Counterspeech

Nandini Jain
IIIT Delhi
nandini22316@iiitd.ac.in

Prithu Raj Singh
IIIT Delhi
prithu22381@iiitd.ac.in

Suhani Kalyani
IIIT Delhi
suhani22511@iiitd.ac.in

Abstract

Hate speech on social media has emerged as a pressing concern due to its potential to spread misinformation and cause psychological harm. Traditional moderation techniques, such as content removal, often conflict with freedom of speech. In contrast, counterspeech is a constructive response aimed at challenging hate by offering a non-censorial alternative. This paper presents a two-phase framework for intent-specific counterspeech generation. In the first phase, we pre-train a BART model using the COBRACorpus to enrich its pragmatic understanding of hate speech, followed by fine-tuning on the IntentCONANv2 dataset to align with specific counterspeech intents (informative, denouncing, questioning, and positive). In the second phase, we refine responses using COT prompting with the Mistral-7B-Instruct model to improve coherence, reduce toxicity, and enhance intent alignment. Evaluation through both automatic metrics (BLEU, ROUGE-L, BERT-F1, Detoxify) and human judgment (Category Accuracy and Conditioned CS) demonstrates that our approach outperforms baselines, achieving higher fluency and relevance in generated counterspeech.

1 Introduction

With growing social media presence in our daily lives, hate speech has turned out to be a significant issue on online platforms. It results in misinformation, discrimination and polarization.It results in psychological harm to targeted individuals and communities. Conventional techniques of moderation such as content removal can be seen as a curb/block to our freedom of speech and expression. Conversely, counterspeech has proven to be a viable, non-censorial option that seeks to counteract hate by replying with positive, non-hateful messages. Counterspeech has been found to be an effective way to combat hostile speech, impose respectful dialogue, and discourage future spread of hate speech.

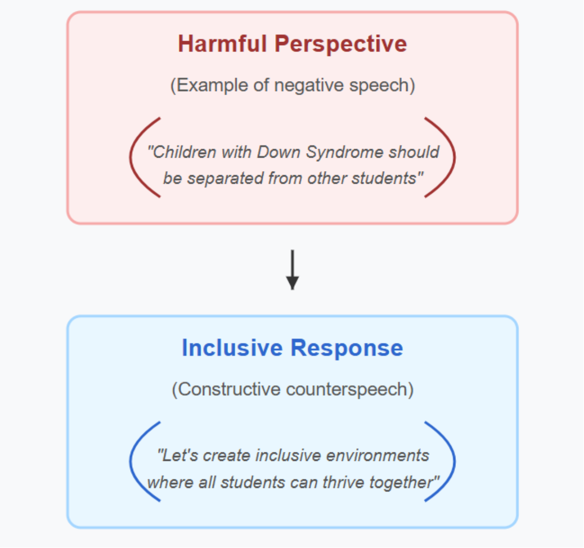

Harmful Perspective
(Example of negative speech)
"Children with Down Syndrome should be separated from other students"
Inclusive Response
(Constructive counterspeech)
"Let's create inclusive environments where all students can thrive together"

Figure 1: An example of positive counterspeech generated for given hate speech

However, generating effective counterspeech is a nuanced task. It requires not only grammatical fluency and coherence, but also an understanding of the context, the emotional undertone of the original message, and the strategic intent behind the response be it to inform, question, denounce or offer positivity. This paper addresses these challenges by proposing a model for corresponding intent counterspeech generation, which aims to produce responses aligned with specific communicative goals. Our objective is to develop a system that can automatically generate meaningful, intent-matched counterspeech that contributes to safer and more respectful online spaces.

2 Related Work

2.1 Counterspeech Datasets

Counterspeech research relies on diverse datasets to analyze, model, and evaluate effective responses to hate speech. Among these, IntentCONAN (Gupta et al., 2023) extends the Multi-Target CONAN dataset (Fanton et al., 2021) by introducing intent-based annotations for counterspeech, categorizing responses into five key intents: informative (INF), question (QUE), denouncing (DEN), humor (HUM), and positive (POS). Unlike earlier frameworks like Benesch et al. (2016) or Mathew et al. (2019), which used nine and seven intent categories respectively, IntentCONAN consolidates semantically similar intents to address data scarcity, ensuring practicality for NLP applications. Human annotated datasets, such as CONAN (Chung et al., 2019), focuses more on Islamophobic hate speech and its corresponding counterspeech, while Mathew et al.'s (2019) Twitter dataset examines organic counterspeech in response to racism, sexism, and xenophobia. These datasets are crucial for training models to evaluate intervention effectiveness and understanding the discourse in combating online hate.

2.2 Counterspeech Generation Methods

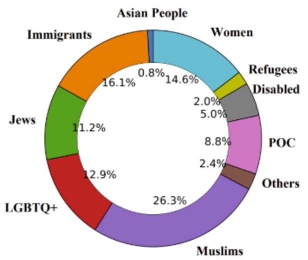
Recent advancements in counterspeech generation have introduced sophisticated methods to produce intent-aware and non-toxic responses to hate speech. Hengle et al. (2024) proposed CoARL, a multi-phase framework that employs sequential multi-instruction tuning to teach models about the intents, reactions, and harms associated with offensive statements. This is followed by learning task-specific low-rank adapter weights for generating intent-conditioned counterspeech, and finally, reinforcement learning is used to fine-tune outputs for effectiveness and non-toxicity. CoARL demonstrated improvements in intent-conformity and argument quality, outperforming existing benchmarks, including prominent LLMs like ChatGPT.

Gupta et al. (2023) introduced QUARC, a two-stage framework for intent-conditioned counterspeech generation. QUARC utilizes vector-quantized representations for each intent category and incorporates PerFuMe, a novel fusion module that integrates intent-specific information into the model. Evaluation results indicated that QUARC outperformed several baselines by an average of 10% across various metrics, with human evaluations supporting the generation of more appropriate responses compared to comparative systems.

3 Dataset

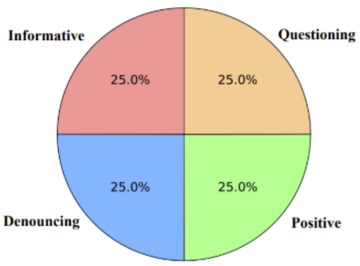
We use the **IntentCONANv2** dataset. It is a large-scale dataset containing **13,952 counterspeech instances** categorized into four distinct intents: **positive (POS)**, **informative (INF)**, **questioning (QUE)** and **denouncing (DEN)** . Notably, this version excludes the humor intent that was present in the original dataset. A key improvement in IntentCONANv2 is the inclusion of an average of four counterspeech responses per hate speech, compared to the original average of two. The train dataset consisted of 9.53k rows, the validation dataset consisted of 1.47k rows and the test dataset consisted of 2.97k rows.

We have also used the **COBRACorpus** dataset for this task, which is a large-scale, context-aware resource designed to facilitate nuanced understanding of offensive or biased language. Unlike traditional hate speech datasets, COBRACorpus emphasizes the importance of social and situational context in interpreting the harms and intentions behind potentially offensive statements. The dataset contains approximately 33,000 such statements, each paired with machine-generated contextual scenarios and free-text explanations detailing perceived offensiveness, implied biases, speaker intent, and listener reactions. This enables a more comprehensive approach to counterspeech by grounding responses in real-world social dynamics. By modeling these contextual factors, the dataset supports the development of systems that move beyond context-agnostic detection and toward a deeper, more empathetic understanding of online discourse.



(a) HS Target distribution

Figure 2: Hatespeech Target distribution in our dataset



(b) CS Intent distribution

Figure 3: Types of intent for CS in our dataset

4 Methodology

Our methodology involves a two-phase pipeline designed to effectively generate intent-aligned and contextually appropriate counterspeech while addressing key limitations such as toxicity and generic output.

Phase 1 : Generation of pragmatic explanations as pre-training and intention-conditioned fine tuning

The initial phase aimed at enriching the base model's comprehension of hate speech through pragmatic contextualization, followed by fine-tuning to produce intent-specific counterspeech.

Pragmatic Pre-training: We leveraged the COBRACORPUS dataset, which annotates hate speech instances across multiple pragmatic dimensions. For our purposes, we utilized **four key pragmatic dimensions: intent, target group, power dynamics, and implication**. A BART-base model was fine-tuned on this dataset to predict these pragmatic explanations. This pre-training step was critical to enabling the model to understand deeper implicit contexts and nuanced meanings behind hateful speech which could be seen by the improved scores.

Intent-Conditioned Fine-Tuning: Following pragmatic pre-training, we conducted a second stage of fine-tuning using the IntentCONAN v2 dataset, specifically curated for intent-conditioned counterspeech generation. Here, the inputs consisted of hate speech instances concatenated explicitly with their intended counterspeech type (informative, denouncing, questioning, positive). This structured input format aimed to enforce the generation of counterspeech explicitly aligned with predefined intents.

Phase 2: Chain-of-Thought Refinement using Mistral-7B Despite effective intent conditioning in Phase 1, preliminary evaluations revealed persistent issues, notably moderate toxicity and instances where generated counterspeech inadvertently shifted blame to other groups. Moreover, BART's outputs occasionally lacked fluency and diversity.

To mitigate these limitations, we introduced a second refinement phase utilizing the Mistral-7B-Instruct language model. We employed a detailed chain-of-thought prompting strategy designed to explicitly instruct the model to:

1. **Identify and mitigate toxicity** within the initially generated counterspeech.

2. **Critically evaluate and enhance intent alignment**, ensuring outputs closely adhered to their specified purpose.

3. **Increase fluency and contextual relevance**, promoting natural and engaging counterspeech.

This prompted refinement stage significantly improved the overall quality, reducing toxicity and enhancing both relevance and coherence of generated counterspeech responses.

This prompted refinement stage significantly improved the overall quality, reducing toxicity and enhancing both relevance and coherence of generated counterspeech responses.

5 Experimental Setup and Results

5.1 Models

For our Baseline, we generated intent specific counterspeeches using both **pre-trained and fine-tuned language models**. Our pipeline included four main approaches:

1. To compare the performance of contemporary large language models, we employed **LLaMA 3.3 70B** in both **zero-shot** and **few-shot** settings. This allowed us to evaluate the model's generative ability on counterspeech tasks without **any task-specific fine-tuning**.

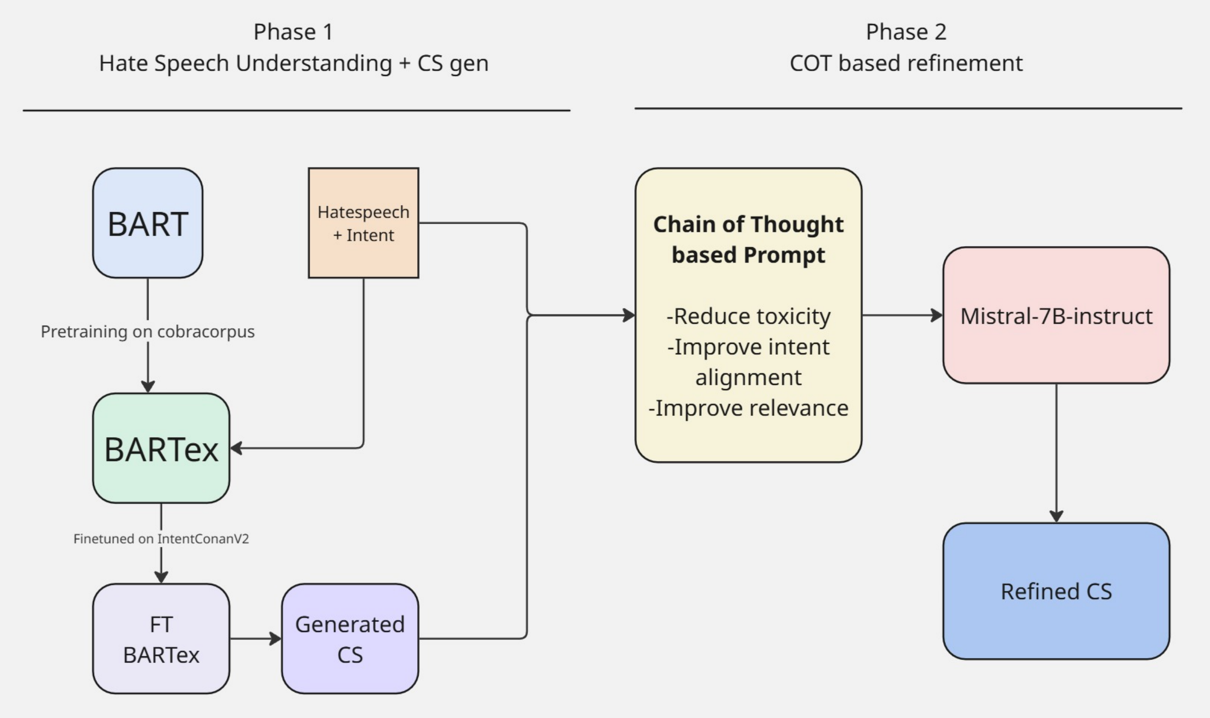


Figure 4: *Overview of the proposed two-phase counterspeech generation pipeline. In Phase 1, BART is first pretrained on CobraCorpus to develop pragmatic awareness, followed by fine-tuning on the intent-conditioned IntentConanV2 dataset to produce intent-specific counterspeech (BARTex). In Phase 2, the generated responses are further refined using a Chain-of-Thought (CoT) prompting approach with Mistral-7B-Instruct, aimed at reducing toxicity, improving intent alignment, and enhancing contextual fluency.*

2. We adopted **DialoGPT** and **BART** models as used in *Rishabh et al.*[1] and **fine-tuned** them on our dataset **IntentCONANv2** to better align their generations with intent-conditioned counterspeech response.

3. We also created a **human evaluation subset** consisting of **40** balanced instances, sampled across different counterspeech types **(10 from each intent)**, to ensure a fair comparison between the models. We used this set for human evaluation.

Our best model was the **Bart** model. We fine tuned this model on the COBRACorpus dataset for 3 epochs on the 16GB P100 GPU to improve its generation. We then further finetuned this model for 4 epochs on our IntentConan dataset.

Further, we used a smaller language model **Mistral-7B-Instruct** to refine the generations received from the Bart model.

5.2 Results

We evaluate our model using a combination of automatic and human evaluation metrics. For automatic evaluation, we employ **BLEU (B)**, **ROUGE-L (R)**, and **BERT-F1 Score (BF)** to assess the similarity between the generated counterspeech and the reference responses. These metrics capture different aspects of textual similarity: **BLEU** measures n-gram overlap, **ROUGE-L** focuses on the longest common subsequence, and **BERTScore** leverages contextual embeddings for a more semantic comparison.

To ensure the generated responses are non-toxic, we use the Detoxify model to compute a **Toxicity (T)** score.

Given the subjective and context-sensitive nature of counterspeech generation, we also conduct human evaluation on a designated subset of the data. Following the methodology of *Rishabh et al.* [1], we use two key human evaluation metrics:

Conditioned Counterspeech (CC), which assesses whether the generated output is a relevant and appropriate response to the given hate speech, and **Category Accuracy (CA)**, which evaluates whether the response aligns with the intended counterspeech category. Condtioned Counterspeech was measured using a **five-point Likert scale**: **1:** The counterspeech is incoherent, irrelevant, or completely off-topic with respect to the hate speech. **2:**

Table 1: Example of Counterspeech Refinement

Hate Speech: *"Women are basically childlike, they remain this way most of their lives. Soft and emotional. It has devastated our once great patriarchal civilizations."*
Generated CS (BARTex): *"Women are not childlike and they deserve respect. They have made significant contributions to our society"*
Refined CS (CoT with Mistral-7B): *"Women are not inherently childlike, and it is not fair to generalize that they remain this way most of their lives. In fact, women have made significant contributions to our societies and civilizations."*

The response barely relates to the hate speech but lacks logical consistency or appropriateness. **3:** The counterspeech shows moderate relevance but may miss important context. **4:** The response is largely coherent and relevant, with minor issues in phrasing or alignment. **5:** The counterspeech is fully coherent, contextually appropriate, and directly counters the given hate speech effectively. Category Accuracy is a **binary metric (Yes/No)** that checks whether the generated counterspeech **correctly reflects the desired intent category** (e.g., informative, denouncing questioning, positive) as specified during generation.

6 Observations

For automatic evaluation metrics -

1. **BLEU and ROUGE-L** steadily improve as the model architecture becomes more informed and intent-aligned. The +Explain model gets the highest scores for both metrics, which means its outputs have the most word and phrase overlap (lexical similarity) with the correct (reference) responses . This shows that training the model with pragmatic explanations helps it produce more similar and relevant text.

2. **BERT-F1**, which captures semantic similarity, is highest for +Explain (0.877), followed closely by +CoT Refine (0.875), indicating strong semantic alignment with human references.

3. **Toxicity** is a major differentiator:

   While Llama (zero shot and few shot) and BART maintain reasonably low toxicity, the **+CoT Refine** model **significantly reduces toxicity (0.009)**, outperforming all baselines. This demonstrates the **effectiveness of Chain-of-Thought prompting in detoxification**.

For human evaluated metrics -

1. **Category Accuracy (CA)** -
   (a) **Llama few shot** sees a substantial jump (0.825), showing that **few-shot learning enhances category awareness**.
   (b) **+CoT Refine** achieves the **best alignment (0.875)**, suggesting that the refinement phase successfully enforces intent-matching during generation.

2. **Conditioned Counterspeech (CC) -**
   (a) +CoT Refine again outperforms all others (4.775), which shows its **superiority in generating coherent, context-sensitive, and impactful responses**.
   (b) DialoGPT performs the worst (3.175), indicating its **generic or off-topic nature** in this domain.

Overall ,

1. The **+Explain** model clearly benefits from **pragmatic pretraining**, which improves textual similarity and semantic alignment, but its toxicity is not significantly lower than BART or even DialoGPT.

2. The **+CoT Refine** model performs better across **all human-centric metrics** and **Toxicity**, showing that **chain-of-thought prompting enables critical reflection, detoxification, and stronger intent alignment**.

3. **BART** provides a strong baseline but lacks robustness in coherence (CC) and intent specificity (CA) without additional conditioning.

Table 2: Comparative Results

Model	B ↑	R ↑	BF ↑	T ↓	CA ↑	CC ↑
DialoGPT	0.105	0.126	0.854	0.052	0.475	3.175
Llama zs	0.135	0.163	0.866	0.014	0.625	4.200
Llama fs	0.138	0.168	0.868	0.118	0.825	4.525
BART	0.152	0.194	0.872	0.017	0.700	3.675
+Explanation Gen	**0.192**	**0.202**	**0.877**	0.050	0.750	3.725
+CoT Refine	0.153	0.167	0.875	**0.009**	**0.875**	**4.775**

7 Conclusion

The experimental results show the effectiveness of our **two-phase methodology** for intent-conditioned counterspeech generation. Each phase contributes distinct and complementary benefits to the overall pipeline:

- **Phase 1 (+Explain)**, which incorporates pragmatic explanation pretraining, significantly enhances the **interpretability** of model outputs and improves **intent-awareness**. By conditioning the generation process on nuanced socio-linguistic features such as speaker intent, power dynamics, and target group, this phase ensures that the model produces responses aligned with the communicative goals of counterspeech (e.g., denouncing, informing, questioning, or offering positivity).

- **Phase 2 (+CoT Refine)**, which applies chain-of-thought prompting through the Mistral-7B-Instruct model, further **polishes** the generated outputs. This phase leads to notable improvements in **fluency**, **contextual relevance**, and **semantic alignment**, while also achieving significant **reduction in toxicity**. Importantly, it enhances **human satisfaction**, as reflected in higher Category Accuracy (CA) and Conditioned Counterspeech (CC) scores in our human evaluations.

Across both automatic metrics (BLEU, ROUGE-L, BERTScore, Toxicity) and human evaluations, our model demonstrates **strong, comparable performance** to powerful baselines like LLaMA-70B (in few-shot settings), while also offering greater controllability and lower toxicity. This shows that open instruction-tuned LLMs can also offer a **scalable and effective** solution for generating constructive counterspeech.

8 Future Work

While our two-phase framework demonstrates significant improvements in generating intent-aligned, low-toxicity counterspeech, there remain several avenues for further exploration and enhancement.

8.1 Multi-Intent Counterspeech Generation

Our current model generates counterspeech conditioned on a single specified intent. However, real-world responses often blend multiple communicative goals, for example, a message may simultaneously inform and denounce harmful claims. A promising extension would be the development of models capable of generating *multi-intent counterspeech*, either by combining intents explicitly during input conditioning or by dynamically identifying relevant intent combinations. This could lead to more nuanced and contextually powerful responses that better mirror human communication strategies.

8.2 Multilingual Counterspeech Systems

Hate speech is a global phenomenon, and many languages lack well-developed moderation tools. Expanding the current model to support *multilingual counterspeech generation* would significantly increase its societal impact. This could involve fine-tuning large multilingual language models or leveraging cross-lingual transfer learning approaches to generate culturally and linguistically appropriate counterspeech in low-resource settings. Creating intent-tagged multilingual datasets akin to Intent-CONANv2 would be a valuable step in this direction.

Acknowledgments

We would like to express or sincere gratitude to Professor Md Shad Akhtar for his invaluable guidance, constructive feedback, and constant support throughout the course. His expertise and mentorship played a crucial role in shaping the direction and quality of this work. We are also thankful to our Teaching Assistant, Shivam Kumar, for his timely clarifications, helpful discussions during various stages of the project.

References

[1] Hengle, Amogh, Vishvak Murahari, Anirudh Srinivasan, and Tanmay Sinha. *CoARL: A Multi-Phase Framework for Intent-Aware and Non-Toxic Counterspeech Generation.* arXiv preprint arXiv:2305.13776, 2023.

[2] Gupta, Rishabh, Megha Srivastava, and Tanmoy Chakraborty. *QUARC: Intent-Conditioned and Quality-Aware Counterspeech Generation via Representation Fusion.* In Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, pages 6702–6717. NAACL 2024.

[3] Gupta, Rishabh, Megha Srivastava, and Tanmoy Chakraborty. *IntentCONAN: A Pragmatic Perspective on Intent-aware Counterspeech Generation.* Findings of the Association for Computational Linguistics: ACL 2023, pages 6346–6362.

[4] Aswini123. *IntentCONANv2 Dataset.* Available at: https://huggingface.co/datasets/Aswini123/IntentCONANv2

[5] Zhou, Xuhui, Hao Zhu, Akhila Yerukola, Thomas Davidson, Jena D. Hwang, Swabha Swayamdipta, and Maarten Sap.
COBRA Frames: Contextual Reasoning about Effects and Harms of Offensive Statements. Available at: https://aclanthology.org/2023.findings-acl.392/