"Four Heads Are Better Than One": QUADRA for Intent-Aware Counterspeech Generation

Mehul Agarwal
IIIT Delhi
New Delhi, India
mehul222294@iiitd.ac.in

Noel Abraham Tiju
IIIT Delhi
New Delhi, India
noel22338@iiitd.ac.in

Rahul Ramesh Omalur
IIIT Delhi
New Delhi, India
rahul22392@iiitd.ac.in

Abstract

We explore intent-specific counterspeech generation to tackle hate speech online. Using the IntentCONAN v2 dataset—with 9,532 training examples balanced across four rhetorical intents (Informative, Denouncing, Questioning, and Positive)—we propose **QUADRA**, a modular framework with a shared HateBERT encoder and intent-specific BART decoders. QUADRA investigates three fusion mechanisms (Linear, Shared, and Cross Attention) to effectively combine hate speech embeddings with intent representations. For evaluation, we introduce **DialoRank**, a zero-shot DialoGPT-based ranking method that assesses responses by intent relevance. Results show that QUADRA outperforms DialoGPT and GPS baselines across lexical and semantic metrics, with **SharedFusion** achieving the best performance (ROUGE-1: 0.251, METEOR: 0.158, BERTScore F1: 0.871). Our findings highlight the effectiveness of intent conditioning and multi-decoder architectures in generating contextually appropriate and rhetorically impactful counterspeech.

1 Introduction

The widespread adoption of online platforms has amplified the reach and impact of hate speech. Conventional moderation strategies such as content removal or user bans often fall short—raising concerns over censorship and failing to address the deeper issues of harmful discourse. In contrast, *counterspeech* has emerged as a promising alternative: civil, persuasive responses aimed at challenging hate, encouraging constructive dialogue, and influencing both perpetrators and bystanders.

While recent advancements in language generation have enabled automated counterspeech generation, most existing systems focus on producing a single generic response per hate speech instance. This overlooks the importance of rhetorical intent, which plays a critical role in shaping the reception and effectiveness of counterspeech. Prior work (1) has shown that communities respond differently to various styles of counterspeech—such as informative rebuttals, empathetic messages, or critical questioning—highlighting the need for more context-sensitive generation strategies.

In this work, we present **QUADRA** (QUAD-based Response Architecture), a modular framework for intent-aware counterspeech generation. QUADRA is designed to produce multiple stylistically distinct responses aligned with communicative intents and select the most contextually appropriate one. By leveraging specialized encoders and decoders alongside a reranking mechanism, our approach encourages diversity, relevance, and rhetorical alignment in generated outputs.

We evaluate QUADRA on the IntentCONAN v2 dataset and benchmark it against leading systems such as DialoGPT and the Generate-Prune-Select (GPS) framework. Our experiments show that QUADRA achieves significant improvements in both semantic and lexical evaluation metrics, as well as intent classification accuracy, demonstrating the value of intent-driven generation in counterspeech modeling.

2 Related Work

Counterspeech generation has emerged as a critical area in NLP, with various models aiming to generate diverse, context-aware, and non-toxic responses. Two major architectures that inform our work are the **Generate-Prune-Select (GPS)** framework and Microsoft's **DialoGPT**.

2.1 Generate-Prune-Select (GPS)

The GPS framework proposed by (4). adopts a modular pipeline to ensure response diversity and quality. It consists of:

- **Generate:** A VAE-based module generates diverse candidate responses conditioned on hate speech inputs.

- **Prune:** A lightweight classifier filters ungrammatical or incoherent outputs.
- **Select:** Retrieval-based techniques such as TF-IDF are used to select the most contextually relevant response.

While GPS ensures modularity and interpretability, our model aims to unify generation and filtering via end-to-end trainable fusion modules, thereby improving fluency and contextual sensitivity in a single pass.

2.2 DialoGPT

DialoGPT (3) is a large-scale generative model fine-tuned on Reddit conversations. As an autoregressive transformer-based language model derived from GPT-2, DialoGPT is optimized for generating human-like dialogue responses. Its multi-turn training structure captures long-range conversational dependencies effectively.

DialoGPT serves as our foundational baseline. However, it often suffers from bland or inconsistent outputs in emotionally charged or domain-specific settings such as counterspeech. Inspired by its robust generation capability, we incorporate fusion architectures to enrich contextual and semantic alignment, improving over DialoGPT in both lexical diversity and relevance.

3 Dataset Description

We use the IntentCONAN v2 dataset, a curated subset of the original IntentCONAN benchmark designed for training intent-aware counterspeech generation models. Each counterspeech is explicitly labeled with a communicative intent that reflects the rhetorical strategy used to counter the hate speech.

In IntentCONAN v2, only four intents are included:

Category	Training	Validation	Testing
# Entries	9532	2971	1470
# Unique HS	2383	1097	900
Informative	2383	747	369
Denouncing	2383	742	366
Positive	2383	741	370
Questioning	2383	741	365

Table 1: Distribution of dataset statistics across training, validation, and testing sets.

- **Informative:** Offers factual corrections to dispel misinformation.
- **Denouncing:** Explicitly condemns the hate speech or its ideology.
- **Question:** Asks critical questions to expose flawed reasoning or assumptions.
- **Positive:** Responds with empathy, encouragement, or kindness.

An additional intent, **Humour**, is present in the original *IntentCONAN* dataset and involves using wit or irony to de-escalate and disarm hate speech. However, this intent is not included in the *IntentCONAN v2* subset used in our study.

Metric	DialoGPT	GPS
Category Accuracy	0.681	0.754
ROUGE-1	0.130	0.176
ROUGE-2	0.003	0.030
ROUGE-L	0.105	0.132
METEOR	0.040	0.116
BS (Precision)	0.791	0.240
BS (Recall)	0.808	0.121
BS (F1)	0.799	0.180

Table 2: Evaluation metrics comparing DialoGPT and updated GPS results.

4 Baselines

For the baselines, we fine-tuned **DialoGPT-small** (3), the **GPS model** (4), and included the reported results from the **QUARC** model (1) for comparison. The models were evaluated using a combination of lexical and semantic metrics. Table 2 presents the evaluation results for all three models, where DialoGPT serves as a generation-based baseline, GPS represents a retrieval-generation pipeline, and QUARC is the current state-of-the-art two-stage intent-conditioned generator.

5 Methodology

Our approach, QUADRA, is a modular architecture for intent-aware counterspeech generation with four specialized heads: informative, denouncing, humorous, and positive. Input hate speech is first processed by **HateBERT: Hate Speech Embedding Layer** to extract hate-specific semantic embeddings. Simultaneously, four intent-specialized BART encoders (2) (**QUAD-BART: Intent-Specific Hate Representation**), each trained exclusively on hate speech matching their target intent, create intent-specific representations. A **Fusion Module: Enriching Hate with Intent** then combines the HateBERT embedding with each intent representation to create enriched contextual vectors, which are passed to intent-specific BART decoders (**Multi-Decoder Response Generator**) to generate candidate responses. Finally, **DialoRank: Intent-Aware Reranking Module** selects the most appropriate counterspeech based on relevance, tone, and alignment.

5.1 Input Representation

Our framework distinguishes between hate speech and counterspeech inputs to better align with the respective encoder and decoder architectures. Hate speech is tokenized using the **HateBERT** tokenizer, ensuring compatibility with the HateBERT encoder, which captures nuanced linguistic cues in hateful content. Counterspeech is tokenized using the **BART** tokenizer to align with the decoder's vocabulary and structure for effective generation.

Each data sample in the `IntentCONAN v2` dataset includes:

- **Hate speech tokens** with attention masks for the encoder.
- **Counterspeech tokens** as targets for generation.
- **Intent label** indicating the rhetorical strategy.
- **Full intent set** for multi-intent evaluation scenarios.
- **Raw hate speech text** used during inference and prompting.

Only the tokenized hate speech, counterspeech, and intent label are used for model training. The remaining fields support evaluation and qualitative analysis.

5.2 Feature Encoder using HateBERT

To encode hate speech for intent-aware generation, we use **HateBERT** (6), a RoBERTa-based model pre-trained on toxic Reddit communities. HateBERT has demonstrated superior performance over general-purpose models such as BERT and RoBERTa on hate speech detection benchmarks, including OLID (F1: 0.86 compared to 0.83/0.81) and HateXplain, due to its domain-specific pretraining. This makes it particularly adept at capturing the subtle, implicit, and veiled forms of hate speech that are common in real-world social media scenarios.

In our framework, HateBERT encodes the hate speech into rich contextual embeddings, which are then projected into four distinct representations, each aligned with a target communicative intent—*Informative*, *Questioning*, *Denouncing*, and *Positive*. These intent-conditioned vectors guide their respective decoders, facilitating the generation of contextually relevant and rhetorically appropriate counterspeech.

5.3 QUAD Bart Decoder Network

The proposed `CounterSpeechNetwork` generates intent-specific counterspeech by combining hate speech and intent embeddings. It consists of:

- A shared `FeatureEncoder` to extract hate speech and intent representations.
- Four specialized `BART` decoders, each corresponding to one of the intents: *informative*, *questioning*, *denouncing*, and *positive*.
- A fusion mechanism (`Linear`, `Shared`, or `Cross Attention`) to combine hate speech and intent embeddings before decoding.

During training, the network uses teacher forcing with cross-entropy loss to train each decoder independently. At inference, all decoders generate candidate responses, which are then scored by a pretrained `DialoGPT`-LLM to select the most appropriate output.

5.3.1 Fusion Mechanisms

- **Linear Fusion:** Simply concatenates the hate speech and intent embeddings and projects them through a linear layer to match the `BART` input dimensions. Each decoder has its own fusion layer.
- **Cross Attention Fusion:** Utilizes a shared cross-attention module where hate speech embeddings act as queries (Q) and intent embeddings as keys (K) and values (V). Multi-head attention enables fine-grained alignment between modalities. The attention mechanism is defined as:

  $$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^\top}{\sqrt{d_k}}\right) V$$

  where

  $$Q = W_q h, \quad K = W_k e, \quad V = W_v e$$

  are linear projections of the hate speech embedding h and intent embedding e, and d_k is the head dimension. The attended output is passed through residual connections and a feedforward network:

  $$z = \text{LayerNorm}\left(\text{Attention}(Q, K, V) + e\right)$$
  $$\text{Output} = \text{LayerNorm}\left(\text{FFN}(z) + z\right)$$

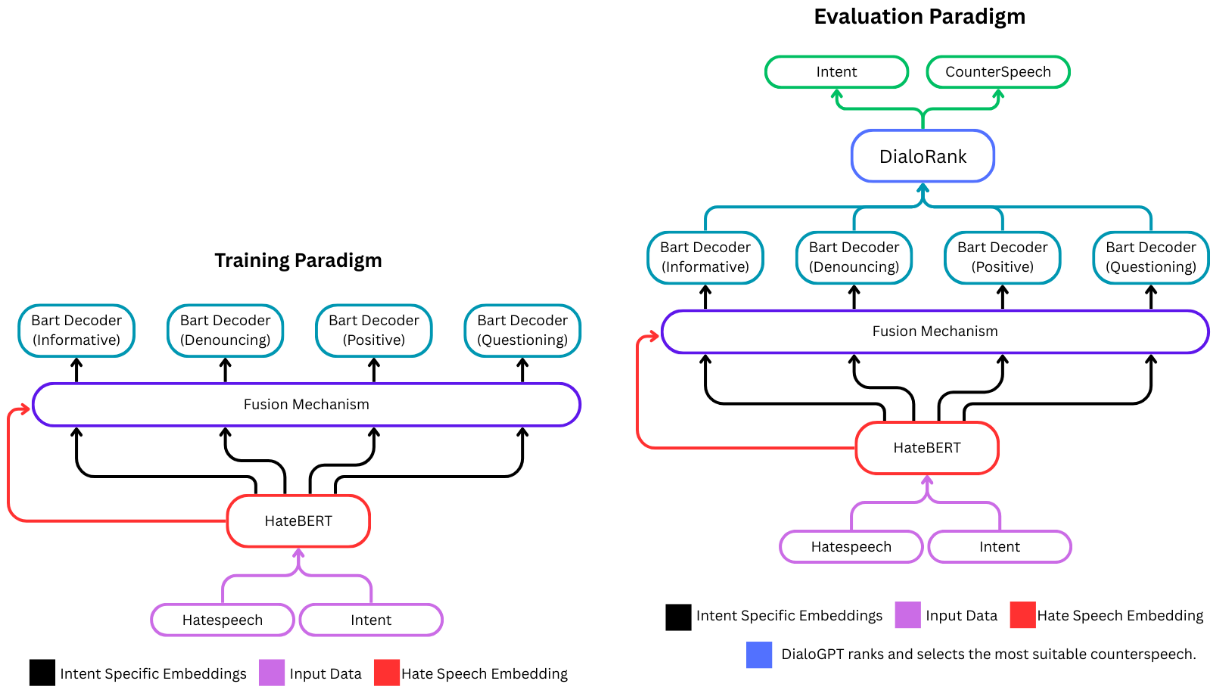

Training Paradigm
Bart Decoder (Informative)
Bart Decoder (Denouncing)
Bart Decoder (Positive)
Bart Decoder (Questioning)
Fusion Mechanism
HateBERT
Hatespeech
Intent
■ Intent Specific Embeddings
■ Input Data
■ Hate Speech Embedding
Evaluation Paradigm
Intent
CounterSpeech
DialoRank
Bart Decoder (Informative)
Bart Decoder (Denouncing)
Bart Decoder (Positive)
Bart Decoder (Questioning)
Fusion Mechanism
HateBERT
Hatespeech
Intent
■ Intent Specific Embeddings
■ Input Data
■ Hate Speech Embedding
■ DialoGPT ranks and selects the most suitable counterspeech.

Figure 1: Overview of the proposed framework. Left: Training with HateBERT and intent-specific BART decoders. Right: Evaluation using DialoRank to select the most suitable counterspeech.

This fusion mechanism, inspired by Transformer architectures (7), is shared across all decoders to enforce consistency and reduce parameter overhead.

- **Shared Fusion:** Inspired by (8), this module computes dual cross-attentive embeddings which are then merged via gated interpolation. Two gating functions, G_{HS} and G_{IE}, are learned using sigmoidal activations:

$$G_{\text{HS}} = \sigma(W_{\text{HS}} E_{\text{HS}})$$
$$G_{\text{IE}} = \sigma(W_{\text{IE}} E_{\text{IE}})$$

These gates control the contribution of each modality's attended features. The fusion outputs are computed as:

$$F_1 = G_{\text{HS}} \cdot A_{\text{HS}\to\text{IE}} + (1 - G_{\text{HS}}) \cdot A_{\text{IE}\to\text{HS}}$$
$$F_2 = G_{\text{IE}} \cdot A_{\text{HS}\to\text{IE}} + (1 - G_{\text{IE}}) \cdot A_{\text{IE}\to\text{HS}}$$
$$F_{\text{HS}} = G_{\text{HS}} \cdot E_{\text{HS}} + (1 - G_{\text{HS}}) \cdot A_{\text{HS}\to\text{IE}}$$
$$F_{\text{IE}} = G_{\text{IE}} \cdot E_{\text{IE}} + (1 - G_{\text{IE}}) \cdot A_{\text{IE}\to\text{HS}}$$

A shared, learnable combination of these four components defines the final fused representation:

$$F_{\text{shared}} = \alpha_1 F_1 + \alpha_2 F_2 + \beta_1 F_{\text{HS}} + \beta_2 F_{\text{IE}}$$

This formulation ensures both fine-grained alignment and high-level semantic consistency across hate speech and intent modalities.

5.4 DialoRank: LLM-Based Response Ranking

To evaluate and select the most effective counterspeech among the generated intent-specific candidates, we employ `DialoGPT-small` as a judge model in a process we call **DialoRank**. For each sample, the model generates four counterspeech responses—one per intent—which are decoded and embedded into a prompt alongside the original hate speech.

DialoGPT is then prompted to assign a score from 1 to 10, reflecting the appropriateness and effectiveness of each response. The highest-scoring intent is selected as the predicted category. If this predicted intent matches any of the gold labels associated with the hate speech, it is considered a correct prediction. This allows us to compute an intent-level accuracy metric, which reflects how well the LLM ranks the generated responses according to human-like judgment.

6 Experimental Setup

6.1 DialoGPT

For the DialoGPT baseline, we fine-tuned the `microsoft/DialoGPT-small` model using the Hugging Face `Trainer` API. The training configuration was adopted from the setup described in (1).

Training Arguments:

- Epochs: `20`
- Batch size: `32`
- Learning rate: `8e-5`
- Weight decay: `0.03`
- Save strategy: `epoch`
- Logging steps: `10`
- Output directory: `./dialogpt_logs`

6.2 GPS

The GPS (Generate-Prune-Select) model was reimplemented and trained with the hyperparameter configuration adapted from the original paper (4).

Training Arguments:

- Epochs: `10`
- Batch size: `8`
- Learning rate: `1e-5`
- Hidden size (Encoder/Generator): `512`
- Latent dimension (`n_z`): `100`
- Word dropout: `0.5`
- Highway layers: `2`
- Embedding size: `100`
- Vocabulary size: `12000`
- Output size: `30000`
- Special tokens: `<unk>`, `<pad>`, `<sos>`, `<eos>`

6.3 Counter Speech Network

We trained the **Counter Speech Network** using a multi-decoder BART architecture, with each decoder tailored to one of the four intents: *informative*, *questioning*, *denouncing*, and *positive*. A fusion layer combines hate speech and intent embeddings before decoding.

The model was trained using the AdamW optimizer with the following hyperparameters:

Training Configuration:

- Epochs: `10`
- Batch size: `32`
- Learning rate: `5e-5`
- Optimizer: `AdamW`
- Max sequence length: `50`
- Hidden dimension: `768`
- Encoder output dimension: `256`
- Input dimension: `128`

7 Evaluation Metrics

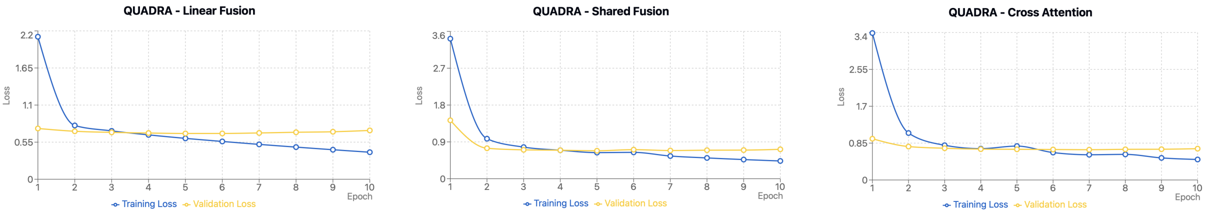
To comprehensively evaluate our model and baseline systems, we use both lexical and semantic metrics.

- **ROUGE-1, ROUGE-2, ROUGE-L:** These are recall-oriented metrics that measure the overlap between the generated counterspeech and the reference response. ROUGE-1 and ROUGE-2 capture unigram and bigram overlap respectively, while ROUGE-L captures the longest common subsequence.

- **METEOR:** A harmonic mean of unigram precision and recall, with additional alignment based on stemming and synonymy, making it more tolerant to paraphrasing than ROUGE.

- **BERTScore (BS):** We compute BERTScore precision, recall, and F1 using contextual embeddings from a pretrained language model to measure the semantic similarity between generated and reference texts.

- **Category Accuracy:** This metric assesses the model's ability to generate the most suitable intent-conditioned counterspeech. For each hate speech instance, the model produces four responses—one for each intent: `informative`, `questioning`, `denouncing`, and `positive`. These are evaluated by a pretrained **DialoGPT-small** model, referred to as **DialoRank**. DialoRank rates each response on a scale from 1 to 10 based on its appropriateness and effectiveness. The intent corresponding to the highest-scoring response is considered the predicted category. If this prediction matches any of the gold-standard intents associated with the hate speech instance, it is marked as correct. The final Category Accuracy is computed as the proportion of such correct predictions across all test samples.



Linear Fusion

Shared Fusion

Cross Attention

Figure 2: Visual comparison of generation performance for different fusion architectures.

Model	R1	R2	RL	M
Linear Fusion	0.250	0.064	0.175	0.154
Shared Fusion	**0.251**	**0.065**	**0.176**	**0.158**
Cross Attention	0.242	0.061	0.171	0.152
DialoGPT	0.130	0.003	0.105	0.040
GPS	0.176	0.030	0.132	0.116

Table 3: Text generation metrics across all models.

Model	BS(P)	BS(R)	BS(F1)	CA
Linear Fusion	0.869	**0.871**	0.870	0.752
Shared Fusion	**0.871**	0.870	**0.871**	0.751
Cross Fusion	0.870	0.869	0.870	0.752
DialoGPT	0.791	0.808	0.799	0.681
GPS	0.240	0.121	0.180	**0.754**

Table 4: Semantic similarity and classification metrics across all models. BS(P), BS(R), and BS(F1) refer to BERTScore Precision, Recall, and F1, respectively.

Metric	QUARC
ROUGE-1 (R1)	0.25
ROUGE-2 (R2)	0.08
ROUGE-L (RL)	0.24
METEOR (M)	0.22
Semantic Similarity (SS)	0.77
BERTScore (BS)	0.89
Category Accuracy (CA)	0.70

Table 5: Performance of **QUARC** on the **IntentCONAN** dataset with five communicative intents (1).

7.1 Discussion

Our results demonstrate the effectiveness of intent-aware modeling via **QUADRA**, a multi-decoder framework for counterspeech generation. All three fusion mechanisms—**Linear**, **Cross-Attention**, and **Shared Fusion**—consistently outperformed strong baselines (GPS, DialoGPT) across semantic and lexical metrics.

Shared Fusion achieved the highest overall performance, with a BERTScore-F1 of **0.871**, ROUGE-1 of **0.251**, ROUGE-L of **0.176**, and METEOR of **0.158**, outperforming DialoGPT by over **+7 BERTScore** and **+12 ROUGE-1**, and GPS by a large margin across all metrics. Notably, QUADRA matched GPS's top Category Accuracy (CA) of **0.75**, while surpassing it in fluency and relevance.

Compared to QUARC on IntentCONAN, QUADRA also improved CA from **0.70 to 0.75**, highlighting its strength in intent selection. Though DialoRank (based on DialoGPT) offered strong semantic matching, its zero-shot nature lacked consistent intent alignment—further validating QUADRA's intent conditioning approach.

Overall, **Shared Fusion + QUADRA** performed best by adaptively balancing hate speech and intent features through gated interpolation, leading to more coherent and intent-aligned responses. Unlike **Linear Fusion**, which lacked modulation, and **Cross-Attention**, which sometimes diluted intent focus, Shared Fusion offered a controlled yet flexible integration—crucial for generating rhetorically effective counterspeech.

8 Conclusion

We presented **QUADRA**, an intent-aware counterspeech framework with four BART decoders aligned to rhetorical intents and fused with HateBERT embeddings. Through structured fusion and zero-shot evaluation via **DialoRank**, QUADRA consistently outperforms strong baselines on semantic and classification metrics.

Our findings emphasize the importance of intent conditioning and cross-modal alignment for generating persuasive counterspeech. DialoRank further offers a scalable, interpretable alternative to human evaluation.

Future directions include more expressive classification heads, improved fusion alignment, and extensions to nuanced intents and multi-turn dialogues.

References

[1] Rishabh Gupta, Shaily Desai, Manvi Goel, Anil Bandhkavi, Tanmoy Chakraborty, and Md Shad Akhtar. 2023. *Counterspeeches up my sleeve! Intent Distribution Learning and Persistent Fusion for Intent-Conditioned Counterspeech Generation*. arXiv:2305.13776.

[2] Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. BART: Denoising Sequence-to-Sequence Pre-training for Natural Language Generation, Translation, and Comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 7871–7880, 2020. `https://arxiv.org/abs/1910.13461`.

[3] Yizhe Zhang, Siqi Sun, Michel Galley, Yen-Chun Chen, Chris Brockett, Xiang Gao, Jianfeng Gao, Jingjing Liu, and Bill Dolan. 2019. *DialoGPT: Large-Scale Generative Pretraining for Conversational Response Generation*. arXiv:1911.00536.

[4] Wanzheng Zhu and Suma Bhat. 2021. *Generate, Prune, Select: A Pipeline for Counterspeech Generation*. arXiv:2106.01625.

[5] 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*. arXiv:2403.10088.

[6] Tommaso Caselli, Valerio Basile, Jelena Mitrović, and Michael Granitzer. *HateBERT: Retraining BERT for Abusive Language Detection in English*. In Proceedings of the 2021 Workshop on Abusive Language Online, arXiv preprint arXiv:2010.12472.

[7] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in Neural Information Processing Systems (NeurIPS)*, pages 5998–6008, 2017.

[8] Ming Jiang and Shaoxiong Ji. Cross-Modality Gated Attention Fusion for Multimodal Sentiment Analysis. *arXiv preprint arXiv:2208.11893*, 2022. `https://arxiv.org/abs/2208.11893`.

Appendix

Model Weights

The pre-trained weights for all the models used in this project — including **DialoGPT FineTuned**, **Linear Fusion**, **Shared Fusion**, and **Cross Attention** — are available for download.

Download Model Weights on Google Drive

QUARC Model Implementation

The official implementation of the **QUARC** model is publicly available on GitHub:
Repository: https://github.com/LCS2-IIITD/quarc-counterspeech

The model was introduced in the following paper:

> *Counterspeeches up my sleeve! Intent Distribution Learning and Persistent Fusion for Intent-Conditioned Counterspeech Generation*
> **Rishabh Gupta, Shaily Desai, Manvi Goel, Anil Bandhakavi, Tanmoy Chakraborty, and Md. Shad Akhtar**
> arXiv:2305.13776