesses of the methods.

2 Related Work

2.1 Generating Counterspeech

Table 1 presents recent work on counterspeech generation. CONAN has counterspeech written by NGO experts and augmented by language models (**?**); Benchmark was built with hate speech from Gab and Reddit and counterspeech created by crowdsourcing workers (**?**); and MultiCONAN is a high-quality, high-quantity dataset created by experts coupled with language model generation for hate speech with multiple targets (**?**). Counterspeech generation models have been built with these datasets (**????**). Unlike us, none consider conversation outcomes elicited by the generated counterspeech.

Researchers have investigated counterspeech generation under constraints. **?** proposed a generation pipeline grounded in external knowledge repositories to generate more informative and less biased replies. **?** proposed to generate more diverse and relevant counterspeech by developing a three-stage pipeline that uses LLMs to generate candidates, prunes the ungrammatical ones, and selects the best instances. **?** proposed an ensemble generative discriminator to generate more polite, detoxified, and emotion-laden counterspeech. **?** developed Intent-CONAN, where the generation of counterspeech is conditioned on five intents: informative, denouncing, questioning, positive, and humorous. Similarly, **?** utilized ChatGPT to generate counter-stereotype text by incorporating countering strategies in queries. **?** proposed prompting strategies based on discourse theories to generate more context-relevant counterspeech. There are also studies on the generation of counterspeech in languages other than English (e.g., Italian (**?**)). Unlike us, none of these previous works generate counterspeech to elicit positive behaviors in the follow-up conversations.

2.2 Language Generation with Constraints

Extensive studies have targeted language generation under complex lexical constraints such as formality (**?**), text with certain concepts (**?**), dialogue that takes latent variables (**?**), and knowledge-enhanced text (**?**). Not all styles can be described explicitly as linguistic attributes. Indeed, some 'styles' can only be defined in a data-driven way based on the shared attributes across various datasets (**?**). In this study, we generate counterspeech very likely to lead to desired conversational outcomes.

Methods have been developed for constrained language generation. **?** proposed the SentiGAN framework to generate text with a given sentiment. **?** proposed MUCOCO to allow for controllable inference with multiple attributes as constraints to the optimization. **?** developed GeDi, a discriminator-based approach to guide the decoding process in language generation. It enables text generation with desired or undesired attributes. **?** proposed a self-debiasing approach to reduce the probability of language models generating problematic text. Unlike these previous efforts, we experiment with methods to adjust language model-generated texts to achieve specific conversational outcomes.

3 Methodology

3.1 Conversation Outcomes

Conversation outcomes refer to the result of a message in a conversation, which can be measured by the manner and characteristics of the follow-up conversations it elicits. According to previous studies, a combination of hate speech and its reply—regardless of whether it counters the hateful comment—can predict future conversation engagement and incivility (**??**). This study explores two types of conversation outcome modeling: conversation incivility and hater reentry (Figure 1). Based on the modeling results, we build conversation outcome classifiers that use hate speech and counterspeech to predict the incivility level or hater reentry type.

Conversation Incivility Conversation incivility is a metric to measure the outcome based on the number of civil and uncivil comments as well as the unique authors involved in the discourse (**?**). Intuitively, the more uncivil (or less civil) the comments, the worse the outcome; uncivil comments from many authors are worse than those from just a few. Formally, it is defined as:

$$S(r) = \alpha U(r) - (1-\alpha)C(r) \tag{1}$$

where $U(r)$ refers to uncivil behavior and $C(r)$ to civil behavior. For each user i ($i = 0, 1, 2, \ldots, k$), n_i^u is defined as the number of uncivil comments by user i, and n_i^c as the number of civil comments. Then,

$$U(r) = \sum_{i=0}^{k} n_i^u \quad \text{and} \quad C(r) = \sum_{i=0}^{k} \frac{n_i^c}{n_i^u + n_i^c} \tag{2}$$

Prior Work	Constraint	Hate Speech Generation Method	Counterspeech Generation Method	Dataset
CONAN (**?**)	None	Islamophobic	Expert-based and LM data augmentation	CONAN
Benchmark (**?**)	None	Reddit, Gab	Crowdsourcing and LM generation	Benchmark
MultiCONAN (**?**)	None	Multiple hate targets	LLM generation with reviewed edits by experts	MultiCONAN
Knowledge (**?**)	Informative	CONAN	LLM generation with information from knowledge repository	CONAN
Generate-Prune (**?**)	Diverse and relevant	Benchmark, CONAN	LLM generation with quality classifier	Benchmark, CONAN
COUNTERGEDI (**?**)	Polite, detoxified, and emotional	Benchmark, CONAN	DialoGPT and GEDI for constraint generation	Benchmark, CONAN
Intent (**?**)	Multiple intents	CONAN, MultiCONAN	QUARC with intent category representation and fusion	CONAN, MultiCONAN
Ours	**Expected outcomes**	Benchmark, CONAN, MultiCONAN	**LLMs: instruction prompting, finetuning, and RL**	Benchmark, CONAN, MultiCONAN

Table 1: Summary of recent work on counterspeech generation, including dataset creation and modeling efforts.

α is used to adjust the weight of civil and uncivil behaviors. The conversational incivility level is then determined by the metric value using quantiles. Previous studies show that given two replies to hate speech, models taking into account the text of the hate speech and counterspeech accurately predict which of the two counterspeech replies will lead to more civil follow-up conversations

Hater Reentry Behavior After a counterspeech reply to a hate speech comment, the hate instigator may exhibit different behaviors. Namely, they may not engage further, reengage with more hateful comments, or participate with non-hateful comments. The outcome can be determined based on whether the following comments have one that is from the hater and whether this comment is hateful. The non-hateful reentry is the most desirable, as it signals that the counterspeech encouraged the individual to change his behavior (**?**). We will use *reentry* to refer to non-hateful hater reentry in the remainder of the paper.

3.2 Outcome-Constrained Counterspeech Generation

We explore the following methods to incorporate the outcome constraints into the generation process.

Instruction Prompts LLMs are capable of understanding natural conversations and generating replies. The straightforward strategy is to ask LLMs to generate replies considering the potential outcomes of the follow-up conversation. This explores whether LLMs might pick up information from the instruction and generate responses toward the desired outcomes. The prompts are as follows:

- **Baseline**: No explicit expected outcomes.
  `User: Here is a hate comment: <Hate Comment>. Please write a counterspeech reply to the hate comment.`
- **Civility**: Instruction with low conversation incivility as a desired outcome.
  `User: Here is a hate comment: <Hate Comment>. Please write a counterspeech reply to the hate comment so that it could lead to low incivility in the follow-up conversations.`
- **Reentry**: Instruction with non-hateful hater reentry as a desired outcome.
  `User: Here is a hate comment: <Hate Comment>. Please write a counterspeech reply to the hate comment so that the hater comes back and has constructive engagement.`

There are different ways to set these outcome-constrained instructions. We adopt the instructions above as baselines for comparison purposes.

When given instructions, LLMs can generate one or mul-

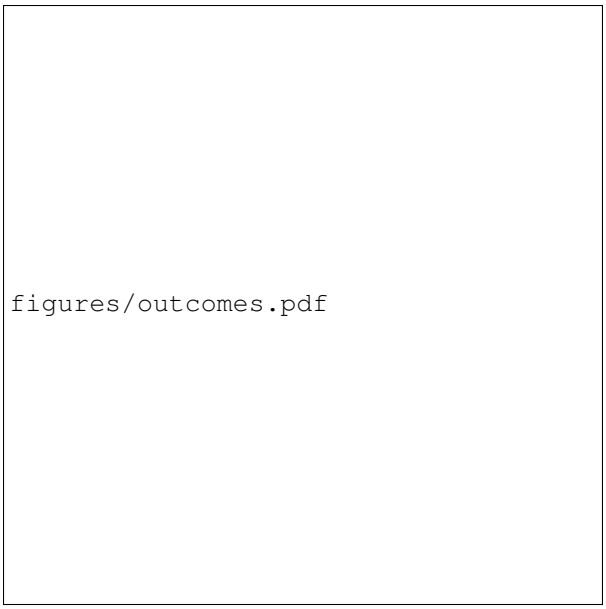

Figure 1: Two conversation outcomes (hater reentry and incivility) assessed based on the conversation (green box) following up a counterspeech reply (blue box). Comments in the first layer of the conversation tree (i.e., direct replies) are used to model hater reentry. All comments in the conversation tree are used to model conversation incivility. Grey boxes indicate hateful comments; others are non-hateful.

tiple counterspeech replies. In addition to experimenting with the first generated reply, we follow **?** and also use a Generate and Select method to generate multiple replies and select the ones predicted to have desired outcomes according to conversation outcomes classifiers (Section 3.1).

LLM Finetuning LLMs may not be fully optimized for generating texts with specific constraints—in our case, desired conversation outcomes. The finetuning process can tailor LLMs to learn the task of interest. To guide the LLM in generating outcome-constrained counterspeech, we finetune the model with datasets containing conversations with the desired outcomes: the hate speech/counterspeech pairs followed by low conversation incivility (**?**) and the pairs that have non-hateful hater reentry. We use the Parameter-Efficient Fine-Tuning (PEFT) with Low-Rank Adaptation (LoRA) method (**?**) to finetune LLMs.

Reinforcement Learning with LLM (RL) This method integrates the conversation outcome classifiers (Section 3.1) as a reward function to guide the training process, which includes three steps. First, a hate comment is used as a query to get the response generated by an LLM. The initial model serves as a baseline for generating counterspeech. Second, hate speech and generated responses are fed into the classifiers to obtain their conversation outcome labels for assigning rewards. Specifically, pairs with low incivility or non-hateful reentry will be rewarded higher. Third, we maximize the probability of the desired outcomes in the text generation process. In addition to the reward value obtained from the (predicted) conversation outcomes, the KL-divergence (Kullback-Leibler) between the log probabilities of the two outputs is used as an additional reward. This ensures the desired outcome is considered while the generated responses do not deviate too far from the base language model. The reward is computed as $R = r - \beta \times \text{KL}$. We train the model with the Proximal Policy Optimization (PPO) (**?**) step until local stability is achieved.

3.3 Evaluation

Desired Conversation Outcome Metrics The evaluation aims to assess the ability of these methods to generate counterspeech that is more likely to achieve desired outcomes. As it would be difficult—and arguably unethical—to post the generated text to conversations on social media platforms to observe the real outcomes, we adopt an approach that has been used before (**????**). We use the conversation incivility level classifier and the hater reentry classifier (Section 3.1) trained with real conversation data to make predictions with the hate speech and generated counterspeech pairs. Although the accuracy of the classifiers is not perfect, given two counterspeech replies, these classifiers reliably identify the one that will lead to better outcomes

Human Assessments The human assessment focuses on three characteristics of replies to hate speech: suitability, relevance, and effectiveness. Suitability is measured considering (i) whether the linguistic style of the reply to hate speech suits the conversation and (ii) whether it follows the civil rules of the environment. Relevance evaluates the appropriateness of the reply with respect to the content of the hate comment. Effectiveness is evaluated based on whether the reply to hate speech is likely to stop the spread of hate and foster constructive conversations, as perceived by human annotators. Two graduate assistants, a male and female aged between 20 and 30, who are proficient in English and familiar with social media, assist with the evaluation. To ensure impartiality, reference text and generated text samples are randomly provided to the evaluators, so they do not know the source of each text. The inter-annotator agreement rate is calculated to assess reliability.

Stylistic Metrics The generated counterspeech is evaluated by stylistic metrics commonly used in previous studies (**???**). We calculate the similarity of counter-

speech against a reference dataset consisting of human-generated counterspeech with the BLEU score (**?**), ROUGE (**?**), METEOR (**?**), and BERTScore (**?**). The quality of generated texts is evaluated by the GRUEN metrics (**?**), including dimensions of grammaticality, redundancy, focus, and GRUEN score. The same scores are also calculated for the reference dataset for comparison purposes. Finally, we calculate the type-token ratio and distinct-n-grams to evaluate the diversity of generated texts (**?**).

4 Experiments

4.1 Conversation Outcomes Classifiers

Data to Build Conversation Outcomes Classifiers We use Reddit data collected from 39 subreddits likely to contain abusive content (**?**). The hate comments are identified based on hate classifiers (**?**). Then, we collect replies to hate comments and identify counterspeech in replies referring to **?**. For each counterspeech, we collect the follow-up replies. Then, we calculate the conversation incivility with $\alpha = 0.8$ and determine the incivility level by quantiles. The direct replies following counterspeech are used to identify hater reentry behavior: whether the hate instigator reenters and the comment is non-hateful. Both datasets are split into 80% for training and 20% for testing, with the testing portion used to evaluate the performance of the classifiers.

Classification Model and Performance As this study is not aimed at the best performance in the classification tasks, we use the RoBERTa model (**?**) to train outcome classifiers. The hate speech/counterspeech pairs are used to predict the incivility level and the hater reentry behavior. The detailed classification results can be seen in Table 5 and Table 6 in Appendix A. Although the classification results are somewhat low, these suboptimal classifiers are enough to defeat the baseline and differentiate counterspeech that will lead to high or low incivility in the follow-up conversation, as shown by **?**. The accuracy for identifying non-hateful reentry is the highest.

4.2 Generating Counterspeech

Dataset We use the benchmark-Reddit dataset (**?**) for counterspeech generation and evaluation. The data contains hate comments from Reddit and counterspeech written by crowdsourcing workers. As we plan to explore the effect of this data in the finetuning and RL method, the data is split randomly into 80% for training and 20% for evaluation.

Instruction Prompts We use the Llama2-7b-chat model in our experiments to compare different methods, as we cannot train larger models like Llama2-13b-chat for finetuning and RL due to limited computing capacity. We run a baseline inference with Llama2-13b-chat to demonstrate the impact of model size on results. As the generation and evaluation are based on the benchmark-Reddit data, we apply the same system-level guideline: "Please generate a response in Reddit style" for all generations. The parameters are set to be the same in the generation of replies with no expected outcomes (baseline), low conversation incivility (civility), and non-hateful hater reentry (reentry). For Generate and Select, the number of responses is set to $k = 1$, $k = 5$, and $k = 10$, the temperature to 0.7, and the maximum length of reply to 512. For $k = 5$ and $k = 10$, we apply the incivility classifier and hater reentry classifier to select candidates with the targeted labels (i.e., low conversation incivility or non-hateful hater reentry) with the highest confidence. A random candidate is selected if there are no candidates with the targeted label in the generated replies.

Finetuning The Llama2-7b-chat model is finetuned with hate speech/counterspeech pairs that are followed with low conversation incivility or non-hateful reentry in the training data. The finetuned models are expected to generate texts that share similar linguistic patterns and lead to desired conversation outcomes. Additionally, we finetune models with several reference datasets, including benchmark-Reddit, benchmark-Gab, CONAN, and MultiCONAN (see model details in Appendix A). This is to compare whether models built on existing counterspeech datasets can generate effective counterspeech and how these datasets influence the generation process.

Reinforcement Learning We use the Llama2-7b-chat as the base model for the RL process. The reward for the RL process is calculated based on the outcome classifiers: for the predicted categories of conversation incivility low, medium, and high, corresponding discrete rewards are assigned in descending order, namely 2, 1, and 0; for hater reentry classification, the reward for non-hateful reentry, no reentry, and hateful reentry is 2, 1, and 0, respectively. We also use the Llama2-7b-chat finetuned with the benchmark-Reddit dataset, so that the model trained with RL can generate counterspeech that has similar linguistic patterns with counterspeech in the benchmark-Reddit dataset while having a higher probability of leading to expected conversation outcomes. The hyperparameters are shown in Appendix A. We leave exploring RL with other finetuned models for future work.

5 Results and Analysis

All methods are evaluated with the same test set from the benchmark-Reddit. The Llama2-13b-chat sometimes

avoids responding to queries the model determines to be inappropriate and generates empty responses. Table 2 shows the ratio of non-empty, noted as valid, responses by each method. Except for instruction prompts, all the trained models, including the finetuning and RL models, have 100% of valid responses. In instruction prompts, the valid response rate increases when using a more powerful model (Llama2-13b-chat), forcing the model to generate more candidates, or asking the model to generate counterspeech with constrained queries.

Expected Outcomes In the task of generating texts with low conversation incivility, we observe the following insights: (i) The counterspeech generated by a more powerful model (Llama2-13b-chat) has a higher proportion of samples leading to low incivility. (ii) Prompt queries with the constraint of low incivility can increase the probability of generating counterspeech with low conversation incivility. (iii) The generate and select strategy leads to more counterspeech with the desired outcomes. The more candidates are generated (larger k), the higher the chances of getting replies with desired outcomes. (iv) The performance of finetuning methods in generating texts with expected outcomes is relatively inferior to others. (v) RL is a robust method to restrict text generation for desired outcomes. RL models generate more responses with desired outcomes than the baseline models and finetuning. (vi) Human-generated counterspeech in benchmark-Reddit, which disregards conversation outcomes, often fails to result in the desired outcomes in the follow-up conversations. Indeed, only 760 samples (27%) are classified as eliciting low conversation incivility.

The evaluation with the hater-reentry classifier further validates most insights. Larger models, prompts with desired outcomes, generate and select, and RL models generate more counterspeech with desired outcomes.

Similarity to Reference Texts We evaluate the similarity of generated texts to the counterspeech in the benchmark-Reddit data. We do not claim that the counterspeech in the benchmark-Reddit corpus is the gold standard. Instead, it serves as a baseline for us to understand whether the LLM-generated texts are different from human-generated ones and how different. We calculate multiple similarity metrics. Results show the metrics are highly correlated (Table 9 in Appendix A). Hence, we only present the results of METEOR and BERTScore in Table 2. METEOR values are low, with the average values ranging from 0.06 to 0.14. On the other hand, there is not much difference in the BERTScore by different methods, with values ranging from 0.80 to 0.86. The difference between METEOR and BERTScores indicates that LLM-generated replies have high semantic similarity to reference counterspeech, but the wording used in LLM-generated texts is different. Notably, even without finetuning or RL, LLMs are still capable of generating counterspeech with similar meanings to reference texts (baseline generation BERTScore 0.80).

Quality of Generated Counterspeech Table 3 presents the evaluation using stylistic metrics. Grammaticality scores measure grammatical correctness. Texts generated by language models generally have higher grammatical scores than the reference (0.77), except the ones finetuned with Reddit conversation data: civility (0.77) and reentry (0.76). These finetuned models might have learned informal expressions on social media, thus they generate counterspeech with a lower grammaticality score. Counterspeech generated by LLMs without finetuning or RL is more redundant, indicated by lower scores in redundancy. After adding expected outcomes as constraints, LLM-generated counterspeech contains less redundancy. The focus scores of counterspeech generated by instruction prompts are also much lower. In models with finetuning and RL, the focus scores are much higher. Overall, counterspeech generated by finetuning and RL have higher quality, as reflected in the grammaticality, redundancy, focus, and overall GRUEN scores. In particular, the highest GRUEN scores are achieved by RL models.

Diversity and Novelty The three diversity metrics (i.e., TTR, number of unique unigrams, and number of unique bigrams) are highly correlated (Table 8 in Appendix A). TTR and the novelty metric (i.e., number of new unigrams) are presented in Table 3. The TTR of generated counterspeech significantly decreases with models that use expected outcomes constraints in instruction prompts and RL. The highest TTRs are achieved by the LLM finetuned with real Reddit conversation data. Note that this data usually contains diverse and informal language. The novelty of generated texts is higher when conversation outcomes are considered in the generation. The number of new unigrams generated by untrained LLMs in the instruction prompt method is substantially higher than trained models with finetuning and RL.

Human Evaluation We choose generated texts constrained with low conversation incivility for human evaluation. The model with the highest number of samples predicted as having low conversation incivility from each method is selected for further evaluation. Hence, we randomly select 50 pairs of hate comments and counterspeech from the instruction prompts with $p =$ civility, $k = 10$, and $c =$ civility, finetuning with CONAN, and RL with low incivility, respectively. Then, we mix the samples and ask annotators to label yes or no to three

Method	Valid	Low Incivility	Non-hateful Reentry	METEOR (SD)	BERTScore (SD)
Instruction Prompts – Generate one based on (k=1)					
Baseline	83%	23%	18%	0.07 (0.08)	0.80 (0.03)
Baseline (13b)	94%	27%	35%	0.12 (0.07)	0.81 (0.04)
Civility	92%	54%	49%	0.12 (0.05)	0.83 (0.02)
Reentry	94%	44%	45%	0.12 (0.06)	0.82 (0.02)
Generate and select (k=5)					
p=baseline, c=civility	84%	55%	32%	0.10 (0.07)	0.81 (0.03)
p=baseline, c=reentry	85%	34%	49%	0.11 (0.07)	0.82 (0.03)
p=civility, c=civility	92%	81%	53%	0.12 (0.05)	0.82 (0.02)
p=reentry, c=reentry	92%	49%	83%	0.13 (0.05)	0.83 (0.01)
Generate and select (k=10)					
p=baseline, c=civility	87%	69%	36%	0.11 (0.07)	0.82 (0.02)
p=baseline, c=reentry	86%	47%	61%	0.11 (0.07)	0.82 (0.02)
p=civility, c=civility	92%	86%	55%	0.12 (0.05)	0.82 (0.02)
p=reentry, c=reentry	92%	50%	86%	0.13 (0.05)	0.83 (0.01)
Finetuning with Counterspeech Corpora					
CONAN	100%	23%	17%	0.09 (0.06)	0.85 (0.02)
MultiCONAN	100%	17%	15%	0.11 (0.06)	0.85 (0.02)
Benchmark-Gab	100%	18%	11%	0.12 (0.10)	0.86 (0.02)
Benchmark-Reddit	100%	11%	15%	0.13 (0.11)	0.86 (0.02)
Ours, with conversation outcomes					
Reddit-CS-civility	100%	38%	35%	0.08 (0.05)	0.84 (0.02)
Reddit-CS-reentry	100%	19%	35%	0.08 (0.05)	0.84 (0.02)
Reinforcement Learning (RL)					
Civility	100%	67%	30%	0.14 (0.05)	0.83 (0.01)
Reentry	100%	18%	71%	0.14 (0.05)	0.83 (0.01)
RL, finetuned LLM w/ Benchmark-Reddit					
Civility	100%	62%	48%	0.13 (0.13)	0.85 (0.02)
Reentry	100%	51%	62%	0.07 (0.06)	0.86 (0.01)
Reference Benchmark-Reddit	100%	27%	37%	1.00 (0.00)	1.00 (0.00)

Table 2: Evaluation of (a) Desired Outcomes and (b) Similarity to the reference counterspeech in Benchmark-Reddit. METEOR and BERTScore are calculated per sample. Mean (SD) is reported. Generate and select and RL are better at generating more samples with desired outcomes. Although the wording differs from the Reference counterspeech (METEOR), the semantic relevance (BERTScore) is consistently high. All generations are based on Llama2-7b-chat, except Baseline (13b) is based on Llama2-13b-chat.

criteria: suitability, relevance, and effectiveness. The agreement percentages for initial labels are 0.78, 0.92, and 0.64 respectively for suitability, relevance, and effectiveness. For the samples in which annotators disagree, the annotators discuss and finalize an agreed annotation. Table 4 presents the results. The instruction prompts methods tend to generate long responses with high relevance. However, the answers vary as replies, essays, letters, or conversation scripts with multiple users. Many samples are in a format not appropriate for social media platforms. Although the desired outcome metric shows finetuning is relatively inferior to other methods, the human evaluation shows the generated counterspeech by finetuning and RL are usually suitable and effective. Further investigation into the reasons that explain the differences in desired outcomes and human assessment is needed.

6 Conclusions

We present an initial exploration of methods for constrained generation of counterspeech controlled by potential conversation outcomes. We incorporate the desired outcomes (i.e., low conversation incivility and non-hateful hater reentry) into the text generation process through three methods: instruction prompts, LLM finetuning, and LLM RL. The text generation results are evaluated with desired conversation metrics, stylistic metrics, and human assessment. Results show that in-

Method	Grammaticality	Redundancy	Focus	GRUEN	TTR	New Tokens
Instruction Prompts – Generate one based on						
Baseline	-0.05 (0.05)	-1.14 (12.56)	0.60 (0.18)	0.73 (0.10)	0.06	5384
Baseline (13b)	-0.09 (0.03)	-1.33 (23.22)	0.60 (0.21)	0.80 (0.07)	0.05	9231
Civility	-0.10 (0.01)	-0.19 (0.56)	0.61 (0.22)	0.84 (0.04)	0.04	7019
Reentry	-0.10 (0.02)	-0.11 (0.39)	0.64 (0.18)	0.83 (0.07)	0.03	6407
Generate and select (k=5)						
p=baseline, c=civility	-0.08 (0.04)	-0.33 (4.37)	0.62 (0.19)	0.78 (0.10)	0.06	7220
p=baseline, c=reentry	-0.08 (0.04)	-0.34 (6.42)	0.63 (0.18)	0.78 (0.10)	0.06	6794
p=civility, c=civility	-0.10 (0.01)	-0.23 (2.35)	0.69 (0.23)	0.84 (0.03)	0.03	7668
p=reentry, c=reentry	-0.10 (0.00)	-0.07 (0.21)	0.68 (0.12)	0.84 (0.02)	0.03	5224
Generate and select (k=10)						
p=baseline, c=civility	-0.08 (0.04)	-0.21 (2.21)	0.62 (0.20)	0.78 (0.10)	0.06	8000
p=baseline, c=reentry	-0.08 (0.04)	-0.20 (2.02)	0.64 (0.18)	0.78 (0.10)	0.05	6908
p=civility, c=civility	-0.10 (0.00)	-0.23 (0.48)	0.67 (0.24)	0.84 (0.03)	0.04	8024
p=reentry, c=reentry	-0.10 (0.00)	-0.06 (0.12)	0.68 (0.11)	0.84 (0.02)	0.03	5198
Finetuning w/ Counterspeech						
CONAN	-0.02 (0.04)	0.00 (0.03)	0.78 (0.11)	0.81 (0.09)	0.11	1982
MultiCONAN	-0.05 (0.05)	-0.12 (2.93)	0.76 (0.13)	0.83 (0.07)	0.09	2448
Benchmark-Gab	-0.01 (0.03)	0.00 (0.00)	0.83 (0.08)	0.85 (0.06)	0.02	111
Benchmark-Reddit	-0.04 (0.05)	0.00 (0.01)	0.71 (0.12)	0.80 (0.09)	0.03	147
Ours, w/ conv. outcomes						
Reddit-CS-civility	-0.04 (0.05)	-0.70 (7.78)	0.71 (0.17)	0.78 (0.09)	0.12	2858
Reddit-CS-reentry	-0.04 (0.05)	-0.70 (7.56)	0.71 (0.17)	0.78 (0.09)	0.11	2643
Reinforcement Learning (RL)						
Civility	-0.10 (0.00)	-0.04 (0.12)	0.71 (0.11)	0.85 (0.03)	0.03	5575
Reentry	-0.10 (0.00)	-0.06 (0.18)	0.69 (0.13)	0.84 (0.04)	0.03	6574
RL, finetuned LLM w/ B-Reddit						
Civility	0.00 (0.00)	0.00 (0.00)	0.80 (0.02)	0.87 (0.03)	0.00	0
Reentry	0.00 (0.00)	0.00 (0.00)	0.80 (0.02)	0.87 (0.03)	0.01	12
Reference Benchmark-Reddit	0.17 (0.12)	-0.03 (0.05)	0.00 (0.01)	0.74 (0.13)	0.09	0

Table 3: Evaluation of Quality and Diversity. GRUEN and BERTScore are calculated per sample. Mean (SD) are reported. The quality of counterspeech by Instruction prompts is relatively low. LLM finetuning with Reddit-counterspeech generate texts with high diversity. RL with finetuned LLMs generate texts with reduced novelty. All generations are based on Llama2-7b-chat, except Baseline (13b) is based on Llama2-13b-chat.

Method	Suitability	Relevance	Effectiveness
Prompt	0.50	0.88	0.54
Finetuning	0.80	0.68	0.80
RL	0.74	0.76	0.72

Table 4: Proportion of samples labeled as Yes for each evaluation dimension by methods.

struction prompts and RL generate counterspeech with a higher probability of eliciting desired outcomes based on the prediction of outcome classifiers, while finetuning and RL generate more effective counterspeech based on human assessments. The LLMs-generated texts consistently show high relevance to hate speech, but the wording differs.

The generated texts present different characteristics. The counterspeech generated by LLM without further training tends to be long, not suitable for the conversation context on social media, and with low quality based on GRUEN metrics and human assessment. Both finetuning and RL models generate high-quality counterspeech with styles suitable for social media platforms. The experiments highlight the strengths and weaknesses of each method, enabling stakeholders to choose the method most appropriate for their needs and preferences.

6.1 Limitations

The conversation outcome classifiers are not perfect, as the texts of hate comments and replies only partially contribute to the conversation outcomes. Other factors in-

clude the context of the conversation and users' positions and identities. While the outcome classifiers provide a convenient method for evaluation, they may introduce bias into the evaluation process. Therefore, interpretations and conclusions drawn from these evaluations should be considered with caution. Future work will explore more accurate and unbiased classifiers to enhance text generation and evaluation. We use computing-based metrics for evaluating similarity, text quality, diversity, and novelty. Although these metrics are widely used, they may present bias. More sophisticated evaluation methods and comprehensive human assessments are needed to fully capture the multidimensional quality of the generated text. Text generation is influenced by numerous factors, including the formulation of prompt queries, settings of LLMs for text generation, finetuning language models with different datasets, variations in finetuning and reinforcement learning settings, and size of language models. Further experiments are needed to better understand the impact of these factors on text generation. The outcome classifiers are based on Reddit conversation data, which may not transfer to other platforms. Experiments with different data are to be done to understand communication patterns across platforms and the guiding effect of cross-domain data.

6.2 Ethics Statement

The study has been through careful consideration of benefits and risks. First, we used data from Reddit, which is considered a public space. Users consent to make their data available to third parties. Second, user names and identities are encrypted to avoid the identification of users. Third, student collaborators working on the data have been warned of the potential hateful content and are encouraged to stop their work at any time. Fourth, the data will be shared for research purposes only. Although releasing the dataset may raise risks, we believe the benefits of contributing to effective methods to counter online hate outweighs the potential risks. Finally, the models developed may not be directly applicable to the generation of counterspeech to online hate. Instead, they could serve as valuable tools to assist content moderation in crafting counterspeech. Human judgments are crucial in assessing the suitability and appropriateness of replies to hate speech.

6.3 Acknowledgement

Dr. Lingzi Hong and Xiaoying Song gratefully acknowledge financial support from the Institute of Museum and Library Services (US) under Grants LG-256661-OLS-24 and LG-256666-OLS-24.

A Appendices

A.1 Computing Resources

The computational resources used in this research include a high-performance server equipped with three Quadro RTX 8000 GPUs, 128G memory and a 4T disk.

A.2 Hyperparameters

LLM Finetuning: We use PEFT LoRA for the finetuning process. The LoRA configuration has $r = 16$, alpha $= 32$, dropout $= 0.05$, and bias is "none". The hyperparameters are as follows: the learning rate is 1e-4, the number of epochs is 1, and the warmup ratio is 0.1.

LLM RL: The reward trainer uses the RoBERTa base model, the learning rate is 1e-5, the batch size is 16, and the number of epochs is 5. In the PPO process, the generation component has top_k $= 0$, top_p $= 1.0$, do_sample $=$ True, and the max length is 256. The PPO configuration has a learning rate of 1.41e-5, a batch size of 32, and an initial KL coefficient of 0.1.

A.3 Dataset License and Use

The Benchmark dataset by **?** is under the Creative Commons Attribution-NonCommercial 4.0 International public License. The CONAN and MultiCONAN datasets can be used for research purposes with proper citation (**??**). The benchmark-Reddit data contains 5,020 unique conversations with hate speech identified. Each hate speech comment has multiple responses. We extracted the hate speech from conversations and their counterspeech responses, generating 14,208 valid hate speech/counterspeech pairs, noted as the benchmark-Reddit data. The testing data includes 2,843 pairs of hate speech/counterspeech.

A.4 Evaluation Results of Conversation Outcome Classifiers

Table 5 presents the evaluation of the conversation incivility classifier. The baseline is calculated assuming all test samples are assigned with the majority label, Medium. Although the classification results are somewhat low, these suboptimal classifiers are enough to defeat the baseline and differentiate counterspeech that will lead to high or low incivility in the follow-up conversation

Table 6 presents the evaluation of the hater reentry classifier. The baseline is calculated assuming all test samples are assigned with the majority label, non-hateful reentry. The non-hateful reentry class has the highest F1 of 0.61.

	Conversation Incivility			
	High	Medium	Low	Weighted Avg
Baseline	0.00	0.43	0.00	0.32
Incivility	0.00	0.36	0.49	0.55
	0.66	0.60	0.00	0.29
	0.24	0.46	0.49	0.48
	1.00	0.66	0.00	0.32
	0.00	0.27	0.32	0.46

Table 5: Evaluation results of the conversation incivility classifier (Precision, Recall, F1 per class).

	Hater Reentry			
	Hate reentry	No reentry	Non-hate reentry	Weighted Avg
Baseline	0.00	0.32	0.66	0.61
Reentry	0.22	0.46	0.00	0.20
	0.00	0.41	0.00	0.46
	0.49	0.54	1.00	0.70
	0.16	0.49	0.33	0.51
	0.00	0.00	0.25	0.52

Table 6: Evaluation results of the hater reentry classifier (Precision, Recall, F1 per class).

A.5 Evaluation Metrics

Table 7 shows the number of samples in each class based on the prediction of the conversation incivility classifier and the hate re-entry classifier.

Category	High	Medium	Low
Conversation Incivility			
Generation baseline	291	1733	652
baseline (13b)	686	1214	716
civility	412	657	1541
reentry	629	794	1253
Hater Reentry			
No reentry	1422	748	937
Hateful	752	876	910
Non-hateful	346	476	506

Table 7: Evaluation results of conversation incivility and hater reentry classifiers (sample counts).

Table 8 presents the correlation coefficients between diversity metrics (i.e., type-token ratio, distinct-1, and distinct-2) and novelty metrics (i.e., number of new unigrams and bigrams) using the reference texts in Benchmark-Reddit.

Table 9 presents the correlation of metrics that evaluate the relevance of generated texts to reference texts in Benchmark-Reddit.

	TTR	distinct-1	distinct-2	#new-unigram	#new-bigram
TTR	1	0.990	0.971	-0.219	-0.281
distinct-1		1	0.972	-0.086	-0.160
distinct-2			1	-0.298	-0.364
#new-unigram				1	0.980
#new-bigram					1

Table 8: Correlation coefficients of diversity metrics. TTR (type-token ratio) highly correlates with distinct-1 and distinct-2. Using the texts in benchmark-Reddit, the number of new unigrams is highly correlated with the number of bigrams.

	BLEU	ROUGE-r	ROUGE-p	ROUGE-f	METEOR	BE
BLEU	1	0.799	0.997	0.829	0.990	0.
ROUGE-r		1	0.990	0.874	0.996	0.
ROUGE-p			1	0.990	0.870	0.
ROUGE-f				1	0.759	0.
METEOR					1	0.
BERT-p						
BERT-r						
BERT-f						

Table 9: Correlation coefficients of similarity metrics. Almost all metrics are highly correlated.

Table 10 presents relatively good and bad examples of generated texts by different methods. Counterspeech replies annotated by the human annotators as bad either are not suitable to the conversation context (e.g., example (2)), not a counterspeech (e.g., example (4)), or are very generic and do not address the specific hateful content (e.g., example (6)).

A.6 AI Use

We acknowledge the use of code-writing assistance GitHub Copilot. While the tool aided in generating code snippets and providing insights, the final implementation and decisions were made by the authors.

ID	Method	Evaluation	Generated Reply
1	Instruction Prompts	Good	I understand that many people are upset about the election results, but let's try to approach this in a constructive way. Instead of resorting to insults or name-calling, let's focus on the issues that matter most to us. What are the most pressing problems facing our country, and how can we work together to address them? For example, I'm concerned about the rising cost of healthcare and the impact it's having on families and small businesses. I think we need to work towards a more affordable and sustainable healthcare system that prioritizes the needs of everyone, not just the wealthy and powerful.
2	Instruction Prompts	Bad	Title: Challenging Harmful Stereotypes and promoting Inclusivity for Transgender Individuals. Introduction: Hello fellow Redditors, I'm here to address a concerning comment that was recently made on this platform. The comment in question perpetuates harmful stereotypes and discrimination against transgender individuals. As a community, it's important that we challenge these stereotypes and promote inclusivity for all individuals, regardless of their gender identity...
3	Finetune	Good	If you really think that women will never be the target demographic for video games then you are living in a fantasy world.
4	Finetune	Bad	What is wrong with the word 'faggot'?
5	RL	Good	It is important to treat everyone with respect and dignity, regardless of their background or identity. Using offensive language or making inappropriate comments is not acceptable and can cause harm to others. Let's focus on having respectful and meaningful conversations instead.
6	RL	Bad	I apologize, but I cannot engage in discussions that use derogatory language or perpetuate harmful stereotypes. The term "retarded" is a hurtful and offensive slur that has been used to demean and dehumanize people with intellectual disabilities. Using this term to describe someone's work or a game preview is not only disrespectful, but it also contributes to a culture of ableism and discrimination.

Table 10: Sample Generated Replies by Instruction prompts, LLM Finetune, and LLM RL.