Counter with Evidence! A Multi-Agent Memory Efficient Reasoning Framework for Hate Category Informed Counterspeech Generation

Sujoy Nath* **Aswini Kumar*** **Tanmoy Chakraborty**
Indian Institute of Technology Delhi, India
{sujoynathofficial, aswinikumarpadhi1995}@gmail.com
tanchak@iitd.ac.in

Abstract

Counterspeech effectively neutralizes the impact of online hate. Although prior work explores automated counterspeech generation, it largely emphasizes stylistic control while treating hate speech as homogeneous, overlooking that distinct forms of abuse require fundamentally different counterspeech strategies. To address this gap, we introduce FIRE (**F**actuality **I**nformed Multi-Agent **RE**asoning Framework) that first decomposes hate speech into one of the five distinct categories (misinformation, stereotype, conspiracy, dehumanizing, non-factual), and then maps it to a targeted counterspeech style. To facilitate FIRE, we curate FactualCS, a novel dataset of 4,784 instances that provides the annotations regarding hate categories, reasoning traces, and evidence mappings, which are critical elements for grounded generation that are missing in prior work. A comprehensive evaluation across 28 baseline configurations demonstrates that FIRE significantly surpasses existing methods, despite using compact agents (<2B). FIRE achieves a $\sim 12\%$ and $\sim 11\%$ improvement in factual and category-specific accuracy, respectively, while simultaneously reducing toxicity by $\sim 11\%$ relative to the strongest baselines. Further human evaluation confirms that responses generated by FIRE are significantly preferred over the strongest baselines, underscoring its effectiveness for real-world deployment. These findings show that decomposing the underlying intent of hate speech is essential for generating safe, effective, and contextually precise counterspeech.[1] [2]

1 Introduction

Social media platforms serve as the primary infrastructure for global communication (Bonaldi et al., 2024; Gennaro et al., 2025). However, this connectivity is frequently weaponized to target vulnerable communities with Hate Speech (HS), causing severe psychological distress (Awal et al., 2021; Davidson et al., 2017). To address online hate, counterspeech (CS) comes out as an efficient approach, which aims to neutralize hostility through polite, factually grounded engagement rather than suppression (Benesch, 2016; Wright et al., 2017).

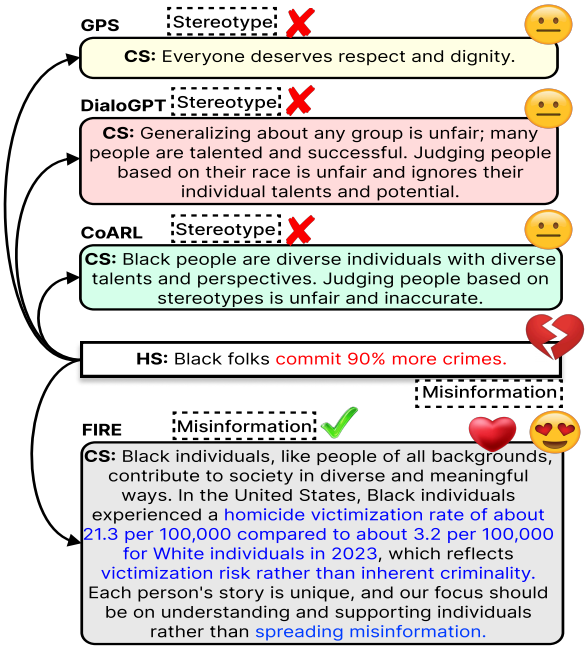


Figure 1: A qualitative comparison demonstrating the efficacy of FIRE against baselines; our framework generates superior quality response by accurately identifying the specific characteristics of the target hate speech.

Given the sheer volume of online content, manual intervention is unscalable, making automated CS generation a critical research trajectory (Gupta et al., 2023). Early approaches treated this as a text-to-text translation task, mapping hateful inputs to safe replies (Chung et al., 2019; Mathew et al., 2019). Although subsequent research has moved beyond simple translation to control stylis-

* Equal contribution.

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

[2] Our codebase is available at: https://github.com/C0mRD/Counter_with_evidence

Proceedings of the 2026 Conference on Empirical Methods in Natural Language Processing, pages 37435–37458
October 24-29, 2026 ©2026 Association for Computational Linguistics

tic attributes of counterspeech like politeness or emotion (Saha et al., 2022), these methods typically homogenize hate speech. By treating diverse abusive content as a uniform entity, they miss the critical requirement that different types of hate demand distinct logical refutations (Baider, 2023).

Motivation and Crisis. The primary limitation of current methodologies is their failure to account for the semantic diversity of hate speech, which is not uniform. Online hate manifests through distinct logical structures such as misinformation (verifiable false claims) (Wardle, 2024), conspiracies (allegations of secret plots) (Jolley et al., 2020), stereotypes (oversimplified group generalizations) (Vargas et al., 2023), dehumanization (denial of human dignity or animalistic metaphors) (Mendelsohn et al., 2020), and non-factual assertions (subjective expressions of disgust or opinion) (Wardle, 2024). Addressing these diverse forms of hate speech requires specific counter-strategies targeted to address that particular type of hate (Buerger, 2021). For instance, a verifiable false claim necessitates factual evidence, whereas a dehumanizing slur requires moral reframing. Current models, by failing to distinguish between these categories, often produce a generalized responses that fail to address the core premise of the abuse.

Need of an Agentic Framework. However, implementation of such specific strategies reveals the architectural limitations of monolithic Large Language Models (LLMs). A single LLM often struggles to simultaneously perform deep semantic reasoning (e.g., distinguishing *misinformation* from a *stereotype*) and factual verification while maintaining stylistic coherence (Liu et al., 2023). Without explicit task decomposition, these models lack the capacity to plan their logical strategy, frequently hallucinating facts or reverting to safe, context-agnostic responses (Hengle et al., 2024). To address this, we move to an *agentic* paradigm. By employing specialized agents, we decouple the problem, the system first diagnoses the specific nature of the abuse before attempting to generate an appropriate counter-response. This mimics human cognitive decomposition, ensuring that responses are planned and factually grounded before they are articulated (Cima et al., 2025). Crucially, training agents to reason rather than just mimic requires data that goes beyond the simple input-output pairs found in existing datasets (Saha et al., 2024; Chung et al., 2019). These resources lack the intermediate *reasoning traces*, the explicit mapping between hate categories, the search query required to debunk them, and the retrieved evidence, necessary to supervise an agent's logic.

Our Major Contributions. To bridge these gaps, we introduce the Factuality Informed Multi-Agent Reasoning Framework (`FIRE`), which prioritizes the semantic decomposition of abuse. `FIRE` employs specialized agents (<2B), to decouple the reasoning phase from the generation phase. This allows the system to explicitly identify one of five hate categories (Misinformation, Stereotype, Conspiracy, Dehumanization, Non-factual). This modularity ensures the response strategy matches the specific abuse type. To support `FIRE`, we develop `FactualCS`, a dataset comprising 4,784 hate-counterspeech pairs across 14 target communities. Unlike prior resources, `FactualCS` includes dense annotations for hate categories, reasoning traces, and evidence mappings. This solves the data scarcity issue for grounded counterspeech, enabling models to learn *why* a response is effective, not just *what* to say. As demonstrated in Figure 1, `FIRE` outperforms baselines like GPS (Zhu and Bhat, 2021), DialoGPT (Zhang et al., 2020), and CoARL (Hengle et al., 2024), which often produce context-agnostic text. Extensive evaluation shows that `FIRE` achieves state-of-the-art performance, matching larger models like LLaMA-3.1-8B while significantly improving factuality and category-specific accuracy.

2 Related Work

The study of counterspeech generation has evolved through a synergy of dataset curation and architectural advancements. Early efforts relied on expert-crafted, static corpora like CONAN (Chung et al., 2019) and Multi-Target CONAN (Fanton et al., 2021), which provided high-quality but limited coverage. Recent benchmarks such as CrowdCounter (Saha et al., 2024), IntentCONANv2 (Hengle et al., 2024), and MultiCONAN (Padhi et al., 2025) introduce annotations for intent and strategy that support the modeling of diverse response strategies. Methodologically, counterspeech generation has progressed from basic sequence-to-sequence translation tasks (Qian et al., 2019; Mathew et al., 2019) toward frameworks prioritizing strategic control. Recognizing that a single generic response is often insufficient, researchers developed pipelines like GPS (Zhu and Bhat, 2021) and CounterGeDi (Saha

et al., 2022), which steer generation toward specific tones such as politeness or detoxification. Moving beyond mere text fluency, recent strategies actively dismantle biases rather than generically denouncing hate (Gupta et al., 2023; Fraser et al., 2023; Mun et al., 2023). However, monolithic LLMs' struggles with reasoning and hallucination have catalyzed a shift toward Agentic AI (Acharya et al., 2025) to better preserve human-like authenticity and empathy (Mun et al., 2024). Unlike isolated prompting, multi-agent frameworks employ iterative reasoning and task decomposition to handle the evolving complexity of hate speech (Chen et al., 2025). By facilitating self-correction and integrating external knowledge via Retrieval Augmented Generation (RAG) (Lewis et al., 2021), these systems significantly reduce hallucinations and ensure interventions are factually grounded (Anik et al., 2025; Podolak et al., 2024). Our work builds upon this frontier, leveraging agentic decomposition to enable complex, evidence-based social strategies that are difficult to achieve with single-model architectures (Song et al., 2025).

Target	SH	MH	CH	DH	NH	Total
Women	178	12	3	166	285	644
Men	29	3	1	14	15	62
LGBT+	117	34	2	110	137	400
Muslim	271	101	22	90	99	583
Jews	167	35	96	109	36	443
IMGT	278	275	15	103	166	837
PoC	232	17	5	172	55	481
AP	15	1	0	5	1	22
Disable	41	1	0	73	5	120
Individual	55	30	26	57	84	252
Political	56	16	13	34	41	160
Occupation	6	5	1	7	8	27
OR	10	7	0	12	9	38
Others	115	44	28	172	356	715
Total	**1570**	**581**	**212**	**1124**	**1297**	**4784**
Train	1269	471	181	919	1072	3912
Val	123	50	15	87	108	383
Test	178	60	16	118	117	489

Table 1: Distribution of Hate Speech Types Across Targeted Groups: Stereotype Hate (SH), Misinformation Hate (MH), Conspiracy Hate (CH), Dehumanization Hate (DH), Non-factual Hate (NH) [3]

3 The `FactualCS` Dataset

To be truly effective, counterspeech must dismantle the core premises of the hate speech it addresses. Since abusive content varies significantly, ranging from false claims to attacks on human dignity, the rebuttal must be equally specialized. A response designed to counter a stereotype, for instance, may be ineffective against a conspiracy theory. Unfortunately, existing resources have largely overlooked this nuance, typically focusing on simple input-output pairings or the stylistic tone of the reply, while neglecting the specific semantic category of the hate itself. This limitation leaves models blind to the underlying nature of the abuse they are trying to neutralize. Driven by this gap, we develop `FactualCS` [4], a specialized dataset designed to ground responses in the specific characteristics of the hate speech.

Six annotators (4 male, 2 female) voluntarily participated in the annotation process, which followed a multi-stage protocol designed to establish a shared understanding of the task before full-scale annotation. Prior to annotation, the annotators reviewed relevant resources on online harassment and counterspeech and discussed the hate-speech taxonomy and corresponding annotation criteria. They then completed a 500-instance pilot phase, independently annotating hate-speech type, annotation rationale, target group, retrieval query, and supporting evidence, where applicable. Following the independent annotation, cases exhibiting disagreement were collaboratively examined to clarify challenging category boundaries, particularly between stereotypes and misinformation, dehumanizing language and other forms of non-factual hate, and cases requiring empirical verification. Insights from the pilot phase were subsequently formalized into a common set of operational guidelines specifying category definitions, evidence requirements, and decision rules for ambiguous or overlapping cases. These guidelines were then followed during the main annotation phase for the remaining instances, with hate-speech type, rationale, target group, query formulation, and evidence collection manually annotated. Further details of the complete annotation procedure, category-specific criteria, and decision rules are provided in Appendix A.4. The main annotation phase achieved a mean pairwise Cohen's κ of approximately 0.915 across the six annotators for 5-way categorization (Table 7). `FactualCS` comprises $4,784$ unique instances, each explicitly annotated with one of five functionally distinct categories: misinforma-

[3] OR: Other Religions, AP: Affected Person, PoC: Person of Color

[4] Dataset available at: https://huggingface.co/datasets/Aswini123/FactualCS

tion, stereotype, dehumanization, conspiracy theories, and non-factual claims (refer to Table 9 and Appendix A.5 for an analysis of their functional boundaries). By explicitly mapping these categories, we provide the necessary context for models to learn not only **what** to say, but also **why** a particular response is appropriate for a specific form of abuse. This structured approach moves beyond simple text generation, offering a roadmap for logical refutation. Further details regarding the dataset statistics and the annotation guidelines are provided in Table 1 and Appendices A.7, and A.4.

4 Proposed Methodology

In this section, we describe the design of our novel agentic framework FIRE for generating counterspeech. By concentrating on two crucial mechanisms, we designed FIRE to overcome known limitations: (i) ensuring categorical accuracy, where the model adapts its response to the particular type of abuse found in the input, and (ii) enforcing quality alignment, which directs the generation to match the constructive, non-toxic nature of human-written response. Figure 2 provides an overview of this entire workflow.

Problem Formulation

We define the task of counterspeech generation as a conditional text generation problem. Let $\mathcal{D} = \{d_1, d_2, ...dn\}$ represent our dataset, consisting of n samples in which each data point is denoted as a tuple $(h_i, t_i, g_i, r_i, q_i, e_i, c_i)$. In this framework, $h_i \in \mathcal{H}$ serves as the sole input source (hate speech). The goal is to generate the target counterspeech $c_i \in \mathcal{C}$. However, to ensure the response is grounded and logically sound, the model must also generate a set of latent auxiliary variables: the hate category $t_i \in \mathcal{T}$, the target group $g_i \in \mathcal{G}$, the reasoning trace $r_i \in \mathcal{R}$, the search query $q_i \in \mathcal{Q}$, and the supporting evidence $e_i \in \mathcal{E}$. We denote this auxiliary context collectively as $\mathcal{Z} = \mathcal{T} \times \mathcal{G} \times \mathcal{R} \times \mathcal{Q} \times \mathcal{E}$. Consequently, our objective is to learn a generation function $\psi : \mathcal{H} \to \mathcal{Z} \times \mathcal{C}$, which maximizes the joint probability of generating both the intermediate context and the final response, formulated as $(z_i, c_i) \sim \psi(\cdot|h_i)$.

We address this problem by decomposing counterspeech generation into two stages utilizing Qwen3-1.7B (Yang et al., 2025) in our agents, chosen for their superior reasoning capabilities (Joshi, 2025). In the first stage, the Hatespeech Analyst (<2b) leverages a memory module to identify the hate type, target, and reasoning, dynamically triggering a web search for evidence when factual claims are detected. In the second stage, the Counterspeech Generator (<2b) synthesizes the response with these outputs, including retrieved examples and evidence, to produce a response with a tone adapted to the specific hate category. Crucially, as the agents are not fine-tuned, FIRE relies on the memory module to derive specialization entirely from retrieved examples and structured guidance.

Memory Module

FIRE maintains a compact memory constructed from all annotated training instances. Each instance i contributes a record $m_i = (h_i, t_i, g_i, r_i, q_i, e_i, c_i)$, where only the hate speech h_i is used for learning representations. A lightweight encoder w_θ of size 22M maps h_i to an embedding $u_i = w_\theta(h_i)$, and the memory $M = (m_i, u_i)$ forms a fixed retrieval index. The encoder is trained to cluster instances sharing the same hate type while separating those from different types. Given a minibatch of embeddings u_i and corresponding HS type labels t_i, we compute the pairwise similarity matrix with temperature τ, $s_{ij} = cos(u_i, u_j)/\tau$. For each anchor i, positives are all other samples $j \neq i$ with matching hate type $t_j = t_i$. The supervised contrastive loss used to train the encoder is,

$$\mathcal{L} = -\frac{1}{|\mathcal{I}|} \sum_{i \in \mathcal{I}} \frac{1}{|\mathcal{P}(i)|} \sum_{p \in \mathcal{P}(i)} \log \frac{\exp(s_{ip})}{\sum_{a \in \mathcal{A}(i)} \exp(s_{ia})}$$

where $\mathcal{P}(i)$ is the set of positive examples for i, $\mathcal{A}(i)$ contains all other samples in the batch except i, and $\mathcal{I}$ is the set of samples that have at least one positive example in the batch. Although $|\mathcal{P}(i)|$ varies across batches, normalization by $|\mathcal{P}(i)|$ ensures uniform contributions. This objective encourages embeddings of the same hate type to remain close while pushing apart embeddings of different types.

At retrieval time, for a new input h, the encoder computes $u = w_\theta(h)$, and cosine similarities are evaluated against all memory entries. For each hate type $t \in \mathcal{T}$, we define the index set $\mathcal{I}_t = \{i : t_i = t\}$ and compute similarities $s_i = \frac{u^\top u_i}{\|u\|\|u_i\|}$ for all $i \in \{1, \ldots, |M|\}$. The top-k most similar indices within hate type t are obtained as $\text{TopK}_t(h) = \arg\max_{J \subset \mathcal{I}_t, |J|=k} \sum_{j \in J} s_j$, and we compute a type-level score by averaging these top-

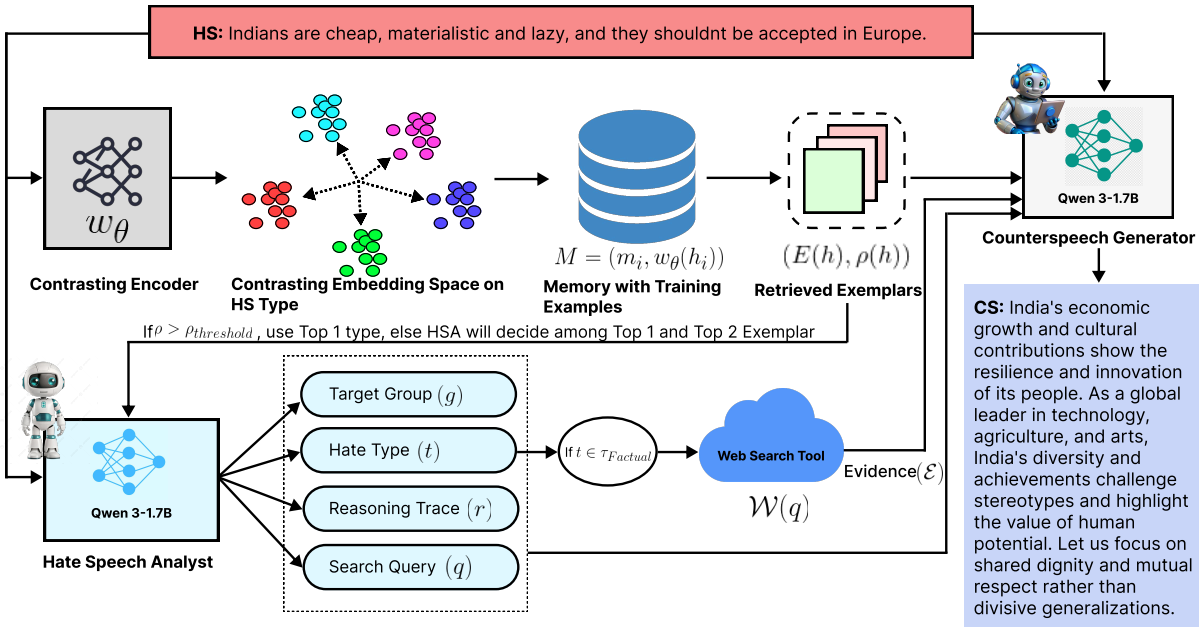


Figure 2: **An overview of the proposed FIRE framework that follows the two-phase approach.** The Memory Module utilizes a trained contrasting encoder (w_θ) to retrieve top-k exemplars (E) and compute a purity score (ρ). In Phase 1 (HEAL), the Hate Speech Analyst (HSA) leverages these exemplars to predict structured latent variables (g, t, r, q). If the detected hate type contains factual claims ($t \in \mathcal{T}_{\text{factual}}$), a Web Search Tool is triggered to retrieve supporting evidence ($\mathcal{E}$). In Phase 2 (CARE), the Counterspeech Generator (CSG) synthesizes the final response by conditioning on the original input, the structured analysis from HSA, the retrieved evidence, and the stylistic patterns from the memory exemplars.

k similarities: $\mathcal{S}_t(h) = \frac{1}{k}\sum_{i \in \text{TopK}_t(h)} s_i$. The predicted hate type is $\hat{t} = \arg\max_{t \in \mathcal{T}} \mathcal{S}_t(h)$, which determines the exemplars to retrieve: $E(h) = \{m_i : i \in \text{TopK}_{\hat{t}}(h)\}$. These retrieved entries provide reasoning patterns that guide both the Hate Speech Analyst and Counter Speech Generator. Importantly, the encoder w_θ is the only trained module in FIRE, preserving modularity and avoiding any fine-tuning of language models.

Phase 1: Hate Explanation with Analysis and Lookup (HEAL)

Hate Speech Analyst (HSA). Given hate speech h, retrieved exemplars $E(h)$, and a purity score $\rho(h)$ measuring retrieval confidence, the analyst generates the target group g, hate type t, reasoning trace r, and search query q according to the distribution $P_{\text{HSA}}(g, t, r, q \mid h, E, \rho)$. The purity score quantifies the homogeneity of the k-nearest neighbors in the memory. Let $\mathcal{N}_k(h) = \arg\max_{J \subset \{1,\ldots,|M|\}, |J|=k} \sum_{j \in J} s_j$ be the k-nearest neighbors of h. The purity is computed as:

$$\rho(h) = \frac{1}{k} \sum_{i \in \mathcal{N}_k(h)} \mathbb{F}[t_i = \hat{t}]$$

where $\mathbb{F}[\cdot]$ is the indicator function. When $\rho(h) \geq \rho_{\text{thresh}}$, the analyst performs maximum a posteriori estimation using exemplars from the single most confident hate type $\hat{t}$. When $\rho(h) < \rho_{\text{thresh}}$, indicating ambiguity, exemplars from both the top-two predicted hate types $\hat{t}_1$ and $\hat{t}_2$ are provided to the analyst, which must then explicitly choose between them and justify its decision. Formally, the structured output is sampled as $(g, t, r, q) \sim P_{\text{HSA}}(\cdot \mid h, E, \rho)$, where the prompt construction function encapsulates all contextual information. This design guides HSA through both exemplar-based reasoning and explicit uncertainty cues, significantly improving its capability despite its small size.

Websearch Tool. For hate types that can be better addressed with proper factual evidence, the search query q generated by the analyst triggers an external websearch tool. We define the set of factual hate types as $\mathcal{T}_{\text{factual}} = \{\text{misinformation}, \text{stereotype}, \text{conspiracy}\}$. The websearch function $\mathcal{W} : \mathcal{Q} \to \mathcal{E} \cup \{\emptyset\}$ takes a query q and returns evidence snippets. The evidence retrieval process is formalized as:

$$e = \begin{cases} \mathcal{W}(q) & \text{if } t \in \mathcal{T}_{\text{factual}} \\ \emptyset & \text{otherwise} \end{cases}$$

where $\mathcal{W}(q)$ executes the web query q and returns factual evidence snippets $e \in \mathcal{E}$. For all non-factual forms of hate, we set $e = \emptyset$. A manual audit of 100 retrieved instances confirms our websearch tool effectively suppresses misinformation, with 92% of evidence originating from institutional or mainstream informational sources (Appendix A.13).

Phase 2: Category-aware Agentic Response Engine (CARE)

Counterspeech Generator (CSG). The counterspeech generator produces the final counterspeech c conditioned on all structured information extracted by the HSA, the retrieved exemplars, and the evidence. Let $\mathcal{Z} = (g, t, r, E, e)$ denote the complete conditioning context. We model the generator with a conditional distribution:

$$P_{\text{CSG}}(c \mid h, \mathcal{Z}) = \prod_{i=1}^{|c|} P_\phi(c_i \mid c_{<i}, h, \mathcal{Z})$$

where $c = (c_1, \ldots, c_{|c|})$ is the counterspeech token sequence and $c_{<i}$ represents all preceding tokens. The contextual variables in $\mathcal{Z}$ collectively shape the generation: hate type t activates type-specific response strategies, exemplars E supply stylistic patterns from similar cases, and evidence e introduces factual grounding when required. As the model parameters remain frozen, all task specialization emerges from how the conditioning context $\mathcal{Z}$ modifies the generation distribution, enabling hate type-aware and evidence-based responses without finetuning. Due to resource and accessibility constraints, we focus on open-weight models for reproducible and cost-effective evaluation; however, our model-agnostic framework can be extended to proprietary models via API integration, with systematic comparison left for future work.

5 Experimental Setup

5.1 Baselines

To evaluate our approach, we compare against several baselines: GPS (Zhu and Bhat, 2021) (a three-stage pipeline), fine-tuned DialoGPT (Zhang et al., 2020) (for contextual coherence), CoARL (Hengle et al., 2024) (conversational strategy), and the state-of-the-art HiPPrO (Padhi et al., 2025) (multi-attribute prefix learning). Furthermore, we evaluate six foundational architectures, encoder-decoders (**BART**$_{Large}$ (Lewis et al., 2020), **FLAN-T5**$_{XL}$ (Chung et al., 2024a)) and decoder-only models (**GPT2**$_{XL}$, **Llama-3.1-Instruct**$_{8B}$ (Dubey et al., 2024), **Mistral Instruct**$_{7B}$ (Jiang et al., 2023), and **Qwen-2.5-Instruct**$_{7B}$ (Qwen et al., 2025)), under zero-shot (ZS) without task specific training, and few-shot (FS) utilizing five in-context exemplars. We also evaluate retrieval-based (RB) standard RAG baselines that condition generation on the top retrieved exemplar (including its category and evidence), providing single-pass models with equivalent information access to explicitly isolate the performance gains of our multi-agent decomposition (Appendix A.10). Furthermore, for supervised fine-tuning (SFT), we implement Low-Rank Adaptation (LoRA) (Hu et al., 2021) fine-tuning with rank 16 on all model architectures.

5.2 Evaluation Metrics

Evaluating counterspeech generation is challenging due to its context-dependent nature and multiple valid responses for a single HS instance (Chung et al., 2024b). Our evaluation framework employs multi-dimensional metrics across lexical, semantic, and counterspeech-specific dimensions. We measure lexical similarity using **METEOR (M)** (Banerjee and Lavie, 2005), which incorporates synonym matching, and also report **ROUGE** scores (Lin, 2004). Semantic relevance is captured through **BERTScore (BS)** (Zhang et al., 2019) using contextualized embeddings, and **Cosine Similarity (CoSim)** (Reimers and Gurevych, 2019) between sentence embeddings. We assess linguistic quality through **Repetition Rate (R)** (Cettolo et al., 2014) (measuring redundancy via unique token ratios), **Diversity (Div)** (ratio of unique to total tokens), and **Novelty (Nov)** (lexical divergence from reference text). **Toxicity (T)** is measured using Detoxify (Hanu and Team, 2020) to ensure respectful counterspeech. **Category Accuracy (CatAcc)** verifies both correct hate speech classification and proper counterspeech strategy execution. We require a strict three-way match: the dataset gold label, the model's predicted hate category, and the corresponding counter-strategy detected in the final response by an independent $RoBERTa_{Large}$ (Liu et al., 2019) classifier. **Factuality Score (FSc)** assesses factual correctness using BART-MNLI (Chen et al., 2018) to verify if generated claims are supported by the retrieved gold evidence. Crucially, to fairly measure inherent hallucination, non-retrieval baselines generate responses without access to FIRE's evidence, but all model outputs are evaluated against the same gold facts.

Model	P/A	ROUGE ↑			BS ↑	M ↑	R ↓	CoSIM ↑	T ↓	Nov ↑	Div ↑	CatAcc ↑	FSc ↑
		R1	R2	RL									
GPS	–	0.103	0.032	0.129	0.831	0.179	0.257	0.472	0.035	0.489	0.643	0.422	0.467
DialoGPT	–	0.191	0.087	0.188	0.863	0.210	0.231	0.518	0.021	0.618	0.765	0.563	0.529
CoARL	–	0.207	0.101	0.202	0.865	0.213	0.212	0.549	0.019	0.649	0.788	0.604	0.579
HiPPrO	–	0.235	0.104	0.217	0.879	0.227	0.198	0.574	0.018	0.651	0.809	0.622	0.605
BART Large	ZS	0.063	0.018	0.072	0.823	0.083	0.274	0.318	0.041	0.487	0.742	0.108	0.198
GPT2 XL	ZS	0.057	0.016	0.028	0.788	0.064	0.346	0.147	0.061	0.414	0.687	0.138	0.182
Flan-t5-xl	ZS	0.104	0.042	0.123	0.842	0.092	0.247	0.382	0.029	0.524	0.731	0.273	0.332
Llama-3.1-8B-instruct	ZS	0.269	0.060	0.177	0.876	0.216	0.163	0.528	0.022	0.646	0.837	0.219	0.455
Mistral-7B-instruct	ZS	0.277	0.068	0.180	0.877	0.206	0.145	0.521	0.021	0.636	0.855	0.204	0.472
Qwen-2.5-7B-instruct	ZS	0.263	0.058	0.173	0.876	0.181	0.151	0.545	0.056	0.659	0.849	0.189	0.626
BART Large	FS	0.141	0.029	0.102	0.835	0.123	0.243	0.392	0.032	0.529	0.763	0.174	0.262
GPT2 XL	FS	0.072	0.021	0.039	0.823	0.083	0.308	0.251	0.049	0.462	0.723	0.204	0.254
Flan-t5-xl	FS	0.173	0.061	0.182	0.859	0.142	0.219	0.474	0.024	0.561	0.764	0.346	0.413
Llama-3.1-8B-instruct	FS	0.265	0.058	0.186	0.879	0.173	0.150	0.535	0.022	0.667	0.860	0.324	0.520
Mistral-7B-instruct	FS	0.268	0.067	0.191	0.871	0.179	0.131	0.521	0.038	0.662	0.859	0.356	0.548
Qwen-2.5-7B-instruct	FS	0.255	0.056	0.183	0.879	0.170	0.165	0.535	0.061	0.641	0.835	0.338	0.681
BART Large	RB	0.193	0.091	0.164	0.842	0.158	0.217	0.448	0.026	0.563	0.804	0.237	0.303
GPT2 XL	RB	0.079	0.028	0.044	0.841	0.093	0.282	0.314	0.041	0.503	0.742	0.259	0.321
Flan-t5-xl	RB	0.234	0.073	0.193	0.862	0.183	0.210	0.521	0.019	0.603	0.773	0.382	0.472
Llama-3.1-8B-instruct	RB	0.302	0.078	0.203	0.873	0.230	0.133	0.538	0.019	0.644	0.847	0.366	0.612
Mistral-7B-instruct	RB	0.298	0.080	0.202	0.872	0.221	0.133	0.523	0.020	0.654	0.857	0.413	0.709
Qwen-2.5-7B-instruct	RB	0.264	0.062	0.182	0.878	0.190	0.128	0.548	0.038	0.655	0.849	0.379	0.715
BART Large	SFT	0.236	0.081	0.175	0.867	0.192	0.183	0.419	0.021	0.558	0.817	0.227	0.227
GPT2 XL	SFT	0.080	0.030	0.059	0.825	0.067	0.135	0.130	0.045	0.517	0.825	0.260	0.260
Flan-t5-xl	SFT	0.304	0.104	0.235	0.876	0.238	0.203	0.548	0.018	0.669	0.797	0.508	0.508
Llama-3.1-8B-instruct	SFT	**0.316**	**0.109**	**0.239**	0.880	0.239	0.130	0.558	0.018	0.650	0.860	0.623	0.623
Mistral-7B-instruct	SFT	0.313	0.107	0.231	0.879	0.237	0.129	0.542	0.020	0.651	0.861	0.620	0.763
Qwen-2.5-7B-instruct	SFT	0.307	0.104	0.230	0.878	0.224	0.127	0.556	0.019	0.665	0.862	0.632	0.810
FIRE-(2x1.7B) (Ours)	–	0.294	0.071	0.197	**0.886**	**0.247**	**0.118**	**0.574**	**0.016**	**0.730**	**0.873**	**0.702**	**0.969**
- without Websearch Tool	–	0.277	0.060	0.179	0.879	0.224	0.133	0.535	0.017	0.708	0.866	0.664	0.832
- without Memory	–	0.262	0.043	0.163	0.874	0.214	0.152	0.490	0.018	0.692	0.847	0.599	0.795
- without HSA	–	0.231	0.034	0.148	0.868	0.189	0.162	0.410	0.020	0.656	0.837	0.507	0.62
$\Delta_{FIRE(Ours)-BestBaseline}$		↓ 0.022	↓ 0.038	↓ 0.042	↑ 0.006	↑ 0.008	↓ 0.009	↑ 0.016	↓ 0.002	↑ 0.061	↑ 0.011	↑ 0.07	↑ 0.159

Table 2: Comparing FIRE with baselines across various evaluation metrics. Here, ↑ (*resp.* ↓) denotes that higher (*resp.* lower) is better. **Bold** (*resp.* underline) indicates the best (*resp.* second-ranked) performance. Here P/A represents different strategies used: ZS for Zero Shot Prompting, FS for Few Shot Prompting, RB for Retrieval-Based and SFT represent the Supervised Finetuning.

6 Experimental Results

This section presents a comprehensive empirical analysis that evaluates the effectiveness of FIRE in comparison to existing state-of-the-art techniques.

6.1 Quantitative Results

Table 2 presents the quantitative evaluation across all metrics, where FIRE achieves the best performance across nine of thirteen categories.

Semantic and Lexical Alignment. FIRE demonstrates superior semantic understanding and contextual relevance while maintaining effective surface-level matching. It leads in all deep semantic metrics, achieving the highest BERTScore (0.886), METEOR (0.247), and CoSIM (0.574), which indicates strong semantic alignment and lexical matching. These gains are supported by competitive ROUGE scores (R1: 0.294, R2: 0.071, RL: 0.197), confirming that the model retains key content while improving semantic depth.

Response Quality and Safety. FIRE produces content that is significantly safer, more natural, and linguistically diverse than baselines. This is evidenced by a marked improvement in safety, with Toxicity scores dropping to 0.016 (11.1% lower). The responses are also less repetitive and more varied, indicated by a Repetition Rate of 0.118 (7.1% lower than the best baseline), alongside increases in Novelty (0.730, 9.1% higher) and Diversity (0.873,

Metric	T-statistic	p-value	Significant	Outperform
FIRE vs LLaMA-3.1-8B-Instruct (SFT)				
ROUGE-1	-9.41	$2.05E{-}19$	Yes	No
ROUGE-2	-9.75	$1.28E{-}20$	Yes	No
ROUGE-L	-11.19	$5.63E{-}26$	Yes	No
BERTScore	-11.33	$1.48E{-}26$	Yes	No
METEOR	-7.35	$8.80E{-}13$	Yes	No
RepetitionRate	0.35	0.727	No	Yes
CoSIM	-4.75	$2.74E{-}06$	Yes	No
Toxicity	0.36	0.717	No	Yes
Novelty	12.72	$3.84E{-}32$	Yes	Yes
Diversity	-0.35	0.727	No	Yes
CatAcc	6.84	$1.10E{-}11$	Yes	Yes
FactualScore	14.92	$2.10E{-}34$	Yes	Yes
FIRE vs LLaMA-3.1-8B-Instruct (FS)				
ROUGE-1	4.31	$1.99E{-}05$	Yes	Yes
ROUGE-2	2.53	0.012	Yes	Yes
ROUGE-L	0.97	0.331	No	Yes
BERTScore	3.14	$1.80E{-}03$	Yes	Yes
METEOR	11.33	$1.53E{-}26$	Yes	Yes
RepetitionRate	11.78	$2.57E{-}28$	Yes	Yes
CoSIM	2.91	$3.90E{-}03$	Yes	Yes
Toxicity	-1.45	0.149	No	Yes
Novelty	6.45	$2.69E{-}10$	Yes	Yes
Diversity	11.78	$2.57E{-}28$	Yes	Yes
CatAcc	9.27	$4.80E{-}20$	Yes	Yes
FactualScore	18.63	$7.40E{-}41$	Yes	Yes

Table 3: Statistical comparison of FIRE against LLaMA-3.1-8B-Instruct under supervised fine-tuning (SFT) and few-shot (FS) settings using paired t-tests. Positive T-statistic values indicate better performance by FIRE. Statistical significance is determined at $p < 0.05$.

1.3% higher), reflecting a richer vocabulary.

Strategic and Factual Integrity. We observe a substantial boost in Category Accuracy (0.702, 11.1% higher), reflecting better strategy alignment. Most significantly, the model demonstrates superior reliability with a Factual Score of 0.969 (12.2% higher), validating its ability to maintain factual accuracy without sacrificing coherence.

Overall Performance. Despite using smaller agents (<2B), FIRE proves highly effective against baselines; its sequential activation of two 1.7B agents reduces peak VRAM by 75% (∼4GB vs. ∼16GB for monolithic 8B models) while maintaining comparable end-to-end inference latency (Appendix A.8). Against SFT models, which outperform ZS, FS, and RB approaches, FIRE achieves superior performance across key metrics, demonstrating effectiveness in generating factually accurate, and contextually appropriate responses.

6.2 Statistical Significance Testing

To determine whether the observed performance differences between FIRE and best performing baseline model LLAMA-3.1-8B-INSTRUCT are statistically meaningful, we conduct paired two-tailed t-tests across all evaluation metrics. Tests are performed separately for the supervised fine-tuning (SFT) and few-shot (FS) settings, comparing model outputs on identical test instances. Statistical significance is assessed at a threshold of $p < 0.05$.

As shown in Table 3, the SFT comparison reveals that FIRE achieves statistically significant improvements in factuality-oriented and task-specific metrics, including Novelty ($t = 12.72, p < 10^{-31}$), Category Accuracy ($t = 6.84, p < 10^{-11}$), and FactualScore ($t = 14.92, p < 10^{-33}$). In contrast, LLAMA-3.1-8B-INSTRUCT attains higher scores on surface-level overlap and semantic similarity metrics such as ROUGE and BERTScore. Differences in Repetition Rate and Toxicity are not statistically significant, indicating comparable fluency and safety characteristics under supervised fine-tuning.

In the FS setting, FIRE consistently and significantly outperforms the baseline across nearly all metrics. Notable gains are observed in both generation quality and factual effectiveness, including ROUGE-1, ROUGE-2, METEOR, BERTScore, Category Accuracy ($t = 9.27, p < 10^{-19}$), and FactualScore ($t = 18.63, p < 10^{-40}$). Improvements in Novelty, Diversity, and Repetition Rate further suggest that FIRE produces more informative and less repetitive responses in the few-shot regime. Differences in Toxicity are not statistically significant, indicating that performance gains do not come at the expense of increased harmfulness.

The statistical analysis confirms that the observed improvements are robust and unlikely to be due to random variation.

6.3 Ablation Study

Our ablation study assesses the effects of each component in FIRE (Table 2). To understand the importance of websearch tool in our framework we remove it, which results in a significant drop in Factual Score (FSc) from 0.969 to 0.832 (14.1% decrease), demonstrating that availability of supporting information is critical for maintaining factual accuracy in counterspeech. Additionally, Category Accuracy drops from 0.702 to 0.664 (5.4% decrease). When the memory is removed, the category accuracy falls to 0.599 (14.5% below full model), while CoSIM drops from 0.574 to 0.490 (14.6% decrease), indicating reduced contextual coherence. BERTScore decreases from 0.886 to 0.874, and Repetition Rate increases from 0.118

Models on comparison	Metrics			
	ICS ↑	Ad ↑	CoRl ↑	ArgE ↑
FIRE vs Llama-3.1-8b (SFT)	0.91	0.89	0.86	0.93
FIRE vs HiPPrO	0.96	0.94	0.95	0.97
FIRE vs CoARL	0.98	0.96	0.96	0.98

Table 4: Results of the human evaluation study, where responses generated by FIRE are shown against those produced by (a) Llama-3.1-8b-instruct, (b) HiPPrO, and (c) CoARL. The results are reported in terms of Win Rate %, indicating the % of instances where FIRE outperforms the respective baselines.

to 0.152, suggesting less diverse and more repetitive outputs. When the HSA is removed, simulating a single-pass scenario with equivalent tool and memory access, FSc and CatAcc drop by 36.0% and 27.7%. This proves multi-agent decomposition is functionally necessary to manage cognitive load. Overall, each FIRE component is crucial, web search drives factuality, memory ensures coherence, and the HSA enables targeted strategy. Furthermore, Appendix A.9's purity score distribution proves FIRE robustly mitigates HSA error cascading by resolving ambiguous classifications before generation.

6.4 Human Evaluation

Previous studies (Jones et al., 2024; Hengle et al., 2025) highlight that automatic metrics often fail to capture the real-world effectiveness of counterspeech, motivating the need for human evaluation. We therefore conducted a dedicated human study comparing responses generated by FIRE against strong baselines (LLaMA-3.1-8B-Instruct (SFT), HiPPrO, CoARL). We followed (Hengle et al., 2024) for our human evaluation. Evaluation was performed on a randomly sampled (seed = 2) subset of model outputs, with balanced coverage across hate categories. A total of 30 expert annotators (aged 20–35, balanced gender) with relevant backgrounds ranked paired responses from 1 (Best) to 4 (Worst) across four criteria: Independent Counterspeech (ICS), Adequacy (Ad), Contextual Relevance (CoRl), and Argumentative Effectiveness (ArgE) (Appendix A.14). Results are reported using win rates, representing the fraction of instances where FIRE was preferred over the baseline. As shown in Table 4, FIRE consistently outperforms all comparison models across every metric. Against the best baseline method, LLaMA-3.1-8B-Instruct SFT, FIRE achieves strong win rates across ICS (0.91), Ad (0.89), CoRl (0.86), and ArgE (0.93), indicating clear human preference for its fluency, contextual grounding, and argumentative strength.

Type	Content
Hate Speech	No muslims are compatible with any Western society.
Model	**Generated Counterspeech**
Llama-3.1-8b-instruct	Muslims are a diverse group and integration varies across countries. Many Muslims contribute positively to Western societies.
FIRE (Ours)	A Muslim leader like London Mayor Sadiq Khan has integrated into Western societies while upholding Islamic values, demonstrating that compatibility is possible through mutual respect and dialogue. Diversity enriches societies, and stereotypes ignore the complexity of individual experiences. Understanding Islamophobia requires recognizing that cultural and religious differences foster mutual growth, not division.

Table 5: Error analysis between the response generated by our method FIRE vs the best baseline.

7 Error Analysis

Table 5 shows the qualitative disparity in counterspeech between best performing baseline Llama-3.1-8B-Instruct and FIRE. Responding to the claim 'No Muslims are compatible with any Western society', the baseline offers a hedged rebuttal; its statement that 'integration varies across countries' unintentionally validates the stereotype by implying conditional legitimacy. Conversely, FIRE directly invalidates the universal generalization through concrete counter-evidence ('London Mayor Sadiq Khan'). By reframing compatibility as 'mutual respect', FIRE moves beyond the baseline's passive neutrality to provide a constructive, factually grounded counterspeech.

8 Conclusion

In this work, we introduce FIRE, a hierarchical agentic framework, and FactualCS, a dataset targeting different hate speech categories. By decomposing generation into intent analysis and grounded response synthesis, FIRE enables categorically precise and factually supported interventions. Using compact (<2B) models without any finetuning, FIRE substantially outperforms established baselines and matches larger state-of-the-art LLMs, demonstrating that collaborative, reasoning-centric agents provide an efficient and effective approach to safe, context-aware counterspeech.

Limitation

Our research presents several limitations that warrant consideration. Firstly, while `FactualCS` targets five distinct categories of abuse, the dataset is not exhaustive, potentially limiting generalization to other forms of hate speech or intersectional attacks not covered within our 14 target communities. Secondly, our reliance on small language models (<2B), though efficient, may constrain the depth of reasoning compared to larger state-of-the-art systems, particularly when decoding highly ambiguous linguistic nuances. Additionally, the agentic workflow introduces a dependency on external tools; web search failures can propagate errors into the final response, compromising factual grounding. Further, We also acknowledge that the annotation team was relatively small and consisted primarily of researchers with technical expertise; although this expertise was valuable for applying the operational taxonomy and evidence-retrieval criteria, the annotators did not necessarily represent the lived experiences or cultural backgrounds of all target communities covered by FactualCS. Consequently, culturally specific, implicit, coded, or community-dependent expressions of hate may remain underrepresented or may be subject to interpretation bias. Furthermore, our current evaluation is restricted to English-language datasets, which may limit applicability to the diverse linguistic and cultural nuances of hate speech in other regions. Lastly, while we demonstrate high factual accuracy, our evaluation primarily focuses on immediate response quality rather than long-term conversational impact or the potential for conflict escalation. Future work could address these gaps by expanding the taxonomy of hate types, adapting our inherently language-agnostic architecture to diverse cultural contexts via multilingual models and localized search APIs, and incorporating longitudinal studies on user interactions.

Ethics Statement

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

Acknowledgment

Tanmoy Chakraborty acknowledges the support of the Anusandhan National Research Foundation (DST/INT/USA/NSF-DST/Tanmoy/P2/2024), Google GCP Grant, and Rajiv Khemani Young Faculty Chair Professorship in Artificial Intelligence.

References

Deepak Bhaskar Acharya, Karthigeyan Kuppan, and B Divya. 2025. Agentic ai: Autonomous intelligence for complex goals–a comprehensive survey. *IEEe Access*.

Anirban Saha Anik, Xiaoying Song, Elliott Wang, Bryan Wang, Bengisu Yarimbas, and Lingzi Hong. 2025. Multi-agent retrieval-augmented framework for evidence-based counterspeech against health misinformation. *arXiv preprint arXiv:2507.07307*.

Md. Rabiul Awal, Rui Cao, Roy Ka-Wei Lee, and Sandra Mitrovic. 2021. Angrybert: Joint learning target and emotion for hate speech detection. In *Advances in Knowledge Discovery and Data Mining - 25th Pacific-Asia Conference, PAKDD 2021, Virtual Event, May 11-14, 2021, Proceedings, Part I*, volume 12712 of *Lecture Notes in Computer Science*, pages 701–713. Springer.

Fabienne Baider. 2023. Accountability issues, online covert hate speech, and the efficacy of counter-speech. *Politics and Governance*, 11(2):249–260.

Satanjeev Banerjee and Alon Lavie. 2005. Meteor: An automatic metric for mt evaluation with improved correlation with human judgments. In *Proceedings of the acl workshop on intrinsic and extrinsic evaluation measures for machine translation and/or summarization*, pages 65–72.

Valerio Basile, Cristina Bosco, Elisabetta Fersini, Debora Nozza, Viviana Patti, Francisco Manuel Rangel Pardo, Paolo Rosso, and Manuela Sanguinetti. 2019. SemEval-2019 task 5: Multilingual detection of hate speech against immigrants and women in Twitter. In *Proceedings of the 13th International*

Workshop on Semantic Evaluation, pages 54–63, Minneapolis, Minnesota, USA. Association for Computational Linguistics.

Susan Benesch. 2016. Considerations for successful counterspeech. Dangerous Speech Project.

Helena Bonaldi, Yi-Ling Chung, Gavin Abercrombie, and Marco Guerini. 2024. NLP for counterspeech against hate: A survey and how-to guide. In *Findings of the Association for Computational Linguistics: NAACL 2024*, pages 3480–3499, Mexico City, Mexico. Association for Computational Linguistics.

Catherine Buerger. 2021. Counterspeech: A literature review. *Available at SSRN 4066882*.

Mauro Cettolo, Nicola Bertoldi, and Marcello Federico. 2014. The repetition rate of text as a predictor of the effectiveness of machine translation adaptation. In *Proceedings of the 11th Conference of the Association for Machine Translation in the Americas: MT Researchers Track*, pages 166–179.

Jinkun Chen, Sher Badshah, Xuemin Yu, and Sijia Han. 2025. Static sandboxes are inadequate: Modeling societal complexity requires open-ended co-evolution in llm-based multi-agent simulations. *arXiv preprint arXiv:2510.13982*.

Qian Chen, Xiaodan Zhu, Zhen-Hua Ling, Diana Inkpen, and Si Wei. 2018. Neural natural language inference models enhanced with external knowledge. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Association for Computational Linguistics.

Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Yunxuan Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, et al. 2024a. Scaling instruction-finetuned language models. *Journal of Machine Learning Research*, 25(70):1–53.

Yi-Ling Chung, Gavin Abercrombie, Florence Enock, Jonathan Bright, and Verena Rieser. 2024b. Understanding counterspeech for online harm mitigation. In *Northern European Journal of Language Technology, Volume 10*, pages 30–49.

Yi-Ling Chung, Elizaveta Kuzmenko, Serra Sinem Tekiroğlu, and Marco Guerini. 2019. Conan-counter narratives through nichesourcing: a multilingual dataset of responses to fight online hate speech. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2819–2829.

Lorenzo Cima, Alessio Miaschi, Amaury Trujillo, Marco Avvenuti, Felice Dell'Orletta, and Stefano Cresci. 2025. Contextualized counterspeech: Strategies for adaptation, personalization, and evaluation. In *Proceedings of the ACM on Web Conference 2025*, pages 5022–5033.

Thomas Davidson, Dana Warmsley, Michael Macy, and Ingmar Weber. 2017. Automated hate speech detection and the problem of offensive language. *arXiv preprint arXiv:1703.04009*.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, et al. 2024. The llama 3 herd of models. *arXiv e-prints*, pages arXiv–2407.

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. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 3226–3240.

Kathleen Fraser, Svetlana Kiritchenko, Isar Nejadgholi, and Anna Kerkhof. 2023. What makes a good counter-stereotype? evaluating strategies for automated responses to stereotypical text. In *Proceedings of the First Workshop on Social Influence in Conversations (SICon 2023)*, pages 25–38, Toronto, Canada. Association for Computational Linguistics.

Gloria Gennaro, Laurenz Derksen, Aya Abdelrahman, Emma Broggini, Mariya Alexandra Green, Victoria Andrea Haerter, Elia Heer, Isabel Heidler, Fiona Kauer, Han-Nuri Kim, et al. 2025. Counterspeech encouraging users to adopt the perspective of minority groups reduces hate speech and its amplification on social media. *Scientific Reports*, 15(1):22018.

Rishabh Gupta, Shaily Desai, Manvi Goel, Anil Bandhakavi, Tanmoy Chakraborty, and Md. Shad Akhtar. 2023. Counterspeeches up my sleeve! intent distribution learning and persistent fusion for intent-conditioned counterspeech generation. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5792–5809, Toronto, Canada. Association for Computational Linguistics.

Laura Hanu and Unitary Team. 2020. Detoxify. *Github*.

Amey Hengle, Aswini Padhi, Sahajpreet Singh, Anil Bandhakavi, Md Shad Akhtar, and Tanmoy Chakraborty. 2024. Intent-conditioned and non-toxic counterspeech generation using multi-task instruction tuning with RLAIF. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 6716–6733, Mexico City, Mexico. Association for Computational Linguistics.

Amey Hengle, Aswini Kumar Padhi, Anil Bandhakavi, and Tanmoy Chakraborty. 2025. CSEval: Towards automated, multi-dimensional, and reference-free counterspeech evaluation using auto-calibrated LLMs. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association*

for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers), pages 5402–5419, Albuquerque, New Mexico. Association for Computational Linguistics.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2021. Lora: Low-rank adaptation of large language models.

Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. 2023. Mistral 7b.

Daniel Jolley, Rose Meleady, and Karen M Douglas. 2020. Exposure to intergroup conspiracy theories promotes prejudice which spreads across groups. *British Journal of Psychology*, 111(1):17–35.

Jaylen Jones, Lingbo Mo, Eric Fosler-Lussier, and Huan Sun. 2024. A multi-aspect framework for counter narrative evaluation using large language models. *arXiv preprint arXiv:2402.11676*.

Satyadhar Joshi. 2025. A comprehensive review of qwen and deepseek llms: Architecture, performance and applications. *Performance and Applications (May 15, 2025)*.

Brendan Kennedy, Mohammad Atari, Aida Mostafazadeh Davani, Leigh Yeh, Ali Omrani, Yehsong Kim, Kris Coombs, Shreya Havaldar, Gwenyth Portillo-Wightman, Elaine Gonzalez, et al. 2018. The gab hate corpus: A collection of 27k posts annotated for hate speech. *PsyArXiv. July*, 18.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2020. 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*, pages 7871–7880.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. 2021. Retrieval-augmented generation for knowledge-intensive nlp tasks.

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

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

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach.

Binny Mathew, Punyajoy Saha, Hardik Tharad, Subham Rajgaria, Prajwal Singhania, Suman Kalyan Maity, Pawan Goyal, and Animesh Mukherjee. 2019. Thou shalt not hate: Countering online hate speech. In *Proceedings of the international AAAI conference on web and social media*, volume 13, pages 369–380.

Julia Mendelsohn, Yulia Tsvetkov, and Dan Jurafsky. 2020. A framework for the computational linguistic analysis of dehumanization. *Frontiers in artificial intelligence*, 3:55.

Ioannis Mollas, Zoe Chrysopoulou, Stamatis Karlos, and Grigorios Tsoumakas. 2022. Ethos: a multi-label hate speech detection dataset. *Complex & Intelligent Systems*, 8(6):4663–4678.

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*, pages 9759–9777.

Jimin Mun, Cathy Buerger, Jenny T Liang, Joshua Garland, and Maarten Sap. 2024. Counterspeakers' perspectives: Unveiling barriers and ai needs in the fight against online hate. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems*, pages 1–22.

Aswini Kumar Padhi, Anil Bandhakavi, and Tanmoy Chakraborty. 2025. Counterspeech the ultimate shield! multi-conditioned counterspeech generation through attributed prefix learning. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 27643–27663, Vienna, Austria. Association for Computational Linguistics.

Jakub Podolak, Szymon Łukasik, Paweł Balawender, Jan Ossowski, Jan Piotrowski, Katarzyna Bakowicz, and Piotr Sankowski. 2024. Llm generated responses to mitigate the impact of hate speech. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 15860–15876.

Jing Qian, Anna Bethke, Yinyin Liu, Elizabeth Belding, and William Yang Wang. 2019. A benchmark dataset for learning to intervene in online hate speech. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 4755–4764.

Qwen, :, An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jingren Zhou, Junyang Lin, Kai Dang,

Keming Lu, Keqin Bao, Kexin Yang, Le Yu, Mei Li, Mingfeng Xue, Pei Zhang, Qin Zhu, Rui Men, Runji Lin, Tianhao Li, Tianyi Tang, Tingyu Xia, Xingzhang Ren, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yu Wan, Yuqiong Liu, Zeyu Cui, Zhenru Zhang, and Zihan Qiu. 2025. Qwen2.5 technical report.

Nils Reimers and Iryna Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. *arXiv preprint arXiv:1908.10084*.

Punyajoy Saha, Abhilash Datta, Abhik Jana, and Animesh Mukherjee. 2024. Crowdcounter: A benchmark type-specific multi-target counterspeech dataset. *arXiv preprint arXiv:2410.01400*.

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

Xiaoying Song, Sujana Mamidisetty, Eduardo Blanco, and Lingzi Hong. 2025. Assessing the human likeness of ai-generated counterspeech. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 3547–3559.

Gemini Team, Rohan Anil, Sebastian Borgeaud, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M Dai, Anja Hauth, Katie Millican, et al. 2023. Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*.

Francielle Vargas, Isabelle Carvalho, Ali Hürriyetoğlu, Thiago Pardo, and Fabrício Benevenuto. 2023. Socially responsible hate speech detection: Can classifiers reflect social stereotypes? In *Proceedings of the 14th international conference on recent advances in natural language processing*, pages 1187–1196.

Claire Wardle. 2024. A conceptual analysis of the overlaps and differences between hate speech, misinformation and disinformation. *Department of Peace Operations (DPO). Office of the Special Adviser on the Prevention of Genocide (OSAPG). United Nations*.

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

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, et al. 2025. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q Weinberger, and Yoav Artzi. 2019. Bertscore: Evaluating text generation with bert. *arXiv preprint arXiv:1904.09675*.

Yizhe Zhang, Siqi Sun, Michel Galley, Yen-Chun Chen, Chris Brockett, Xiang Gao, Jianfeng Gao, Jingjing Liu, and William B Dolan. 2020. Dialogpt: Large-scale generative pre-training for conversational response generation. In *Proceedings of the 58th annual meeting of the association for computational linguistics: system demonstrations*, pages 270–278.

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

A Appendix

A.1 Hatespeech Selection

To construct the `FactualCS` dataset, we aggregated hate speech text from four established datasets to capture diverse dynamics of online abuse: *IntentCONANv2* (Hengle et al., 2024) (conversational abuse), *HatEval* (Basile et al., 2019) (identity-targeted attacks), *ETHOS* (Mollas et al., 2022) (discrimination and violence), and the *Gab Hate Corpus* (Kennedy et al., 2018) (extreme dehumanization). Our selection process involved multiple filtering stages to ensure dataset quality. First, we removed redundant hate speech instances across the five datasets, reducing the initial collection to $8,000$ unique samples. Next, we filtered out samples constrained by brevity (those under 5 words) that lacked sufficient context for meaningful analysis, retaining only self-expressive hate speech that clearly articulated harmful intent or targeted content. We extracted 2,384, 1,500, 550, 350, samples from IntentCONANv2, HatEval, ETHOS, Gab Hate Corpus respectively.

A.2 Annotator Demographics

Six annotators volunteered in the annotation process, comprising 4 male and 2 female annotators. All annotators possess strong computational backgrounds with expertise in natural language processing and hate speech analysis. The team includes researchers with prior publications in areas of content moderation, computational social science, and counter-narrative generation, ensuring they brought both technical proficiency and domain knowledge to the annotation task.

	Annotator 1	Annotator 2	Annotator 3	Annotator 4	Annotator 5	Annotator 6
Annotator 1	1.0000	0.9220	0.9500	0.9260	0.9420	0.9540
Annotator 2	0.9220	1.0000	0.9440	0.9240	0.9240	0.9520
Annotator 3	0.9500	0.9440	1.0000	0.9480	0.9520	0.9760
Annotator 4	0.9260	0.9240	0.9480	1.0000	0.9280	0.9560
Annotator 5	0.9420	0.9240	0.9520	0.9280	1.0000	0.9560
Annotator 6	0.9540	0.9520	0.9760	0.9560	0.9560	1.0000

Table 6: Inter-annotator Agreement Coefficients for 500 samples in Pilot Annotation Phase

	Annotator 1	Annotator 2	Annotator 3	Annotator 4	Annotator 5	Annotator 6
Annotator 1	1.0000	0.8860	0.9200	0.9000	0.9120	0.9300
Annotator 2	0.8860	1.0000	0.9140	0.8900	0.8900	0.9240
Annotator 3	0.9200	0.9140	1.0000	0.9240	0.9280	0.9540
Annotator 4	0.9000	0.8900	0.9240	1.0000	0.9000	0.9300
Annotator 5	0.9120	0.8900	0.9280	0.9000	1.0000	0.9300
Annotator 6	0.9300	0.9240	0.9540	0.9300	0.9300	1.0000

Table 7: Inter-annotator Agreement Coefficients in Main Annotation Phase for Hate Speech Type, Reasoning, Target Group, Query, and collected evidence

	Annotator 1	Annotator 2	Annotator 3	Annotator 4	Annotator 5	Annotator 6
Annotator 1	1.0000	0.8960	0.9180	0.9080	0.9140	0.9240
Annotator 2	0.8960	1.0000	0.9180	0.9000	0.8980	0.9200
Annotator 3	0.9180	0.9180	1.0000	0.9180	0.9240	0.9380
Annotator 4	0.9080	0.9000	0.9180	1.0000	0.9060	0.9240
Annotator 5	0.9140	0.8980	0.9240	0.9060	1.0000	0.9220
Annotator 6	0.9240	0.9200	0.9380	0.9240	0.9220	1.0000

Table 8: Inter-annotator Agreement Coefficients for generated counterspeech

A.3 Annotation Process

A.3.1 Pilot Annotation Phase

Before undertaking full-scale annotation, we conducted a pilot phase with 500 hate speech samples to establish annotation protocols and achieve inter-annotator consensus. During the proprietary phase, all six annotators first independently annotated each sample, after which they collaboratively reviewed and refined the annotations through mutual discussion and deliberation. For each hate speech instance, annotators identified and documented: (1) the hate speech text, (2) hate speech type, (3) reasoning behind choosing this hate type, (4) target group, (5) contextual query, (6) supporting evidence (when applicable), (7) counterspeech response, (8) source URL, (9) source date, and (10) source author. We classified hate speech into five categories: misinformation, stereotype, conspiracy theory, dehumanizing language, and non-factual claims. For hate speech types requiring factual verification specifically misinformation, stereotypes, and conspiracy theories, annotators conducted web searches to locate credible evidence and documented all relevant source details through manual annotation. Initial inter-annotator agreement (IAA) scores revealed substantial disagreement among annotators. The primary sources of divergence included: (1) distinguishing between stereotype and misinformation, (2) determining the threshold for dehumanizing language versus harsh criticism, and (3) identifying when stereotypes required explicit factual evidence versus contextual refutation. Through iterative discussion sessions and collaborative review of contentious cases, annotators developed a shared understanding of category boundaries and annotation standards. These deliberations led to the formulation of comprehen-

sive annotation guidelines (detailed in Appendix A.4), which operationalized each hate speech type with clear definitional criteria and decision rules. By the conclusion of the pilot phase, IAA scores exceeded 0.9 across all annotation dimensions (See Table 6), and the 500 annotated samples formed our gold-standard reference dataset.

A.3.2 Main Annotation Phase

Following the pilot phase, the six annotators proceeded to annotate the remaining dataset using the established guidelines detailed in Appendix A.4. All annotation components: hate speech type identification, reasoning behind hate speech type choice, target group detection, query formulation, and evidence collection, were completed manually to ensure high-quality labels. Inter-annotator agreement during the main phase remained consistently high around 91.55%, with scores reported in Table 7, demonstrating that the guidelines successfully transferred to the broader dataset. However, generating counterspeech responses for thousands of samples proved prohibitively time-intensive for manual annotation. To address this efficiency challenge while maintaining quality, we employed Gemini-2.0-Flash (Team et al., 2023) to generate counterspeech responses. We provided the model with few-shot examples from our 500 sample gold-standard dataset along with the manually annotated hate speech attributes (type, reasoning, group, query, evidence) as context for generation. This hybrid approach allowed us to scale counterspeech generation while grounding outputs in carefully curated examples. To validate counterspeech quality, annotators evaluated a random sample of generated responses across multiple dimensions including appropriateness, factual accuracy, and persuasiveness. Inter-annotator agreement scores for counterspeech quality assessment are presented in Table 8, indicating that the generated counterspeech met our quality standards and exhibited consistency with the manually authored examples from the pilot phase.

A.4 Procedure and Annotation Criteria

Before beginning the annotation process, all annotators thoroughly reviewed comprehensive resources on addressing online harassment and counterspeech strategies, including the field guide on "addressing online harassment" [5]. This preparatory phase involved extensive discussions with the annotators to deepen their understanding of hate speech taxonomy and effective counter-narrative generation. These dialogues ensured that the annotators were well-equipped with the necessary knowledge and context, enabling them to effectively contribute to the project with a shared conceptual framework.

A.4.1 Hate Speech Type Definitions and Criteria

The annotation process required annotators to classify each hate speech instance into one of five mutually exclusive categories. The following criteria were adhered to consistently throughout the annotation process.

Misinformation: Misinformation refers to hate speech containing factually incorrect or misleading claims that can be empirically verified against credible sources. This category encompasses false statements presented as facts that target or disparage specific groups, regardless of whether the misinformation was spread intentionally or unintentionally. The annotation criteria for this:

- The hate speech must contain at least one specific, verifiable factual claim that is demonstrably false.
- The false claim must directly contribute to the hateful characterization of the target group.
- Annotators must identify the precise false claim within the hate speech.
- Evidence contradicting the misinformation must be drawn from authoritative sources such as peer-reviewed research, government statistics, reputable news organizations, or fact-checking institutions.
- The evidence must be current and contextually relevant to the claim being refuted.

Stereotype: Stereotype refers to hate speech that perpetuates oversimplified, generalized, or prejudiced characterizations of social groups. These are fixed, commonly held beliefs about groups that reduce individual complexity to monolithic attributes, often negative or limiting in nature. The annotation criteria for this:

- The hate speech must attribute characteristics, behaviors, or traits to an entire group based solely on group membership.

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

HS Type	Description	Example	Ideal Response Type	Factual?
Misinformation	Relies on false or misleading factual claims about a group	Immigrants increase crime rates.	Fact-based rebuttal using data, statistics, authoritative sources	Yes
Conspiracy	Misinformation mixed with large-scale conspiracy framing	Jews control the world's banks.	Debunk narrative patterns, use cognitive inoculation strategies	Yes
Stereotype	Cultural or historical generalizations applied to a group	Women can't handle leadership.	Reframing, factual correction, highlighting counterexamples	Yes
Dehumanizing	Use of animalistic or violent metaphors	Refugees are parasites.	Moral, legal, and social norm reminders	No
Non-factual	Hatred or disgust without factual claims	Muslims are disgusting.	Empathy-based counterspeech, reinforcing social norms	No

Table 9: Hate Speech Types based on Factuality, Response Strategies, and Factual Rebuttal Requirements

- The characterization must be reductive, denying individual variation within the group.
- Evidence to counter stereotypes should include: statistical data showing group diversity, research findings on within-group variation, documented counter-examples, or scholarly work debunking the stereotype.
- For cultural or behavioral stereotypes, evidence may include anthropological studies, sociological research, or representative examples demonstrating group heterogeneity.

How Stereotype is different from Misinformation? Stereotypes make generalized claims about group characteristics; misinformation makes specific false factual claims. Stereotypes may not be empirically testable in the same way (e.g., "X group is lazy" vs. "X group has 90% unemployment"). When a stereotype includes a verifiable false statistic, classify as misinformation.

Conspiracy Theory: Conspiracy theory refers to hate speech promoting unfounded explanatory narratives that attribute events, policies, or social phenomena to secret plots by powerful actors, typically targeting specific groups as the alleged conspirators. These narratives reject mainstream explanations in favor of hidden malevolent coordination. The annotation criteria for this:

- The hate speech must allege a secret, coordinated plan by a specific group to achieve nefarious goals.
- The conspiracy must lack credible evidentiary support and reject established explanations.
- The alleged conspiracy must involve extraordinary claims of coordination, secrecy, or power.
- Evidence must demonstrate: (a) the lack of credible support for the conspiratorial claim, and (b) when possible, factual explanations for the events or phenomena in question.
- Evidence sources should include: investigative journalism, expert analysis, academic debunking, or fact-checking reports.

Dehumanizing Language: Dehumanizing language refers to hate speech that depicts individuals or groups as subhuman, animalistic, objects, or fundamentally less than human. This category focuses on linguistic markers that strip target groups of human dignity, moral status, or basic humanity. The annotation criteria for this:

- The hate speech must use language that explicitly or implicitly denies the full humanity of the target group.
- Dehumanization can take multiple forms: animalistic (e.g., vermin, animals), mechanistic (e.g., tools, objects), biological (e.g., disease, infestation), or exclusionary (e.g., not deserving of human rights).
- This category focuses on the nature of the language used rather than factual claims.
- Evidence is typically not required for this category, as it concerns linguistic framing rather than empirical claims.
- However, contextual information about why such language is harmful and its historical usage may be included.

Non-Factual Claims: Non-factual claims refer to hate speech expressing opinions, value judgments, moral condemnations, or abstract assertions

that cannot be empirically verified through evidence but still promote harmful attitudes, prejudice, or discrimination toward target groups. The annotation criteria for this:

- The hate speech must express subjective evaluations, opinions, or value judgments rather than factual claims.
- The statements are not amenable to fact-checking but nonetheless promote hatred or prejudice. These may include: aesthetic judgments, moral condemnations, predictions about the future, personal preferences framed as group characteristics, or abstract value-laden assertions.
- Counterspeech for this category typically involves explaining why such rhetoric is harmful, its potential effects, or providing counter-narratives that promote inclusive values.

A.4.2 Additional Annotation Requirements

Reasoning Behind Hate Speech Type Choice For each classification decision, annotators must provide explicit reasoning explaining why the hate speech was assigned to a particular category. This reasoning should reference specific textual elements and explain how they meet the definitional criteria for the chosen category. This metadata helps in decision-making processes and facilitates review and quality control.

Target Group Identification Annotators must identify the specific social group(s) targeted by the hate speech. Target groups may be defined by: race, ethnicity, religion, nationality, gender, sexual orientation, disability status, or other protected characteristics. When multiple groups are targeted, all should be documented.

Query Formulation For hate speech types requiring evidence (misinformation, stereotype, conspiracy theory), annotators must formulate clear, focused search queries that would retrieve relevant counter-evidence. Queries should be specific enough to yield targeted results but broad enough to capture relevant sources. Example: For "Muslims commit most terrorist attacks," a query might be "global terrorism statistics by perpetrator group."

Evidence Documentation When evidence is required, annotators must document:

- Source URL: Complete web address of the evidence source.
- Source Date: Publication or last updated date.
- Source Author/Organization: Individual author or publishing organization.
- Relevant Excerpt: Brief quote or summary of the evidence that directly counters the hate speech claim.
- Credibility Assessment: Brief note on why the source is authoritative (e.g., peer-reviewed, government statistics, reputable news organization)

A.4.3 Edge Cases and Decision Rules

Multiple Category Applicability: When hate speech could fit multiple categories, annotators should prioritize based on the most prominent or harmful element. The decision hierarchy typically follows: (1) Dehumanizing language (if explicit dehumanization is present), (2) Conspiracy theory (if elaborate coordination is alleged), (3) Misinformation (if specific false facts are central), (4) Stereotype (if generalization is primary), (5) Non-factual claims (as default for subjective assertions).

Ambiguous Cases: When categorization is uncertain, annotators should document the ambiguity in the reasoning field and, during the pilot phase, flag for group discussion. In the main phase, annotators should apply the guidelines as consistently as possible and note any uncertainties.

Intersectional Targeting: When hate speech targets multiple groups simultaneously (e.g., "Muslim women"), annotators should document all target groups and consider how intersecting identities may compound the harm.

A.5 `FactualCS` Taxonomy Distinctions

To ensure robust categorization and address the diverse semantic nature of online abuse, our taxonomy is grounded in established social science literature. Rather than treating hate speech as a homogeneous entity, we define five mutually exclusive categories that act as functionally distinct logical triggers, each necessitating a fundamentally different counterspeech strategy as summarized in Table 9. The core distinction relies on separating claims that are empirically verifiable from those based purely on subjective affect or linguistic framing.

Misinformation versus Stereotype: While both categories fall under factual hate, they operate on different levels of specificity. Following the distinctions outlined by (Wardle, 2024) and (Vargas et al., 2023), misinformation relies on specific, falsifiable events or fabricated statistics (e.g., "Immigrants increase crime rates"), requiring a direct, empirical fact-check using authoritative data. In contrast, stereotypes make broad, reductive cultural or historical generalizations about a group's traits or abilities (e.g., "Women can't handle leadership"). Stereotypes are countered not merely with isolated statistics, but by highlighting group heterogeneity and providing specific counterexamples to reframe the narrative.

Stereotype versus Non-Factual Hate: A critical boundary exists between stereotypes and non-factual hate, addressing potential ambiguities in classification. The distinction relies entirely on the presence of a propositional claim. Stereotypes, despite being generalizations, contain a premise that can be logically debated and factually refuted. Conversely, non-factual hate consists of subjective expressions of disgust, slurs, or opinions without any verifiable assertion (e.g., "Muslims are disgusting"). Because non-factual hate contains no empirical claim to disprove, attempting to "fact-check" it results in a logical mismatch. Therefore, non-factual expressions trigger an entirely different strategic response based on empathy and social norm reinforcement.

Conspiracy Theories versus Misinformation: Incorporating the definitions from (Jolley et al., 2020), conspiracy theories are differentiated from standard misinformation by their structural narrative. They allege secret, large-scale coordination by powerful actors (e.g., "Jews control the world's banks"). Because these narratives often reject mainstream evidence as part of the cover-up, effective counterspeech requires debunking the narrative patterns and utilizing cognitive inoculation strategies, rather than simply presenting contradictory facts.

Dehumanization versus General Non-Factual Hate: Drawing on the framework of (Mendelsohn et al., 2020), dehumanization is separated from general non-factual hate due to its specific linguistic severity. It utilizes animalistic, mechanistic, or violent metaphors (e.g., "Refugees are parasites") to actively strip target groups of their human dignity and moral status. This extreme form of abuse requires counterspeech that explicitly reminds the user of moral, legal, and social norms to re-establish the target's humanity. By strictly delineating these boundaries, `FactualCS` ensures that models learn the underlying intent of the abuse, allowing for precise, strategy-aligned generation.

Beyond defining these boundaries, it is crucial to justify the specific granularity of our five-category taxonomy compared to coarser or finer alternatives. A coarser categorization scheme (such as a simple binary Hate versus Not-Hate classification) is fundamentally insufficient because it treats distinct logical flaws homogeneously. For example, grouping misinformation with dehumanization forces a model into "logical mismatching," where it might attempt to empirically fact-check a subjective slur or provide a normative moral opinion against a falsifiable statistic. The empirical necessity of this specific decomposition is proven in our ablation study (Section 6.3); when the Hate Speech Analyst is removed ("Without HSA") and the model must generate counterspeech without categorical gatekeeping, the Factual Score severely drops from 0.969 to 0.620 (a 35% decrease). Conversely, a finer-grained taxonomy would introduce redundant complexity and data sparsity without fundamentally altering the required tool usage or reasoning paths. Our five categories represent the optimal resolution, sufficiently detailed to trigger the correct core reasoning strategy (empirical fact-checking, logical rebuttal, or normative reframing) while remaining computationally efficient for compact agents.

A.6 Advantage of `FactualCS`

`FactualCS` advances counterspeech dataset construction by addressing critical gaps in existing resources and introducing evidence and strategy aware counter-narratives for combating online hate. Unlike prior datasets that primarily emphasize rhetorical or emotional responses, `FactualCS` grounds counterspeech in verifiable evidence with explicit source attribution and claim–rebuttal linkage, making it effective against misinformation, stereotype, and conspiracy driven hate. Importantly, it is not limited to factual refutation: through a fine-grained hate speech taxonomy, `FactualCS` enables strategy-specific counterspeech for non-factual and dehumanizing cases, where value-based reframing and rehumanization are more appropriate. Each instance is further enriched with contextual metadata such as target groups, reasoning, and evidence provenance. By aggregating hate speech from multiple established corpora and

employing a rigorous hybrid human–AI annotation pipeline with high inter-annotator agreement (>0.8), `FactualCS` achieves both diversity and quality at scale, resulting in a reliable and practically useful resource for hate speech type aware counterspeech generation.

A.7 Dataset Statistics

We perform a comprehensive statistical analysis of `FactualCS` to characterize the distribution of hate speech types and the corresponding necessity for external evidence. The dataset consists of a total of 4,784 instances, which are partitioned into a training set of 3,912 samples, a validation set of 383 samples, and a test set of 489 samples, as detailed in Figure 3a. To prevent class imbalance during model training and evaluation, we employed a stratified splitting strategy, ensuring that the proportional representation of each hate speech category remains consistent across all three splits (Figure 3b). In terms of class distribution, the dataset is semantically diverse: Stereotype Hate (SH) constitutes the largest portion at 32.8%, followed by Non-factual Hate (NH) at 27.1% and Dehumanization Hate (DH) at 23.5%, with Misinformation (MH) and Conspiracy Hate (CH) accounting for 12.1% and 4.4% respectively (Figure 3d). A defining feature of our dataset is the strategic inclusion of reasoning traces for evidence retrieval. As shown in Figure 3e, 47.0% of the instances utilize the web search tool to ground the counterspeech. Crucially, this tool usage is strictly correlated with the hate category; as illustrated in Figure 3c, factual categories (SH, MH, CH) predominantly trigger web searches to debunk claims, whereas abstract categories (DH, NH) rely on logical and moral refutation without requiring external data.

A.8 Computational Efficiency Analysis

In this study, the claim of being "memory-efficient" refers to the FIRE framework's ability to achieve state-of-the-art results using compact models without requiring the hardware footprint of larger, monolithic LLMs. As detailed in Table 10, FIRE utilizes two 1.7B parameter models (Hate Speech Analyst and Counterspeech Generator). Because these agents operate in distinct, sequential stages, the system only needs to load one 1.7B model into memory at any given time. This architectural choice reduces the peak hardware requirement from approximately 16GB of VRAM (typical for 7B–8B parameter models) to roughly 4GB of VRAM, representing a 75% reduction in peak memory consumption.

Furthermore, while FIRE involves multiple agentic steps and external tool usage, the overall inference time remains highly competitive. The slight overhead in FIRE's end-to-end latency (2.5 – 3.0 seconds) is primarily attributable to the web search API round-trip during the HEAL phase. Despite this network latency, the framework's runtime remains comparable to single-pass generations from 7B and 8B baselines, proving that multi-agent decomposition can yield superior factual performance and significant memory savings without introducing prohibitive latency bottlenecks.

Model	Params	Peak VRAM	Latency
FIRE (Sequential)	2×1.7B	∼4GB	2.5–3.0s
Llama-3.1 8B-Instruct	8B	∼16GB	2.0–2.8s
Qwen-2.5 7B-Instruct	7B	∼16GB	1.8–2.5s
Mistral 7B-Instruct	7B	∼16GB	1.7–2.4s

Table 10: Computational efficiency profiling comparing peak memory requirements and inference latency between FIRE and single-pass baselines.

A.9 Hate Speech Analyst: Purity Score and Error Cascading

To address potential concerns regarding error propagation, where an initial misclassification by the Hate Speech Analyst (HSA) cascades into an incorrect downstream response, we analyze the distribution of the HSA's Purity Score ($\mathcal{P}$). This score acts as a confidence proxy by quantifying the homogeneity of the k-nearest neighbors retrieved from memory. A predefined threshold ($\tau = 0.62$) dictates whether the agent accepts the prediction with high confidence or triggers an uncertainty-aware "think again" mechanism. Based on an evaluation of 482 test samples, the purity score successfully isolates reliable predictions from ambiguous ones. As shown in Table 11, $\mathcal{P}$ exceeds the threshold in 88.4% of instances, corresponding to a high accuracy of 73.2%. Conversely, in the 11.6% of cases where $\mathcal{P} < 0.62$, accuracy drops to 48.2%. This 25.0% gap confirms that $\mathcal{P}$ is a highly effective, correlated signal for downstream error likelihood.

Furthermore, the relationship between purity and accuracy follows a clear monotonic trend: accuracy increases steadily from 46.9% in the 0.4–0.6 range to 76.0% in the high-confidence 0.8–1.0 range (peaking at 82.7% when $\mathcal{P} = 1.0$). By successfully flagging the 11.6% of instances that are highly am-

biguous and forcing the agent to explicitly reason over the top-two predicted categories, the framework actively intercepts and mitigates cascading errors prior to the final generation phase.

Condition	Samples (%)	Accuracy
High Confidence ($\mathcal{P} \geq 0.62$)	88.4% (426)	73.2%
Ambiguous ($\mathcal{P} < 0.62$)	11.6% (56)	48.2%
Overall	**100%** (482)	**70.3%**

Table 11: Distribution and accuracy breakdown of HSA predictions based on the purity score threshold.

A.10 Promting Strategies

A.10.1 ZeroShot Prompt

We used the following system prompt for the Zero-Shot setting of baselines:

> **System Instruction**: You are a hate speech analyst. Analyze the hate text and identify the target group, category of hate (misinfo_hate, nonfactual_hate, conspiracy_hate, stereotype_hate, dehumanization_hate), and give reasoning behind your decision in `analyst_reasoning` field. Generate concise and factual search queries and evidence (a single evidence) that will help verify or fact-check the hate speech. Generate a concise, factual counterspeech (a single counterspeech text) in response to the given hate text using its target group, hate type, and reasoning. Use retrieved evidence when available to correct misinformation or challenge harmful narratives. You output ONLY valid JSON. No markdown, no explanation, no commentary. Only return JSON with fields: `target_group`, `hs_type`, `analyst_reasoning`, `retrieval_query`, `retrieved_evidence`, `counterspeech`.
>
> **User Input**: [HATE_SPEECH]
> Analyze the hate speech and return ONLY JSON. Do NOT add ````json` or explanations.

A.10.2 FewShot Prompt

For the Few-Shot setting, we utilized the same system instruction as the Zero-Shot setting but provided 5 in-context examples to guide the model's reasoning and formatting:

> **System Instruction**: [Same as Zero-Shot]
>
> **Examples**:
>
> 1. **User**: "Weak beta male, hope he turns into a tranny after this traumatic experience, haha"
> **Assistant**: {
> "target_group": "transgender people",
> "hs_type": "dehumanization_hate",
> "analyst_reasoning": "This statement expresses a desire for harm to befall someone and uses a derogatory term for transgender people. It aims to strip them of their humanity by associating them with weakness and trauma.",
> "retrieval_query": null,
> "retrieved_evidence": null,
> "counterspeech": "Wishing harm on someone and using derogatory terms against transgender people is unacceptable and harmful. Transgender individuals deserve respect and dignity, like everyone else."
> }
> 2. **User**: "The Democrat Party has ONE MISSION - to destroy the USA and if YOU press the Dummy lever - you are helping... LINK"
> **Assistant**: {
> "target_group": "Democrat Party",
> "hs_type": "misinfo_hate",
> "analyst_reasoning": "This statement makes a false claim about the Democrat Party's mission. It needs fact-checking to determine accuracy.",
> "retrieval_query": "Democrat Party mission statement",
> "retrieved_evidence": "The Democratic Party states its mission is to support working families, equal opportunity, and a healthy future for all Americans.",
> "counterspeech": "The Democratic Party's mission focuses on supporting working families and ensuring

equal opportunity, not destroying the USA."
}

3. **User**: "Women are complaining because they cant get highpaying jobs. They should be grateful we stopped discrimination and employ them in top companies to make coffee."

   Assistant: {
   "target_group": "women",
   "hs_type": "stereotype_hate",
   "analyst_reasoning": "This statement relies on the stereotype that women are only suitable for menial jobs and that they are ungrateful despite progress in gender equality. It diminishes their capabilities and aspirations.",
   "retrieval_query": "gender pay gap statistics",
   "retrieved_evidence": "In 2022, women earned 84 cents for every dollar earned by men among full-time workers.",
   "counterspeech": "Women have significantly contributed to various fields throughout history; Marie Curie, for example, pioneered radioactivity research. The gender pay gap persists, but women's ambitions go far beyond stereotypical roles; they aim to lead and innovate, deserving equal opportunities and recognition."
   }

4. **User**: "If women dont stop complaining, then men will find a solution. If a woman does not submit, then there will be war."

   Assistant: {
   "target_group": "women",
   "hs_type": "nonfactual_hate",
   "analyst_reasoning": "This statement promotes male dominance and threatens violence against women who do not conform to traditional gender roles. It does not contain factual information but expresses a harmful opinion.",
   "retrieval_query": null,
   "retrieved_evidence": null,
   "counterspeech": "I understand your frustration; it's tough when people don't see eye-to-eye. Let's strive for mutual respect and open communication in our interactions."
   }

5. **User**: "We are ruled by a small Jewish elite that exploit the weak."

   Assistant: {
   "target_group": "Jewish people",
   "hs_type": "conspiracy_hate",
   "analyst_reasoning": "This statement promotes the antisemitic conspiracy theory that Jewish people secretly control governments and exploit others for their benefit.",
   "retrieval_query": "Jewish people controlling the world",
   "retrieved_evidence": "The claim that Jewish people secretly control the world is a long-running antisemitic conspiracy theory.",
   "counterspeech": "The claim that a Jewish elite rules and exploits is a conspiracy theory. This antisemitic trope is rooted in fabricated texts like 'The Protocols of the Elders of Zion'. These lies have been repeatedly discredited. Jews account for approximately 0.2 percent of the global population."
   }

User Input: "[HATE_SPEECH]"
Analyze the hate speech and return ONLY JSON. Do NOT add ````json` or explanations.

A.10.3 Retrival-Based Prompt

In the Retrieval-Based setting, we dynamically select the most similar example from the training set using cosine similarity on sentence embeddings (all-mpnet-base-v2). This retrieved exemplar is prepended to the prompt context.

System Instruction: [Same as Zero-Shot]

Retrieved Exemplar (Top-1): "[RETRIEVED_HS_EXAMPLE]" **Assistant**: { "target_group": "...", "hs_type": "...", ... }

User Input: "[HATE_SPEECH]"
Analyze the hate speech and return ONLY JSON.

A.11 Computing Information

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

A.12 Hyper-parameter Information

- Precision: BF16
- Temperature: 0.0 (Greedy decoding)
- Baselines Batch Size: 16
- Baselines SFT Learning rate: 2e-4
- Max New Tokens: 512
- Retrieval Encoder: sentence-transformers/all-mpnet-base-v2

A.13 Manual Audit of Websearch Tool Quality

To directly address concerns regarding the reliability and potential retrieval of harmful or fake web evidence, we conducted a manual audit on a randomly sampled subset of 100 evidence instances retrieved during the `HEAL` phase. We categorized the sources into four tiers: Institutional/Academic, Mainstream Informational, Social-Media-Style, and Potentially Unreliable.

As detailed in Table 12, the combination of the DuckDuckGo safe-search API and our framework's strict filtering effectively suppresses low-quality content. The vast majority of retrieved evidence (92%) originates from reliable domains, with 58% coming from institutional or academic sources (e.g., WHO, PubMed/NCBI, ScienceDirect) and 34% from mainstream informational websites (e.g., Wikipedia). Only 6% originated from social-media-style platforms (e.g., Reddit), and a mere 2% were judged as potentially low-quality or weakly reliable. This low incidence of content from malicious sources confirms that `FIRE`'s retrieval mechanism acts as a robust factual safeguard rather than an amplifier of misinformation. Furthermore, as a secondary defense against the 2% of potentially unreliable sources that bypass the safe-search filter, the Counterspeech Generator's system prompt explicitly instructs the agent to critically evaluate retrieved text and strictly prohibits the propagation of any hate speech claims.

Source Category	Retrieval Frequency
Institutional / Academic	58%
Mainstream Informational	34%
Social-Media-Style	6%
Potentially Low-Quality	2%

Table 12: Manual audit of 100 retrieved web search instances categorized by source reliability.

A.14 Human Evaluation Protocol

To ensure rigorous subjective evaluation, we employed 30 expert annotators who were provided with formal definitions of hate speech and counterspeech. The annotators were tasked with a comparative ranking of model outputs from 1 (Best) to 4 (Worst) across four distinct dimensions: Independent Counterspeech (ICS), Adequacy/Fluency (Ad), Contextual Relevance (CoRl), and Argumentative Effectiveness (ArgE). Following the ranking phase, responses were categorized into Best, Above Average, Average, and Worst. To calibrate the annotators and ensure consistency, we provided explicit anchor examples prior to the task. For instance, a "Good/Best" response for Argumentative Effectiveness directly refutes the underlying propositional claim using logical evidence without resorting to counter-attacks. Conversely, a "Bad/Worst" response either hallucinates facts, utilizes toxic language, or completely misidentifies the target group's protected characteristic. This structured calibration ensured high reliability in the final comparative rankings presented in Section 6.4. The exact instructions provided to annotators were as follows:

Core Definitions

Hate Speech (HS): Communication that attacks or uses pejorative/discriminatory language against a person or group based on identity factors (religion, race, descent, gender, etc.).

Counterspeech (CS): A tactic to counter hate speech or misinformation by presenting an alternative narrative, using empathy, and challenging hate narratives without censorship or reciprocal hate.

Anchor Example (HS): "All Muslims are terrorists."

Anchor Example (CS): "This statement is a gross generalization that is not supported by facts. The vast majority of Muslims are peaceful..."

Evaluation Metrics (Ranked 1 [Best] to 4 [Worst])

- **Independent CS (IC):** Measures if the CS stands alone meaningfully without context.

Bad: "No, they're not." *Good:* "This statement is a gross generalization not supported by facts."

- **Fluency (F):** Measures grammatical soundness and coherence.
  Bad: "Muslims are good people." *Good:* "The vast majority of Muslims are peaceful and law-abiding citizens."

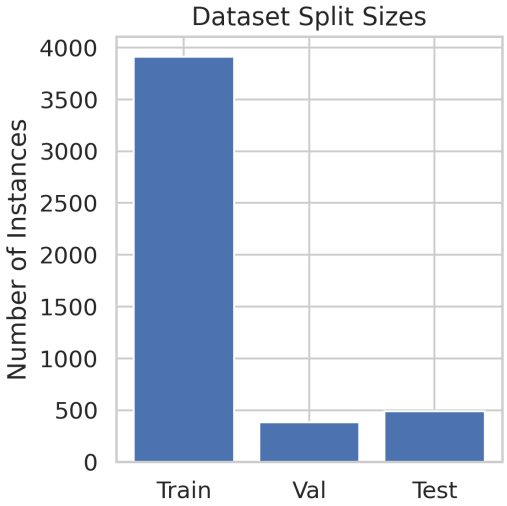
- **Contextual Relevance (CR):** Addresses central/implicit aspects of the HS (Subject, False Claim, Stereotype).
  Bad: "You're wrong." *Good:* "It is unfair and unjust to label an entire group of people based on the actions of a few extremists."

- **Argumentative Effectiveness (AE):** Provides a coherent, persuasive, logically sound argument.
  Bad: "No, they're not." *Good:* "This statement is a gross generalization... We should focus on building bridges between different communities."

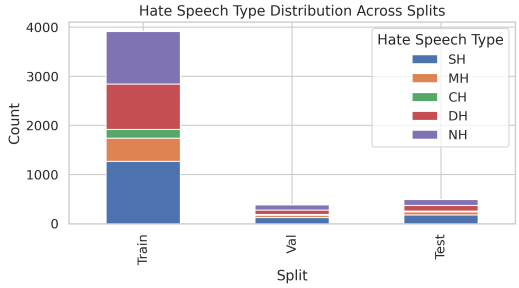
Example Annotation Interface

For each section, annotators were given an HS statement and four CS generations (A, B, C, D) and asked to rank them for each of the four metrics. A real task instance is shown below:

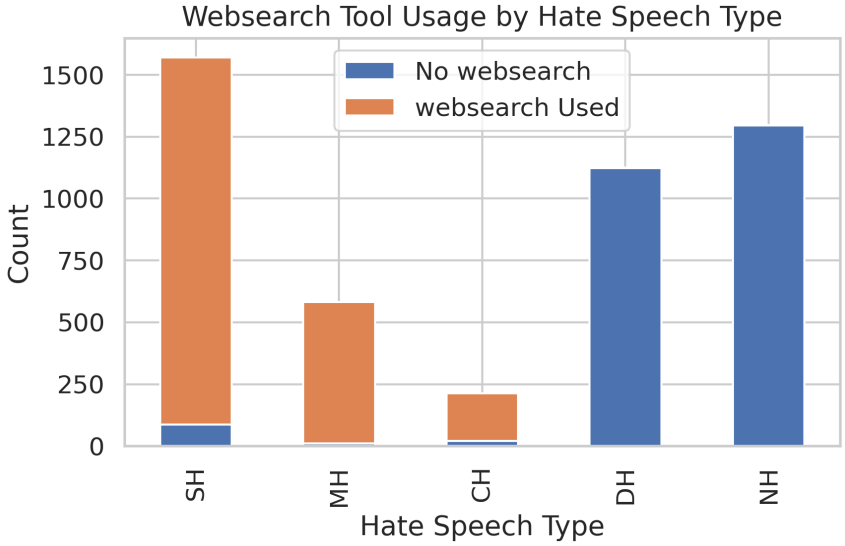
> **HS:** Muslim are invading our country.
>
> **CS-A:** Muslims enrich our country with their diverse culture, traditions, and values. Their presence brings beauty and unity, making our nation stronger and more vibrant...
>
> **CS-B:** Muslims are contributing wonderfully to our society, spreading a culture full of rich history, unique traditions, and incredible cuisine that brings joy...
>
> **CS-C:** The claim that Muslims are invading our country is false and it spreads misinformation about immigration and demographic changes. It stokes fear and hatred by suggesting that a specific group is deliberately undermining the native population. But the reality is shaped by complex social and economic factors, not hostile intent...
>
> **CS-D:** I am a muslim and i am not invading your country. i am a muslim and i am not invading your country. i am a muslim and i am not invading your country...



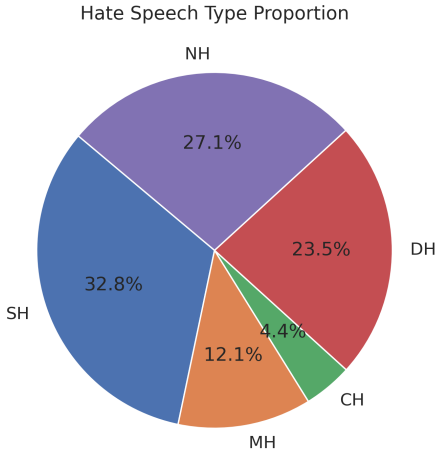
(a) Dataset split sizes across Train, Val, and Test.



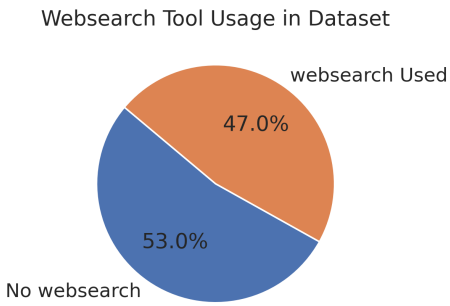
(b) Distribution of Hate Speech types across splits.



(c) Analysis of Websearch Tool usage conditioned on specific Hate Speech categories.



(d) Overall proportion of Hate Speech types.



(e) Overall distribution of Websearch Tool usage.