Outcome-Constrained Large Language Models for Countering Hate Speech

Lingzi Hong
University of North Texas
lingzi.hong@unt.edu

Pengcheng Luo
Peking University
luopc@pku.edu.cn

Eduardo Blanco
University of Arizona
eduardoblanco@arizona.edu

Xiaoying Song
University of North Texas
XiaoyingSong@my.unt.edu

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 is seldom 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 (Buerger, 2021), promoting productive user engagement (Miškolci et al., 2020), and educating online users (Blaya, 2019).

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 (Qian et al., 2019) and human-in-the-loop strategies (Chung et al., 2021). 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 (Saha et al., 2022)), or embedded with knowledge (Chung et al., 2021). 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 (Mun et al., 2024). 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 (Horawalavithana et al., 2022), reentry behaviors (Wang et al., 2021), and the rise of hate speech (Liu et al., 2018). 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 coun-

Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing, pages 4523–4536
November 12-16, 2024 ©2024 Association for Computational Linguistics

Prior Work	Constraint	Hate Speech	Generation Method
CONAN (Chung et al., 2019)	None	Islamophobic	Expert-based and LM data augmentation
Benchmark (Qian et al., 2019)	None	Reddit, Gab	Crowdsourcing and LM generation
MultiCONAN (Fanton et al., 2021)	None	Multiple hate targets	LLM generation with review/edits by experts
Knowledge (Chung et al., 2021)	Informative	CONAN	LLM generation with information from knowledge repository
Generate-Prune (Zhu and Bhat, 2021)	Diverse and relevant	Benchmark, CONAN	LLM generation with quality classifier
COUNTERGEDI (Saha et al., 2022)	Polite, detoxified, and emotional	Benchmark, CONAN	DialoGPT and GEDI for constraint generation
Intent (Gupta et al., 2023)	Multiple intents	CONAN, MultiCONAN	QUARC with intent category representation and fusion
Ours	Expected outcomes	Benchmark, CONAN, MultiCONAN	LLMs: instruction prompting, finetuning, and RL

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

terspeech 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

Generating Counterspeech Table 1 presents recent work on counterspeech generation. CONAN has counterspeech written by NGO experts and augmented by language models (Chung et al., 2019); Benchmark was built with hate speech from Gab and Reddit and counterspeech created by crowdsourcing workers (Qian et al., 2019); and MultiCONAN is a high-quality, high-quantity dataset created by experts coupled with language model generation for hate speech with multiple targets (Fanton et al., 2021). Counterspeech generation models have been built with these datasets (Halim et al., 2023; Tekiroğlu et al., 2020, 2022; Bonaldi et al., 2024). Unlike us, none consider conversation outcomes elicited by the generated counterspeech.

Researchers have investigated counterspeech generation under constraints. Chung et al. (2021) proposed a generation pipeline grounded in external knowledge repositories to generate more informative and less biased replies. Zhu and Bhat (2021) 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. Saha et al. (2022) proposed an ensemble generative discriminator to generate more polite, detoxified, and emotion-laden counterspeech. Gupta et al. (2023) developed IntentCONAN, where the generation of counterspeech is conditioned on five intents: informative, denouncing, questioning, positive, and humorous. Similarly, Fraser et al. (2023) utilized ChatGPT to generate counter-stereotype text by incorporating countering strategies in queries. Hassan and Alikhani (2023) 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 (Chung et al., 2020)). Unlike us, none of these previous works generate counterspeech to elicit positive behaviors in the follow-up conversations.

Language Generation with Constraints Extensive studies have targeted language generation under complex lexical constraints such as formality (Jin et al., 2022), text with certain concepts (Lu et al., 2022), dialogue that takes latent variables (Bao et al., 2020), and knowledge-enhanced text (Yu et al., 2022a). 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

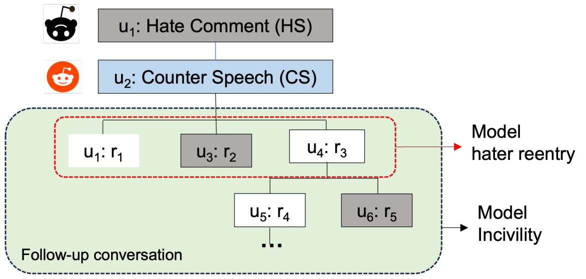

Figure 1: Two conversation outcomes (hater ressntry and incivility0 assessed based on the conversation (green box) following up a counterspeech reply (blue box). Comments in the first layer of the conversation tree (i.e., direct replies) are used to model hater reentry. All comments in the conversation tree are used to model conversation incivility. Grey boxes indicate hateful comments; others are non-hateful.

various datasets (Mou and Vechtomova, 2020). In this study, we generate counterspeech very likely to lead to desired conversational outcomes.

Methods have been developed for constrained language generation. Wang and Wan (2018) proposed the SentiGAN framework to generate text with a given sentiment. Kumar et al. (2021) proposed MUCOCO to allow for controllable inference with multiple attributes as constraints to the optimization. Krause et al. (2021) developed GeDi, a discriminator-based approach to guide the decoding process in language generation. It enables text generation with desired or undesired attributes. Schick et al. (2021) 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 (Liu et al., 2018; Yu et al., 2024). 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 (Yu et al., 2024). 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 $S(r) = \alpha U(r) - (1-\alpha)C(r)$, where $U(r)$ refers to uncivil behavior and $C(r)$ to civil behavior. For each user i ($i = 0, 1, 2, ..., k$), n_{ui} is defined as the number of uncivil comments by user i, and n_{ci} as the number of civil comments. Then, $U(r) = \sum_{i=1}^{k} \sqrt{n_{ui}}$ and $C(r) = \sum_{i=1}^{k} \sqrt{n_{ci}}$. 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 (Yu et al., 2024, binary classification, F1=0.66–0.75). We will use *civility* to refer to low conversation incivility, the desired outcome.

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 (Baider, 2023). 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 multiple counterspeech replies. In addition to experimenting with the first generated reply, we follow (Zhu and Bhat, 2021) 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 (Yu et al., 2022b) and the pairs that have non-hateful hater reentry. We use the Parameter-Efficient Fine-Tuning (PEFT) with Low-Rank Adaptation (LoRA) method (Hu et al., 2021) 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 $R = r - \beta * \text{KL}$. We train the model with the Proximal Policy Optimization (PPO) (Schulman et al., 2017) 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 (Saha et al., 2022; Tekiroğlu et al., 2022; Halim et al., 2023; Gupta et al., 2023). 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 (Yu et al., 2024, binary classification, F1=0.66–0.75). Thus, they serve as a proxy to compare counterspeech generated by different methods. Additionally, we conduct human assessments for reliability purposes.

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 (Chung et al., 2021; Zhu and Bhat, 2021; Tekiroğlu et al., 2022). We calculate the similarity of counterspeech against a reference dataset consisting of human-generated counterspeech with the BLEU score (Chen and Cherry, 2014), ROUGE (Lin, 2004), METEOR (Banerjee and Lavie, 2005), and BERTScore (Zhang et al., 2019). The quality of generated texts is evaluated by the GRUEN metrics (Zhu and Bhat, 2020), 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 (Fanton et al., 2021).

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 (Vidgen et al., 2021). The hate comments are identified based on hate classifiers (Qian et al., 2019). Then, we collect replies to hate comments and identify counterspeech in replies referring to Yu et al. (2022b). 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 (Liu et al., 2019) 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 6 in A.4. 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 (Yu et al., 2024). The accuracy for identifying non-hateful reentry is the highest.

4.2 Generating Counter Speech

Dataset We use the benchmark-Reddit dataset (Qian et al., 2019) 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 fine-tuned models are expected to generate texts that share similar linguistic patterns and lead to desired conversation outcomes. Additionally, we fine-tune models with several reference datasets, including benchmark-Reddit, benchmark-Gab, CONAN, and MultiCONAN (see model details in A.2). 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 Llama-2-7b-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 A.2. 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 Llama2-7b-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 the A.5). 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 0.8).

Quality of Generated Counterspeech Table 3 presents the evaluation using stylistic metrics. Grammaticality scores measure grammatical correctness. Texts generated by language models gen-

	Valid (%)	Desired Outcomes: Civility (%)	Desired Outcomes: Reentry (%)	Similarity: METEOR	Similarity: BERTScore
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%	41%	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%	48%	0.09 (0.06)	0.85 (0.02)
MultiCONAN	100%	22%	48%	0.11 (0.06)	0.85 (0.02)
Benchmark-Gab	100%	10%	43%	0.12 (0.10)	0.86 (0.02)
Benchmark-Reddit	100%	11%	42%	0.13 (0.11)	0.86 (0.02)
Ours, with conversation outcomes					
Reddit-CS-civility	100%	18%	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%	77%	71%	0.14 (0.05)	0.83 (0.01)
Reentry	100%	67%	62%	0.14 (0.05)	0.83 (0.01)
RL, finetuned LLM w/ Benchmark-Reddit					
Civility	100%	30%	48%	0.13 (0.13)	0.85 (0.02)
Reentry	100%	18%	57%	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.

erally 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

	Text Quality				Diversity	Novelty
	Grammaticality	Focus	Redundancy	GRUEN	TTR	New Tokens
Instruction Prompts						
Generate one based on						
Baseline	0.73 (0.10)	-0.05 (0.05)	-1.14 (12.56)	0.60 (0.18)	0.06	5384
Baseline (13b)	0.80 (0.07)	-0.09 (0.03)	-1.33 (23.22)	0.60 (0.21)	0.06	9231
Civility	0.84 (0.04)	-0.10 (0.01)	-0.19 (0.56)	0.61 (0.22)	0.03	7019
Reentry	0.83 (0.07)	-0.10 (0.02)	-0.11 (0.39)	0.64 (0.18)	0.03	6407
Generate and select (k=5)						
p=baseline, c=civility	0.78 (0.10)	-0.08 (0.04)	-0.33 (4.37)	0.62 (0.19)	0.06	7220
p=baseline, c=reentry	0.78 (0.10)	-0.08 (0.04)	-0.34 (6.42)	0.63 (0.18)	0.05	6794
p=civility, c=civility	0.84 (0.03)	-0.10 (0.01)	-0.23 (2.35)	0.59 (0.23)	0.04	7668
p=reentry, c=reentry	0.84 (0.02)	-0.10 (0.00)	-0.07 (0.21)	0.68 (0.12)	0.03	5224
Generate and select (k=10)						
p=baseline, c=civility	0.79 (0.09)	-0.08 (0.04)	-0.27 (2.27)	0.62 (0.20)	0.06	8000
p=baseline, c=reentry	0.80 (0.09)	-0.08 (0.04)	-0.20 (2.02)	0.64 (0.18)	0.05	6908
p=civility, c=civility	0.84 (0.03)	-0.10 (0.00)	-0.23 (0.48)	0.57 (0.24)	0.04	8024
p=reentry, c=reentry	0.84 (0.02)	-0.10 (0.00)	-0.06 (0.12)	0.68 (0.11)	0.03	5198
Finetuning w/ Counterspeech						
CONAN	0.81 (0.09)	-0.02 (0.04)	0.00 (0.03)	0.78 (0.11)	0.11	1982
MultiCONAN	0.83 (0.07)	-0.05 (0.05)	-0.12 (2.93)	0.76 (0.13)	0.09	2448
Benchmark-Gab	0.85 (0.06)	-0.01 (0.03)	0.00 (0.00)	0.83 (0.08)	0.02	111
Benchmark-Reddit	0.80 (0.09)	-0.04 (0.05)	0.00 (0.01)	0.77 (0.12)	0.03	147
Ours, w/ conv. outcomes						
Reddit-CS-civility	0.78 (0.09)	-0.04 (0.05)	-0.70 (7.78)	0.71 (0.17)	0.12	2858
Reddit-CS-reentry	0.78 (0.09)	-0.04 (0.05)	-0.70 (7.56)	0.71 (0.17)	0.11	2643
Reinforcement Learning (RL)						
Civility	0.85 (0.03)	-0.10 (0.00)	-0.04 (0.12)	0.71 (0.11)	0.03	5575
Reentry	0.84 (0.04)	-0.10 (0.00)	-0.06 (0.18)	0.69 (0.13)	0.03	6574
RL, finetuned LLM w/ B-Reddit						
Civility	0.80 (0.02)	0.00 (0.00)	0.00 (0.00)	0.80 (0.02)	0.00	0
Reentry	0.87 (0.03)	0.00 (0.00)	0.00 (0.00)	0.87 (0.03)	0.01	12
Reference						
Benchmark-Reddit	0.77 (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.

number of unique bigrams) are highly correlated (Table 8 in A.5). 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 $p = civility$, $k = 10$,

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.

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 criteria: suitability, relevance, and effectiveness. The agreement percentages for initial labels are 0.78, 0.92, and 0.64 respectively for suitability, quality, 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 *instruction 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.

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 include 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 fine-tuning 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.

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 HS.

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.

References

Fabienne Baider. 2023. Accountability issues, online covert hate speech, and the efficacy of counter-speech. *Politics and Governance*, 11(2):249–260.

Satanjeev Banerjee and Alon Lavie. 2005. Meteor: An automatic metric for mt evaluation with improved correlation with human judgments. In *Proceedings of the acl workshop on intrinsic and extrinsic evaluation measures for machine translation and/or summarization*, pages 65–72.

Siqi Bao, Huang He, Fan Wang, Hua Wu, and Haifeng Wang. 2020. Plato: Pre-trained dialogue generation model with discrete latent variable. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 85–96.

Catherine Blaya. 2019. Cyberhate: A review and content analysis of intervention strategies. *Aggression and violent behavior*, 45:163–172.

Helena Bonaldi, Yi-Ling Chung, Gavin Abercrombie, and Marco Guerini. 2024. Nlp for counterspeech against hate: A survey and how-to guide. *arXiv preprint arXiv:2403.20103*.

Catherine Buerger. 2021. # iamhere: Collective counterspeech and the quest to improve online discourse. *Social Media+ Society*, 7(4):20563051211063843.

Boxing Chen and Colin Cherry. 2014. A systematic comparison of smoothing techniques for sentence-level bleu. In *Proceedings of the ninth workshop on statistical machine translation*, pages 362–367.

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

Yi-Ling Chung, Serra Sinem Tekiroglu, and Marco Guerini. 2020. Italian counter narrative generation to fight online hate speech. In *Proceedings of the Seventh Italian Conference on Computational Linguistics (CLIC-it 2020)*, volume 2769.

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.

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.

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

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

Sadaf MD Halim, Saquib Irtiza, Yibo Hu, Latifur Khan, and Bhavani Thuraisingham. 2023. Wokegpt: Improving counterspeech generation against online hate speech by intelligently augmenting datasets using a novel metric. In *2023 International Joint Conference on Neural Networks (IJCNN)*, pages 1–10. IEEE.

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

Sameera Horawalavithana, Nazim Choudhury, John Skvoretz, and Adriana Iamnitchi. 2022. Online discussion threads as conversation pools: predicting the growth of discussion threads on reddit. *Computational and Mathematical Organization Theory*, pages 1–29.

Edward J Hu, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, Weizhu Chen, et al. 2021. Lora: Low-rank adaptation of large language models. In *International Conference on Learning Representations*.

Di Jin, Zhijing Jin, Zhiting Hu, Olga Vechtomova, and Rada Mihalcea. 2022. Deep learning for text style transfer: A survey. *Computational Linguistics*, 48(1):155–205.

Ben Krause, Akhilesh Deepak Gotmare, Bryan McCann, Nitish Shirish Keskar, Shafiq Joty, Richard Socher, and Nazneen Fatema Rajani. 2021. Gedi: Generative discriminator guided sequence generation. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 4929–4952.

Sachin Kumar, Eric Malmi, Aliaksei Severyn, and Yulia Tsvetkov. 2021. Controlled text generation as continuous optimization with multiple constraints. *Advances in Neural Information Processing Systems*, 34:14542–14554.

Chin-Yew Lin. 2004. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, pages 74–81.

Ping Liu, Joshua Guberman, Libby Hemphill, and Aron Culotta. 2018. Forecasting the presence and intensity of hostility on instagram using linguistic and social features. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 12.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Ximing Lu, Sean Welleck, Peter West, Liwei Jiang, Jungo Kasai, Daniel Khashabi, Ronan Le Bras, Lianhui Qin, Youngjae Yu, Rowan Zellers, et al. 2022. Neurologic a* esque decoding: Constrained text generation with lookahead heuristics. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 780–799.

Jozef Miškolci, Lucia Kováčová, and Edita Rigová. 2020. Countering hate speech on facebook: The case of the roma minority in slovakia. *Social Science Computer Review*, 38(2):128–146.

Lili Mou and Olga Vechtomova. 2020. Stylized text generation: Approaches and applications. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: Tutorial Abstracts*, pages 19–22.

Jimin Mun, Cathy Buerger, Jenny T Liang, Joshua Garland, and Maarten Sap. 2024. Counterspeakers' perspectives: Unveiling barriers and ai needs in the fight against online hate. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*, pages 1–22.

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

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

Timo Schick, Sahana Udupa, and Hinrich Schütze. 2021. Self-diagnosis and self-debiasing: A proposal for reducing corpus-based bias in nlp. *Transactions of the Association for Computational Linguistics*, 9:1408–1424.

John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. 2017. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*.

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

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.

Bertie Vidgen, Dong Nguyen, Helen Margetts, Patricia Rossini, and Rebekah Tromble. 2021. Introducing cad: the contextual abuse dataset. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2289–2303.

Ke Wang and Xiaojun Wan. 2018. Sentigan: Generating sentimental texts via mixture adversarial networks. In *Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence (IJCAI-18)*, pages 4446–4452.

Lingzhi Wang, Xingshan Zeng, Huang Hu, Kam-Fai Wong, and Daxin Jiang. 2021. Re-entry prediction for online conversations via self-supervised learning. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 2127–2137.

Wenhao Yu, Chenguang Zhu, Zaitang Li, Zhiting Hu, Qingyun Wang, Heng Ji, and Meng Jiang. 2022a. A survey of knowledge-enhanced text generation. *ACM Computing Surveys*, 54(11s):1–38.

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

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

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q Weinberger, and Yoav Artzi. 2019. Bertscore: Evaluating text generation with bert. In *International Conference on Learning Representations*.

Wanzheng Zhu and Suma Bhat. 2020. Gruen for evaluating linguistic quality of generated text. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 94–108.

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.

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 Qian et al. (2019) 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 (Chung et al., 2019; Fanton et al., 2021). 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 (Yu et al., 2024, binary classification, F1=0.66–0.75). 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.

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.

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.

Table 10 presents relatively good and bad examples of generated texts by different methods[1]. 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.

[1]The examples in this paper contain hateful content. We cannot avoid it due to the nature of our work.

	High			Medium			Low			Weighted Average		
	P	R	F1	P	R	F1	P	R	F1	P	R	F1
Baseline	0.00	0.00	0.00	0.49	1.00	0.66	0.00	0.00	0.00	0.24	0.49	0.32
Incivility	0.43	0.32	0.36	0.55	0.66	0.60	0.32	0.27	0.29	0.46	0.48	0.46

Table 5: Evaluation results of the conversation incivility classifier.

	Hate reentry			No reentry			Non-hate reentry			Weighted Average		
	P	R	F1	P	R	F1	P	R	F1	P	R	F1
Baseline	0.00	0.00	0.00	0.00	0.00	0.00	0.49	1.00	0.66	0.16	0.33	0.22
Reentry	0.32	0.20	0.25	0.52	0.41	0.46	0.54	0.70	0.61	0.49	0.51	0.46

Table 6: Evaluation results of the hater reentry classifier.

Category	Model	Conversation Incivility			Hater Reentry		
		High	Mediun	Low	No reentry	Hateful	Non-hateful
Generation	baseline	291	1733	652	1422	748	506
	baseline(13B)	686	1214	776	752	937	987
	civility	412	657	1547	876	346	1394
	reentry	629	794	1253	910	476	1290
Prompt and Select	p=baseline k=5 c=civility	195	855	1566	1117	595	904
	p=civility k=5 c=civility	134	176	2306	849	253	1514
	p=baseline k=5 c=reentry	415	1240	961	771	443	1402
	p=reentry k=5 c=reentry	914	312	1390	64	186	2366
	p=baseline k=10 c=civility	114	537	1965	1070	511	1035
	p=civility k=10 c=civility	73	100	2443	828	222	1566
	p=baseline k=10 c=reentry	444	994	1178	511	371	1734
	p=reentry k=10 c=reentry	890	295	1431	25	160	2431
LLM Finetune	civility	953	1298	592	881	954	1008
	reentry	939	1417	487	731	1152	960
	CONAN	1429	752	662	438	1031	1374
	MultiCONAN	1386	835	622	559	931	1353
	Benchmark-Reddit	1775	757	311	510	1149	1184
	Benchmark-Gab	1974	585	284	533	1076	1234
LLM TRL	civility	239	423	2181	292	540	2011
	reentry	481	461	1901	408	661	1774
	bm_reddit_ft_civility	66	1917	860	448	1036	1359
	bm_reddit_ft_reentry	1212	1130	501	222	992	1629
Reference	benchmark_reddit	1245	838	760	683	1117	1043

Table 7: Evaluation results of conversation incivility and hater reentry classifiers.

	TTR	distinct-1	distinct-2	#new_unigram	#new_bigram
TTR	1	**0.990**	**0.971**	-0.219	-0.298
distinct-1		1	**0.972**	-0.287	-0.364
distinct-2			1	-0.086	-0.160
#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	BERT-p	BERT-r	BERT-f
BLEU	1	0.799	0.997	0.990	0.973	0.870	0.983	0.952
ROUGE-r		1	0.829	0.874	0.916	0.466	0.866	0.649
ROUGE-p			1	0.996	0.983	0.848	0.985	0.939
ROUGE-f				1	0.995	0.806	0.99	0.914
METEOR					1	0.759	0.988	0.884
BERT-p						1	0.835	0.974
BERT-r							1	0.938
BERT-f								1

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

ID	Method	Evaluation	Hate Comment	Generated Reply
1	Instruction Prompts	Good	Wow. Can't believe anyone would actually vote for that cunt.	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	what was this dumb cunt thinking? "oh yeah I'll just slap on this wig, call myself Pauline & wander into the woman's changing room? Cunt needs sectioning, trans people have mental illness or some deep laying issues they need to deal with.	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	Why are these companies so retarded that they think this kind of PR is gonna help them? Women aren't the target gaming demographic and probably never will be.	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 a faggot bot. Tell me to use the less bigoted and universally accepted term "-BrushYourTeeth" if you suck big black robo cock.	What is wrong with the word 'faggot'?
5	RL	Good	Allo, ya queer bitch! 'Ow's yer sex life?	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	Nathan Grayson wrote that retarded article, but to be fair Kotaku's hands on preview of the game didn't have much good to say about it.	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*.