JOURNAL OF E-LEARNING AND KNOWLEDGE SOCIETY
Vol. 21, No. 1 (2026), pp. 75-90

Beyond automation: the collaborative role of AI in countering Hate Speech. A scoping review

Antonella Cosola[a,1], Marinella Paciello[b], Alessandro Pollini[a], Francesca D'Errico[c]

[a]*Uninettuno Telematic International University – Rome (Italy)*

[b]*Sapienza, University of Rome – Rome (Italy)*

[c]*University of Bari "Aldo Moro"– Bari (Italy)*

(submitted: 31/3/2026; accepted: 23/7/2026; published: 17/8/2026)

Abstract

The unregulated proliferation of hate speech, defined as any expression of beliefs that incite hatred, discrimination, or violence against individuals or social groups (Fortuna & Nunes, 2018), in social media has called into question the efficacy of removal-based content moderation and has driven the need to explore alternative approaches. AI-generated counterspeech could be a scalable and cost-effective intervention (Ashida & Komachi, 2022). Counterspeech is defined as a direct response to hate speech aimed at refuting or undermining it (Wachs et al., 2023). Research in this growing field is fragmented across computational linguistics and social sciences, lacking systematic methods to define and evaluate effective counterspeech, revealing a gap between AI's technical capabilities and evidenced social impact.

This review aims to examine evidence from these two fields to identify the differing conceptualizations of counterspeech, map the collaborative approaches used for counterspeech generation and evaluation, and examine the alignment between stated research goals (e.g., reducing hate speech in digital environments) and the operational measures employed to identify them.

Only a few studies evaluate counterspeech in real-world interactive settings or measure long-term impact. We identify some key misalignments, such as the reliance on text metrics and human or artificial annotators as proxies for complex social interventions, instead of integrating psychological and behavioral measures. We propose a linked approach to counterspeech that links theoretical frameworks, procedures, measurements and research goals from the perspective of both psychology and computational linguistics to identify promising directions for future interdisciplinary collaboration and more socially meaningful AI-assisted interventions.

KEYWORDS: Counter-speech, Artificial Intelligence, Large Language Models, Human Augmentation.

DOI
https://doi.org/10.20368/1971-8829/1136378

CITE AS
Cosola, A., Paciello, M., Pollini, A., & D'Errico, F. (2026). Beyond automation: the collaborative role of AI in countering Hate Speech. A scoping review. *Journal of e-Learning and Knowledge Society*, *21*(1), 75-90.
https://doi.org/10.20368/1971-8829/1136378

1. Introduction

The choice to implement artificial intelligence (AI) in content moderation on social media is likely driven by the surge of negative polarizing content. Hate speech is defined as

> *"language that attacks or diminishes, that incites violence or hate against groups, based on specific characteristics such as physical appearance, religion, descent, national or ethnic origin, sexual orientation, gender identity or other, and can manifest itself with different linguistic styles, even in subtle forms or when humor is used" (Fortuna & Nunes, 2018, p.5).*

Deleting content and banning users tends to disperse hatred across less regulated platforms rather than resolving it (Horta Ribeiro et al., 2023), resulting in being now outdated in the current landscape. In recent

1 corresponding author - email: a.cosola@students.uninettunouniversity.net

years, therefore, researchers have begun to explore the potential of online counterspeech, defined as a direct response to hate speech aimed at refuting or undermining hate speech (Wachs et al. 2023). Counterspeech may be a more effective response to hateful content as it is not constrained by the single platform rules and doesn't undermine the principles of free expression (Zhu & Bhat, 2021).

Writing effective counterspeech is time-consuming and requires expertise, making human-only moderation costly (Chung et al., 2024; Sportelli et al., 2025). Large Language Models (LLMs) can be a solution, being easily applicable, less costly at scale (Ashida & Komachi, 2022), and preventing human strain from prolonged exposure to harmful content (Chung et al., 2024). Although AI-generated counterspeech can be a useful content moderation tool, it also presents some weaknesses, such as a lack of argumentative richness (Bonaldi et al., 2024), struggle with deeper reasoning (Mun et al., 2023), perceived inauthenticity by the users (Bar et al., 2024), and risk of generating hallucinated or non-factual information (Wang et al., 2024).

Critically, counterspeech evaluation still relies on traditional natural language processing (NLP) metrics. Also, LLM evaluators may favor AI-generated counterspeech (Zubiaga et al., 2024). Another critical issue is the absence of a metric that captures counterspeech quality or effectiveness, creating a misalignment with users' behavioral change. Computational research on AI-generated counterspeech is advancing rapidly, while psychosocial research lags behind. This misalignment limits insight into psychosocial outcomes, particularly as tool development lacks robust theoretical grounding (Chung et al., 2024). Consequently, unaddressed risks remain: fully automated approaches may promote diffusion of responsibility (Darley & Latané, 1968) and moral disengagement (Bandura, 1996), harming the social norms that bystander intervention relies on (Mathew et al., 2019).

Adopting a psychosocial lens, this review assesses how psychological and social dimensions are integrated into AI-generated counterspeech conceptualization, creation, and evaluation, ensuring a common object of study across terminological differences.

2. Review Methods

This review was conducted in accordance with the PRISMA-ScR (Preferred Reporting Items for Systematic reviews and Meta-Analyses) guidelines (Tricco et al., 2018) to ensure transparency and reproducibility. The protocol was not registered.

2.1 Search Strategy

The search was conducted in January 2026 on Scopus, Web of Science, and arXiv. Given the rapidly evolving nature of AI research, we included relevant pre-prints from the arXiv repository, even beyond the search date.

The search string, illustrated in Table 1, was designed to combine key concepts related to counterspeech (e.g., counterarguments, counternarratives) keywords meant to search the tools (e.g., AI, large language models) and terms related to human-AI collaboration (e.g., human-AI collaboration, human augmentation).

2.2 Selection Process

The selection process followed the PRISMA-Scr flowchart phases, as illustrated in Figure 1. References were initially imported into Excel for duplicate removal and the following screening process.

The initial search yielded 1184 records. After duplicate removal, 1100 remained. We excluded 730 studies published before January 1, 2017, as this marks the introduction of transformer models, which now underpin modern AI and Natural Language Processing.

Of the remaining, 608 were excluded during title and abstract screening as irrelevant, leaving 121 articles for full-text review. Two independent reviewers achieved moderate initial agreement (κ=0.608). After discussion with a third reviewer, we proceeded with the full-text assessment.

In the full-text assessment, an additional 60 studies were excluded due to a theoretical approach that was not equivalent to counterspeech because they presented a theoretical concept of counterspeech that did not align with the operational definition employed in this review, not generating counterspeech, being inadequate (book chapters or shared tasks overviews), or not being in the English language. The remaining 61 were further assessed: we excluded studies lacking human involvement in generation or counterspeech evaluation, those where researchers self-evaluated counterspeech, and kept studies with external participants as reviewers, such as crowdworkers, recruited evaluators, and real social media users (n=18). We also excluded one dataset evaluation paper, one due to a different theoretical conceptualization, and one non-English paper. Including 5 additional studies not returned by the keyword search, the final total is 45 papers. The full reference list is available in Appendix A.

For each included study, the following data were extracted:

a) bibliographic details and disciplinary affiliation;

b) operational definition of counterspeech used;

c) collaborative human-AI generation method;

d) evaluation methods.

Table 1 - Databases and citation indexes searched.

Database	Search query	Filters	Results
Scopus	TITLE-ABS-KEY ((“counterspeech” OR “counter speech” OR “counter narrative” OR “counter-narrative” OR “counter argument” OR “counterargument” OR “counter-argument” OR “counter reply” OR “counter-stereotype” OR “counter-aggression” OR “counter hate speech” OR “countering hate speech” OR “counter-terrorism” OR “counter-hate speech”) AND (“AI” OR “generative AI” OR “large language model” OR “LLM” OR “GPT” OR “natural language generation” OR “NLP” OR “language model” OR “artificial intelligence” OR “generation” OR “automation” OR “human-AI collaboration” OR “human augmentation”))	2000-2026	685
arXiv	“counterspeech” OR “counter speech” OR “counter narrative” OR “counter-narrative” OR “counter argument” OR “counterargument” OR “counter-argument” OR “counter reply” OR “counter-stereotype” OR “counter-aggression" OR “counter hate speech” OR “countering hate speech" OR “counter-terrorism” OR “counter-hate speech” AND “AI” OR “generative AI” OR “large language model” OR “LLM” OR “GPT” OR “natural language generation” OR “NLP" OR “language model” OR “artificial intelligence” OR “generation” OR “automation” OR “human-AI collaboration” OR “human augmentation”	Computer science (cs); 2000–2026	104
Web of Science	TS = ((“counterspeech" OR “counter speech” OR “counter narrative” OR “counter-narrative” OR “counter argument” OR “counterargument” OR “counter-argument” OR “counter reply” OR “counter-stereotype” OR “counter-aggression” OR “counter hate speech” OR “countering hate speech” OR “counter-terrorism” OR “counter-hate speech”) AND (“AI” OR “generative AI” OR “large language model” OR “LLM” OR “GPT” OR “natural language generation” OR “NLP” OR “language model” OR “artificial intelligence” OR “generation” OR “automation” OR “human-AI collaboration” OR “human augmentation”))	2000-2026	401

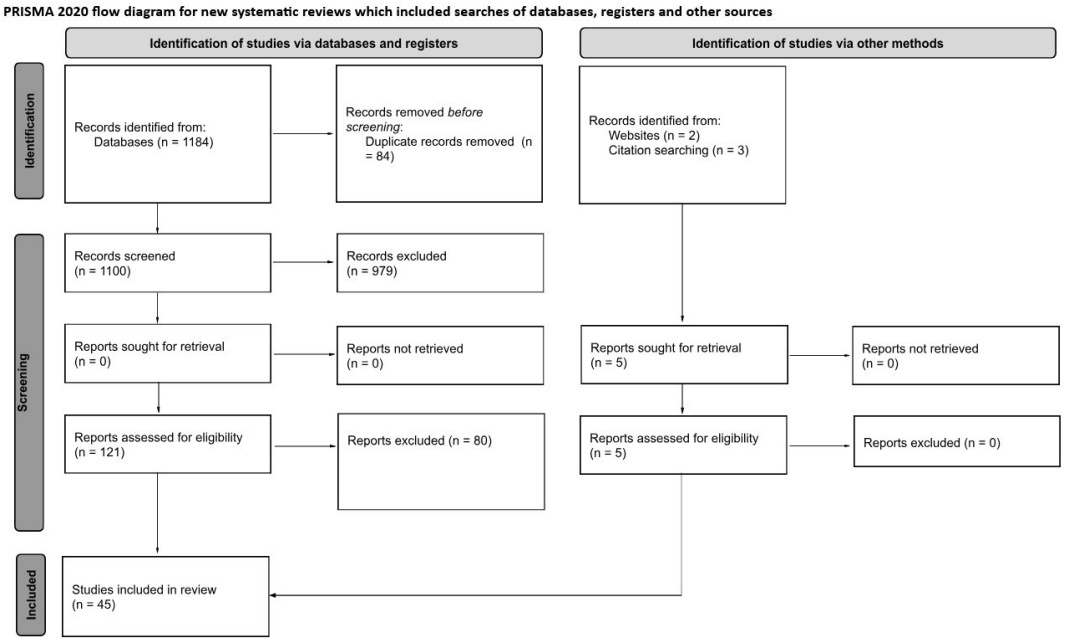

Figure 1 - PRISMA flow diagram showing the identification, screening, eligibility, and inclusion phases of the studies selected for this review.

We summarized and grouped the types of counterspeech conceptualizations, human-AI generation methods, evaluation methods, and broad findings.

2.3 Eligibility Criteria

Studies were included if they:

a) were published from January 1, 2017 onwards;
b) addressed counterspeech (or synonym terms, e.g., counternarratives and counter-arguments operationalized as counterspeech) as a direct response to counter or mitigate online hate speech;
c) employed computational or human-AI assisted methods for counterspeech generation and/or analysis, experimental/computational methods for evaluation; AI-generated counterspeech with human collaborators involved in at least one phase (development, testing, assessment, or writing);
d) included a systematic evaluation through computational metrics, human judgment, or behavioral/contextual measures;
e) were original research articles or conference proceedings.

Studies were excluded if they:

a) did not generate counterspeech, as the AI system was used only for hate speech/ counterspeech detection or classification;
b) employed AI only to generate counterspeech without human intervention;
c) conceptualized counterspeech as unrelated to online hate speech;
d) described shared task systems without humans employed in any phase of the counterspeech generation process;
e) were not empirical (theoretical papers, opinion pieces, editorials, book chapters, framework proposals without empirical validation, literature reviews, benchmark papers, and shared task proceedings);
f) were duplicate publications, only the most comprehensive or recent version was retained;
g) were not available as full-text;
h) were not written in the English language.

3. Results

This section provides a descriptive synthesis of the 45 included studies. The findings are organized according to the three objectives that guided this review: conceptualizations of counterspeech, collaborative generation methods, and evaluation approaches.

3.1 Counterspeech conceptualizations

As noted by Chung et al. (2024), computer science research often employs "counterspeech", "counternarratives", and "counterarguments" interchangeably, despite their distinct nuanced meanings in the social sciences. Benesch (2016) defines counternarratives as efforts to shift collective beliefs through institutional action, while counterspeech refers to spontaneous, citizen-based responses to hateful content (Garland et al., 2022). Counterarguments, in turn, challenge the factual or logical validity of specific claims and can function as a counterspeech strategy. Despite these distinctions, the three concepts share a common theoretical core as communicative acts responding to harmful discourse.

Many of the following taxonomies take as a base Benesch et al. (2016) original taxonomy, which takes into consideration eight non-exclusive counterspeech strategies. In Table 2, we list the main taxonomies found in the reviewed studies.

The absence of a shared taxonomy hinders cross-study comparison and prevents isolating which dimensions drive effectiveness.

3.2 Counterspeech generation methods: humans and AI collaboration

In this section, we focus on the collaborative generation strategies adopted, referred to as Human in the Loop (HITL), that reject the notion that counterspeech can or should be fully delegated to AI.

By mapping this technical landscape, we aim to identify dominant practices and emerging innovations.

In the studies reviewed, HITL manifests in different configurations. The majority of them take humans into account, reviewing AI-generated responses before publication. A smaller number of studies involve humans as trainers (Bengoetxea et al., 2024; Bonaldi et al., 2022; Fanton et al., 2021; Mun et al., 2023), providing feedback that shapes model behavior over time.

In the Author-Reviewer pipeline, an LLM generates counterspeech candidates while human experts filter and refine them (Ashida & Komachi, 2022; Fanton et al., 2021; Wang et al., 2024). This collaborative approach has been shown to drastically reduce the time experts need to create training datasets, from approximately 480 seconds per pair to 49-76 seconds (Tekiroğlu et al., 2020).

Two studies adopted HITL approaches to develop counterspeech datasets in non-English languages.

Table 2 - Counterspeech taxonomies. The classification aligns with the multidimensional framework proposed by Wang et al. (2026), which differentiates between sentence type, tone, persuasion strategies (factual information and logical reasoning) and affective strategies (influencing emotions, social bonds, peripheral cues).

Strategy	Definition	Sub-strategies	Representative studies
Cognitive	Aim to influence beliefs through factual information and logical reasoning	Presentation of facts; correction; counter-examples; pointing out contradictions; explanations; targeting stereotypes (alternate groups, alternate qualities, counterexamples, external factors, broadening, general denouncing)	Benesch et al. (2016); Fraser et al. (2023); Hassan and Alikhani (2023); Mun et al. (2023)
Affective	Aim to influence emotions, empathy, and social identification	Empathy; identification with the target; emphasizing positive qualities; emotional support	Benesch et al. (2016); Bilewicz et al. (2021); Fraser et al. (2023)
Normative	Appeal to social norms, shared values, and potential consequences	Denouncing speech; invoking norms; warning of consequences; descriptive and prescriptive norms	Benesch et al. (2016); Bilewicz et al. (2021); Fraser et al. (2023); Gupta et al. (2024)
Dialogic	Engage the speaker through interaction and reflective exchange	Critical questions; probing; clarification; dialogic contrast	Hassan and Alikhani (2023); Fraser et al. (2023)
Rhetorical	Rely on communicative style and expressive choices rather than content	Humor; tone; use of visual media; rhetorical framing; rhetorical appeal (logos, ethos, pathos)	Alyahya and Aldayel (2025); Benesch et al. (2016); Fraser et al. (2023); Mathew et al. (2019)
Transformative	Reformulate or replace harmful content rather than directly opposing it	Alternative speech; microinterventions (making harm visible, disarming, educating, seeking support)	Ashida and Komachi (2022); Lee et al. (2024)

Prasannan et al. (2025) used commercial models to generate Malayalam hate-counterspeech pairs, with two rounds of validation by native speakers: first for naturalness and coherence, then for contextual appropriateness and ethical sensitivity. Sahoo et al. (2024) developed a Hindi, Indian, and English counternarrative dataset where annotators first wrote and validated counternarratives, then an mGPT model generated additional instances, which were validated and modified before inclusion.

Chung and Bright (2024) developed an iterative adversarial collection procedure to improve counterspeech generation models. The authors conducted four rounds of data collection. In each round, a model (DynaCounter) was used to generate counterspeech, and its outputs were examined to identify failure cases. To generate adversarial examples (challenging cases where the model performed poorly), the researchers collaborated with ten trained members of a civil society organization specializing in online abuse countering.

Recent studies have explored collaborative interventions. Wu et al. (2025) first prompted the model to describe the hate speech content it was responding to, ensuring accurate understanding. A human reviewer then evaluated this description and corrected it if not accurate. In the second phase, the generated counterspeech was reviewed and manually revised by human collaborators, which could modify expressions, adjust wording, or rephrase entire sentences to better align with communicative goals. A co-writing approach was designed by Ding et al. (2025). They introduce CounterQuill, a model that educates users through reflection and structured support. This system is organized in three sessions: learning about hate and counter speech; brainstorming, identifying harmful patterns and generating initial ideas; co-writing, where users refine their counterspeech, with the system offering suggestions and revisions that users can accept, modify, or reject.

There is also a shared consensus that humans should also be employed as safeguards (Bonaldi et al., 2024; Hassan & Alikhani, 2023). In these studies, humans are usually employed to monitor experiments in real-time, ensuring that interventions stay within ethical guidelines and do not escalate hostility (Bär et al., 2024; Podolak et al., 2024).

3.3 Counterspeech evaluation methods

The generation strategies reviewed above reveal increasingly sophisticated technologies, yet sophistication does not guarantee effectiveness. In the following section, we map the evaluation landscape, distinguishing between automatic metrics, human judgments, LLM evaluations, and psychosocial measures.

3.3.1 Linguistic Metrics

A significant portion of the studies we reviewed relies on automatic linguistic metrics to evaluate AI-generated counterspeech. These metrics allow researchers to assess large volumes of generated text without the time and cost of human evaluation. In Table 3, we outline relevant metrics used to assess the quality of the generated counterspeech. Automatic measures can address structural properties of text such as length, grammaticality, vocabulary size, and expansion. However, these measures are not specific to counterspeech evaluation and were not listed.

Automatic metrics enable large-scale evaluation, but face significant limitations. BLEU and ROUGE correlate poorly with human judgments in creative tasks (Chung et al., 2024) and assume a "gold standard" reference (problematic for counterspeech, where multiple valid responses can exist). Diversity metrics capture lexical variation but not its communicative function. Toxicity detectors may miss subtle harm or over-flag legitimate content. Most critically, no automatic metric addresses whether counterspeech achieves its intended social effects.

3.3.2 Human Evaluation Metrics

Beyond automatic metrics, a subset of studies incorporates human evaluation, where trained annotators or crowdworkers rate counterspeech on critical dimensions, acting as proxies for real social media users. The assessed measures are presented in Table 4.

Similar to automated metrics, human evaluation metrics that focused on general text quality were excluded from this review, as they do not directly measure counterspeech properties.

This heterogeneity in human evaluation practices mirrors the methodological fragmentation already documented, reinforcing the need for shared standards that enable systematic comparison across studies.

Table 3 - Automatic linguistic metrics for counterspeech.

Dimension	Description	Representative Metrics	Relevance
Semantic metrics	Measure semantic overlap between generated counterspeech and human-written reference texts	BLEU, ROUGE, METEOR, GLEU; BERTScore (Zhang et al., 2019; Hengle et al., 2024); BARTScore (Hengle et al., 2025); MoverScore; Semantic Similarity (Gupta et al., 2023)	Indicates how closely AI output approximates natural human language
Linguistic quality and readability	Assess grammatical correctness, coherence, and accessibility of the generated text	GRUEN (Zhu & Bhat, 2021); Flesch Reading Ease (Cima et al., 2025)	Influences perceived credibility, clarity, and the likelihood of message comprehension by target audiences
Diversity and novelty	Evaluate lexical and structural variety to avoid repetitive or generic outputs	Self-BLEU; Entropy; Novelty (Wang & Wan, 2018; Tekiroglu et al., 2022; Fanton et al., 2022; Gupta et al., 2023); Repetition Rate (Zubiaga et al., 2024)	Repetitive language may reduce perceived authenticity and persuasiveness in interpersonal or public health interventions
Toxicity and safety	Detect harmful, aggressive, or toxic content in generated counterspeech	Toxicity scores (Hengle et al., 2024; Wang et al., 2024b; Chung and Bright, 2024); Safety Guardrails (Bonaldi et al., 2024)	Essential for ensuring that counterspeech does not unintentionally perpetuate harm or escalate conflict
Effectiveness	Quantify the extent to which counterspeech counters hateful content through refutation, factual support, or reduction in hate intensity	Success Rate of Countering (Jiang et al., 2025); Pro/Con scores, Argument Quality (Hengle et al., 2024); PD-Score (Hengle et al., 2025); Hate Mitigation Value (Wang et al., 2024b)	Aim to determine whether AI-generated messages can meaningfully reduce hate or change attitudes

Table 4 - Human evaluation metrics of AI-generated counterspeech.

Dimension	Definition	Representative Ratings	Relevance
Offensiveness	Whether the generated response contains content that may cause harm or be perceived as offensive	Offensiveness (Ashida & Komachi, 2022)	Most generated text rated as safe, though variability in human perception exists; some outputs deemed offensive
Aggressiveness	Degree of confrontational, inflammatory, or hostile content, including personal attacks	Aggressiveness (Hengle et al., 2025); emotional tone assessment (Prasannan et al., 2025)	Low aggressiveness is prioritized to avoid escalating interactions; cultural appropriateness matters (e.g., Malayalam community context)
Politeness / Suitability	Adherence to social conventions of respect and courtesy; whether the response follows civil rules and avoids attacking the hater personally	Politeness (Song et al., 2024); Suitability (Bonaldi et al., 2024); Suitableness (Chung et al., 2020, 2021)	High politeness serves as a proxy for safe, socially acceptable responses and may reduce defensive reactions from recipients
Relevance	Whether the response addresses the correct topic and target group identified in the hate speech	Relevance (Bonaldi et al., 2024); contextual relevance (Hengle et al., 2024; Kumar et al., 2025)	High relevance increases persuasiveness; attention regularization techniques have shown improvements
Specificity	How well the response engages with the specific content of the hate speech, including logical strength, precision, and depth	Cogency (Bonaldi et al., 2024); correspondence (Lee et al., 2024); relatedness, specificity, richness, coherence (Baez Santamaria et al., 2024)	Specific, detailed responses are more likely to be perceived as thoughtful and effective than generic ones
Stance	Degree to which the response explicitly or implicitly opposes the hateful content	Stance: agreeing, neutral, disagreeing (Ashida & Komachi, 2022)	Most AI models demonstrate a clear stance against hate, which is essential for counterspeech interventions
Informativeness	Extent to which the response provides facts, data, or contextual information that challenges hate speech claims	Informativeness (Ashida & Komachi, 2022; Bonaldi et al., 2024; Chung et al., 2020, 2021); phrase guidance (Lee et al., 2024)	High informativeness does not guarantee factual accuracy; hallucinations (factual errors) remain a concern
Factuality	The accuracy and reliability of real-world data, statistics, and verifiable evidence incorporated into a counterspeech response	Truthfulness (Wang et al., 2024); Factuality (Wilk et al., 2025)	Factual inconsistencies or errors can compromise the authority of the source and lower the perceived persuasiveness of counterspeech
Effectiveness	The actual or perceived social, psychological, or persuasive impact of the counterspeech intervention	Effectiveness (Baez Santamaria et al., 2024; Zheng et al., 2023); persuasiveness (Wilk et al., 2025); improvement in speaker's hateful perception (Lee et al., 2024); argumentative effectiveness (Hengle et al., 2024; Kumar et al., 2025)	Effectiveness is conceptualized in multiple ways (emotional, persuasive, argumentative), reflecting its complexity as a primary outcome of interest for psychological interventions

3.3.3 LLM-as-a-judge

The limitations of conventional evaluation methods have prompted researchers to explore an alternative: using LLMs as evaluators (referred to as LLM-as-a-judge), which employs models to assess generated text directly. It achieves stronger alignment with human assessments than traditional metrics (Hengle et al., 2025), while offering the scalability of automated approaches. Commercial models can be used for this purpose (Zubiaga et al., 2024), but new research adopts specialized evaluation models, including PandaLM (Bennie et al., 2024) and JudgeLM (Zhu et al., 2023).

Chung et al. (2024) have criticized the validity and reliability of using LLMs as NLG evaluators. Specifically, there is an issue with how automated metrics are evaluated. In response, Hengle et al. (2025) developed CSeval, a dataset of human judgments on counterspeech from five AI models across four dimensions: contextual relevance, aggressiveness, argument coherence, and suitableness. Using these ratings, they built Auto-CSEval, an LLM-based evaluator employing Chain of Thought (CoT) prompting. This method consists of asking the model to reason step-by-step about each quality dimension

before producing a score. Auto-CSEval outperformed both traditional metrics and other LLM evaluators in correlation with human judgments.

While LLM judges show improved human alignment, research identifies several critical biases: other works (Bennie et al., 2025) underline how LLM judges like JudgeLM (Zhu et al., 2023) frequently assign higher scores to AI-generated responses than to human-preferred or human-edited ones. They suggest that JudgeLM may favor AI-generated text due to stylistic and syntactic features, rather than substantive quality such as logical or factual accuracy.

3.3.4 Psychosocial Metrics

Unlike evaluation, psychosocial metrics investigate the outcomes of generated counterspeech. The following paragraph presents the examined measures.

- User engagement. Studies agree that counterspeech presence increases discussion engagement. Podolak et al. (2024) tested the effectiveness of LLM-generated (GPT-4) counterspeech on Twitter. The counterspeech was relevant and accurate in responses, since the model could retrieve information from 23 verified articles. Interventions with LLM-generated responses significantly reduced user engagement with the hate speech tweets. For original tweets that had received at least ten views, engagement decreased by over 20. The study of Wu et al. (2025) found, across three studies with 809 participants, that people were more willing to engage when counterspeech was present, regardless of the source. Empathy-based counterspeech significantly increased user engagement intention compared to fact-based counterspeech. Moreover, participants were more willing to engage with AI-generated counterspeech when the AI identity was not disclosed. When AI identity was revealed, trust in both human and AI counterspeakers decreased.
- Post deletion and toxic language. Bär et al. (2024) measured behavioural changes of 2,664 users after being randomly assigned to contextualized and non-contextualized counterspeech on Twitter/X. Post deletion rate tracked whether the author of the original hate speech removed their post; future toxicity levels measured whether the perpetrator adopts less toxic language in subsequent posts, offering a longitudinal indicator of possible educational or deterrent effects. Non-contextualized counterspeech employing a warning of consequences strategy significantly reduced online hate speech. Contextualized counterspeech was proven ineffective and even backfired. The authors attribute the backfire to perceived deception from the undisclosed AI identity, not the counterspeech strategy itself.
- Reduction of verbal aggression. Bilewicz et al. (2021) estimated the impact of counterspeech interventions on users' engagement in personal attacks or acts of verbal violence within a Reddit community. The study found a significant reduction in verbal aggression following the intervention: the proportion of verbally aggressive comments decreased from pre to post-intervention. Norm-inducing, disapproval, and empathizing counterspeech strategies were all effective in reducing verbal aggression compared to the control group. However, the norm-inducing intervention did not differ significantly in effectiveness from either the disapproval or the empathizing interventions.
- Counterspeech participation. Wang et al. (2026) exposed 52 bystanders to their chatbot Civilbot, designed according to their three-dimensional framework (sentence type, tone, strategic intent). It was perceived as credible and normatively appropriate, but its persuasiveness was limited by superficial reasoning that was too AI-like. Cognitive strategies outperformed affective ones on perceived quality and acceptance, with positive tone amplifying this effect, while negative tone risked normalizing aggression and raising ethical concerns (Cecconi et al., 2020). Questions could stimulate reflection when genuine, but became hostile if rhetorical. Participants distinguished between their expectations of Civilbot (objective, normative) and their own human behavior (using a negative tone for emotional venting). Mismatches between context and strategy weakened the impact.
- Persuasiveness. Alyahya & Aldayel (2025) found that counterspeech employing reason as a persuasion mode is associated with obtaining more supportive replies than those using emotion or credibility. Interestingly, in their study, humans tend to lean towards reason, while machine-generated counterspeech tends to exhibit an emotional persuasion mode. Mun et al. (2023) found that strategy effectiveness varies: providing alternative, higher-quality information was chosen as convincing twice as frequently as appealing to a different group's perspective, even though both require similar levels of reasoning. Broadening was the preferred strategy, while generic denouncement was the least preferred. Cima et al. (2025) demonstrated that contextualized counterspeech significantly outperformed generic counterspeech in terms of persuasiveness.

Notably, Hong et al. (2024) developed conversation outcome classifiers that predict what happens after counterspeech is deployed. Conversation incivility captures the overall tone of the discussion, based on the number of civil and uncivil comments (Poggi et al.,

2013) and the unique authors contributing each. Hater reentry tracks whether the original hate speech author returns to the conversation and whether their reentry is hateful or non-hateful. However, this analysis was conducted retrospectively on a benchmark dataset of Reddit conversations and not in a real intervention.

A closer look at psychosocial findings reveals both convergences and divergences. Studies in real contexts capture actual behaviors, whereas controlled context studies can face an intention-behavior gap. Additionally, real context studies can measure long-term network dynamics, where counterspeech could dampen the spread of negativity across a user's neighboring accounts.

Lastly, almost all studies assume that the target of counterspeech is the hater, but several findings (Bilewicz et al., 2021; Wu et al., 2026) suggest that the main effect may be on bystanders, implying designs targeted within the community rather than on individual change.

4. Discussion and conclusions

Counterspeech is a social intervention aimed at countering hate speech; its effectiveness depends on prosocial effects on both the hater and bystanders, such as changes in behavior, norms, attitudes, and empathy toward victims. When AI generates such interventions, key questions arise: what makes counterspeech effective? How does an artificial sender affect perception? And what happens to digital communities when machines act as social actors? However, reducing the effectiveness to the machine alone is an oversimplification.

The preceding sections have presented our findings in response to the three objectives that guided this review: conceptualizations of counterspeech, collaborative approaches to generation and evaluation, and alignment between goals and measures. In this discussion, we synthesize the findings in our review, highlight research gaps, and propose future directions.

The explorative findings of this review reveal a theoretical and methodological fragmented picture. Conceptualizations vary widely; taxonomies often blend tone, rhetorical strategies, and normative functions. Rather than treating this heterogeneity as a limitation alone, we propose that future research could leverage these dimensions to be distinguished and operationalized. This review proposed a six-dimension taxonomy (see Table 2). However, this taxonomy is purely illustrative. Moving forward, we argue that the field would benefit from a collective effort to establish a shared conceptual framework. Generation methods differ across studies, and evaluation metrics are often incomparable, limiting synthesis. Automated and human metrics typically focus on formal text properties, failing to capture psychological and behavioral outcomes.

Human evaluations attempt to fill this gap, but are usually applied to decontextualized texts; only a few studies assess counterspeech in real-world contexts (Bär et al., 2024; Bilewicz et al., 2021; Podolak et al., 2024). There is a structural disconnect between LLM capabilities (now more context-aware and nuanced) and psychosocial implementation. Traditional metrics correlate poorly with human judgment (Bonaldi et al., 2024; Cima et al., 2025), reflecting a systematic misalignment between objectives and measures. To address the latter, researchers have turned to LLMs as evaluators, achieving better scalability and correlation with human judgment (Hengle et al., 2025). However, validity concerns remain (Chung et al., 2024). JudgeLM often favors AI-generated over human-preferred responses (Bennie et al., 2025), suggesting susceptibility to stylistic cues rather than substantive quality. Future directions should include defining and reliably measuring effective counterspeech through integrated computational and psychosocial expertise, and expanding non-English datasets for culturally aware evaluation.

We argue that computational metrics alone are insufficient for evaluating counterspeech effectiveness. They should be complemented by psychosocial measures that can target different mechanisms of change (normative, cognitive or affective), therefore strategy-specific. Moreover, LLM judges cannot replace human evaluators as they can't assess psychosocial effectiveness. We therefore recommend a hybrid pipeline where LLMs serve as initial evaluators and humans (ideally real users) remain the decisive standard.

Research on the social identity of chatbots as counterspeakers remains scarce. Recent studies (for instance, Wang et al., 2026) show differing expectations for AI versus human counterspeakers, but systematic evaluation of whether users accept AI-generated counterspeech is lacking, leaving its effectiveness unclear. Moreover, identity transparency is also an issue. Wu et al. (2025) found that nondisclosure increases engagement, while disclosure reduces it. Bär et al. (2024) observed that revealing AI involvement led to backfire effects and conversational aggression. Nondisclosure may boost short-term engagement but risks eroding trust; disclosure may provoke hostility. Future research should address this ethical problem by investigating how different degrees of transparency can be successfully implemented in real life context.

Human in the Loop (HITL) approaches vary: humans as reviewers filtering AI outputs (Ashida & Komachi, 2022; Fanton et al., 2021), humans as trainers providing feedback to improve models over time (Bengoetxea et al., 2024; Bonaldi et al., 2022), as in an Author-Reviewer pipeline (Tekiroğlu et al., 2020; Wu

et al., 2025), and leveraging human expertise to identify model weaknesses and generate targeted training data (Chung & Bright, 2024). Genuine collaboration in counterspeech writing has only been employed by Ding et al. (2025) with their chatbot CounterQuill. The most common human role is still that of a database curator, reviewer, or proofreader. There is a lack of collaboration evaluation in terms of the psychological and educational effectiveness of counterspeech. We align with Chung et al. (2024) in stating that counterspeech generation tools should be deployed as suggestion tools to assist humans rather than replace them.

Future research should prioritize systematizing effective counterspeech through a unified interdisciplinary framework, essential for moving from fragmented efforts toward cumulative knowledge. Unexplored outcomes also deserve attention, including cognitive and affective changes, across platforms and communities. Additionally, research should examine how users perceive AI counterspeakers across contexts and cultures, which identities enhance or reduce effectiveness, how human-AI dynamics shape community norms, and how to balance transparency with impact, all questions essential for ethically grounded system design. A priority should be collaborative alternatives to full automation: AI systems that support human empowerment, serving as tools for augmentation rather than substitution. As AI takes on social roles in social media, we must promote not only technical progress but also morally responsible and informed use.

Finally, we acknowledge several limitations of this study. Given the rapid evolution of AI research, we chose to include preprints. Although these have not yet undergone formal peer review, we carefully examined their content and relevance for the purposes of this scoping review. Furthermore, the substantial heterogeneity in conceptualizations, generation methods, and evaluation metrics across the 45 included studies precluded any quantitative synthesis or meta-analysis. Consequently, our findings should be interpreted as a mapping of the field rather than as a definitive assessment of effectiveness.

References

Ashida, M., & Komachi, M. (2022). Towards Automatic Generation of Messages Countering Online Hate Speech and Microaggressions. In K. Narang, A. Mostafazadeh Davani, L. Mathias, B. Vidgen, & Z. Talat (A c. Di), Proceedings of the Sixth Workshop on Online Abuse and Harms (WOAH) (pp. 11–23). Association for Computational Linguistics.

Baez Santamaria, S., Gomez Adorno, H., & Markov, I. (2024). Contextualized Graph Representations for Generating Counter-Narratives against Hate Speech. In Y. Al-Onaizan, M. Bansal, & Y.-N. Chen (A c. Di), Findings of the Association for Computational Linguistics: EMNLP 2024 (pp. 7664–7674). Association for Computational Linguistics.

Bandura, A., Barbaranelli, C., Caprara, G. V., & Pastorelli, C. (1996). Mechanisms of moral disengagement in the exercise of moral agency. Journal of Personality and Social Psychology, 71, 364–374.

Bär, D., Maarouf, A., & Feuerriegel, S. (2024). Generative AI may backfire for counterspeech (arXiv:2411.14986). arXiv.

Benesch, S. (2016). Counterspeech on Twitter: A Field Study.

Bennie, M., Zhang, D., Xiao, B., Cao, J., Liu, C. X., Meng, J., & Tripp, A. (2025). PANDA -- Paired Anti-hate Narratives Dataset from Asia: Using an LLM-as-a-Judge to Create the First Chinese Counterspeech Dataset (arXiv:2501.00697). arXiv.

Bilewicz, M., Tempska, P., Leliwa, G., Dowgiałło, M., Tańska, M., Urbaniak, R., & Wroczyński, M. (2021). Artificial intelligence against hate: Intervention reducing verbal aggression in the social network environment. Aggressive Behavior, 47(3), 260–266.

Bonaldi, H., Damo, G., Ocampo, N. B., Cabrio, E., Villata, S., & Guerini, M. (2024). Is Safer Better? The Impact of Guardrails on the Argumentative Strength of LLMs in Hate Speech Countering (arXiv:2410.03466). arXiv.

Bonaldi, H., Dellantonio, S., Tekiroğlu, S. S., & Guerini, M. (2022). Human-Machine Collaboration Approaches to Build a Dialogue Dataset for Hate Speech Countering. In Y. Goldberg, Z. Kozareva, & Y. Zhang (A c. Di), Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing (pp. 8031–8049). Association for Computational Linguistics.

Cecconi, C., Poggi, I., & D'Errico, F. (2020). Schadenfreude: malicious joy in social media interactions. Frontiers in Psychology, 11, 558282.

Cima, L., Miaschi, A., Trujillo, A., Avvenuti, M., Dell'Orletta, F., & Cresci, S. (2025). Contextualized Counterspeech: Strategies for Adaptation, Personalization, and Evaluation. WWW - Proc. ACM Web Conf., 5022–5033.

Chung, Y.-L., Abercrombie, G., Enock, F., Bright, J., & Rieser, V. (2024). Understanding Counterspeech for Online Harm Mitigation. ArXiv.org.

Chung, Y.-L., & Bright, J. (2024). On the Effectiveness of Adversarial Robustness for Abuse Mitigation

with Counterspeech. In K. Duh, H. Gomez, & S. Bethard (A c. Di), Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers) (pp. 6988–7002). Association for Computational Linguistics.

Darley, J. M., & Latane, B. (1968). Bystander intervention in emergencies: Diffusion of responsibility. Journal of Personality and Social Psychology, 8(4, Pt.1), 377–383.

Ding, X., Ping, K., Gunturi, U. S., Carik, B., Stil, S., Wilhelm, L. T., Daryanto, T., Hawdon, J., Lee, S. W., & Rho, E. H. (2025). Designing Human-AI Collaboration to Support Learning in Counterspeech Writing (arXiv:2410.03032). arXiv.

Fanton, M., Bonaldi, H., Tekiroğlu, S. S., & Guerini, M. (2021). Human-in-the-Loop for Data Collection: A Multi-Target Counter Narrative Dataset to Fight Online Hate Speech. In C. Zong, F. Xia, W. Li, & R. Navigli (A c. Di), 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) (pp. 3226–3240). Association for Computational Linguistics.

Fortuna, P., & Nunes, S. (2018). A Survey on Automatic Detection of Hate Speech in Text. ACM Computing Surveys, 51(4), 1–30.

Garland, J., Ghazi□Zahedi, K., Young, J. G., Hébert□Dufresne, L., & Galesic, M. (2022). Impact and dynamics of hate and counter□speech online. EPJ Data Science, 11(3):3.

Hassan, S., & Alikhani, M. (2023). DisCGen: A Framework for Discourse-Informed Counterspeech Generation (arXiv:2311.18147). arXiv.

Hengle, A., Padhi, A. K., Bandhakavi, A., & Chakraborty, T. (2025). CSEval: Towards Automated, Multi-Dimensional, and Reference-Free Counterspeech Evaluation using Auto-Calibrated LLMs. In L. Chiruzzo, A. Ritter, & L. Wang (A c. Di), 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) (pp. 5402–5419). Association for Computational Linguistics.

Hengle, A., Padhi, A., Singh, S., Bandhakavi, A., Akhtar, M. S., & Chakraborty, T. (2024). Intent-conditioned and Non-toxic Counterspeech Generation using Multi-Task Instruction Tuning with RLAIF. In K. Duh, H. Gomez, & S. Bethard (A c. Di), Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers) (pp. 6716–6733). Association for Computational Linguistics.

Hong, L., Luo, P., Blanco, E., & Song, X. (2024). Outcome-Constrained Large Language Models for Countering Hate Speech (arXiv:2403.17146). arXiv.

Horta Ribeiro, M., Hosseinmardi, H., West, R., & Watts, D. J. (2023). Deplatforming did not decrease Parler users' activity on fringe social media. PNAS Nexus, 2(3).

Kumar, A., Bandhakavi, A., & Chakraborty, T. (2025). Counterspeech the ultimate shield! Multi-Conditioned Counterspeech Generation through Attributed Prefix Learning (arXiv:2505.11958). arXiv.

Mathew, B., Saha, P., Tharad, H., Rajgaria, S., Singhania, P., Maity, S. K., Goyal, P., & Mukherje, A. (2019). Thou shalt not hate: Countering Online Hate Speech (arXiv:1808.04409). arXiv.

Mun, J., Allaway, E., Yerukola, A., Vianna, L., Leslie, S.-J., & Sap, M. (2023). Beyond Denouncing Hate: Strategies for Countering Implied Biases and Stereotypes in Language (arXiv:2311.00161). arXiv.

Podolak, J., Łukasik, S., Balawender, P., Ossowski, J., Piotrowski, J., Bakowicz, K., & Sankowski, P. (2024). LLM generated responses to mitigate the impact of hate speech. In Y. Al-Onaizan, M. Bansal, & Y.-N. Chen (A c. Di), Findings of the Association for Computational Linguistics: EMNLP 2024 (pp. 15860–15876). Association for Computational Linguistics.

Poggi, I., D'Errico, F., & Vincze, L. (2013). Comments by words, face and body. Journal on Multimodal User Interfaces, 7(1), 67-78.

Prasannan, P., Kumaresan, P. K., Rajiakodi, S., Subalalitha, C. N., & Chakravarthi, B. R. (2025). Counter-speech generation for homophobic and transphobic social media content in Malayalam. Social Network Analysis and Mining, 15(1), 87.

Sahoo, N. R., Beria, G. P., & Bhattacharyya, P. (2024). IndicCONAN: A Multilingual Dataset for Combating Hate Speech in Indian Context. Proceedings of the AAAI Conference on Artificial Intelligence, 38(20), 22313–22321.

Sportelli, C., Cicirelli, P. G., Paciello, M., Corbelli, G., & D'Errico, F. (2025). "Let's Make the Difference!" Promoting Hate Counter□Speech in Adolescence Through Empathy and Digital Intergroup Contact. Journal of Community & Applied Social Psychology, 35(1), e70028

Tekiroğlu, S. S., Chung, Y.-L., & Guerini, M. (2020). Generating Counter Narratives against Online Hate

Speech: Data and Strategies. In D. Jurafsky, J. Chai, N. Schluter, & J. Tetreault (A c. Di), Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (pp. 1177–1190). Association for Computational Linguistics.

Tricco, AC, Lillie, E, Zarin, W, O'Brien, KK, Colquhoun, H, Levac, D, Moher, D, Peters, MD, Horsley, T, Weeks, L, Hempel, S et al. PRISMA extension for scoping reviews (PRISMA-ScR): checklist and explanation. Ann Intern Med. 2018,169(7):467-473. doi: 10.7326/M18-0850

Wachs, S., Valido, A., Espelage, D. L., Castellanos, M., Wettstein, A., & Bilz, L. (2023). The relation of classroom climate to adolescents' countering hate speech via social skills: A positive youth development perspective. Journal of Adolescence, 95(6), 1127–1139.

Wang, H., Pan, Y., Song, X., Zhao, X., Hu, M., & Zhou, B. (2024). F^2RL: Factuality and Faithfulness Reinforcement Learning Framework for Claim-Guided Evidence-Supported Counterspeech Generation. In Y. Al-Onaizan, M. Bansal, & Y.-N. Chen (A c. Di), Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing (pp. 4457–4470). Association for Computational Linguistics.

Wang, M., Ma, S., Li, N., Zhang, P., Li, C., Gu, N., & Lu, T. (2026). Echoes of Norms: Investigating Counterspeech Bots' Influence on Bystanders in Online Communities (arXiv:2603.03687). arXiv.

Wu, C., Wang, Y., Zhang, Y., Wang, H., & Pang, Y. (2025). Confront hate with AI: How AI-generated counter speech helps against hate speech on social Media? Telematics and Informatics, 101, 102304.

Zhu, W., & Bhat, S. (2021). Generate, Prune, Select: A Pipeline for Counterspeech Generation against Online Hate Speech (arXiv:2106.01625). arXiv.

Zhu, L., Wang, X., & Wang, X. (2023, October 26). JudgeLM: Fine-tuned Large Language Models are Scalable Judges. arXiv.Org.

Zubiaga, I., Soroa, A., & Agerri, R. (2024). A LLM-based Ranking Method for the Evaluation of Automatic Counter-Narrative Generation. In Y. Al-Onaizan, M. Bansal, & Y.-N. Chen (A c. Di), Findings of the Association for Computational Linguistics: EMNLP 2024 (pp. 9572–9585). Association for Computational Linguistics.

Appendix A

Alyahya, G., & Aldayel, A. (2025). *Hatred Stems from Ignorance! Distillation of the Persuasion Modes in Countering Conversational Hate Speech* (arXiv:2403.15449). arXiv. https://doi.org/10.48550/arXiv.2403.15449

Ashida, M., & Komachi, M. (2022). Towards Automatic Generation of Messages Countering Online Hate Speech and Microaggressions. In K. Narang, A. Mostafazadeh Davani, L. Mathias, B. Vidgen, & Z. Talat (Eds.), *Proceedings of the Sixth Workshop on Online Abuse and Harms (WOAH)* (pp. 11–23). Association for Computational Linguistics. https://doi.org/10.18653/v1/2022.woah-1.2

Baez Santamaria, S., Gomez Adorno, H., & Markov, I. (2024). Contextualized Graph Representations for Generating Counter-Narratives against Hate Speech. In Y. Al-Onaizan, M. Bansal, & Y.-N. Chen (Eds.), *Findings of the Association for Computational Linguistics: EMNLP 2024* (pp. 7664–7674). Association for Computational Linguistics. https://doi.org/10.18653/v1/2024.findings-emnlp.450

Bär, D., Maarouf, A., & Feuerriegel, S. (2024). *Generative AI may backfire for counterspeech*

(arXiv:2411.14986). arXiv. https://doi.org/10.48550/arXiv.2411.14986

Bengoetxea, J., Chung, Y.-L., Guerini, M., & Agerri, R. (2024). Basque and Spanish Counter Narrative Generation: Data Creation and Evaluation. In N. Calzolari, M.-Y. Kan, V. Hoste, A. Lenci, S. Sakti, & N. Xue (Eds.), *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)* (pp. 2132–2141). ELRA and ICCL. https://aclanthology.org/2024.lrec-main.192/

Bennie, M., Zhang, D., Xiao, B., Cao, J., Liu, C. X., Meng, J., & Tripp, A. (2025). *PANDA -- Paired Anti-hate Narratives Dataset from Asia: Using an LLM-as-a-Judge to Create the First Chinese Counterspeech Dataset* (arXiv:2501.00697). arXiv. https://doi.org/10.48550/arXiv.2501.00697

Bilewicz, M., Tempska, P., Leliwa, G., Dowgiałło, M., Tańska, M., Urbaniak, R., & Wroczyński, M. (2021). Artificial intelligence against hate: Intervention reducing verbal aggression in the social network environment. *Aggressive Behavior*, *47*(3), 260–266. https://doi.org/10.1002/ab.21948

Bonaldi, H., Attanasio, G., Nozza, D., & Guerini, M. (2023). Weigh Your Own Words: Improving Hate Speech Counter Narrative Generation via Attention Regularization. In Y.-L. Chung, H. Bonaldi, G. Abercrombie, & M. Guerini (Eds.), *Proceedings of the 1st Workshop on CounterSpeech for Online Abuse (CS4OA)* (pp. 13–28). Association for Computational Linguistics. https://aclanthology.org/2023.cs4oa-1.2/

Bonaldi, H., Damo, G., Ocampo, N. B., Cabrio, E., Villata, S., & Guerini, M. (2024). *Is Safer Better? The Impact of Guardrails on the Argumentative Strength of LLMs in Hate Speech Countering* (arXiv:2410.03466). arXiv. https://doi.org/10.48550/arXiv.2410.03466

Bonaldi, H., Dellantonio, S., Tekiroğlu, S. S., & Guerini, M. (2022). Human-Machine Collaboration Approaches to Build a Dialogue Dataset for Hate Speech Countering. In Y. Goldberg, Z. Kozareva, & Y. Zhang (Eds.), *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing* (pp. 8031–8049). Association for Computational Linguistics. https://doi.org/10.18653/v1/2022.emnlp-main.549

Chen, G., Cheng, L., Luu, A. T., & Bing, L. (2024). Exploring the Potential of Large Language Models in Computational Argumentation. In L.-W. Ku, A. Martins, & V. Srikumar (Eds.), *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)* (pp.

2309–2330). Association for Computational Linguistics. https://doi.org/10.18653/v1/2024.acl-long.126 Chung, Y.-L., & Bright, J. (2024). On the Effectiveness of Adversarial Robustness for Abuse Mitigation with

Counterspeech. In K. Duh, H. Gomez, & S. Bethard (Eds.), *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)* (pp. 6988–7002). Association for Computational Linguistics. https://doi.org/10.18653/v1/2024.naacl-long.386

Chung, Y.-L., Tekiroğlu, S. S., & Guerini, M. (2020). Italian Counter Narrative Generation to Fight Online Hate Speech. In J. Monti, F. Dell'Orletta, & F. Tamburini (Eds.), *Proceedings of the Seventh Italian Conference on Computational Linguistics (CLiC-it 2020)* (pp. 82–88). CEUR Workshop Proceedings. https://aclanthology.org/2020.clicit-1.15/

Chung, Y.-L., Tekiroğlu, S. S., & Guerini, M. (2021). Towards Knowledge-Grounded Counter Narrative Generation for Hate Speech. *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, 899–914. https://doi.org/10.18653/v1/2021.findings-acl.79

Cima, L., Miaschi, A., Trujillo, A., Avvenuti, M., Dell'Orletta, F., & Cresci, S. (2025). Contextualized Counterspeech: Strategies for Adaptation, Personalization, and Evaluation. *WWW - Proc. ACM Web Conf.*, 5022–5033. https://doi.org/10.1145/3696410.3714507

Das, M., Pandey, S., Sethi, S., Saha, P., & Mukherjee, A. (2024). Low-Resource Counterspeech Generation for Indic Languages: The Case of Bengali and Hindi. In Y. Graham & M. Purver (Eds.), *Findings of the Association for Computational Linguistics: EACL 2024* (pp. 1601–1614). Association for Computational Linguistics. https://doi.org/10.18653/v1/2024.findings-eacl.111

Ding, X., Ping, K., Gunturi, U. S., Carik, B., Stil, S., Wilhelm, L. T., Daryanto, T., Hawdon, J., Lee, S. W., & Rho, E. H. (2025). *Designing Human-AI Collaboration to Support Learning in Counterspeech Writing* (arXiv:2410.03032). arXiv. https://doi.org/10.48550/arXiv.2410.03032

Doğanç, M., & Markov, I. (2023). From Generic to Personalized: Investigating Strategies for Generating Targeted Counter Narratives against Hate Speech. In Y.-L. Chung, H. Bonaldi, G. Abercrombie, & M. Guerini (Eds.), *Proceedings of the 1st Workshop on CounterSpeech for Online Abuse (CS4OA)* (pp. 1–12). Association for Computational Linguistics. https://aclanthology.org/2023.cs4oa-1.1/

Fanton, M., Bonaldi, H., Tekiroğlu, S. S., & Guerini, M. (2021). Human-in-the-Loop for Data Collection: A Multi-Target Counter Narrative Dataset to Fight Online Hate Speech. In C. Zong, F. Xia, W. Li, & R. Navigli (Eds.), *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)* (pp. 3226–3240). Association for Computational Linguistics.

https://doi.org/10.18653/v1/2021.acl-long.250

Furman, D., Torres, P., Rodríguez, J., Letzen, D., Martinez, M., & Alemany, L. (2023). High-quality argumentative information in low resources approaches improve counter-narrative generation. In H. Bouamor, J. Pino, & K. Bali (Eds.), *Findings of the Association for Computational Linguistics: EMNLP 2023* (pp. 2942–2956). Association for Computational Linguistics. https://doi.org/10.18653/v1/2023.findings-emnlp.194

Gupta, R., Desai, S., Goel, M., Bandhakavi, A., Chakraborty, T., & Akhtar, Md. S. (2023). Counterspeeches up my sleeve! Intent Distribution Learning and Persistent Fusion for Intent-Conditioned Counterspeech Generation. In A. Rogers, J. Boyd-Graber, & N. Okazaki (Eds.), *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)* (pp. 5792–5809). Association for Computational Linguistics. https://doi.org/10.18653/v1/2023.acl-long.318

Hassan, S., & Alikhani, M. (2023). *DisCGen: A Framework for Discourse-Informed Counterspeech Generation* (arXiv:2311.18147). arXiv. https://doi.org/10.48550/arXiv.2311.18147

Hengle, A., Padhi, A. K., Bandhakavi, A., & Chakraborty, T. (2025). CSEval: Towards Automated, Multi-Dimensional, and Reference-Free Counterspeech Evaluation using Auto-Calibrated LLMs. In L. Chiruzzo, A. Ritter, & L. Wang (Eds.), *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)* (pp. 5402–5419). Association for Computational Linguistics. https://doi.org/10.18653/v1/2025.naacl-long.279

Hengle, A., Padhi, A., Singh, S., Bandhakavi, A., Akhtar, M. S., & Chakraborty, T. (2024). Intent-conditioned and Non-toxic Counterspeech Generation using Multi-Task Instruction Tuning with RLAIF. In K. Duh, H. Gomez, & S. Bethard (Eds.), *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)* (pp. 6716–6733). Association for Computational Linguistics. https://doi.org/10.18653/v1/2024.naacl-long.374

Hong, L., Luo, P., Blanco, E., & Song, X. (2024). *Outcome-Constrained Large Language Models for Countering Hate Speech* (arXiv:2403.17146). arXiv. https://doi.org/10.48550/arXiv.2403.17146

Jiang, S., Tang, W., Chen, X., Tang, R., Wang, H., & Wang, W. (2025). ReZG: Retrieval-augmented zero-shot counter narrative generation for hate speech. *Neurocomputing*, *620*, 129140. https://doi.org/10.1016/j.neucom.2024.129140

Kumar, A., Bandhakavi, A., & Chakraborty, T. (2025). *Counterspeech the ultimate shield! Multi-Conditioned Counterspeech Generation through Attributed Prefix Learning* (arXiv:2505.11958). arXiv. https://doi.org/10.48550/arXiv.2505.11958

Lee, S., Jung, D., Park, C., Lee, S., & Lim, H. (2024). *Alternative Speech: Complementary Method to Counter-Narrative for Better Discourse* (arXiv:2401.14616). arXiv. https://doi.org/10.48550/arXiv.2401.14616

Mun, J., Allaway, E., Yerukola, A., Vianna, L., Leslie, S.-J., & Sap, M. (2023). *Beyond Denouncing Hate: Strategies for Countering Implied Biases and Stereotypes in Language* (arXiv:2311.00161). arXiv. https://doi.org/10.48550/arXiv.2311.00161

Piot, P., & Parapar, J. (2025). Decoding Hate: Exploring Language Models' Reactions to Hate Speech. *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)*, 973–990. https://doi.org/10.18653/v1/2025.naacl-long.45

Podolak, J., Łukasik, S., Balawender, P., Ossowski, J., Piotrowski, J., Bakowicz, K., & Sankowski, P. (2024). LLM generated responses to mitigate the impact of hate speech. In Y. Al-Onaizan, M. Bansal, & Y.-N. Chen (Eds.), *Findings of the Association for Computational Linguistics: EMNLP 2024* (pp. 15860–15876). Association for Computational Linguistics. https://doi.org/10.18653/v1/2024.findings-emnlp.931

Prasannan, P., Kumaresan, P. K., Rajiakodi, S., Subalalitha, C. N., & Chakravarthi, B. R. (2025). Counter-speech generation for homophobic and transphobic social media content in Malayalam. *Social Network Analysis and Mining*, *15*(1), 87. https://doi.org/10.1007/s13278-025-01507-x

Saha, P., Datta, A., Jana, A., & Mukherjee, A. (2024). *CrowdCounter: A benchmark type-specific multi-target counterspeech dataset* (arXiv:2410.01400). arXiv. https://doi.org/10.48550/arXiv.2410.01400

Sahoo, N. R., Beria, G. P., & Bhattacharyya, P. (2024). IndicCONAN: A Multilingual Dataset for Combating Hate Speech in Indian Context. *Proceedings of the AAAI Conference on Artificial Intelligence*, *38*(20), 22313–22321. https://doi.org/10.1609/aaai.v38i20.30237

Song, X., Mamidisetty, S., Blanco, E., & Hong, L. (2024). *Assessing the Human Likeness of AI-Generated Counterspeech* (arXiv:2410.11007). arXiv. https://doi.org/10.48550/arXiv.2410.11007

Tekiroğlu, S. S., Bonaldi, H., Fanton, M., & Guerini, M. (2022). Using Pre-Trained Language Models for Producing Counter Narratives Against Hate Speech: A Comparative Study. In S. Muresan, P. Nakov, & A. Villavicencio (Eds.), *Findings of the Association for Computational Linguistics: ACL 2022* (pp. 3099–3114). Association for Computational Linguistics. https://doi.org/10.18653/v1/2022.findings-acl.245

Tekiroğlu, S. S., Chung, Y.-L., & Guerini, M. (2020). Generating Counter Narratives against Online Hate Speech: Data and Strategies. In D. Jurafsky, J. Chai, N. Schluter, & J. Tetreault (Eds.), *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics* (pp. 1177–1190). Association for Computational Linguistics. https://doi.org/10.18653/v1/2020.acl-main.110

Wang, H., Pan, Y., Song, X., Zhao, X., Hu, M., & Zhou, B. (2024). F^2RL: Factuality and Faithfulness Reinforcement Learning Framework for Claim-Guided Evidence-Supported Counterspeech Generation. In Y. Al-Onaizan, M. Bansal, & Y.-N. Chen (Eds.), *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing* (pp. 4457–4470). Association for Computational Linguistics. https://doi.org/10.18653/v1/2024.emnlp-main.255

Wang, H., Tian, Z., Song, X., Zhang, Y., Pan, Y., Tu, H., Huang, M., & Zhou, B. (2024). Intent-Aware and Hate-Mitigating Counterspeech Generation via Dual-Discriminator Guided LLMs. In N. Calzolari, M.-Y. Kan, V. Hoste, A. Lenci, S. Sakti, & N. Xue (Eds.), *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)* (pp. 9131–9142). ELRA and ICCL. https://aclanthology.org/2024.lrec-main.800/

Wang, M., Ma, S., Li, N., Zhang, P., Li, C., Gu, N., & Lu, T. (2026). *Echoes of Norms: Investigating Counterspeech Bots' Influence on Bystanders in Online Communities* (arXiv:2603.03687). ArXiv. https://doi.org/10.48550/arXiv.2603.03687

Wilk, B., Shomee, H. H., Maity, S. K., & Medya, S. (2025). Fact-based Counter Narrative Generation to Combat Hate Speech. *Proceedings of the ACM on Web Conference 2025, WWW '25*, 3354–3365. https://doi.org/10.1145/3696410.3714718

Wu, C., Wang, Y., Zhang, Y., Wang, H., & Pang, Y. (2025). Confront hate with AI: How AI-generated counter speech helps against hate speech on social Media? *Telematics and Informatics*, *101*, 102304. https://doi.org/10.1016/j.tele.2025.102304

Zheng, Y., Ross, B., & Magdy, W. (2023). What Makes Good Counterspeech? A Comparison of Generation Approaches and Evaluation Metrics. In Y.-L. Chung, H. Bonaldi, G. Abercrombie, & M. Guerini (Eds.), *Proceedings of the 1st Workshop on CounterSpeech for Online Abuse (CS4OA)* (pp. 62–71). Association for Computational Linguistics. https://aclanthology.org/2023.cs4oa-1.5/

Zhu, W., & Bhat, S. (2021). *Generate, Prune, Select: A Pipeline for Counterspeech Generation against Online Hate Speech* (arXiv:2106.01625). arXiv. https://doi.org/10.48550/arXiv.2106.01625

Zubiaga, I., Soroa, A., & Agerri, R. (2024). A LLM-based Ranking Method for the Evaluation of Automatic Counter-Narrative Generation. In Y. Al-Onaizan, M. Bansal, & Y.-N. Chen (Eds.), *Findings of the Association for Computational Linguistics: EMNLP 2024* (pp. 9572–9585). Association for Computational Linguistics. https://doi.org/10.18653/v1/2024.findings-emnlp.559