

# Fact-based Counter Narrative Generation to Combat Hate Speech

Brian Wilk  
University of Illinois Chicago  
Chicago, Illinois, USA  
bwilk26@uic.edu

Suman Kalyan Maity  
Missouri University of Science and Technology  
Rolla, Missouri, USA  
smaity@mst.edu

Homaira Huda Shomee  
University of Illinois Chicago  
Chicago, Illinois, USA  
hshome2@uic.edu

Sourav Medya  
University of Illinois Chicago  
Chicago, Illinois, USA  
medya@uic.edu

## Abstract

Online hatred has become an increasingly pervasive issue, affecting individuals and communities across various digital platforms. To combat hate speech in such platforms, counter narratives (CNs) are regarded as an effective method. In recent years, there has been growing interest in using generative AI tools to construct CNs. However, most of the generative models produce generic responses to hate speech and can hallucinate, reducing their effectiveness. To address the above limitations, we propose a counter narrative generation method that enhances CNs by providing non-aggressive, fact-based narratives with relevant background knowledge from two distinct sources, including a web search module. Furthermore, we conduct a comprehensive evaluation using multiple metrics, including LLM-based measures for persuasion, factuality, and informativeness, along with human and traditional NLP evaluations. Our method significantly outperforms baselines, achieving an average factuality score of 0.915, compared to 0.741, 0.701, and 0.69 for competitive baselines, and performs well in human evaluations.

## CCS Concepts

- **Computing methodologies** → **Natural language generation;**
- **Human-centered computing** → **Empirical studies in collaborative and social computing.**

## Keywords

Hate speech, Counter narrative, Fact-based narrative, Large language model

## ACM Reference Format:

Brian Wilk, Homaira Huda Shomee, Suman Kalyan Maity, and Sourav Medya. 2025. Fact-based Counter Narrative Generation to Combat Hate Speech. In *Proceedings of the ACM Web Conference 2025 (WWW '25)*, April 28-May 2, 2025, Sydney, NSW, Australia. ACM, New York, NY, USA, 12 pages. <https://doi.org/10.1145/3696410.3714718>

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than the author(s) must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from [permissions@acm.org](mailto:permissions@acm.org).

WWW '25, April 28-May 2, 2025, Sydney, NSW, Australia

© 2025 Copyright held by the owner/author(s). Publication rights licensed to ACM.

ACM ISBN 979-8-4007-1274-6/25/04

<https://doi.org/10.1145/3696410.3714718>

## 1 Introduction

The rise of social media has profoundly reshaped society, revolutionizing communication and allowing people to share knowledge, opinions, and experiences with a global audience almost instantaneously [18, 35]. These platforms have democratized information, giving a voice to individuals who might otherwise remain unheard, and have played a critical role in driving social movements, advocating for marginalized groups, and connecting people across geographic and cultural divides [29, 35]. However, while social media has brought numerous benefits, it has also given rise to a number of serious social issues. Among the most concerning is the spread of hate speech.

Hate speech thrives in the virtual environment of social media, where anonymity and distance from real-life consequences empower individuals to express harmful, abusive, and often discriminatory rhetoric without fear of accountability [24, 47]. The viral nature of social media allows toxic messages to spread rapidly, reaching vast audiences very quickly [21]. What might have once been an isolated comment in a small community can now go viral, exposing thousands or even millions of people to hateful rhetoric. The ease with which this content spreads and its broader reach allows hate speech to gain momentum quickly and become ingrained within online communities thereby normalizing abusive behavior and making it increasingly difficult to address.

**Widespread effects of online hate speech.** The surge in online abuse and hate poses a complex and pervasive threat to societal security. Victims of hate speech frequently experience profound psychological harm, such as anxiety, depression, and social isolation [1, 40]. Continuous exposure to toxic rhetoric erodes a person's sense of safety and self-worth, potentially leading to long-lasting mental health issues. This impact is particularly devastating for marginalized groups, who are often disproportionately targeted by online abuse. In these cases, hate speech exacerbates existing social inequalities, deepening the psychological toll. The repercussions of online hate are not confined to the digital space. Online hate speech often serves as a precursor to real-world violence, as extremist ideologies and hateful rhetoric spread across social media, inciting or inspiring acts of physical aggression [22, 25, 26]. The unchecked proliferation of hate speech on social media platforms threatens not only individuals but the broader fabric of society. It fosters division, fuels fear, and undermines the mutual respect and understanding essential for maintaining social cohesion.

**Counter narrative as promising combat strategy.** In response to the increasing prevalence of online hate, counter narrative

**Table 1: Comparison with previous works across various criteria of counter narrative generation. ‘Yes’ (‘No’) suggests that the particular criterion is present (absent) in the paper. The table suggests that TKGCN [7] as a competitive baseline. Our experiments (Section 4) show that our method outperforms TKGCN [7] in almost all settings.**

Paper	New Dataset	External Knowledge		Inject Facts	Eval. with Different LLMs	Group Eval.	Human Eval.
		Knowledge Repo.	Web Search				
MTCO[10]	Yes	No	No	No	No	No	Yes
RAUCG [15]	No	Yes	No	Yes	No	Yes	Yes
TKGCN [7]	Yes	Yes	No	Yes	Yes	Yes	Yes
GPS [46]	No	No	No	No	No	No	Yes
<b>Ours</b>	Yes	Yes	Yes	Yes	Yes	Yes	Yes

has emerged as a proactive and promising strategy [4, 33]. Unlike traditional methods such as content moderation, which primarily focus on removing harmful material post-dissemination, counter narrative strategy takes a more dynamic approach by preventing the spread of hate while also safeguarding free speech [4, 17, 33]. This strategy involves directly confronting harmful narratives with positive counterarguments, empowering individuals and communities to challenge toxic content as it arises. By interrupting the momentum of hate, counter narrative fosters more inclusive online discussions. Beyond simply mitigating the damage caused by on-line hate, this approach encourages a culture of empathy, respect, and constructive communication, facilitating a shift towards more positive interactions in digital spaces.

The existing literature on counter narrative generation faces several limitations. Despite several attempts of counter narrative generation, obtaining high-quality, effective and factual counter narrative remains a significant challenge. Only a few methods focus on the factual accuracy of counter narratives, which is crucial for effectively exposing the flaws in hate speech. This poses a particular issue as many state-of-the-art techniques rely on pre-trained LLMs, which are prone to generating hallucinated or inaccurate information. These factual inconsistencies can compromise the credibility of counter narratives and foster mistrust, reducing its impact. Additionally, there is a notable gap in the thorough evaluation of generated counter narratives, with limited attention given to assessing its factual accuracy, relevance, and overall efficacy in combating hate speech.

**Our contributions.** To address these limitations, we propose a novel method for generating fact-based counter narratives. Table 1 shows the major differences with the existing works. Our main contributions are as follows.

- *Novel Framework.* We improve the effectiveness of counter narratives (CNs) generated by large language models (LLMs), which often suffer from being generic. To enhance the quality of CNs, we propose a framework that responds with non-aggressive and fact-based feedback, by incorporating relevant background knowledge from two distinct sources.
- *Comprehensive Evaluation.* We conduct a comprehensive evaluation of the counter narratives using multiple metrics. We use LLM based metrics that measure the persuasion, factuality, and informativeness of our CNs, as well as target group-wise evaluation, human evaluation and traditional NLP metrics.

- *Results.* Our method significantly outperforms the baselines across several measures and settings. Our generated CNs achieve an average factuality score of 0.915, compared to the three competitive baselines achieving just 0.741, 0.701, and 0.69. Additionally, our generated CNs also perform well in terms of human evaluation.

- *Code.* We have made the codebase publicly available here: <https://github.com/000brian/counternarratives>

## 2 Related Work

### 2.1 Counter Narrative Dataset Creation

Numerous efforts have been made to create counter-narrative data. Mathew et al. [23] introduced the first dataset, which was created by annotating YouTube comments to identify counter narrative in response to hate speech. Qian et al. [31] published two large-scale hate speech intervention datasets. These datasets include conversations collected from social media platforms like Reddit<sup>1</sup> and Gab<sup>2</sup>, labeled as hate speeches, with 40K intervention responses written by 900 Mechanical Turk workers. Recognizing that the meaning of certain statements can change depending on context, Yu et al. [41] examined the role of conversational contexts in the annotation and detection of hate and counter speeches, releasing a context-aware dataset. Hundreds of non-expert annotators were invited to label Reddit comments with and without context as hate speech, counter narrative, or neutral speech. Chung et al. [6] introduced the first high-quality multilingual counter narrative dataset, CONAN, which contains around 6K pairs of hate speech and counter narratives in English, French, and Italian. This dataset was meticulously curated by over 100 well-trained NGO experts, requiring more than 500 person-hours to complete.

To reduce the manual labor of generating data, Tekiroğlu et al. [37] proposed an author-reviewer framework in which GPT-2 functions as the “author”, generating initial counter narrative responses. Experts then act as “reviewers”, tasked with filtering, refining, and post-editing these machine-generated responses. Fanton et al. [10] proposed a human-in-the-loop data collection method where experts post-edit generated counter narratives, iteratively refining the dataset to enhance the generative model’s output. Over 18 weeks, this method resulted in 5,000 pairs of hate speech and counter narratives. Bonaldi et al. [5] also utilized this author-reviewer pipeline but focused specifically on collecting counter

<sup>1</sup><https://www.reddit.com/>

<sup>2</sup><https://gab.com/>

narratives in multi-turn dialogues, further enhancing the quality and scope of the dataset.

## 2.2 Counter Narrative Generation

Several efforts have focused on generating counter narratives using Natural Language Generation (NLG) techniques. Qian et al. [31] evaluated the performance of various basic generative models, such as sequence-to-sequence models, for generating counter narratives and found that they often lacked relevance and variety, with many responses being generic or irrelevant. To address these limitations, Zhu et al. [46] proposed a three-module pipeline called Generate, Prune, Select (GPS), which enhances both data diversity and relevance. The method incorporates a retrieval-based selection mechanism. Chung et al. [7] integrated generative models with information retrieval techniques, leveraging external knowledge to enrich counter narratives and reduce hallucinations—situations. Jiang et al. [16] utilized stance consistency, semantic overlap, and relevance to hate speech in constructing a knowledge repository from the ChangeMyView subreddit<sup>3</sup>. This repository is then employed to provide knowledge-augmented counter narratives.

Several works have explored additional stylistic aspects of counter narratives to improve its effectiveness. To enhance specificity, Bonaldi et al. [5] applied two attention-based regularization techniques, incorporating a broader context during both the training and generation phases. Furman et al. [11] emphasized the argumentative structure within hate speech to guide the generation of responses. Saha et al. [32] focused on controlling multiple stylistic dimensions simultaneously, including politeness, detoxification, and emotional tone in the generated counter narratives. Gupta et al. [13] proposed a two-stage framework that conditions counter narrative generation on five different strategies: informative, denouncing, questioning, positive, and humorous approaches. These methods allow for contextually appropriate counter narratives, tailored to the nature of the hate speech being addressed.

The most common approach for generating counter narratives involves fine-tuning a pre-trained language model on a dedicated counter narrative dataset in a relatively low-computation setup (e.g., [14, 31, 36]). However, recent advances have enabled more flexible methods, such as generating counter narratives using few-shot learning techniques [3, 8, 11, 38]. Additionally, one-shot and zero-shot prompting can be useful for counter narrative generation with minimal training data [27, 44]. While these approaches significantly reduce the need for large-scale annotated datasets, making the generation process more accessible, the most of the counter narratives are not fact-based. *In this paper, we build a method to produce counter narrative that are fact-based and thus, more effective.*

## 3 Our Method

Large language models (LLMs) often generate generic or repetitive counter narratives (CNs), reducing their effectiveness in combating hate speech [9]. A more effective approach is to respond with non-aggressive, fact-based narratives that incorporates relevant background knowledge. This not only increases the relevance of the counter narrative to the specific hate speech but also enhances

diversity by avoiding repetition [2]. Our proposed framework addresses this by integrating three key components: (i) *document retrieval*, (ii) *document summarization*, and (iii) *counter narrative generation*. Figure 1 illustrates the workflow of these steps. We discuss each component in detail in the following subsections.

### 3.1 Document Retrieval

The first step involves retrieving relevant documents, which are subsequently used to craft fact-based CNs. Towards this objective, we generate targeted queries via LLM prompts.

**3.1.1 Query Generation.** This step focuses on generating queries to be used in the document retrieval task. We employ step-back prompting [43] to generate queries, where the model is prompted to first consider what information is necessary to construct a response. In our setting, it is crucial that the base LLM understands the goal of the query. For every instance of hate speech, we generate three specific queries denoted as  $Q = \{Q_1, Q_2, Q_3\}$  to help us retrieve the related documents containing facts and statistics to counter the hate speech. We use GPT4o<sup>4</sup> with the following prompt. Instead of using key phrases, we include the full instance of the hate speech to ensure that the model becomes aware of the entire context.

#### Prompt for Query Generation

You are responding with counter speech to the hate speech {hatespeech}.  
Generate 3 precise queries you would research in order to generate counterspeech for this hate speech.

**3.1.2 Document Retrieval.** Next, we aim to retrieve relevant documents from a knowledge repository. In this step, we use the Newsroom dataset [12] as our primary knowledge base, which includes 1.3 million articles from 38 different news publications. However, a limitation of the Newsroom dataset is that it only includes data from 1998 to 2017. To effectively combat hate speech with current information, it is crucial to integrate more recent datasets. Since facts and statistics are constantly evolving, relying solely on outdated data risks producing ineffective responses. To address this, we also incorporate the Tavily<sup>5</sup> online web search API to retrieve recent documents. This is especially designed for RAG (Retrieval-Augmented Generation) systems [19] and LLMs. It retrieves the most relevant information from the sources based on a query. Thus, we ensure that our generated counter narratives are grounded with the recent information.

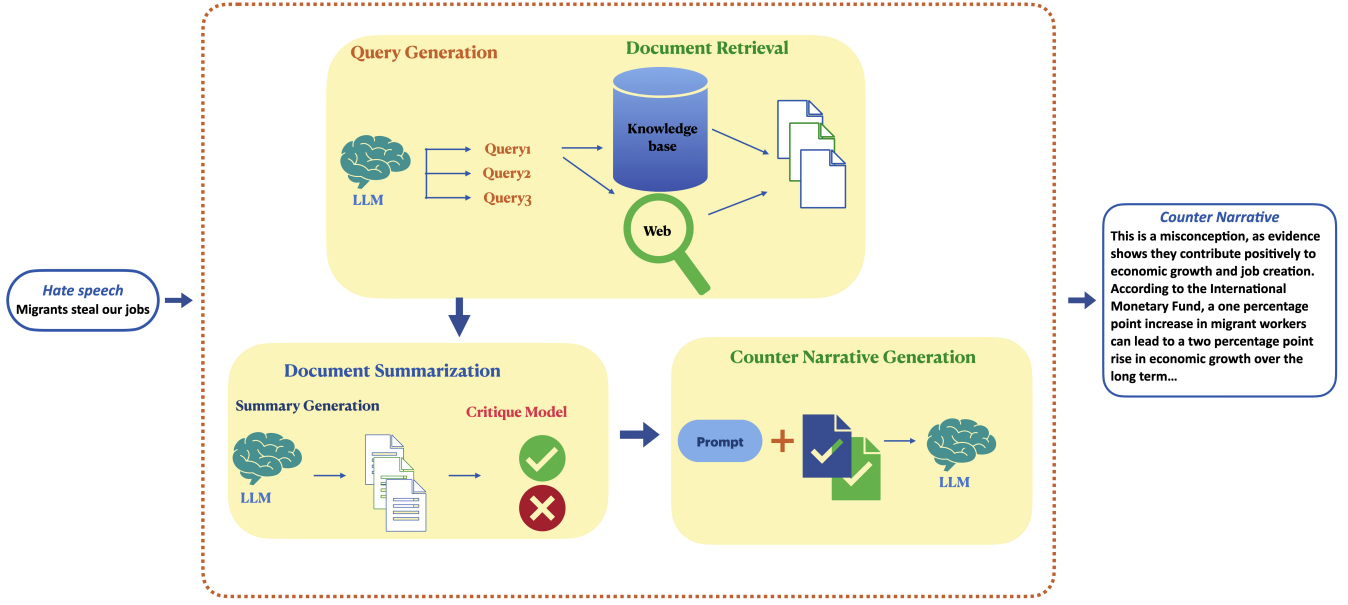
To manage and access large amounts of data, we depend on the low dimensional representations or vector embeddings space. By transforming both our knowledge base and queries into vector embeddings, we can perform similarity-based search to find the most relevant documents for each query. We use Chroma<sup>6</sup> to store and retrieve vector embeddings. First, we generate a vector store from our knowledge base. Then, for each query, it is converted to vector embeddings, which are used to retrieve relevant documents

<sup>3</sup><https://www.reddit.com/r/changemyview/>

<sup>4</sup><https://openai.com/index/hello-gpt-4o/>

<sup>5</sup><https://tavily.com/>

<sup>6</sup><https://www.trychroma.com/>



**Figure 1: Our pipeline for generating counter narratives involves three key stages: (1) Query Generation and Document Retrieval, where three queries are generated to retrieve relevant documents from a knowledge base and web search; (2) Document Summarization, where the retrieved documents are summarized, and a critique model determines whether they are relevant; and (3) Counter Narrative Generation, where the relevant documents are incorporated into a prompt to generate the final counter narrative.**

from the knowledge repository. Chroma uses *hierarchical navigable small world (HNSW)* [20] as an approximate nearest neighbor search algorithm. In this algorithm, documents are indexed into a hierarchical structure of graphs and edges. Connection between documents are based on their similarity, such as euclidean distance, cosine similarity, or inner product. Given a query, the algorithm starts the search at the top layer to find the nearest neighbor and go into deeper layer if the similarity is not sufficient. For each instance of hatespeech, we generate 3 queries as described in Section 3.1.1. We retrieve two documents per query from both Newsroom and online search which produces a total of 6 documents. The retrieved documents are denoted as  $D = \{D_{1N}, D_{1W}, D_{2N}, D_{2W}, D_{3N}, D_{3W}\}$  where, for the  $i^{\text{th}}$  query,  $D_{iN}$  and  $D_{iW}$  refer to the documents from Newsroom and the web respectively.

### 3.2 Document Summarization

We obtain relevant documents in the previous step (Section 3.1.2). The challenge with these retrieved documents is that they are lengthy, and can sometimes be irrelevant and noisy. Before injecting information that might not accurately align with the hate speech and subsequent CNs, we need to verify the relevance. So, we split this process into two steps: first, we apply techniques to summarize the key points from the relevant documents, and then we use the critique model to filter and select the relevant documents.

**3.2.1 Summary Generation.** The documents in the retrieved set  $D$  are usually lengthy. Besides, LLMs also have a certain limit for

tokens that are used for prompts. Additionally, not all paragraphs in the documents contain relevant facts needed for our generation task. To address this, we generate summaries focusing on key facts, and statistics from the documents. We employ GPT-4o to summarize the contents as it has shown a great potential in generating automated text summarization in various domain such as food summary [34], discharge summary for a patient [39], news summary [28], etc. Please see Appendix A.3 for the detailed prompts. This process produces the set of summarized documents  $D_S = \{D_{1NS}, D_{1WS}, D_{2NS}, D_{2WS}, D_{3NS}, D_{3WS}\}$  where  $D_{iNS}$  refers to the summarized document from Newsroom and  $D_{iWS}$  refers to the document summary retrieved from the web for the  $i^{\text{th}}$  query.

**3.2.2 Critique Model.** As the critique model, we utilize GPT-4o with a specific prompt. The prompt includes the query used to generate the documents, the retrieved documents, the identified hate speech, and the purpose of the documents, which is to counter the hate speech. The output of the model (GPT-4o) for a document is binary with "yes" or "no". The detailed prompt used for this process is provided in Appendix A.3. After processing with the critique model, we derive a subset of relevant documents  $D_C$  where the final document set can be of any number between one to six. Generally speaking, the critique model is used for filtering the irrelevant ones. Now there is a possibility that the critique model might yield an empty set when no relevant documents have been found. However, after testing with multiple hate speech instances,

we observe that this scenario is rare. The critique model always returns at least one relevant document for the task in all cases.

#### Example of Hate Speech (HS) & Counter Narratives (CNs)

[t] **HS:** Gays are spreading AIDS all around the world because they are promiscuous and fornicate with strangers, we should prevent them to make blood donations.

**CN (MTCO):** We can simply disallow donations from those who had unprotected sex or a high number of partners, instead of making sexual orientation the basis as in your claim.

**CN (TKGCN):** The real issue is unprotected sex. Regardless of sexual orientation, it is dangerous for anyone to have unprotected sex that could lead to a life threatening disease.

**CN (Ours) :** The notion that gay individuals are responsible for spreading AIDS is a harmful stereotype not supported by scientific evidence. Studies have shown that HIV transmission can be effectively prevented when the HIV-positive partner is on antiretroviral treatment, regardless of sexual orientation ([source: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC8063169/](https://www.ncbi.nlm.nih.gov/pmc/articles/PMC8063169/)). Furthermore, the Red Cross implements rigorous safety protocols to ensure that all eligible individuals, regardless of sexual orientation, can safely donate blood ([source: https://www.redcrossblood.org/donate-blood/dlp/share-your-health.html](https://www.redcrossblood.org/donate-blood/dlp/share-your-health.html)). Discriminatory practices based on sexual orientation hinder efforts to combat HIV/AIDS and ignore the fact that transmission is related to specific behaviors, not sexual orientation itself.

### 3.3 Counter Narrative Generation

To effectively counter hate speech, our final step involves generating counter narratives that directly address the underlying misinformation, prejudice, or harmful rhetoric present in the hate speech instance. These counter narratives aim to promote fact-based corrections. For this generation process, we employ several LLMs including GPT-4o<sup>7</sup>, GPT-3.5<sup>8</sup>, Gemini 1.5 Flash<sup>9</sup>, Claude Sonnet 3.5<sup>10</sup>, and Mistral 7B<sup>11</sup>. The detailed prompt is provided in Appendix A.3. The task of the model is to generate a counter narrative that is both factually accurate and aligned with the goal of promoting positive, alternative perspectives. The prompt  $p = \{h_{si}, D_{Ci}, instruct\}$  includes the hate speech instance  $h_{si}$ , the summarized documents  $D_{Ci}$  and instructions for generating a counter narrative of  $i^{th}$  instance of the dataset. After running evaluations based on toxicity, factuality, persuasiveness, and informativeness as shown in Table 2, we select GPT-4o as the preferred model. It consistently produces non-toxic and coherent counter narratives, outperforming the other models in our trial. Examples of hatespeech and counter narratives generated by Multi-Target CONAN (MTCO) [10], TKGCN [7], and

ours are shown (left). Note that, counter narratives generated by our model has more facts and it cites the sources (shown in green). We find that on average a complete run takes 75 seconds. The expensive steps are the document retrieval (average 11 seconds) and the summarization (48 seconds) steps. One can lower the processing time by limiting the amount of documents retrieved.

## 4 Experimental Results

In this section, we discuss the experimental set up and provide results on various measures. More specifically,

- We use various evaluation metrics to validate our counter narrative (CNs) generations in terms of NLP measures, fact-based measures, and linguistic quality. We also perform human evaluation of our generated CNs.
- Since our pipeline consists of multiple stages or components, we assess the quality of each step individually. Furthermore, we evaluate the CNs for individual target group. Finally, we make some interesting observations between the original hate speech and our constructed CNs.

### 4.1 Set up

**Datasets.** We use the Multi-Target CONAN (MTCO) [10] dataset for our analysis. This dataset has hate speech and counter narrative pairs in English, targeting multiple groups. It has been collected using a Human-in-the-Loop approach to ensure quality and relevance. The dataset contains 5,003 hate speech and counter narrative pairs, addressing various hate targets such as people with disabilities, Jews, LGBT+ , Muslims, migrants, people of color (POC), and women. A few examples are shown in Appendix A.1.

**Baselines.** Our first baseline is the Multi-Target CONAN (MTCO) [10] dataset. This paper proposes a human-in-the-loop approach to generate counter narratives. As seen in the examples in Table 7, the CNs produced by this method are not fact-based. We also compare our results with TKGCN [7], where counter narratives are similarly grounded in knowledge from a repository and GPS [46] which generates multiple candidate samples and refines them using a retrieval-based selection mechanism. However, our method outperforms these baselines in almost all settings. Another relevant work, RAUCG [15] has not made the code and data publicly available, and therefore we have not included this baseline for our experiments.

### 4.2 Evaluation of Counter Narrative Generation on Different Measures

We evaluate the quality of our counter narrative generation with various measures in this experiment. Table 2 presents all the results along with three main baselines: TKGCN [7], MTCO [10], and GPS [46]. Table 8 (see Appendix) shows the differences in metrics between our method and the baselines is statistically significant. We also show variations with different LLMs for the last step where we generate the counter narratives (CNs).

We would like to emphasize the fact that it is challenging to evaluate the CNs because of the lack of ground truth. Thus, we evaluate these from different measures that are practical for CNs. We describe each measure along with results in details as follows.

<sup>7</sup><https://openai.com/index/hello-gpt-4o/>

<sup>8</sup><https://platform.openai.com/docs/models/gpt-3-5-turbo>

<sup>9</sup><https://deepmind.google/technologies/gemini/flash/>

<sup>10</sup><https://docs.anthropic.com/en/docs/welcome>

<sup>11</sup><https://mistral.ai/news/announcing-mistral-7b/>

**Table 2: Evaluation of the counter narratives with different measures. We compare our proposed model with three main baselines: TKGCN [7], MTCO [10] and GPS [46]. We also show variations with different LLMs for the last step where we generate the counter narratives. "Tox.", "Fact.", "Pers.", "Inf.", and "LQ" refer to toxicity, factuality, persuasiveness, informativeness, and linguistic quality, respectively. For toxicity, lower values indicate better performance, whereas for the other metrics higher values are better. The best scores are highlighted in bold. For BLEU and BERTScore we compare all the CNs with the given CNs from MTCO. Our method outperforms others especially in terms of factuality, persuasiveness, and informativeness.**

	Model	Tox. HS	CN	Fact.	Pers.	Inf.	LQ	BLEU	BERTScore
<b>Baselines</b>	<b>MTCO [10]</b>	0.501	0.161	0.701	0.620	0.661	0.791	–	–
	<b>TKGCN [7]</b>	0.501	0.201	0.741	0.673	0.695	0.845	<b>0.046</b>	0.847
	<b>GPS [46]</b>	0.501	0.152	0.696	0.634	0.655	0.786	0.026	0.843
<b>Our Methods</b>	<b>Gemini</b>	0.501	0.145	0.892	0.845	0.808	0.849	0.030	0.847
	<b>Claude</b>	0.501	0.159	0.905	0.858	0.874	0.854	0.024	0.840
	<b>Mistral</b>	0.501	0.125	0.866	0.824	0.764	0.831	0.036	0.850
	<b>GPT3.5</b>	0.501	<b>0.094</b>	0.866	0.820	0.773	<b>0.859</b>	0.034	<b>0.851</b>
<b>Our Main Model</b>	<b>GPT4o</b>	0.501	0.131	<b>0.915</b>	<b>0.859</b>	<b>0.879</b>	0.832	0.02	0.827

**Toxicity.** Toxicity in hate speech is expected to be high. In contrary, toxicity does not help in counter narrative (CN) for obvious reasons. Our objective is to ensure that our created CN exhibits lower toxicity. To validate this, we utilize the Perspective API<sup>12</sup> developed by Jigsaw and Google. This is based on the BERT-based model and Convolutional Neural Networks (CNNs). For a given text, it returns a probability score ranging from 0 to 1, where 0 indicates the text is non-toxic, and 1 means it is highly likely to be toxic. In Table 2, we present the toxicity levels of both hate speech and generated counter narratives using different LLMs. The hatespeech has a toxicity score of 0.50, where the lowest toxicity score for the CN is 0.094, produced by GPT3.5. The highest score, 0.20, comes from the counter narratives proposed by [7]. Our proposed model achieves a score of 0.13 which is a significant improvement over the existing works.

**Factuality, Persuasiveness, and Informativeness.** This is one of our major experiments. We follow the method used by [15] to measure factuality, persuasiveness and informativeness. We use GPT 3.5 to score these three criteria in the range of 0 to 1. We aim to have higher scores in all the three metrics as our objective is to generate CNs that are informative, factually reliable and most importantly persuasive. Here, persuasiveness means how the CNs are effective to influence the readers opinion. Table 2 shows that our CNs receive the highest scores across all three metrics: 0.91 for factuality, 0.85 for persuasiveness, and 0.87 for informativeness. In comparison, the method in [7] produces lower scores such as 0.74, 0.67, and 0.69 for these metrics respectively.

**Linguistic Quality.** Most NLP metrics (e.g., BLEU, BERTScore) rely on reference texts to evaluate generated text. GRUEN [45], however, is a reference-less linguistic quality metric that assesses text quality based on several aspects: grammar, non-redundancy, focus, and coherence. It evaluates whether the text is grammatically correct, free of unnecessary repetition, maintains topic relevance within paragraphs, and flows coherently between sentences. GRUEN provides a holistic quality score. Table 2 shows that our

approach with GPT-3.5 achieves the highest score (0.85) where as our main model produces 0.83. The slight difference may be due to the fact that GPT-3.5 does not include any source while injecting facts to the CNs where our main model does the same with relevant sources. We also provide examples in Table 9 (Appendix A.2).

**BLEU-2.** We also aim to evaluate the similarity of our CNs with the original CNs in the Multi-Target CONAN (MTCO) [10] dataset. The BLEU [30] score is commonly used to assess such quality. BLEU-2 focuses on 2-gram precision, meaning it evaluates how well pairs of consecutive words (bigrams) in the candidate text match those in the reference. Here, we chose MTCO [10] as our reference text. The score ranges from 0 to 1, with higher scores indicating closer alignment with the reference text. As shown in Table 2, our CNs do not exhibit significant bigram overlap with reference CNs. This is expected as our CNs incorporate a moderate number of factual statements, which naturally reduces bigram similarity but enhances the informative value of the generated CNs. In comparison, the baseline TKGCN is less factual and shows a higher bigram similarity.

**BERTScore [42].** This is used to capture the semantic similarity between the generated text and reference texts by utilizing contextual embeddings. This evaluation focuses on how well the meanings of the generated text align with the reference text—which is Multi-Target CONAN (MTCO) [10] data in our case—rather than relying solely on exact word matches. This metric helps us to validate that our counter narratives (CN) maintain a level of semantic similarity, even though the inclusion of factual content may cause slight deviations from the reference. As shown in Table 2, all models achieve relatively similar scores, though our method produces the lowest one as our CNs aim to have other components such as factuality.

### 4.3 Human Evaluation

Human evaluation is widely regarded as the gold standard for assessing the quality of text generation especially by generative models such as LLMs. To validate our counter narratives (CNs), we randomly chose three samples and ask 29 evaluators to evaluate CNs based on the followings. **(1) Factuality:** Do the counter narratives include facts and statistics? **(2) Coherence:** Are the

<sup>12</sup><https://perspectiveapi.com/>

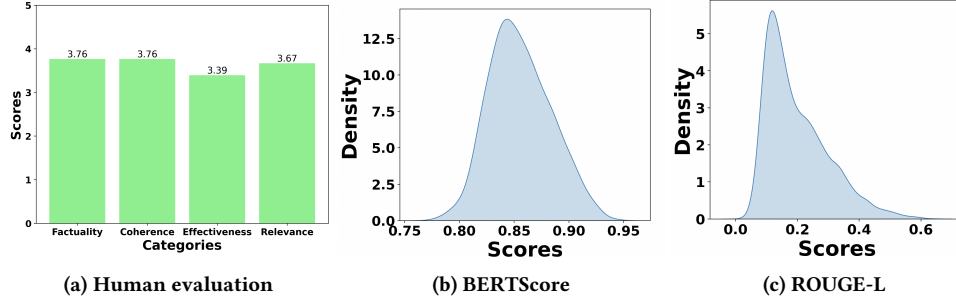


Figure 2: (a) Results of human evaluation across four categories: Factuality, Coherence, Effectiveness, and Relevance. In most of them the output are more than 3.5 out of 5, indicating the high quality of the generated counter narratives. (b-c) Comparison of BERTScore scores and ROUGE-L scores between documents and summaries. Low scores for ROUGE-L suggest that the summaries do not retain similar words as the original documents. However, the summaries retain the semantic meaning of the original documents as indicated by the high BERTScore values.

counter narratives logically structured and easy to understand? (3) **Effectiveness**: Is the counter narrative likely to be persuasive and change someone’s perspective on the issue? (4) **Relevance**: Are the facts and information presented in the counter narrative directly related to the content of the hate speech? Our study was determined to not involve direct human subjects research and was granted an exemption by the Institutional Review Board.

The users are asked to rate each counter narrative on a scale of 1 to 5, with 5 indicating excellent performance and 1 indicating poor performance. Figure 2a shows the results of human evaluation. In every category, the output are more than 3.5 out of 5 except for *effectiveness*. The slightly lower score in this category may indicate that while the counter narratives are coherent and factual, the users—who evaluate these—think that (or are unsure about whether) these counter narrative may not always have an effect on shifting opinions or persuading the audience. The 29 evaluators represent diverse demographics, including nationalities from Bangladesh, China, India, Iran, the United States, and Vietnam, with both female and male evaluators. Their professions include a mix of researchers and non-researchers. This diversity ensures a balanced evaluation by incorporating varied cultural perspectives, which is crucial for assessing the global impact and relevance of CNs.

#### 4.4 Component-wise Evaluation

Our framework has multiple steps and, in these experiments, we evaluate the quality of the individual steps.

**4.4.1 Document Retrieval.** In order to inject facts, our framework has a document retrieval component that obtains six documents per hate speech, summarizes them, and assesses their relevance using a critic model. The critic model filters out the summaries that are irrelevant for countering hate speech, ensuring only pertinent information is utilized in the counter narrative generation. In Table 3, we present the average amount of summaries that are irrelevant per hate speech. The low numbers indicate that the documents retrieved are mostly relevant as low amount of filtering is needed. The low scores across each target group show that the retrieval step consistently retrieves relevant documents.

Table 3: Number of document summaries filtered out per target group. For each hate speech, six documents are retrieved and summarized. Lower scores of filtering here suggest that the documents retrieved are indeed relevant to counter the hate speech.

Target	Summaries Filtered
DISABLED	0.074
JEWS	0.144
LGBT+	0.175
MIGRANTS	0.156
MUSLIMS	0.106
POC	0.097
WOMEN	0.045
OTHER	0.156

**4.4.2 Summarization.** In our framework, the document summarization is an important step. In this experiment, we aim to capture the similarity between the original document and its corresponding generated summary. In Figure 2b and 2c, we present the BERTScore and ROUGE-L between every document retrieved and its summary. The scores are relatively low for ROUGE-L, showing little n-gram overlap due to paraphrasing or alternative wording. However, BERTScore performs generally well, showing that the core semantic meaning is still retained in our summaries. This implies that our summarization approach effectively captures the main ideas while expressing them in different words.

#### 4.5 Quantitative Analysis for Target Groups

This experiment aims to analyze the counter narratives in each target group. In Table 4, we show a detailed comparison of the CNs generated by our model against those from the Multi-Target CONAN (MTCO) [10] dataset as it is our competitive baseline. This dataset [10] has eight target groups, and here, we present target-group wise measures. The metrics include the average number of words and sentences, factuality scores, and toxicity scores. Our model consistently generates longer CNs, as measures by the higher word and sentence counts across all target groups. It shows that

**Table 4: Target group-wise evaluation of counter narratives. The results show the average number of words, sentences, factuality score, and toxicity score for each target group in both our produced CNs and the original CNs in the Multi-Target CONAN (MTCO) dataset. The best results on factuality and toxicity are in bold. We observe higher word and sentence counts in ours due to factual information. In most of the target groups, our produced CNs achieve lower (better) toxicity scores.**

Target	Words		Sentences		Factuality		Toxicity	
	Ours	MTCO	Ours	MTCO	Ours	MTCO	Ours	MTCO
DISABLED	113.77	27.81	2.67	1.63	<b>0.909</b>	0.700	<b>0.060</b>	0.088
JEWS	125.55	27.96	4.57	2.10	<b>0.959</b>	0.644	0.229	<b>0.204</b>
LGBT+	114.60	26.23	3.68	1.80	<b>0.924</b>	0.650	<b>0.117</b>	0.217
MIGRANTS	117.42	29.17	5.34	2.24	<b>0.902</b>	0.688	<b>0.083</b>	0.084
MUSLIMS	122.91	23.81	4.71	1.74	<b>0.901</b>	0.675	<b>0.112</b>	0.178
POC	117.34	27.08	3.94	2.00	<b>0.928</b>	0.731	0.207	<b>0.201</b>
WOMEN	119.60	28.70	4.54	2.00	<b>0.900</b>	0.708	<b>0.121</b>	0.162
OTHER	115.14	23.91	4.15	1.50	<b>0.909</b>	0.734	<b>0.102</b>	0.129

**Table 5: Similarity evaluation between one CN compared with every HS in the target group. Lower scores (best are in bold) suggest that the CNs are less generic and less directly aligned with the hate speech within each category.**

Model	BLEU	ROUGE-L	BERTscore
TKGCN [7]	0.0064	0.0826	0.8406
MTCO [10]	0.0100	0.0964	0.8560
<b>Ours</b>	<b>0.0037</b>	<b>0.0516</b>	<b>0.8226</b>

our model includes more detailed and comprehensive responses. Additionally, our model achieves higher factuality scores which implies its ability to produce fact-based CNs. In terms of toxicity, our model generally has lower or similar scores compared to the MTCO dataset. This shows that even though our CNs are longer and more detailed, they do not become more harmful or hateful.

## 4.6 Interesting Observations

**4.6.1 Are the generated CNs generic?** One of the limitations of the existing counter narrative methods is that the produced counter narratives tend to be generic, which makes them less effective. We aim to show that the CNs—produced by our method—are unique and specifically tailored to each instance of hate speech. In Table 5, we present the BLEU, ROUGE-L, and BERTScore between one CN and every hate speech in a specific target group, and then average those scores. We repeat the process for each data point and report the average score. The results show that the baselines MTCO [10] and TKGCN [7] have higher similarity between the CN and hate speech, whereas our model has less similarity. Since BERTScore accounts for contextual similarity, it still produces a high score, though it is the lowest compared to the other works.

**4.6.2 How similar are the HS and the corresponding CN?** To answer this, we also present the BLEU, ROUGE-L, and BERTScore values between every pair of counter narratives (CNs) and hate speech in Table 6. The scores are notably low which shows that the CNs are not simply repeating words or n-grams from the hate speech. It

**Table 6: Similarity evaluation between HS and CN. The low scores (best results are in bold) in our CNs indicate that the words from the HS are not retained in our CNs, validating that our model constructs more diverse CNs.**

Model	BLEU	ROUGE-L	BERTscore
TKGCN [7]	0.0468	0.1275	0.8523
MTCO [10]	0.0719	0.1712	0.8698
<b>Ours</b>	<b>0.0268</b>	<b>0.0840</b>	<b>0.8370</b>

also validates that our model generates more diverse and context-specific counter narratives, rather than relying on direct overlaps with the original text.

## 5 Conclusion

In this paper, we have designed a novel method for generating fact-based counter narratives. By integrating relevant background knowledge from multiple sources and focusing on non-aggressive language and facts, our approach significantly improves the quality and effectiveness of CNs in combating hate speech. Our comprehensive evaluation demonstrates the strength of our method across various metrics, including persuasion, factuality, and informativeness. Both LLM-based assessments and human evaluations show that our approach consistently outperforms competitive baselines. These results underscore the potential of our framework to significantly enhance counter narrative generation and its practical application in addressing online hate speech at scale.

**Limitations and Future Work.** Our work assumes that the hatred on social media is primarily expressed through text. However, it is important to recognize that hate can also appear in images, which remains a significant challenge to address. Additionally, future research could explore the creation of multi-response counter-narratives.

## Acknowledgments

We acknowledge the UIUC HACC Cluster and the NSF ACCESS UIUC NCSA Cluster (Ref: ELE230014) for their in-kind support and computational resources for this work.

## References

- [1] 2023. The Impact of Hate Crime and Discrimination on Mental Health. <https://www.stopthateuk.org/2023/08/30/the-impact-of-hate-crime-and-discrimination-on-mental-health/>
- [2] Nami Akazawa, Serra Sinem Tekiroğlu, and Marco Guerini. 2023. Distilling implied bias from hate speech for counter narrative selection. In *Proceedings of the 1st Workshop on CounterSpeech for Online Abuse (CS4OA)*. 29–43.
- [3] Mana Ashida and Mamoru Komachi. 2022. Towards automatic generation of messages countering online hate speech and microaggressions. In *Proceedings of the Sixth Workshop on Online Abuse and Harms (WOAH)*. 11–23.
- [4] Susan Benesch. 2014. Countering dangerous speech: New ideas for genocide prevention. Available at SSRN 3686876 (2014).
- [5] 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*. Association for Computational Linguistics, 8031–8049. <https://doi.org/10.18653/v1/2022.emnlp-main.549>
- [6] Yi-Ling Chung, Elizaveta Kuzmenko, Serra Sinem Tekiroğlu, and Marco Guerini. 2019. CONAN - Counter Narratives through Nicheourcing: a Multilingual Dataset of Responses to Fight Online Hate Speech. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, 2819–2829. <https://doi.org/10.18653/v1/P19-1271>
- [7] Yi-Ling Chung, Serra Sinem Tekiroğlu, and Marco Guerini. 2021. Towards knowledge-grounded counter narrative generation for hate speech. *arXiv preprint arXiv:2106.11783* (2021).
- [8] Mekseline Doğanç and Ilia Markov. 2023. From generic to personalized: Investigating strategies for generating targeted counter narratives against hate speech. In *Proceedings of the 1st Workshop on CounterSpeech for Online Abuse (CS4OA)*. 1–12.
- [9] Mekseline Doğanç and Ilia Markov. 2023. From generic to personalized: Investigating strategies for generating targeted counter narratives against hate speech. In *Proceedings of the 1st Workshop on CounterSpeech for Online Abuse (CS4OA)*. 1–12.
- [10] Margherita Fanton, Helena Bonaldi, Serra Sinem Tekiroğlu, and Marco Guerini. 2021. Human-in-the-loop for data collection: a multi-target counter narrative dataset to fight online hate speech. *arXiv preprint arXiv:2107.08720* (2021).
- [11] Damián Furman, Pablo Torres, José Rodríguez, Diego Letzen, Maria Martinez, and Laura Alemany. 2023. High-quality argumentative information in low resources approaches improve counter-narrative generation. In *Findings of the Association for Computational Linguistics: EMNLP 2023*. Association for Computational Linguistics, Singapore, 2942–2956. <https://doi.org/10.18653/v1/2023.findings-emnlp.194>
- [12] Max Grusky, Mor Naaman, and Yoav Artzi. 2018. Newsroom: A dataset of 1.3 million summaries with diverse extractive strategies. *arXiv preprint arXiv:1804.11283* (2018).
- [13] 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)*. Association for Computational Linguistics, Toronto, Canada, 5792–5809. <https://doi.org/10.18653/v1/2023.acl-long.318>
- [14] Sadaf MD Halim, Saquib Irtiza, Yibo Hu, Latifur Khan, and Bhavani Thuraisingham. 2023. Wokeypt: Improving counterspeech generation against online hate speech by intelligently augmenting datasets using a novel metric. In *2023 International Joint Conference on Neural Networks (IJCNN)*. IEEE, 1–10.
- [15] Shuyi Jiang, Wenyi Tang, Xingshu Chen, Rui Tanga, Haizhou Wang, and Wenxian Wang. 2023. Raucg: Retrieval-augmented unsupervised counter narrative generation for hate speech. *arXiv preprint arXiv:2310.05650* (2023).
- [16] Shuyi Jiang, Wenyi Tang, Xingshu Chen, Rui Tanga, Haizhou Wang, and Wenxian Wang. 2023. Raucg: Retrieval-augmented unsupervised counter narrative generation for hate speech. *arXiv preprint arXiv:2310.05650* (2023).
- [17] Svetlana Kiritchenko, Isar Nejadgholi, and Kathleen C Fraser. 2021. Confronting abusive language online: A survey from the ethical and human rights perspective. *Journal of Artificial Intelligence Research* 71 (2021), 431–478.
- [18] Amanda Lenhart, Kristen Purcell, Aaron Smith, and Kathryn Zickuhr. 2010. Social Media & Mobile Internet Use among Teens and Young Adults. Millennials. *Pew internet & American life project* (2010).
- [19] Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, et al. 2020. Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in Neural Information Processing Systems* 33 (2020), 9459–9474.
- [20] Yu A Malkov and Dmitry A Yashunin. 2018. Efficient and robust approximate nearest neighbor search using hierarchical navigable small world graphs. *IEEE transactions on pattern analysis and machine intelligence* 42, 4 (2018), 824–836.
- [21] Binny Mathew, Ritam Dutt, Pawan Goyal, and Animesh Mukherjee. 2019. Spread of hate speech in online social media. In *Proceedings of the 10th ACM conference on web science*. 173–182.
- [22] Binny Mathew, Anurag Illendula, Punyajoy Saha, Soumya Sarkar, Pawan Goyal, and Animesh Mukherjee. 2020. Hate begets hate: A temporal study of hate speech. *Proceedings of the ACM on Human-Computer Interaction* 4, CSCW2 (2020), 1–24.
- [23] 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*, Vol. 13. 369–380.
- [24] Mainack Mondal, Leandro Araújo Silva, and Fabricio Benevenuto. 2017. A measurement study of hate speech in social media. In *Proceedings of the 28th ACM conference on hypertext and social media*. 85–94.
- [25] Karsten Müller and Carlo Schwarz. 2018. Making America hate again? Twitter and hate crime under Trump. *SSRN Electronic Journal* (2018).
- [26] Karsten Müller and Carlo Schwarz. 2021. Fanning the flames of hate: Social media and hate crime. *Journal of the European Economic Association* 19, 4 (2021), 2131–2167.
- [27] Jimin Mun, Emily Allaway, Akhila Yerukola, Laura Vianna, Sarah-Jane Leslie, and Maarten Sap. 2023. Beyond Denouncing Hate: Strategies for Countering Implied Biases and Stereotypes in Language. In *Findings of the Association for Computational Linguistics: EMNLP 2023*. Association for Computational Linguistics, Singapore, 9759–9777. <https://doi.org/10.18653/v1/2023.findings-emnlp.653>
- [28] Hilário Oliveira and Rafael Dueire Lins. 2024. Assessing Abstractive and Extractive Methods for Automatic News Summarization. In *Proceedings of the ACM Symposium on Document Engineering* 2024. 1–10.
- [29] Esteban Ortiz-Ospina. 2019. Over 2.5 billion people use social media. This is how it has changed the world. *World Economic Forum Agenda* (2019). <https://www.weforum.org/agenda/2019/10/rise-of-social-media/>
- [30] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*. 311–318.
- [31] Jing Qian, Anna Bethke, Yinyin Liu, Elizabeth Belding, and William Yang Wang. 2019. A benchmark dataset for learning to intervene in online hate speech. *arXiv preprint arXiv:1909.04251* (2019).
- [32] Punyajoy Saha, Kanishk Singh, Adarsh Kumar, Binny Mathew, and Animesh Mukherjee. 2022. CounterGeDi: A controllable approach to generate polite, detoxified and emotional counterspeech. *arXiv preprint arXiv:2205.04304* (2022).
- [33] Carla Schieb and Mike Preuss. 2016. Governing hate speech by means of counterspeech on Facebook. In *66th ica annual conference, at fukuoka, japan*. 1–23.
- [34] Yiwen Shi, Ping Ren, Jing Wang, Biao Han, Taha ValizadehAslani, Felix Agbavor, Yi Zhang, Meng Hu, Liang Zhao, and Hualou Liang. 2023. Leveraging GPT-4 for food effect summarization to enhance product-specific guidance development via iterative prompting. *Journal of biomedical informatics* 148 (2023), 104533.
- [35] Waralak V. Siricharoen. 2023. Social Media as Communication–Transformation Tools. In *Information Systems for Intelligent Systems*, Chakchai So-In, Narendra D. Londhe, Nityesh Bhatt, and Meelis Kitsing (Eds.). Springer Nature Singapore, Singapore, 1–11.
- [36] Serra Sinem Tekiroğlu, Helena Bonaldi, Margherita Fanton, and Marco Guerini. 2022. Using Pre-Trained Language Models for Producing Counter Narratives Against Hate Speech: a Comparative Study. In *Findings of the Association for Computational Linguistics: ACL 2022*. Association for Computational Linguistics, Dublin, Ireland, 3099–3114. <https://doi.org/10.18653/v1/2022.findings-acl.245>
- [37] Serra Sinem Tekiroğlu, Yi-Ling Chung, and Marco Guerini. 2020. Generating Counter Narratives against Online Hate Speech: Data and Strategies. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, 1177–1190. <https://doi.org/10.18653/v1/2020.acl-main.110>
- [38] Maria Estrella Vallecillo-Rodríguez, Arturo Montejo-Raéz, and Maria Teresa Martín-Valdivia. 2023. Automatic counter-narrative generation for hate speech in Spanish. *Procesamiento del Lenguaje Natural* 71 (2023), 227–245.
- [39] Ethan Waisberg, Joshua Ong, Mouayad Masalkhi, Sharif Amit Kamran, Nasif Zaman, Prithul Sarker, Andrew G Lee, and Alireza Tavakkoli. 2023. GPT-4: a new era of artificial intelligence in medicine. *Irish Journal of Medical Science* (1971-) 192, 6 (2023), 3197–3200.
- [40] Mark Walters. 2014. Repairing the harms of hate crime: A restorative justice approach. In *The Routledge International Handbook on Hate Crime*. Routledge, 400–410.
- [41] Xinchun Yu, Eduardo Blanco, and Lingzi Hong. 2022. Hate Speech and Counter Speech Detection: Conversational Context Does Matter. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. 5918–5930. <https://doi.org/10.18653/v1/2022.naacl-main.433>
- [42] Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q Weinberger, and Yoav Artzi. 2019. BERTscore: Evaluating text generation with bert. *arXiv preprint arXiv:1904.09675* (2019).
- [43] Huaixiu Steven Zheng, Swaroop Mishra, Xinyun Chen, Heng-Tze Cheng, Ed H Chi, Quoc V Le, and Denny Zhou. 2023. Take a step back: Evoking reasoning via abstraction in large language models. *arXiv preprint arXiv:2310.06117* (2023).
- [44] Yi Zheng, Björn Ross, and Walid Magdy. 2023. What makes good counterspeech? a comparison of generation approaches and evaluation metrics. In *Proceedings of*

the 1st Workshop on CounterSpeech for Online Abuse (CS4OA). 62–71.

- [45] Wanzheng Zhu and Suma Bhat. 2020. GRUEN for evaluating linguistic quality of generated text. *arXiv preprint arXiv:2010.02498* (2020).
- [46] 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*. Association for Computational Linguistics, 134–149. <https://doi.org/10.18653/v1/2021.findings-acl.12>
- [47] PG Zimbardo. 1969. The human choice, Individuation, reason, and order versus deindividuation, impulse, and chaos. In *Nebraska symposium on motivation/University of Nebraska Press*.

## A Appendix

### A.1 Dataset Description

The Multi-Target CONAN (MTCO) [10] dataset consists of hate speech and corresponding counter-narratives across multiple target groups. In Table 7, we show target group-wise examples of the dataset that includes disabled, Jews, LGBT+, Muslims, migrants, people of color (POC), and women.

### A.2 Examples

We provide examples from our generation process. Table 9 shows the overall pipeline, including the generated query, retrieved documents, critique model, document summary, and counter-narrative generation process. We also provide some examples of generated CNs using different LLMs as follows.

- **HS:** Allowing gays to donate blood compromises patient safety because they are having unprotected anal sex all the time.
- **Gemini:** This is a harmful and unfounded generalization. The FDA’s new blood donation guidelines focus on individual risk assessment, not blanket bans based on sexual orientation. Countries like Britain and Italy have successfully implemented individual risk assessment for all donors, demonstrating that allowing gay individuals to donate blood does not compromise safety. It’s important to remember that sexual behavior is diverse and not limited to any one group, and making assumptions about individuals based on stereotypes is harmful and inaccurate.
- **GPT-3.5:** This statement perpetuates harmful stereotypes about gay individuals and ignores the fact that blood donation policies are based on evidence-based risk assessments, not blanket assumptions. Implementing individual risk assessments for all donors, regardless of sexual orientation, is crucial in maintaining blood safety while promoting inclusivity.
- **GPT-4o:** The notion that allowing gay individuals to donate blood compromises safety is unfounded, as countries like Britain and Italy have successfully implemented individual risk assessments for all donors, ensuring safety without discrimination (source: The Economist). The FDA’s 2023 policy update in the U.S. also reflects this approach, focusing on individual risk rather than blanket deferrals, which aligns with advancements in HIV testing and reduces discrimination (source: JAMA Health Forum). Furthermore, the stereotype that gay men engage in unprotected sex "all the time" is misleading and harmful, as data shows a decrease in new HIV infections among gay and bisexual men from 2015 to 2019 (source: CDC). Evidence suggests that lifting bans on

gay blood donations could increase the blood supply significantly, potentially saving over a million lives annually, without compromising safety (source: The Williams Institute).

### A.3 Prompts

From query generation to counter narrative generation, we utilize LLMs to produce outputs. We experiment with multiple prompts, and in this section, we present the prompts that yield the best results for the corresponding tasks. The instructions are written in black, additional information or contexts highlighted in red. Number or sentence specifications are marked in blue.

#### Prompt for Query Generation

You are responding with counter speech to the hate speech {hatespeech}.

Generate 3 precise queries you would research in order to generate counterspeech for this hate speech.

#### Prompt for Summary Generation

Summarize this document in 2-3 sentences in a way that answers the query and retain information relevant for countering the hate speech. Keep any relevant statistics, and sources.

Query: {query}

Hate speech: {hs}

Document: {document}

#### Prompt for Critique Model

You are a grader assessing the relevance of a retrieved document for use in constructing counterspeech to hate speech.

Here is the retrieved document: {document}

Here is the hate speech: {hs}

Here is the query used to retrieve that document: {question}

If the document is generally relevant to countering the hate speech, grade it as relevant. Grade it as irrelevant if it is generally irrelevant to countering the hate speech, or if it seems to agree with the hate speech.

#### Prompt for CN Generation

Generate counterspeech to this hate speech, utilizing the following information fetched from various sources. Utilize the information to build your narrative, and cite sources when possible.

Context: {documents}

Hate speech: {hs}

Keep your counter speech to 4 sentences.

**Table 7: Target group-wise examples from Multi-Target CONAN (MTCO) dataset**

Target	HS	CN
<b>DISABLED</b>	Downies may hinder the productivity of other workers because they are not able to interact in a normal way. They should be isolated.	There are various levels of Down syndrome. Some are suitable for working with other people in "normal workplaces".
<b>JEWS</b>	Global Jewry did not want Brexit and are doing everything they can to sabotage it.	What makes you think that? Do you have any facts to back this up?
<b>LGBT+</b>	Why don't gay people seek a cure?	Have you ever talked to a homosexual person? It could help you understand many things, including that sexual orientation is fundamental to a person's identity, not a matter of whim or ill health.
<b>MIGRANTS</b>	Immigrants just don't want to integrate. Let's kick them out!	Immigrants are integrating very well into our society and it seems that they have reached a better level of English than other people.
<b>MUSLIMS</b>	Apples and oranges are not the same and cultures have never been 'equal'. Muslims are just misogynist!	You can find outdated teaching in all major religions but it doesn't mean most people follow those parts, a similar example from Christianity: Ephesians 5:22, "Wives, submit to you husbands as to the Lord"
<b>POC</b>	I hate negros: they killed a cop!	So are you blaming all of them? Lets say you are in a group of 10 people, 1 is a killer, does that make you ALL killers? Are you serious?
<b>WOMEN</b>	A woman's place is with her husband or his family.	There are many cases where the wife works and provides for the family, making a valuable contribution. It is a source of pride for modern men to be married to these women.
<b>OTHER</b>	Gypsy rape our women. They are criminals.	Forgive me but I don't understand how rape can be a Gypsy characteristic. Rather it is the product of toxic misogynistic culture.

**Table 8: P-values of paired t-test results for different models. The low values suggest the differences are statistically significant.**

Model	Factuality	Persuasiveness	Informativeness	Toxicity	Linguistic Quality	BLEU	BERTScore
<b>MTCO [10]</b>	$2.16 \times 10^{-162}$	$8.88 \times 10^{-217}$	$7.44 \times 10^{-158}$	$5.29 \times 10^{-31}$	$1.54 \times 10^{-32}$	–	–
<b>TKGCN [7]</b>	$8.38 \times 10^{-26}$	$6.22 \times 10^{-27}$	$1.43 \times 10^{-20}$	$1.24 \times 10^{-21}$	$3.43 \times 10^{-17}$	$2.84 \times 10^{-6}$	$1.69 \times 10^{-22}$

#### A.4 Statistical Significance

In Table 8, we present a paired t-test of the metrics for the results in Table 2 of our baselines and our method. The results here demonstrate that our results are statistically significant ( $p < 0.05$ ). We use

a paired t-test as it allows us to account for variability within the pairs and determine if the observed differences in the metrics are statistically meaningful.

**Table 9: Example of Generated CN along with all the steps mentioned in our pipeline. As the retrieved documents are lengthy, few lines of the documents are shown. D1N, D1W denotes documents from Newsroom and web search, and S1N, S1N denotes summaries of the documents D1N, and D1W, respectively.**

<b>HS</b>	Immigrants are overrunning our country, and are here illegally	
<b>Generated Query</b>	<p><b>Q1:</b> What are the economic contributions of immigrants to the country, including both documented and undocumented immigrants?</p> <p><b>Q2:</b> What are the legal pathways and challenges for immigrants seeking to enter and reside in the country?</p> <p><b>Q3:</b> What are the current statistics on immigration rates compared to historical data to assess the claim of overrunning?</p>	
<b>Retrieved Documents</b>	<b>Docu-</b>	<p><b>D1N:</b> The consensus of the economic analysis is that the taxes paid by immigrants and their descendants exceed the benefits they receive. Even with undocumented immigrants, the Internal Revenue Service estimates that they paid almost \$50 billion in federal taxes between 1996 to 2003, including payroll and Social Security...</p> <p><b>D1W:</b> Immigrants make significant contributions to our economy on virtually every front - including on tax revenue, where they contribute \$458.7 billion to state, local, and federal taxes in 2018. If the U.S. provided a pathway to permanent residency and citizenship, it would increase Americans' income by \$791 billion and generate \$184 billion in additional state and local and federal tax revenue from currently undocumented immigrants, and add more than 200,000 jobs to the U.S. economy per year...</p> <p><b>D2N:</b> The immigration bill being drafted by Congress has bipartisan support on three broad concepts - a pathway to citizenship for the undocumented, streamlining legal immigration and more stringent enforcement of the laws against hiring illegal workers...</p> <p><b>D2W:</b> CAP analysis shows that practices and policies that facilitate legal pathways can reduce irregular migration and encourage prospective migrants to seek entry into the United States through such ...</p>
<b>Document Summary</b>		<p><b>S1N:</b> Immigrants, both documented and undocumented, make significant economic contributions to the U.S. economy. The Internal Revenue Service estimates that undocumented immigrants paid nearly \$50 billion in federal taxes from 1996 to 2003, and their purchasing power supports hundreds of thousands of U.S. jobs. Despite some working off the books, the overall economic impact of immigrants is positive, as highlighted by the Council on Foreign Relations, which warns that failing to implement effective immigration policies could weaken the U.S. economy and national security.</p> <p><b>S1W:</b> Immigrants, both documented and undocumented, play a vital role in the U.S. economy, contributing \$458.7 billion in taxes in 2018 alone. Providing a pathway to permanent residency and citizenship for undocumented immigrants could further boost the economy by increasing Americans' income by \$791 billion, generating an additional \$184 billion in tax revenue, and creating over 200,000 jobs annually. These contributions highlight the positive economic impact of immigrants, countering the narrative that they are overrunning the country.</p> <p><b>S2N:</b> The immigration bill being drafted by Congress aims to address three main areas: creating a pathway to citizenship for undocumented immigrants, streamlining the legal immigration process, and enforcing laws against hiring illegal workers. However, the legal immigration system is complex, with multiple pathways and categories, each with limited visas and specific criteria, leading to significant backlogs. For instance, a Migration Policy Institute study noted that as of November 2012, over 4.4 million people with approved visa petitions were awaiting processing, primarily from family-based categories, and it could take 19 years to clear these backlogs without additional visas, highlighting the challenges in reforming the system to accommodate both current and new applicants.</p> <p><b>S2W:</b> The CAP analysis highlights that establishing and promoting legal pathways for immigration can significantly reduce irregular migration by encouraging prospective migrants to enter the United States through official channels. Legal pathways include options such as family reunification, employment-based visas, and humanitarian programs, which provide structured and lawful means for immigrants to reside in the country. By expanding and streamlining these processes, the U.S. can effectively manage immigration and counter the misconception that immigrants are predominantly entering the country illegally.</p>
<b>Critique Model</b>		<b>S1N:</b> relevant, <b>S1W:</b> relevant, <b>S2N:</b> relevant, <b>S2W:</b> relevant
<b>Counter (CN)</b>	<b>Narrative</b>	The notion that immigrants are "overrunning" the country is not supported by data, as unauthorized immigrants make up only 28% of the foreign-born population, and states like California, Illinois, and New York have seen declines in their unauthorized populations (Pew report). Immigrants, both documented and undocumented, contribute significantly to the U.S. economy, paying \$458.7 billion in taxes in 2018 alone, and their presence is projected to increase Americans' income by \$791 billion if given a pathway to citizenship (PDF Immigration Facts). Furthermore, the U.S. net immigration rates show only a modest increase, with 9.27 million net immigrants projected from 2022 to 2024, which is comparable to previous decades (Visual Capitalist). By expanding legal pathways and reforming immigration policies, the U.S. can effectively manage immigration and harness the positive economic impact immigrants bring (CAP Analysis).