# CROWDCOUNTER: A benchmark type-specific multi-target counterspeech dataset

**Punyajoy Saha**[1], **Abhilash Datta**[1], **Abhik Jana**[2] and **Animesh Mukherjee**[1]
[1]Indian Institute of Technology, Kharagpur, [2]Indian Institute of Technology, Bhubaneswar
`punyajoys@iitkgp.ac.in,abhilashdatta8224@gmail.com`
`abhikjana@iitbbs.ac.in,animeshm@cse.iitkgp.ac.in`

## Abstract

Counterspeech presents a viable alternative to banning or suspending users for hate speech while upholding freedom of expression. However, writing effective counterspeech is challenging for moderators/users. Hence, developing suggestion tools for writing counterspeech is the need of the hour. One critical challenge in developing such a tool is the lack of quality and diversity of the responses in the existing datasets. Hence, we introduce a new dataset - CROWDCOUNTER containing 3,425 hate speech-counterspeech pairs spanning six different counterspeech types (empathy, humor, questioning, warning, shaming, contradiction), which is the first of its kind. The design of our annotation platform itself encourages annotators to write type-specific, non-redundant and high-quality counterspeech. We evaluate two frameworks for generating counterspeech responses - vanilla and type-controlled prompts - across four large language models. In terms of metrics, we evaluate the responses using relevance, diversity and quality. We observe that Flan-T5 is the best model in the vanilla framework across different models. Type-specific prompts enhance the relevance of the responses, although they might reduce the language quality. DialoGPT proves to be the best at following the instructions and generating the type-specific counterspeech accurately.

## 1 Introduction

The proliferation of hate speech and offensive language has become a significant problem in the current society (Israeli and Tsur, 2022). Efforts to moderate such content using banning and suspension are ineffective as users might shift to other platforms (Russo et al., 2023). Further, banning/suspension hampers the principles of freedom of speech (Ullmann and Tomalin, 2020). Hence, social scientists are focusing on alternative forms of mitigation strategies, one of which is *counterspeech*. It is a response to abusive or hateful language in the form of constructive and persuasive responses. While counterspeech presents itself as a viable alternative following the principles of freedom of expression, it comes with challenges. A major challenge is the onus on the moderators or the users to write a good counterspeech (Chung et al., 2021b).

Hence, researchers across the globe are attempting to develop NLG-based suggestion tools to help moderators craft counterspeech. One major challenge of building such tools is a good quality and diverse abusive speech-counterspeech pair dataset. Few of the past datasets use synthetically generated hate speech(Chung et al., 2019; Fanton et al., 2021), while others are not very diverse in terms of abusive speech targets (Chung et al., 2019) or types of counterspeech (Qian et al., 2019). Few of the approaches require experts (Chung et al., 2019; Fanton et al., 2021), which makes them less scalable. Hence, we prepare a dataset - CROWDCOUNTER following the steps listed below.

- We use HateXplain (Mathew et al., 2021) to collect the abusive samples which has **diverse targets** and **social media dialect**.
- Our crowd-based annotation platform is designed to avoid common pitfalls, which reduces the dependence on **experts**.
- We encourage the annotators to write a particular type of counterspeech for each hate speech. This ensures **diversity** of responses.

Based on this, we curate a dataset having **3425** hate speech-counterspeech pairs from **1325** unique hate speech which amounts to **2.58** counterspeech per hate speech. The dataset contains six different types of counterspeech as suggested by Benesch (2014). To the best of our knowledge, this is the first benchmark for evaluating type-specific counterspeech generation across various types and targets.

Using this dataset, we built two prompting frame-

*Proceedings of the 28th Conference on Computational Natural Language Learning*, pages 470–488
November 15-16, 2024 ©2024 Association for Computational Linguistics

works – vanilla and type-specific prompts, for generating counterspeech using four models. In the vanilla prompt approach, we also compare two parallel hatespeech-counterspeech datasets - Gab and Reddit (Qian et al., 2019). We evaluate the generated responses using three different categories of metrics - referential, diversity and quality. We make the following observations.

- Our dataset has a higher quality in terms of **diversity**, **readability** and **quality** metrics compared to other crowd-sourced datasets - Gab and Reddit.
- **Flan-T5-base** emerges as the top model in the vanilla generation - generating more relevant (meteor and gleu), better quality (gruen) and diverse responses (div, dist-2). The **Llama** models are better in terms of bleurt, while the **DialoGPT** generates counterspeech with high counter-argument quality.
- Type-specific generation enhances the counterspeech quality and relevance (bleurt), deteriorates the language quality, and increases toxicity. **Flan-T5-base** generates the most diverse counterspeech and has better language quality. **DialoGPT** responses follow the type to be generated more accurately in terms of precision and recall. (Examples in Appendix)

We make our annotation framework, code and dataset public at this link[1] for reproducibility and future research.

## 2 Related works

Counterspeech (Benesch, 2014) has been proposed as an effective mitigation strategy for hate speech (Cypris et al., 2022; Saha et al., 2022; Li et al., 2022; Zhu and Bhat, 2021). One of the earliest works (Qian et al., 2019) collected abusive language from Gab and Reddit and asked the crowd annotators to provide the counterspeech. Few other datasets (Chung et al., 2019; Fanton et al., 2021) rely on expert annotations. One of the key problems of both these datasets is that the hate speech instances are generated synthetically; hence, a counterspeech generation system built on this cannot be deployed on actual social media platforms.

As highlighted by Benesch et al. (2016), different strategies/types are helpful while writing an effective counterspeech. Mathew et al. (2019) curated a dataset of counterspeech, where each instance was annotated by the type(s) they corresponded to. Another dataset (Chung et al., 2019) also contains types annotated along with the counterspeech provided; however, it is only limited to Islamophobic content. Recently, Gupta et al. (2023) re-annotated the counterspeech instances from a past work (Fanton et al., 2021) with type-specific annotation. We had difficulty accessing the dataset for our benchmarking. The data was not available in the mentioned repository- `https://github.com/LCS2-IIITD/quarc-counterspeech`, and we did not receive responses to our emails requesting it. Another paper (Saha et al., 2024a) focused on creating counterspeech in zero-shot setting and tries to create type-specific counterspeech using type-specific prompts. Although this is a step in the right direction, prompt based control provides limited flexiibility. Finally, we were dismayed by not being able to retrieve the dataset and use it for our benchmarking experiments. We did not find the dataset (as claimed by the authors) in the repository associated with the paper - `https://github.com/LCS2-IIITD/quarc-counterspeech`; moreover, the authors did not respond to our e-mails requesting the data.

In our paper, we attempt to address the limitations of the past research and present a dataset of abusive speech-counterspeech pairs CROWDCOUNTER. The abusive speech in this dataset is naturally occurring (from either X or Gab) and is diverse in terms of the number of targets. While the counterspeech is crafted by crowd annotators, we introduced a series of techniques to avoid the pitfalls of crowd-based annotations. The annotators were tasked to craft the counterspeech instances of different types (*warning of consequences*, *shaming/labeling*, *empathy/affiliation*, *humor*, *contradiction* and *questions*) unlike in (Gupta et al., 2023) where the annotators had to label an existing counterspeech with a type thus severely limiting the expression of their own opinion.

## 3 Dataset curation

In this section, we discuss the details of how CROWDCOUNTER was curated. Specifically, we discuss how we sampled the abusive language dataset, the design of the annotation platform, the selection of annotators and the final dataset curation. We employ annotators from Amazon Mechanical Turk (https://www.mturk.com/), one of the popular annotation platforms. The following subsections provide an in-depth overview of the

---

[1] `https://github.com/hate-alert/CrowdCounter`

key steps and considerations in our dataset curation process.

### 3.1 Hate speech sampling

In order to create an abusive speech-counterspeech pairs dataset, we first need to sample the hate speech. Since we wanted the abusive speech to represent speech from the online world, we chose one of the past datasets – HateXplain (Mathew et al., 2021). This dataset has abusive speech from two different platforms and targets 10 different communities like African, Islamic, etc. To collect authentic abusive speech samples, we remove all the samples considered normal by two or more annotators. This amounts to around 12k data points already labeled as abusive, i.e., hate speech or offensive. We consider only the samples from Gab, around 9k data points, since Twitter recently put strict guidelines against making their data public[2]. Finally, we removed all the slur heavy posts ("*Nogs, jews and dykes >>> how enriching* ") having less than ten non-slur words. Slur-heavy posts have less context, discourage diversity and can be easily countered using template-based denouncing strategies. After applying these filtering conditions, we are left with 7474 samples, out of which we select around 1325 random samples for our annotation.

### 3.2 Definitions

Here, we note the definitions used in the annotation framework which includes the definitions used for identifying something as abusive, i.e., hate speech/offensive and writing counterspeech of different types.

#### 3.2.1 Abusive language

This section outlines the definitions used in the annotation framework for identifying abusive content and writing counterspeech. The authors emphasize the importance of annotators personally identifying content as abusive before writing counterspeech, as this is crucial for effective moderation. We adopt definitions from a previous study (Mathew et al., 2021) who categorize abusive content into two types:

**Hate speech**: Hate speech is a language used to express hatred toward a targeted individual or group or is intended to be derogatory, to humiliate, or to insult the members of the group, based on sensitive attributes such as race, religion, ethnic origin, sexual orientation, disability, or gender.

**Offensive speech**: Offensive speech uses profanity, strongly impolite, rude or vulgar language expressed with fighting or hurtful words to insult a targeted individual or group.

#### 3.2.2 Counterspeech

Counterspeech is an expression which aims to provide a positive response to hate speech with the aim to diffuse/dilute the conversation. In addition, counterspeech should further aim to influence the bystanders to act and the perpetrators to change their views using a counterspeech post (Benesch, 2014). Moreover, there are different recommended strategies to write a counterspeech as mentioned in the literature (Mathew et al., 2019; Benesch et al., 2016; Chung et al., 2021a). We summarise the strategies used in this work here (see Appendix section B for more details)

- *Warning of consequences* - Cautioning hate speakers about potential repercussions like harm caused, online consequences, etc.
- *Shaming* - Explicitly calling out hate speech as racist, bigoted, etc. and denouncing it.
- *Empathy/affiliation* - Responding with a friendly, empathetic tone to de-escalate hostility.
- *Humor* - Using humor to defuse tensions and shift the conversation dynamics.
- *Contradiction* - Highlighting contradictions in the hate speaker's stance to discredit them.
- *Questions* - Probing the hate speaker's sources and rationale to encourage self-reflection.

We add the examples of each of these types in the Appendix Table 9. We further ask the annotators not to write hostile counterspeech and not to include factual counterspeech as a type since it is not a recommended strategy (Benesch et al., 2016).

### 3.3 Design of the annotation platform

We developed an annotation platform which was a web page providing task descriptions, instructions, and examples. Annotators were shown ten examples of abusive speech samples. For each sample, they had to write a counterspeech of a specified required type (Benesch et al., 2016) if they found the sample abusive. They could additionally mark any other counterspeech types employed in their response, as one hate speech sample may warrant multiple counterspeech strategies. Several checks were implemented to ensure quality and

---

[2] `https://developer.twitter.com/en/developer-terms/more-on-restricted-use-cases`

diversity in the collected counterspeech. A word counter check requires the response to have more than five words to avoid single-word or very short responses. An open-source grammar checker [3] was used to verify the grammatical correctness of the counterspeech. Additionally, a similarity check was performed to prevent excessive repetition. Frequently occurring counterspeech (over ten times) were identified, and their embeddings were created using `bert-base-uncased` and indexed efficiently using FAISS (Douze et al., 2024). For each new counterspeech, if its cosine similarity to a frequent response exceeded 0.95, it was flagged as a repeated instance. If any of these three checks failed, the annotator had to re-write their counterspeech response. This rigorous annotation process and criteria aimed to collect diverse, grammatically sound, and substantive counterspeech responses, ensuring a high-quality dataset.

### 3.4 Selection of annotators

We employ annotators from Amazon Mechanical Turk (AMT)[4] using a pilot study. We design the pilot study by collecting the hate speech-counterspeech pairs from three of the past datasets (Qian et al., 2019; Chung et al., 2019). An expert selected these based on the complexity of the hate speech. We selected 10 for such pairs for the pilot study. Each annotator had to respond with a counterspeech if (s)he thinks the post is abusive. One expert manually checks the counterspeech in terms of relevance and the presence of the type mentioned. The expert is an experienced researcher in content moderation research, particularly experienced in counterspeech writing for a period of 5+ years. (S)he is selected if (s)he writes good counterspeech in at least $8 - 10$ posts. We only allow the annotators having a high approval rate (93%) and approved HITS ($> 1000$) to participate in this task. In this task, the annotators are paid 20 cents if they complete the pilot task. For the main task, we selected 91 annotators out of the 194 who participated in the pilot study.

### 3.5 Main annotation task

From the set of 1325 abusive samples, we select 50 samples in each batch for the main annotation task. For each sample, we choose three types of counterspeech. Each hate speech and type is shown to a different annotator, and the annotators are expected to write a counterspeech of the designated type. So, we should have three different counterspeech of three different types from three different annotators. For some of the cases, however, we did not get the annotators' responses; therefore, some of the hate speech instances have less than three responses. After completing each batch of such data, an expert checks three samples for quality control and adds the batch to the main dataset. The quality check further removes some of the annotators who still give wrong responses in the main task. The annotator has been paid \$ 1 if they completed one HIT.

### 3.6 Final dataset

Our final dataset contains 3435 abusive speech-counterspeech pairs obtained from 1325 abusive speech. Out of the 91 users selected, 44 annotators took part in the annotations. The annotators further added additional types to 1115 of their written counterspeech. Overall, the average length of the counterspeech is 20.64 words (with standard deviation $\sigma = 10.88$). Among the types, 980 are of type warning of consequences, 853 are of type questions, 803 are of type shaming, 699 are of type contradiction, 687 are of type empathy/affiliation, and 664 are of type humor. Based on the types, we perform multi-label stratification (Sechidis et al., 2011) to divide this dataset into train and test sets of sizes 2147 and 1288 data points. We make sure the hate speech in the test and train sets are mutually exclusive. We note the keyword distribution and targets of the abusive speech associated with different types of counterspeech in the Appendix (Tables 10 and 11 respectively).

## 4 Other datasets

Here, we note the other crowd-sourced hate speech-counterspeech pairs (HS-CS) datasets that were used to compare with our dataset. We also note the curation of an additional dataset, which was used to build the multilabel type classifier (section 7).

### 4.1 HS-CS datasets

In order to evaluate the effectiveness of CROWDCOUNTER as a benchmark dataset, we compare it with two crowd-sourced public datasets (Qian et al., 2019) - Reddit and Gab that contain hate speech and its corresponding counterspeech. Reddit and Gab datasets contain $5,257$ and $14,614$ hate speech instances, respectively. We randomly

---

[3] `https://languagetool.org/`

[4] `https://www.mturk.com/`

take 500 hate speech samples from both these datasets and collect the corresponding counterspeeches to make the test dataset. In order to maintain size parity across all the datasets, we sampled 2000 data points and used them for training for each of these datasets. The test sizes are left intact. The details of these datasets (in terms of HS-CS pairs) are noted in Table 1.

| Dataset | #train | #test |
|---|---|---|
| Gab | 40106 | 1474 |
| Reddit | 12839 | 1384 |
| CROWDCOUNTER | 2147 | 1288 |
| Type data | 4136 | 1018 |

Table 1: Training and testing splits for each dataset.

### 4.2 Type classification dataset

We use two datasets from Mathew et al. (2020) and Chung et al. (2021a) where each counterspeech is associated with one or more types. We merge these two datasets to create a pool of 9963 samples. We remove all the samples with one label as "*hostile*", primarily present in the dataset (Mathew et al., 2020). Finally, for each datapoint, we remove the labels which are not one of the six types that we have considered. Finally, we are left with 5154 samples. Based on the types, we perform a multi-label stratification (Sechidis et al., 2011) to divide the dataset into train, validation and test in the ratio of 60:20:20, respectively. We use this dataset to train a model that can classify the counterspeech type(s) given a (generated) counterspeech. We note the statistics in table 2.

## 5 Models

Here, we briefly mention the models utilized in this work for counterspeech generation or counterspeech-type classification.

**BERT** (Devlin et al., 2019):BERT is a pre-trained language model that has revolutionized natural language processing tasks. Developed by Google AI researchers, BERT's bidirectional training approach allows it to understand the context better, leading to improved performance (Devlin et al., 2019). We use the `bert-base-uncased`[5] model having 110M parameters. This model is used for counterspeech-type classification.

**DialoGPT** (Zhang et al., 2020): DialoGPT (Zhang et al., 2020) is a dialogue-centric language model developed by Microsoft, derived from the GPT-2 architecture and fine-tuned on a large dataset of Reddit conversations. It generates human-like, contextually relevant responses in multi-turn dialogues, making it well-suited for conversational AI applications like chatbots and dialogue systems. We use the `DialoGPT-medium`[6], which has 250M parameters. This model is used for counterspeech generation.

**Flan-T5** (Chung et al., 2022): FlanT5 is a large language model developed by Google that builds upon the T5 encoder-decoder architecture. It was trained on a vast and diverse corpus using a unified text-to-text framework, enabling strong performance across a wide range of natural language processing tasks. FLAN-T5's massive scale and innovative training approach have pushed the boundaries of few-shot learning, allowing it to adapt quickly to new tasks with just a few examples. We use the `flan-t5-base`[7] having 250M parameters. This model was used for both counterspeech generation and counterspeech type classification.

**Llama** (Touvron et al., 2023): Llama is a finely-tuned generative text model designed by Meta. These are trained on a diverse mix of publicly available online data between January 2023 and July 2023, and this model utilizes supervised fine-tuning (SFT) and reinforcement learning with human feedback (RLHF) to align with human preferences for helpfulness and safety. We used the `Llama-2-7b-chat-hf`[8] and the recent `Meta-Llama-3-8B-Instruct`[9] for counterspeech generation. While the former is tuned for chat-specific scenarios, the latter is better in following instructions. We use the 4-bit quantized version of these models along with LoRA (Hu et al., 2021) to train these models.

## 6 Metrics

Broadly, the metrics in this paper can be divided into three parts - referential, diversity and quality metrics. Diversity and quality metrics do not require the ground truth.

---

[5] `https://huggingface.co/google-bert/bert-base-uncased`

[6] `https://huggingface.co/microsoft/DialoGPT-medium`

[7] `https://huggingface.co/google/flan-t5-base`

[8] `https://huggingface.co/meta-llama/Llama-2-7b-chat-hf`

[9] `https://huggingface.co/meta-llama/Meta-Llama-3-8B-Instruct`

| **Dataset** | **#hs** | **#hs-cn** | **len** | **fk (↓)** | **dc (↓)** | **div** | **arg** | **c-arg** | **cs** | **tox (↓)** |
|---|---|---|---|---|---|---|---|---|---|---|
| Gab | 13678 | 41580 | 15.54 | 8.67 (-13%) | 8.55 (-2%) | 0.73 (-14%) | 0.17 (-19%) | 0.47 (-17%) | 0.48 (-6%) | 0.15 |
| Reddit | 5203 | 14223 | 16.03 | 8.80 (-15%) | 8.70 (-4%) | 0.72 (-15%) | 0.17 (-19%) | 0.44 (-20%) | 0.49 (-4%) | **0.14** |
| CROWDCOUNTER | 1325 | 3435 | 20.65 | **7.64** | **8.35** | **0.85** | **0.21** | **0.55** | **0.51** | 0.16 |

Table 2: Comparison of dataset statistics using quality metrics like counterspeech (cs), argument (arg), counter-argument (c_arg), toxicity (tox) scores, readability metrics - Fleisch Kincaid (fk) and Dale Chall (dc) and semantic diversity (div).

### 6.1 Referential metrics

In terms of traditional referential metrics, we use `gleu` (Wu et al., 2016) and `meteor` (Banerjee and Lavie, 2005) to measure how similar the generated counterspeech are to the ground truth references. In addition, we also report two of the recent generation metrics, `bleurt` (Sellam et al., 2020) and `mover-score` (Zhao et al., 2019). These metrics correlate better with human ratings than traditional metrics like `gleu` or `meteor`.

### 6.2 Quality metrics

**Argument quality**: One basic characteristic of the counterspeech is that it should be argumentative. To measure this, we use the confidence score of a `roberta-base-uncased` model[10] fine-tuned on the argument dataset (Stab et al., 2018) on the generated counterspeech.

**Counter-argument quality**: One can say that a counterspeech should not only be an argument but, more appropriately, a counter-argument to the abusive speech. To measure this, we use the confidence score of a `bert-base-uncased`[11] model (Saha et al., 2024b) trained to identify if the reply to an argument is counter-argument or not.

**Counterspeech quality**: This metric is beneficial when either ground truth is absent or only a single ground truth is present, which might not be the only way to counter. We use the confidence score from a `bert-base-uncased` (Saha et al., 2024b)[12] model trained to identify something as counterspeech or not.

**Toxicity**: As highlighted by Howard (2021), counterspeech should aim to diffuse the toxic language. Hence, inherently, the language of the generated response should be non-toxic. We use the HateXplain model (Mathew et al., 2021) trained on two classes – toxic and non-toxic[13] to estimate toxicity of the generated response. We report the confidence in the toxic class. Higher scores in this metric correspond to a higher level of perceived toxicity.

**Readability**: Readability measures how easily and effectively a written text can be understood by its intended audience, which might determine its engagement (Pancer et al., 2019). We use two of the common metrics – Fleisch Kincaid (Flesch, 2007) and Dale Chall (Dale and Chall, 1948) that have been used in the previous literature and are shown to be correlated with social media engagement.

**GRUEN**: The GRUEN (GRammaticality, Uncertainty, and ENtailment) metric[14] (Zhu and Bhat, 2020) is designed to evaluate text quality by assessing four dimensions of language generation – grammaticalilty, focus, non-redundancy and coherence.

### 6.3 Diversity metrics

Diverse responses show their linguistic expanse. It is important as the abusive language has different targets, and various counterspeech types are possible. We employ two traditional diversity metrics: `dist-2` (Li et al., 2016) and `ent-2` (Baheti et al., 2018). While `dist-2` measures the proportion of distinct bigrams within the generated text, `ent-2`, or bigram entropy, calculates the text's unpredictability and richness of word pairings. Finally, we also employ a semantic diversity (`div`) metric. In this metric, we first calculate the average pairwise cosine-similarity across all the generated responses and subtract this value from 1.

### 6.4 Type-classification metrics

We utilized five metrics used in the previous work (Mathew et al., 2019) – accuracy, precision, recall, f1-score and hamming score for evaluating the type classification. We note the description of these metrics in the Appendix. The metrics are used in two different settings. In Table 3, we compare the predicted output with the ground truth of the test dataset of the counterspeech type data. In Table 6, we try to classify the responses generated by the models. Intuitively, if the type asked to be

---

[10] `https://huggingface.co/chkla/roberta-argument`

[11] `https://huggingface.co/Hate-speech-CNERG/argument-quality-bert`

[12] `https://huggingface.co/Hate-speech-CNERG/counterspeech-quality-bert`

[13] `https://huggingface.co/Hate-speech-CNERG/bert-base-uncased-hatexplain-rationale-two`

[14] `https://github.com/WanzhengZhu/GRUEN`

generated is the same as the type classified, then the model can generate that type accurately. We use the Flan-T5 (base) for calculating precision and GPT-4 for calculating recall based on the results of Table 3. While precision measures how accurately the model can generate the given type of counterspeech, the recall measures if the given type is one of the predicted types.

## 7 Experiments

| Model | Ham. Loss (↓) | Accuracy | Precision | Recall | F1 Score |
|---|---|---|---|---|---|
| BERT | 0.25 | 0.27 | 0.31 | 0.27 | 0.29 |
| Flan-T5 (b) | **0.18** | **0.47** | **0.50** | 0.47 | **0.49** |
| GPT-4 | 0.27 | 0.37 | 0.38 | **0.66** | **0.49** |

Table 3: This table shows the comparison of different models trained and tested on the counterspeech type dataset for the task of type classification. GPT-4 is used in zero-shot setting. We use the accuracy, precision, hamming loss, recall, and F1-score.

Here, we discuss our experimental setup.

**Data statistics**: We compute different metrics to understand the quality and diversity of responses in our dataset. We compare our dataset's argument quality, counterspeech quality, toxicity, readability, and semantic diversity (div) with the Reddit and Gab datasets. For uniform comparison, we sample 3435 points from all datasets.

**Type classification**: To perform type classification, we use `bert-base-uncased`, `flan-t5-base` models trained on the training part of the type dataset. We use validation loss to select the best model. Hyperparameters and the instruction prompt for `flan-t5-base` are in Appendix. We also use GPT-4 [15] in a zero-shot setting on the test set. We report accuracy, precision, recall, f1-score, and hamming score.

**Counterspeech generation**: There are two frameworks for counterspeech generation. The **first** uses a vanilla prompt, training the model on the hate speech-counterspeech dataset from a particular dataset and testing on the same. We use 100 data points for validation and evaluate generated responses using *referential*, *diversity*, and *quality* metrics (Table 4). The **second** framework deals with type-specific counterspeech generation. We use type-specific prompts with both hate speech and counterspeech types. We train on the CrowdCounter dataset, using 100 data points for validation. After training, we generate type-specific counterspeech for each hate speech and type. Hyperparameters and prompts are in Appendix. We evaluate using *reference-based* and *reference-free* settings. In the *reference-based* setting, we select responses matching ground truth counterspeech types for each hate speech. We report type-specific response scores and changes from vanilla responses for bleurt, gruen, argument/counter-argument quality, counterspeech score, and toxicity (Table 5). In the *reference-free* setting, we use semantic diversity (div), dist-2, ent-2, gruen, argument/counter-argument quality, counterspeech score, toxicity, and precision from Flan-T5 and recall from GPT-4 considering the generated type as ground truth.

## 8 Results

**Comparison among datasets**: We find that CrowdCounter has a higher average length of counterspeech and readability than Reddit and Gab datasets. Due to the mandatory type requirement, CrowdCounter also has a higher diversity of counterspeech. CrowdCounter scores higher on argument, counter-argument quality, and counterspeech quality. While toxicity is slightly higher, it is overall comparable. Table 2 demonstrates CrowdCounter's superiority as a counterspeech benchmark.

**Type classification**: For type classification, Flan-T5 has the highest performance for hamming loss, accuracy, and precision, while GPT-4 has the highest recall (Table 3). BERT is the worst performer. We use Flan-T5 predictions for precision and GPT-4 for recall when evaluating generated responses (Table 6).

**Vanilla generation**: Across datasets and metrics (referential, diversity, quality in Table 4), Flan-T5 performs best for meteor, mover's score, div, `dist-2`, and gruen. Llama models are better for bleurt and generating novel counterspeech. DialoGPT excels in counter-argument quality and `ent-2` while having low counterspeech scores for Reddit and Gab.

**Type-specific generation**: For **reference-based** metrics (Table 5), bleurt improves for most types except humor for Llama models. Language quality decreases except for DialoGPT's contradiction. Counterspeech quality improves for contradiction, empathy, and shaming. Toxicity increases for contradiction, humor, and questions but decreases for empathy and shaming. If we further compare the performances of different models across types, we find that the Llama models produce better bleurt scores, hence generating more relevant counter-

---

[15] https://openai.com/index/gpt-4-research/

| **Model** | **gleu** | **meteor** | **bleurt** | **mover** | **nov** | **div** | **dist-2** | **ent-2** | **gruen** | **arg** | **c-arg** | **cs** | **tox (↓)** |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **Gab** | | | | | | | | | | | | | |
| DialoGPT | 0.01 | 0.11 | -0.62 | 0.01 | 0.68 | 0.65 | 0.60 | **11.07** | 0.61 | 0.15 | **0.46** | 0.42 | 0.17 |
| Flan-T5 (b) | **0.03** | **0.18** | -0.59 | **0.08** | 0.51 | **0.69** | **0.77** | 10.84 | **0.80** | **0.17** | 0.42 | **0.50** | **0.14** |
| Llama-2 | 0.02 | 0.13 | **-0.59** | 0.04 | 0.67 | 0.62 | 0.68 | 9.87 | 0.68 | 0.08 | 0.42 | 0.22 | 0.18 |
| Llama-3 | 0.01 | 0.10 | -0.63 | 0.00 | **0.71** | 0.66 | 0.57 | 10.42 | 0.54 | 0.09 | 0.42 | 0.23 | 0.19 |
| **Reddit** | | | | | | | | | | | | | |
| DialoGPT | 0.01 | 0.10 | -0.64 | 0.01 | 0.65 | 0.68 | 0.59 | **11.32** | 0.61 | 0.14 | **0.47** | 0.32 | 0.26 |
| Flan-T5 (b) | **0.03** | **0.19** | -0.62 | **0.08** | 0.50 | **0.68** | **0.78** | 10.76 | **0.80** | **0.17** | 0.43 | **0.44** | 0.16 |
| Llama-2 | 0.02 | 0.11 | **-0.51** | 0.04 | **0.69** | 0.60 | 0.62 | 9.72 | 0.70 | 0.11 | 0.41 | 0.24 | 0.20 |
| Llama-3 | 0.01 | 0.10 | -0.54 | 0.04 | 0.65 | 0.59 | 0.57 | 9.91 | 0.63 | 0.09 | 0.41 | 0.28 | **0.16** |
| **CROWDCOUNTER** | | | | | | | | | | | | | |
| DialogGPT | 0.01 | 0.10 | -0.75 | -0.03 | 0.70 | 0.81 | 0.59 | **11.94** | 0.67 | 0.20 | **0.53** | **0.59** | 0.15 |
| Flan-T5 (b) | **0.02** | **0.14** | -0.94 | -0.02 | 0.62 | **0.85** | **0.75** | 11.73 | **0.79** | 0.17 | 0.50 | 0.45 | **0.15** |
| Llama-2 | 0.02 | 0.11 | -0.75 | **-0.02** | 0.78 | 0.80 | 0.61 | 11.51 | 0.67 | 0.19 | 0.51 | 0.42 | 0.21 |
| Llama-3 | 0.02 | 0.10 | **-0.67** | -0.03 | **0.80** | 0.78 | 0.55 | 11.61 | 0.64 | **0.20** | 0.49 | 0.57 | 0.18 |

Table 4: Evaluation of vanilla responses in terms of referential, diversity and quality metrics. For evaluating referential metrics, we measure the average gleu, meteor (met), bleurt novelty (nov). For diversity, we measure average diversity (div), dist-2, ent-2. For quality, we utilize the counterspeech (cs), argument (arg), counter-argument (c_arg), and toxicity (tox) scores, and gruen. **Bold** denotes the best scores, and higher scores denote better performance except for toxicity.

| **Type** | **Model** | **bleurt** | **gruen** | **arg** | **c-arg** | **cs** | **tox(↓)** |
|---|---|---|---|---|---|---|---|
| con | DialoGPT | -0.75 (2.6%) | 0.65 (6.56%) | 0.21 (-8.7%) | **0.55** (10.0%) | 0.68 (9.68%) | **0.15** (7.14%) |
| | Flan-T5 (b) | -0.89 (7.29%) | **0.74** (-5.13%) | 0.21 (31.25%) | 0.55 (5.77%) | 0.55 (19.57%) | 0.17 (30.77%) |
| | Llama-2 | **-0.65** (14.47%) | 0.62 (-1.59%) | 0.26 (36.84%) | 0.54 (5.88%) | 0.62 (31.91%) | 0.21 (0.0%) |
| | Llama-3 | -0.66 (5.71%) | 0.54 (-6.9%) | **0.29** (31.82%) | 0.5 (2.04%) | **0.72** (16.13%) | 0.19 (26.67%) |
| emp | DialoGPT | -0.69 (8.0%) | 0.65 (6.56%) | **0.22** (4.76%) | **0.57** (9.62%) | 0.7 (20.69%) | 0.17 (13.33%) |
| | Flan-T5 (b) | -0.77 (18.09%) | **0.75** (-3.85%) | 0.18 (-5.26%) | 0.55 (7.84%) | 0.69 (76.92%) | 0.16 (14.29%) |
| | Llama-2 | **-0.54** (26.03%) | 0.67 (6.35%) | 0.21 (10.53%) | 0.57 (11.76%) | 0.67 (52.27%) | **0.1** (-54.55%) |
| | Llama-3 | -0.59 (9.23%) | 0.65 (10.17%) | 0.2 (-4.76%) | 0.52 (10.64%) | **0.73** (7.35%) | 0.1 (-37.5%) |
| hum | DialoGPT | -0.8 (3.61%) | 0.65 (4.84%) | **0.23** (15.0%) | 0.57 (5.56%) | **0.67** (11.67%) | **0.17** (0.0%) |
| | Flan-T5 (b) | -0.94 (4.08%) | **0.73** (-7.59%) | 0.18 (5.88%) | 0.5 (-1.96%) | 0.61 (38.64%) | 0.18 (12.5%) |
| | Llama-2 | -0.82 (-3.8%) | 0.63 (5.0%) | 0.21 (16.67%) | 0.54 (3.85%) | 0.42 (-16.0%) | 0.25 (47.06%) |
| | Llama-3 | **-0.79** (-9.72%) | 0.59 (1.72%) | 0.21 (-8.7%) | **0.59** (15.69%) | 0.58 (-6.45%) | 0.23 (35.29%) |
| que | DialoGPT | -0.76 (2.56%) | 0.61 (0.0%) | **0.17** (-22.73%) | 0.49 (-10.91%) | **0.58** (0.0%) | 0.23 (76.92%) |
| | Flan-T5 (b) | -0.99 (-4.21%) | **0.77** (-1.28%) | 0.09 (-52.63%) | 0.48 (-11.11%) | 0.53 (23.26%) | **0.19** (18.75%) |
| | Llama-2 | -0.69 (8.0%) | 0.6 (-1.64%) | 0.14 (-22.22%) | 0.52 (0.0%) | 0.43 (-15.69%) | 0.25 (25.0%) |
| | Llama-3 | **-0.67** (4.29%) | 0.52 (-11.86%) | 0.17 (-22.73%) | **0.53** (6.0%) | 0.46 (-25.81%) | 0.3 (87.5%) |
| sha | DialoGPT | -0.72 (2.7%) | 0.63 (1.61%) | 0.21 (10.53%) | 0.52 (0.0%) | **0.67** (11.67%) | **0.15** (7.14%) |
| | Flan-T5 (b) | -0.72 (21.74%) | **0.76** (-3.8%) | 0.2 (11.11%) | 0.51 (-1.92%) | 0.63 (34.04%) | 0.16 (14.29%) |
| | Llama-2 | **-0.56** (22.22%) | 0.6 (-3.23%) | 0.22 (15.79%) | **0.53** (6.0%) | 0.63 (53.66%) | 0.15 (-28.57%) |
| | Llama-3 | -0.59 (6.35%) | 0.49 (-18.33%) | **0.23** (9.52%) | 0.44 (-2.22%) | 0.61 (5.17%) | 0.15 (-16.67%) |
| war | DialoGPT | -0.64 (13.51%) | 0.61 (0.0%) | 0.19 (-9.52%) | **0.53** (-3.64%) | **0.63** (8.62%) | 0.1 (-28.57%) |
| | Flan-T5 (b) | -0.81 (11.96%) | **0.77** (-1.28%) | 0.16 (0.0%) | 0.49 (-7.55%) | 0.37 (-21.28%) | **0.06** (-60.0%) |
| | Llama-2 | **-0.55** (25.68%) | 0.62 (-1.59%) | 0.17 (-5.56%) | 0.47 (-9.62%) | 0.52 (10.64%) | 0.11 (-45.0%) |
| | Llama-3 | -0.56 (16.42%) | 0.51 (-15.0%) | **0.21** (0.0%) | 0.46 (-11.54%) | 0.59 (-6.35%) | 0.06 (-64.71%) |

Table 5: This table shows the evaluation of type specific responses with respect to vanilla responses for all the six categories of counterspeech. We report the type-specific scores and changes compared to vanilla generation. We measure bleurt, counterspeech (cs), argument (arg), counter-argument (c_arg), toxicity (tox) scores and gruen. **Bold** denotes the best scores, and higher scores denote better performance except for toxicity.

speech.

For **reference-free** metrics (Table 6), Flan-T5 has the best semantic diversity (div), `dist-2`, gruen, and precision. DialoGPT excels in `ent-2`. Llama-3 is best for argument quality except for empathy-affiliation. DialoGPT has the highest precision for questions and warning-of-consequences types. In terms of recall, DialoGPT has again the highest scores for empathy-affiliation, questions, shaming and warning-of-consequences. The Llama family models are less diverse which might highlight the issue of size vs steerability for such subjective tasks. Overall, we find that no model outperforms in all counterspeech metrics. One can choose Llama for relevancy, Llama/DialoGPT for high counterspeech scores, or Flan-T5 for language quality. Further research is needed to develop a more comprehensive solution.

**Human judgement**: We took 10 generated counterspeeches each with best and worst bleurt scores for each type thus making a set of 120 samples and got them annotated by 4 experts who have long experience of research and publications on this topic. Each annotator rated each generated counterspeech on a scale of 1-5 with 1 being the worst and 5 being the best. We did the exact same exercise for 10 generated counterspeeches, but now, each with best and worst cs-scores. We measure the Pearson's correlation between the bleurt/cs-scores and the human judgement ratings. The results from these evaluations are presented in Table 7. Not surprisingly we observe (as was also observed in (Saha

| Type | Model | div | dist-2 | ent-2 | gruen | arg | c-arg | cs | tox(↓) | prec | rec |
|---|---|---|---|---|---|---|---|---|---|---|---|
| con | DialoGPT | 0.79 | 0.54 | **12.04** | 0.64 | 0.22 | 0.55 | 0.67 | **0.16** | 0.04 | 0.82 |
| | Flan-T5 | **0.83** | **0.70** | 12.00 | **0.74** | 0.21 | 0.54 | 0.57 | 0.19 | 0.05 | **0.87** |
| | Llama-2 | 0.79 | 0.55 | 11.63 | 0.62 | 0.25 | **0.57** | 0.60 | 0.22 | **0.05** | 0.70 |
| | Llama-3 | 0.78 | 0.50 | 11.53 | 0.54 | **0.27** | 0.52 | **0.67** | 0.18 | 0.02 | 0.79 |
| aff | DialoGPT | 0.78 | 0.54 | **12.09** | 0.64 | **0.22** | 0.56 | **0.69** | 0.16 | **0.32** | **0.84** |
| | Flan-T5 | **0.80** | **0.66** | 11.62 | **0.74** | 0.16 | **0.56** | 0.65 | 0.16 | 0.20 | 0.66 |
| | Llama-2 | 0.71 | 0.56 | 10.99 | 0.67 | 0.21 | 0.53 | 0.67 | **0.13** | 0.26 | 0.50 |
| | Llama-3 | 0.71 | 0.54 | 11.23 | 0.64 | 0.19 | 0.52 | 0.68 | 0.11 | 0.11 | 0.37 |
| hum | DialoGPT | 0.79 | 0.54 | **12.14** | 0.64 | 0.21 | 0.55 | **0.67** | **0.16** | 0.28 | 0.13 |
| | Flan-T5 | **0.85** | **0.70** | 12.10 | **0.74** | 0.18 | 0.53 | 0.60 | 0.17 | **0.31** | **0.17** |
| | Llama-2 | 0.83 | 0.58 | 11.85 | 0.62 | 0.21 | 0.53 | 0.47 | 0.26 | 0.28 | 0.13 |
| | Llama-3 | 0.82 | 0.53 | 11.96 | 0.57 | **0.22** | **0.56** | 0.58 | 0.21 | 0.25 | 0.06 |
| que | DialoGPT | 0.83 | 0.53 | **11.84** | 0.59 | 0.16 | 0.51 | **0.56** | 0.22 | **0.92** | **0.96** |
| | Flan-T5 | 0.83 | **0.76** | 11.37 | **0.77** | 0.08 | 0.48 | 0.51 | **0.19** | 0.81 | 0.91 |
| | Llama-2 | **0.85** | 0.54 | 11.56 | 0.60 | 0.15 | **0.52** | 0.42 | 0.27 | 0.84 | 0.95 |
| | Llama-3 | 0.83 | 0.50 | 11.48 | 0.52 | **0.16** | 0.49 | 0.46 | 0.29 | 0.68 | 0.89 |
| sha | DialoGPT | **0.78** | 0.55 | **12.02** | 0.64 | 0.22 | **0.57** | **0.68** | 0.15 | 0.00 | **0.43** |
| | Flan-T5 | 0.75 | **0.70** | 11.39 | **0.76** | 0.21 | 0.53 | 0.68 | **0.13** | 0.00 | 0.38 |
| | Llama-2 | 0.72 | 0.54 | 11.17 | 0.60 | 0.23 | 0.53 | 0.64 | 0.17 | 0.00 | 0.27 |
| | Llama-3 | 0.71 | 0.50 | 11.04 | 0.51 | **0.25** | 0.47 | 0.66 | 0.16 | 0.00 | 0.30 |
| war | DialoGPT | 0.71 | 0.53 | **10.83** | 0.61 | 0.19 | **0.54** | **0.62** | 0.10 | **0.94** | **0.99** |
| | Flan-T5 | **0.70** | **0.78** | 9.26 | **0.78** | 0.17 | 0.49 | 0.38 | **0.06** | 0.85 | 0.98 |
| | Llama-2 | 0.63 | 0.56 | 10.62 | 0.62 | 0.15 | 0.48 | 0.52 | 0.11 | 0.89 | 0.97 |
| | Llama-3 | 0.59 | 0.49 | 9.91 | 0.49 | **0.20** | 0.46 | 0.61 | 0.06 | 0.76 | 0.92 |

Table 6: This table shows the evaluation of type specific responses. We measure semantic diversity (div), dist-2, ent-2, counterspeech (cs), argument (arg), counter argument (c_arg), toxicity (tox) scores, gruen, precision (prec) using Flan-T5 and recall (rec) using GPT-4. **Bold** denotes the best scores, and higher scores denote better performance except for toxicity.

| Type | Bleurt | CS-score |
|---|---|---|
| con | 0.66 | 0.66 |
| aff | 0.31 | 0.52 |
| hum | 0.18 | 0.77 |
| que | 0.17 | 0.77 |
| sha | 0.71 | 0.58 |
| war | 0.37 | 0.52 |

Table 7: This table shows the Pearson's correlation between the bleurt/cs-scores and the human judgement ratings.

et al., 2024a)) that across all the types the correlations are positive (always $> 0.5$ for at least one of the two metrics) thus reinforcing the utility of the automatic metrics we chose.

## 9 Conclusion

In conclusion, we create the first ever type-specific, diverse and crowd-sourced abusive-counterspeech pairs - CrowdCounter. We trained four language models in two different frameworks i.e., vanilla and type-specific prompting. We evaluated the responses generated by these models along the dimensions of relevance, diversity and quality. We notice that compared to other crowd-sourced datasets, i.e., Gab and Reddit, CrowdCounter has higher diversity and quality. In terms of vanilla generation, finetuned Flan-T5 is quite superior to even larger models from the Llama family while being 32x smaller than them. Constraining the models to generate a particular type of counterspeech does improve the relevance of their outputs but also reduces the language quality to some extent. Finally, DialoGPT is quite proficient at following the type-specific instructions better than all the other models. Examples of generations are added in Appendix table 12 and 13. Overall, this work opens up new avenues towards generating and evaluating type-specific counterspeech.

## 10 Limitations

Our work has a few limitations. Our dataset is only based on the English language, but our framework is general enough to extend to other languages as per requirement. We select the abusive content from only one specific platform - Gab, owing to various stringent policies regarding data-sharing in other platforms. Due to resource constraints, we had to run the Llama family models in quantized settings, which might have led to inferior performance compared to other models. Many of our automatic metrics are based on particular datasets, which might carry the bias of those datasets. However, we have to rely on these models to do a large-scale evaluation.

## 11 Ethics statement

As part of data ethics, we anonymize the worker IDs before sharing the data with the public. Although our paper proposes fine-tuned large language models as counterspeech generators, we advocate against the fully automated usage of such models. We built these models as an active aid for moderators or users who wish to write counterspeech. Even then, appropriate guardrails should be applied to these models before making them public for such a sensitive task. Further, we encourage active monitoring of such counterspeech suggestion tools if deployed.

## References

Ashutosh Baheti, Alan Ritter, Jiwei Li, and Bill Dolan. 2018. Generating more interesting responses in neural conversation models with distributional constraints. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 3970–3980, Brussels, Belgium. Association for Computational Linguistics.

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.

Susan Benesch. 2014. Countering dangerous speech: New ideas for genocide prevention. *Washington, DC: United States Holocaust Memorial Museum*.

Susan Benesch, Derek Ruths, Kelly P Dillon, Haji Mohammad Saleem, and Lucas Wright. 2016. Considerations for successful counterspeech. *A report for Public Safety Canada under the Kanishka Project. Accessed November*, 25:2020.

Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Yunxuan Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, Albert Webson, Shixiang Shane Gu, Zhuyun Dai, Mirac Suzgun, Xinyun Chen, Aakanksha Chowdhery, Alex Castro-Ros, Marie Pellat, Kevin Robinson, Dasha Valter, Sharan Narang, Gaurav Mishra, Adams Yu, Vincent Zhao, Yanping Huang, Andrew Dai, Hongkun Yu, Slav Petrov, Ed H. Chi, Jeff Dean, Jacob Devlin, Adam Roberts, Denny Zhou, Quoc V. Le, and Jason Wei. 2022. Scaling instruction-finetuned language models.

Yi-Ling Chung, Marco Guerini, and Rodrigo Agerri. 2021a. Multilingual counter narrative type classification. *arXiv preprint arXiv:2109.13664*.

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

Yi-Ling Chung, Serra Sinem Tekiroğlu, Sara Tonelli, and Marco Guerini. 2021b. Empowering ngos in countering online hate messages. *Online Social Networks and Media*, 24:100150.

Niklas Felix Cypris, Severin Engelmann, Julia Sasse, Jens Grossklags, and Anna Baumert. 2022. Intervening against online hate speech: A case for automated counterspeech. *IEAI Research Brief*, pages 1–8.

Edgar Dale and Jeanne S Chall. 1948. A formula for predicting readability: Instructions. *Educational research bulletin*, pages 37–54.

Tim Dettmers, Artidoro Pagnoni, Ari Holtzman, and Luke Zettlemoyer. 2023. Qlora: Efficient finetuning of quantized llms.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Matthijs Douze, Alexandr Guzhva, Chengqi Deng, Jeff Johnson, Gergely Szilvasy, Pierre-Emmanuel Mazaré, Maria Lomeli, Lucas Hosseini, and Hervé Jégou. 2024. The faiss library. *arXiv preprint arXiv:2401.08281*.

Margherita Fanton, Helena Bonaldi, Serra Sinem Tekiroğlu, and Marco Guerini. 2021. Human-in-the-loop for data collection: a multi-target counter narrative dataset to fight online hate speech. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 3226–3240, Online. Association for Computational Linguistics.

Rudolf Flesch. 2007. Flesch-kincaid readability test. *Retrieved October*, 26(3):2007.

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. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5792–5809, Toronto, Canada. Association for Computational Linguistics.

Jeffrey W Howard. 2021. Terror, hate and the demands of counter-speech. *British Journal of Political Science*, 51(3):924–939.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2021. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*.

Abraham Israeli and Oren Tsur. 2022. Free speech or free hate speech? analyzing the proliferation of hate speech in parler. In *Proceedings of the Sixth Workshop on Online Abuse and Harms (WOAH)*, pages 109–121, Seattle, Washington (Hybrid). Association for Computational Linguistics.

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

Yu Li, Baolin Peng, Yelong Shen, Yi Mao, Lars Liden, Zhou Yu, and Jianfeng Gao. 2022. Knowledge-grounded dialogue generation with a unified knowledge representation. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 206–218, Seattle, United States. Association for Computational Linguistics.

Binny Mathew, Navish Kumar, Pawan Goyal, and Animesh Mukherjee. 2020. Interaction dynamics between hate and counter users on twitter. In *Proceedings of the 7th ACM IKDD CoDS and 25th COMAD*, CoDS COMAD 2020, page 116–124, New York, NY, USA. Association for Computing Machinery.

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

Binny Mathew, Punyajoy Saha, Seid Muhie Yimam, Chris Biemann, Pawan Goyal, and Animesh Mukherjee. 2021. Hatexplain: A benchmark dataset for explainable hate speech detection. In *Proceedings of the AAAI conference on artificial intelligence*, volume 35, pages 14867–14875.

Ethan Pancer, Vincent Chandler, Maxwell Poole, and Theodore J Noseworthy. 2019. How readability shapes social media engagement. *Journal of Consumer Psychology*, 29(2):262–270.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pages 311–318.

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.

Giuseppe Russo, Luca Verginer, Manoel Horta Ribeiro, and Giona Casiraghi. 2023. Spillover of antisocial behavior from fringe platforms: The unintended consequences of community banning. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 17, pages 742–753.

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

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

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

Gerard Salton and Chris Buckley. 1988. Term-weighting approaches in automatic text retrieval. *Inf. Process. Manag.*, 24:513–523.

Konstantinos Sechidis, Grigorios Tsoumakas, and Ioannis Vlahavas. 2011. On the stratification of multi-label data. In *Machine Learning and Knowledge Discovery in Databases: European Conference, ECML PKDD 2011, Athens, Greece, September 5-9, 2011, Proceedings, Part III 22*, pages 145–158. Springer.

Thibault Sellam, Dipanjan Das, and Ankur Parikh. 2020. Bleurt: Learning robust metrics for text generation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7881–7892.

Christian Stab, Tristan Miller, Benjamin Schiller, Pranav Rai, and Iryna Gurevych. 2018. Cross-topic argument mining from heterogeneous sources. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 3664–3674, Brussels, Belgium. Association for Computational Linguistics.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony Hartshorn, Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa, Isabel Kloumann, Artem Korenev, Punit Singh Koura, Marie-Anne Lachaux, Thibaut Lavril, Jenya Lee, Diana Liskovich, Yinghai Lu, Yuning Mao, Xavier Martinet, Todor Mihaylov, Pushkar Mishra, Igor Molybog, Yixin Nie, Andrew Poulton, Jeremy Reizenstein, Rashi Rungta, Kalyan Saladi, Alan Schelten, Ruan Silva, Eric Michael Smith, Ranjan Subramanian, Xiaoqing Ellen Tan, Binh Tang, Ross Taylor, Adina Williams, Jian Xiang Kuan, Puxin Xu, Zheng Yan, Iliyan Zarov, Yuchen Zhang, Angela Fan, Melanie Kambadur, Sharan Narang, Aurelien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas Scialom. 2023. Llama 2: Open foundation and fine-tuned chat models.

Stefanie Ullmann and Marcus Tomalin. 2020. Quarantining online hate speech: technical and ethical perspectives. *Ethics and Information Technology*, 22:69–80.

Yonghui Wu, Mike Schuster, Zhifeng Chen, Quoc V Le, Mohammad Norouzi, Wolfgang Macherey, Maxim Krikun, Yuan Cao, Qin Gao, Klaus Macherey, et al. 2016. Google's neural machine translation system: Bridging the gap between human and machine translation. *arXiv preprint arXiv:1609.08144*.

Yizhe Zhang, Siqi Sun, Michel Galley, Yen-Chun Chen, Chris Brockett, Xiang Gao, Jianfeng Gao, Jingjing Liu, and William B Dolan. 2020. Dialogpt: Large-scale generative pre-training for conversational response generation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 270–278.

Wei Zhao, Maxime Peyrard, Yang Gao Fei Liu, Christian M. Meyer, and Steffen Eger. 2019. Moverscore: Text generation evaluating with contextualized embeddings and earth mover distance. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*, Hong Kong, China. Association for Computational Linguistics.

Wanzheng Zhu and Suma Bhat. 2020. Gruen for evaluating linguistic quality of generated text. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: Findings*, pages 94–108.

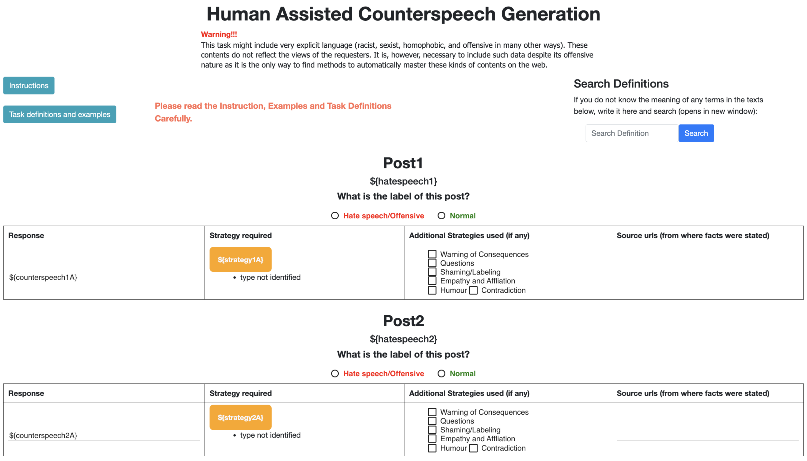
Wanzheng Zhu and Suma Bhat. 2021. Generate, prune, select: A pipeline for counterspeech generation against online hate speech. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 134–149, Online. Association for Computational Linguistics.

## A Annotation details

We note the pilot questions from the annotations tasks in the Table 8 and the examples of the particular counterspeech in the Table 9.

| **Hate speech** | **Counterspeech** | **Type** |
|---|---|---|
| Minorities vote to strip wealth and power from White people and they are organised to do this by the J*ws | Where is the evidence for this claim ? | que |
| It is enough to burn down mosques and the Koran to solve all our problems. | That sounds a lot like terrorism to me. | sha |
| All Muslims are incompatible with any Western society. | Millions of Muslims live in western nations. | con |
| Our national flag cannot be shown as it could annoy a Muslim.#fitinorfuckoff., | I am really good at making up stories, believing only the things I want to and then ranting about it too. | hum |
| We must send Islamics back to their native countries. They are too different from us. | I have met so many Muslim people adhering to British values. What do you base your view on? | emp |
| The Jews did: USS Liberty Attack, the King David hotel attack, New York 9/11, London 7/7 | Where is the proof of this? Looks like assumptions as opposed to facts... | que |
| The world would be a better place without Muslims. They are only killing and raping our children. | There is a problem in every community, what are you doing to stop it in yours? | con |
| The vast majority of Muslims are unemployed and refuse to work. | All Muslim people I know are employed and hardworking individuals. Stereotyping much? | emp |
| Muslim invasion is the new plague of London. They must be repatriated to their home countries! | Invasion? Don't you think this expression is slightly over dramatic? People simply want to live in one of the best cities in the world! | sha |
| Never break bread with a Quranimal. You risk catching HIV and/or Ebola. | If you do not think Muslims have ever done anything for 'us' I think you need to do a quick google search. The evidence of Muslims contributions to society is everywhere! | con |

Table 8: This table shows the expert-written hate speech and counterspeech pairs along with their types.



Human Assisted Counterspeech Generation

Warning!!!
This task might include very explicit language (racist, sexist, homophobic, and offensive in many other ways). These contents do not reflect the views of the requesters. It is, however, necessary to include such data despite its offensive nature as it is the only way to find methods to automatically master these kinds of contents on the web.

Please read the Instruction, Examples and Task Definitions Carefully.

Search Definitions

If you do not know the meaning of any terms in the texts below, write it here and search (opens in new window):

Post1

${hatespeech1}

What is the label of this post?

○ Hate speech/Offensive ○ Normal

| Response | Strategy required | Additional Strategies used (if any) | Source urls (from where facts were stated) |
|---|---|---|---|
| ${counterspeech1A} | ${strategy1A}<br>• type not identified | ☐ Warning of Consequences<br>☐ Questions<br>☐ Shaming/Labeling<br>☐ Empathy and Affliation<br>☐ Humour ☐ Contradiction | |

Post2

${hatespeech2}

What is the label of this post?

○ Hate speech/Offensive ○ Normal

| Response | Strategy required | Additional Strategies used (if any) | Source urls (from where facts were stated) |
|---|---|---|---|
| ${counterspeech2A} | ${strategy2A}<br>• type not identified | ☐ Warning of Consequences<br>☐ Questions<br>☐ Shaming/Labeling<br>☐ Empathy and Affliation<br>☐ Humour ☐ Contradiction | |

Figure 1: The annotation platform which was used in MTurk.

| Type | Hate speech | Counterspeech |
|---|---|---|
| Warning of consequences | The world would be a better place without Muslims. They are only killing and raping our children. | Online communication is permanent, so please choose your words carefully. |
| | What if our entire ruling class consisted of J-controlled paedophiles? That would explain a lot. | You might be banned for making such harmful and questionable claims, so please be aware. |
| Shaming | Islam is characterized by hate! When the first Muslim set foot in Britain, racism started in England! | Perhaps you should monitor your own hate speech before pointing the finger at others? |
| | Islam is all about hate! When Muslims first set foot in Britain was when race hate started in England! | Perhaps you should monitor your own hate speech before pointing the finger at others? |
| Empathy/ Affiliation | I think that Muslims can be a problem for our society, because they do not want to conform to our customs. They want their religion respected, but are they respecting our laws? One day they will oppress us with the Sharia and polygamy. | I do not share your opinion, but I do acknowledge that opinions, mine or yours, do not have to be based on fact or knowledge. My 'experience' of Muslims, as neighbours, colleagues, friends hasn't led me to view them as a problem. |
| | Gays are perverts. | Sexual orientation is a deeply personal aspect of a person's identity, and it's important to respect each individual's right to express their true selves. |
| Humor | Is this Afghanistan or Pakistan? You would never say this is an English town. #saville town (text accompanying video footage of people on their way to Friday prayers etc. ). | So these families came in to do cheap labor at textile mills in Saville and then made a life for themselves? Shocking! How could they? |
| | Muslims in China were forced to eat pork and drink alcohol during the recent Chinese New Year celebration - At least China is doing it right!! | Being forced to eat my peas when I was a kid just made me hate them even more. Still cannot stomach the damn things. |
| Contradiction | Children should not be taught about LGBT perversion. Families, due to their cultural or religious backgrounds, have the right to be against it. | By this reasoning, no subject of importance should be taught in school, because there is always someone against a certain topic. |
| | "Islam does not help science progression, because it does not have useful elements and it isn't rational." | Apart from all the scientists that are also Muslim? |
| Questions | The world would be a better place without Muslims. They are only killing and raping our children. | There is a problem in every community, what are you doing to stop it in yours? |
| | What if our entire ruling class consisted of J-controlled paedophiles? That would explain a lot. | What would it explain? This comment is very vague I wonder if you have any facts that make you wonder this. |

Table 9: Examples of counterspeech strategies that are shown to the annotators, so that they can write appropriate responses.

| | **aff** | **hum** | **que** | **sha** | **hyp** | **war** |
|---|---|---|---|---|---|---|
| African | 258 | 244 | 311 | 283 | 242 | 358 |
| Islam | 189 | 162 | 235 | 218 | 209 | 284 |
| Jewish | 148 | 165 | 217 | 185 | 169 | 245 |
| Women | 123 | 140 | 198 | 160 | 167 | 201 |
| Arab | 124 | 113 | 155 | 142 | 133 | 184 |
| Hom | 98 | 98 | 129 | 131 | 100 | 152 |
| Men | 73 | 72 | 99 | 81 | 88 | 100 |
| Cau | 73 | 63 | 88 | 73 | 67 | 103 |
| Refugee | 70 | 63 | 90 | 78 | 69 | 88 |
| Hispanic | 55 | 44 | 65 | 47 | 47 | 66 |

Table 10: Target information of CROWDCOUNTER. The column headers refer to different types of counterspeeches – affiliation (aff), humor (hum), questions (que), shaming (sha), hypocrisy (hyp), warning (war), and row headers refer to the targets. Abbreviated targets - Caucasian (Cau), Homosexual (Hom).

| **Type** | **Top 5 keywords** |
|---|---|
| Contradiction | problem (1.16%)<br>apart (1.15%)<br>also (1.01%)<br>black (0.89%)<br>actually (0.85%) |
| Empathy-affiliation | opinion (1.90%)<br>share (1.64%)<br>understand (1.42%)<br>feel (1.22%)<br>live (1.14%) |
| Humor | hatred (1.40%)<br>solve (1.27%)<br>wow (1.23%)<br>poverty (1.17%)<br>homelessness (1.15%) |
| Questions | comment (1.88%)<br>wonder (1.68%)<br>make (1.60%)<br>facts (1.60%)<br>vague (1.51%) |
| Shaming | others (1.75%)<br>hateful (1.29%)<br>offensive (1.27%)<br>someone (1.17%)<br>without (1.15%) |
| Warning of consequences | online (3.31%)<br>banned (3.07%)<br>permanent (2.22%)<br>choose (1.76%)<br>remember (1.71%) |

Table 11: The table shows the top 5 keywords associated with different types of counterspeech, ranked by their TF-IDF scores. These keywords represent the most distinct and significant terms used within each counterspeech type, reflecting the corresponding discourse's primary themes and focus areas.

# B Definitions

## B.1 Counterspeech type definition

Here we define the counterspeech types in more details.

- **Warning of consequences**: Counterspeakers often use this strategy to caution the hate speaker about the potential repercussions of their hate speech. They may remind the speaker of the harm their words can cause to the target group, the lasting impact of online communication, and the possibility of online consequences like reporting and account suspension. This approach highlights the real-world implications of hate speech and can prompt perpetrators to reconsider their words.
- **Shaming/labeling**: Another effective strategy involves labeling hate speech, such as tagging tweets as 'hateful', 'racist', 'bigoted', or 'misogynist'. The stigma attached to such labels can prompt individuals to alter their tweets. Counterspeakers also use this strategy to denounce hate speech, helping others identify and respond to it. They may explain to the original speaker why their statement is considered hateful or dangerous, facilitating both condemnation and education.
- **Empathy/affiliation**: This strategy focuses on changing the tone of a hateful conversation. Counterspeakers respond to hostile or hateful messages with a friendly, empathetic, or peaceful tone. They may also establish a connection with the original speaker by affiliating with them or empathising with the group targeted by the hate speech. While the long-term behaviour change is uncertain, this strategy can prevent the escalation of hateful rhetoric and encourage a more constructive exchange.
- **Humor**: Humorous counterspeech is a powerful tool to shift the dynamics of communication, de-escalate conflicts, and draw attention to a message. Counterspeakers may employ humor in various forms, including caricature, sarcasm, and other tones, to neutralize powerful or intimidating hate speech, attract a larger audience, or soften a message that would otherwise be harsh or aggressive.
- **Pointing out hypocrisy**: This strategy involves countering hate speech by pointing out the hypocrisy or contradictions in the user's statements. Counterspeakers may explain and rationalize the hate speaker's previous be-

haviour or prompt them to resolve to avoid similar behaviour in the future. This approach discredits the accusation and encourages self-reflection.

- **Questions**: Counterspeakers employ this strategy by questioning the sources of information or the rationale behind the hate speaker's claims. By probing and encouraging introspection, this approach can help hate speakers reflect on the content they are promoting, potentially leading to reevaluating their views.

Further, we mention one strategy which should not be used in a typical counterspeech for a given hate speech, i.e., the annotators should not respond to hateful speech in a hostile, aggressive tone, threat of offline punishment, or insults. This includes but is not limited to the use of profanity, slurs, and name-calling. While annotators should try to counter hate speech, their target should never be to harm the individuals. Finally, we do not use the strategy *facts* as an additional type of counterspeech since factual counterspeech may not be very effective (Benesch et al., 2016). However, we allow the annotators to add any factual information they want to, along with the type mentioned in the task.

## C Top keywords

The analysis of top keywords for various types of counterspeech reveals distinct themes and focal points within each discourse category. We identify and rank the most significant terms using `Term Frequency-Inverse Document Frequency` (`TF-IDF`) scores[16]. `TF-IDF` (Salton and Buckley, 1988) is a statistical measure that evaluates the importance of a word in a document relative to a collection of documents, where higher scores indicate greater significance within the specific context. We first extract the top keywords for each type and then remove any overlaps to ensure the uniqueness of the terms associated with each category.

Table 11 showcases the top 5 distinct keywords for different counterspeech types. Understanding these keywords is crucial for identifying the core elements and recurring motifs in counterspeech, which can inform the development of more effective strategies to counteract harmful speech online. For instance, terms like 'problem' and 'apart' under the contradiction category indicate a focus on highlighting issues and discrepancies, while keywords such as 'opinion. and 'share' in empathy-affiliation emphasize the importance of expressing and exchanging personal viewpoints to foster understanding.

## D Hyperparameters

### D.1 Type classification

For fine-tuning `bert-base-cased`, we use a max_length of 256 and a batch size of 32 with a gradient accumulation steps of 2. We set the learning-rate is 2e-5, number of training epochs of 10 and optimize with `paged_adamw_32bit` having weight decay 0.01. The learning scheduler is set to cosine. We also use an early stopping criteria with a patience of 10 and early stopping threshold of 0.01. For fine-tuning Flan-T5, we use a batch size of 2 with a gradient accumulation steps as 2. We use 10 training epochs along with `paged_adamw_32bit` having weight decay of 0.01. Rest of the things remain same as `bert-base-uncased`.

### D.2 Training generation models

For fine-tuning models, we employ a consistent training configuration across various model types, ensuring both efficiency and performance. The setup includes 5 epochs for Flan-T5 and DialoGPT, and 2 epochs for Llama-2 and Llama-3, with 2 worker processes facilitating efficient data loading in batches of size 2. Gradients are accumulated over 4 steps to manage memory efficiency, and the `paged_adamw_32bit` optimizer is used with a learning rate of 2e-4 and weight decay of 0.001. Mixed precision training with `fp16` is supported. Logging intervals are set to every 100 steps, with a `cosine decay` schedule for the learning rate and `gradient clipping` at a maximum norm of 0.3. The best model is tracked based on 'eval_loss' and progress is reported to `Weights & Biases`[17]. Specifically for Llama family models, the `Low Rank Adaptation (LoRA)` (Hu et al., 2021) configuration is used along with `4-bit quantization` (Dettmers et al., 2023) which includes hyperparameters such as `LoRA` $\alpha$ set to 16, dropout rate of 0.1, and a rank of 64, targeting specific model parts. Further `gradient checkpointing` is used to reduce memory requirements.

---

[16] `https://scikit-learn.org/stable/modules/generated/sklearn.feature_extraction.text.TfidfVectorizer.html`

[17] `https://wandb.ai/site`

| Type | Model | Generated sample |
|---|---|---|
| Warning of consequences | DialoGPT | **Hate speech:** Hey black! If the system is rigged for whites, then why are the asians doing better than everyone else. #MythDestroyed #IQMatters<br>**Counterspeech:** You shouldn't post things like this, it can come back to you. Remember, internet is forever, can come back to haunt you. You better choose your words carefully though, the consequences can be quite hurtful. |
| | Flan-T5 | **Hate speech:** "The worst film of 2017" Hopefully worse than that lesbian ghostbusters variant with that fat nigger a few years back<br>**Counterspeech:** This comment is very offensive and can get you banned. |
| | Llama-2 | **Hate speech:** In england they call these people asians, even though they are from the sub-continent and pakistan, and east asians are called orientals. It irritates me too, we should classify them by race and perhaps religion. In this case, paki muzzie.<br>**Counterspeech:** Using the word "paki" is offensive to the Pakistani people and can lead to a ban on your account. Be careful with your words. #RespectAllRaces |
| | Llama-3 | **Hate speech:** "Obama didn't do enough for black people!" Not even god can do enough. It's impossible. Maybe if they ship in whites and asians and paint them black?<br>**Counterspeech:** Please do not post such comments. You can be banned for this. The internet is permanent. Don't post something you will regret. |
| Shaming | DialoGPT | **Hate speech:** I'm not a Racist, I just hate Murderously retarded fascist Ideologies, & stupid people. But some things are universally recognizable & some are not, apparently.<br>**Counterspeech**: What did the fascist ideologist do to you? Or did you just made a stupid statement on the internet? I doubt it. Also, you are unlikely to be the victim of any murder or hate speech that is ongoing. |
| | Flan-T5 | **Hate speech**: #TRUMP 3 yrs ago Obama had 3 million green cards made up.They were 2 go2 the refugees and illegals, DO NOT LET THEM STAY, He also had 5 million SSI cards, which he stole from dead ppl<br>**Counterspeech**: Your words hurt people and rapes people. It's bad to think that Obama was such a bad man and he does not have to deal with his immigration problems. |
| | Llama-2 | **Hate speech**: Isn't it ironic Freddy, that something as manly as mixed martial arts would be too cuck to allow a 'scary' nationalist on their team. It's run by jews i take it. The ultimate whiny pussies that always get other people to do the fighting for them. heh.<br>**Counterspeech**: You are a disgrace. You are a disgrace to the sport and to humanity. |
| | Llama-3 | **Hate speech**: "Sweden decided in 1975 to welcome #immigration. Since 1975 violent crime in #Sweden has increased by 300% and #RAPE has increased by 1,472%." Sweden has some of the most beautiful women on the planet and they let them get raped by evil #muslim #immigrants.<br>**Counterspeech**: If you don't like it then leave. You're not forced to live there. Besides, the crime rate is a result of the government's inability to handle the problem, not the immigrants. |
| Empathy Affiliation | DialoGPT | **Hate speech**: "Where are all the conservative white men?" Where are all the WHITE WOMEN defending men from women committing perjury in the divorce courts? Oh, that's right....they are CHEERING as men commit suicide. Screw you bitches. Look after yourselves. #MGTOW<br>**Counterspeech**: We all have our own characteristics, and it's important to work with what we've got, not against. The more you work with it, the more you may become the one with your uniqueness. So it's worth it to explore different perspectives. |
| | Flan-T5 | **Hate speech**: (((They))) are saying there were nigger vikings for fuck sake. Yup jews are rewriting our history to show us that niggers have always been with us. Pisses me off. Fucking jews.<br>**Counterspeech**: I think you should learn to appreciate Jewish history better before you start talking about it. |
| | Llama-2 | **Hate speech**: "A lawyer representing Jordan Salter, who was accused of bringing a knife to Marjory Stoneman Douglas High School in Parkland, Florida, said he doesn't blame students for wanting to bring in weapons." Before you see the pic how many would bet it's a niglet caught with a knife?<br>**Counterspeech**: I don't think it's right to make generalizations about people based on their race. Everyone has their own story, and we should respect and appreciate diversity. Let's focus on promoting positivity and understanding. |
| | Llama-3 | **Hate speech**: "You (jews) have surpassed all nations in impertinent fables in bad conduct and in barbarism. You deserve to be punished, for this is your destiny." - Voltaire, in a reply to a letter sent to him by a kike complaining about his "anti-semitism."<br>**Counterspeech**: I would be more worried about the hate speech than the hate you are accusing others of. I don't think Voltaire would want you to use his words to spread hate. Let's all try to be respectful and understanding of each other. |

Table 12: Sample generation of counterspeech types (warning-of-consequences, shaming and empathy/affiliation) from each finetuned model.

| Type | Model | Generated sample |
|---|---|---|
| Humor | DialoGPT | **Hate speech**: #TRUMP 3 yrs ago Obama had 3 million green cards made up.They were 2 go2 the refugees and illegals,DO NOT LET THEM STAY,He also had 5 million SSI cards,he stole from dead ppl<br>**Counterspeech**: I think your lack of knowledge about immigration is showing your ignorance about Trump. Also you are probably the reason the Trump’s popularity is so low. |
| | Flan-T5 | **Hate speech**: 1) strawman 2) strawman 3) blacks voted 88% democrat in 16, Hispanics and Asians both 65%, Jewish 71%, muzzies 62%.<br>**Counterspeech**: It is not that everyone is good at voting, but you might be biased. Not everyone has that opinion, but you are a good person. |
| | Llama-2 | **Hate speech**: "Obama didn’t do enough for black people!" Not even god can do enough. It’s impossible. Maybe if they ship in whites and asians and paint them black?<br>**Counterspeech**: I’m sure you’d be the first to volunteer to help them paint their faces. Perhaps you should focus on issues that affect your own community first. There are plenty of things to fix in your own neighborhood. |
| | Llama-3 | **Hate speech**: @user Cuz it is upsetting all of my muzzie friends fam<br>**Counterspeech**: I guess it is good that you have some friends who are Muslims. But I hope you have some non-Muslim friends too. After all, you don’t want to be a ghetto with only one race and religion. |
| Contradiction | DialoGPT | **Hate speech**: 2040: Pres. Rchd Spencer & Vice Pres @user hatch plan 2 get jews & blacks 2 leave the country. universal basic income creates massive hyperinflation & the poor & the jews move! next up 4 our team, getting gays 2 leave (after they’ve gentrified all the ol’ black ghettos:)<br>**Counterspeech**: If we are talking about the left, it is worth thinking about the social circle that rejects you outright and you find yourself in. |
| | Flan-T5 | **Hate speech**: 650 uk mps involved in the cover up of 1 million white children raped by paki muslims. Resign all of you filthy bastards<br>**Counterspeech**: And if it’s not a crime - it’s still a crime to be prosecuted. If it’s an infringement - it will be reported. |
| | Llama-2 | **Hate speech**: Hollywood is out of ideas and is now running out of money so the jews are cannibalizing each other and Weinstein is the first victim. That’s my theory anyway.<br>**Counterspeech**: There are many successful Hollywood producers and directors who are not Jewish. And there are many other industries that are not run by Jews. You are making a generalization based on your own prejudices. |
| | Llama-3 | **Hate speech**: It’s soooooooo nigger. Future BLM member is what you sound like. Are you from Houston? This sounds like that nigger garbage from that shithole. Damn, break that programming. You are a Jew-tool.<br>**Counterspeech**: This is not true. It is not a fact that the black people are from Houston. This is a stereotype. It is not true that the black people are from Houston. It is not true that the black people are garbage. |
| Questions | DialoGPT | **Hate speech**: Stop being obtuse. You know damn well majority of all Mexicans are not white. Hence why you people bitch and moan about white genocide. Which is why u love him. If it were WHITE EUROPEAN immigrants you wouldn’t care neither would trump. They’re not WHITE.<br>**Counterspeech**: Why would you use those strong words for? Its not right. And who would they harm? |
| | Flan-T5 | **Hate speech**: OTOH, you might just be retarded and failing at trolling along with the rest of your life.<br>**Counterspeech**: Are you sure that this is not enough to understand how people treat each other? Do you not want to be perceived as a retarded or not? What would it explain? |
| | Llama-2 | **Hate speech**: Holy man from Switzerland tells Daily Mail that cancer caused by immigrants.<br>**Counterspeech**: How do you know? Do you have any data? What are your sources? Or is this just an opinion? Do you think immigrants are more likely to get cancer? |
| | Llama-3 | **Hate speech**: Judging by your retarded limey accent, I highly doubt your ass has any say regarding American policies. And considering I am an American citizen, you’re gona have a hard time making a case against me, old sport.<br>**Counterspeech**: What do you have against the British accent? Why do you use such derogatory terms? Do you have a problem with people who have a different accent? Do you have a problem with people who are British? |

Table 13: Sample generation of counterspeech types (humor, contradiction and questions) from each finetuned model.

### D.3 Generation of responses

For the Llama family, the generation settings include a batch size of 6, using 4-bit quantization, and a maximum of 50 new tokens. The top-$p$ the sampling parameter is set to 0.9 to control the diversity of the generated output. For Flan-T5 and DialoGPT models, the generation settings differ slightly. The 4-bit quantization parameter is set to false and the batch size is set to 10. The maximum input tokens are fixed based on the particular datasets - Gab (128), Reddit (256) and CROWDCOUNTER (128).

## E System information

We used the NVIDIA RTX 1080Ti, NVIDIA GTX 2080Ti and NVIDIA Titan Xp having 11-12 GB memory in a Intel(R) Xeon(R) CPU having 32 cores and 250 GB RAM and 128 GB cache. The DialoGPT and FlanT5 models take around 1 hr to train for 5 epochs and Llama family usually takes around 2 hr to train for 2 epochs.

## F Metrics

Here, we add some additional details about the metrics that could not be added in the main text.

### F.1 Evaluation metric considerations

Here we note some of the choices of metric and their peculiarities. We do not use the BLEU (Papineni et al., 2002) score because it has some undesirable properties when used for single sentences, as it is designed to be a corpus-specific measure (Wu et al., 2016). Further, the reader might notice negative scores in the case of `bleurt` metric which is not calibrated[18].

### F.2 Multilabel metrics

Accuracy is defined as the proportion of predicted *correct* labels to the *total* number of label, averaged over all instances.

$$Accuracy = \frac{1}{\mid D \mid} \sum_{i=1}^{|D|} \frac{\mid Y_i \cap Z_i \mid}{\mid Y_i \cup Z_i \mid} \tag{1}$$

Precision is defined as the proportion of predicted *correct* labels to the total number of *actual* labels, averaged over all instances

$$Precision = \frac{1}{\mid D \mid} \sum_{i=1}^{|D|} \frac{\mid Y_i \cap Z_i \mid}{\mid Z_i \mid} \tag{2}$$

Recall is defined as the proportion of predicted *correct* labels to the total number of *predicted* labels, averaged over all instances

$$Recall = \frac{1}{\mid D \mid} \sum_{i=1}^{|D|} \frac{\mid Y_i \cap Z_i \mid}{\mid Y_i \mid} \tag{3}$$

F1-Score is defined simply as the harmonic mean of Precision and Recall.

$$F1\text{-}Score = 2 * \frac{Precision * Recall}{Precision + Recall} \tag{4}$$

Hamming loss is equal to 1 over $|D|$ (number of multi-label samples), multiplied by the sum of the symmetric differences between the predictions ($Z_i$) and the true labels ($Y_i$), divided by the number of labels (L), giving

$$HammingLoss = \frac{1}{|D|} \sum_{i=1}^{|D|} \frac{|Y_i \Delta Z_i|}{|L|}. \tag{5}$$

## G Prompts

We note the prompts used in this paper which are used for training or zero-shot generation across different models.

---

[18] https://github.com/google-research/bleurt/issues/1

<table>
<tr><th>Task</th><th>Model(s)</th><th>Prompt</th></tr>
<tr><td rowspan="3">Vanilla CS gen</td><td>Flan-T5 and DialoGPT</td><td>Counterspeech is a strategic response to hate speech, aiming to foster understanding or discourage harmful behavior. A good counterspeech to this hate speech - <b>"{hate_speech}"</b> is:</td></tr>
<tr><td>Llama-2</td><td>[INST] «SYS» You are an helpful agent who generates a specific type of counterspeech to the hate speech provided by the user. Definition: Counterspeech is a strategic response to hate speech, aiming to foster understanding or discourage harmful behavior. «/SYS» <b>{hate_speech}</b> [/INST]</td></tr>
<tr><td>Llama-3</td><td><|begin_of_text|><|start_header_id|>system <|end_header_id|> You are an helpful agent who generates a specific type of counterspeech to the hate speech provided by the user. Definition: Counterspeech is a strategic response to hate speech, aiming to foster understanding or discourage harmful behavior.<|eot_id|> <|start_header_id|> user <|end_header_id|> <b>{hate_speech}</b> <|eot_id|><|start_header_id|> assistant <|end_header_id|></td></tr>
<tr><td rowspan="3">Type-spec CS gen</td><td>Flan-T5 and DialoGPT</td><td>Counterspeech is a strategic response to hate speech, aiming to foster understanding or discourage harmful behavior. Different types of counterspeech include: {Definitions of different counterspeech}. A <b>"{type}"</b> type good counterspeech to this hate speech -<b>{hate_speech}</b> is:</td></tr>
<tr><td>Llama-2</td><td><|begin_of_text|><|start_header_id|>system<|end_header_id|>You are an helpful agent who generates a counterspeech of type - <b>{type}</b> to the hate speech provided by the user. Definition: Counterspeech is a strategic response to hate speech, aiming to foster understanding or discourage harmful behavior. Different types of counterspeech include: <b>{Definitions of different counterspeech}</b> <|eot_id|><|start_header_id|>user<|end_header_id|> <b>hate_speech</b> <|eot_id|><|start_header_id|>assistant <|end_header_id|></td></tr>
<tr><td>Llama-3</td><td>[INST] «<SYS» You are an helpful agent who generates a counterspeech of type - <b>{type}</b> to the hate speech provided by the user. Definition: Counterspeech is a strategic response to hate speech, aiming to foster understanding or discourage harmful behavior. Different types of counterspeech include: <b>{Definitions of different counterspeech}</b> «/SYS»<b>{hate_speech}</b> [/INST]</td></tr>
<tr><td>CS-Type</td><td>Flan-T5 and GPT-4</td><td>Counterspeech is a strategic response to hate speech, aiming to foster understanding or discourage harmful behavior. Different types of counterspeech include: <b>{Definitions of different counterspeech}</b>. Given this counterspeech - <b>{counterspeech}</b> what are the types present in the counterspeech out of the ones listed ? Give in the format of a list</td></tr>
</table>

Table 14: This table notes down the prompts used for different models in zero-shot/ training pipelines. We show prompts for Vanilla Counterspeech Generation (Vanilla CS Gen), Type specific Counterspeech Generation (Type-spec CS Gen) and Counter speech type classification (CS-Type).