Echoes of Norms: Investigating Counterspeech Bots' Influence on Bystanders in Online Communities

Mengyao Wang
Fudan University
Shanghai, China
mengyaowang23@m.fudan.edu.cn

Shuai Ma
Institute of Software
Chinese Academy of Sciences
Beijing, China
mashuai@iscas.ac.cn

Nuo Li
Fudan University
Shanghai, China
linuo@fudan.edu.cn

Peng Zhang[*]
Fudan University
Shanghai, China
zhangpeng_@fudan.edu.cn

Chenxin Li
Fudan University
Shanghai, China
24300740004@m.fudan.edu.cn

Ning Gu
Fudan University
Shanghai, China
ninggu@fudan.edu.cn

Tun Lu[*]
Fudan University
Shanghai, China
lutun@fudan.edu.cn

Abstract

Counterspeech offers a non-repressive approach to moderate hate speech in online communities. Research has examined how counterspeech chatbots restrain hate speakers and support targets, but their impact on bystanders remains unclear. Therefore, we developed a counterspeech strategy framework and built *Civilbot* for a mixed-method within-subjects study. Bystanders generally viewed Civilbot as credible and normative, though its shallow reasoning limited persuasiveness. Its behavioural effects were subtle: when performing well, it could guide participation or act as a stand-in; when performing poorly, it could discourage bystanders or motivate them to step in. Strategy proved critical: cognitive strategies that appeal to reason, especially when paired with a positive tone, were relatively effective, while mismatch of contexts and strategies could weaken impact. Based on these findings, we offer design insights for mobilizing bystanders and shaping online discourse, highlighting when to intervene and how to do so through reasoning-driven and context-aware strategies.

CCS Concepts

• **Human-centered computing → Empirical studies in collaborative and social computing**.

Keywords

Online communities, Hate speech, Counterspeech, Bystanders, Chatbot

ACM Reference Format:
Mengyao Wang, Shuai Ma, Nuo Li, Peng Zhang, Chenxin Li, Ning Gu, and Tun Lu. 2026. Echoes of Norms: Investigating Counterspeech Bots' Influence on Bystanders in Online Communities. In *Proceedings of the 2026 CHI Conference on Human Factors in Computing Systems (CHI '26), April 13–17, 2026, Barcelona, Spain.* ACM, New York, NY, USA, 23 pages. https://doi.org/10.1145/3772318.3791797

1 Introduction

Online communities aspire to foster a diverse, open, and vibrant space for public, yet this ideal is increasingly undermined by the spread of hate speech. Hate speech is commonly defined as targeted, harmful, and weaponized forms of expression against specific groups [80]. It inflicts profound emotional and psychological harm on its victims, ranging from anxiety and self-blame to suicidal ideation, and can even trigger real-world violence [2, 3, 12]. A particularly troubling aspect is how easily hate speech spreads: when people are exposed to malicious or antisocial comments, they become more likely to produce similar negativity [3, 15], even if they are usually not highly aggressive [70]. Within communities, hate speech propagates imitation, perpetuates prejudice, and reinforces stereotypes [80], ultimately fuelling stigmatization [51] and structural injustice [14], which can further divide society. Traditional content moderation relies on restrictive interventions imposed by platforms to counteract hate speech. However, such methods often risk the over-removal of empowering discourse, forcing platforms to adopt cautious and conservative classification thresholds, which in turn further overlook implicit forms of hate speech [36]. Moreover, users can easily migrate to platforms with looser moderation policies, limiting the effectiveness of isolated interventions [80]. In light of the limitations of heavy-handed moderation, counterspeech has emerged as a widely recognized alternative—non-repressive, socially grounded, and scalable—that seeks not to suppress but to enrich online discourse.

Counterspeech is typically defined as a direct reply to hateful content [7, 71, 80], which can take forms such as challenging or

[*] Corresponding authors.

CHI '26, Barcelona, Spain

ACM ISBN 979-8-4007-2278-3/26/04
https://doi.org/10.1145/3772318.3791797

condemning the attack [37, 81], expressing solidarity and support [37], or even attempting to shift perspectives [62]. Unlike content moderation, counter speech is rooted in a liberal tradition, emphasizing "speech against speech" rather than "power against speech" [14, 47]. Its potential lies not only in addressing individual instances of hate but also in influencing broader social norms that frame hate as unacceptable, thereby altering the overall tone of community dialogue [9, 11]. Within such a mechanism, online spaces need not serve as arenas for the spread of hostility but can evolve into fairer and safer environments for communication [69, 80]. Moreover, counterspeech may generate a "positive contagion effect": expressions that denounce incivility or encourage respectful dialogue can themselves inspire similar responses, creating cycles of positive imitation and role modelling [54, 73].

In recent years, the rapid advancement of generative AI has provided a technical foundation for the automation of counterspeech. A growing body of work has constructed hate-counterspeech corpora [1, 33, 52, 85, 89], trained classification and generation models [8, 17, 32, 52, 68, 91], and introduced enhancement strategies such as incorporating background knowledge [1, 84], contextual information [6, 18, 25], and complex argumentative structures [32, 35, 68], etc. These approaches have enabled counterspeech chatbots to achieve more persuasive argumentation and contextually appropriate responses in online community settings, amplifying potential social impact. However, existing research on counterspeech chatbots has primarily focused on curbing hate speakers [6, 11, 35] or providing support for targets of hate [60, 61], while paying far less attention to bystanders—arguably the most crucial group in terms of both size and influence. Bystanders are not neutral observers: they often lean toward those who oppose hate speech, forming a "silent majority" that substantially shapes the community's prevailing attitudes and social norms [80]. They may become potential counterspeakers, yet their silence can also be interpreted as implicit tolerance of hate. Prior research suggests that public counterspeech can disrupt the perception that "most people tolerate hate", thereby weakening the spiral of silence [57], and may also trigger herd effects [5] that motivate more users to speak out. However, it remains unclear whether counterspeech chatbots can exert comparable social influence, as systematic empirical evidence is still limited.

To this end, we focus on the social influence of counterspeech chatbots on bystanders in online communities and how these effects ripple into the broader normative climate. We therefore pose the following research questions:

- RQ1: To what extent do bystanders endorse the chatbot's counterspeech and show changes in their behavioral tendencies (e.g., perceived reason strength, credibility, and confidence in countering)?
- RQ2: How do different types of chatbot' counterspeech shape bystanders' endorsement of the chatbot's responses and change of their behavioral tendencies?

To establish a structured foundation for studying counterspeech, we developed a unified framework along three dimensions: sentence type (question and non-question), tone (positive and negative), and strategic intent, which includes cognitive strategies (e.g., highlighting truth) and affective strategies (e.g., denouncing hate speakers). The combination of dimensions produced eight distinct counterspeech strategies. Based on this framework, we built Civilbot, a prototype chatbot that generates context-aware counterspeech across these strategies. We then conducted a mixed-methods within-subject experiment with participants recruited from online communities who were interested in sensitive topics, generally silent in public discussions, yet opposed to hate speech. Each participant chose eight topics and, for each, read a hate-speech post, observed a counterspeech reply, and completed pre- and post-exposure attitude measures. Sessions were randomly assigned to the eight strategies, so everyone experienced all strategies. We assessed perceived counterspeech quality (e.g., convincing and strong reasons), subjective acceptance (e.g., credibility, importance, overall agreement), and behavioural tendencies (e.g., confidence in countering, willingness to participate), and complemented these metrics with semi-structured interviews exploring bystanders' detailed perceptions of chatbot roles, and potential effects on community norms, etc.

Our findings show that Civilbot's counterspeech shapes bystanders' perceptions of both the counterspeech itself and the bot, and also influences the overall climate of online communities—even affecting subtle behavioural tendencies. Civilbot is generally viewed as credible and as signalling community norms, though its shallow reasoning constrains its persuasiveness. Behaviourally, its impact is subtle and sometimes mixed: it can guide and encourage bystanders, substitute for their own response, dampen participation when it performs poorly, or prompt users to step in out of frustration with its limitations. Beyond persuasion, it also contributes to the broader community atmosphere, such as by helping to cool down emotional intensity or provide additional information for reflection. Strategy proves decisive: cognitive strategies are typically more effective than affective ones; tone may influence behavioural tendencies but relies on specific contexts; and sentence forms, when paired with other strategies, can either stimulate reflection or trigger resistance. These findings offer design insights for Civilbot to engage bystanders and help mediate a hostile community climate: Civilbot must determine when to intervene and why its intervention is needed, and it must also decide how to intervene through reasoning-driven motivation, information-enabled argument, context-adaptive strategy, and extending modalities beyond text.

Our work makes the following contributions:

- To the best of our knowledge, this is the first study that examines the mechanisms through which counterspeech chatbots influence bystanders in online hate incidents.
- We develop an experimental framework for counterspeech that incorporates three key dimensions (sentence type, tone, and strategic intent). Based on this framework, we build a chatbot prototype (Civilbot) to empirically examine its influence on bystanders' attitudes and behavioural intentions.
- We provide design insights for future counterspeech chatbots that explicitly consider the role of bystanders in shaping responses to hate speech within online communities.

2 Related Work

2.1 Origins and Functions of Counterspeech

The idea of mitigating the harms of hate speech through counterspeech originates from debates on freedom of expression. Rather than silencing people, this perspective argues that harms can be reduced by responding in constructive ways [14]. In contrast, coercive suppression often entails moral costs: it may further restrict the expressive freedom of marginalized groups, thereby reinforcing exclusion and resentment [48]. Algorithmic moderation is now the dominant response, but platform-level opacity, over-moderation, and exclusion from decision-making have raised concerns about fairness and legitimacy [36, 55]. Even human moderation often functions as an ex-post control [72]. Against this backdrop, counterspeech has re-emerged as an alternative that promotes social justice: the way to counter falsehood is not suppression but exposure, debate, and persuasion, allowing truth to prevail in open contestation [13, 47].

Subsequent research has examined how counterspeech affects online social interactions. It can trigger contagion effects and act as implicit cues of community norms: for example, polite responses increase willingness to engage [34], while metacommunication (calling out incivility) promotes civility [54]. Emotional contagion and imitation also play a role, with users adapting to the affective tone and behaviour of others [46, 73]. These dynamics alter perceptions of norms—both descriptive (what is common) and injunctive (what is appropriate)—and thereby influence whether hate appears tolerated or unwelcome [11, 69].

The social impact of counterspeech is shaped not only by its content but also by the identity of the counterspeaker. Counterspeech can originate from targets of hate, bystanders, or non-targeted users, and may be delivered by ordinary members or authority figures [14]. Interventions from influential or high-status members are more likely to be adopted, while "outsiders" without established identities exert weaker impact [7, 73]. Moreover, factors such as race or follower count can further moderate influence [56].

In sum, prior work highlights the promise of counterspeech in shaping community norms through mechanisms such as contagion, emotion, and identity influence. Inspired by this perspective, AI chatbots may serve as consistent and scalable counterspeakers, activating normative cues and emotional dynamics in ways that differs from human interventions.

2.2 Strategies and Automated Generation of Counterspeech

As scholarly attention to counterspeech has grown, researchers have proposed diverse taxonomies of strategies. Early work on Twitter identified eight non-exclusive forms, including factual correction, highlighting contradictions, warning of consequences, expressing identification, denouncing hate, using media, humor, or particular tones [64]. Later studies refined these categories by distinguishing positive versus hostile tones [53], grouping responses into informative, denouncing, questioning, positive, and humorous intents [32], or emphasizing empathetic, consequence-warning, and polite formulations [6, 66]. Other work has drawn on argument structures, speech acts, and psychological mechanisms such as normative influence and empathy induction [25, 68]. Surveys summarize these efforts along broader axes such as active versus passive, positive versus negative style, and responses to explicit versus implicit hate [14].

Parallel to this conceptual work, rapid advances in AI have enabled the automated generation of counterspeech. Researchers have constructed a range of datasets—from Twitter and online articles [1, 52] to large domain-specific corpora such as WokeCorpus [33] and Reddit-based annotations [85]. These resources support models that generate counterspeech [8, 32, 68], experiment with large language models [33, 84], and incorporate contextual information such as argumentation, psychology, or personalization [6, 18]. At the same time, HCI studies have explored human–AI co-creation frameworks and design guidelines for AI-assisted counterspeech [23, 55].

Overall, prior work has developed diverse but fragmented taxonomies of strategies, and demonstrated the feasibility of automated generation. Yet the lack of a unifying framework limits comparability across approaches. Building on these foundations, our study introduces a structured strategy framework and empirically examines how chatbot-mediated counterspeech influences online communities.

2.3 Social Impact of Chatbot-Generated Counterspeech

Existing research on the social impact of counterspeech primarily focuses on two directions: constraining hate speakers and supporting targets of hate. Studies on hate speakers examine both reflections on past hateful behaviours—such as deleting hateful comments [6, 35]—and potential changes in future behaviours, including shifts in the toxicity of expressed opinions [6] or reductions in hate speech and aggression [11, 35]. These findings, derived from data analysis and controlled experiments, have also informed design implications for counterspeech chatbot to affect hate speakers. Research on targets of hate, in contrast, investigates how this identity influences their attitudes and behaviours of countering hate. For instance, exposure to online hate can become a key motivator for sustained participation in counterspeech [61]. Moreover, users whose identities closely align with those of the targets are more likely to perceive counterspeech as a feasible response and to actively engage in it [60].

In contrast, HCI research emphasizes the crucial role of bystanders—the "silent majority" of a community who shape mainstream norms [80]. The value of interventions stems from their ability to trigger the observer effect, prompting users to align self-expression with community expectations [65]. Even when targeting violators, these visible interventions function as a "deterrence" mechanism [41]. Bystanders learn descriptive norms by observing others' behaviours and their consequences, and learn injunctive norms through the explicit behavioural guidelines presented in the interventions [69]. Studies have designed interventions to stimulate prosocial bystander behaviour. For instance, interface designs simulating "under observation" contexts, e.g., displaying audience size metrics [22] or notifying relevant groups that bystanders have viewed cyberbullying content [78], have been shown to heighten accountability. Similarly, improving moderation transparency via public post-removal explanations helps bystanders understand acceptability boundaries, fostering norm-aware contributions [40] and significantly boosting engagement levels [41]. As a similarly information-rich intervention, counterspeech has garnered increasing attention regarding its impact on bystanders. While strategies like condemnation or distraction may yield limited behavioural intent changes [42], perspective-focused strategies may reduce the spread and amplification of hate speech on platforms [27]. Furthermore, effective counterspeech not only empowers bystanders to actively challenge hate but also reduces the likelihood of future hateful content creation by both bystanders and hate speakers [20].

While prior work has shown how chatbot-generated counterspeech can restrain hate speakers and support targets, its effects on bystanders remain largely unexplored—even though the value of bystander intervention has been extensively validated in HCI research. Investigating how chatbot-mediated counterspeech interacts with this silent majority is therefore crucial, both to fill a theoretical gap and to derive design insights for chatbots that promote anti-hate and constructive online discourse.

3 Constructing a Framework of Counterspeech Strategies

3.1 Methodology of constructing strategy framework

To synthesize commonly used counterspeech strategies from prior research, we conducted a literature review on Google Scholar focusing on studies related to hate speech interventions, counterspeech generation, and strategy typologies. Building upon existing labels of counterspeech strategies reported in the literature [6, 14, 25, 32, 53, 64, 66, 68] and guided by inductive coding methods [79], we developed a unified framework. Initially, one author reviewed the literature and extracted text segments pertaining to counterspeech strategies. Two authors then independently identified and annotated preliminary strategy types based on these excerpts. Ambiguities and disagreements were resolved through cross-validation and group discussions, which included semantic refinements to minimize overlap and redundancy among categories. For example, the label "question" appeared in prior studies with different meanings. In some cases, it referred to counter question form [16, 17, 32, 53, 66, 68]; in others, it denoted challenges to the credibility of information sources or claims underlying hate speech [66]. To reduce confusion, we treated the first as "question" and the second as part of the "pointing out hypocrisy or contradictions" sub-strategy. This process resulted in the identification of 20 initial sub-strategies.

We then inductively consolidated all initial sub-strategies into eight mutually exclusive categories: *question*, *non-question*, *positive tone*, *negative tone*, *rebutting falsehoods*, *highlighting truth*, *denouncing hate speakers*, and *supporting targets of hate*. Guided by the Elaboration Likelihood Model (ELM) [59], the latter four strategies were further synthesized into two overarching types: *cognitive strategy* and *affective strategy*. The final classification framework comprises three dimensions—sentence type (question vs. non-question), tone (positive vs. negative), and strategic intent (cognitive vs. affective influence). The resulting framework is both theoretically grounded and practically oriented, facilitating structured experimental design.

3.2 Classification Dimensions

The framework comprises three dimensions: Sentence Type, Tone, and Strategic Intent.

3.2.1 Sentence Type. This dimension distinguishes between questions and non-questions. Questions are highlighted as a specific strategy that formulates counterspeech in interrogative form to challenge hate speakers, including approaches variously labelled as counter questions or questioning [16, 17, 32, 53, 66, 68]. Given that "question" is frequently identified as a distinct strategy type in counterspeech research, and that studies underscore its role in fostering critical thinking [21], we consider it necessary to include sentence type as an independent dimension. All other utterances—such as statements, imperatives, or exclamations—are classified as non-questions.

3.2.2 Tone. We conceptualize tone in counterspeech as a spectrum but, for experimental controllability, adopt a binary classification of positive versus negative tone. Positive tone encompasses cooperative and friendly expressions, aligning with prior work on perceived interaction quality with conversational agents [76]. Examples include polite, empathetic, or detoxified responses [16, 17, 32, 53, 63, 67, 68]. Negative tone, by contrast, involves more confrontational or critical expressions, often manifested through sarcasm or humour [16, 17, 32, 35, 64, 66], whose intensity may vary depending on the utterance.

3.2.3 Strategic Intent. The intent dimension is divided into cognitive and affective strategies. This classification builds on inductive synthesis of existing strategy typologies and is informed by the ELM [59], which distinguishes cognitive, affective, and behavioural persuasion routes. Since behavioural persuasion in ELM refers to effects triggered by individuals' own actions—which falls outside the scope of this study and is not reflected in current strategy types—we exclude it here.

Cognitive strategies aim to shift bystanders' cognition and can be further divided into rebutting falsehoods and highlighting truth [14]. Rebutting falsehoods involves exposing false, contradictory, or hypocritical elements in hate speech [16, 64, 66, 68, 81], such as rejecting abusive premises [81] or questioning information sources and underlying claims [66]. Saha et al. further extended this category by incorporating Walton's argumentation schemes—Means for Goal, Goal from Means, Source Knowledge, Source Authority, and Rule or Principle [68]—to enrich argumentative methods. Highlighting truth seeks to change cognition by presenting facts [16, 32, 64, 68] or citing arguments from online sources [1]. Recommendations to conduct additional verification such as doing more research [63] also fall into this category. Additionally, warning about potential online or offline consequences of hate speech [6, 35, 64, 66, 68] is included, as such warnings supplement information rather than directly refute hate.

Affective strategies aim to trigger change by engaging emotions, primarily including denouncing hate speakers and supporting targets of hate. Denouncing hate speakers labels speech as hateful, dangerous, or biased [16, 64, 68, 81], often by pointing out hate-related keywords [63, 66] or warning about their inappropriateness [63], thereby eliciting shame in the hate speaker. Supporting targets of hate includes expressing solidarity with the targeted group

Table 1: Framework of Counterspeech Strategies

Dimension	Sub-dimension	Methods & Explanation	Example
Sentence Type	**Question**	Counterspeech framed as questions to challenge hate speakers, prompt reflection, or expose contradictions.	“If this stereotype were true, how do you explain the many successful Uyghur entrepreneurs in tech and finance?”
	Non-question	Includes statements, imperatives, exclamations; directly presents counter-arguments without question framing.	“Your claim ignores census data showing that crime rates are not higher in immigrant communities.”
Tone	**Positive**	Friendly, cooperative, polite, or constructive tone; designed to reduce defensiveness and encourage dialogue.	“I see why you might think that, but here’s another study that tells a different story.”
	Negative (Sarcasm / Hostile)	Confrontational, mocking, or hostile tone; can provoke shame or resistance but may discourage repetition.	“Right, because decades of peer-reviewed research are clearly just made up for fun.”
Strategic Intent	**Cognitive strategy**	Aims to change beliefs through reasoning, evidence, and fact-checking, encouraging rational evaluation. [**Rebutting Falsehoods**] Identifying and refuting false claims, contradictions, or unreliable sources. [**Highlighting Truth**] Presenting accurate facts, verified evidence, or pointing to credible sources; sometimes warning of real consequences.	“You said immigrants ‘don’t pay taxes,’ but IRS data (2022) shows immigrant households contribute over $330 billion annually in taxes.” / “WHO reports confirm vaccines save 4–5 million lives every year. Spreading misinformation only increases public health risks.”
	Affective strategy	Aims to trigger emotional reactions, such as shame (toward hate speakers) or empathy (toward targets). [**Denouncing Hate Speakers**] Labelling the statement as hateful, biased, or harmful to evoke accountability. [**Supporting Targets**] Expressing solidarity, empathy, or defence of the targeted group to humanize them and restore dignity.	“This remark is racist and fuels dangerous stereotypes that have led to violence offline.” / “I stand with Muslim women who choose how they dress—their voices matter more than your prejudice.”

[16, 64, 66], voicing support for specific entities [81], or showing empathy toward the target group [6, 66]. The goal here is to make hate speakers aware of the harm caused to others [35]. Compared to denouncing hate speakers, this approach emphasizes empathy rather than shame.

It is worth noting that although some studies propose other strategies such as moral qualities, identity traits, or values [68], we did not incorporate them into the current framework. Because the literature review indicates that most counterspeech research builds upon, adjusts, and reorganizes earlier classic frameworks [64] around a relatively stable and reusable set of common strategies. This aligns with the aim of our study: to construct a foundational framework for examining how different strategies influence bystanders. Furthermore, in our experimental design, we explicitly defined the chatbot’s identity to avoid confounding anthropomorphic factors. Issues related to chatbot identity, persona, and more complex strategy and expression will be further discussed in Section 6.

Accordingly, the final framework centers on three binary dimensions—sentence type, tone, and strategic intent—forming a $2 \times 2 \times 2$ taxonomy that yields eight distinct strategy combinations (see Table 1). This structure underpins the design of the experimental stimuli. Additionally, each dimension is linked to a set of sub-strategies that function as optional, randomized concrete methods during counterspeech generation. Specifically, within the strategic intent dimension, we only adopt the two primary classifications—cognitive and affective strategies—where further sub-classifications serve as optional, concrete operational methods.

4 Experimental Design

4.1 Counter Hate Transcript Design

To support our strategy-based counterspeech experiment, we constructed a dataset of hate speech from the peer-reviewed Chinese bias corpus CDIAL-BIAS DATASET [90]. This dataset is sourced from Zhihu, a widely used Chinese platform known for discussions on diverse social issues, where biased and hateful expressions are prevalent. The corpus covers four major topics of social controversy (gender, race, region, and occupation) and includes multiple subtopics targeting specific groups, offering well-structured content. For example, each entry contains a question and an answer, with the question serving as the context. To extract hate speech from the dataset, we leveraged its existing annotations and selected entries marked as expressing bias or prejudice, which are more likely to contain hateful content. We then applied an operational definition of hate speech—weaponized expressions targeting specific social groups that may cause emotional or psychological harm

[80]—to further refine the data. Using this definition, we designed prompts (see Appendix A) and applied the Qwen-Turbo model to re-screen entries while annotating the targeted groups (e.g., women, immigrants). Two authors then independently reviewed the outputs to determine whether each entry qualified as hate speech and to identify the target groups. Disagreements were resolved through discussion. Ultimately, 27 representative hate speech entries across four categories (gender, race, region, and occupation) were retained as the foundation for strategy-based experimentation. These entries were categorized into several subtypes targeting specific groups (e.g., factory workers), with each subtype comprising multiple distinct questions.

We then used the GPT-5 model and the default decoding settings provided by the API to generate counterspeech texts for the chatbot, adopting an iterative and incremental prompt-engineering process [30] to ensure that each response adhered to one—and only one—of the eight counterspeech strategies. Building on the strategy framework and definitions introduced in Section 3, we first designed prompts and refined them iteratively. For example, the tone label hostile was revised to sharply critical or emotionally intense to avoid excessively toxic outputs. The final prompt included three components (see Appendix A): (1) the role definition of the model as a counterspeech generation expert and its goal of producing responses aligned with one of the eight strategy combinations; (2) detailed definitions of each counterspeech dimension and sub-strategy from the perspective of a counterspeaker; and (3) the hate speech text, presented in the format "Hate speech: … Counterspeech: …". To stabilize the generation format, we constructed a small set of eight counterspeech examples that illustrate the expected strategic features. These examples were collaboratively drafted and refined by the authors and remained fixed across all prompts. We adopted a standard few-shot prompting setup to ensure that model outputs conformed consistently to the intended strategy.

To verify that the generated responses correctly implemented the target strategies, two authors independently annotated the initial outputs across three dimensions of the counterspeech framework: sentence type, tone, and strategic intent. As each dimension featured two categories, this created a total of eight distinct target strategies. Inter-rater reliability (IRR) was assessed by calculating Fleiss's Kappa ($\kappa = 0.82$) and classification accuracy (84.42%) across the three dimensions in parallel. The resulting κ value indicated "almost perfect" agreement [26], which, alongside the high accuracy, established the authors' reliability for subsequent tasks. Responses inconsistent with the target strategy in at least one dimension were flagged for iterative regeneration. For example: "Have you considered how hurtful such comments might be to girls who strive for independence and self-worth? Let's try to see things from their perspective and respond with understanding rather than blame." Although this response satisfies the requirements for a positive tone and an affective strategic intent, it mixes question and non-question forms, failing to meet the strategy's requirement for a purely question sentence type. Therefore, it was flagged for regeneration. This process continued until both authors confirmed complete alignment across all three dimensions, and eight responses for each of the eight strategy types were collected for every hate speech entry. To better simulate a realistic browsing experience in online communities, we also retrieved neutral responses corresponding to the "q"(question) entries in the hate speech dataset from Zhihu. We first collected a broad set of candidate comments via web crawling. Then, two independent researchers then screened the comments to ensure they contained no hateful language, emotional tone, or overt stance-taking. All data were anonymized to protect privacy. Finally, five neutral responses were selected for each question.

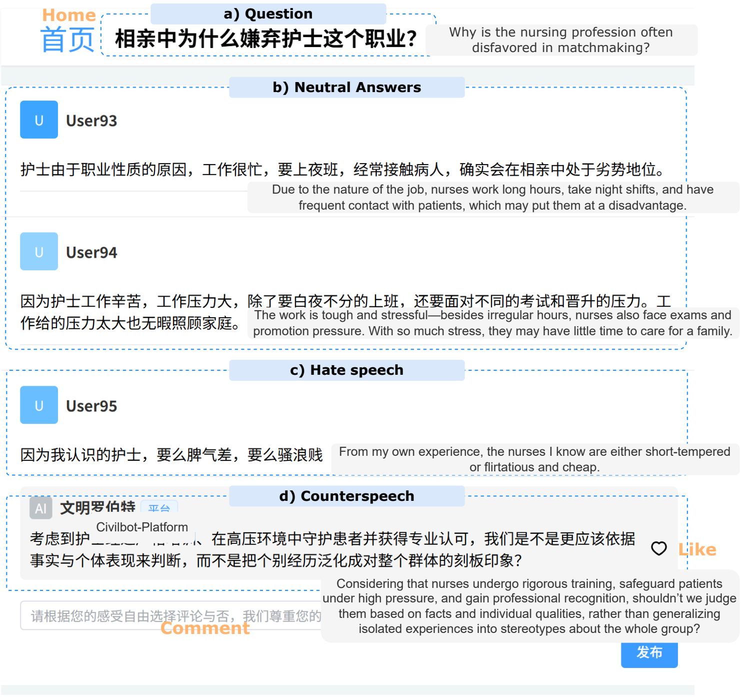
4.2 Bias Mitigation-Counter Hate Transcript Visual Design

We selected Zhihu as the design reference for our simulated discussion platform, since our hate speech dataset was originally collected from Zhihu and thus naturally contains the platform's question-answer structure. The system used a Vue [1] + Element Plus [2] front end and a lightweight Node.js back end deployed on a remote server. The platform provided two main interfaces: a question-browsing page and a Zhihu-style question page. Following [30], we made several design decisions in our Zhihu-style question page visual interface to mitigate potential biases, as illustrated in Figure 1:

- Username: In each topic transcript, the hate speakers were assigned different usernames (e.g., User13 for misogynistic hate speech, User12 for hate speech against nurses) to prevent any carry-over effects from prior utterances.
- Colour choice: All user avatars were presented in shades of blue, with only tonal variations, to minimize potential bias stemming from avatar imagery or colour-related metaphors.
- Removal of interactions: All answer-level interactions, including upvotes, downvotes, and comments, were removed to eliminate implicit cues about community norms that could bias participants.
- Answer order: For each question, we displayed a random number (1-5) of neutral answers. The set of neutral responses for each question was fixed across participants to ensure consistency. Neutral answers were always shown first, followed by the hate speech, simulating the experience of unexpectedly encountering hate speech while browsing a topic and providing the necessary contextual buildup for the scenario. To isolate the effect of the counterspeech without potential interference from any subsequent neutral answers, no additional neutral answers were shown afterward, and all measurements were completed immediately after participants viewed the hate speech and the counterspeech.
- Timing: After participants encountered the hate speech and completed the first survey measure, counterspeech shows up. This design prevented participants from overlooking counterspeech or being influenced by it prematurely. A second survey measure was then triggered once participants had viewed the counterspeech but before they left the page, capturing their immediate reactions.
- Counterspeech robot username: To avoid anthropomorphisation and gendered connotations [30], we did not personify the moderation bot. Instead, we named it Civilbot (文明

[1] https://vuejs.org/

[2] https://element-plus.org/



相亲中为什么嫌弃护士这个职业?

Why is the nursing profession often disfavored in matchmaking?

User93

护士由于职业性质的原因，工作很忙，要上夜班，经常接触病人，确实会在相亲中处于劣势地位。

Due to the nature of the job, nurses work long hours, take night shifts, and have frequent contact with patients, which may put them at a disadvantage.

User94

因为护士工作辛苦，工作压力大，除了要白夜不分的上班，还要面对不同的考试和晋升的压力。工作给的压力太大也无暇照顾家庭。

The work is tough and stressful—besides irregular hours, nurses also face exams and promotion pressure. With so much stress, they may have little time to care for a family.

User95

因为我认识的护士，要么脾气差，要么骚浪贱

From my own experience, the nurses I know are either short-tempered or flirtatious and cheap.

AI 文明罗伯特 平台

Civilbot-Platform

考虑到护[illegible]、在高压环境中守护患者并获得专业认可，我们是不是更应该依据事实与个体表现来判断，而不是把个别经历泛化成对整个群体的刻板印象?

Considering that nurses undergo rigorous training, safeguard patients under high pressure, and gain professional recognition, shouldn't we judge them based on facts and individual qualities, rather than generalizing isolated experiences into stereotypes about the whole group?

请根据您的感受自由选择评论与否，我们尊重您的

Figure 1: Sample interface of the simulated discussion platform, showing: (a) an excerpted question; (b) neutral answers; (c) a hate-speech post. After participants complete the pre-test questionnaire, the interface displays (d) a counterspeech message. Participants may "Like" the counterspeech or post their own comments in response to the hate-speech post. After finishing all interactions, they click "Home", complete the post-test questionnaire, and return to the homepage.

罗伯特 in Chinese). To further ensure transparency, we appended the tag platform to its username, clarifying that it was an official platform-moderation agent rather than a real user.

- Reply length: we adopted a soft-approximate approach by prompting GPT-5 to generate one short paragraph per speech act (detailed in Appendix A). The resulting length variance allowed us to balance response lengths across different strategies.

4.3 Participants

Given that bystanders lack an accessible sampling frame and cannot be directly identified through platform-level data, we adopted convenience sampling [29] by posting recruitment information in WeChat to reach individuals who were easy to access and willing to participate. We supplemented this with limited snowball sampling [10], which is commonly used when the target population is behaviour-defined and distributed within social networks. We recruited a total of 58 participants through WeChat, of which 52 were available for the quantitative analysis (demographics and individual differences shown in Table 2). Prior to the formal experiment, all participants completed a pre-survey reporting their demographic information, social media usage habits, topic interests, and behavioural tendencies in response to hate incidents. We adopted a survey [30] for participant inclusion and used self-report method [42] to assess bystander roles. Specifically, the process was as follows: Participants' inclusion as active users of online communities was determined by collecting their social media usage

Table 2: Summary of Participant Demographics and Covariates ($N = 52$)

Covariates	Value Type	Values/Distribution
Gender	Categorical	Female ($n = 27$) | Male ($n = 25$)
Age	Continuous	Range (18–27) Mean = 20.17, Std. = 2.52
Education Level	Ordinal	High School ($n = 1$) Undergraduate ($n = 40$) Graduate ($n = 11$)
Major	Nominal	Law ($n = 11$) | Economics ($n = 10$) Engineering ($n = 10$) | History ($n = 7$) Literature ($n = 7$) | Science ($n = 7$)
AI Literacy	Continuous (Likert)	Scale: 1 (Very Low) to 5 (Very High) Mean = 3.10, Std. = 1.29
AI Use	Continuous (Likert)	Scale: 1 (Never) to 5 (Daily) Mean = 4.37, Std. = 0.87

Note: The last two covariates (AI Literacy and AI Use) are self-reported on a 5-point Likert scale.

frequency using a 5-point Likert scale. Subsequently, we assessed three key constructs, each measured by a separate 5-point Likert scale: interest in sensitive issues ("How interested are you in sensitive issues such as gender, race, region and so on?"), counter-engagement ("How often do you challenge or rebut hate speech?"), and hate speech endorsement ("How often do you like, share, or publish hate speech?"). Furthermore, in the formal experiment, bystander roles were reconfirmed through both the measurement of empathy in the pre-survey of each session and subsequent interviews. Based on these responses, we screened for our target group: potential bystanders, individuals who are active in online communities and interested in sensitive issues, but who typically remain silent and do not endorse hate speech.

The required sample size was determined via a G*Power analysis for a repeated-measures ANOVA (within-subject design, 8 conditions). Based on the empirically derived median effect size [75] ($Cohen'sf = 0.175$), and a systematic review of reported effect sizes at CHI by [58] (where the small–medium range of the type "human-centered computing" is approximately 0.10-0.26, making 0.175 close to the median of this interval), we adopted $f = 0.175$ as the planned effect size for sample size estimation. This analysis indicated a minimum required sample size of 48 participants to achieve 80% statistical power at $\alpha = .05$ [19]. Among the 58 participants we recruited, 5 completed a pilot study that informed revisions to the questionnaire, and 1 dataset was excluded due to incompleteness of all sessions, leaving 52 valid participants for analysis. The study protocol was approved by the university's Institutional Review Board. Participants were informed that they could withdraw at any time and would have access to psychological support if needed. The entire study lasted approximately 45 minutes and consisted of four phases (see Fig. 2). Participants who completed all sessions received a compensation of 50 RMB.

4.4 Experiment procedure

Our study procedure contains four phases (see Fig. 2). In Phase A (Pre-survey), beyond the screening items, we included a self-developed strategy preference questionnaire to better understand participants and provide cues for subsequent experiments and interviews. This introduced three dimensions of counterspeech strategies and their bipolar expressions; participants indicated their preferences using a five-point scale. The main experiment comprised Phase B (Introduction) and Phase C (Experiment sessions). In Phase B, participants provided informed consent and read task instructions, which included a content warning for potentially offensive material. They then proceeded to Phase C, where participants freely selected eight questions of interest from the 27 items available. To mitigate potential topic effects and ensure comprehensive content coverage, the study incorporated three primary controls: The presentation order of the questions was randomized per participant, controlling for sequence effects. Second, participants freely chose their eight questions after browsing the full pool, ensuring broad topic engagement based on personal interest. Third, the strategy presentation order was also randomized, preventing the fixed coupling of any specific counterspeech strategy with a particular question. Each selected question corresponded to one session, and across the eight sessions, participants were exposed to all strategy types without repetition. Within each session, participants first encountered a piece of hate speech and completed the first survey measure. They were then presented with counterspeech generated by Civilbot, to which they could respond (e.g., by liking counterspeech or commenting on hate speakers). After clicking the "Home" button, the system displayed the second survey measure. In Phase D (Post-survey), participants engaged in semi-structured interviews. They were asked to reflect on their session experiences, discuss their subjective perceptions of Civilbot, elaborate on how Civilbot and these strategies affect their behaviours, and share their views on the role of Civilbot in shaping community social norms.

4.5 Measurements

During the initial scale design, we included multiple constructs to assess bystanders' evaluations before and after exposure to counterspeech, such as willingness to participate [54], willingness to counterspeak [76], counterspeech efficacy [82], and empathy [76]. A pilot study (N=5) gathered qualitative feedback, which indicated item redundancy and the absence of a direct measure of counterspeech persuasiveness. These issues increased respondent fatigue and compromised measurement precision. We therefore streamlined items to reduce repetition and incorporated the well-validated Perceived Argument Strength Scale [88] to better capture argument quality. The final framework consists of three dimensions (detailed in Appendix A): (1) perceived counterspeech quality (e.g., convincing and strong reasons), (2) subjective acceptance (e.g., credibility, importance, overall agreement), and (3) behavioural tendencies (e.g., confidence in countering, willingness to participate). Items for pre-post comparison (e.g., confidence in countering, willingness to participate) were measured at both time points to track changes, serving as a self-comparison baseline for assessing the intervention's effects. Items directly evaluating counterspeech were measured only post-exposure to avoid repetition. Empathy was assessed only in the pre-survey to exclude hate-endorsing respondents. Since the original scales were in English, we applied back-translation to ensure the conceptual validity of their Chinese version.

4.6 Data Analysis

To address the research questions, we employed a mixed-methods approach combining quantitative survey analysis with qualitative content analysis. For the quantitative component, we first calculated mean scores for each item to identify general trends in the effectiveness of all counterspeech strategies, providing a quantitative backdrop for subsequent qualitative analysis of RQ1. For RQ2, one-way ANOVA tests were conducted to examine the main effects of each of the three dimensions: intent, tone, and sentence type. Due to the conceptual distinctness of these dimensions, separate one-way ANOVAs were employed instead of MANOVA. Where significant main effects were identified, two-way ANOVA was performed to investigate potential interaction effects, followed by simple effects analysis where appropriate to uncover more granular patterns [28]. Additionally, we conducted exploratory pairwise comparisons between all eight strategy groups via pairwise t-tests. These analyses were intended only to complement the ANOVA results by highlighting potential patterns that could inform our qualitative interpretation, rather than to establish definitive effects.

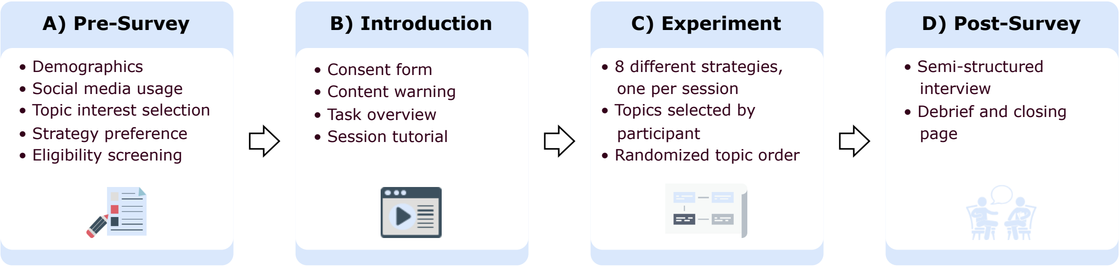


Figure 2: The overall experiment procedure, including four phases: (A) Pre-survey, (B) Introduction, (C) Experiment sessions, (D) Post-survey.

Therefore, we did not apply corrections for multiple comparisons. The qualitative analysis of open-ended feedback followed the thematic analysis approach [45], involving an iterative process wherein two researchers independently coded the data, identified recurring patterns, and through discussion consolidated them into salient themes reflecting participants' perceptions of Civilbot effectiveness, authenticity, contextual appropriateness, etc.

5 Results

5.1 RQ1: Counterspeech Chatbot's Overall Influence

Overall, counterspeech delivered by Civilbot had a complex yet meaningful influence on bystanders. On the attitudinal level, it was generally perceived as credible and normative, reaffirming that hate speech is unacceptable and offering comfort to silent users. Yet its reasoning was often seen as shallow or "*too AI-like*", limiting persuasiveness and its ability to mobilize engagement. On the behavioural level, its effects were subtle and sometimes contradictory—acting alternately as guidance, substitution, negative modelling, or reverse motivation, depending on users' self-efficacy and motivation. At the community level, Civilbot contributed to shaping the climate by informing bystanders, cooling their emotions, and encouraging them critical thinking. Together, these findings highlight both the potential and the limits of Civilbot in supporting bystander engagement and mediate online community climate.

5.1.1 Perceived as Normative but Limited in Persuasiveness. Bystanders basically perceived Civilbot as credible and normative but not strongly persuasive. Quantitatively (see Fig. 4), it received moderate ratings on *credibility* (3.33/5) and *overall agreement* (3.31/5), while scoring lower on *convincing reason*, *strong reason*, and *importance*, with *strong reason* in particular falling below the passing threshold (2.84/5).

Participants often described Civilbot as a transmitter of community norms (P20, P15, P5, P17, P48, P54) and even a representative of ethical values, especially for those with higher AI literacy. For some, its presence helped counterbalance despair in hostile comment sections: "*If the comment section is full of irrational racist remarks, you might feel disillusioned about the world. But when AI [Civilbot] shows up, it reassures me that normal human values still exist*" (P21, P49, P54, P58). Others emphasized that because chatbots are perceived as relatively objective, their interventions carry stronger normative weight than those of ordinary users: "*People usually won't argue with a robot. If it says something is problematic, most will accept it as correct*" (P30). Some participants even felt that siding with Civilbot placed them on moral high ground (P7, P12, P17). At the same time, views on its identity were divided: while some dismissed AI outputs as superficial recombinations (P1), others considered them a form of "*collective intelligence*" (P6, P21), or simply cared more about content than source (P5, P27).

However, participants rated Civilbot low on perceived quality of counterspeech, largely because its arguments were perceived as formulaic, off-topic, or shallow (P1, P4, P5, P8, P11, P26, P27, P29, P31, P37, P45, P58). As one participant put it, "*It didn't explain the causes of the issue or respond to the actual question—just gave a detached piece of advice*" (P26). Several emphasized that effective counterspeech requires engaging with the speaker's underlying logic rather than replying at the surface level. As P6 noted, "*To refute someone, you first need to know their reasoning ... it's about uncovering their motives and addressing their logic, rather than just throwing back a generic line.*" Such perspectives highlight that participants expected Civilbot to probe the motives behind hateful speech, exposing logical gaps or critical points that could make counterspeech more persuasive. Additionally, some criticized its "*AI-like*" phrasing and formal tone (P8, P19, P41), which clashed with the informal style of hate speech, making responses feel distant or even unintentionally ridiculous (P5, P7, P13, P21, P22, P24, P26, P39, P40, P46). Several highlighted that counterspeech should adapt to the norms and dominant hate narratives of each platform (P7, P17, P18, P24, P34), and one even suggested adopting more implicit, moderate expressions aligned with Chinese cultural traditions (P10).

Interestingly, participants sometimes rate low due to perceived inappropriateness of rebutting certain posts. When hate speakers were merely sharing personal opinions with limited hostility, participants felt that counterspeech could appear excessive or misdirected, thereby creating unnecessary tension in an otherwise normal discussion space—"*It really felt like attacking someone*" (P4). Others noted cases where Civilbot misinterpreted a post and delivered a counterspeech based on that misunderstanding (P21). Preferences also diverged: while some participants favoured addressing factual inaccuracies (P23), others argued that biased or emotionally charged comments likewise warranted a response (P11, P12).

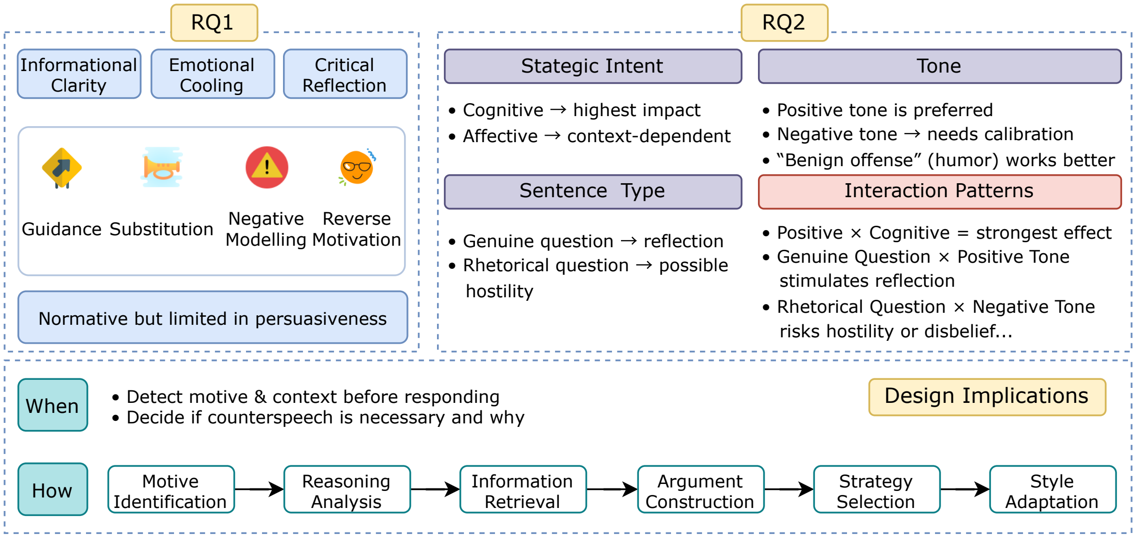
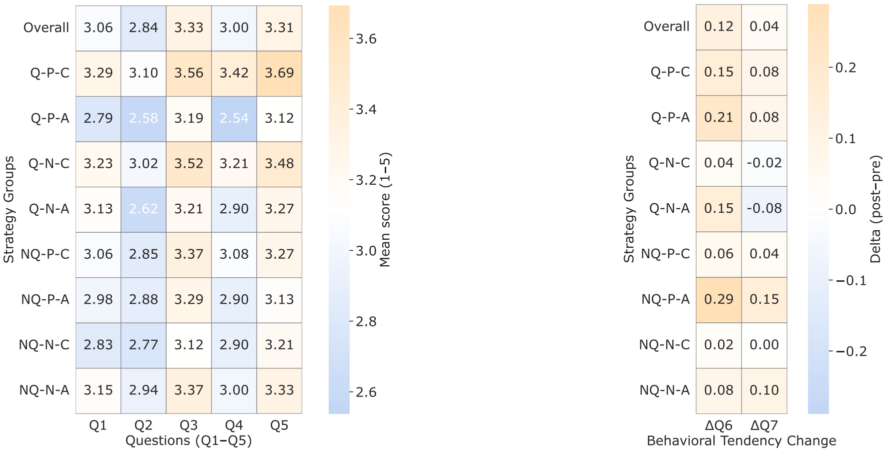

RQ1
Informational Clarity
Emotional Cooling
Critical Reflection
Guidance
Substitution
Negative Modelling
Reverse Motivation
Normative but limited in persuasiveness
RQ2
Stategic Intent
• Cognitive → highest impact
• Affective → context-dependent
Tone
• Positive tone is preferred
• Negative tone → needs calibration
• "Benign offense" (humor) works better
Sentence Type
• Genuine question → reflection
• Rhetorical question → possible hostility
Interaction Patterns
• Positive × Cognitive = strongest effect
• Genuine Question × Positive Tone stimulates reflection
• Rhetorical Question × Negative Tone risks hostility or disbelief...
Design Implications
When
• Detect motive & context before responding
• Decide if counterspeech is necessary and why
How
Motive Identification
Reasoning Analysis
Information Retrieval
Argument Construction
Strategy Selection
Style Adaptation

Figure 3: Overview of results for RQ1–R3. RQ1 shows overall effects on bystanders, Civilbot's roles for them, and perceived community-level mechanisms; RQ2 presents strategy-level effects and key interaction patterns; Design implications illustrate design insights of Civilbot from the perspective of bystanders, organized by when to counterspeak and how to counterspeak.



(a) Mean Score of Q1-5

(b) Mean Score of Delta Q6-7

Figure 4: Heatmap of the correlation between mean scores of different variables and the eight strategy groups. On the y-axis, Q/NQ indicates question or non-question sentence type, P/N indicates positive or negative tone, and C/A indicates cognitive or affective strategic intent. Variables on the x-axis: Q1 (convincing reason), Q2 (strong reason), Q3 (credibility), Q4 (importance), Q5 (overall agreement), Q6 (confidence in countering) and Q7 (willingness to participant)

> In summary, participants generally viewed Civilbot as a legitimate normative voice, yet its persuasiveness was constrained by weak reasoning, rigid expression, and limited sensitivity to context. These findings highlight the need for future designs to balance normative authority with adaptive context-awareness.

5.1.2 Mixed Behavioural Effects: Guidance, Substitution, Negative Modelling, and Reverse Motivation. On the behavioural dimension, Civilbot showed only a modest positive effect on confidence in countering (Q6) and willingness to participate (Q7). This limited impact is understandable: behavioural change is shaped by complex factors such as knowledge, empathy toward targets, and communication habits, and is more likely to evolve over time. Interviews further revealed that Civilbot's influence on bystanders was subtle and often paradoxical, ranging from a guidance effect ("Civilbot did well, so I can join in"), to a substitution effect ("Civilbot did well, so I don't need to act"), a negative modeling effect ("Civilbot did poorly, so I lack confidence to respond"), and even a reverse motivation effect ("Civilbot did poorly, so I should step in to supplement"). These patterns are closely tied to the varied reasons bystanders choose silence in the first place.

As a **guidance**, Civilbot provided alternative perspectives that encouraged reflection and sometimes lowered the threshold for participation. Participants who felt under-informed or cautious about debating hate speech noted that chatbot interventions helped keep the conversation alive: "*Its comments can serve as a starting point, attracting more people to join in and keeping an anti-hate atmosphere*" (P10). Others felt inspired by specific arguments, which provided cognitive scaffolding for their own contributions (P40, P55). For instance, in a scenario involving hate speech accusing Koreans of being stingy and arrogant, P27 noted: "*It reminded me not to generalize about all Koreans, and then I thought of examples to use in my own counterspeech.*" Similarly, Civilbot's responses sparked empathy toward targeted groups (P4) or curiosity about hate speakers' motives (P6, P52), thereby motivating engagement. In a case attacking nurses for alleged poor conduct, P4 reflected: "*At first I thought nurses had nothing to do with me, but after Civilbot's response, I realized how unfair it was, especially remembering their effort during COVID-19*". In these instances, Civilbot's intervention served to remind participants of their social accountability or offered heuristics on how to construct an effective counterspeech.

Civilbot also acted as a **substitution**. For participants who opposed hate speech but feared conflict, its presence reduced the pressure to respond personally. Many described "*liking counterspeech comments*" as a low-cost way to register opposition without direct confrontation (P4, P7, P11, P16, P17, P18, P24, P32, P48). "*I don't want to be the pioneer or opinion leader. I just want my like to show the hate speaker they're wrong, and I hope the counterspeech comments get tens of thousands of likes while the hate speech gets only a few*" (P4). Others even wished Civilbot could serve as their spokesperson, absorbing personal viewpoints and expressing them on their behalf (P10, P57). Participants highlighted that, unlike humans, chatbots never tire, remain emotionally unaffected, and can consistently respond to recurring hate topics—an advantage for long-term engagement (P2, P5, P18, P36). For many, Civilbot's presence provided reassurance that they were part of a larger anti-hate majority (P20, P33, P54, P56, P57).

At the same time, Civilbot sometimes functioned as a **negative modelling** or **reverse motivation**. When its interventions were ineffective, some participants perceived it as a negative modelling and reported decreased confidence in their own ability to intervene. For example, regarding hate speech labeling Asian Americans as "weak", P23 noted: "*AI knows this topic better than me, yet it performs poorly… I don't know what to do.*" Yet others felt compelled to "step in" precisely because Civilbot's contribution fell short: "*After reading its comment, I wanted to add my own—not because it was good, but because it missed the point, and I couldn't resist correcting it*" (P31). In this way, Civilbot inadvertently created a kind of psychological safety net, acting as a first responder that reduced the perceived risk of joining in. This divergence likely stems from the interplay between the participant's self-efficacy in countering and their perceived efficacy of Civilbot. When Civilbot performs poorly, participants with higher self-confidence—often linked to topic familiarity, writing capability, or the perceived weakness of the hate speech—feel motivated to correct the error. In contrast, those with lower confidence may interpret Civilbot's failure as a signal of the task's difficulty, leading to withdrawal.

Finally, participants noted that the behavioural impact of counterspeech may be constrained by external factors such as content moderation. Some worried that over-policing could mistakenly punish counterspeech (P9), underscoring the need for clearer distinctions between hate speech and counterspeech. At the same time, they emphasized that Civilbot's role is not to compel or entice every bystander to actively respond. It establishes a reliable baseline of opposition to hate, providing essential reassurance and symbolic justice even for bystanders who prefer indirect forms of participation, such as liking a counterspeech comment or simply observing its presence (P24).

> In summary, Civilbot's behavioural influence is complex and nuanced. It can inspire reflection and participation, substitute for silent bystanders, undermine confidence through weak responses, or motivate corrective action when it fails. Its value lies not in mobilizing all audiences to counterspeak, but in sustaining a minimal yet consistent countervoice that anchors community norms and offers bystanders a safer space to position themselves.

5.1.3 Shaping Community Climate through Informational Clarity, Emotional Cooling, and Critical Reflection. Beyond individual actions, participants emphasized Civilbot's broader role in shaping community climate. They valued its ability to provide timely information and perspectives, preventing hateful misinformation from misleading uninformed bystanders, especially vulnerable groups such as teenagers (P9, P15, P20, P26, P32). As one participant noted, "*Its greatest value is helping neutral people realize the truth*" (P32). Because the information was presented as reasoning rather than instructions, some participants felt it was more credible—they could follow the logic themselves and thus trust their own conclusion (P23), and several even reported learning laws, theories, or statistics from its comments (P13, P20, P21, P30, P56).

Civilbot also helped stabilize emotions and discourse. Counterspeech could cool overheated reactions, interrupt escalating disputes, and prevent large-scale mobilization of hate, partly because hate speakers were unlikely to engage directly with a chatbot (P24, P29, P52). For some (P20, P56), this provided emotional relief: "*At first I believed the hateful comment and felt angry at Koreans, but Civilbot's response calmed me down and reminded me of a more positive perspective*".

Finally, Civilbot fostered critical reflection and dialogue by presenting opposing viewpoints, which encouraged bystanders to contribute. As one participant observed, "*When two opposite opinions are put forward, people naturally start to discuss around them*" (P19). In this way, counterspeech functioned not only as correction but also as a catalyst for community-level discussion (P22, P28, P43, P47, P56).

> In summary, Civilbot shaped the community climate by (1) offering information that supports recognition of hate and misleading cognition, (2) cooling emotions and preventing escalation, and (3) stimulating dialogue through diverse perspectives.

> These functions positioned it less as a debater and more as a balancer of knowledge, emotion, and reflection, thereby helping consolidate community values at a broader level.

5.2 RQ2: Influences between Counterspeech Strategies

RQ2 revealed differentiated effects of counterspeech strategies across perceived quality, subjective acceptance, and behavioural tendencies. Among the three measurements, intent proved most decisive: cognitive strategies generally outperformed affective ones, though the latter retained situational value, especially when sequenced adaptively—for instance, a positive affective move could ease emotional tension before a cognitive argument. Tone shaped behaviour more than perception. Positive tone enhanced confidence in countering and willingness to participate, while negative tone required careful calibration, working best as "benign offense" (e.g., humor) rather than direct attack. Question forms showed significant effects overall but carried nuanced potential: genuine questions could stimulate reflection, whereas rhetorical ones risked hostility or disbelief. Importantly, interaction effects revealed that positive tone amplifies the strength of cognitive strategies, while affective strategies often falter in this register.

5.2.1 Strategic Intent Shapes Perceived Quality, Subjective Acceptance, and Behavioural Tendencies. Participants in interviews consistently valued cognitive strategies that encouraged reasoning through offering facts and correcting misconceptions. These strategies aligned with their expectation that Civilbot should provide knowledge to prevent the spread of harmful narratives and to stimulate reasoned debate. This finding was also consistent with the quantitative results, which indicated that strategic intent significantly influenced perceived quality ($F = 18.59, p < 0.001$, Cohen's $f = 0.21$) and subjective acceptance ($F = 24.47, p < 0.001$, Cohen's $f = 0.24$) (see Table 3). Suggestions for improvement included digging deeper into the motives of hate (P7, P21, P23), using logical and focused arguments (P1, P6, P11, P12, P17, P29, P30, P34, P37, P52), combining theory with concrete cases (P5, P9, P17, P19, P22, P24, P27, P30, P32, P33, P36), and even incorporating external links (P22, P23, P28, P31) or visual aids (P9). As one participant put it, "*Charts would catch more attention*" (P9).

By contrast, affective strategies elicited more nuanced and sometimes polarized reactions. For some, personal preference dictated rejection of emotional appeals, as they favored objective debate (P4, P9). However, these participants occasionally rated affective counterspeech highly when it successfully evoked empathy through meaningful associations—such as recalling the sacrifices of nurses during COVID-19 (P4), connecting to media portrayals (P9), or invoking shared human values like mutual respect (P5, P40, P53). These examples suggest that affective strategies can work when they trigger authentic connections rather than relying on generic moralizing. Other participants, however, found affective strategies expressed by a digital chatbot less acceptable, perceiving them as hollow or even absurd (P42, P45). For example, statements like "I stand with X" were dismissed as meaningless because Civilbot "*is just a void machine*" without the capacity for real-world solidarity: "*I don't care who it stands with; it's just typing words. If a real person says it, I believe they mean it. But the robot? It changes nothing*" (P9). Such responses highlight that emotional expressions risk backfiring by unintentionally reminding participants of Civilbot's non-human identity, which may ultimately undermine credibility instead of fostering closeness. Similar scepticism extended to first-person pronouns, which some saw as dissonant when used by an AI-driven chatbot (P7).

> Taken together, participants suggested that cognitive and affective strategies are not mutually exclusive but rather complementary. Cognitive approaches are necessary when clear misconceptions must be corrected (P18, P20), but affective appeals may be effective in engaging hostile speakers or indifferent bystanders (P5). When paired with a negative tone, affective strategies were considered more persuasive if delivered after factual reasoning, so as not to appear as mere venting (P26). Ultimately, participants envisioned an adaptive combination of strategies, tailored to the audience and context, as the most effective form of counterspeech (P18, P20).

5.2.2 Tone Influences Behavioural Tendencies, but Its Effect Depends on Context. Overall, participants tended to favour a positive tone, despite conceding that a negative tone was more effective at capturing attention. Conditions such as Q-P-A and NQ-P-A scored relatively high in behavioural items (see Fig. 4). This pattern, as revealed by the qualitative data, suggests potential reasons. First, if Civilbot counters hate with hostility, it risks normalizing aggression and even raising ethical concerns about "*a machine attacking a human*", especially when the hateful comment is not strongly malicious (P9, P11, P21, P22, P26, P29, P55). Second, negative tones may contribute to hate speakers' resistance, escalate conflicts, and inadvertently harm innocent bystanders or cautious victims who are already vulnerable in public discussions (P6, P11, P14, P16, P20, P22, P23, P30, P32, P43). However, positive tones are not a universal solution: when paired with extreme hate, they risk seeming absurdly mismatched and failing to engage anyone (P9, P28, P38).

Some participants emphasized that negative tones, while potentially effective, should be employed with restraint and caution. A degree of sharpness could help capture attention, convey emotion such as a sense of justice, and assert community norms. As P13 explained: "*A stronger tone like this can directly bring more people who are watching into an emotion like yours, and it can make them choose to stand on the side with a stronger tone towards you. They may be more inclined to believe it.*" Yet when overused, negativity risked souring the atmosphere and discouraging participation in discussions (P7, P10, P21, P25, P56, P57).

> Ultimately, participants called for context-sensitive tone management. When hate is relatively weak, a gentle tone can reassure bystanders that meaningful dialogue is possible (P5). A balanced mix of positivity and negativity was considered ideal, drawing attention without succumbing to toxic dynamics. Some highlighted "benign offense", such as humorous sarcasm, as a particularly effective middle ground: "*Although I [Civilbot] am attacking you, the 'attack' is in quotes—it's playful, so you can attack me back. That makes it a closer form of exchange*" (P17).

Table 3: Main Effects of the Three Counterspeech Dimensions across the Three Measurements

Measurement Direction	Factor (Counterspeech Dimension)	F-value	p-value	$Cohen's f$
Perceived Quality	Sentence	9.096	**0.003****	0.148
	Tone	2.659	0.104	0.080
	Intent	18.589	<**0.001*****	0.212
Subjective Acceptance	Sentence	5.189	**0.023***	0.112
	Tone	0.313	0.576	0.027
	Intent	24.474	<**0.001*****	0.243
Behavioural Tendency	Sentence	0.011	0.915	0.005
	Tone	1.662	0.198	0.063
	Intent	0.564	0.453	0.037

Note: * $p < .05$, ** $p < .01$, *** $p < .001$.

Table 4: Interaction Effects between Counterspeech Dimensions across the Three Measurements

Measurement Direction	Interaction (Counterspeech Dimension combination)	F-value	p-value	$Cohen's f$
Perceived Quality	Sentence × Intent	0.022	0.882	0.007
	Sentence × Tone	0.472	0.492	0.034
	Intent × Tone	4.756	**0.030***	0.107
Subjective Acceptance	Sentence × Intent	0.188	0.665	0.021
	Sentence × Tone	0.860	0.354	0.046
	Intent × Tone	1.071	0.301	0.051

Note: * $p < .05$, ** $p < .01$, *** $p < .001$.

Table 5: Simple Effects Analysis of Intent at Different Levels of Tone (Perceived Quality)

Comparison	F-value	p-value	$Cohen's f$
Positive-Affective vs Positive-Cognitive	24.198	<**0.001*****	0.343
Negative-Affective vs Negative-Cognitive	2.079	0.151	0.100

Note: * $p < .05$, ** $p < .01$, *** $p < .001$.

5.2.3 *Question Forms May Encourage Reflection but Risk Backfiring.*

Question-based sentence types were identified by participants as a potentially effective counterspeech strategy, and specific observations were made regarding the further influence of the question's precise form and matched tone. This finding was also consistent with the quantitative results, showing that sentence type significantly affected two measurements: perceived quality ($F = 9.10$, $p = 0.002$, Cohen's$f = 0.15$) and subjective acceptance ($F = 5.19$, $p = 0.023$, Cohen's$f = 0.11$). Genuine questions were valued for inviting reflection and broadening dialogue, often pairing well with positive tones to create a persuasive, approachable style (P7, P26, P42, P53). By presenting multiple possibilities and adding new information, such questions could encourage critical thinking among bystanders. In contrast, rhetorical or confrontational questions functioned less as invitations to reasoning and more as attacks. These were frequently perceived as rude or untrustworthy, yet when combined with negative tones, they conveyed emotional intensity and appeared forceful (P9, P20, P47).

Participants also noted that questions could help uncover the logic or motives behind hate speech, exposing inconsistencies (P15, P22, P28, P46). However, others warned that probing too deeply might backfire: hate speakers could use the opportunity to cite further evidence, reinforcing stereotypes if counterspeech failed to respond effectively (P22, P56). This illustrates the double-edged nature of questioning. While critical inquiry carries risks, participants stressed that its value lies less in "defeating" hate speakers than in sustaining constructive discussion. From this perspective, withholding responses for fear of giving opponents more space is counterproductive, as counterspeech can still guide bystanders and prevent dialogue from being dominated by hate. Therefore, Civilbot is expected not just to confront, but to supplement, guide, and mediate.

> In short, quantitative results showed significant effects, and qualitative analysis revealed that questions can either stimulate critical reflection or undermine credibility, depending on how they are designed. For Civilbot, questions work best as tools for providing information, guiding reflection, and sustaining dialogue, rather than as blunt instruments of confrontation.

5.2.4 Positive Tone Amplifies the Advantage of Cognitive Strategies.

Participants pointed to a potential relationship between tone and cognitive strategies. Cognitive strategies delivered in a negative tone risked being misinterpreted as emotional venting. While a negative tone could attract attention (P7, P9, P10, P21, P57), this attention might not be directed towards the reasoning itself, and might even be counterproductive. As P26 noted: "*You need to be mindful of your [Civilbot's] tone. If you come in strongly with resentment, others may not have the patience to read what you specifically said, believing it's just venting.*" This perceived departure from neutrality, akin to "*personal expression*", might further undermine Civilbot's credibility as a symbol of community social norms (P16, P22). Conversely, the combination of a positive tone with cognitive strategies raised the hope of constructive discourse (P5): it clarifies or provided information as material for discussion while the positive attitude signalled the potential for further communication rather than confrontation. As P33 pointed out: "*Reasoning need to be delivered with a rational attitude.*" These findings are consistent with the Interaction analyses (see Table 4), which revealed that in perceived quality, a significant interaction was observed ($F = 4.76, p = 0.030$, Cohen's $f = 0.11$). Furthermore, subsequent Simple-effect tests echoed this pattern: under positive tone, cognitive strategies substantially outperformed affective ones ($F = 24.20, p < 0.001$, Cohen's $f = 0.34$), whereas under negative tone the two did not differ ($F = 2.08, p = 0.150$, Cohen's $f = 0.10$). Participants explained that positive-tone cognitive counterspeech was perceived as more objective, fair, and conducive to discussion (P12, P16, P22). In contrast, affective strategies often struggled to evoke genuine empathy in a positive register (as discussed in Section 5.2.2), which might weaken their persuasive power.

> Overall, these findings indicate that cognitive strategies are amplified by positive tone, which reinforces their credibility and constructive potential. Affective strategies, by contrast, do not enjoy the same boost and may come across as superficial when expressed positively. Under negative tone, strong emotional expression tends to overshadow strategic intent, blurring distinctions between cognitive and affective strategies. For Civilbot, this underscores the importance of aligning intent with tone: negative tone must be used with caution—while it can sometimes blunt strategic differences and undermine persuasion, it may also, in certain contexts, intensify affective influence, whereas positive tone should be prioritized to maximize the impact of cognitive counterspeech.

5.2.5 Exploratory Analyses: Positive-Affective Questions, Non-Questions, and Human-AI Dynamics.

Pairwise t-tests across the eight strategy groups revealed several exploratory patterns (see Fig.5), which should be read as suggestive rather than causal.

In perceived quality and subjective acceptance, Q-P-A (positive-tone affective questions) performed worse than most other groups, with lower ratings of quality, credibility, importance, and agreement. Participants often perceived such utterances as idealized appeals that lacked authentic emotional resonance. On the one hand, positive-tone affective strategies were seen as too mild for condemnation yet insufficient for empathy. On the other hand, positive-tone questions leaned toward open-ended inquiry without adding substantive reasoning. As a result, this combination was generally weak, though some still valued its normative stance when responding to mild hate speech: "*at least it conveys the right idea*" (P9, P20, P21, P27). Overall, NQ-N-A (non-question, negative, affective) performed relatively worst, yet relatively positive on willingness to participate (Fig. 4). This apparent contradiction aligns with earlier findings in Section 5.1.2 that poorly received counterspeech can serve as negative exemplars, motivating participants to craft their own counterspeech.

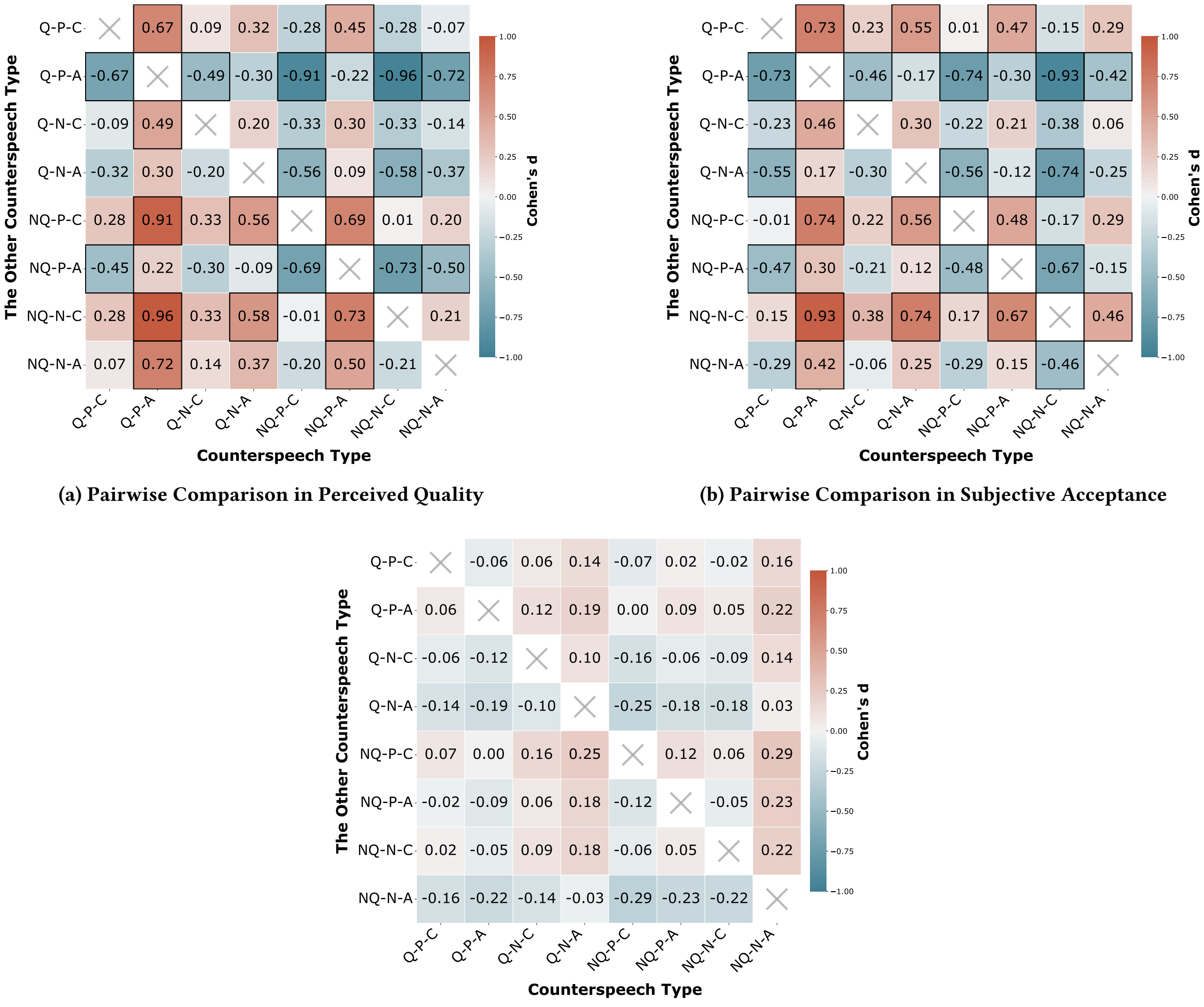
Additional analyses of likes and participant-generated counterspeeche further nuanced this picture (see Fig. 6). Q-P-A again received the fewest likes, consistent with its weak reception. Interestingly, participants' own counterspeech showed a preference of negative tone, diverging from their expectations of Civilbot. This reflects Civilbot's dual role: as a non-human chatbot and a symbolic actor of social norms, it is expected to remain objective and norm-affirming, avoiding slightly harsh negative tone. However, it is contrasted with human counterspeech, where negative tone is often valued as a tool of justice and emotional release. For participants, harsh counterspeech serve pragmatic purposes—raising the cost of hate (e.g., by making it risky or unpleasant to speak), disrupting hate speakers' goals of silencing targets, and in some cases even shifting or suppressing hate speakers (P5, P7, P16, P17, P19, P20, P24, P26). As one participant explained, "*my own harsher counterspeech helps me vent, but I'd want AI to play a supporting role*" (P24).

6 Discussion

In this section, we integrate the findings and derive design implications for Civilbot. Specifically, we discuss two core dimensions: when to intervene—that is, whether counterspeech should be deployed and for what reasons; and how to intervene—that is, how counterspeech reasoning, evidence, and organization should be constructed. We also discuss the role of identity in the influence of chatbot counterspeech, and conclude with limitations and future directions.

6.1 Boundaries of Chatbot Counterspeech: When and Why to Intervene

Our results highlight participants' views on the scope of counterspeech and its complementarity with content moderation. In cases of extreme hate speech, counterspeech was often perceived as ineffective or even harmful. This is because posts containing slurs or large-scale harassment often leave little room for discussion; instead, counterspeech may risk reinforcing their salience, distract attention from urgent responses [14], and even accelerate their spread due to platform recommendation algorithms. As P4 noted, "*There's no point in arguing with such extreme attacks. They're just clowns*". In these cases, removal or suspension was considered more appropriate, a judgment many participants voiced explicitly from a bystander standpoint. In other words, while counterspeech offers greater flexibility, it is ill-suited for deeply motivated or extreme hate. Conversely, in mild or borderline cases, counterspeech was not always necessary, and when used, needed to be carefully matched in tone. Hate expressed out of ignorance or self-deprecating humour, if met with overly critical counterspeech, could push users toward



(a) Pairwise Comparison in Perceived Quality

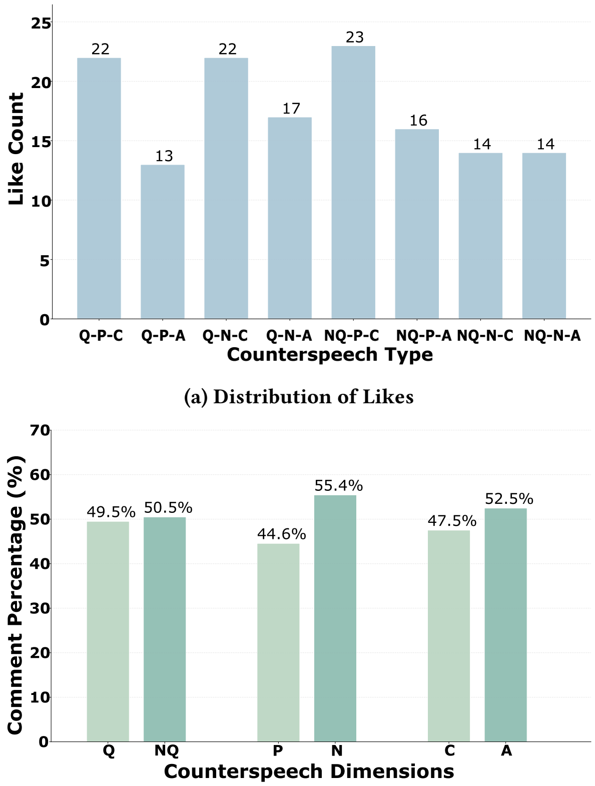
(b) Pairwise Comparison in Subjective Acceptance

(c) Pairwise Comparison in Behavioural Tendencies

Figure 5: Heatmap of pairwise paired t-tests between counterspeech types across the three questionnaire measures. Significant comparisons are outlined with bold borders. Cell values represent effect sizes (Cohen's d). Perceived quality = mean(Q1–2); subjective acceptance = mean(Q3–5); behavioral tendencies = mean(ΔQ6–7) (Δ = post − pre).

opposition and harm the community climate. As P7 explained, "*(If I were that hate speaker,) I might just be expressing my view without strong malice, but Civilbot framed me as if I had bad intentions. I would feel wronged, or uncomfortable.*" Several participants described this reaction while observing others' exchanges, indicating it was a bystander impression rather than only the hate speakers' concern. This suggests the importance of distinguishing between different stages of stigmatization [51]: initial labelling versus entrenched separation motives call for different responses. For users repeatedly engaging in hate, counterspeech alone is unlikely to change their behaviours, and punitive measures may be warranted (P9, P52, P58).

Within the space where counterspeech is considered appropriate, participants also cautioned against "countering every hate speech". Excessive automation could discourage genuine user engagement. Instead, counterspeech was seen as most valuable when directed at high-visibility and potentially disruptive posts, those that often set the tone of discussion and accelerate diffusion. As P22 put it, "*Only high-attention hate speech needs to be countered, because it already exerts a significant influence.*" This emphasis on visibility came from participants speaking as third-party observers, highlighting what bystanders notice. In such cases, counterspeech was perceived not only as a corrective signal superior to deletion but also as a way to neutralize the dominance of hate. Participants further suggested a proactive role for Civilbot at the early-warning stage—for example, predicting when a hateful post is likely to trigger toxic replies or attract huge attention, and intervening before escalation. As P5 proposed, "*Civilbot could be preventive, responding before harmful*



(a) Distribution of Likes

(b) Distribution of Comments

Figure 6: Distribution of participants' interactions across different counterspeech strategies.

speech causes damage." Prior research on conversational structure for toxicity detection [70], early signals of antisocial behaviour [86], and the PMCR framework [31] offer useful references for identifying intervention opportunities. Building on these approaches, Civilbot could further adopt methods from explainable toxicity detection [49] to justify "*why counterspeech is needed here*" as P35 suggested, thereby enhancing credibility for bystanders.

Overall, decisions about whether to counter hinge on two critical factors: the allocation of attention and the risk of conflict. On the one hand, counterspeech may unintentionally draw more traffic to hate speech [14], underscoring the need to prioritize high-attention or high-risk content. On the other hand, avoiding counterspeech due to possible conflicts risks falling into "negative peace" [44], where harmful ideas spread unchallenged. The desirable balance lies in using a constructive tone and strategic design to transform attacks into dialogue. As P5 reflected, "*When the hateful comment wasn't too emotional and Civilbot also responded in a mild tone, it gave me confidence that the hate speaker's view could actually be changed. It felt like we could have an exchange and reach a warmer, more peaceful outcome.*"

6.2 Designing Effective Counterspeech: From Reasoning to Style Adaptation

Drawing on experimental feedback, we propose an operational workflow for counterspeech: motive identification ⟶ reasoning analysis ⟶ information retrieval ⟶ argument construction ⟶ strategy selection ⟶ style adaptation. This workflow outlines the design space of Civilbot in answering the question of how to counter, raising three core considerations: what to counter, what evidence to use, and how to organize the response. In the organizational dimension, we further discuss strategy selection, stylistic adaptation, and opportunities beyond text.

6.2.1 Identifying What to Counter: Motives and Weak Points. Findings in Section 5.1.1 underscore the importance of uncovering the motives behind hate speech. Motives not only shape whether a counterspeech is needed but also determine how it should be constructed. For example, when hate stems from cognitive limitations, cognitive strategies—such as prompting critical reflection through questions—can be effective. When driven by negative emotions rooted in personal experience, affective strategies may help elicit more positive affect. If hate is merely attention-seeking, a combination of appropriate negative tone and condemnatory strategies may be more suitable. As P20 observed, different problems demand different approaches: "*If it's ignorance, use rational data; if it's emotional venting, emotion works better.*"

Ignoring these motives risks superficial responses that fail to reach the core, weakening persuasiveness and even making bystanders persuaded by harmful content. In discussions on social issues, participants also emphasized the need for Civilbot to demonstrate sufficient depth of reasoning, without which it would not be taken seriously. As P4 noted, Civilbot's replies sometimes felt "*too vague*", failing to directly address the hate speaker's claims. In short, effective counterspeech must go beyond surface-level wording to engage with the hate speaker's motives, experiences, and reasoning, revealing implicit meanings. This calls for Civilbot to develop deeper interpretive capacities—for instance, handling implicit communication [50] and employing argument schemes [68] to identify hidden assumptions and logical fallacies that can serve as entry points for counterspeech.

Prior work has suggested that personalizing counterspeech with limited user information [25] could further expose motives and enhance persuasion. Yet such personalization raises ethical concerns and technical costs, warranting caution. At the same time, some participants worried that focusing too heavily on the hate speaker's motives might alienate bystanders or targets of hate. As P23 warned, "*The closer you get to the hate speaker's motive, the farther you may be from the other readers.*" Our interpretation is that motive analysis should not be about appeasing the hate speaker but about sharpening the persuasiveness of counterspeech, especially toward bystanders.

Moreover, as P20 emphasized, Civilbot's role is not necessarily to confront hate speakers individually but also to "*assert and maintain community positions*". Counterspeech should therefore aim to resist hate while sustaining constructive discussions on social issues. From this perspective, Civilbot's focus on motives is not merely about "defeating" hate speakers, but about communicating community norms, preserving deliberative climates, and—in certain cases—creating pathways for re-entry of those who move away from hate [38].

6.2.2 *Enhancing Arguments through Information Support.*

After analyzing hate speech, Civilbot must determine its arguments and supporting evidence. Results in 5.1.3 show that grounding counterspeech in information support was widely seen as promising. Participants expected Civilbot to integrate theoretical knowledge from psychology, sociology, or law, as well as authoritative data and credible news examples, thereby enhancing credibility and significance while helping Civilbot articulate positions and substantiate claims. From a technical standpoint, augmenting counterspeech with external knowledge bases or real-time retrieval was viewed as feasible, aligning with the paradigm of knowledge-guided generation [16]: retrieving relevant information first, then generating enhanced counterspeech. Another pathway is to leverage curated counter-hate datasets (e.g., candidate arguments extracted from online articles [1]) or fine-tune models with high-quality human-authored counterspeech, making Civilbot's output more faithful to authentic user discourse.

Notably, participants already rated Civilbot as moderately credible, likely because it mainly voiced general perspectives rather than relying heavily on factual references. However, as more factual arguments are incorporated, knowledge augmentation will be essential—both to improve persuasiveness and to mitigate hallucination risks [1]. Some participants further stressed that Civilbot should provide facts rather than opinions, enabling viewers to reason independently and thereby reinforcing credibility. This highlights the necessity of complementing counterspeech with factual evidence.

6.2.3 *Choosing Effective Strategies in Context.*

From sentence type to tone to strategic intent, results in Section 5.2 indicate that context-sensitive strategies are crucial.

For **sentence type**, questions showed the greatest potential when combined with cognitive strategies: by offering arguments and evidence, they expanded reasoning space and, supported by positive tone, fostered a guiding atmosphere that encouraged critical thinking among both hate speakers and bystanders. Conversely, when paired with negative tone, questions resembled affective strategies of reproach or denunciation. While such forms attracted attention and satisfied some bystanders' sense of justice, their implicit insincerity often suppressed reflection and provoked resistance. Regarding **tone**, negative expressions were found effective in certain contexts but carried risks. As a non-human agent, Civilbot using negativity could unintentionally normalize hostility or weaken community climate. Emotional countering may therefore be better suited for human users. For Civilbot, a safer design is to prioritize positive or neutral tone while enabling adaptive modulation: in less extreme cases, constructive tone conveys hope for dialogue; in more intense exchanges, restrained negativity or humor-driven benign offense can achieve balance without eroding atmosphere. On **strategic intent**, results underscored the centrality of cognitive strategies: fact- and reasoning-based counterspeech was key to persuasion and, when paired with positive tone, conveyed objectivity and neutrality. Affective strategies, however, retained supplementary value—empathy-based appeals could ease tension before reasoning took effect, or intensify condemnation after. Together, these findings suggest Civilbot's strategy selection must achieve higher contextual sensitivity, integrating not only the hate speech itself but also broader cues from conversational background, participant interaction, and community climate.

Beyond the strategies synthesized here, broader frameworks are also instructive. For instance, persuasion-oriented approaches [68] include value-based and structure-based strategies. In our framing, argument schemes are incorporated in Section 6.2.2 under reasoning, while value-oriented strategies are discussed in Section 6.2.4 as part of Civilbot's role design. Other categories such as denouncing or positive tone naturally align with this section's discussion of contextualized choices.

6.2.4 *Adapting Style to Communities and Roles.*

Once counterspeech is equipped with clear motives, solid arguments, and strategic combinations, the next challenge is adapting to specific platforms and cultural contexts. Different platforms host distinct user groups and content norms, shaping prevailing values and hate discourses. For instance, a sexist remark toward women might be challenged on Rednote but endorsed on HoopCHINA. Community conventions, expressive styles, and moderation policies also vary. Thus, Civilbot must tailor its counterspeech to the linguistic style and dominant hate topics of each platform, ensuring it integrates with community norms while steering value transmission. More broadly, culture shapes preferences of effective counterspeech: for example, the Chinese tradition of indirectness and metaphor suggests that counterspeech should align with social communication norms to avoid being perceived as biased [55].

Participants further emphasized the potential of role diversification. Civilbot need not remain solely an "enforcer" but could also act as a mediator or knowledge provider. Some even suggested that engaging with high-quality comments might contribute more to constructive debate than directly confronting hate. This implies that Civilbot's role should adapt to situational demands and strategy choices. In our experiment, we deliberately used a neutral avatar and username to avoid over-anthropomorphizing. Yet, profile design (e.g., avatars) could be leveraged for role embodiment, enhancing style and positioning [43], potentially boosting Civilbot's influence and even enabling it to function as an opinion leader. Such role flexibility was noted by several participants specifically from a bystander viewpoint, who valued seeing Civilbot model diverse constructive roles within the community. Role diversification may also balance two competing needs: serious, restrained expression to uphold norms, and more flexible, creative expression to attract attention and foster an anti-hate climate. Future work could explore persona-guided generation, such as configuring dynamic roles through datasets like PersonaChat or leveraging dialogue history for richer role expressions [16].

6.2.5 *Beyond text: Emerging Modalities.*

Currently, only text-based counterspeech is supported by Civilbot, but participants widely envisioned richer modalities. Visualizations, for example, can present arguments more intuitively and capture attention more constructively than negative tones. At the same time, hate is often spread through images [24], such as malicious memes, posing new detection and response challenges. Integrating multimodal generation—combining text with images—thus represents a promising direction. Beyond text and images, broader design opportunities are opened by interactive digital narratives (IDN) and related formats such as video games and VR/AR/XR [74]. Anti-hate narratives can be delivered through these media in immersive, participatory environments, extending Civilbot's reach beyond textual dialogue and community threads into richer experiential domains.

6.3 Identity in Counterspeech: How It Shapes Chatbot Intervention

Our findings reveal the impact of Civilbot on bystanders when employing various counterspeech strategies. It is crucial to clarify that this impact stems neither solely from the "strategy itself" nor merely from the fact that it is "chatbot-mediated," but rather from the interplay of both. As noted in Section 5.2.2, bystanders perceive Civilbot differently from humans even when identical tones are used. For instance, when Civilbot adopted a hostile tone, participants (e.g., P5, P9, P24, P26) raised ethical concerns regarding "*machines attacking humans*". P24 explicitly preferred taking personal ownership of intense negative expressions rather than delegating them to Civilbot. Similarly, affective strategies (e.g., "I stand with X") triggered resistance among some participants (P7, P9). Civilbot's limited agency made such statements prone to being perceived as over-commitment. Moreover, the use of first-person pronouns or empathetic phrasing paradoxically highlighted the agent's non-human identity. Strategies originally intended to bridge psychological distance [87] instead accentuated identity differences in this sensitive social context, creating a sense of detachment.

These phenomena point to a fundamental issue: the identity of the counterspeaker shapes the intervention's efficacy. This aligns with existing literature suggesting that high-status members or those with clear commitment are more likely to have their counterspeech mimicked [7, 73], and that factors like race and follower count shape a speaker's community influence [56]. In our study, where the counterspeaker is an AI chatbot, the implications of identity are more nuanced. **On the positive side**, Civilbot, as a non-human agent, offers a baseline response free from social and reputational costs, serving as a psychological safety net similar to AI in creative tasks [77]. This baseline response may encourage bystander engagement by lowering entry thresholds (reflecting the reverse motivation noted in Section 5.1.2) or by serving as an "icebreaker". Furthermore, as shown in section 5.1.3, Civilbot might de-escalate emotions. While partly due to strategy, section 5.1.1 suggests identity played a role: participants (e.g., P30) pointed that people were disinclined to argue with a chatbot, preventing conflict escalation. Additionally, the perception of AI as objective may contribute to this effect. **On the negative side**, disclosing the bot identity can trigger bias, which varies depending on the perceived level of control [4]. For instance, some participants in Section 5.1.1 dismissed the AI's output as "*mindless stitching*". This echoes the theoretical distinction that different identities imply different commitments (what it is expected to do) and beliefs (what it is believed capable of doing) [83]. Consequently, the same expressions may yield divergent effects depending on the speaker's identity.

In summary, these insights suggest two key design implications. First, autonomous counterspeech bots (like Civilbot) and AI systems that assist users in writing counterspeech [55] should be treated as distinct design paradigms occupying different social ecological niches. In AI-assisted systems, the speaker remains human; thus, design should focus on collaboration, human-centricity, and even personalization to provide authenticity, emotional support, and empowerment. Conversely, for autonomous bots, the speaker is a non-human agent. Design must therefore prioritize the social perception of the chatbot, managing its role, persona, and strategy selection to optimize its influence on groups like bystanders. Second, the role of identity in counterspeech warrants systematic future research. For example, studies could manipulate perceived identity (perceived as human or bot) alongside conversational style (human or bot), similar to the Ideabot in creativity task [39]. Furthermore, P23 suggested the possibility of a "*hybrid identity*", where Civilbot's content is known to be partially authored by humans without explicitly disclosing the source of each message, which may be a meaningful direction of future research.

6.4 Limitations and Future Directions

This study employed a lab study to examine Civilbot's impact on bystanders, striving to simulate authentic browsing contexts, yet several limitations remain.

- First, the lab environment may suppress natural behaviours: for instance, P19 noted that the lack of privacy led them to remain silent. While anonymity could theoretically mitigate this, in our study researcher presence was unavoidable due to the need for follow-up interviews based on the interaction process. We sought to compensate by eliciting potential comments through interviews. Future work could explore more covert or automated data collection methods to reduce external interference.
- Second, this was a short-term experiment. This design balanced experimental control with participant burden, allowing focus on immediate effects. However, long-term behavioural trends and community dynamics may manifest additional variations. Future studies should conduct longitudinal field study, incorporating natural interactions, such as having hate speech appear in a natural comment sequence, to evaluate Civilbot's sustained impact more comprehensively.
- Third, to isolate strategy impact, methods were employed to mitigate context effects, including offering a pool of questions for selection, randomizing question and strategy order, and anonymizing all posted answers. Nevertheless, we recognize that different contexts (e.g., the hate speech, the question topic, and the hate speaker) may necessitate distinct counterspeech strategies, leading to varied effects. As noted in Sections 6.1 and 6.2.3, the decision to counter and the choice of strategy are linked to the hate speech's tone, intent, and other factors. Furthermore, Section 6.2.1 underscores the importance of discerning the underlying motivations of the hate speech. Future work, therefore, requires a deeper exploration of the compound effect of context × strategy, ultimately aiming to realize a context-sensitive adaptive mechanism for counterspeech strategies.
- Fourth, the study focused on common counter questions, categorizing sentence types as "question" versus "non-question" to highlight the potential of questions in fostering critical thinking. This simplification overlooks potential differences

from statements, exclamations, and other sentence forms. Future work could compare sentence types at finer granularity and incorporate role-based strategies such as value to support more nuanced modelling of counterspeech.

- Fifth, the study concentrated on the Chinese context. Yet communication norms differ across languages and cultures, which may affect counterspeech reception. Cross-linguistic and cross-cultural studies are needed to assess the generalizability of Civilbot.
- Finally, Civilbot relied on pre-generated content rather than real-time detection and generation. While this ensured experimental control, it differs from deployment scenarios. Future research should explore real-time detection, context-aware generation, and integration with platform mechanisms to enhance practical applicability.

7 Ethical Consideration

During the conduct of this study, we prioritized ethical considerations for all participants. The study protocol was reviewed and approved by the university's Institutional Review Board (IRB), with all required materials submitted according to regulations. Participants were fully briefed on the study's purpose, procedures, potential risks, and their rights, and provided informed consent before participation. They were explicitly informed that participation was voluntary and that they could withdraw at any time without penalty. Psychological support was made available if participants experienced discomfort. All collected data were anonymized, securely stored, and transmitted using encryption to ensure privacy and protect participants' identities. Compensation was provided to participants who completed all sessions.

Additionally, to illustrate Civilbot's counterspeech, we included a small number of real hate-speech examples with minimal redaction in this paper. These excerpts were selected solely for research transparency and are not intended to perpetuate harmful language. Readers are advised that these examples may contain sensitive content.

8 Conclusion

Counterspeech is proposed as a non-repressive, socially grounded response to online hate incidents, complementing traditional moderation by enriching rather than restricting public discourse. We constructed a unified framework of common counterspeech strategies across sentence type, tone, and strategic intent. Building on this framework, we designed Civilbot, a prototype chatbot capable of generating diverse counterspeech responses, and conducted a mixed-methods, within-subjects experiment to examine its influence on bystanders. Our findings show that Civilbot shaped bystander responses at multiple levels. It was generally perceived as credible and norm-affirming, though its shallow reasoning constrained persuasiveness. Behaviourally, its influence was subtle and sometimes contradictory—providing guidance, substitution, negative modelling, or reverse motivation—while also extending beyond persuasion to fostering community climate. Strategy proved decisive: cognitive strategies outperformed affective ones; tone might influence behavioural tendencies but required contextual calibration; and question forms, when combined with other strategies, could either stimulate critical reflection or provoke resistance. Taken together, these results point to counterspeech design that centres on cognitive strategies while flexibly combining styles in a context-sensitive manner. These findings inform design directions for future counterspeech chatbots. Beyond deciding when to intervene, effective design must also consider how to structure reasoning, integrate credible evidence, and adapt style to community norms. Expanding beyond text into multimodal formats and flexible role configurations can further enhance impact. Ultimately, such chatbots hold potential to mobilize bystanders and cultivate healthier discursive environments against online hate.

9 Acknowledgments of the Use of AI

We used AI, in particular large language models (LLMs), in the following ways: (1) Dataset annotation and filtering for hate speech: Qwen-Turbo was used to re-screen entries and tag targeted groups. (2) Counterspeech generation: GPT-5 was employed to generate Civilbot responses according to the eight predefined counterspeech strategies, using iterative prompt engineering to refine outputs. Details of these usages are provided in Section 4 and Appendix A. Additionally, the full manuscript was AI-assisted for language polishing and grammar checking only; all content decisions, analyses, and interpretations were made by the authors. Authors take full responsibility for the output and use of AI in this paper.

Acknowledgments

This research was supported by National Natural Science Foundation of China (NSFC) under the Grant No. 62372113. Peng Zhang is a faculty of College of Computer Science and Artificial Intelligence, Fudan University. Tun Lu is a faculty of College of Computer Science and Artificial Intelligence, Shanghai Key Laboratory of Data Science, and Silver-X MOE Philosophy & Social Sciences Laboratory, Fudan Institute on Aging.

References

[1] Abdullah Albanyan, Ahmed Hassan, and Eduardo Blanco. 2023. Finding Authentic Counterhate Arguments: A Case Study with Public Figures. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, Houda Bouamor, Juan Pino, and Kalika Bali (Eds.). Association for Computational Linguistics, Singapore, 13862–13876. doi:10.18653/v1/2023.emnlp-main.855

[2] Ana Aleksandric, Hanani Pankaj, Gabriela Mustata Wilson, and Shirin Nilizadeh. 2023. Sadness, Anger, or Anxiety: Twitter Users' Emotional Responses to Toxicity in Public Conversations. doi:10.48550/arXiv.2310.11436 arXiv:2310.11436

[3] Ana Aleksandric, Sayak Saha Roy, Hanani Pankaj, Gabriela Mustata Wilson, and Shirin Nilizadeh. 2024. Users' Behavioral and Emotional Response to Toxicity in Twitter Conversations. In *Proceedings of the International AAAI Conference on Web and Social Media*, Vol. 18. 29–42.

[4] Zahra Ashktorab, Casey Dugan, James Johnson, Qian Pan, Wei Zhang, Sadhana Kumaravel, and Murray Campbell. 2021. Effects of Communication Directionality and AI Agent Differences in Human-AI Interaction. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems (CHI '21)*. Association for Computing Machinery, New York, NY, USA, 1–15. doi:10.1145/3411764.3445256

[5] Michelle Baddeley. 2010. Herding, Social Influence and Economic Decision-Making: Socio-Psychological and Neuroscientific Analyses. *Philosophical Transactions of the Royal Society B: Biological Sciences* 365, 1538 (Jan. 2010), 281–290. doi:10.1098/rstb.2009.0169

[6] Dominik Bär, Abdurahman Maarouf, and Stefan Feuerriegel. 2024. Generative AI May Backfire for Counterspeech. doi:10.48550/arXiv.2411.14986 arXiv:2411.14986

[7] Susan Benesch. 2014. Countering Dangerous Speech: New Ideas for Genocide Prevention. doi:10.2139/ssrn.3686876 social science research network:3686876

[8] Michael Bennie, Demi Zhang, Bushi Xiao, Jing Cao, Chryseis Xinyi Liu, Jian Meng, and Alayo Tripp. 2025. PANDA – Paired Anti-hate Narratives Dataset

from Asia: Using an LLM-as-a-Judge to Create the First Chinese Counterspeech Dataset. doi:10.48550/arXiv.2501.00697 arXiv:2501.00697

[9] George Berry and Sean J. Taylor. 2017. Discussion Quality Diffuses in the Digital Public Square. In *Proceedings of the 26th International Conference on World Wide Web*. International World Wide Web Conferences Steering Committee, Perth Australia, 1371–1380. doi:10.1145/3038912.3052666

[10] Patrick Biernacki and Dan Waldorf. 1981. Snowball Sampling: Problems and Techniques of Chain Referral Sampling. *Sociological Methods & Research* 10, 2 (Nov. 1981), 141–163. doi:10.1177/004912418101000205

[11] Michał Bilewicz, Patrycja Tempska, Gniewosz Leliwa, Maria Dowgiałło, Michalina Tańska, Rafał Urbaniak, and Michał Wroczyński. 2021. Artificial Intelligence against Hate: Intervention Reducing Verbal Aggression in the Social Network Environment. *Aggressive Behavior* 47, 3 (May 2021), 260–266. doi:10.1002/ab.21948

[12] 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*, Kevin Duh, Helena Gomez, and Steven Bethard (Eds.). Association for Computational Linguistics, Mexico City, Mexico, 3480–3499. doi:10.18653/v1/2024.findings-naacl.221

[13] David Bromell. 2022. Counter-Speech Is Everyone's Responsibility. In *Regulating Free Speech in a Digital Age: Hate, Harm and the Limits of Censorship*, David Bromell (Ed.). Springer International Publishing, Cham, 191–215.

[14] Bianca Cepollaro, Maxime Lepoutre, and Robert Mark Simpson. 2023. Counterspeech. *Philosophy Compass* 18, 1 (2023), e12890. doi:10.1111/phc3.12890

[15] Justin Cheng, Michael Bernstein, Cristian Danescu-Niculescu-Mizil, and Jure Leskovec. 2017. Anyone Can Become a Troll: Causes of Trolling Behavior in Online Discussions. In *Proceedings of the 2017 ACM Conference on Computer Supported Cooperative Work and Social Computing (CSCW '17)*. Association for Computing Machinery, New York, NY, USA, 1217–1230. doi:10.1145/2998181.2998213

[16] Yi-Ling Chung, Gavin Abercrombie, Florence Enock, Jonathan Bright, and Verena Rieser. 2023. Understanding Counterspeech for Online Harm Mitigation. doi:10.48550/arXiv.2307.04761 arXiv:2307.04761

[17] Yi-Ling Chung, Elizaveta Kuzmenko, Serra Sinem Tekiroglu, 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*, Anna Korhonen, David Traum, and Lluís Màrquez (Eds.). Association for Computational Linguistics, Florence, Italy, 2819–2829. doi:10.18653/v1/P19-1271

[18] 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 (WWW '25)*. Association for Computing Machinery, New York, NY, USA, 5022–5033. doi:10.1145/3696410.3714507

[19] Jacob Cohen. 1992. Statistical Power Analysis. *Current Directions in Psychological Science* 1, 3 (1992), 98–101. jstor:20182143

[20] Niklas Felix Cypris. 2024. *The Effectiveness of Counterspeech in Mitigating Online Hate: Insights From a Multi-Method Investigation.* Ph. D. Dissertation. Technische Universität München.

[21] Valdemar Danry, Pat Pataranutaporn, Yaoli Mao, and Pattie Maes. 2023. Don't Just Tell Me, Ask Me: AI Systems That Intelligently Frame Explanations as Questions Improve Human Logical Discernment Accuracy over Causal AI Explanations. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems (CHI '23)*. Association for Computing Machinery, New York, NY, USA, 1–13. doi:10.1145/3544548.3580672

[22] Dominic DiFranzo, Samuel Hardman Taylor, Franccesca Kazerooni, Olivia D. Wherry, and Natalya N. Bazarova. 2018. Upstanding by Design: Bystander Intervention in Cyberbullying. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems (CHI '18)*. Association for Computing Machinery, New York, NY, USA, 1–12. doi:10.1145/3173574.3173785

[23] Xiaohan Ding, Kaike Ping, Uma Sushmitha Gunturi, Buse Carik, Sophia Stil, Lance T. Wilhelm, Taufiq Daryanto, James Hawdon, Sang Won Lee, and Eugenia H. Rho. 2025. CounterQuill: Investigating the Potential of Human-AI Collaboration in Online Counterspeech Writing. doi:10.48550/arXiv.2410.03032 arXiv:2410.03032

[24] Daisy Dixon. 2022. Artistic (Counter) Speech. *The Journal of Aesthetics and Art Criticism* 80, 4 (Sept. 2022), 409–419. doi:10.1093/jaac/kpac038

[25] Mekselina Doğanç and Ilia Markov. 2023. From Generic to Personalized: Investigating Strategies for Generating Targeted Counter Narratives against Hate Speech. In *Proceedings of the 1st Workshop on CounterSpeech for Online Abuse (CS4OA)*, Yi-Ling Chung, Helena Bonaldi, Gavin Abercrombie, and Marco Guerini (Eds.). Association for Computational Linguistics, Prague, Czechia, 1–12.

[26] Joseph L. Fleiss. 1971. Measuring Nominal Scale Agreement among Many Raters. 76, 5 (1971), 378–382. doi:10.1037/h0031619

[27] Gloria Gennaro, Laurenz Derksen, Aya Abdelrahman, Emma Broggini, Mariya Alexandra Green, Victoria Andrea Haerter, Elia Heer, Isabel Heidler, Fiona Kauer, Han-Nuri Kim, Benjamin Landry, Alessio Levis, Jiazhen Li, Şevval Şimşir, Iva Srbinovska, Robin Anna Vital, Karsten Donnay, Fabrizio Gilardi, and Dominik Hangartner. 2025. Counterspeech Encouraging Users to Adopt the Perspective of Minority Groups Reduces Hate Speech and Its Amplification on Social Media. *Scientific Reports* 15, 1 (July 2025), 22018. doi:10.1038/s41598-025-05041-w

[28] Ellen Girden. 1992. *ANOVA*. SAGE Publications, Inc. doi:10.4135/9781412983419

[29] Jawad Golzar, Shagofah Noor, and Omid Tajik. 2022. Convenience Sampling. *International Journal of Education & Language Studies* 1, 2 (Dec. 2022), 72–77. doi:10.22034/ijels.2022.162981

[30] Jarod Govers, Eduardo Velloso, Vassilis Kostakos, and Jorge Goncalves. 2024. AI-Driven Mediation Strategies for Audience Depolarisation in Online Debates. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems (CHI '24)*. Association for Computing Machinery, New York, NY, USA, 1–18. doi:10.1145/3613904.3642322

[31] Nitesh Goyal, Leslie Park, and Lucy Vasserman. 2022. "You Have to Prove the Threat Is Real": Understanding the Needs of Female Journalists and Activists to Document and Report Online Harassment. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems (CHI '22)*. Association for Computing Machinery, New York, NY, USA, 1–17. doi:10.1145/3491102.3517517

[32] 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)*, Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki (Eds.). Association for Computational Linguistics, Toronto, Canada, 5792–5809. doi:10.18653/v1/2023.acl-long.318

[33] Sadaf MD Halim, Saquib Irtiza, Yibo Hu, Latifur Khan, and Bhavani Thuraisingham. 2023. WokeGPT: Improving Counterspeech Generation Against Online Hate Speech by Intelligently Augmenting Datasets Using a Novel Metric. In *2023 International Joint Conference on Neural Networks (IJCNN)*. 1–10. doi:10.1109/IJCNN54540.2023.10191114

[34] Soo-Hye Han and LeAnn M. Brazeal. 2015. Playing Nice: Modeling Civility in Online Political Discussions. *Communication Research Reports* 32, 1 (Jan. 2015), 20–28. doi:10.1080/08824096.2014.989971

[35] Dominik Hangartner, Gloria Gennaro, Sary Alasiri, Nicholas Bahrich, Alexandra Bornhoft, Joseph Boucher, Buket Buse Demirci, Laurenz Derksen, Aldo Hall, Matthias Jochum, Maria Murias Munoz, Marc Richter, Franziska Vogel, Salomé Wittwer, Felix Wüthrich, Fabrizio Gilardi, and Karsten Donnay. 2021. Empathy-Based Counterspeech Can Reduce Racist Hate Speech in a Social Media Field Experiment. *Proceedings of the National Academy of Sciences* 118, 50 (Dec. 2021), e2116310118. doi:10.1073/pnas.2116310118

[36] David Hartmann, Amin Oueslati, Dimitri Staufer, Lena Pohlmann, Simon Munzert, and Hendrik Heuer. 2025. Lost in Moderation: How Commercial Content Moderation APIs Over- and Under-Moderate Group-Targeted Hate Speech and Linguistic Variations. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems (CHI '25)*. Association for Computing Machinery, New York, NY, USA, 1–26. doi:10.1145/3706598.3713998

[37] Bing He, Caleb Ziems, Sandeep Soni, Naren Ramakrishnan, Diyi Yang, and Srijan Kumar. 2022. Racism Is a Virus: Anti-Asian Hate and Counterspeech in Social Media during the COVID-19 Crisis. In *Proceedings of the 2021 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM '21)*. Association for Computing Machinery, New York, NY, USA, 90–94. doi:10.1145/3487351.3488324

[38] Lingzi Hong, Pengcheng Luo, Eduardo Blanco, and Xiaoying Song. 2024. Outcome-Constrained Large Language Models for Countering Hate Speech. doi:10.48550/arXiv.2403.17146 arXiv:2403.17146

[39] Angel Hsing-Chi Hwang and Andrea Stevenson Won. 2021. IdeaBot: Investigating Social Facilitation in Human-Machine Team Creativity. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems (CHI '21)*. Association for Computing Machinery, New York, NY, USA, 1–16. doi:10.1145/3411764.3445270

[40] Shagun Jhaver, Amy Bruckman, and Eric Gilbert. 2019. Does Transparency in Moderation Really Matter? User Behavior After Content Removal Explanations on Reddit. *Proc. ACM Hum.-Comput. Interact.* 3, CSCW (Nov. 2019), 150:1–150:27. doi:10.1145/3359252

[41] Shagun Jhaver, Himanshu Rathi, and Koustuv Saha. 2024. Bystanders of Online Moderation: Examining the Effects of Witnessing Post-Removal Explanations. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems (CHI '24)*. Association for Computing Machinery, New York, NY, USA, 1–9. doi:10.1145/3613904.3642204

[42] Yue Jia and Sandy Schumann. 2025. Tackling Hate Speech Online: The Effect of Counter-Speech on Subsequent Bystander Behavioral Intentions. *Cyberpsychology: Journal of Psychosocial Research on Cyberspace* 19, 1 (2025).

[43] Ji-Youn Jung, Sihang Qiu, Alessandro Bozzon, and Ujwal Gadiraju. 2022. Great Chain of Agents: The Role of Metaphorical Representation of Agents in Conversational Crowdsourcing. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems (CHI '22)*. Association for Computing Machinery, New York, NY, USA, 1–22. doi:10.1145/3491102.3517653

[44] David Jurgens, Eshwar Chandrasekharan, and Libby Hemphill. 2019. A Just and Comprehensive Strategy for Using NLP to Address Online Abuse. doi:10.48550/arXiv.1906.01738 arXiv:1906.01738

[45] Sameena Khokhar, Habibullah Pathan, Arsalan Raheem, and Abdul Malik Abbasi. 2020. Theory Development in Thematic Analysis: Procedure and Practice. 3, 3 (2020), 423–433. doi:10.47067/ramss.v3i3.79

[46] Adam D. I. Kramer, Jamie E. Guillory, and Jeffrey T. Hancock. 2014. Experimental Evidence of Massive-Scale Emotional Contagion through Social Networks. *Proceedings of the National Academy of Sciences* 111, 24 (June 2014), 8788–8790. doi:10.1073/pnas.1320040111

[47] Rae Langton. 2018. Blocking as Counter-Speech. *New work on speech acts* 144 (2018), 156.

[48] Maxime Lepoutre. 2021. *Democratic Speech in Divided Times.* Oxford University Press.

[49] Yaqiong Li, Peng Zhang, Hansu Gu, Tun Lu, Siyuan Qiao, Yubo Shu, Yiyang Shao, and Ning Gu. 2025. DeMod: A Holistic Tool with Explainable Detection and Personalized Modification for Toxicity Censorship. *Proc. ACM Hum.-Comput. Interact.* 9, 2 (May 2025), CSCW061:1–CSCW061:24. doi:10.1145/3710959

[50] Claire Liang, Julia Proft, Erik Andersen, and Ross A. Knepper. 2019. Implicit Communication of Actionable Information in Human-AI Teams. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems (CHI '19).* Association for Computing Machinery, New York, NY, USA, 1–13. doi:10.1145/3290605.3300325

[51] Bruce G. Link and Jo C. Phelan. 2001. Conceptualizing Stigma. *Annual Review of Sociology* 27, 1 (Aug. 2001), 363–385. doi:10.1146/annurev.soc.27.1.363

[52] Binny Mathew, Navish Kumar, Ravina, Pawan Goyal, and Animesh Mukherjee. 2018. Analyzing the Hate and Counter Speech Accounts on Twitter. doi:10.48550/arXiv.1812.02712 arXiv:1812.02712

[53] 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. *Proceedings of the International AAAI Conference on Web and Social Media* 13 (July 2019), 369–380. doi:10.1609/icwsm.v13i01.3237

[54] Rocío Galarza Molina and Freddie J. Jennings. 2018. The Role of Civility and Metacommunication in Facebook Discussions. *Communication Studies* 69, 1 (Jan. 2018), 42–66. doi:10.1080/10510974.2017.1397038

[55] 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 (CHI '24).* Association for Computing Machinery, New York, NY, USA, 1–22. doi:10.1145/3613904.3642025

[56] Kevin Munger. 2017. Tweetment Effects on the Tweeted: Experimentally Reducing Racist Harassment. *Political Behavior* 39, 3 (Sept. 2017), 629–649. doi:10.1007/s11109-016-9373-5

[57] Elisabeth Noelle-Neumann. 1974. The Spiral of Silence a Theory of Public Opinion. *Journal of communication* 24, 2 (1974), 43–51.

[58] Anna-Marie Ortloff, Florin Martius, Mischa Meier, Theo Raimbault, Lisa Geierhaas, and Matthew Smith. 2025. Small, Medium, Large? A Meta-Study of Effect Sizes at CHI to Aid Interpretation of Effect Sizes and Power Calculation. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems.* 1–28.

[59] Richard E. Petty and John T. Cacioppo. 1986. *The Elaboration Likelihood Model of Persuasion.* Springer New York, New York, NY, 1–24.

[60] Kaike Ping, James Hawdon, and Eugenia H Rho. 2025. Perceiving and Countering Hate: The Role of Identity in Online Responses. *Proc. ACM Hum.-Comput. Interact.* 9, 2 (May 2025), CSCW147:1–CSCW147:28. doi:10.1145/3711045

[61] Kaike Ping, Anisha Kumar, Xiaohan Ding, and Eugenia Rho. 2024. Behind the Counter: Exploring the Motivations and Barriers of Online Counterspeech Writing. doi:10.48550/arXiv.2403.17116 arXiv:2403.17116

[62] 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)*, Kentaro Inui, Jing Jiang, Vincent Ng, and Xiaojun Wan (Eds.). Association for Computational Linguistics, Hong Kong, China, 4755–4764. doi:10.18653/v1/D19-1482

[63] 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)*, Kentaro Inui, Jing Jiang, Vincent Ng, and Xiaojun Wan (Eds.). Association for Computational Linguistics, Hong Kong, China, 4755–4764. doi:10.18653/v1/D19-1482

[64] Derek Ruths Ruths, Haji Mohammed Saleem Saleem, Kelly P. Dillon Dillon, Lucas Wright Wright, and Susan Benesch Benesch. 2016. *Counterspeech on Twitter: A Field Study.* Technical Report. Dangerous Speech Project, Washington, DC USA.

[65] Koustuv Saha, Pranshu Gupta, Gloria Mark, Emre Kiciman, and Munmun De Choudhury. 2024. Observer Effect in Social Media Use. In *Proceedings of the CHI Conference on Human Factors in Computing Systems.* ACM, Honolulu HI USA, 1–20. doi:10.1145/3613904.3642078

[66] Punyajoy Saha, Abhilash Datta, Abhik Jana, and Animesh Mukherjee. 2024. CrowdCounter: A Benchmark Type-Specific Multi-Target Counterspeech Dataset. doi:10.48550/arXiv.2410.01400 arXiv:2410.01400

[67] Punyajoy Saha, Kanishk Singh, Adarsh Kumar, Binny Mathew, and Animesh Mukherjee. 2022. CounterGeDi: A Controllable Approach to Generate Polite, Detoxified and Emotional Counterspeech. In *Thirty-First International Joint Conference on Artificial Intelligence*, Vol. 6. 5157–5163. doi:10.24963/ijcai.2022/716

[68] Sougata Saha and Rohini Srihari. 2024. Consolidating Strategies for Countering Hate Speech Using Persuasive Dialogues. doi:10.48550/arXiv.2401.07810 arXiv:2401.07810

[69] Julia Sasse and Jens Grossklags. 2023. Breaking the Silence: Investigating Which Types of Moderation Reduce Negative Effects of Sexist Social Media Content. *Proc. ACM Hum.-Comput. Interact.* 7, CSCW2 (Oct. 2023), 327:1–327:26. doi:10.1145/3610176

[70] Martin Saveski, Brandon Roy, and Deb Roy. 2021. The Structure of Toxic Conversations on Twitter. In *Proceedings of the Web Conference 2021 (WWW '21).* Association for Computing Machinery, New York, NY, USA, 1086–1097. doi:10.1145/3442381.3449861

[71] Carla Schieb and Mike Preuss. 2016. Governing hate speech by means of counterspeech on Facebook. (2016), 1–23.

[72] Charlotte Schluger, Jonathan P. Chang, Cristian Danescu-Niculescu-Mizil, and Karen Levy. 2022. Proactive Moderation of Online Discussions: Existing Practices and the Potential for Algorithmic Support. *Proc. ACM Hum.-Comput. Interact.* 6, CSCW2 (Nov. 2022), 370:1–370:27. doi:10.1145/3555095

[73] Joseph Seering, Robert Kraut, and Laura Dabbish. 2017. Shaping Pro and Anti-Social Behavior on Twitch Through Moderation and Example-Setting. In *Proceedings of the 2017 ACM Conference on Computer Supported Cooperative Work and Social Computing (CSCW '17).* Association for Computing Machinery, New York, NY, USA, 111–125. doi:10.1145/2998181.2998277

[74] Cláudia Silva. 2023. Fighting Against Hate Speech: A Case for Harnessing Interactive Digital Counter-Narratives. In *Interactive Storytelling*, Lissa Holloway-Attaway and John T. Murray (Eds.). Springer Nature Switzerland, Cham, 159–174. doi:10.1007/978-3-031-47655-6_10

[75] Nicolas Sommet, David L. Weissman, Nicolas Cheutin, and Andrew J. Elliot. 2023. How Many Participants Do I Need to Test an Interaction? Conducting an Appropriate Power Analysis and Achieving Sufficient Power to Detect an Interaction. *Advances in Methods and Practices in Psychological Science* 6, 3 (July 2023), 25152459231178728. doi:10.1177/25152459231178728

[76] Carmela Sportelli, Paolo Giovanni Cicirelli, Marinella Paciello, Giuseppe Corbelli, and Francesca D'Errico. 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 (2025), e70028. doi:10.1002/casp.70028

[77] Minhyang (Mia) Suh, Emily Youngblom, Michael Terry, and Carrie J Cai. 2021. AI as Social Glue: Uncovering the Roles of Deep Generative AI during Social Music Composition. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems (CHI '21).* Association for Computing Machinery, New York, NY, USA, 1–11. doi:10.1145/3411764.3445219

[78] Samuel Hardman Taylor, Dominic DiFranzo, Yoon Hyung Choi, Shruti Sannon, and Natalya N. Bazarova. 2019. Accountability and Empathy by Design: Encouraging Bystander Intervention to Cyberbullying on Social Media. *Proc. ACM Hum.-Comput. Interact.* 3, CSCW (Nov. 2019), 118:1–118:26. doi:10.1145/3359220

[79] David R. Thomas. 2003. A General Inductive Approach for Qualitative Data Analysis. (2003).

[80] Stefanie Ullmann and Marcus Tomalin (Eds.). 2024. *Counterspeech: Multidisciplinary Perspectives on Countering Dangerous Speech.* Taylor & Francis.

[81] Bertie Vidgen, Austin Botelho, David Broniatowski, Ella Guest, Matthew Hall, Helen Margetts, Rebekah Tromble, Zeerak Waseem, and Scott Hale. 2020. Detecting East Asian Prejudice on Social Media. doi:10.48550/arXiv.2005.03909 arXiv:2005.03909

[82] Sebastian Wachs, Norman Krause, Michelle F. Wright, and Manuel Gámez-Guadix. 2023. Effects of the Prevention Program "HateLess. Together against Hatred" on Adolescents' Empathy, Self-efficacy, and Countering Hate Speech. *Journal of Youth and Adolescence* 52, 6 (June 2023), 1115–1128. doi:10.1007/s10964-023-01753-2

[83] Mengyao Wang, Jiayun Wu, Shuai Ma, Nuo Li, Peng Zhang, Ning Gu, and Tun Lu. 2025. Adaptive Human-Agent Teaming: A Review of Empirical Studies from the Process Dynamics Perspective. doi:10.48550/arXiv.2504.10918 arXiv:2504.10918 [cs]

[84] Brian Wilk, Homaira Huda Shomee, Suman Kalyan Maity, and Sourav Medya. 2025. Fact-Based Counter Narrative Generation to Combat Hate Speech. In *Proceedings of the ACM on Web Conference 2025 (WWW '25).* Association for Computing Machinery, New York, NY, USA, 3354–3365. doi:10.1145/3696410.3714718

[85] Xinchen Yu, Eduardo Blanco, and Lingzi Hong. 2022. Hate Speech and Counter Speech Detection: Conversational Context Does Matter. doi:10.48550/arXiv.2206.06423 arXiv:2206.06423

[86] Justine Zhang, Jonathan P. Chang, Cristian Danescu-Niculescu-Mizil, Lucas Dixon, Yiqing Hua, Nithum Thain, and Dario Taraborelli. 2018. Conversations Gone Awry: Detecting Early Signs of Conversational Failure. doi:10.48550/arXiv.1805.05345 arXiv:1805.05345

[87] Qiaoning Zhang, Matthew L Lee, and Scott Carter. 2022. You Complete Me: Human-AI Teams and Complementary Expertise. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems (CHI '22)*. Association for Computing Machinery, New York, NY, USA, 1–28. doi:10.1145/3491102.3517791

[88] Xiaoquan Zhao, Andrew Strasser, Joseph N. Cappella, Caryn Lerman, and Martin Fishbein. 2011. A Measure of Perceived Argument Strength: Reliability and Validity. *Communication Methods and Measures* 5, 1 (March 2011), 48–75. doi:10.1080/19312458.2010.547822

[89] Yi Zheng, Björn Ross, and Walid Magdy. 2023. What Makes Good Counterspeech? A Comparison of Generation Approaches and Evaluation Metrics. In *Proceedings of the 1st Workshop on CounterSpeech for Online Abuse (CS4OA)*, Yi-Ling Chung, Helena Bonaldi, Gavin Abercrombie, and Marco Guerini (Eds.). Association for Computational Linguistics, Prague, Czechia, 62–71.

[90] Jingyan Zhou, Jiawen Deng, Fei Mi, Yitong Li, Yasheng Wang, Minlie Huang, Xin Jiang, Qun Liu, and Helen Meng. 2022. Towards Identifying Social Bias in Dialog Systems: Framework, Dataset, and Benchmark. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, Yoav Goldberg, Zornitsa Kozareva, and Yue Zhang (Eds.). Association for Computational Linguistics, Abu Dhabi, United Arab Emirates, 3576–3591. doi:10.18653/v1/2022.findings-emnlp.262

[91] Wanzheng Zhu and Suma Bhat. 2021. Generate, Prune, Select: A Pipeline for Counterspeech Generation against Online Hate Speech. doi:10.48550/arXiv.2106.01625 arXiv:2106.01625

A Appendix A: Questionnaire

The final questionnaire assessed bystanders' responses on three dimensions: (1) perceived quality, (2) subjective acceptance, and (3) behavioural tendencies (pre–post). Items were translated from validated English scales and refined through pilot testing; Chinese items were back-translated to ensure conceptual equivalence. Table 6 lists the full bilingual items used in the study.

Table 6: Questionnaire Items (Chinese/English)

Dimension	Item (Chinese)	Item (English)
Perceived Quality	文明罗伯特的反驳理由具有说服力。	The reason given in this Civilbot counterspeech is convincing.
	文明罗伯特的反驳理由是强有力的。	The Civilbot counterspeech gives a strong reason.
Subjective Acceptance	文明罗伯特的反驳理由是可信的。	The reason given in this Civilbot counterspeech is believable.
	文明罗伯特的反驳内容提出了我认为重要的理由。	The counterspeech by Civilbot provides a reason I consider important.
	总体上，我同意文明罗伯特的反驳内容。	Overall, I agree with the Civilbot counterspeech.
Behavioural Tendencies (pre & post)	我有信心应对这种仇恨言论。	I am confident that I can respond to such hate speech.
	我想参与相关讨论。	I want to participate in the related discussion.

Appendix B. Prompt Design

B.1 Hate-Speech Screening Prompt

Table 7: Prompt used for the initial automatic screening of hate-speech candidates.

Instruction
Definition of hate speech: • It is a weaponized statement • Targeted at a specific social group • Likely to cause emotional or psychological harm • Includes hostility, dehumanization, or group contempt Return in JSON with keys: • `label`: 1 (hate) or 0 (not hate) • `justification`: one-sentence explanation • `target_group`: targeted group (e.g., women, LGBTQ, immigrants), or "N/A" • `issue`: main topic (e.g., "STEM vs humanities") Now evaluate: Context: "{q}" Comment: "{a}"

B.2 Counterspeech Generation Prompt

To generate counterspeech with balanced response lengths across rhetorical strategies, we adopted a *soft-approximate approach*. Instead of imposing a strict character or token limit, we instructed GPT-5 to produce *one short paragraph per speech act*, allowing mild natural variance in length while keeping outputs concise. Under the model's default sampling configuration, we applied conditional re-sampling only when outputs were clearly too long or too short, which maintained a moderate and balanced token-length distribution. Table 8 shows the exact prompt used to elicit the counterspeech responses.

Table 8: Prompt for generating counterspeech sentences in Chinese with soft-approximate length control.

Instruction

You are a counterspeech generation expert for online hate speech in Chinese. Task: Generate exactly **ONE short paragraph** (one speech act) in Chinese matching 100% the requested rhetorical dimensions. If any rule is broken, regenerate. Keep the response concise but allow natural length variation to support soft-approximate balance across strategies.

RULES

(1) Sentence Type
- Q: must be a question only, ending with "？" or Chinese question particles; no statements.
- Non-Q: must be a declarative sentence only; no question marks or question words.

(2) Tone
- Positive: friendly, cooperative, supportive.
- Negative: sarcastic, mocking, critical, emotionally intense.

(3) Strategy Intent
- Cognitive:
  - ▹ Rebut falsehoods (e.g., highlight hypocrisy, logical flaws, unreliable sources),
  - ▹ Highlight truth (e.g., provide facts, suggest proper action, warn of consequences).
- Affective:
  - ▹ Denounce perpetrators (e.g., explicitly identify hate, evoke shame, raise alarm),
  - ▹ Support targets (e.g., express empathy, solidarity, or emotional validation).

Input Variables:
Type: {stype} / {tone} / {intent}
Hatespeech: {hate_speech}
Counterspeech: