Counterspeech the ultimate shield! Multi-Conditioned Counterspeech Generation through Attributed Prefix Learning

Aswini Kumar[1], **Anil Bandhakavi**[2], **Tanmoy Chakraborty**[1]
[1]IIT Delhi, India, [2]Logically.ai
{ aswinikumarpadhi1995@gmail.com, }
anil@logically.ai, tanchak@iitd.ac.in

Abstract

Counterspeech has proven to be a powerful tool to combat hate speech online. Previous studies have focused on generating counterspeech conditioned only on specific strategies (single attributed). However, a holistic approach considering multiple attributes simultaneously can yield more nuanced and effective responses. Here, we introduce `HiPPrO`, **Hi**erarchical **P**refix learning with **Pr**eference **O**ptimization, a novel two-stage framework that utilizes the effectiveness of attribute-specific prefix embedding spaces hierarchically optimized during the counterspeech generation process in the first phase. Thereafter, we incorporate both reference and reward-free preference optimization to generate more constructive counterspeech. Furthermore, we extend `IntentCONANv2` by annotating all $13,973$ counterspeech instances with emotion labels by five annotators. `HiPPrO` leverages hierarchical prefix optimization to integrate these dual attributes effectively. An extensive evaluation demonstrates that `HiPPrO` achieves a $\sim 38\%$ improvement in strategy conformity and a $\sim 3\%$, $\sim 2\%$, $\sim 3\%$ improvement in Rouge-1, Rouge-2, and Rouge-L, respectively, compared to several baseline models. Human evaluations further substantiate the superiority of our approach, highlighting the enhanced relevance and appropriateness of the generated counterspeech. This work underscores the potential of multi-attribute conditioning in advancing the efficacy of counterspeech generation systems.[1] Our code is available on Github and dataset is open-sourced on Hugging-face.

1 Introduction

The rise of the Internet has transformed social media platforms into hostile grounds for hateful comments targeting specific communities. Hate speech (HS) (Awal et al., 2021; Chakraborty and Masud, 2022; Masud et al., 2021a, 2023) carries offensive statements that leverage stereotypes to spread harmful content. In contrast, counterspeech (CS) (Benesch et al., 2016a; Wright et al., 2017; Singh et al., 2024) involves constructive responses to counteract hate speech by promoting positive dialogue, thus mitigating online hostility while preserving diverse perspectives (Schieb and Preuss, 2016; Chandrasekharan et al., 2017). As hateful comments surge (Leetaru, 2019; Masud et al., 2021b, 2022; Yadav et al., 2024), depending only on human-generated counterspeech becomes impractical. In this context, machine learning models appear essential for analyzing and generating counterspeech, offering a promising solution for automating the process. By leveraging this, content moderators can efficiently counter online hate (Parker and Ruths,

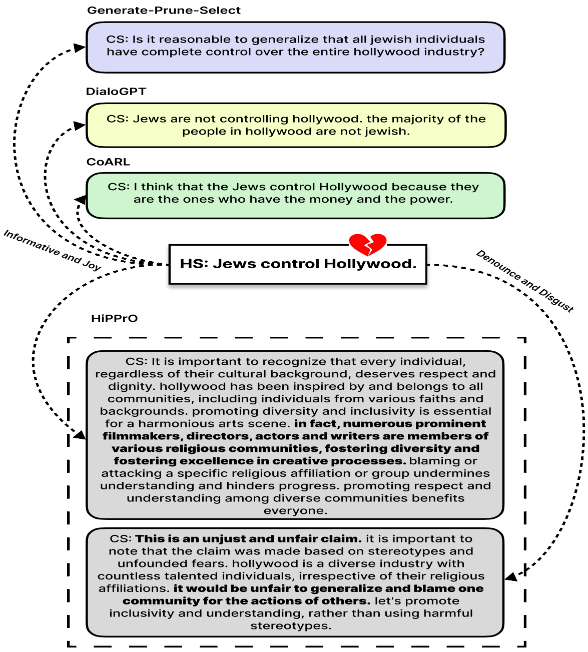


Figure 1: An illustration of the output of existing methods in generating effective counterspeech, `HiPPrO` (ours) demonstrates superior performance in producing high-quality, effective, multi-attributed counterspeech for a given hate speech without relying on instructional prompts.

[1] *Warning: The materials presented in this paper might be disturbing or offensive.*

Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 27643–27663
July 27 - August 1, 2025

2023; Garg et al., 2023; Chung et al., 2023). Consequently, several interests (Mathew et al., 2019; Qian et al., 2019; Fanton et al., 2021a; Bonaldi et al., 2022; Hengle et al., 2024) have intensified in the development of counterspeech generation methods.

The conceptualization of CS generation has evolved from a simple sequence-to-sequence problem to a more nuanced approach, acknowledging the diverse and context-dependent nature of hate speech (Chung et al., 2019; Mathew et al., 2019; Sheng et al., 2020; Parker and Ruths, 2023; Chung et al., 2023). This paradigm shift has fostered the development of advanced generative models incorporating stylistic and condition-guided elements, such as politeness, joyfulness, and detoxification, to create more effective counter-narratives (Saha et al., 2022; Sheng et al., 2020). Recent research has introduced strategy-specific CS generation, where established strategies guide the generation process to combat hate speech (Gupta et al., 2023; Hengle et al., 2024; Benesch et al., 2016b).

Motivation: A significant portion of online hate speech consists of short, abusive statements (Benesch et al., 2016a), and CS has shown potential in effectively countering such harmful content. While conventional approaches generate a single response per hate speech instance (Zhu and Bhat, 2021; Qian et al., 2019), recent studies highlight the benefits of tailoring CS to specific attributes for generating more diverse responses (Gupta et al., 2023; Hengle et al., 2024). In practical scenarios, hateful comments often include multiple user intentions, necessitating the development of counterspeech that effectively addresses diverse attributes, resulting in more comprehensive and effective responses. For instance, single-attribute approaches typically produce one-dimensional responses by focusing solely on strategy (e.g., being informative) while neglecting the emotional harmony required for persuasive communication. Real-world hate speech frequently needs a range of emotional responses, making it imperative to generate counterspeech that balances factual accuracy with emotional engagement.

The integration of Large Language Models (LLMs) for various text generation tasks has become increasingly popular (Yang et al., 2024), but training these models is resource intensive. To address this, parameter-efficient fine-tuning (PEFT) techniques, such as tunable prefixes (Li and Liang, 2021), have gained popularity. These techniques involve adding task-specific continuous vectors (key-value pairs) to transformer layers while keeping the rest of the model unchanged. Recent studies (Liu et al., 2023a) showed that these vectors excel in generating conditional text by capturing hidden implied relationships during training. In this work, we propose generating multi-attribute guided counterspeech generation through a hierarchical approach to learning continuous prefix vectors, enabling more varied and contextually relevant responses. We provide an example in Figure 1, where our method, `HiPPrO`, is compared to leading models like Generate-Prune-Select (GPS) (Zhu and Bhat, 2021), DialoGPT (Zhang et al., 2020b), and COARL (Hengle et al., 2024). While traditional models produce semantically sound responses, they often lack nuance and struggle with complex contextual relationships, as mentioned in (Benesch et al., 2016a). Inspired by Liu et al. (2023b), `HiPPrO` enables to generate more effective and contextually relevant counterspeech by analyzing hate speech and multiple user intentions, resulting in more impactful and persuasive responses.

Our Contribution: This study introduces an advanced pipeline for counterspeech generation, addressing the implicit nature of hate speech with responses aligned to multiple attributes. We focus on four primary strategies– 'positive', 'informative', 'questioning', and 'denouncing.' and five emotion categories –'anger', 'disgust', 'joy', 'sadness', and 'surprise'. Furthermore, we curate `MultiCONAN`, the largest collection of strategy-emo-specific counterspeech, with $13,952$ responses countering $3,487$ instances of hate speech. We propose `HiPPrO`, a novel two-phase framework that first learns attribute-specific prefix embeddings (key-value pairs) and then applies preference tuning to generate constructive, non-toxic responses. Comprehensive evaluations using automated metrics and human evaluations demonstrate that `HiPPrO` consistently outperforms existing methods in CS generation across various criteria and achieves comparable performance with state-of-the-art LLMs like GPT-3.5 and GPT-4.

2 Related Work

The evolution of CS datasets has progressed from crowdsourced collections to expert-curated, strategy-specific compilations (Qian et al., 2019; Chung et al., 2019; Fanton et al., 2021a; Gupta et al., 2023; Hengle et al., 2024), while CS generation techniques have advanced from basic sequence-to-sequence models to sophisticated multi-phase pipelines with attribute control (Zhu and Bhat, 2021; Saha et al., 2019). These advancements have significantly improved the nuance and effectiveness of automated CS in promoting constructive dialogue (Benesch et al., 2016c). Concurrently, parameter-efficient fine-tuning methods like Prefix Tuning (Li and Liang, 2021) and prompt tuning (Lester et al., 2021) have emerged, modifying inputs while preserving language model parameters. Recent developments in preference tuning include RLHF's application to instruction-following tasks (Ouyang et al., 2022a), Direct Preference Optimization (DPO) (Rafailov et al., 2023), and a unified approach omitting both reward and reference models (Hong et al., 2024), addressing challenges in scalability and model sensitivity.

3 The `MultiCONAN` Dataset

Due to the superior quality of counterspeech (see Appendix 9.3), we selected `IntentCONANv2` (Hengle et al., 2024) as the foundation for our work. `IntentCONANv2`, while valuable, has limitations that restrict its ability to fully capture the diversity and complexity of counterspeech. Its dependence on a single attribute, 'strategy,' oversimplifies responses and fails to account for the emotional tone, which plays a critical role in shaping effective counterspeech. For instance, an informative counterspeech can convey vastly different emotional tones, such as joy or sadness, which are not captured in the original dataset. Building upon this, we introduce `MultiCONAN`, an enhanced version featuring additional emotion class labels for each counterspeech instance. `MultiCONAN` includes $13,973$ CS instances of `IntentCONANv2`, each tagged with one of five emotion classes: Anger (AN), Joy (JO), Disgust (DI), Sad (SA), and Surprise (SU) (See Table 1). The added emotion annotations enhance analysis granularity and facilitate the development of models that integrate both strategy and emotional context. This annotation framework enables exploration of how emotional tone interacts with strategic strategy and supports the creation of nuanced CS generation models (Gupta et al., 2023; Hengle et al., 2024). `MultiCONAN` thus serves as a valuable resource for advancing research in CS generation, aiming to produce contextually appropriate and emotionally resonant responses. For detailed information on the annotation process and the statistics of the data set, refer to Appendices 9.1,9.2, and 9.4, respectively.

4 Proposed Methodology

In this section, we elaborate on the inner workings and structural components of `HiPPrO`, a novel automated counterspeech generation framework. Here, we explain how our designed model can address the previous challenges by (i) generating multi-attribute conditioned counterspeech that can address hateful comments through *semantic relevancy*, and (ii) aligning it with qualitative human-generated responses through a reward and reference-model-free approach to ensure that the generation is *contrastive* (Figure 2).

Task Formulation

We use our curated `MultiCONAN` dataset for generating multi-attribute counterspeech generation. Considering our dataset as $\mathcal{D} = \{(h_1, i_1, e_1, c_1), \ldots, (h_n, i_n, e_n, c_n)\}$, where $h_i \in \mathcal{H}$ is the i-th hate speech statement, $c_i \in \mathcal{C}$ is the counterspeech corresponding to h_i, and $i_i \in \mathcal{I}$ and $e_i \in \mathcal{E}$ are the strategy and emotion categories of c_i, respectively. Our objective is to learn a stochastic counterspeech generation function $\psi : \mathcal{H} \times \mathcal{I} \times \mathcal{E} \rightarrow \mathcal{C}$, such that $c_i \sim \psi(\cdot|h_i, i_i, e_i)$.

We address this problem by decomposing the counterspeech generation task into two phases. In the first phase, we focus on learning the prefix vectors for individual attributes and capturing their conditional dependencies using a two-step method. Initially, we learn the prefix vector for the strategy category and collect the optimal prefix vectors. Subsequently, we add another prefix vector initialized with the previously learned values and optimized it for both the strategy and emotion categories while keeping the model parameters from step one fixed. In the second phase, we employ a reward and reference model-free alignment method called the odds ratio preference optimization algorithm (Hong et al., 2024). During this, we treat the counterspeech generated in phase one as the rejected data column and the actual ground truth counterspeech as the selected data column. For all our experiments, we utilize FLAN-T5 (Chung et al., 2022) as the base model due to its robust reasoning and multi-task learning capabilities.

	Counterspeech strategy and Emotion																				
Target	**IN and AN**	**IN and DI**	**IN and JO**	**IN and SA**	**IN and SU**	**DE and AN**	**DE and DI**	**DE and JO**	**DE and SA**	**DE and SU**	**PO and AN**	**PO and DI**	**PO and JO**	**PO and SA**	**PO and SU**	**QU and AN**	**QU and DI**	**QU and JO**	**QU and SA**	**QU and SU**	**Total**
Muslim	236	269	248	149	19	221	420	232	33	9	187	55	622	54	1	397	147	24	9	337	3669
Women	108	159	161	71	10	151	181	152	14	10	59	44	378	26	1	258	91	15	11	134	2034
LGBT+	93	171	129	48	8	140	208	85	12	4	45	36	335	31	2	190	96	11	15	137	1796
Jews	123	113	94	53	10	109	206	54	14	9	73	32	251	36	1	151	67	5	5	164	1570
Refugee	10	8	37	12	3	10	22	33	5	0	5	4	55	6	0	27	11	3	1	28	280
AP	7	10	6	5	1	9	16	4	0	0	4	1	22	2	0	11	4	1	2	11	116
IMGT	99	101	260	87	15	147	137	244	29	5	64	22	441	32	3	242	76	30	14	200	2248
Disable	30	51	65	28	1	33	75	59	7	1	12	17	125	18	1	100	18	6	7	42	696
PoC	62	122	68	47	3	80	148	54	13	7	45	13	223	20	1	122	67	8	5	100	1208
Others	25	23	28	11	2	28	37	16	6	2	19	7	57	5	1	38	19	3	1	28	356
Total	793	1027	1096	511	72	928	1450	933	133	47	513	231	2509	230	11	1536	596	106	70	1181	13973
Train	505	656	686	325	39	605	912	544	84	30	294	142	1594	132	7	978	381	68	47	772	8801
Dev	122	158	161	90	10	152	247	140	29	6	75	32	425	49	3	235	87	11	10	159	2201
Test	166	213	249	96	23	171	291	249	20	11	144	57	490	49	1	323	128	27	13	250	2971

Table 1: Counterspeech distribution in `MultiCONAN` across various multi-attribute combinations, categorized by target groups (see abbreviations in Section 3).

Phase 1: Hierarchical Prefix Optimization (HIPO)

In this section, we mathematically formalize the concept of hierarchical prefix learning. The prefix vectors, represented by tunable key-value pairs (Vaswani et al., 2017), are introduced in two sub-phases. In the first sub-phase, we add $|V_{\mathcal{I}}|$ virtual prefixes with dimension d across l layers. The prefix adapters, $\mathcal{F}_\alpha$ and $\mathcal{F}_\beta$, introduce task-specific continuous vectors, $\alpha, \beta \in \mathbb{R}^{|V_{\mathcal{I}}| \times l \times 2d}$ to the encoder and decoder, respectively, guiding counterspeech generation according to user strategy.

We maximize the expected log-likelihood and collect the optimal prefix adapters $\mathcal{F}_{\alpha^*}, \mathcal{F}_{\beta^*}$ for the strategy-guided counterspeech generation, where the hate speech and strategy are sampled from $\mathcal{D}$, as denoted by,

$$\mathcal{F}_{\alpha^*}, \mathcal{F}_{\beta^*} = \underset{\alpha,\beta}{\operatorname{argmax}}\, \mathbb{E}_{(h,i \sim \mathcal{D})} \log(\mathcal{F}_\alpha[\pi_\theta^{ENC}(X(h,i); \theta, \alpha)], \mathcal{F}_\beta[\pi_\theta^{DEC}(X(c); \theta, \beta)]) \tag{1}$$

where π_θ^{ENC} and π_θ^{DEC} are the encoder and decoder part of the model, respectively. The input $X = [h; i; c]$, where $(h, i, c) \in \mathcal{D}$. We further add another set of adapters, $\mathcal{F}_\gamma$, and $\mathcal{F}_\delta$, on the encoder and decoder side, which add trainable prefix vectors $\gamma, \delta \in \mathbb{R}^{|V_{\mathcal{E}}| \times l \times 2d}$ on top of the previously trained $\mathcal{F}_{\alpha^*}, \mathcal{F}_{\beta^*}$, where $|V_{\mathcal{E}}|$ is the number of virtual tokens for second sub-phase. These prefix vectors are responsible for guiding the counterspeech generation with strategy and emotion-specific attributes. Here, the reformulated input is $X' = [h; i; e; c]$, where $(h, i, e, c) \in \mathcal{D}$. During training, we freeze the model parameters θ and the strategy-specific prefix parameters α, β and maximize the expected log-likelihood to get optimal prefix adapters $\mathcal{F}_{\gamma^*}, \mathcal{F}_{\delta^*}$ for optimal γ^*, δ^* given by,

$$\mathcal{F}_{\gamma^*}, \mathcal{F}_{\delta^*} = \underset{\gamma,\delta}{\operatorname{argmax}}\, \mathbb{E}_{(h,i,e \sim \mathcal{D})} \log(\mathcal{F}_\gamma[\mathcal{F}_{\alpha^*}[\pi_\theta^{ENC}(X'(h,i,e); \theta, \alpha^*, \gamma)]], \mathcal{F}_\delta[\mathcal{F}_{\beta^*}[\pi_\theta^{DEC}(X'(c); \theta, \beta^*, \delta)]]]) \tag{2}$$

Here, all $\alpha^*, \beta^*, \gamma^*$ and δ^* are responsible for the strategy and emotion-conditioned counterspeech generation. Now, our model is ready to advance to Phase 2, where we apply a preference-tuning approach to all pre-trained prefix vectors. This step further optimizes the model to align counterspeech responses more closely with human-generated outputs.

Phase 2: Preference Optimization

In this phase, we focus on optimizing our model to produce counterspeech instances that are both effective and non-toxic by employing a reward and reference-free preference tuning approach. Inspired by Hong et al. (2024), we use the odd ratio method to align the output with the ground truth preferences. This method self-penalizes the model output probabilities using its corresponding odds, denoted as,

$$Odds_{\alpha,\beta,\gamma,\delta}(y|x) = \frac{\pi_{\theta,\alpha,\beta,\gamma,\delta}(y|x)}{1 - \pi_{\theta,\alpha,\beta,\gamma,\delta}(y|x)} \tag{3}$$

Now this ratio has an interesting characteristic – it boils down to less than one when the desired

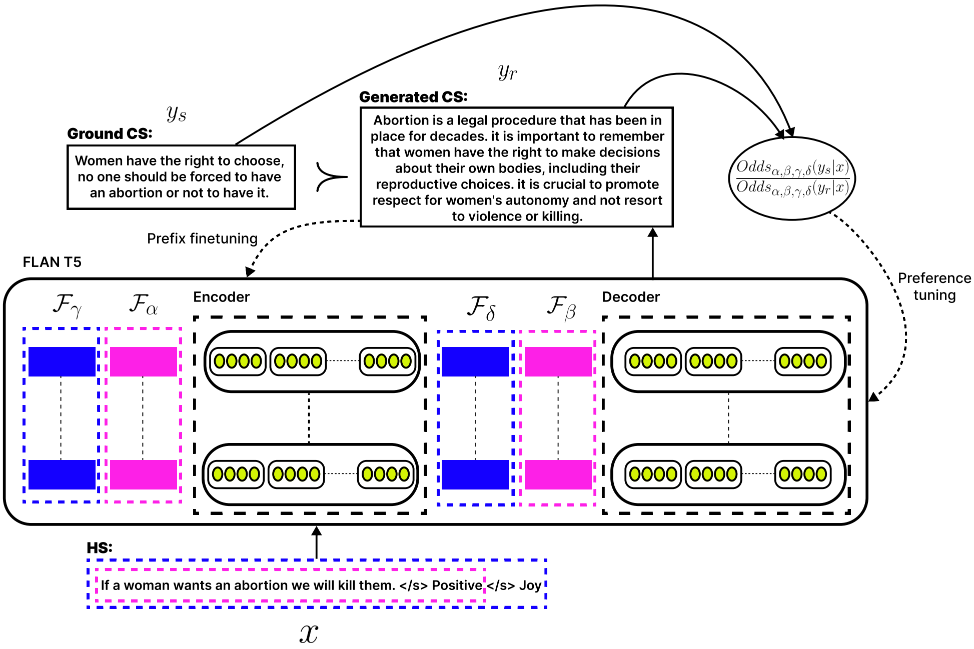


Figure 2: Our proposed model, HiPPrO, follows a two-phase pipeline. In Phase 1, we train the prefix parameters $(\alpha, \beta, \gamma, \delta)$ associated with prefix adapters $(\mathcal{F}_\alpha, \mathcal{F}_\beta)$ on encoder side and $(\mathcal{F}_\gamma, \mathcal{F}_\delta)$ on the decoder side. For α and β, the model is trained with hate speech and strategy, separated by </s> as input. While training γ and δ, the optimal parameters, α^* and β^*, are kept fixed, and it includes hate speech, strategy, and emotion, separated by </s> as input x, this is also used as a prompt input during the second phase. In Phase 2, we apply preference tuning using an odd ratio loss, where the ground-truth counterspeech serves as the chosen candidate y_s, and the model-generated counterspeech is treated as the rejected candidate y_r.

probability is less than its odds, which eventually penalizes the model output and gives a smaller number. The odd ratio is the ratio between two odds of two independent events. The odd ration OR is given by,

$$OR_{\alpha,\beta,\gamma,\delta}(y_s, y_r) = \frac{Odds_{\alpha,\beta,\gamma,\delta}(y_s|x)}{Odds_{\alpha,\beta,\gamma,\delta}(y_r|x)} \quad (4)$$

Here, y_s denotes the ground-truth counterspeech, which is human-generated, and y_r denotes the model-generated counterspeech, which is trained during the first stage. $OR_{\alpha,\beta,\gamma,\delta}(y_s, y_r)$ becomes high when the numerator is higher than the denominator, and hence $\pi_{\theta,\alpha,\beta,\gamma,\delta}(y_s|x)$.

Let us consider the preference tuning dataset $\mathcal{D}' = \{(h_1, y_{s1}, y_{r1}), \ldots, (h_n, y_{sn}, y_{rn})\}$, where y_{si} and y_{ri} are the i^{th} ground-truth response and model-generated response for a hate speech h_i. The final loss is the expectation of prefix-tuned loss and odds ratio loss over the data samples sampled from $\mathcal{D}'$. As shown in Equation 5, the final loss for preference optimization J_{final} is the combined loss for finetuned loss ($J_{finetuned}$) and odds ratio loss J_{OR} weighted by a factor ϵ.

$$J_{final} = \mathbb{E}_{(x,y_s,y_r \sim \mathcal{D}')}[J_{finetuned} + \epsilon J_{OR}] \quad (5)$$

The $J_{finetuned}$ is simply the negative log-likelihood loss obtained in the prefix-tuning step, and J_{OR} is given by,

$$J_{OR} = -\log(\sigma(\log(\frac{Odds_{\alpha,\beta,\gamma,\delta}(y_s|x)}{Odds_{\alpha,\beta,\gamma,\delta}(y_r|x)}))) \quad (6)$$

where σ denotes the sigmoid function. The final objective is to minimize J_{final} by updating the continuous prefix vectors so that the model outputs can align more with the ground-truth counterspeech. For detailed information regarding computational resources and hyperparameter settings, please refer to Appendix 9.9 and Appendix 9.10, respectively. [2]

5 Experimental Setup

5.1 Baselines

To evaluate the efficacy of various models, we investigate **Generate Prune Select (GPS)** (Zhu and Bhat, 2021), a three-stage pipeline encompassing autoencoding, grammatical filtering, and response selection. Furthermore, we optimize **DialoGPT** (Zhang et al., 2020b) to produce contextually coherent responses. Additionally, we incorporate **CoARL** (Hengle et al., 2024), the state-of-the-art strategy-conditioned CS generation approach. Furthermore, we explore prefix tuning

[2]The model was trained for up to 50 epochs using a callback, with a fixed training batch size and a learning rate of 4 and $1 \times e^{-4}$, respectively.

Method	Prompt/Adapter	ROUGE (↑)			M (↑)	BS (↑)	CoSim (↑)	SC (↑)	EC (↑)	TC (↑)	T (↓)
		R1	R2	RL							
GPS	–	0.089	0.011	0.075	0.055	0.840	0.287	0.247	0.292	0.394	0.146
DialoGPT	–	0.164	0.050	0.108	0.198	0.749	0.536	<u>0.714</u>	0.571	0.969	0.309
CoARL	–	0.156	0.035	0.126	0.110	0.860	0.440	0.380	0.468	0.938	0.258
Vanilla FLAN-T5$_{\text{XXL}}$	ZS	0.176	0.045	0.145	0.118	0.866	0.479	0.391	0.586	**0.989**	0.341
Vanilla FLAN-T5$_{\text{XXL}}$	FS	0.174	0.042	0.142	0.116	0.864	0.456	0.352	0.478	0.976	0.301
GPT-3.5-Turbo	ZS	0.239	0.052	0.160	**0.249**	0.868	**0.585**	0.533	<u>0.783</u>	<u>0.978</u>	**0.0007**
GPT-3.5-Turbo	FS	<u>0.242</u>	<u>0.059</u>	<u>0.161</u>	**0.249**	0.868	0.564	0.309	0.527	0.888	<u>0.012</u>
GPT-4	ZS	0.221	0.037	0.145	0.229	0.864	<u>0.583</u>	0.545	**0.790**	0.976	0.030
GPT-4	FS	0.226	0.039	0.149	0.213	0.868	0.554	0.454	0.515	0.932	0.014
FLAN-T5$_{\text{XXL}}$	Retrieval-based	0.188	0.031	0.132	0.160	0.862	0.494	0.379	0.549	0.907	0.097
GPT2$_{\text{XL}}$	Retrieval-based	0.113	0.011	0.074	0.154	0.821	0.366	0.309	0.441	0.754	0.103
Llama-3.1-8B-Instruct	Retrieval-based	0.128	0.026	0.086	0.197	0.828	0.462	0.347	0.678	0.891	0.096
Mistral-7B-Instruct-v0.2	Retrieval-based	0.166	0.033	0.107	0.224	0.847	0.530	0.356	0.782	0.932	0.068
DeepSeek-R1-Distill-Llama-8B	Retrieval-based	0.132	0.027	0.086	0.210	0.830	0.471	0.373	0.751	0.890	0.059
Vanilla FLAN-T5 $_{\text{XXL}}$	PrefixTuning	0.229	0.052	0.158	0.222	<u>0.870</u>	0.539	0.470	0.666	0.901	0.024
Vanilla BART $_{\text{Large}}$	PrefixTuning	0.207	0.042	0.131	0.226	0.861	0.226	0.249	0.453	0.915	0.030
Vanilla GPT2 $_{\text{XL}}$	PrefixTuning	0.155	0.029	0.122	0.084	0.820	0.487	0.453	0.455	0.937	0.837
Vanilla Llama 3.1 Instruct $_{\text{8B}}$	PrefixTuning	0.168	0.043	0.135	0.093	0.854	0.515	0.552	0.592	0.903	0.605
Vanilla Mistral Instruct $_{\text{7B}}$	PrefixTuning	0.170	0.046	0.140	0.117	0.858	0.510	0.516	0.597	0.920	0.645
DeepSeek-R1-Distill-Llama-8B	PrefixTuning	0.035	0.0002	0.034	0.010	0.808	0.181	0.261	0.289	0.582	0.023
DeepSeek-llm-7b-chat	PrefixTuning	0.040	0.001	0.035	0.027	0.805	0.144	0.261	0.311	0.484	0.019
HiPPrO $_{\text{VT=3}}$ (Ours)	PrefixTuning	**0.273***	**0.082***	**0.199***	<u>0.242</u>	**0.879***	0.567	**0.929***	0.706	0.897	0.087
- With out ORPO	PrifixTuning	0.272	0.081	0.198	0.241	0.879	0.567	0.928	0.705	0.896	0.111
- With DPO	PrefixTuning	0.272	0.081	0.198	0.240	0.879	0.566	0.928	0.685	0.895	0.089
- HIPO $_{\text{VT=5}}$	PrifixTuning	0.275	0.081	0.200	0.241	0.880	0.578	0.921	0.656	0.916	0.096
- HIPO $_{\text{VT=7}}$	PrefixTuning	0.273	0.084	0.199	0.24	0.879	0.573	0.937	0.643	0.888	0.084
- HIPO $_{\text{VT=10}}$	PrefixTuning	0.271	0.081	0.197	0.244	0.878	0.564	0.905	0.630	0.858	0.077
$\Delta_{\text{HiPPrO(Ours)}-\text{BestBaselineMethod}}$		↑ 0.034	↑ 0.023	↑ 0.038	↓ 0.007	↑ 0.009	↓ 0.018	↑ 0.384	↓ 0.084	↓ 0.092	↓ 0.086

Table 2: Comparing HiPPrO with baselines across various evaluation metrics. Here, ↑ (*resp.* ↓) denotes that higher (*resp.* lower) is better. **Bold** (*resp.* <u>underline</u>) indicates the best (*resp.* second-ranked) performance. $*$ shows our model *significantly* outperforms ($p < 0.05$) the best baselines – GPT-3.5-Turbo ZS and FS (see Appendix 9.5).

on **Vanilla FLAN-T5**$_{\text{XXL}}$, and **BART**$_{\text{Large}}$ (excluding HiPPrO) and conduct experiments with HIPO (without preference optimization) using different virtual token (VT) sizes (VT = $3, 5, 7, 10$), VT = 3 emerging as the optimal configuration in terms of both parameter efficiency ($589,824$ trainable parameters, 0.0052% of total model parameters) and performance metrics. We consider VT = 3 for preference tuning. With same setup, we experiment with decoder-only models like **GPT2**$_{\text{XL}}$, **Llama 3.1 Instruct**$_{\text{8B}}$ (Grattafiori et al., 2024), **Mixtral Instruct**$_{\text{7B}}$ (Jiang et al., 2024), **DeepSeek-R1-Distill-Llama-8B** and **DeepSeek-llm-7b-chat** (DeepSeek-AI et al., 2025). We also utilize DPO to enhance our evaluation further. Our comprehensive assessment encompasses zero-shot and few-shot performances on three LLMs: **Vanilla FLAN-T5**$_{\text{XXL}}$ (Chung et al., 2022), **GPT-3.5-Turbo** (ChatGPT), and **GPT-4** (Ouyang et al., 2022b) (see Appendix 9.6 and 9.7) and simple retrieval-based methods using state-of-the-art open-source LLMs, employing 'faiss'(Douze et al., 2025) as the retrieval method to retrieve the top five training counterspeech examples (see Appendix 9.8).

5.2 Evaluation Metrics

Evaluating CS generation presents challenges due to its dynamic nature, diverse response possibilities, and absence of standardized metrics (Chung et al., 2023), prompting our framework to utilize comprehensive, multi-dimensional evaluation metrics. These evaluation metrics include lexical similarity, semantic similarity or relevance, strategy conformity, emotion conformity, target conformity, and toxicity score. *Lexical similarity* is evaluated using **Rouge** (Lin, 2004) and **Meteor (M)** (Banerjee and Lavie, 2005), which quantify the linguistic alignment between generated and reference texts. *Semantic relevance* is assessed with **cosine similarity (CoSim)** (Reimers and Gurevych, 2019) and **BERTScore (BS)** (Zhang et al., 2020a), ensuring that generated CS engages meaningfully with the primary topic of hate speech. A low relevance score implies a lack of topical coherence, where the CS fails to adequately address the primary subject of hate speech. We also evaluate the effectiveness of incorporating strategic, emotional resonance, and target alignment through **Strategy Conformity (SC)** (Gupta et al., 2023), **Emotion Conformity (EC)**, and **Target Conformity (TC)**, respectively. These metrics are particularly valuable in scenarios where ground-truth counterspeech is unavailable for unseen hate speech instances. We first train three distinct RoBERTa-large models on our dataset to measure the SC, EC, and TC scores. The models achieve testing accuracy of 0.86, 0.75, and 0.88, respectively, before being considered as evaluation metrics. ***Toxicity*** (T)[3] levels of generated CS using the (Hanu and Unitary team, 2020) library, ensure that our approach promotes respectful and safe communication.

6 Experimental Results

This section presents a comprehensive empirical analysis that systematically evaluates the efficacy of `HiPPrO` in comparison to existing state-of-the-art techniques.

6.1 Quantitative Results

Table 2 demonstrates the quantitative evaluation across various metrics. `HiPPrO` shows a notable improvement over baselines across several evaluation metrics. `HiPPrO` achieves 0.273 ROUGE-1, 0.082 ROUGE-2, 0.199 ROUGE-L, substantially higher than GPS, DialoGPT, and CoARL, with an average improvement of 0.117 in ROUGE-1, 0.05 in ROUGE-2 and 0.096 in ROUGE-L. This indicates that

`HiPPrO`'s counterspeech better aligns with reference content in terms of coverage and detail. In BERTScore, `HiPPrO` scores 0.879, surpassing GPS (0.840), DialoGPT (0.749), and CoARL (0.860), GPT-4 Few-Shot (FS) (0.868) and GPT-3.5 Few-Shot (0.868), which indicates `HiPPrO`'s proficiency in preserving the underlying meaning of the original CS. `HiPPrO` exhibits a slight decrease in CoSim compared to models like GPT-3.5-Turbo Zero-Shot (ZS) with a score of 0.585, indicating a potential trade-off between contextual relevance and semantic alignment with reference responses.

The generated CS should ideally be a balanced generation of both strategy and emotional attributes, ensuring that responses not only address the harmful content effectively but also align with strategy and emotional tone along with the target audience. However, the evaluation results indicate that it is a challenge for many models to achieve this balance. For instance, while GPT-3.5-Turbo ZS demonstrates a low SC score of 0.533, its EC score is significantly high at 0.783, highlighting the difficulty in generating CS that is both strategy-aligned and emotionally resonant. Similarly, GPT-4 ZS shows a high EC score of 0.790, yet its SC score of 0.545 suggests that even more advanced LLMs struggle to maintain emotional alignment in the zero-shot and few-shot scenarios. In contrast, `HiPPrO` demonstrates a better balance, achieving an SC score of 0.929 and an EC score of 0.706. `HiPPrO`'s SC score is substantially higher, indicating its superior ability to generate strategy-aligned CS while still maintaining a better level of emotional resonance. This suggests that `HiPPrO` is more adept at maintaining the stability between these two critical attributes during the generation process.

The relationship between TC and BS provides insights into how models generate counterspeech; a high TC score with a relatively low BS indicates that the CS is target-specific but may lack deeper semantic alignment with the original CS. For instance, GPT-3.5-Turbo in the ZS setting achieves a TC score of 0.978 and a BS score of 0.868, suggesting that while it effectively mentions the target group, it may produce overly generic responses that fail to address nuanced aspects of hate speech. In contrast, `HiPPrO` manages to strike a better balance by achieving a TC score of 0.897, which, while slightly lower than GPT-4 few-shot, is complemented by a BS of 0.879, the highest among all models. This indicates that `HiPPrO` is not only attentive to the target group but also maintains a strong semantic connection to the original content, making its CS both specific and contextually relevant. Furthermore, when examining the toxicity scores, `HiPPrO` achieves a respectable score of 0.087, which, although slightly higher than GPT-3.5-Turbo Zero-Shot, still indicates a significantly low presence of harmful language comparable to all other baselines. This suggests that `HiPPrO` effectively balances generating contextually rich CS while keeping the content non-toxic and constructive. Statistical analyses reveal that our model

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

significantly outperforms both GPT-3.5 ZS and FS across most metrics, with exceptions in meteor, CoSim, and TC scores. Please refer to Appendix (Section 9.5) for additional information about statistical significance tests.

6.2 Ablation Study

Our ablation study assesses the effects of several components of HiPPrO (Table 2). One such study involves fine-tuning HiPPrO without the ORPO component. The results show that the omission of ORPO leads to a decrease in performance in almost all evaluation metrics, with a notable reduction in the toxicity score by 0.024. This suggests that ORPO contributes significantly to enhancing both the overall performance and the generation of less toxic counterspeech. Additionally, we fine-tune HiPPrO with DPO instead of ORPO and observe a slight degradation in performance. Although the difference is marginal, the key advantage of ORPO over DPO lies in ORPO's ability to operate without a reference model, which DPO requires. Additionally, we investigate how different numbers of VT in HiPPrO impact its performance, testing configurations with VT= $5, 7$, and 10. The results indicate that there is no notable performance gain after VT=5. While higher virtual tokens lead to a slight reduction in the toxicity score, the improvements are insufficient across other metrics, suggesting no significant returns with increased token counts.

Our ablation studies also include evaluations of Vanilla FLAN-T5$_{\text{XXL}}$ and Vanilla BART$_{\text{Large}}$, both with prefix tuning, without hierarchical learning, providing further insights into HiPPrO's effectiveness. We observe that while these models offer competitive performance, they do not surpass HiPPrO in several key metrics. For Vanilla FLAN-T5$_{\text{XXL}}$, prefix tuning yields ROUGE-1, ROUGE-2, and ROUGE-L scores of 0.229, 0.052, and 0.158, respectively. These scores are slightly lower than those of HiPPrO, indicating that while the prefix-tuned FLAN-T5$_{\text{XXL}}$ performs well, HiPPrO's output is more detailed and comprehensive. However, the SC scores for Vanilla FLAN-T5$_{\text{XXL}}$ and BART$_{\text{Large}}$ are 0.470 and 0.249, respectively, while their EC scores are 0.666 and 0.453, respectively. HiPPrO achieves higher SC and EC scores (0.929 and 0.706 respectively), highlighting the limitations of simple prefix-tuning methods in effectively aligning CS with multiple attributes.

Models on comparison	Metrics			
	ICS ↑	Ad ↑	CoRl ↑	ArgE ↑
HiPPrO vs CoARL	0.96	0.98	0.93	0.97
HiPPrO vs GPT-4 (FS)	0.91	0.85	0.89	0.89
HiPPrO vs GPT-3.5 (FS)	0.89	0.87	0.87	0.90

Table 3: Results of the human evaluation study, where responses generated by HiPPrO are shown against those produced by (a) CoARL, (b) GPT-4 (FS), and (c) GPT-3.5 (FS). The results are reported in terms of Win Rate %, indicating the % of instances where HiPPrO outperforms the respective baselines.

6.3 Human Evaluation

Previous studies (Jones et al., 2024; Wang et al., 2023; Hengle et al., 2025) emphasize the need for a dual evaluation framework, as automatic metrics show weak correlation with human judgments of counterspeech effectiveness. A comprehensive human evaluation was conducted on a random subset of 30 responses from the top-performing CS generation methods (HiPPrO, CoARL, Few-Shot GPT-4, and GPT-3.5 Turbo) with a random seed value of 1, ensuring uniform distribution across strategies and emotions. A diverse panel of 35 experts in NLP and social sciences (aged 20-35, 45% male, 55% female) evaluated and ranked these responses based on several key metrics. We followed Hengle et al. (2024) for our human evaluation. The evaluation framework consists of five key metrics: Independent Counterspeech (ICS) to gauge the response's self-sufficiency; Adequacy (Ad) to assess its linguistic quality; Contextual Relevance (CoRl) to measure its responsiveness to hate speech components; and Argumentative Effectiveness (ArgE) to evaluate its carefulness and convincing. We present the comparative performance of HiPPrO against leading methods through *Win Rate* scores (see Table 3). For ICS, HiPPrO outperforms the baselines with win rates of 0.91 and 0.89 over GPT-4 and GPT-3.5, showing its ability to generate CS that operates effectively without needing extra context. Regarding Ad, HiPPrO achieves higher scores of 0.85 and 0.87, reflecting the superior grammatical accuracy and fluency of CS. In terms of CoRl, HiPPrO's win rates of 0.89 and 0.87 highlight its strength in addressing crucial aspects of hate speech, such as targeted biases. Finally, for ArgE, HiPPrO leads with scores of 0.89 and 0.90, indicating its effectiveness in delivering compelling and well-structured CS. These results collectively show HiPPrO's robust performance across all metrics compared to the baseline models.

7 Conclusion

This study presented HiPPrO, a novel two-stage framework for generating controllable, multi-attributed counterspeech. The initial stage comprised a hierarchical learning process, where the model acquired attribute-specific prefixes, thereby guiding the LLM towards targeted counterspeech generation. The subsequent stage involved refining the outputs to enhance their human-like quality and non-toxicity, employing a reward- and reference-free alignment approach. Additionally, we introduced the multiCONAN dataset with strategy- and emotion-specific counterspeech. An extensive evaluation, incorporating a range of quantitative and qualitative measures, demonstrated HiPPrO's superiority over multiple baselines.

Limitation

Our research presents several limitations that warrant consideration. Firstly, the dataset utilized for hate speech and counterspeech is not comprehensive, potentially omitting various forms and targets of online hate. Secondly, the framework's reliance on pre-trained models may introduce inherent biases or inaccuracies stemming from these source models. Additionally, the evaluation metrics employed do not fully align with human perceptions of counterspeech quality, thereby failing to capture the intricate nuances of natural language interactions. Moreover, our framework does not address the possibility of feedback loops or escalation that could arise following the generation of counterspeech, which may influence the long-term effectiveness and impact of our approach. Lastly, while efforts were made to maintain high-quality annotations for counterspeech, it is conceivable that our dataset may not match the caliber of those annotated by more experienced operators from NGOs, such as those found in the Multi-Target CONAN (Fanton et al., 2021b) dataset. Future research could mitigate these limitations by expanding and diversifying the dataset, enhancing the evaluation criteria, and incorporating dialogue modeling into the framework. Our primary focus was on generating effective and non-toxic counterspeech, following prior work (Hengle et al., 2024). While we did not explicitly analyze potential biases in this study, we acknowledge that large language models trained on social media data can amplify biases. Addressing and mitigating such biases is indeed a critical area of research.

Ethics Statement

We recognize the sensitivity required in addressing online hate speech and acknowledge the ethical and moral complexities inherent in conducting research in this area. This initiative serves as an initial attempt to compile a comprehensive and varied collection of counterspeech responses for each instance of hate speech encountered. We understand that algorithms developed for automated counterspeech may generate responses that fail to accurately convey the intended meanings, highlighting the urgent need to better integrate real-world knowledge into these systems. Despite the potential of generative algorithms, there remains a critical necessity for a robust and diverse database of counterspeech to ensure consistently favorable outcomes. Furthermore, while fully operational counterspeech algorithms have yet to be realized, organizations such as United Against Hate play a crucial role in mitigating the prevalence of hate speech in online environments.

8 Acknowledgments

We extend our gratitude to the central HPC facility (Padum) at IIT Delhi for computing. We also sincerely thank Logically and Anusandhan National Research Foundation (DST-NSF Grant: DST/INT/USA/NSF-DST/Tanmoy/P-2/2024) for financial support. Tanmoy acknowledges the support of Rajiv Khemani Young Faculty Chair Professorship in Artificial Intelligence.

References

Chittaranjan Andrade. 2019. The p value and statistical significance: Misunderstandings, explanations, challenges, and alternatives. *Indian Journal of Psychological Medicine*, 41(3):210–215.

Md Rabiul Awal, Rui Cao, Roy Ka-Wei Lee, and Sandra Mitrović. 2021. Angrybert: Joint learning target and emotion for hate speech detection. In *Advances in Knowledge Discovery and Data Mining*, pages 701–713, Cham. Springer International Publishing.

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.

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.

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

Pooja Bhojraj, Pratibha Mudgal, and Sipali Panda. 2024. Submental intubation in a case of panfacial fractures in a paediatric patient: A case report. *Sri Lankan Journal of Anaesthesiology*, 32(2).

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.

Tanmoy Chakraborty and Sarah Masud. 2022. Nipping in the bud: detection, diffusion and mitigation of hate speech on social media. *SIGWEB Newsl.*, 2022(Winter).

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).

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.

Jacob Cohen. 1960. A coefficient of agreement for nominal scales. *Educational and psychological measurement*, 20(1):37–46.

DeepSeek-AI, Aixin Liu, Bei Feng, Bing Xue, Bingxuan Wang, Bochao Wu, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chenyu Zhang, Chong Ruan, Damai Dai, Daya Guo, Dejian Yang, Deli Chen, Dongjie Ji, Erhang Li, Fangyun Lin, Fucong Dai, Fuli Luo, Guangbo Hao, Guanting Chen, Guowei Li, H. Zhang, Han Bao, Hanwei Xu, Haocheng Wang, Haowei Zhang, Honghui Ding, Huajian Xin, Huazuo Gao, Hui Li, Hui Qu, J. L. Cai, Jian Liang, Jianzhong Guo, Jiaqi Ni, Jiashi Li, Jiawei Wang, Jin Chen, Jingchang Chen, Jingyang Yuan, Junjie Qiu, Junlong Li, Junxiao Song, Kai Dong, Kai Hu, Kaige Gao, Kang Guan, Kexin Huang, Kuai Yu, Lean Wang, Lecong Zhang, Lei Xu, Leyi Xia, Liang Zhao, Litong Wang, Liyue Zhang, Meng Li, Miaojun Wang, Mingchuan Zhang, Minghua Zhang, Minghui Tang, Mingming Li, Ning Tian, Panpan Huang, Peiyi Wang, Peng Zhang, Qiancheng Wang, Qihao Zhu, Qinyu Chen, Qiushi Du, R. J. Chen, R. L. Jin, Ruiqi Ge, Ruisong Zhang, Ruizhe Pan, Runji Wang, Runxin Xu, Ruoyu Zhang, Ruyi Chen, S. S. Li, Shanghao Lu, Shangyan Zhou, Shanhuang Chen, Shaoqing Wu, Shengfeng Ye, Shengfeng Ye, Shirong Ma, Shiyu Wang, Shuang Zhou, Shuiping Yu, Shunfeng Zhou, Shuting Pan, T. Wang, Tao Yun, Tian Pei, Tianyu Sun, W. L. Xiao, Wangding Zeng, Wanjia Zhao, Wei An, Wen Liu, Wenfeng Liang, Wenjun Gao, Wenqin Yu, Wentao Zhang, X. Q. Li, Xiangyue Jin, Xianzu Wang, Xiao Bi, Xiaodong Liu, Xiaohan Wang, Xiaojin Shen, Xiaokang Chen, Xiaokang Zhang, Xiaosha Chen, Xiaotao Nie, Xiaowen Sun, Xiaoxiang Wang, Xin Cheng, Xin Liu, Xin Xie, Xingchao Liu, Xingkai Yu, Xinnan Song, Xinxia Shan, Xinyi Zhou, Xinyu Yang, Xinyuan Li, Xuecheng Su, Xuheng Lin, Y. K. Li, Y. Q. Wang, Y. X. Wei, Y. X. Zhu, Yang Zhang, Yanhong Xu, Yanhong Xu, Yanping Huang, Yao Li, Yao Zhao, Yaofeng Sun, Yaohui Li, Yaohui Wang, Yi Yu, Yi Zheng, Yichao Zhang, Yifan Shi, Yiliang Xiong, Ying He, Ying Tang, Yishi Piao, Yisong Wang, Yixuan Tan, Yiyang Ma, Yiyuan Liu, Yongqiang Guo, Yu Wu, Yuan Ou, Yuchen Zhu, Yuduan Wang, Yue Gong, Yuheng Zou, Yujia He, Yukun Zha, Yunfan Xiong, Yunxian Ma, Yuting Yan, Yuxiang Luo, Yuxiang You, Yuxuan Liu, Yuyang Zhou, Z. F. Wu, Z. Z. Ren, Zehui Ren, Zhangli Sha, Zhe Fu, Zhean Xu, Zhen Huang, Zhen Zhang, Zhenda Xie, Zhengyan Zhang, Zhewen Hao, Zhibin Gou, Zhicheng Ma, Zhigang Yan, Zhihong Shao, Zhipeng Xu, Zhiyu Wu, Zhongyu Zhang, Zhuoshu Li, Zihui Gu, Zijia Zhu, Zijun Liu, Zilin Li, Ziwei Xie, Ziyang Song, Ziyi Gao, and Zizheng Pan. 2025. Deepseek-v3 technical report.

Giovanni Di Leo and Francesco Sardanelli. 2020. Statistical significance: p value, 0.05 threshold, and applications to radiomics—reasons for a conservative approach. *European Radiology Experimental*, 4:18.

Matthijs Douze, Alexandr Guzhva, Chengqi Deng, Jeff Johnson, Gergely Szilvasy, Pierre-Emmanuel Mazaré, Maria Lomeli, Lucas Hosseini, and Hervé Jégou. 2025. The faiss library.

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.

Joseph L Fleiss. 1971. Measuring nominal scale agreement among many raters. *Psychological bulletin*, 76(5):378.

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.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, Arun Rao, Aston Zhang, Aurelien Rodriguez, Austen Gregerson, Ava Spataru, Baptiste Roziere, Bethany Biron, Binh Tang, Bobbie Chern, Charlotte Caucheteux, Chaya Nayak, Chloe Bi, Chris Marra, Chris McConnell, Christian Keller, Christophe Touret, Chunyang Wu, Corinne Wong, Cristian Canton Ferrer, Cyrus Nikolaidis, Damien Allonsius, Daniel Song, Danielle Pintz, Danny Livshits, Danny Wyatt, David Esiobu, Dhruv Choudhary, Dhruv Mahajan, Diego Garcia-Olano, Diego Perino, Dieuwke Hupkes, Egor Lakomkin, Ehab AlBadawy, Elina Lobanova, Emily Dinan, Eric Michael Smith, Filip Radenovic, Francisco Guzmán, Frank Zhang, Gabriel Synnaeve, Gabrielle Lee, Georgia Lewis Anderson, Govind Thattai, Graeme Nail, Gregoire Mialon, Guan Pang, Guillem Cucurell, Hailey Nguyen, Hannah Korevaar, Hu Xu, Hugo Touvron, Iliyan Zarov, Imanol Arrieta Ibarra, Isabel Kloumann, Ishan Misra, Ivan Evtimov, Jack Zhang, Jade Copet, Jaewon Lee, Jan Geffert, Jana Vranes, Jason Park, Jay Mahadeokar, Jeet Shah, Jelmer van der Linde, Jennifer Billock, Jenny Hong, Jenya Lee, Jeremy Fu, Jianfeng Chi, Jianyu Huang, Jiawen Liu, Jie Wang, Jiecao Yu, Joanna Bitton, Joe Spisak, Jongsoo Park, Joseph Rocca, Joshua Johnstun, Joshua Saxe, Junteng Jia, Kalyan Vasuden Alwala, Karthik Prasad, Kartikeya Upasani, Kate Plawiak, Ke Li, Kenneth Heafield, Kevin Stone, Khalid El-Arini, Krithika Iyer, Kshitiz Malik, Kuenley Chiu, Kunal Bhalla, Kushal Lakhotia, Lauren Rantala-Yeary, Laurens van der Maaten, Lawrence Chen, Liang Tan, Liz Jenkins, Louis Martin, Lovish Madaan, Lubo Malo, Lukas Blecher, Lukas Landzaat, Luke de Oliveira, Madeline Muzzi, Mahesh Pasupuleti, Mannat Singh, Manohar Paluri, Marcin Kardas, Maria Tsimpoukelli, Mathew Oldham, Mathieu Rita, Maya Pavlova, Melanie Kambadur, Mike Lewis, Min Si, Mitesh Kumar Singh, Mona Hassan, Naman Goyal, Narjes Torabi, Nikolay Bashlykov, Nikolay Bogoychev, Niladri Chatterji, Ning Zhang, Olivier Duchenne, Onur Çelebi, Patrick Alrassy, Pengchuan Zhang, Pengwei Li, Petar Vasic, Peter Weng, Prajjwal Bhargava, Pratik Dubal, Praveen Krishnan, Punit Singh Koura, Puxin Xu, Qing He, Qingxiao Dong, Ragavan Srinivasan, Raj Ganapathy, Ramon Calderer, Ricardo Silveira Cabral, Robert Stojnic, Roberta Raileanu, Rohan Maheswari, Rohit Girdhar, Rohit Patel, Romain Sauvestre, Ronnie Polidoro, Roshan Sumbaly, Ross Taylor, Ruan Silva, Rui Hou, Rui Wang, Saghar Hosseini, Sahana Chennabasappa, Sanjay Singh, Sean Bell, Seohyun Sonia Kim, Sergey Edunov, Shaoliang Nie, Sharan Narang, Sharath Raparthy, Sheng Shen, Shengye Wan, Shruti Bhosale, Shun Zhang, Simon Vandenhende, Soumya Batra, Spencer Whitman, Sten Sootla, Stephane Collot, Suchin Gururangan, Sydney Borodinsky, Tamar Herman, Tara Fowler, Tarek Sheasha, Thomas Georgiou, Thomas Scialom, Tobias Speckbacher, Todor Mihaylov, Tong Xiao, Ujjwal Karn, Vedanuj Goswami, Vibhor Gupta, Vignesh Ramanathan, Viktor Kerkez, Vincent Gonguet, Virginie Do, Vish Vogeti, Vítor Albiero, Vladan Petrovic, Weiwei Chu, Wenhan Xiong, Wenyin Fu, Whitney Meers, Xavier Martinet, Xiaodong Wang, Xiaofang Wang, Xiaoqing Ellen Tan, Xide Xia, Xinfeng Xie, Xuchao Jia, Xuewei Wang, Yaelle Goldschlag, Yashesh Gaur, Yasmine Babaei, Yi Wen, Yiwen Song, Yuchen Zhang, Yue Li, Yuning Mao, Zacharie Delpierre Coudert, Zheng Yan, Zhengxing Chen, Zoe Papakipos, Aaditya Singh, Aayushi Srivastava, Abha Jain, Adam Kelsey, Adam Shajnfeld, Adithya Gangidi, Adolfo Victoria, Ahuva Goldstand, Ajay Menon, Ajay Sharma, Alex Boesenberg, Alexei Baevski, Allie Feinstein, Amanda Kallet, Amit Sangani, Amos Teo, Anam Yunus, Andrei Lupu, Andres Alvarado, Andrew Caples, Andrew Gu, Andrew Ho, Andrew Poulton, Andrew Ryan, Ankit Ramchandani, Annie Dong, Annie Franco, Anuj Goyal, Aparajita Saraf, Arkabandhu Chowdhury, Ashley Gabriel, Ashwin Bharambe, Assaf Eisenman, Azadeh Yazdan, Beau James, Ben Maurer, Benjamin Leonhardi, Bernie Huang, Beth Loyd, Beto De Paola, Bhargavi Paranjape, Bing Liu, Bo Wu, Boyu Ni, Braden Hancock, Bram Wasti, Brandon Spence, Brani Stojkovic, Brian Gamido, Britt Montalvo, Carl Parker, Carly Burton, Catalina Mejia, Ce Liu, Changhan Wang, Changkyu Kim, Chao Zhou, Chester Hu, Ching-Hsiang Chu, Chris Cai, Chris Tindal, Christoph Feichtenhofer, Cynthia Gao, Damon Civin, Dana Beaty, Daniel Kreymer, Daniel Li, David Adkins, David Xu, Davide Testuggine, Delia David, Devi Parikh, Diana Liskovich, Didem Foss, Dingkang Wang, Duc Le, Dustin Holland, Edward Dowling, Eissa Jamil, Elaine Montgomery, Eleonora Presani, Emily Hahn, Emily Wood, Eric-Tuan Le, Erik Brinkman, Esteban Arcaute, Evan Dunbar, Evan Smothers, Fei Sun, Felix Kreuk, Feng Tian, Filippos Kokkinos, Firat

Ozgenel, Francesco Caggioni, Frank Kanayet, Frank Seide, Gabriela Medina Florez, Gabriella Schwarz, Gada Badeer, Georgia Swee, Gil Halpern, Grant Herman, Grigory Sizov, Guangyi, Zhang, Guna Lakshminarayanan, Hakan Inan, Hamid Shojanazeri, Han Zou, Hannah Wang, Hanwen Zha, Haroun Habeeb, Harrison Rudolph, Helen Suk, Henry Aspegren, Hunter Goldman, Hongyuan Zhan, Ibrahim Damlaj, Igor Molybog, Igor Tufanov, Ilias Leontiadis, Irina-Elena Veliche, Itai Gat, Jake Weissman, James Geboski, James Kohli, Janice Lam, Japhet Asher, Jean-Baptiste Gaya, Jeff Marcus, Jeff Tang, Jennifer Chan, Jenny Zhen, Jeremy Reizenstein, Jeremy Teboul, Jessica Zhong, Jian Jin, Jingyi Yang, Joe Cummings, Jon Carvill, Jon Shepard, Jonathan McPhie, Jonathan Torres, Josh Ginsburg, Junjie Wang, Kai Wu, Kam Hou U, Karan Saxena, Kartikay Khandelwal, Katayoun Zand, Kathy Matosich, Kaushik Veeraraghavan, Kelly Michelena, Keqian Li, Kiran Jagadeesh, Kun Huang, Kunal Chawla, Kyle Huang, Lailin Chen, Lakshya Garg, Lavender A, Leandro Silva, Lee Bell, Lei Zhang, Liangpeng Guo, Licheng Yu, Liron Moshkovich, Luca Wehrstedt, Madian Khabsa, Manav Avalani, Manish Bhatt, Martynas Mankus, Matan Hasson, Matthew Lennie, Matthias Reso, Maxim Groshev, Maxim Naumov, Maya Lathi, Meghan Keneally, Miao Liu, Michael L. Seltzer, Michal Valko, Michelle Restrepo, Mihir Patel, Mik Vyatskov, Mikayel Samvelyan, Mike Clark, Mike Macey, Mike Wang, Miquel Jubert Hermoso, Mo Metanat, Mohammad Rastegari, Munish Bansal, Nandhini Santhanam, Natascha Parks, Natasha White, Navyata Bawa, Nayan Singhal, Nick Egebo, Nicolas Usunier, Nikhil Mehta, Nikolay Pavlovich Laptev, Ning Dong, Norman Cheng, Oleg Chernoguz, Olivia Hart, Omkar Salpekar, Ozlem Kalinli, Parkin Kent, Parth Parekh, Paul Saab, Pavan Balaji, Pedro Rittner, Philip Bontrager, Pierre Roux, Piotr Dollar, Polina Zvyagina, Prashant Ratanchandani, Pritish Yuvraj, Qian Liang, Rachad Alao, Rachel Rodriguez, Rafi Ayub, Raghotham Murthy, Raghu Nayani, Rahul Mitra, Rangaprabhu Parthasarathy, Raymond Li, Rebekkah Hogan, Robin Battey, Rocky Wang, Russ Howes, Ruty Rinott, Sachin Mehta, Sachin Siby, Sai Jayesh Bondu, Samyak Datta, Sara Chugh, Sara Hunt, Sargun Dhillon, Sasha Sidorov, Satadru Pan, Saurabh Mahajan, Saurabh Verma, Seiji Yamamoto, Sharadh Ramaswamy, Shaun Lindsay, Shaun Lindsay, Sheng Feng, Shenghao Lin, Shengxin Cindy Zha, Shishir Patil, Shiva Shankar, Shuqiang Zhang, Shuqiang Zhang, Sinong Wang, Sneha Agarwal, Soji Sajuyigbe, Soumith Chintala, Stephanie Max, Stephen Chen, Steve Kehoe, Steve Satterfield, Sudarshan Govindaprasad, Sumit Gupta, Summer Deng, Sungmin Cho, Sunny Virk, Suraj Subramanian, Sy Choudhury, Sydney Goldman, Tal Remez, Tamar Glaser, Tamara Best, Thilo Koehler, Thomas Robinson, Tianhe Li, Tianjun Zhang, Tim Matthews, Timothy Chou, Tzook Shaked, Varun Vontimitta, Victoria Ajayi, Victoria Montanez, Vijai Mohan, Vinay Satish Kumar, Vishal Mangla, Vlad Ionescu, Vlad Poenaru, Vlad Tiberiu Mihailescu, Vladimir Ivanov, Wei Li, Wenchen Wang, Wenwen Jiang, Wes Bouaziz, Will Constable, Xiaocheng Tang, Xiaojian Wu, Xiaolan Wang, Xilun Wu, Xinbo Gao, Yaniv Kleinman, Yanjun Chen, Ye Hu, Ye Jia, Ye Qi, Yenda Li, Yilin Zhang, Ying Zhang, Yossi Adi, Youngjin Nam, Yu, Wang, Yu Zhao, Yuchen Hao, Yundi Qian, Yunlu Li, Yuzi He, Zach Rait, Zachary DeVito, Zef Rosnbrick, Zhaoduo Wen, Zhenyu Yang, Zhiwei Zhao, and Zhiyu Ma. 2024. The llama 3 herd of models.

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.

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

Amey Hengle, Aswini Kumar, Anil Bandhakavi, and Tanmoy Chakraborty. 2025. Cseval: Towards automated, multi-dimensional, and reference-free counterspeech evaluation using auto-calibrated llms.

Amey Hengle, Aswini Kumar, Sahajpreet Singh, Anil Bandhakavi, Md Shad Akhtar, and Tanmoy Chakroborty. 2024. Intent-conditioned and non-toxic counterspeech generation using multi-task instruction tuning with rlaif.

Jiwoo Hong, Noah Lee, and James Thorne. 2024. Reference-free monolithic preference optimization with odds ratio. *arXiv preprint arXiv:2403.07691*.

Albert Q. Jiang, Alexandre Sablayrolles, Antoine Roux, Arthur Mensch, Blanche Savary, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Emma Bou Hanna, Florian Bressand, Gianna Lengyel, Guillaume Bour, Guillaume Lample, Lélio Renard Lavaud, Lucile Saulnier, Marie-Anne Lachaux, Pierre Stock, Sandeep Subramanian, Sophia Yang, Szymon Antoniak, Teven Le Scao, Théophile Gervet, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. 2024. Mixtral of experts.

Jaylen Jones, Lingbo Mo, Eric Fosler-Lussier, and Huan Sun. 2024. A multi-aspect framework for counter narrative evaluation using large language models. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 2: Short Papers)*, pages 147–168, Mexico City, Mexico. Association for Computational Linguistics.

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

Brian Lester, Rami Al-Rfou, and Noah Constant. 2021. The power of scale for parameter-efficient prompt tuning. In *Proceedings of the 2021 Conference on*

Empirical Methods in Natural Language Processing, pages 3045–3059, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Xiang Lisa Li and Percy Liang. 2021. Prefix-tuning: Optimizing continuous prompts for generation. 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 4582–4597, Online. Association for Computational Linguistics.

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.

Pengfei Liu, Weizhe Yuan, Jinlan Fu, Zhengbao Jiang, Hiroaki Hayashi, and Graham Neubig. 2023a. Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing. *ACM Computing Surveys*, 55(9):1–35.

Yajing Liu, Yuning Lu, Hao Liu, Yaozu An, Zhuoran Xu, Zhuokun Yao, Baofeng Zhang, Zhiwei Xiong, and Chenguang Gui. 2023b. Hierarchical prompt learning for multi-task learning. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10888–10898.

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. 2021a. 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.

Sarah Masud, Subhabrata Dutta, Sakshi Makkar, Chhavi Jain, Vikram Goyal, Amitava Das, and Tanmoy Chakraborty. 2021b. 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, Md. Shad Akhtar, and Tanmoy Chakraborty. 2023. Overview of the hasoc subtrack at fire 2023: Identification of tokens contributing to explicit hate in english by span detection.

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.

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. 2022a. Training language models to follow instructions with human feedback.

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. 2022b. 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.

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.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Stefano Ermon, Christopher D. Manning, and Chelsea Finn. 2023. Direct preference optimization: Your language model is secretly a reward model.

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.

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.

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.

Sahajpreet Singh, Sarah Masud, and Tanmoy Chakraborty. 2024. Independent fact-checking organizations exhibit a departure from political neutrality.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.

Jiaan Wang, Yunlong Liang, Fandong Meng, Zengkui Sun, Haoxiang Shi, Zhixu Li, Jinan Xu, Jianfeng Qu, and Jie Zhou. 2023. Is ChatGPT a good NLG evaluator? a preliminary study. In *Proceedings of the 4th New Frontiers in Summarization Workshop*, pages 1–11, Singapore. Association for Computational Linguistics.

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.

Neemesh Yadav, Sarah Masud, Vikram Goyal, Vikram Goyal, Md Shad Akhtar, and Tanmoy Chakraborty. 2024. Tox-bart: Leveraging toxicity attributes for explanation generation of implicit hate speech.

Jingfeng Yang, Hongye Jin, Ruixiang Tang, Xiaotian Han, Qizhang Feng, Haoming Jiang, Shaochen Zhong, Bing Yin, and Xia Hu. 2024. Harnessing the power of llms in practice: A survey on chatgpt and beyond. *ACM Transactions on Knowledge Discovery from Data*, 18(6):1–32.

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*.

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.

9 Appendix

9.1 Annotation Process

Our annotation process was conducted by a team of five expert annotators with backgrounds in social science and computational linguistics, specializing in online hate speech and counter-narrative generation. Each annotator had a strong foundation in hate speech analysis, having published research papers or completed advanced studies in the field. To ensure consistency and high-quality annotations, we implemented an extensive training program. This included reviewing existing counter-narrative frameworks, discussing annotation guidelines, and participating in exercises to align understanding of emotion categories. This preparatory phase was crucial for achieving reliable inter-annotator agreement and maintaining the integrity of the annotations.

The annotation process itself followed a rigorous three-phase protocol. Initially, all five annotators independently labeled a common set of 250 instances to establish a baseline agreement. For evaluating inter-annotator agreement, we utilized both Cohen's Kappa (Cohen, 1960) and Fleiss' Kappa (Fleiss, 1971). Cohen's Kappa measures pairwise agreement among annotators, finding most values exceeded 0.70, with several above 0.80, indicating high consistency (See Table 6). Instances with significant disagreement were addressed through group discussions to align understanding and resolve discrepancies. In the second phase, larger batches were annotated with periodic cross-validation, where 20% of instances were randomly assigned to multiple annotators to ensure consistency. This phase involved annotating batches of $2,500, 2,800$, and $3,000$ instances, respectively. Table 4 shows the batch-wise intra-annotation agreement results. The final phase allowed independent annotation after achieving strong agreement, as measured by Cohen's Kappa exceeding 0.8. Throughout the process, annotators utilized a custom interface that systematically displayed the existing counterspeech from `IntentCONANv2`, and emotion category options. To further check the annotation quality, we conducted a post-analysis on the intra-annotator agreement for a randomly selected subset of $5,000$ counterspeech instances (see Table 5). This analysis was in response to suggestions for additional quality checks. Our study utilized high-quality counterspeech instances from `IntentCONANv2`, with annotators independently assigning emotion labels based on established guidelines without access to the strategy categories. The resulting uniform distribution of strategy-emotion pairs across target groups was a posterior outcome of this independent process, uninfluenced by pre-existing constraints or biases. These metrics collectively show the robustness and reliability of the `MultiCONAN` dataset.

9.2 Procedure and Annotation Criteria

Before beginning the annotation process, all annotators thoroughly reviewed the field guide on "addressing online harassment" [4]. This preparatory phase involved extensive discussions with the annotators to deepen their understanding of counterspeech. These dialogues ensured that the annotators were well-equipped with the necessary knowledge and context, enabling them to effectively contribute to the project.

Anger: Anger is characterized by intense feelings of displeasure, hostility, or antagonism toward someone or something perceived as a source of harm or wrongdoing[5]. It often manifests in expressions of frustration, outrage, and resentment. Annotators should look for language that conveys aggression, threats, or overt negativity. Examples might include harsh criticism, shouting, or aggressive demands. This emotion is frequently triggered by situations of perceived injustice, insult, or betrayal, and it is crucial for annotators to distinguish it from other negative emotions like disgust or sadness[6].

Disgust: Disgust is an emotion that arises from a strong sense of aversion or repulsion toward something offensive, distasteful, or morally objectionable. This feeling can be directed toward people, behaviors, or ideas that violate social norms or personal values. Annotators should identify language that reflects contempt, disdain, or severe disapproval[7]. Common indicators include expressions of revulsion, condemnation, or derogatory remarks. Disgust often accompanies discussions of taboo subjects or unethical actions, requiring careful attention to the context in which these sentiments are expressed.

Surprise: Surprise is an emotional response to unexpected events or information that deviates from what is anticipated. It can be positive, negative, or neutral, depending on the nature of the unexpected occurrence. Annotators should recognize cues such as exclamations, sudden changes in tone, or language indicating shock or astonishment. This emotion often appears in contexts where new, unforeseen developments are revealed. The intensity of surprise can vary, but key indicators include words and phrases that express sudden realization or discovery[8]. The following annotation guidelines were adhered to consistently throughout the annotation process by our annotators.

Joy: Joy is a positive emotional state characterized by feelings of happiness, pleasure, and contentment. It often results from fulfilling experiences, successful outcomes, or pleasant interactions. Annotators should identify expressions of delight, gratitude, and satisfaction[9]. Language reflecting enthusiasm, laughter, or praise is commonly associated with this emotion. Joy can be expressed in both overt and subtle ways, and it is essential to consider the broader context to accurately capture this emotion in annotations. Annotators should be attentive to both explicit declarations of happiness and more nuanced indications of joy.

Sadness: Sadness is an emotional response to loss, disappointment, or other unfavorable circumstances. It is characterized by feelings of sorrow, melancholy, and despondency. Annotators should look for language that conveys grief, regret, or a sense of hopelessness[10]. Indicators of sadness include expressions of mourning, lamentation, and heartache. This emotion can be associated with personal experiences, broader societal issues, or empathetic responses to others' suffering. Accurate annotation requires sensitivity to the depth and context of the expressed sadness, ensuring that subtle expressions of this emotion are also captured.

Let us consider some hate speech (HS) and counterspeech (CS) instances to illustrate our annotation process,

(HS): The West is heading in the direction of Fascism because Jews are planning to install a world dictator.

(CS): Blaming an entire group for political

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

[5] https://www.paulekman.com/universal-emotions/what-is-anger/

[6] https://www.apa.org/topics/anger

[7] https://www.paulekman.com/universal-emotions/what-is-disgust/

[8] https://www.paulekman.com/universal-emotions/what-is-surprise/

[9] https://www.paulekman.com/universal-emotions/what-is-enjoyment/

[10] https://www.paulekman.com/universal-emotions/what-is-sadness/

Batch	Inter-annotator Agreement				
	Annotator 1	**Annotator 2**	**Annotator 3**	**Annotator 4**	**Annotator 5**
Batch 1 (2500)					
Annotator 1	1.0	0.808783	0.811652	0.782824	0.762829
Annotator 2	0.808783	1.0	0.661458	0.63632	0.612693
Annotator 3	0.811652	0.661458	1.0	0.640035	0.619542
Annotator 4	0.782824	0.63632	0.640035	1.0	0.605464
Annotator 5	0.762829	0.612693	0.619542	0.605464	1.0
Batch 2 (2800)					
Annotator 1	1.0	0.852106	0.813067	0.817208	0.800127
Annotator 2	0.852106	1.0	0.693793	0.693689	0.687102
Annotator 3	0.813067	0.693793	1.0	0.665585	0.654362
Annotator 4	0.817208	0.693689	0.665585	1.0	0.659997
Annotator 5	0.800127	0.687102	0.654362	0.659997	1.0
Batch 3 (3000)					
Annotator 1	1.0	0.942593	0.894511	0.888579	0.85598
Annotator 2	0.942593	1.0	0.840615	0.836788	0.80862
Annotator 3	0.894511	0.840615	1.0	0.796064	0.765929
Annotator 4	0.888579	0.836788	0.796064	1.0	0.758911
Annotator 5	0.85598	0.80862	0.765929	0.758911	1.0

Table 4: Inter-annotator Agreement Coefficients for Each Batch

Annotator	Annotator 1	Annotator 2	Annotator 3	Annotator 4	Annotator 5
Annotator 1	1.0	0.930857	0.889566	0.855752	0.883054
Annotator 2	0.930857	1.0	0.828692	0.798059	0.823054
Annotator 3	0.889566	0.828692	1.0	0.763612	0.786144
Annotator 4	0.855752	0.798059	0.763612	1.0	0.755936
Annotator 5	0.883054	0.823054	0.786144	0.755936	1.0

Table 5: Inter-annotator Agreement of 5000 Counterspeech Instances for Post Quality Assessment

	A1	A2	A3	A4	A5
A1	0.000	0.875	0.883	0.863	0.850
A2	0.875	0.000	0.788	0.734	0.748
A3	0.883	0.788	0.000	0.737	0.785
A4	0.808	0.734	0.737	0.000	0.691
A5	0.844	0.748	0.785	0.691	0.000

Table 6: Cohen's Kappa Matrix for Inter-Annotator Agreement among five annotators: A1, A2, A3, A4, and A5.

movements is baseless and harmful. **(Disgust)**

In the above HS and CS example, the counterspeech, it is evident that the response expresses disgust through its explicit rejection of the notion that an entire group can be held responsible for political movements. Phrases such as 'baseless' and 'harmful' convey strong moral disapproval, which are critical indicators of disgust. The language reflects a profound aversion to the offensive nature of the hate speech, especially in its attempt to generalize blame. By emphasizing the unethical and unfounded nature of these accusations, the counterspeech communicates a sense of repulsion, which supports its classification under the emotion of **disgust.** Similarly let us consider the following,

(HS): Muslims are a problem for all other religions.

(CS): I understand that you may have concerns about different religions, but it's important to remember that diversity is what makes our world so beautiful. instead of seeing Muslims as a problem, let's try to celebrate our differences and find common ground. we all share the same desire for love, peace, and understanding. By embracing empathy and affiliation, we can build bridges of dialogue that promote respect and unity among different faiths. let's work together to create a world where everyone feels valued and accepted. **(Joy)**

In the analysis of the counterspeech, the response clearly conveys joy by emphasizing the beauty of diversity and the positive potential for unity among different faiths. Language such as "diversity is what makes our world so beautiful" and "we all share the same desire for love, peace, and understanding" reflect a sense of happiness and appreciation for shared human values. The counterspeech fosters a joyful perspective by encouraging the celebration of differences and the building of connections through empathy. This focus on mutual respect and the vision of a harmonious, inclusive world aligns the response with the emotion of **joy.**

Following these criteria, annotators meticulously annotated a total of $13,973$ unique counterspeech instances.

9.3 Advantage of IntentCONANv2

`IntentCONANv2` represents a significant advancement over its predecessors, including `IntentCONAN`, by enhancing the quality and structure of counterspeech instances. This dataset builds upon the annotation guidelines established by Gupta et al. (2023) but focuses on improving content quality by increasing token lengths and ensuring a uniform distribution across four strategies: positive, informative, questioning, and denouncing Hengle et al. (2024). The effectiveness of counterspeech is often linked to its level of detail and comprehensiveness, which can be reflected in its length; a higher token count typically indicates a more thorough and nuanced response, better equipped to address and counteract hate speech. We selected `IntentCONANv2` for our annotation purposes due to several key advantages. Firstly, it is a large-scale dataset comprising $13,952$ counterspeech instances, offering a substantial foundation for analysis. Secondly, it addresses limitations of earlier datasets, such as `CONAN` and `MultiTargetCONAN`, by providing more detailed and informative counterspeech that effectively counters the central aspects of hate speech Hengle et al. (2024). The removal of the humorous strategy is also noteworthy, as it mitigates the risk of subjective or offensive content. Furthermore, `IntentCONANv2` ensures a consistent representation of counterspeech, with an average of four instances per hate speech example, compared to the two instances in `IntentCONAN`. Additionally, the dataset emphasizes substantial content, with an average token length of 40.61, reflecting a focus on creating comprehensive responses that are more effective in countering hate speech.

9.4 Statistical Analysis on Dataset

In the `multiCONAN` dataset, the distribution of strategy categories across counterspeech is uniform (see Figure 3a), with a particular focus on the Emotion category. As illustrated in Figure 3b, the distribution of emotion categories within counterspeech instances reveals that the majority fall under the 'Joy' category. This predominance signifies the quality and positivity of counterspeech. Joy-based counterspeech typically involves presenting constructive examples and success stories, which necessitate more elaborate responses to effectively build emotional connections through positive narratives. Constructive arguments, in turn, require detailed explanations of alternative viewpoints, further contributing to the need for comprehensive and nuanced responses. Following 'Joy,' the categories of 'Anger' and 'Disgust' are also notable, though to a lesser extent. The 'Sad' emotion category has a significantly lower count, indicating its rare occurrence in counterspeech. Figure 3c further demonstrates the distribution of both strategy and emotion categories across the training, validation, and testing sets. It is evident that the data splitting is uniform among all categories, ensuring balanced representation and reliable performance assessment across different dataset partitions.

Figure 3d presents the mean token length of counterspeech across various emotion categories. The data shows that the 'Joy' category has a mean token length of approximately 60, indicating that counterspeech with joyful emotions tends to be more elaborate. In contrast, other emotion categories, such as 'Anger,' 'Disgust,' and 'Sad,' have a more uniform distribution of mean token lengths, ranging from 25 to 35. This variation suggests that counterspeech with more tokens is potentially more effective in targeting and neutralizing hateful com-

Metric	T-statistic	p-value	Significant	Outperform
Comparison with GPT-3.5 FS				
SC	63.34	0.0	Yes	Yes
EC	9.31	$1.71E-20$	Yes	Yes
TC	0.96	0.33	No	Yes
BERT Score	17.20	$9.97E-65$	Yes	Yes
METEOR	-2.37	0.018	Yes	No
ROUGE-1	8.56	$1.47E-17$	Yes	Yes
ROUGE-2	9.97	$3.23E-23$	Yes	Yes
ROUGE-L	14.21	$4.57E-45$	Yes	Yes
CoSim	0.66	0.51	No	Yes
Toxicity	21.30	$4.85E-97$	Yes	Yes
Comparison with GPT-3.5 ZS				
SC	38.09	$2.92E-284$	Yes	Yes
EC	-22.54	$4.86E-108$	Yes	No
TC	-13.22	$2.50E-39$	Yes	No
BERT Score	19.46	$7.40E-82$	Yes	Yes
METEOR	-2.43	0.015	Yes	No
ROUGE-1	10.38	$4.95E-25$	Yes	Yes
ROUGE-2	13.77	$1.64E-42$	Yes	Yes
ROUGE-L	15.91	$7.66E-56$	Yes	Yes
CoSim	-3.72	0.0002	Yes	No
Toxicity	-8.31	$1.18E-16$	Yes	No

Table 7: Statistical comparison of our model with GPT-3.5 FS and GPT-3.5 ZS across various metrics using T-test and p-values. Positive T-statistic value indicates that our model performs and $p < 0.05$ shows the significance.

ments. Additionally, Figure 3e illustrates the mean token length across different target groups, showing a consistent and uniform distribution. This consistency ensures that counterspeech instances are equally detailed and explanatory across all target groups, contributing to the robustness and reliability of the `multiCONAN` dataset.

9.5 Statistical Significance Testing

The statistical evaluation of our model against GPT-3.5 Few-Shot (FS) across various metrics highlights significant differences in performance (see Table 7). For the strategy Conformity (SC) score, our model demonstrates a substantial advantage, with a T-statistic of 63.34 and a p-value of 0.0, indicating highly significant results. Similarly, for Emotion Conformity (EC), the T-statistic of 9.31 and a p-value of $1.71E-20$ confirm that our model significantly outperforms GPT-3.5 FS. Metrics such as BERT Score, ROUGE-1, ROUGE-2, and ROUGE-L also showcase strong performance by our model, with all p-values far below the significance threshold ($p < 0.05$). However, for METEOR, the negative T-statistic (-2.37) and a p-value of 0.018 suggest that GPT-3.5 FS slightly outperforms our model in this metric. Interestingly, for Target Conformity (TC) and CoSim, no statistically significant was observed with p-values of 0.33 and 0.51, respectively. Despite these exceptions, the overall results indicate that our model achieves superior performance across most metrics compared to GPT-3.5 FS.

When compared to GPT-3.5 Zero-Shot (ZS), our model exhibits significant improvements in most metrics, as evidenced by the extremely low p-values across strategy Classification (SC), BERT Score, ROUGE metrics (ROUGE-1, ROUGE-2, ROUGE-L), and Toxicity reduction. For instance, SC achieves a T-statistic of 38.09 with a p-value of $2.92E-284$, underscoring the robustness of our model in this task. However, for EC and TC scores, the negative T-statistics (-22.54 and -13.22) indicate that GPT-3.5 ZS performs better in these areas despite their statistical significance. Additionally, METEOR and CoSim also show slight advantages for GPT-3.5 ZS, with respective T-statistics of -2.43 and -3.72. Overall, while there are isolated cases where GPT-3.5 ZS performs better or comparably, our model consistently outperforms it across critical metrics such as BERT Score, ROUGE metrics, SC, and Toxicity reduction, demonstrating its effectiveness in counterspeech generation tasks.

Furthermore, we conduct significance testing to compare our model with various ablation configurations (see Table 8). The results indicate no statistical significance in performance, which suggests that our model exhibits robustness to architectural variations and maintains stable performance across diverse configurations (Di Leo and Sardanelli, 2020; Andrade, 2019; Bhojraj et al., 2024).

9.6 ZeroShot Prompt

We used the following prompt for GPT ZeroShot:

> Generate a counterspeech response to combat hate speech with the following specifications:
>
> - **Intent**: [INTENT]
> - **Emotion**: [EMOTION]
>
> **Guidelines**:
>
> - Address the hate speech respectfully but firmly
> - Maintain the specified emotional tone throughout
> - Focus on constructive dialogue
> - Avoid toxic or inflammatory language

Metric	HIPO_VT_3	HIPO_VT_5	HIPO_VT_7	HIPO_VT_10	HIPO_VT_3_DPO
SC	0.726	0.463	0.099	0.080	1.000
EC	0.716	0.078	0.062	0.094	0.775
TC	1.000	0.009	0.315	0.000	0.966
BERT_score	0.878	0.358	0.806	0.158	0.923
METEOR	0.937	0.885	0.257	0.416	0.873
ROUGE-1	0.816	0.647	0.933	0.535	0.941
ROUGE-2	0.772	0.865	0.435	0.842	0.819
ROUGE-L	0.777	0.643	0.877	0.595	0.868
CoSim	1.000	0.330	0.285	0.565	0.951
Toxicity	0.929	0.137	0.418	0.098	0.995

Table 8: Statistical significance across different ablations

- Align with the specified intent

Hate Speech: “[HATE_SPEECH]"

Counterspeech:

9.7 FewShot Prompt

The prompt we used for GPT FewShot is as follows:

Task: Generate counterspeech responses to hate speech with specific intent and emotion.

Examples:

1. **Hate Speech**: “Muslims are terrorists who destroy peace."
   Intent: Informative
   Emotion: Joy
   Counterspeech: “I'm excited to share that Muslim doctors, teachers, and community leaders contribute positively to our society every day! It's wonderful to see how diverse faiths enrich our communities and promote peace through understanding."
2. **Hate Speech**: “Women don't belong in tech."
   Intent: Questioning
   Emotion: Surprise
   Counterspeech: “Isn't it fascinating how Ada Lovelace, a woman, wrote the first computer program? And have you seen how mixed-gender tech teams consistently outperform others? These surprising facts challenge our assumptions!"
3. **Hate Speech**: “Immigrants are stealing our jobs."
   Intent: Positive
   Emotion: Sad
   Counterspeech: “It saddens me to see how economic anxieties can create divisions in our community. The reality is that immigrants actually create millions of new jobs through their entrepreneurship and hard work. Studies show that immigrant-founded businesses employ over 3.7 million Americans. While I understand the fear of job security, I believe we can work together to build an economy that provides opportunities for everyone. Let's focus on supporting each other and creating more opportunities rather than letting fear divide us."

Now generate a counterspeech for:

Hate Speech: “[HATE_SPEECH]"

Intent: [INTENT]

Emotion: [EMOTION]

Counterspeech:

9.8 Retrival-Based Prompt

Generate [INTENT] and [EMOTION] aligned counterspeech addressing these aspects from the hate speech: Hate speech: [HATESPEECH] Relevant examples: [Examples]

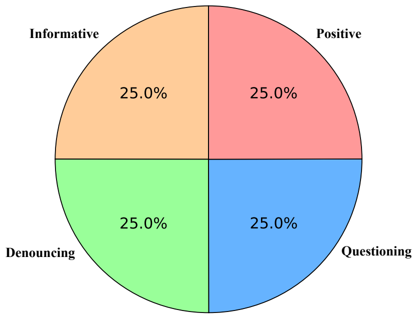
9.9 Computing Information

Our research is conducted on the NVIDIA RTX A100 with 80GB RAM GPU.

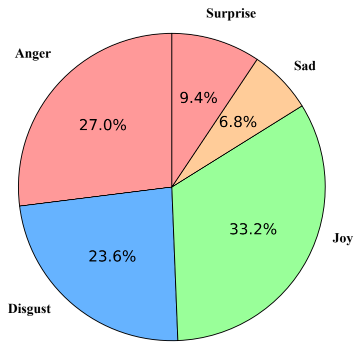
9.10 Hyper-parameter Information

Here we have mentioned the hyper-parameters we used for all our experiments,

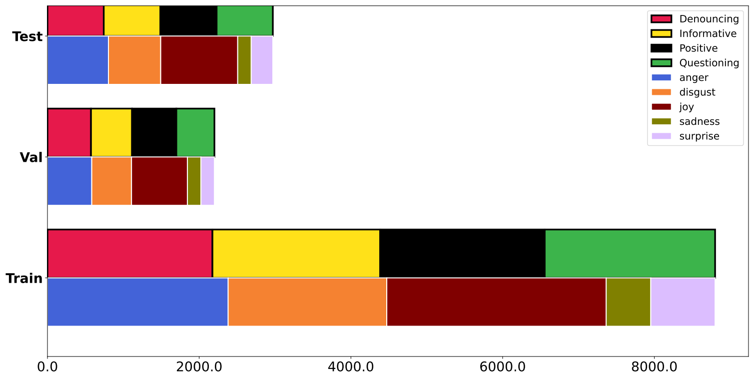
- Batch size: 4
- Learning rate: 1e-4
- Maximum input token: 512
- Maximum output token: 512
- Temperature sampling: Not used
- Max Epoch: 50
- Early stopping used: Yes



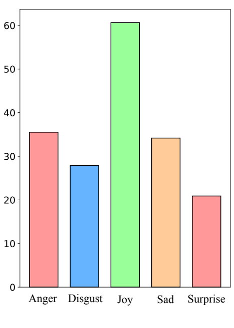
(a) CS, Strategy distribution



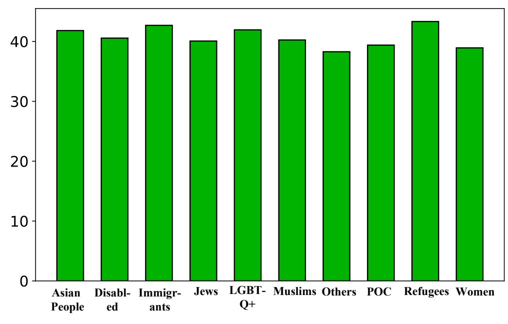
(b) CS, Emotion distribution



(c) CS strategy and CS emotion distribution across train, val, and test.



(d) Emotions, mean token length.



(e) Targets, mean token length.

Figure 3: Visual exploration of various attribute distribution present in the `MultiCONAN` dataset.