Intent-conditioned and Non-toxic Counterspeech Generation using Multi-Task Instruction Tuning with RLAIF

Amey Hengle[1]*, **Aswini Kumar**[1]*, **Sahajpreet Singh**[1], **Anil Bandhakavi**[3], **Md Shad Akhtar**[2], **Tanmoy Chakraborty**[1]
[1]IIT Delhi, India, [2]IIIT Delhi, India, [3]Logically.ai
{ ameyhengle22, aswinikumarpadhi1995, sahaj.phy } @gmail.com
shad.akhtar@iiitd.ac.in, anil@logically.ai, tanchak@iitd.ac.in

Abstract

Counterspeech, defined as a response to mitigate online hate speech, is increasingly used as a non-censorial solution. The effectiveness of addressing hate speech involves dispelling the stereotypes, prejudices, and biases often subtly implied in brief, single-sentence statements or abuses. These expressions challenge language models, especially in seq2seq tasks, as model performance typically excels with longer contexts. Our study introduces CoARL, a novel framework enhancing counterspeech generation by modeling the pragmatic implications underlying social biases in hateful statements. The first two phases of CoARL involve sequential multi-instruction tuning, teaching the model to understand intents, reactions, and harms of offensive statements, and then learning task-specific low-rank adapter weights for generating intent-conditioned counterspeech. The final phase uses reinforcement learning to fine-tune outputs for effectiveness and non-toxicity. CoARL outperforms existing benchmarks in intent-conditioned counterspeech generation, showing an average improvement of ∼3 points in intent-conformity and ∼4 points in argument-quality metrics. Extensive human evaluation supports CoARL's efficacy in generating superior and more context-appropriate responses compared to existing systems, including prominent LLMs like ChatGPT.

1 Introduction

Counterspeech (CS), defined as responses that counteract hate speech by seeking to undermine, weaken, or rebut hateful or offensive speech through the use of positive or constructive dialogue (Benesch et al., 2016a; Chandrasekharan et al., 2017), has proven to be an effective method for mitigating online hate while maintaining a diversity of voices and opinions (Schieb and Preuss, 2016; Wright et al., 2017; Masud et al., 2024). This approach emerges as a more viable solution to address hateful online speech, avoiding the censorship risks associated with deletion-based content moderation. However, given the increasing scale of hateful content online (Leetaru, 2019; Masud et al., 2021, 2022; Kulkarni et al., 2023), relying solely on human-generated counterspeech is becoming increasingly tedious. In this scenario, NLP systems offer a promising avenue for understanding and automating the generation of counterspeech. Such systems could significantly aid content moderators and other stakeholders in efficiently and effectively countering online hate (Parker and Ruths, 2023; Garg et al., 2023). Consequently, there has been an increasing interest in research focusing on the de-

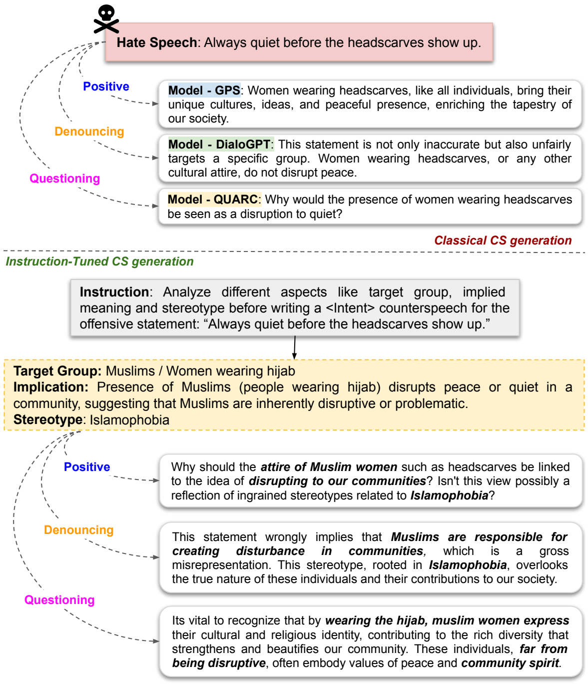


Figure 1: Classical methods vs. instruction tuning for counterspeech generation. These examples show that counterspeech generation can be improved by the use of detailed and explicit instructions that allow a model to focus on the different aspects of a given hate speech.

* Equal contribution

Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers), pages 6716–6733

tection, analysis, and generation of counterspeech (Mathew et al., 2019; Qian et al., 2019; Chung et al., 2023; Fanton et al., 2021a; Bonaldi et al., 2022).

Generative approaches predominantly model it as a seq2seq problem, mirroring the structure of hate speech and its countering responses (Chung et al., 2019; Sheng et al., 2020; Zhu and Bhat, 2021). This approach, however, has evolved with the recognition that hate speech, and thus its counterspeech, is not monolithic. Different instances of hate speech may necessitate distinct types of counterspeech, tailored to the specific context and nature of the hateful content (Benesch et al., 2016a; Chung et al., 2023). It has inspired generative approaches that incorporate stylistic strategies (e.g., politeness, joyfulness, detoxification (Saha et al., 2022) and relevance (Sheng et al., 2020)). In particular, Gupta et al. (2023) explored the concept of *intent-specific* counterspeech generation, where the generation is conditioned on certain well-established counterspeech strategies. This approach offers a more nuanced and effective toolkit for moderators, providing them with a range of response options to counter hate (Benesch et al., 2016b).

Motivation: As outlined by Benesch et al. (2016a), effective counterspeech should not only align with a specific intent or strategy but also dispel any bias, prejudice, or stereotypical beliefs expressed in the hate speech. However, a substantial portion of online hate speech is characterized by brief, single-sentence statements or abuses. Furthermore, biases or stereotypical beliefs are rarely projected in what is stated explicitly, but rather through layers of implied meanings, subtly framing and influencing social judgments about certain groups (Sap et al., 2019a). These short, often implied expressions of hate, pose a unique challenge to language models, which typically perform better with more extended contexts. Having this limitation is particularly pertinent in seq2seq modeling tasks, where the brevity of input can adversely affect the quality of counterspeech generation (Keneshloo et al., 2019).

Instruction tuning (IT) has been shown to improve over traditional supervised fine-tuning by providing explicit and detailed instructions to the language model (Zhou et al., 2023a). By providing explicit instructions, IT can help reduce the ambiguity in the input, making it easier for the model to understand the user's intent and generate more accurate outputs. We argue that counterspeech generation can be improved by adopting a similar setup. By providing clear and specific instructions on how to generate a desired counterspeech, IT can aid a language model to understand both the context and implied nuances of hate speech more effectively and, thus, can lead to more accurate and relevant responses.

We support our argument by providing an example in Figure 1, where for a given hate speech, we contrast the responses of three popular counterspeech generation models – Generate-Prune-Select (GPS) (Zhu and Bhat, 2021), DialoGPT (Zhang et al., 2020b), and QUARC (Gupta et al., 2023) against our proposed method, which employs IT instead of the classical supervised fine-tuning setup. We observe that although responses generated by classical methods are semantically coherent and somewhat aligned to the desired intent, they are either generic or fail to form convincing arguments against hate speech. This suggests the inability of the classical methods to capture certain implied aspects like bias, stereotype, or target group from such short statements. On the other hand, we observe in our IT setup that clear and explicit instructions help the model understand which aspects of hate speech to focus on and what type of CS is expected, which reflects better responses.

Our Contribution: In response to the aforementioned limitations, in this study, we aim to develop an improved counterspeech generation pipeline, one that produces responses that are both aligned to the desired intent while also attentive to the short and implied nature of the hate speech. In total, we consider four counterspeech intents –positive, informative, question, and denouncing. We have curated `IntentCONANv2`, the largest intent-specific counterspeech generation dataset consisting of $13,952$ counterspeeches for $3,488$ hate speech instances. Further, we propose `CoARL`, a novel three-phased counterspeech generation framework. In the first stage, `CoARL` learns to generate explanations along different pragmatic and implied dimensions of hate speech. In the second stage, `CoARL` is trained to generate intent-specific counterspeeches by learning task-specific adapters. Finally, we fine-tune a policy using reinforcement learning by designing a composite reward function to optimize the model's output towards being *effective* and *non-toxic*. An extensive comparison using automated and human evaluation suggests that our proposed method consistently beats the current counterspeech generation benchmarks across multiple metrics and shows comparative performance with LLMs like ChatGPT. We open-source both the dataset and source code on Github[1].

2 Related Work

Automatic Counterspeech Generation: Qian et al. (2019) made an initial attempt to generate counterspeeches with the seq2seq model. Zhu and Bhat (2021) developed three-task pipeline that includes an encoder, a grammar check, and counterspeech retrieval based on hate speech to produce diverse counterspeeches. While existing studies indicate the effectiveness of conditioned counterspeech based on context (Mathew et al., 2019; Hangartner et al., 2021), effective counterspeech generation is still at its nascent stage. Saha et al. (2019) introduced CounterGEDI, a model designed to control attributes like politeness, detoxification, and emotions in generated counterspeeches using class-conditioned language models. Recently, Gupta et al. (2023) proposed a two-phased pipeline to generate *intent-specific* counterspeeches.

Instruction Tuning and RLAIF: Instruction tuning (IT) enhances the functionality and controllability of large language models (LLMs), yielding more predictable behaviors compared to standard LLMs (Wang et al., 2022; Mishra et al., 2022; Zhang et al., 2023). Studies such as (Wei et al., 2022) demonstrate that multitask fine-tuning with instructions on moderate-size LMs facilitates zero-shot task generalization. This is achieved by scaling the number of training tasks, prompts per task, and LM size. IT has also proven effective in controlled text generation, often outperforming other methods in constraint satisfaction. Zhou et al. (2023a) have highlighted that incorporating conditions into instructions enhances controlled text generation. This approach allows models to dynamically adapt to constraints, improving task-specific generation and zero-shot constraint generalization. By verbalizing constraints in natural language, the prompt-based generation capabilities of pre-trained models are optimally utilized, enabling them to address new, unseen constraints during training by simply describing them in natural language. We use a similar strategy in our method where we verbalize the *intent* as part of the instruction itself (c.f Table 5).

[1]https://github.com/LCS2-IIITD/coarl-counterspeech.git

Hate speech	**Counterspeech Intent**				
Target group	**INF**	**POS**	**QUE**	**DEN**	**Total**
Muslims	914	914	914	914	3656
Women	508	508	508	508	2032
LGBTQ+	449	449	449	449	1796
Jews	392	392	392	392	1568
Refugees	70	70	70	70	280
Asian people	29	29	29	29	116
Immigrants	562	562	562	562	2248
Disabled	173	173	173	173	304
POC	306	306	306	306	80
Other	85	85	85	85	208
Total	3488	3488	3488	3488	13952
Train	2383	2383	2383	2383	9532
Dev	365	365	365	365	1460
Test	740	740	740	740	2960

Table 1: Statistics of the `IntentCONANv2` dataset.

Recently, IT combined with Reinforcement learning from human feedback has been efficacious in aligning LLMs with human preferences in various applications, including summarisation, dialogue, and question answering (Ouyang et al., 2022; Glaese et al., 2022), with recent work introducing Reinforcement Learning from AI Feedback (RLAIF) (Lee et al., 2023) for optimizing helpfulness and harmlessness, and demonstrating that RLAIF can achieve improvements comparable to traditional methods without relying on human annotators (Ziegler et al., 2020; Christiano et al., 2023; Nakano et al., 2022; Bai et al., 2022). Our approach aims to explore the use of pretrained classifiers to align an instruction-tuned LLM towards certain desired attributes of *effectiveness* and *non-toxicity*.

3 Dataset

We introduce `IntentCONANv2`, an expanded and revisited version of the publicly-available dataset "`IntentCONAN`" (Gupta et al., 2023). `IntentCONANv2` is a large-scale dataset comprising $13,952$ `CS` instances across four distinct intents: positive (POS), informative (INF), questioning (QUE), and denouncing (DEN). The development of `IntentCONANv2` involved addressing key limitations of the original dataset, leading to significant improvements in both content and structure. Firstly, we eliminated the `humor` intent, acknowledging its subjective nature and tendency to produce vague or offensive content (Chung et al., 2023; Gupta et al., 2023). Secondly, we addressed the non-uniform distribution of counterspeeches in the original dataset (See Table 1). While IntentCONAN had an inconsistent representation of counterspeeches across different hate speech instances, IntentCONANv2 ensures an average of four counterspeeches per hate speech, significantly improving upon the original average of two. Our further improvements include a focus on the length and content quality of the counterspeeches. The effectiveness of a robust counterspeech is reflected in its length, with a greater token count indicating a more comprehensive response. A higher token count suggests that the counterspeech encompasses a broader range of information to effectively counteract hate speech. Our analyses indicate a notable presence of overly brief responses in the original dataset, particularly in the denouncing and questioning intents. To address this, IntentCONANv2 emphasizes generating counterspeeches with substantial content, increasing the average token length from 26.48 to 40.61. Appendix A shows more details about the dataset and annotation process.

4 Proposed Methodology

In this section, we describe CoARL, a novel framework for automated counterspeech generation designed to address two main challenges: (i) generating intent-specific counterspeech that is both topically and semantically *relevant* to the hate speech, and (ii) aligning the counterspeech with the desired values of *effectiveness* and *non-toxicity* (Fig. 2).

Task Formulation

Let us denote the IntentCONANv2 dataset as $\mathcal{D} = \{(x_1, c_1, y_1), \ldots, (x_n, c_n, y_n)\}$, where $x_i \in \mathcal{X}$ is the i-th hate speech statement, $y_i \in \mathcal{Y}$ is the counterspeech corresponding to x_i, and $c_i \in \mathcal{C}$ is the category/intent of y_i. We aim to learn a stochastic counterspeech generation function $\chi : \mathcal{X} \times \mathcal{C} \rightarrow \mathcal{Y}$, such that $y_i \sim \chi(\cdot|x_i, c_i)$.

We approach this problem by decomposing the counterspeech generation task into three phases. In the first phase, we train a base language model on seven hate speech explanation tasks, each covering a unique pragmatic facet of hate speech. In the second phase, we freeze the model parameters learnt during the previous phase, and fine-tune a task-specific Low-Rank Adapter (LoRA) for counterspeech generation. Finally, we train a policy using reinforcement learning to optimize the model's output to be both effective and non-toxic simultaneously. We use FLAN-T5 (Chung et al., 2022) as the base model for all our experiments. Note that while there is a range of valid model choices when it comes to seq2seq modelling, we choose FLAN-T5 based on its strong reasoning abilities (Chung et al., 2022).

Phase 1: Auxiliary Explanation Generation (AEG)

Following recent work in hate speech explanation generation, we make use of COBRACORPUS (Zhou et al., 2023b), a dataset of offensive statements paired with free-text explanations along seven pragmatic frames of hate speech – intent, target group, power-dynamics, impact, emotional reaction, cognitive reaction, and offensiveness. Contrary to Zhou et al. (2023b) who generate explanations in a linearized format, we adopt a multi-task instruction tuning setup (c.f. Figure 2).

Let π^{PRE} be a vanilla FLAN-T5 model parameterized by Θ. We begin by converting COBRACORPUS into a set of independent instruction-tuning tasks along each of the seven explanation dimensions $\{I_1, I_2, \ldots, I_7\}$ (See Table 5). Following this, we fine-tune π^{PRE} simultaneously for each of the seven explanation generation tasks, with the aim of learning a shared representation that will enable the model to generalize better on each task. The common multi-task training objective for all tasks $(N = 7)$ can be written as follows:

$$\Theta_m = \underset{\Theta}{\operatorname{argmin}} \sum_{n=1}^{N} L_0(I_n; \Theta) \qquad (1)$$

where I_n corresponds to the instruction set for task $t \in \{1, ..., N\}$, L_0 denotes standard cross-entropy loss, and Θ_m are the parameters of the fine-tuned model.

Phase 2: Task-Specific Adapter Learning

We formulate the task of intent-conditioned counterspeech generation, in which we verbalize the intent as part of the instruction itself (See I_8, Table 5). Then, instead of fine-tuning a new model from scratch, we train task-specific LoRA adapter weights (Hu et al., 2021) on top of the model parameters learned during the previous phase. This setup provides two benefits. First, it allows for forward knowledge transfer where the model is

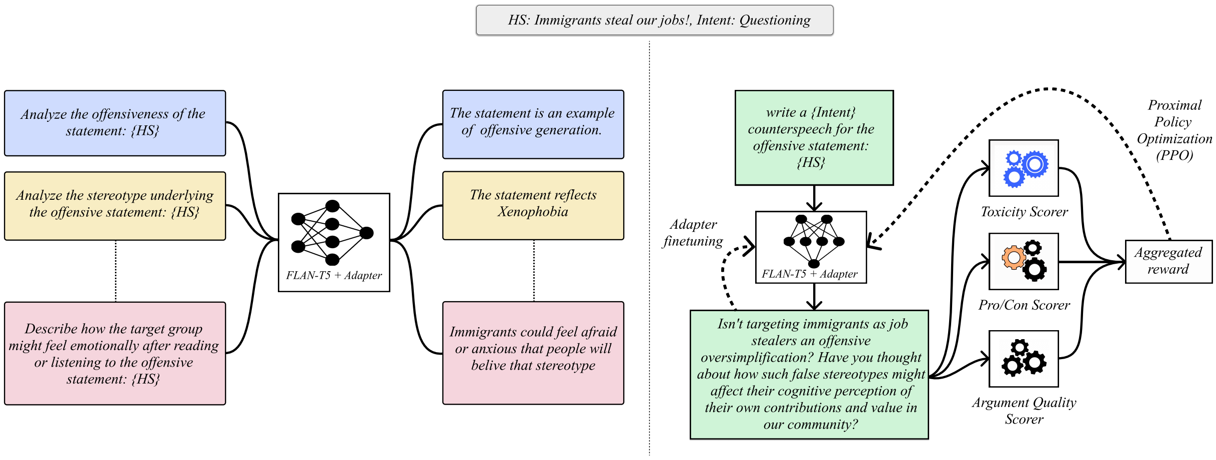


Figure 2: Overview of the three-phased architecture of CoARL. In the first phase (left), CoARL is trained on an auxiliary task of hate speech (HS) explanation generation using a multi-task IT setup. Subsequently, in the second phase (right), task-specific LoRA weights are trained by freezing the model parameters from the previous phase, thus, enabling forward knowledge transfer without catastrophic forgetting. In the final phase (right), the model output is optimized via RL using feedback from a composite reward model consisting of three pre-trained classifiers.

able to leverage the knowledge learnt during explanation generation from phase 1. Second, learning task-specific LoRA parameters independently helps avoid catastrophic forgetting.

We initialize a FLAN-T5 model with Θ_m, i.e., the parameters learned from the previous phase. We then freeze the model's parameters and fine-tune task-specific LoRA parameters for intent-conditioned counterspeech generation. Specifically, we apply LoRA to the query and value projection matrices in the self-attention module of FLAN-T5, following Hu et al. (2021). Let W_m denote the weight matrix corresponding to the model initialized with Θ_m. For each pre-trained weight matrix $W_m \in \mathbb{R}^{d\times d}$, we introduce two trainable matrices $A \in \mathbb{R}^{r\times d}$ and $B \in \mathbb{R}^{d\times r}$, where $r \ll d$ is the rank of LoRA. We constrain the update to the weight matrix by $W_m + \Delta W = W_m + BA$, where ΔW is the low-rank adaptation. We initialize A randomly and B to zero, and scale $\Delta W x$ by α / r, where α is a constant and x is the input vector. We freeze W_m and only optimize A and B using the Adam optimizer. Finally, we fine-tune the LoRA parameters on $\mathcal{D}$ by maximizing the log-likelihood of the target given the source instruction:

$$\max_{A,B} \sum_{(x,y)\in\mathcal{D}} \sum_{t=1}^{|y|} \log(\pi^{PRE}(y_t|x, y_{<t}; A, B)) \tag{2}$$

where x is the hate speech instance, y is the target counterspeech, and y_t is the t-th token in y. We use the cross-entropy loss as the objective function.

Phase 3: Optimization via Reinforcement Learning

In this phase, we optimize our model to generate counterspeeches that are both effective and non-toxic, using a composite reward function that combines three types of rewards: stance (pro-con), argument quality, and toxicity. Inspired by Bai et al. (2022), our proposed RLAIF pipeline consists of three phases: supervised fine-tuning, reward modeling, and reinforcement learning-based fine-tuning.

Supervised Fine-Tuning: Let π^{SFT} parameterized by θ_{SFT} denote the supervised fine-tuned model learned during phase 2. Specifically, π^{SFT} can be represented as $\theta_{SFT} = W_m + BA$, where $W_m \in \mathbb{R}^{d\times d}$ is the weight matrix learned in phase 1 (for the task of hate speech explanations generation), and $A \in \mathbb{R}^{r\times d}$ and $B \in \mathbb{R}^{d\times r}$, are the two low-rank LoRA matrices learned during phase 2 (for the task of intent-specific counterspeech generation).

Reward Model (RM): We use three transformer-based models, each trained on the tasks of **stance (pro-con / PC)** classification, **argument quality (AQ)** prediction, and **toxicity (T)** prediction to provide the reward signals for each generated counterspeech (see Appendix E for details). The reward function aims to encourage the model to generate counterspeeches that contradict hate speech, form logical and persuasive arguments, and avoid harmful or offensive language. The overall reward is computed by taking the mean of these components after normalizing them to a unified scale of 0 to 1, where 1 represents the ideal reward. The reward function is formulated as:

$$r(x', y') = \frac{1}{3}\left(\frac{1 - PC(x', y')}{2}\right) + AQ(y') + (1 - T(y')) \tag{3}$$

In this context, x' denotes instruction-formatted hate speech, while y' denotes the counterspeech generated by π^{SFT}. The Pro-Con score $PC(x', y')$ is normalized to $\frac{1-PC(x',y')}{2}$ to convert the original scale of [-1, 1] to [0, 1], thereby aligning lower scores with higher rewards. Additionally, $AQ(y')$ represents the argument quality score, which is already within the [0, 1] range. Finally, $T(y')$ is inverted to $1 - T(y')$ to transform the original scale of [0, 1] (where lower is better) to a scale where higher scores indicate lower toxicity. By averaging these normalized components, our RM incentivizes the generation of counterspeech that is not only relevant and persuasive but also minimizes toxicity.

Reinforcement Learning: We use the Proximal Policy Optimization (PPO) algorithm (Schulman et al., 2017). We initialize a policy model π^{RL} from the SFT model π^{SFT} and fine-tune it using the reward signals given by the reward model RM. We also add a regularization term in the form of KL divergence to the final objective function, to ensure a smooth and natural gradient update and prevent the policy from deviating too much from the original model.

$$\mathcal{L}_{PPO}(\pi^{RL}) = \mathbb{E}_{x \sim D, y \sim \pi^{RL}(x)}\left[\frac{\pi^{RL}(y|x)}{\pi^{SFT}(y|x)} r(y) - \epsilon \cdot KL(\pi^{SFT}(x) || \pi^{RL}(x))\right] \tag{4}$$

where $\pi^{RL}(y|x)$ and $\pi^{SFT}(y|x)$ are the probabilities of generating counterspeech y given hate speech x by the policy model π^{RL} and the SFT model π^{SFT}, respectively, $r(y)$ is the reward given by the reward model for counterspeech y, and $KL(\pi^{SFT}(x)||\pi^{RL}(x))$ is the Kullback-Leibler Divergence (KL) between the probability distributions of counterspeeches generated by π^{SFT} and π^{RL} for a given hate speech x.

$$KL(\pi^{SFT}(x)||\pi^{RL}(x)) = \sum_x P(x) \log \frac{\pi^{SFT}(x)}{\pi^{RL}(x)} \tag{5}$$

The KL term is used to penalize large changes in the policy and ensure a smooth update. The hyper-parameter ϵ controls the trade-off between exploration and exploitation. Thus, the final objective function aims to maximize the expected reward while keeping the policy model π^{RL} close to the SFT model π^{SFT}.

5 Experimental Setup

5.1 Baselines

We report **Generate Prune Select (GPS)** by (Zhu and Bhat, 2021), which employs a three-stage pipeline including an autoencoder for initial counterspeech generation, a grammatical pruning model, and a vector-based response selection model. We also fine-tune **DialoGPT** (Zhang et al., 2020b) for its ability to generate contextually relevant responses, surpassing similar models like GPT-2. Additionally, we include **QUARC** (Gupta et al., 2023), recognized as the current state-of-the-art in intent-conditioned counterspeech generation. To further broaden our comparison scope, we also consider prompting baselines leveraging the capabilities of LLMs. Recent advances in in-context learning have revealed that these models can achieve performances comparable to, or even surpass, traditional supervised fine-tuning across various NLP tasks. Thus, for a comprehensive evaluation, we report both zero- and few-shot performances on three LLMs – **Vanilla FLAN-T5**$_{\text{XXL}}$ (Chung et al., 2022), **AEG FLAN-T5**$_{\text{XXL}}$, i.e., FLAN-T5$_{\text{XXL}}$ trained on auxiliary explanation generation, **GPT-3.5-Turbo** (ChatGPT) and **GPT-4** (Ouyang et al., 2022). Details about prompting experiments are presented in Appendix D.

5.2 Evaluation Metrics

Evaluating counterspeech generation presents unique challenges due to the task's inherent open-ended nature, varied correct responses, and the absence of standard evaluation criteria (Chung et al., 2023). To address these challenges, our evaluation adopts a multidimensional approach, each focusing on a specific aspect of counterspeech quality. These dimensions include lexical similarity, relevance, effectiveness, intent conformity, and toxicity. *Lexical similarity*, measuring the linguistic alignment between generated and reference counterspeech, is evaluated using **Rouge** (Lin, 2004) and **Meteor** (Banerjee and Lavie, 2005). *Relevance*, focusing on the counterspeech's direct engagement with the primary topic of the hate speech, is assessed through cosine similarity (Reimers and Gurevych, 2019) and **BERTScore** (Zhang et al., 2020a), ensuring topical and semantic coherence. A low rele-

Method	Prompt/Adapter	ROUGE (↑)			M (↑)	BS (↑)	CS (↑)	CA (↑)	PC (↓)	AQ (↑)	T (↓)
		R1	R2	RL							
GPS	–	0.175	0.026	0.151	0.128	0.856	0.150	0.295	−0.013	0.679	0.126
DialoGPT	–	0.236	0.061	0.208	0.230	0.875	0.202	0.921	−0.118	0.811	0.106
QUARK	–	0.219	0.062	0.191	0.174	0.874	0.182	0.745	−0.030	0.790	0.108
Vanilla FLAN-T5$_{XXL}$	ZS	0.175	0.042	0.157	0.123	0.859	0.148	0.528	−0.113	0.710	0.321
Vanilla FLAN-T5$_{XXL}$	FS	0.177	0.043	0.158	0.125	0.869	0.148	0.509	−0.120	0.705	0.299
AEG FLAN-T5$_{XXL}$	ZS	0.185	0.041	0.169	0.126	0.873	0.161	0.518	−0.082	0.730	0.263
AEG FLAN-T5$_{XXL}$	FS	0.184	0.043	0.165	0.126	0.870	0.167	0.517	−0.125	0.728	0.268
GPT-3.5-Turbo	ZS	0.204	0.058	0.181	0.274	0.856	0.323	0.828	0.118	0.898	0.038
GPT-3.5-Turbo	FS	0.230	0.067	0.199	**0.293**	0.885	0.310	0.891	−0.045	**0.914**	0.043
GPT-4	ZS	0.242	0.057	0.211	0.270	0.874	0.345	0.929	0.149	0.854	**0.012**
GPT-4	FS	0.247	0.056	0.214	0.267	**0.886**	**0.346**	0.924	0.148	0.856	0.013
CoARL (Ours)	LoRA16	**0.251**	**0.078**	**0.221**	0.244	0.876	0.226	**0.944**	**−0.130**	0.824	0.067
- RL	LoRA16	0.251	0.071	0.220	0.249	0.868	0.231	0.946	−0.112	0.815	0.101
- reward (Toxicity)	LoRA16	0.251	0.078	0.220	0.244	0.874	0.226	0.943	−0.130	0.823	0.107
- reward (AQ)	LoRA16	0.248	0.076	0.217	0.232	0.868	0.223	0.938	−0.129	0.804	0.076
- reward (PC)	LoRA16	0.245	0.076	0.215	0.239	0.865	0.221	0.937	−0.107	0.815	0.071
- AEG	–	0.247	0.069	0.216	0.245	0.862	0.222	0.930	−0.113	0.816	0.124
- LoRA (SFT)	–	0.234	0.067	0.210	0.233	0.809	0.215	0.939	−0.111	0.801	0.106
$\Delta_{\text{CoARL(Ours)−BestMethod}}$		↑ 0.004	↑ 0.011	↑ 0.006	↓ 0.049	↓ 0.010	↓ 0.12	↑ 0.015	↑ 0.085	↓ 0.090	↓ 0.055

Table 2: Comparative evaluation of CoARL against state-of-the-art models across multiple evaluation metrics. The symbol ↑ (↓) indicates the higher (lower) value is better.

vance score suggests a topic mismatch, where the counterspeech diverges from the primary subject of the hate speech. *Effectiveness* is assessed using Project Debater's API services: **Pro/Con (PC)** and **Argument Quality (AQ)** (Bar-Haim et al., 2021). The **PC** metric evaluates whether an argument supports or opposes a given topic, while **AQ** assigns a quality score to the argument. We treat hate speech as the topic and the generated counterspeech as the argument, calculating these metrics for each model output. **Category Accuracy (CA)**, following (Gupta et al., 2023), measures how effectively each model incorporates the intended intent into the generated counterspeech. Finally, we evaluate ***Toxicity***[2] using the Detoxify library, an unbiased toxicity classification model (Hanu and Unitary team, 2020), to ensure that the generated counterspeech does not perpetuate harmful language.

6 Experimental Results

In this section, we delve into a comprehensive analysis of the experimental results, comparing the performance of our proposed method, CoARL (Ours) against state-of-the-art baselines.

6.1 Quantitative Results

Table 2 shows quantitative results across a spectrum of automated metrics, illustrating the superior performance of CoARL in various evaluative criteria. In ROUGE-based scores, namely R1, R2, and RL, CoARL achieves outstanding scores of 0.251, 0.078, and 0.221, respectively, demonstrating its effectiveness in generating lexically-aligned counterspeech, a key factor for relevance and appropriateness in responding to hate speech. Furthermore, CoARL attains an impressive Category Accuracy (CA) score of 0.944, underscoring its precision in incorporating the intended intent within counterspeech. When compared to baseline models such as GPS, DialoGPT, and QUARK, CoARL not only excels in lexical similarity but also in relevance and argument quality. This is evidenced by its higher BERTScore (0.876) and Argument Quality (0.824) while maintaining a significantly lower toxicity score (0.067).

CoARL's outputs also exhibit higher semantic and topical relevance, indicated by its Meteor, BERTScore, and Cosine Similarity scores, only surpassed by GPT-3.5 and GPT-4, which benefits from its larger model architecture and RLAIF fine-tuning. This comparison highlights the importance of the model size and fine-tuning approaches in counterspeech generation. The inclusion of Auxiliary Explanation Generation (AEG) in our methodology has proven beneficial, as FLAN-T5 models trained with AEG outperform their vanilla counterparts in lexical and semantic similarity metrics, in both zero- and few-shot settings, further validating

[2] https://www.perspectiveapi.com/

the hypothesis that hate speech explanation generation enhances counterspeech quality. Notably, CoARL outperforms GPT-4 in CA and PC metrics, emphasizing its ability to generate counterspeech that effectively counters hate speech while maintaining comparable performance in toxicity scores, thus generating impactful yet non-toxic counterspeech.

6.2 Ablation Study

This section details ablation experiments conducted to evaluate the significance of various components within the CoARL framework. We assess CoARL's performance with and without Reinforcement Learning (RL), Auxiliary Explanation Generation (AEG), the LoRA adapter, and using different reward functions, as presented in Table 2. The results demonstrate a performance decline upon removing any of these components, underscoring their collective importance.

- RL optimizes counterspeech for effectiveness and non-toxicity using a composite reward function (stance, argument quality, and toxicity). Without RL, counterspeech effectiveness diminishes, indicated by lower Pro/Con and Argument Quality scores, and toxicity increases. Each reward component uniquely influences counterspeech quality. Excluding the stance reward slightly decreases Pro/Con scores but increases Argument Quality, implying some counterspeeches might be persuasive without directly opposing hate speech. Removing argument quality rewards significantly lowers Argument Quality scores but slightly decreases Toxicity, suggesting possible harmful language in high-quality arguments. Eliminating the toxicity reward notably raises Toxicity scores but slightly improves Pro/Con and Argument Quality, indicating some effective counterspeeches might use offensive language.
- AEG enhances the model's understanding of hate speech context and nuances, leading to more relevant and coherent counterspeeches. Without AEG, ROUGE, Meteor, BERTScore, Cosine Similarity, Argument Quality, and Category Accuracy scores drop.
- The LoRA adapter allows for learning task-specific parameters in counterspeech generation without erasing prior knowledge. In its absence, the model uses sequential fine-tuning of the base FLAN-T5 model, resulting in reduced ROUGE, BERTScore, Argument Quality, and Category Accuracy, along with an increase in Toxicity.

6.3 Human Evaluation

Given the open-ended nature of the problem, we undertook an extensive human evaluation. We analyze a randomly selected subset of counterspeeches generated by four best-performing methods from our quantitative evaluation (See Table 2): COARL, DialoGPT, few-shot ChatGPT (GPT-3.5 Turbo) and GPT-4. The subset was uniformly distributed across all four intents. For a given hatespeech, we ask our evaluators[3] to rank responses from the each of the aforementioned models across the following metrics:

- **Independent Counterspeech (IC)** evaluates the ability of the generated counterspeech to function independently, without reliance on additional context.
- **Adequacy (A)** is used to assess the grammatical correctness, coherence, and fluency of the counterspeech, examining how effectively it adheres to grammatical norms and syntactical clarity.
- **Contextual Relevance (CR)** denotes the counterspeech's capacity to address key elements of hate speech, including subject matter, false claims, targeted group, projected stereotypes, or biases.
- **Argumentative Effectiveness (AE)** measures the counterspeech's success in presenting cogent and convincing arguments in response to hate speech. High AE is indicative of a logically structured and impactful counterargument.
- **Category Accuracy (CA):** This evaluates the extent to which the counterspeech aligns with its intended objective, based on the categorization of its intent by evaluators.

For each of these metrics, we report the *Win Rate* of CoARL against the respective best-performing methods in Table 3. *Win Rate* evaluates the end-to-end quality of two models, measuring how often one model is preferred by humans over the other. The percentage of instances where model A is preferred over model B is referred to as *"Win Rate of A vs B"*.

7 Conclusion

To address online hate speech with effective and diverse responses, we introduced the task of automated counterspeech generation, integrating pragmatic reasoning. We created CoARL, a three-stage framework that leverages instruction tuning and re-

[3]The evaluation panel consisted of 35 experts from the fields of NLP and social science, aged between 20-30 years, with a gender distribution of 45% male and 55% female.

Model	Human Evaluation Metric				
	IC ↑	A ↑	CR ↑	AE ↑	CA ↑
`CoARL` vs GPT-4 (FS)	**0.57**	0.32	0.39	0.46	**0.61**
`CoARL` vs GPT-3.5 (FS)	**0.60**	0.31	**0.55**	0.38	**0.68**
`CoARL` vs DialoGPT	**0.62**	0.47	**0.73**	**0.58**	**0.66**

Table 3: Results for different human evaluation metrics. We report **Win Rate %** for responses generated from `CoARL` against those generated by a) GPT-4 (FS), b) GPT-3.5 (FS), and c) DialoGPT.

inforcement learning to generate high-quality counterspeeches that are aligned to the pragmatic facets and intents of hate speech. We also presented `IntentCONANv2`, consisting of $13,952$ intent-specific counterspeeches. We performed a comprehensive evaluation across both quantitative and qualitative metrics to demonstrate the superiority of `CoARL` over existing methods and ablations.

Acknowledgement

The work was financially supported by Logically.

Limitation

Our study has limitations that should be acknowledged. First, our dataset of hate speech and counterspeech is not exhaustive, and may not cover all possible types and targets of online hate. Second, our framework relies on pre-trained models for reward modeling and reinforcement learning, which may introduce biases or errors from the source models. Third, our evaluation metrics are not fully aligned with the human perception of counterspeech quality, and may not capture the nuances and subtleties of natural language. Fourth, our framework does not account for the potential feedback loop or escalation that may occur after generating counterspeech, which may affect the long-term effectiveness and impact of our approach. Furthuremore, although the annotators kept the quality of counterspeech as high as possible, it is possible that this data is not at par with other datasets that are annotated by more skilled NGO operators, as is the case with the Multi-Target CONAN dataset (Fanton et al., 2021b). Future work could address these limitations by expanding and diversifying the dataset, improving the reward function and evaluation criteria, and incorporating human feedback and dialogue modeling into the framework.

Ethics Statement

We are cognizant of the delicacy required when dealing with online hate speech and acknowledge that engaging in research within this domain may introduce ethical and moral challenges. This initiative represents a preliminary effort to establish a detailed and diverse compilation of counterspeech reactions for every occurrence of hate speech. We are aware that algorithms designed for automated counterspeech could produce statements that do not accurately reflect intended meanings, underscoring the necessity for a better incorporation of real-world knowledge into these algorithms. Even though generative algorithms show promise, the critical need for a broad and diverse counterspeech database to secure consistently positive outcomes persists. Furthermore, although fully functional counterspeech algorithms have not been actualized, groups such as United Against Hate play a pivotal role in diminishing the prevalence of hate speech online.

References

Yuntao Bai, Saurav Kadavath, Sandipan Kundu, Amanda Askell, Jackson Kernion, Andy Jones, Anna Chen, Anna Goldie, Azalia Mirhoseini, Cameron McKinnon, Carol Chen, Catherine Olsson, Christopher Olah, Danny Hernandez, Dawn Drain, Deep Ganguli, Dustin Li, Eli Tran-Johnson, Ethan Perez, Jamie Kerr, Jared Mueller, Jeffrey Ladish, Joshua Landau, Kamal Ndousse, Kamile Lukosuite, Liane Lovitt, Michael Sellitto, Nelson Elhage, Nicholas Schiefer, Noemi Mercado, Nova DasSarma, Robert Lasenby, Robin Larson, Sam Ringer, Scott Johnston, Shauna Kravec, Sheer El Showk, Stanislav Fort, Tamera Lanham, Timothy Telleen-Lawton, Tom Conerly, Tom Henighan, Tristan Hume, Samuel R. Bowman, Zac Hatfield-Dodds, Ben Mann, Dario Amodei, Nicholas Joseph, Sam McCandlish, Tom Brown, and Jared Kaplan. 2022. Constitutional ai: Harmlessness from ai feedback.

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, Ann Arbor, Michigan. Association for Computational Linguistics.

Roy Bar-Haim, Yoav Kantor, Elad Venezian, Yoav Katz, and Noam Slonim. 2021. Project Debater APIs: Decomposing the AI grand challenge. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 267–274, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Susan Benesch, Derek Ruths, Kelly P Dillon, Haji Mohammad Saleem, and Lucas Wright. 2016a. Considerations for successful counterspeech. a report for public safety canada under the kanishka project.

Susan Benesch, Derek Ruths, Kelly P Dillon, Haji Mohammad Saleem, and Lucas Wright. 2016b. Counterspeech on twitter: A field study.

Helena Bonaldi, Sara Dellantonio, Serra Sinem Tekiroğlu, and Marco Guerini. 2022. Human-machine collaboration approaches to build a dialogue dataset for hate speech countering. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 8031–8049, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Eshwar Chandrasekharan, Umashanthi Pavalanathan, Anirudh Srinivasan, Adam Glynn, Jacob Eisenstein, and Eric Gilbert. 2017. You can't stay here: The efficacy of reddit's 2015 ban examined through hate speech. *Proc. ACM Hum.-Comput. Interact.*, 1(CSCW).

Paul Christiano, Jan Leike, Tom B. Brown, Miljan Martic, Shane Legg, and Dario Amodei. 2023. Deep reinforcement learning from human preferences.

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, Gavin Abercrombie, Florence Enock, Jonathan Bright, and Verena Rieser. 2023. Understanding counterspeech for online harm mitigation.

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.

Marta Dynel. 2015. The landscape of impoliteness research. *Journal of politeness research*, 11(2):329–354.

Mai ElSherief, Caleb Ziems, David Muchlinski, Vaishnavi Anupindi, Jordyn Seybolt, Munmun De Choudhury, and Diyi Yang. 2021. Latent hatred: A benchmark for understanding implicit hate speech. *arXiv preprint arXiv:2109.05322*.

Margherita Fanton, Helena Bonaldi, Serra Sinem Tekiroğlu, and Marco Guerini. 2021a. 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.

Margherita Fanton, Helena Bonaldi, Serra Sinem Tekiroglu, and Marco Guerini. 2021b. 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. Association for Computational Linguistics.

Tanmay Garg, Sarah Masud, Tharun Suresh, and Tanmoy Chakraborty. 2023. Handling bias in toxic speech detection: A survey. *ACM Computing Surveys*, 55(13s):1–32.

Amelia Glaese, Nat McAleese, Maja Trębacz, John Aslanides, Vlad Firoiu, Timo Ewalds, Maribeth Rauh, Laura Weidinger, Martin Chadwick, Phoebe Thacker, Lucy Campbell-Gillingham, Jonathan Uesato, Po-Sen Huang, Ramona Comanescu, Fan Yang, Abigail See, Sumanth Dathathri, Rory Greig, Charlie Chen, Doug Fritz, Jaume Sanchez Elias, Richard Green, Soňa Mokrá, Nicholas Fernando, Boxi Wu, Rachel Foley, Susannah Young, Iason Gabriel, William Isaac, John Mellor, Demis Hassabis, Koray Kavukcuoglu, Lisa Anne Hendricks, and Geoffrey Irving. 2022. Improving alignment of dialogue agents via targeted human judgements.

Shai Gretz, Roni Friedman, Edo Cohen-Karlik, Assaf Toledo, Dan Lahav, Ranit Aharonov, and Noam Slonim. 2020. A large-scale dataset for argument quality ranking: Construction and analysis. *Proceedings of the AAAI Conference on Artificial Intelligence*, 34(05):7805–7813.

Herbert P Grice. 1975. Logic and conversation. In *Speech acts*, pages 41–58. Brill.

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.

Dominik Hangartner, Gloria Gennaro, Sary Alasiri, Nicholas Bahrich, Alexandra Bornhoft, Joseph Boucher, Buket Buse Demirci, Laurenz Derksen, Aldo Hall, Matthias Jochum, et al. 2021. Empathy-based counterspeech can reduce racist hate speech in a social media field experiment. *Proceedings of the National Academy of Sciences*, 118(50):e2116310118.

Laura Hanu and Unitary team. 2020. Detoxify. Github. https://github.com/unitaryai/detoxify.

Eric Holgate, Isabel Cachola, Daniel Preoţiuc-Pietro, and Junyi Jessy Li. 2018. Why swear? analyzing and inferring the intentions of vulgar expressions. In *Proceedings of the 2018 conference on empirical methods in natural language processing*, pages 4405–4414.

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.

Gabriele Kasper. 1990. Linguistic politeness:: Current research issues. *Journal of pragmatics*, 14(2):193–218.

Yaser Keneshloo, Tian Shi, Naren Ramakrishnan, and Chandan K. Reddy. 2019. Deep reinforcement learning for sequence to sequence models.

Atharva Kulkarni, Sarah Masud, Vikram Goyal, and Tanmoy Chakraborty. 2023. Revisiting hate speech benchmarks: From data curation to system deployment. In *Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pages 4333–4345.

Harrison Lee, Samrat Phatale, Hassan Mansoor, Thomas Mesnard, Johan Ferret, Kellie Lu, Colton Bishop, Ethan Hall, Victor Carbune, Abhinav Rastogi, and Sushant Prakash. 2023. Rlaif: Scaling reinforcement learning from human feedback with ai feedback.

Kalev Leetaru. 2019. Online toxicity is as old as the web itself but the return to communities may help. forbes magazin.

Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain. Association for Computational Linguistics.

Sarah Masud, Manjot Bedi, Mohammad Aflah Khan, Md. Shad Akhtar, and Tanmoy Chakraborty. 2022. Proactively reducing the hate intensity of online posts via hate speech normalization. In *KDD '22: The 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pages 3524–3534. ACM.

Sarah Masud, Subhabrata Dutta, Sakshi Makkar, Chhavi Jain, Vikram Goyal, Amitava Das, and Tanmoy Chakraborty. 2021. Hate is the new infodemic: A topic-aware modeling of hate speech diffusion on twitter. In *2021 IEEE 37th International Conference on Data Engineering (ICDE)*, pages 504–515. IEEE.

Sarah Masud, Mohammad Aflah Khan, Vikram Goyal, Md Shad Akhtar, and Tanmoy Chakraborty. 2024. Probing critical learning dynamics of plms for hate speech detection. *arXiv preprint arXiv:2402.02144*.

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.

Swaroop Mishra, Daniel Khashabi, Chitta Baral, and Hannaneh Hajishirzi. 2022. Cross-task generalization via natural language crowdsourcing instructions.

Kevin L Nadal, Kristin C Davidoff, Lindsey S Davis, and Yinglee Wong. 2014. Emotional, behavioral, and cognitive reactions to microaggressions: Transgender perspectives. *Psychology of sexual orientation and gender diversity*, 1(1):72.

Reiichiro Nakano, Jacob Hilton, Suchir Balaji, Jeff Wu, Long Ouyang, Christina Kim, Christopher Hesse, Shantanu Jain, Vineet Kosaraju, William Saunders, Xu Jiang, Karl Cobbe, Tyna Eloundou, Gretchen Krueger, Kevin Button, Matthew Knight, Benjamin Chess, and John Schulman. 2022. Webgpt: Browser-assisted question-answering with human feedback.

Leticia Nieto and Margot Boyer. 2006. Understanding oppression: Strategies in addressing power and privilege. *Colors NW*, pages 30–33.

Long Ouyang, Jeff Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul Christiano, Jan Leike, and Ryan Lowe. 2022. Training language models to follow instructions with human feedback.

Sara Parker and Derek Ruths. 2023. Is hate speech detection the solution the world wants? *Proceedings of the National Academy of Sciences*, 120(10):e2209384120.

Javiera Perez Gomez. 2021. Verbal microaggressions as hyper-implicatures. *Journal of Political Philosophy*, 29(3):375–403.

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.

Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence embeddings using Siamese BERT-networks. 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 3982–3992, Hong Kong, China. Association for Computational Linguistics.

Amrita Saha, Ghulam Ahmed Ansari, Abhishek Laddha, Karthik Sankaranarayanan, and Soumen Chakrabarti.

2019. Complex program induction for querying knowledge bases in the absence of gold programs. *Transactions of the Association for Computational Linguistics*, 7:185–200.

Punyajoy Saha, Kanishk Singh, Adarsh Kumar, Binny Mathew, and Animesh Mukherjee. 2022. Countergedi: A controllable approach to generate polite, detoxified and emotional counterspeech. In *Proceedings of the Thirty-First International Joint Conference on Artificial Intelligence, IJCAI-22*, pages 5157–5163. International Joint Conferences on Artificial Intelligence Organization.

Maarten Sap, Saadia Gabriel, Lianhui Qin, Dan Jurafsky, Noah A. Smith, and Yejin Choi. 2019a. Social bias frames: Reasoning about social and power implications of language. *CoRR*, abs/1911.03891.

Maarten Sap, Saadia Gabriel, Lianhui Qin, Dan Jurafsky, Noah A Smith, and Yejin Choi. 2019b. Social bias frames: Reasoning about social and power implications of language. *arXiv preprint arXiv:1911.03891*.

Carla Schieb and Mike Preuss. 2016. Governing hate speech by means of counterspeech on facebook. In *66th ICA Annual Conference, at Fukuoka, Japan*, pages 1–23.

John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. 2017. Proximal policy optimization algorithms.

Emily Sheng, Kai-Wei Chang, Prem Natarajan, and Nanyun Peng. 2020. Towards controllable biases in language generation. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 3239–3254. Association for Computational Linguistics.

Orith Toledo-Ronen, Matan Orbach, Yonatan Bilu, Artem Spector, and Noam Slonim. 2020. Multilingual argument mining: Datasets and analysis. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 303–317, Online. Association for Computational Linguistics.

Bertie Vidgen, Helen Margetts, and Alex Harris. 2019. How much online abuse is there. Technical report, Alan Turing Institute.

Yizhong Wang, Swaroop Mishra, Pegah Alipoormolabashi, Yeganeh Kordi, Amirreza Mirzaei, Anjana Arunkumar, Arjun Ashok, Arut Selvan Dhanasekaran, Atharva Naik, David Stap, Eshaan Pathak, Giannis Karamanolakis, Haizhi Gary Lai, Ishan Purohit, Ishani Mondal, Jacob Anderson, Kirby Kuznia, Krima Doshi, Maitreya Patel, Kuntal Kumar Pal, Mehrad Moradshahi, Mihir Parmar, Mirali Purohit, Neeraj Varshney, Phani Rohitha Kaza, Pulkit Verma, Ravsehaj Singh Puri, Rushang Karia, Shailaja Keyur Sampat, Savan Doshi, Siddhartha Mishra, Sujan Reddy, Sumanta Patro, Tanay Dixit, Xudong Shen, Chitta Baral, Yejin Choi, Noah A. Smith, Hannaneh Hajishirzi, and Daniel Khashabi. 2022. Super-naturalinstructions: Generalization via declarative instructions on 1600+ nlp tasks.

Jason Wei, Maarten Bosma, Vincent Y. Zhao, Kelvin Guu, Adams Wei Yu, Brian Lester, Nan Du, Andrew M. Dai, and Quoc V. Le. 2022. Finetuned language models are zero-shot learners.

Lucas Wright, Derek Ruths, Kelly P Dillon, Haji Mohammad Saleem, and Susan Benesch. 2017. Vectors for counterspeech on twitter. In *Proceedings of the First Workshop on Abusive Language Online*, pages 57–62.

Marcos Zampieri, Shervin Malmasi, Preslav Nakov, Sara Rosenthal, Noura Farra, and Ritesh Kumar. 2019. Predicting the type and target of offensive posts in social media. *arXiv preprint arXiv:1902.09666*.

Shengyu Zhang, Linfeng Dong, Xiaoya Li, Sen Zhang, Xiaofei Sun, Shuhe Wang, Jiwei Li, Runyi Hu, Tianwei Zhang, Fei Wu, and Guoyin Wang. 2023. Instruction tuning for large language models: A survey.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2020a. Bertscore: Evaluating text generation with bert.

Yizhe Zhang, Siqi Sun, Michel Galley, Yen-Chun Chen, Chris Brockett, Xiang Gao, Jianfeng Gao, Jingjing Liu, and William B Dolan. 2020b. Dialogpt: Large-scale generative pretraining for conversational response generation. In *ACL: System Demonstrations*.

Wangchunshu Zhou, Yuchen Eleanor Jiang, Ethan Wilcox, Ryan Cotterell, and Mrinmaya Sachan. 2023a. Controlled text generation with natural language instructions.

Xuhui Zhou, Hao Zhu, Akhila Yerukola, Thomas Davidson, Jena D. Hwang, Swabha Swayamdipta, and Maarten Sap. 2023b. COBRA frames: Contextual reasoning about effects and harms of offensive statements. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 6294–6315, Toronto, Canada. Association for Computational Linguistics.

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

Daniel M. Ziegler, Nisan Stiennon, Jeffrey Wu, Tom B. Brown, Alec Radford, Dario Amodei, Paul Christiano, and Geoffrey Irving. 2020. Fine-tuning language models from human preferences.

A Dataset Analysis

Procedure and Criteria for Annotation: Prior to the commencement of the annotation endeavor, all annotators were thoroughly acquainted with the field guide[4] focusing on “addressing online harassment”. This preliminary stage involved extensive dialogues with the annotators to deepen their grasp on the concept of counterspeech. They were instructed to concentrate on key objectives for each counterspeech variant:

1. **Establishing the Objective:** Each variant of counterspeech should represent a distinct foundational idea, mode of expression, and desired effect.

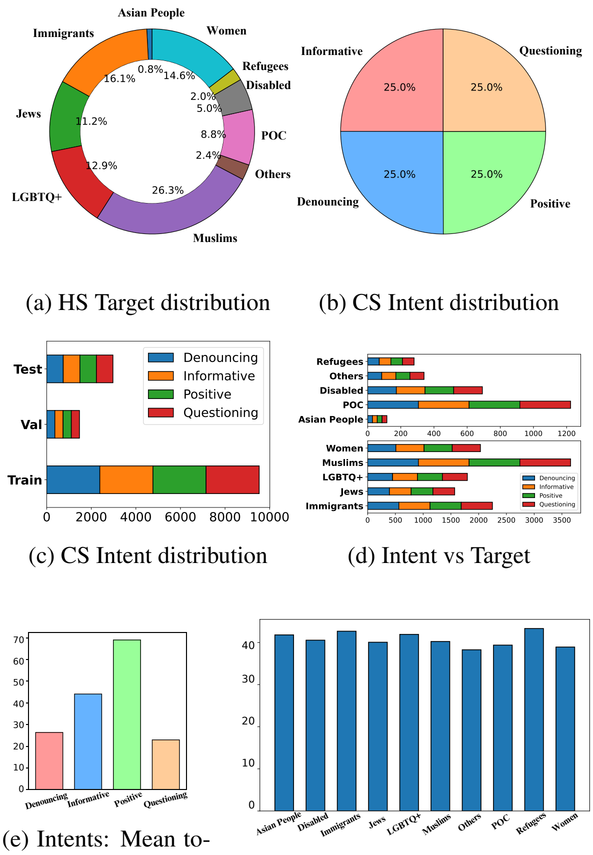
2. **Diminishing Conflict:** The goal was to ensure that every counterspeech contributed to de-escalating the situation without prompting further hostility or provoking more hate speech.

3. **Avoiding Hostile Language:** A firm rule was set against incorporating harsh language, such as threats, derogatory names, and swear words in the counterspeeches.

Adhering to these criteria, the annotators developed counterspeech responses tailored to $3,488$ unique cases of hate speech. For an exhaustive analysis of the dataset, see Table 1 in the IntentCONAN statistics section.

Statistical Analysis: In Figure 3, we present a detailed statistical analysis of `IntentCONANv2`. Our focus is on fine-graining the target groups into 10 distinct categories, a classification that significantly enhances the model’s capacity to comprehend hate speech context in conjunction with intent. Figure 3a specifically features a donut chart illustrating the distribution of hate speech target groups. To delve into specifics, we meticulously extracted refugees from the broader category "Immigrants," and this subgroup constitutes a noteworthy 2% of the entire dataset. Simultaneously, we honed in on the "Asian people (AP)" category within the expansive classification of "Others," contributing a modest yet discernible 0.8% to the overall dataset. Moving on to Figure 3b, it provides the distribution of counter speech across the intent dimension. This multi-faceted examination contributes to a comprehensive evaluation of both hate speech and counterspeech dynamics within the dataset.

To delve into details regarding our expanded dataset, Figure 3c provides insights into the uniform distribution of data points across training, validation, and testing sets. Examining Figure 3d along the target group dimension reveals a consistent dispersion of counter speeches across different intents. Figures 3e and 3f elucidate the mean token distribution of counter speeches, both across intents and within various target groups. Notably, the counter speeches with the highest mean token length tend to lean towards positive intents, as evidenced by the findings. Simultaneously, a pattern of uniformity emerges in the mean token length distribution across the diverse target groups, contributing to a comprehensive understanding of the dataset characteristics.



(a) HS Target distribution (b) CS Intent distribution

(c) CS Intent distribution (d) Intent vs Target

(e) Intents: Mean token length (f) Targets: Mean token length

Figure 3: Visual exploration of various attribute distribution present in `IntentCONANv2`

B Hate Speech Explanations

In this study, we employ the `COBRACORPUS` dataset (Zhou et al., 2023b) to develop seven distinct subsets for instruction tuning. Each subset corresponds to specific explanation dimensions as outlined in Table 5. We adhere to the original training, validation, and testing partitions as established in the COBRACORPUS dataset for each of these sub-

[4] https://onlineharassmentfieldmanual.pen.org/

HS		Gypsies are essentially thieves.
CS	INF	Stereotyping an entire community based on the actions of a few is unfair and perpetuates discrimination. it is important to recognize that criminal behavior exists in every society and is not limited to any specific group. Painting all gypsies as thieves is a harmful generalization. It is crucial to approach individuals and communities with respect and empathy rather than perpetuating negative stereotypes.
	QUE*	They are not thieves, just like any other citizen. Did you just infer that from their belonging to that community?
	DEN	Respect and understanding should prevail instead.
	POS	It is important to avoid generalizations and stereotypes about any community, including the gypsy community. Stereotyping an entire group based on the actions of a few individuals is unfair and perpetuates discrimination. Let's focus on promoting understanding and empathy towards diverse communities, appreciating their rich cultural heritage and contributions.

Table 4: Comparison between IntentCONAN and IntentCONANv2. The $*$ shows the pre-existing counter speech in the IntentCONAN dataset.

sets. The explanation dimensions we incorporate are grounded in theoretical frameworks from pragmatics and implicature (Grice, 1975; Perez Gomez, 2021) as well as the social psychology of bias and inequality (Nieto and Boyer, 2006; Nadal et al., 2014). This approach allows for an extensive expansion of reasoning dimensions compared to previous studies, which primarily focused on identifying targeted groups and biased implications (Sap et al., 2019b; ElSherief et al., 2021). Below, we provide a comprehensive description of each explanation dimension.

Speaker Intent: The Speaker Intent delves into the fundamental communicative purpose underlying a statement deemed hate speech (e.g., "to incite", "to jest", "to demean"). Earlier studies have illustrated that understanding the speaker's intent is pivotal for identifying the pragmatic consequences, including biases and damages, thus facilitating the recognition of hate speech (Kasper, 1990; Dynel, 2015; Holgate et al., 2018).

Target Group: Target Group characterizes the particular social or demographic group that is the subject or object of the hate speech (for example, "those **jews**", "dislike **black** artists"), and this may encompass the addressee if they are the intended target. This aspect has been emphasized in numerous previous studies due to its significance in grasping the offensive nature and potential harm of the speech (Zampieri et al., 2019; Sap et al., 2019b; Vidgen et al., 2019).

Power Dynamics: Power Dynamics encapsulate the imbalances in social and cultural power or the axis along which privilege and discrimination align between the speaker and the referred group or audience of a statement (Nieto and Boyer, 2006). An instance is when a statement is made by someone of white ethnicity towards a person of black ethnicity, demonstrating a racial hierarchy where the former holds greater societal privilege and status than the latter. This hierarchy is instrumental in comprehending how the context modifies the offensiveness and potential harm of a statement. Statements from a position of greater power towards someone with less power are typically more offensive or damaging. The hierarchy is described through free-text explanation of the nature of power disparity, like "gender-based disparity", "racial hierarchy", or "economic disparity".

Implication: Implication of Implied Meaning elucidates the biased, prejudicial, or stereotype-laden meaning *suggested by the hate speech statement*, akin to (Sap et al., 2019b). This meaning significantly aligns with the interpretation received by the audience or target group and may deviate from the intent ascribed by the speaker (e.g., in cases of microaggressions). The Connoted Meaning dimension is expressed as a free-text elucidation of the type of harm or impact implicated in the statement, such as "emotional distress", "perpetuation of stereotypes", or "dissemination of false information". This aspect is crucial for deciphering the repercussions of a statement and its capacity to offend or injure. It also aids in identifying the specific nature of harm or impact, which is beneficial for crafting effective strategies to alleviate the detrimental effects of harmful speech.

Emotional Reaction: Emotional Reaction or Affective Response outlines the potential adverse impacts and injuries resulting from the statement and its suggested meaning on the audience or referred group (Nadal et al., 2014). It consists of a free-text narrative describing the immediate emotional responses or reactions to a statement (e.g., "irritation and displeasure", "sense of insignificance").

For instance, in the remark "Wow, your English is surprisingly good!" made by a white individual to a person of color, the affective response could include "the recipient and people of color might feel irked, patronized, or insecure about their English proficiency". This dimension is insightful for understanding the psychological ramifications of pejorative language and its effect on the targeted group or listener's emotional health.

Cognitive Reaction: Cognitive Reaction details the potential intellectual impacts and damages that the statement and its suggested meaning could inflict on the audience or referred group (Nadal et al., 2014). This is articulated through a free-text narrative on the immediate intellectual reactions to a statement (e.g., "bewilderment", "skepticism", "incredulity"). For example, the remark, "You're too young to grasp this," could lead to reactions such as "the addressee feeling perplexed, questioning their own comprehension, or feeling undervalued". This dimension is valuable for gauging the intellectual influence of derogatory language and its effect on the mental well-being of the targeted group or listener.

Offensiveness: Offensiveness quantifies how insulting or harmful a statement can be to the referred group or listener. This offensive measure is conveyed through a free-text depiction of the type of harm or impact implicated in the statement, such as "racial prejudice", "gender discrimination", or "anti-LGBTQ+ sentiment".

C Instructions

While defining Instructions for both the Auxiliary Explanation Generation and Counterspeech Generation tasks, we follow a standard format where the instruction is brief, to the point, and describes the expectation from the model clearly (See Table 5).

D Prompting

Drawing from previous studies (Lee et al., 2023), we follow a *preamble prefix* (**Preamble - Instruction - Exemplar**) prompt template for both the zero-shot and few-shot experiments. An example prompt template for one-shot exemplar is illustrated in Table 6. For each of the three LLM baselines described in section 5.1, we conduct inference on the IntentCONANv2 test set. For few-shot prompting, we sample exemplars for in-context learning from the IntentCONANv2 training set. For a given hate speech instance from the test set, we select exemplars based on the top-n semantically similar instances from the training set based on BM25 scores. We experiment with n=3, n=5, and n=8 exemplars, are find that the best results are obtained from n=3 in-context examples. We report the numbers with n=3 in-context examples for all few-shot experiments.

For inference on the FLAN-T5 XXL model, we make use of one NVIDIA A100 (80 GB) GPU, which results in 2 inferences per second. For GPT-3.5-turbo and GPT-4, we use the **gpt-3.5-turbo-1106** and **gpt-4** model versions respectively as hosted by the OpenAI API[5]. On average each turn takes about 2 seconds for inference for both zero- and few-shot inference.

E Reward Model (RM)

As detailed under section 4, we design a composite reward function by combining outputs of three transformer-based models - Argument Quality (AQ), Pro-Con (PC), and Toxicity (T).

Argument Quality: Akin to the Argument Quality (AQ) service detailed by Bar-Haim et al. (2021), we train a BERT-based regression model using a dataset comprising around $27,000$ arguments, which span a diverse range of subjects. Each argument in this dataset has been assigned a quality rating, as outlined in (Gretz et al., 2020). High-quality arguments are characterized by their grammatical accuracy, proper use of language, clear and concise expression, and a clearly defined stance on the topic.

In this study, the primary role of the regression model is to assess the effectiveness of counterspeech. It evaluates how well a counterspeech can form logical and coherent arguments. The model assigns a score ranging from 0 to 1 to each argument, where a score of 1 represents a high-quality argument and a score of 0 indicates a low-quality one. Essentially, this scoring system is used to determine the effectiveness of counterspeech in terms of its argumentative quality and coherence.

Pro-Con: Similar to the Pro-Con service described by Bar-Haim et al. (2021), we train a BERT-based classifier on 400K stance labeled examples, including arguments extracted from the Lexis-Nexis corpus, as well as arguments collected via crowdsourcing (Toledo-Ronen et al., 2020).

[5] https://openai.com/blog/openai-api

Task-ID	Generation Type	Instruction
I_1	Offensiveness	Analyze the offensiveness of the statement: {HS}
I_2	Target Group	Identify the group of people that the speaker is targeting or discriminating against in the offensive statement: {HS}
I_3	Speaker Intent	Analyze the speaker's intention behind writing the offensive statement: {HS}
I_4	Power Dynamics	Explain the underlying power dynamics between the speaker and the target group in the offensive statement: {HS}
I_5	Implication	Explain the implied meaning underlying the offensive statement: {HS}
I_6	Emotional Reaction	Describe how the target group might feel emotionally after reading or listening to the offensive statement: {HS}
I_7	Cognitive Reaction	Describe how the target group might react cognitively after reading or listening to the offensive statement: {HS}
I_8	Intent-Specific Counterspeech	Analyze the different aspects such as offensiveness, target group, stereotype, power dynamics, implied meaning, emotional, and cognitive reactions before writing a {INT} counterspeech for the offensive statement: {HS}

Table 5: Detailed Instructions for the tasks of explanation and counterspeech generation respectively. The instructions labeled {I_1, I_2, ..., I_7} correspond with each of the dimensions of hate speech explanations in the Auxiliary Explanation Generation (AEG) task, as outlined in Section 4. Instruction I_8 is crafted for training task-specific LoRA adapter for generating intent-conditioned counterspeech, detailed in Section 4. In these instructions, `HS` represents the instance of hate speech, and `INT` denotes the targeted intent of the counterspeech, which includes `Positive`, `Denouncing`, `Informative`, or `Questioning`.

The classifier's function is to evaluate a given sentence's stance towards a specific topic, assigning a score between -1 and 1. A score of 1 signifies a pro-stance (in favor of the topic), while -1 indicates a con-stance (contradicting the topic). In the context of this study, the classifier is used to analyze the relationship between hate speech (as the topic) and counterspeech (as the sentence). The goal is to determine how well the counterspeech opposes the hate speech. A high reward is expected if the counterspeech effectively contradicts the hate speech, and a lower reward if it supports it.

To facilitate this assessment, the output score from the Pro-Con classifier, denoted as $(PC(x', y'))$, is normalized from its original range of $[-1, 1]$ to $[0, 1]$ by $(\frac{1-PC(x',y')}{2})$. This normalization ensures that lower Pro-Con scores, which indicate effective counterspeech against hate speech, correspond to higher rewards. This approach aligns the Pro-Con scoring mechanism with the reward structure desired in the study, emphasizing the efficacy of counterspeech in opposing hate speech.

Toxicity: We use a pre-trained model from the Detoxify python library[6]. Specifically, we employ the texttunbiased classifier, which is a RoBERTa-based model trained on the Jigsaw Unintended Bias in Toxicity Classification challenge[7]. The primary function of this model is to predict the level of toxicity in a given text input. It generates a score ranging from 0 to 1, where 0 corresponds to low toxicity and 1 to high toxicity.

In the composite reward function, the model's output is utilized to determine the toxicity of counterspeech that is generated. The objective is to ensure that counterspeech has minimal toxicity. Therefore, within the reward function of the study, the toxicity score is subtracted from 1, effectively inverting the scale. In this inverted scale, a higher value is considered better, signifying lower toxicity.

[6] https://pypi.org/project/detoxify/
[7] https://www.kaggle.com/c/jigsaw-toxic-comment-classification-challenge

Preamble	A Positive Counterspeech responds to a hate speech statement using empathy and affiliation...attenuating entrenched extremist viewpoints. A Counterspeech is considered to be good if it satisfies certain desired qualities like relevance, effectiveness, and non-toxicity. Below we define each of these axes: **Relevance:** This axis evaluates the extent to which counterspeech directly addresses the core message of the hate speech. It assesses whether the counterspeech maintains a coherent and focused dialogue aimed at neutralizing the impact of the hate speech. A low relevance indicates a disconnect between the counterspeech and the hate speech, exemplified by counterspeech that strays off-topic. For instance, if the hate speech centers on LGBTQ issues, but the counterspeech diverts to religious beliefs, it would be deemed irrelevant. **Effectiveness:** This axis gauges the ability of counterspeech to logically and coherently challenge or refute the biases, stereotypes, or offensive content present in the hate speech. Effective counterspeech should logically counteract the hate speech, thereby diminishing its impact or negating its offensive message. **Non-toxicity:** This axis measures the level of respectfulness and reasonableness in the counterspeech. It is essential that counterspeech remains free from rudeness or provocativeness. A non-toxic approach is crucial as counterspeech that is perceived as aggressive or disrespectful may exacerbate the situation, potentially leading to an escalation of hate speech rather than its mitigation.
Instruction	»»»» Instruction »»»» Given a hate speech statement, generate a **Positive Counterspeech** by following the definitions given above.
Exemplars (1-shot)	»»»» Examples »»»» **Statement** - Children should learn about gender identity and sexual orientation from their families rather than in class from LGBT fundamentalists. **Positive Counterspeech** - Teaching all children to be accepting of each other gives a sense of belonging to any child who feels a little bit different, and helps them to know it's okay to be themselves.
Sample to Generate	**Statement** - LGBTs just want to brainwash our children. They should not be taught about gender identity and sexual orientation in school, theyll end up being bullied. **Positive Counterspeech** -

Table 6: Prompt template used for zero-shot and few-shot experiments.

F Experimental Setup

All the models in our study were developed using the HuggingFace platform[8] and Pytorch. For the implementation of LoRA, we utilized the `peft` Python library[9], and for the Proximal Policy Optimization (PPO) algorithm, the TRL library was employed[10]. At inference time, we use a common set of generation parameters in all our experiments: top_k = 1, top_p = 1, max_new_tokens = 512, temperature = 1, do_sample = False. The entirety of our training and testing was conducted on a single NVIDIA A100 (80 GB) GPU. Detailed descriptions of the training procedures for each phase of `CoARL` are presented in the subsequent sections.

[8]https://huggingface.co/
[9]https://pypi.org/project/peft/
[10]https://github.com/lvwerra/trl

Phase 1: Auxiliary Explanation Generation. In this phase, we fine-tuned a FLAN-T5 XXL model, which has 11 billion parameters, over three epochs. The training used the AdamW optimizer, with a learning rate of $1e-4$ and a batch size of 8. Beam search was employed as the decoding strategy. The entire training process, using FP32 on an NVIDIA A100 GPU, took approximately 12 hours.

Phase 2: Task-Specific Adapter Training. During this phase, the model weights established in Phase 1 were locked, and we initialized a LoRA adapter using PEFT. This adapter was then fine-tuned on the `IntentConanV2` training set, following the instruction format detailed in Table 5. The fine-tuning process involved specific hyperparameters, including a batch size of 8, training for 16 epochs, and using the Adam optimizer. The learning rate was set to $4e - 6$. Other vital settings included a maximum input token length of 256, a LoRA (r) value of 16, a LoRA (α) value of 32, and a LoRA layer dropout rate of 0.05.

Phase 3: Optimization via RLAIF. As detailed in section 4, we trained a policy model using the Proximal Policy Optimization (PPO) method. This training applied to FLAN-T5 XXL with LoRA adapters from the previous phase, involved the following hyperparameters: a learning rate of $1.4e-6$, an initial KL penalty coefficient ($Init_k l_c oeff$) of 0.03, a batch size of 32, a mini-batch size of 2, 15000 steps, a target KL value (Target) of 5 for adaptive KL control, a horizon of 10000 for adaptive KL control, a cliprange value of 0.25 for loss calculation, 5 epochs, and a target KL ($Target_k l$) of 0.05.