

# Generative AI may backfire for counterspeech

Dominik Bär

LMU Munich

Munich Center for Machine Learning

Munich, Germany

baer@lmu.de

Abdurahman Maarouf

LMU Munich

Munich Center for Machine Learning

Munich, Germany

a.maarouf@lmu.de

Stefan Feuerriegel

LMU Munich

Munich Center for Machine Learning

Munich, Germany

feuerriegel@lmu.de

## Abstract

Online hate speech poses a serious threat to individual well-being and societal cohesion. A promising solution to curb online hate speech is counterspeech. Counterspeech is aimed at encouraging users to reconsider hateful posts by direct replies. However, current methods lack scalability due to the need for human intervention or fail to adapt to the specific context of the post. A potential remedy is the use of generative AI, specifically large language models (LLMs), to write tailored counterspeech messages. In this paper, we analyze whether contextualized counterspeech generated by state-of-the-art LLMs is effective in curbing online hate speech. To do so, we conducted a large-scale, pre-registered field experiment ( $N = 2,664$ ) on the social media platform Twitter/X. Our experiment followed a 2x2 between-subjects design and, additionally, a control condition with no counterspeech. On the one hand, users posting hateful content on Twitter/X were randomly assigned to receive either (a) contextualized counterspeech or (b) non-contextualized counterspeech. Here, the former is generated through LLMs, while the latter relies on predefined, generic messages. On the other hand, we tested two counterspeech strategies: (a) promoting empathy and (b) warning about the consequences of online misbehavior. We then measured whether users deleted their initial hateful posts and whether their behavior changed after the counterspeech intervention (e.g., whether users adopted a less toxic language). We find that non-contextualized counterspeech employing a warning-of-consequence strategy significantly reduces online hate speech. However, contextualized counterspeech generated by LLMs proves ineffective and may even backfire.

## Keywords

hate speech, content moderation, counterspeech, social media, field experiment

**⚠ Warning:** Content in this paper may be upsetting or offensive. Reader discretion is advised

## 1 Introduction

Online hate speech poses a serious threat to individual well-being and societal cohesion. Individuals who experience online hate speech frequently suffer from psychological consequences that negatively affect their mental and physical well-being [18, 30, 48, 60]. Additionally, online hate speech is known to foster hostility between societal groups [50, 51] and may even motivate real-world violence such as witnessed in the 2017 Rohingya genocide in Myanmar [5] and the 2019 Christchurch mosque shooting [56]. Reducing online hate speech is thus a pressing issue for society. Here, we evaluate whether generative AI, specifically large language models

(LLMs), can help in writing counterspeech and thereby reduce hate speech on social media.

Counterspeech refers to direct responses (typically posted publicly) intended to encourage users to reconsider their hateful posts [34]. A benefit of counterspeech is that it does not infringe on users' freedom of speech since no content is removed. Generally, there are different strategies to counter online hate speech. For example, one counterspeech strategy is to promote **empathy** toward the attacked group or individual (e.g., "Imagine how it feels for group X to see people be attacked like this ...") [25, 55]. Another strategy is commonly referred to as **warning-of-consequences** and reminds offenders of social norms and warns of the consequences of online misbehavior (e.g., "This is hate speech! Such posts can damage your personal and professional reputation") [3, 4, 51, 55]. Previous research has demonstrated the effectiveness of counterspeech for reducing online hate speech across multiple field experiments [25, 39, 40, 51]. Later, we build upon the two strategies, which we then automated for contextualized counterspeech through the use of LLMs.

Counterspeech is traditionally implemented via two approaches: (1) manual counterspeech through human counterspeakers [54] or (2) scalable approaches with predefined, yet generic counterspeech messages under an "one-fits-all" paradigm [25, 39, 40, 51]. Manual counterspeech done by human counterspeakers is highly flexible and can be specifically tailored to the context of a hateful post [21]. However, manual counterspeech requires significant human effort and is thus **not** scalable to large social media platforms. Counterspeakers are further exposed to large amounts of online hate speech, which may negatively affect their well-being [54], rendering this approach impractical. In contrast, the "one-fits-all" approach from previous research [25, 39, 40, 51] can be automated and is thus scalable to a wider population of social media users. However, this approach ignores the context of a hateful post, potentially limiting the effectiveness of such a non-contextualized approach [38, 55]. In sum, existing studies have only studied scalable approaches based on a "one-fits-all" paradigm where hateful users received predefined, yet generic counterspeech messages. However, a counterspeech approach that is scalable and contextualized is missing.

Motivated by the above, we study the effectiveness of generative AI in the form of LLMs as a scalable approach for writing contextualized counterspeech. Modern LLMs can generate human-like text tailored to specific contexts [19, 20, 23, 27, 53, 68], which allows to generate counterspeech at scale and further enables to provide counterspeech that is contextualized to a specific topic.

In this paper, we analyze whether contextualized counterspeech generated by LLMs is effective in curbing online hate speech.<sup>1</sup> To do so, we conducted a large-scale, pre-registered field experiment

<sup>1</sup>Code and data are available at [https://osf.io/2nhsm/?view\\_only=247d97d62b994a419e07ff5c5a156f10](https://osf.io/2nhsm/?view_only=247d97d62b994a419e07ff5c5a156f10).

( $N = 2,664$ ) on the social media platform Twitter/X (see Fig. 1 for an overview). A particular strength of our study is its external validity. In contrast to survey or lab experiments that rely on simulated online environments, we provide real-world evidence from actual social media users who posted hate speech. To the best of our knowledge, we are the first to test the effectiveness of contextualized counterspeech generated using LLMs for reducing online hate speech in the field.

Our experiment followed a 2x2 between-subjects design, with an additional control condition without counterspeech. Users posting hateful content on Twitter/X were randomly assigned to receive either (a) contextualized counterspeech or (b) non-contextualized counterspeech. Here, the former is generated through an LLM, while the latter relies on predefined, generic messages. In doing so, we test whether counterspeech is more effective when carefully tailored to the context of the original hate speech post. Additionally, we employed two counterspeech strategies: (a) promoting empathy and (b) warning about the consequences of online misbehavior. Here, we test which strategies are effective and whether the effectiveness may be positively (or negatively) influenced when contextualized counterspeech messages are crafted through an LLM. Eventually, we measured whether users reconsidered their actions (i.e., whether they deleted their initial hateful post) and whether their behavior changed as a response to the intervention (i.e., whether they posted fewer hateful posts and adopted a less toxic language). Thereby, we contribute new insights into the role of LLMs in promoting online civility. Importantly, we later find that the use of LLMs may even backfire and thus call for caution when LLMs are used to improve online safety.

## 2 Related work

Online hate speech is a significant threat to individual well-being and social cohesion [10, 18, 30, 48, 50, 51, 60]. The United Nations defines hate speech as “any kind of communication in speech, writing or behavior, that attacks or uses pejorative or discriminatory language with reference to a person or a group on the basis of who they are, in other words, based on their religion, ethnicity, nationality, race, color, descent, gender or other identity factor.” [61] Previous research has, for example, studied the virality of hate speech [33], the characteristics of users sharing hate speech [45] but also how to detect [2, 41, 45] and curb online hate speech [25, 39, 40, 51].

Curbing online hate speech is challenging [21, 28]. Automated measures, such as content removal and account suspensions, are scalable and can effectively reduce online hate speech [21]. However, automated measures that are not properly calibrated may falsely remove content, which may be perceived as an infringement on individuals’ freedom of speech [22, 24, 38] and thus even spur more hostility [26, 34]. In contrast, manual moderation, such as the removal of problematic content or accounts by human moderators, can be more precise [21]. However, the high prevalence of online hate speech makes such efforts impractical [28]. Research has also shown that manual moderation can negatively impact the well-being of moderators [54], raising ethical concerns about its use. In this paper, we focus on counterspeech, which is seen as a promising approach to address the rise in online hate speech [12, 25, 62].

Counterspeech refers to direct responses intended to encourage users to reconsider their hateful posts [34]. Since no content is removed, a key advantage of counterspeech is that it does not infringe on users’ freedom of speech [34]. Previous research has suggested various strategies for counterspeech, such as empathy or reminding of social norms through warning-of-consequences [25, 34, 38–40, 55]. The effectiveness of counterspeech has been demonstrated in multiple field experiments [25, 39, 40, 51]. For example, counterspeech reminding of social norms reduced religious hate speech [51]. Furthermore, empathy-based counterspeech significantly reduced racist hate on Twitter/X [25, 39]. However, these studies typically follow a “one-fits-all” paradigm where predefined, generic counterspeech messages are sent to all offenders (e.g., “*This post is disrespectful. Please stop posting such hateful content!*”). In other words, this “one-fits-all” approach ignores the context of the underlying hateful post, potentially limiting its persuasiveness.

The emergence of LLMs has greatly improved the quality of automated text generation. LLMs take so-called prompts as inputs and then generate human-like text [19, 20, 23, 27, 53, 68]. LLMs are nowadays applied across various fields. For example, LLMs have been used to write political messages [23], aid mental health support [49], and provide recommendations in the emergency department [65]. Here, we explore the use of LLMs to generate contextualized counterspeech to curb online hate speech. A key strength of LLMs in this setting is that LLMs are scalable to the large volume of hate speech on social media platforms.

Previous research has demonstrated the potential of LLMs to produce counterspeech [7, 13, 14, 44, 58], yet with important limitations. On the one hand, these studies [7, 13, 14, 58] do not evaluate the effectiveness of LLM-generated counterspeech in the field but merely use surveys for evaluation. This is problematic since, even if people report that they are willing to behave civilly, it does not mean they act accordingly. This observation is known as the “intention-behavior gap” and poses a severe limitation when measuring intentions instead of actual behavior [29]. This can lead to inflated reports of civil behavior in surveys that may not translate to real-world social media environments. On the other hand, there is research [44] that analyzes secondary outcomes (e.g., views, likes) and thus fails to measure the effectiveness of counterspeech. Consequently, it remains unclear whether LLM-generated counterspeech can reduce online hate speech on real-world social media platforms.

**Research gap:** To the best of our knowledge, there is no evidence on whether LLM-generated counterspeech can effectively change real-world social media users’ behavior to reduce online hate speech. To close this gap, we conducted a field experiment on Twitter/X to test whether contextualized counterspeech generated by state-of-the-art LLMs is effective in curbing online hate speech.

## 3 Research question

There are good reasons to believe that contextualized counterspeech generated by an LLM is effective in curbing online hate speech. Context generally plays an important role in countering hate speech [38, 55]. For example, educating an aggressor on “why” a post is offensive may be more effective and lead to more lasting behavior change than a generic message (e.g., a user may not be fully aware of why a post is perceived as offensive) [55]. Previous research has

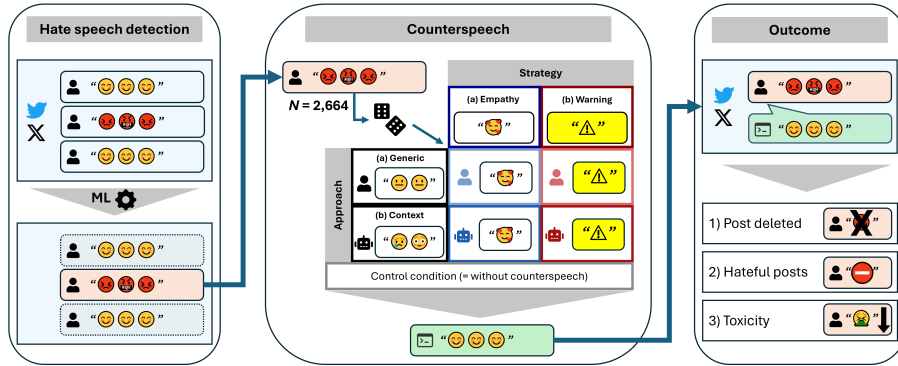


Figure 1: Overview of our field experiment.

demonstrated that LLM-generated messages are generally persuasive across various applications [35] but outside of counterspeech. For example, LLMs can generate messages that successfully mediate between opposing groups [59], decrease conspiracy beliefs [17] and promote civility in online conversations [6, 22]. Thus, it is likely that crafting custom messages through an LLM could also encourage online offenders to reconsider their hateful posts and, therefore, potentially reduce hate speech.

However, there are also reasons why counterspeech generated by LLMs may be ineffective. Generally, whether LLMs are persuasive varies across different use cases [57, 63]. For example, outside of counterspeech, some works ask users to have long discussions with chatbots and then assess whether their beliefs have changed as a result [6, 17, 22, 59]. In contrast, one-time interventions such as counterspeech are minimally invasive and may thus be ineffective. Additionally, studies suggest that contextualized messages are more likely to be identified as LLM-generated, which could lessen their impact compared to non-contextualized messages [23]. In fact, the identity of the source delivering counterspeech is crucial to be effective [39, 51]. Users who recognize that they are interacting with an LLM may resist changing their behavior or feel deceived. There is also evidence that counterspeech may be perceived as intrusive and therefore can backfire and even escalate hostility [34].

Motivated by the above, we evaluate the effectiveness of contextualized counterspeech generated by an LLM in a large-scale, pre-registered field experiment ( $N = 2,664$ ) on Twitter/X. In particular, we seek to answer the following research question (RQ):

**Research question:** *Can contextualized counterspeech generated by an LLM effectively curb online hate speech?*

## 4 Methodology

To evaluate the effectiveness of contextualized counterspeech generated by an LLM to reduce hateful content on social media, we conducted a large-scale, pre-registered<sup>2</sup> field experiment on Twitter/X (www.X.com). In the following, we describe our (1) interventions, (2) experimental procedure, (3) study population, (4) statistical analysis, and (5) ethical considerations.

<sup>2</sup>Pre-registration available at [https://osf.io/38saz/?view\\_only=263687bff9824852b8ed204f257de8d5](https://osf.io/38saz/?view_only=263687bff9824852b8ed204f257de8d5)

### 4.1 Interventions

Our experiment followed a 2x2 between-subjects design where, in addition, we included a control condition with no counterspeech. Overall, we thus randomly assigned users to one of five experimental conditions. Our experimental conditions are shown in Table 1.

**Table 1: Experimental conditions for the field experiment. We randomly assigned users to one of five experimental conditions: Users receive either (a) contextualized LLM-generated counterspeech or (b) non-contextualized counterspeech. We further tested two counterspeech strategies: (a) promoting empathy and (b) warning about the consequences of online misbehavior. Additionally, we used a control condition without intervention (i.e., without counterspeech).**

	Empathy	Warning-of-Consequences
Non-contextualized	GENERIC-EMPATHY	GENERIC-WARNING
Contextualized	CONTEXT-EMPATHY	CONTEXT-WARNING
CONTROL CONDITION (=no counterspeech intervention)		

Our messages for non-contextualized counterspeech are inspired by [25]. The messages either promote empathy (GENERIC-EMPATHY) or warning-of-consequences (GENERIC-WARNING). To avoid a strongly repetitive reply pattern that may irritate users, we used five different counterspeech messages for each non-contextualized condition that are qualitatively the same.<sup>3</sup> The exact messages are reported in Table 5.

To generate contextualized counterspeech, we used Llama-3 70B Chat, a state-of-the-art open-source LLM developed by Meta [1]. Depending on the condition, we prompt the model to generate counterspeech aimed at promoting empathy (CONTEXT-EMPATHY) or warning-of-consequences (CONTEXT-WARNING).<sup>4</sup> Our prompts

<sup>3</sup>Throughout our paper, we label non-contextualized counterspeech with an emoji showing a “human” (👤) to indicate that these messages were crafted by human experts. However, all non-contextualized counterspeech messages are predefined and not customized to address specific posts.





<sup>4</sup>LLMs are known for their ability to generate empathetic content [31]. However, it is unclear whether they can produce convincing warnings. To address this, we conducted an online survey with 500 participants recruited from Prolific www.prolific.com to evaluate perceived differences between human- and LLM-generated warnings. Our

followed best practices in prompt engineering [64] and prior research [6, 17, 23]. The exact prompts are in Table 3. Examples of contextualized counterspeech are in Table 4.

## 4.2 Procedure

Our experimental procedure is as follows (see Fig. 1): We sampled hateful users on Twitter/X by searching for hateful posts using a comprehensive list of keywords (see Table 2 for a full list) via the Twitter/X API v2.<sup>5</sup> We then manually filtered posts by users matching our keywords for hateful content. Note that this step could be automated. We opted for a manual validation for ethical considerations and, in particular, to comply with requirements from our Institutional Review Board (IRB), which allows us to ensure participants’ safety (see our extensive discussion of ethical considerations in Sec. 4.5).

Subsequently, we retrieved the user profiles associated with each hateful post and filtered users according to pre-registered exclusion criteria (see Section 4.3 for details). The remaining users were then randomly assigned to one of the five experimental conditions.

Next, we assigned counterspeech to each user and replied to their hateful post. Of note, each user is only treated once. For users receiving non-contextualized counterspeech (i.e.,  GENERIC-EMPATHY or  GENERIC-WARNING), we randomly selected one of five pre-defined counterspeech messages based on the assigned strategy (see Table 5). For users receiving contextualized counterspeech (i.e.,  CONTEXT-EMPATHY or  CONTEXT-WARNING), we prompted Llama 3 to generate contextualized counterspeech using the corresponding prompt template (see Table 3).

We administered our intervention via multiple human-controlled accounts. The accounts were designed to appear politically neutral and natural to users on Twitter/X, which was inspired by the design in [25]. Each account was assigned a unisex English name, with no disclosure of gender, ethnicity, nationality, or beliefs. Furthermore, to appear as natural users, we regularly posted neutral posts via our accounts (e.g., “*Just witnessed the most breathtaking sunset!*”) and re-posted content from diverse accounts (e.g., NASA, WWF, ESPN). The accounts were created at least 3 months before the start of the experiment. Screenshots of example profiles are in Fig. 5.

Following our intervention, we monitored users’ behavior on Twitter/X for two weeks to assess its effectiveness. Specifically, we analyzed the following three outcome variables:

- (1) **Rate of deleted posts:** A dichotomous variable that indicates whether a user deleted their original hateful post (= 1 if the post was deleted, = 0 otherwise) following our intervention. We later operationalize this by computing the average rate of deleted posts per experimental condition.
- (2) **Number of hateful posts:** The number of hateful posts by a user after the intervention.
- (3) **Relative change in toxicity:** The relative change in toxicity of a user’s posts after the intervention.

Our outcome variables are designed to measure the effect of counterspeech on two distinct psychological processes: (a) whether users reconsidered their action and deleted the original tweet and

results indicate that LLM-generated and human-generated warnings are equally likely to be recognized as such.

<sup>5</sup><https://developer.twitter.com/en/products/twitter-api>

(b) whether users changed their behavior and posted less hate speech but also adapted their tone and engaged in more civil conversations. We chose this approach since it should reflect users’ recent activities on Twitter/X and their immediate reaction to our intervention. Details for each outcome are below:

- **Rate of deleted posts:** To measure whether a user deleted their original post, we queried the respective post via the Twitter/X API two weeks after our intervention. If the post was no longer available, we then used Twitter/X’s compliance API endpoint to confirm whether the post had been actually deleted by the user (i.e., we do not count cases where the user has changed their privacy settings or was suspended by the platform).

- **Number of hateful posts:** To measure the number of hateful posts by users after our intervention, we collected up to the most recent 100 posts from the two weeks following the intervention. We then classified each post as hateful or not using Twitter-roBERTa-base-hate, which is trained on  $\approx 58$  million Twitter posts and fine-tuned for hate speech detection [11]. Eventually, we counted the number of hateful posts by each user.

- **Relative change in toxicity:** To measure the relative change in the toxicity of a user’s posts following the intervention, we also collected up to 100 posts from the two weeks before the intervention. We then computed the average toxicity of posts shared before and after the intervention using Google’s Perspective API [32], which is frequently used by previous research to study the toxicity of online content [8, 9, 67] and yields a toxicity score  $\in [0, +1]$ .

## 4.3 Study population

We recruited  $N = 2,778$  users who posted hateful content on Twitter/X on weekdays between June 7 and July 26, 2024. Users are identified following the procedure described above. Specifically, we searched for hateful posts and retrieved the associated user profiles. Subsequently, each user was randomly assigned to one of the experimental conditions and received the corresponding treatment. Importantly, each user is treated only once.

Following our pre-registration, we excluded users based on the following criteria: (1) We excluded verified accounts (i.e., we excluded organizations with a golden checkmark). (2) We only considered users that posted original posts (i.e., we excluded retweets and replies). (3) We excluded users that had been inactive (i.e., they had shared fewer than 5 posts in the past 7 days) to ensure sufficient data to compare later if users had changed their behavior after our intervention. (4) We excluded users likely to be bot accounts.

As per our pre-registration, we have excluded 99 accounts that altered their privacy settings or were suspended by Twitter/X during the experiment. To check for differential attrition rates between treatment and control conditions, we used a  $\chi^2$ -test [42]. We found no significant differences in attrition rates between treatment and control groups ( $p > 0.1$ ). Following our debrief, we have excluded 15 users who opted out of the study (see section 4.5 for details). Overall, we have thus excluded 114 accounts and, eventually, have  $N = 2,664$  accounts in the subsequent analysis.

## 4.4 Statistical analysis

To statistically compare the effectiveness of the different interventions, we use a linear regression model. Our unit of analysis is users



who posted hate speech. Let  $y_i$  denote one of our three outcome variables, namely, (1) the *rate of deleted posts*, (2) the *number of hateful posts*, and (3) the *relative change in toxicity*, for user  $i$ . Let  $t_i$  denote the intervention received by user  $i$ , and let  $x_i$  refer to a vector of different characteristics belonging to that user (described later). We then estimate the following linear regression model

$$y_i = \alpha + \theta t_i + \beta^T x_i + \varepsilon_i, \quad (1)$$

where  $\alpha$  represents the model intercept,  $\theta$  measures the effect of the intervention  $t_i$ ,  $\beta$  captures the effect of all control variables in  $x_i$  on  $y_i$ , and  $\varepsilon_i$  is the error term. For estimation, we use ordinary least squares regression (OLS) with robust standard errors to account for heteroskedasticity in the error term due to variations in our intervention and control variables. We test whether the coefficients are significantly different from zero using two-sided  $t$ -tests.

Our analysis is split into two: (1) We evaluate the effectiveness of counterspeech compared to the CONTROL CONDITION (=no counterspeech intervention). (2) We study the effectiveness of contextualized counterspeech generated by an LLM compared to non-contextualized counterspeech. Both are as follows:

(1) *Intervention vs. control*: To evaluate the effectiveness of each intervention compared to the control, we separately estimate the linear regression model described above for each type of counterspeech. Specifically, we set  $t_i = 1$  if a user received a counterspeech reply and  $t_i = 0$  for users assigned to the control condition.

(2) *Contextualized vs. non-contextualized*: We compare the effectiveness of contextualized counterspeech generated by an LLM to non-contextualized counterspeech. To do so, we re-estimate the linear regression model from above but set  $t_i = 1$  if a user received contextualized counterspeech and  $t_i = 0$  if a user received non-contextualized counterspeech. We perform this comparison separately for each counterspeech strategy. Hence, we estimate one model to compare 🗨️ CONTEXT-EMPATHY vs. 👤 GENERIC-EMPATHY and a separate model to compare 🗨️ CONTEXT-WARNING vs. 👤 GENERIC-WARNING.

For each regression model described above, we estimate three versions, each with a different outcome variable. Thus, we estimate separate models for (1) the *rate of deleted posts*, (2) the *number of hateful posts*, and (3) the *relative change in toxicity*. The analysis was implemented in R 4.4.1. using the `stats` and `lmtest` packages.

Following our pre-registration, we included a set of pre-treatment covariates to account for variability in the outcome explained by pre-treatment covariates. Specifically, we included a user’s account age (in years), follower count, following count, tweet count, and whether the user subscribes to Twitter/X Premium (= 1 if premium, = 0 otherwise) as indicated by a blue checkmark on a user’s profile page. We further included the number of hateful posts shared by a user in the two weeks before the intervention. To classify whether a post is hateful, we again used the Twitter-roBERTa-base model [11]. Lastly, we included the average toxicity of a user’s posts shared within two weeks before the intervention, measured by Google’s Perspective API [32]. Of note, we only collected up to 100 posts before our interventions, which should reflect the recent activities of users on Twitter/X.

**Robustness checks:** To ensure the robustness of our results, we conducted a series of checks: (1) We re-estimated our analysis using a single model that included separate dummy variables

for each intervention, instead of estimating separate models for each combination of treatment and control. (2) We pooled all observations in the treatment conditions to assess the overall effect of counterspeech compared to the control group. (3) We pooled observations based on the counterspeech strategies (i.e., empathy vs. warning-of-consequences) and re-estimated our regression model to evaluate their overall effects compared to the control. (4) To compare the effectiveness of contextualized and non-contextualized counterspeech, we pooled all users assigned to contextualized counterspeech across both strategies and repeated the analysis. All robustness checks led to consistent findings.

## 4.5 Ethics

Ethics approval (EK-MIS-2024-254) for the field experiment was obtained from the Institutional Review Board at the Faculty of Mathematics, Informatics, and Statistics, LMU Munich. This ethics approval complies with regulations for studies involving human participants at the Faculty of Mathematics, Informatics, and Statistics, LMU Munich. The experimental task, data collection, and analysis closely follow related works involving counterspeech on social media [25, 39, 40]. Our study solely relies on publicly available data and follows common guidelines for ethical research with social media [46]. We only report aggregated and anonymized results to protect users’ privacy.

Ethical considerations were of utmost importance for our study. Following previous studies [25, 39, 40] and ethical guidelines on experimental research on social media [37], we designed our interventions to be minimally invasive and socially beneficial. Specifically, our interventions are designed to mitigate hate speech while preserving users’ right to free expression.

To minimize ethical risks and protect the well-being of all participants, we have further implemented a detailed experimental protocol that includes comprehensive safety measures. This includes explicit guidelines for continuous human monitoring and specific countermeasures. Our experimental protocol ensures (a) the appropriateness of our interventions and (b) immediate actions to guarantee the safety of all participants. For example, we manually ensured that each counterspeech conveyed an appropriate tone, avoided biases, and was culturally sensitive.

Of note, users provide informed consent to receive public replies when they agree to the terms of service at Twitter/X when signing up for the platform [66]. Users agree that they may receive replies from other users when engaging on Twitter/X (Section 3, Twitter/X Terms of Service [66]). As such, our intervention fully complies with users’ informed consent to the use of Twitter/X and aligns with the platform’s goal to combat hate speech without infringing users’ rights to freedom of expression [66].

Our study was carefully designed to ensure users’ privacy. Specifically, our study was compliant with the General Data Protection Regulation (GDPR) of the European Union. We have implemented measures to repeatedly enforce the privacy settings of all participants by frequently calling the Twitter/X Compliance API to check if users have changed their privacy settings and delete private data accordingly. We further abide by the privacy regulations of GDPR and ensure that users can fully opt out of data collection. To do so, we have debriefed users following our experiment. Our debrief

included detailed information on the goals, methods, and interventions of our study. We have further informed participants about their privacy rights concerning GDPR and provided contact addresses for questions. We sent our debrief collectively at the end of the experiment and granted users an extended period to request additional information on the study or have their data removed.

## 5 Results

### 5.1 Effectiveness of counterspeech

To evaluate the effectiveness of counterspeech in reducing online hate speech, we compare each intervention to the CONTROL CONDITION (=no counterspeech intervention) across our outcome variables. Fig. 2 shows the result, which we discuss in the following:

- (1) *Rate of deleted posts*: Fig. 2a shows the average rate of deleted posts across each condition in our experiment. In the control group, on average, 7.13 % of users deleted their original posts. Across all conditions, except 🗨️ CONTEXT-WARNING, fewer users deleted their hateful posts following counterspeech. On average 3.94 % (-3.19 p.p. compared to control), 3.74 % (-3.39 p.p.), and 5.21 % (-1.92 p.p.) of users deleted their hateful posts for 👤 GENERIC-EMPATHY, 🗨️ CONTEXT-EMPATHY, and 🗨️ CONTEXT-WARNING counterspeech, respectively. In contrast, counterspeech based on 🗨️ GENERIC-WARNING resulted in an average of 7.72 % (+0.59 %) of users deleting their hateful posts, indicating that users are encouraged to remove hateful content when receiving non-contextualized warning-of-consequences counterspeech.

- (2) *Number of hateful posts*: The average number of hateful posts shared by each user within two weeks after the intervention for each experimental condition is shown in Fig. 2b. In the control group, users shared an average of 9.07 hateful posts in the two weeks following the intervention. Users shared fewer hateful posts following counterspeech when receiving 🗨️ CONTEXT-EMPATHY or 🗨️ GENERIC-WARNING counterspeech. Specifically, users shared an average of 8.18 (-0.89 compared to control) and 8.04 (-1.03) hateful posts for 🗨️ CONTEXT-EMPATHY, and 🗨️ GENERIC-WARNING counterspeech, respectively. In contrast, 👤 GENERIC-EMPATHY and 🗨️ CONTEXT-WARNING counterspeech resulted in an average of 9.20 (+0.13) and 9.16 (+0.09) hateful posts, respectively. These results suggest that hostility increased among users who received either non-contextualized empathy-based counterspeech or contextualized LLM-generated warning-of-consequences counterspeech in the two weeks following the intervention.

- (3) *Relative change in toxicity*: The mean relative change in toxicity of users' posts within 2 weeks after our intervention for each experimental condition is shown in Fig. 2c. On average, toxicity increased by 3.44 % for users that did not receive any counterspeech (i.e., CONTROL CONDITION (=no counterspeech intervention)). Non-contextualized counterspeech led to a reduction in toxicity: Users that received 👤 GENERIC-EMPATHY (2.99 %) and 🗨️ GENERIC-WARNING (1.88 %) counterspeech are, on average, less toxic (-0.45 p.p., and -1.56 p.p. compared to the control, respectively). In contrast, LLM-generated counterspeech led to an increase in toxicity for 🗨️ CONTEXT-EMPATHY (9.74 %) and 🗨️ CONTEXT-WARNING (4.54 %) by, on average, +6.30 p.p. and +1.10 p.p. compared to the control. Overall, this suggests that LLM-generated counterspeech increases toxicity regardless of the counterspeech strategy.

### 5.2 Regression analysis

To statistically compare the effectiveness of the different interventions, we use a linear regression model. Our regression analysis is split in two: (1) First, we evaluate the effectiveness of the different counterspeech interventions compared to the CONTROL CONDITION (=no counterspeech intervention). (2) Second, we study the effectiveness of LLM-generated counterspeech compared to non-contextualized counterspeech. As before, we estimated the treatment effect of our counterspeech compared to the CONTROL CONDITION (=no counterspeech intervention) across our three outcomes (see Fig. 3 and Fig. 4):

- (1) *Rate of deleted posts*: Fig. 3a shows the treatment effects on the rate of users deleting their posts following our intervention vs. the control. In line with our descriptive analysis, empathy-based counterspeech negatively affects the likelihood of users deleting their hateful posts. All else equal, users who received 👤 GENERIC-EMPATHY and 🗨️ CONTEXT-EMPATHY counterspeech were, on average, 2.62 percentage points ( $p = 0.055$ ) and 2.89 percentage points ( $p = 0.0345$ ) less likely to delete their posts, respectively. We also observe a positive coefficient for 🗨️ GENERIC-WARNING, yet this effect is not statistically significant ( $\theta = -0.83\%$ ;  $p = 0.608$ ).

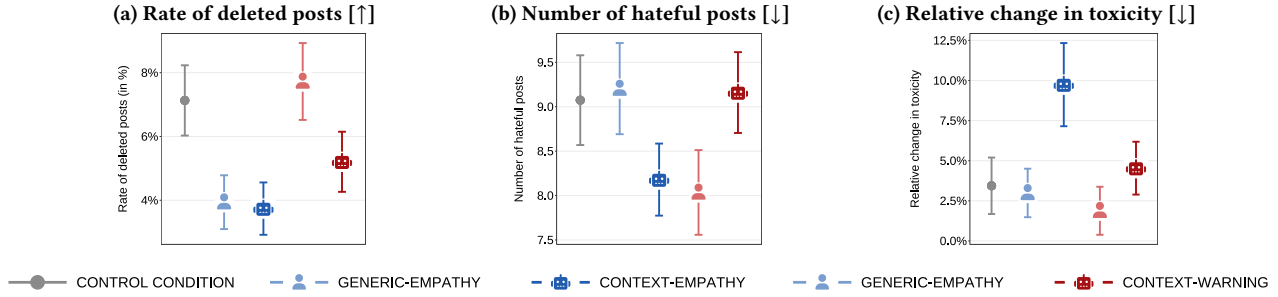
- (2) *Number of hateful posts*: The effect of counterspeech on the number of hateful posts shared by users after the intervention is shown in Fig. 3b. Compared to the control, we observe a negative coefficient for both non-contextualized and LLM-generated counterspeech across both strategies (i.e., both empathy and warning-of-consequences). This effect is statistically significant for 🗨️ GENERIC-WARNING, where users shared, all else equal, on average, 1.03 fewer hateful posts ( $p = 0.022$ ) after receiving non-contextualized warning-of-consequences counterspeech.

- (3) *Relative change in toxicity*: Fig. 3c presents the estimated effects of counterspeech on the relative change in the toxicity of a user's posts. We do not observe a statistically significant effect of counterspeech on the relative change in toxicity across all experimental conditions compared to the control. However, the negative coefficients for 👤 GENERIC-EMPATHY and 🗨️ GENERIC-WARNING, alongside the positive coefficients for 🗨️ CONTEXT-EMPATHY and 🗨️ CONTEXT-WARNING, suggest a potential adverse effect of LLM-generated counterspeech.

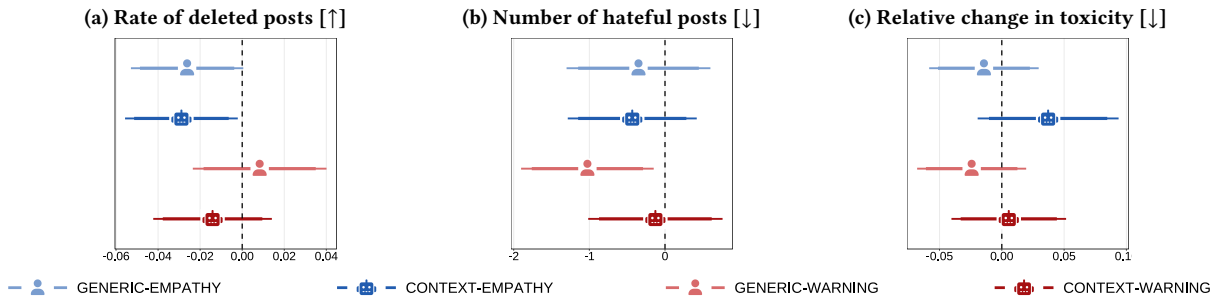
### 5.3 Deep-dive: contextualized vs. non-contextualized counterspeech

Our descriptive analysis revealed that contextualized counterspeech generated by LLMs led to worse outcomes compared to non-contextualized counterspeech. We thus use the linear regression model from above to statistically compare contextualized counterspeech vs. non-contextualized counterspeech across our outcome variables. The results are shown in Fig. 4

- (1) *Rate of deleted posts*: When comparing non-contextualized to contextualized counterspeech for *rate of deleted posts*, we do not find any significant effects (see Fig. 4a). Nevertheless, the negative coefficients for 🗨️ CONTEXT-EMPATHY and 🗨️ CONTEXT-WARNING indicate that contextualized counterspeech generated by LLMs may be less effective than non-contextualized counterspeech in reducing online hate speech.



**Figure 2:** Average (a) rate of deleted posts, (b) number of hateful posts after the intervention, and (c) relative change in toxicity and standard errors (bars) by experimental condition. [↑] ([↓]) indicates that a [positive] ([negative]) outcome is associated with an [increase] ([decrease]) in the outcome values.



**Figure 3:** Treatment effect of an intervention relative to the CONTROL CONDITION (=no counterspeech intervention) for (a) *Rate of deleted posts*, (b) *Number of hateful posts*, and (c) *Relative change in toxicity*. Shown are the estimated coefficients from our linear regression model (symbol) as well as 95% (thin), and 90% (thick) confidence intervals. [↑] ([↓]) indicates that a [positive] ([negative]) outcome is associated with an [increase] ([decrease]) in the outcome values.

- (2) *Number of hateful posts*: Here, we do not find a statistically significant difference for empathy, when comparing non-contextualized and contextualized counterspeech. However, we find a positive and statistically significant coefficient for CONTEXT-WARNING ( $p = 0.032$ ). Hence, all else equal, CONTEXT-WARNING increased the number of hateful posts shared within the two weeks following the intervention, on average, by 0.84 posts compared to GENERIC-WARNING. As such, contextualized warning-of-consequences increases online hostility compared to non-contextualized warning-of-consequences.

- (3) *Relative change in toxicity*: The treatment effect of contextualized vs. non-contextualized counterspeech is shown in Fig. 4c. We find a positive and statistically significant coefficient for CONTEXT-EMPATHY compared to GENERIC-EMPATHY ( $p = 0.048$ ). All else equal, CONTEXT-EMPATHY led to an increase in toxicity by 2.80 percentage points, on average, compared to GENERIC-EMPATHY. While we also observe a positive coefficient for CONTEXT-WARNING, this effect is not statistically significant at common significance thresholds.

#### 5.4 Additional analysis: Counterspeech for Twitter/X Premium users

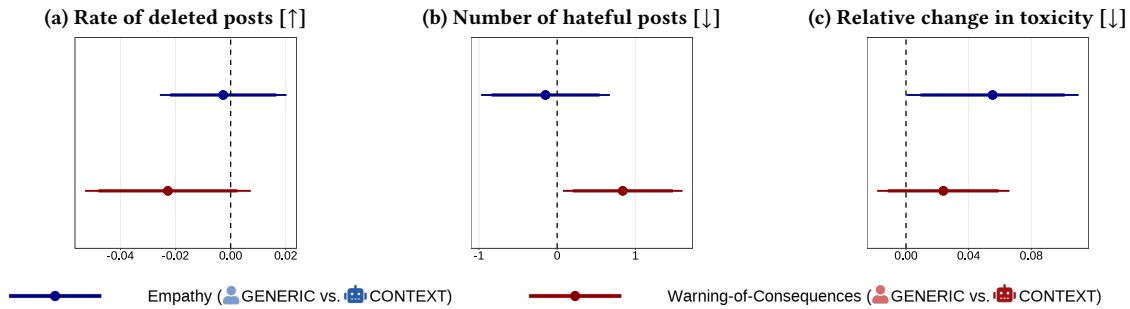
Hateful users who subscribe to Twitter/X Premium are less likely to have their content removed by the platform, and their posts are algorithmically boosted [15]. Hence, we evaluate whether our

intervention is effective for Twitter/X Premium users. To do so, we re-estimate the regression model from our main analysis, adding an interaction term between the treatment and Twitter/X Premium subscription status (= 1 if subscribed, = 0 otherwise).

We find no significant interaction between our intervention and Twitter/X Premium subscription status for the *rate of deleted posts*. However, Twitter/X Premium users who received CONTEXT-EMPATHY counterspeech shared significantly more hateful posts ( $p = 0.012$ ) than non-subscribers. Additionally, Premium users exhibited higher toxicity levels when receiving GENERIC-EMPATHY ( $p = 0.049$ ) and CONTEXT-EMPATHY ( $p = 0.024$ ) counterspeech. This suggests that empathetic counterspeech, particularly when LLM-generated, may backfire for Premium users. Importantly, all treatment effects remain consistent with our primary analysis across all models and dependent variables, except for CONTEXT-WARNING vs. GENERIC-WARNING and the *number of hateful posts*, which is no longer significant ( $p = 0.107$ ).

## 6 Discussion

**Relevance:** Online hate speech poses a serious threat to societal cohesion and individual well-being [18, 30, 48, 50, 51, 60] and can even incite real-world violence [62]. Hence, curbing online hate speech is a crucial challenge for society. In this paper, we evaluate the effectiveness of contextualized counterspeech generated by an LLM in reducing online hate speech through a large-scale, pre-registered field experiment.



**Figure 4: Treatment effect of contextualized vs. non-contextualized counterspeech for (a) Rate of deleted posts, (b) Number of hateful posts, and (c) Relative change in toxicity.** Shown are the estimated coefficients from our linear regression model (dot) measuring the relative effect of generic (👤 Non-contextualized) vs. contextualized (🗨️ Contextualized) counterspeech for the respective strategy as well as 95 % (thin), and 90 % (thick) confidence intervals. [↑] ([↓]) indicates that a [positive] ([negative]) outcome is associated with an [increase] ([decrease]) in the outcome values.

**Summary of findings:** Our field experiment offers only limited evidence that counterspeech can significantly reduce online hate speech. While we find that 🗨️ GENERIC-WARNING leads to a slight but statistically significant reduction in the sharing of hateful posts compared to the control, we observe only weak directional evidence or even adversarial effects for other counterspeech strategies and outcomes. In particular, for both 👤 GENERIC-EMPATHY and 🗨️ CONTEXT-EMPATHY, we even see a significantly lower rate of deleted posts (and for Twitter/X Premium users even more hateful posts and increased toxicity), indicating a negative outcome.

Our results contrast with previous research reporting that counterspeech is effective [25, 39, 40, 51]. Given that our study design and non-contextualized messages are inspired by prior work [25], this discrepancy may be attributed to changes in the ecosystem of Twitter/X, which is reported to have become more hostile and toxic [15, 16]. This shift could make it increasingly difficult to persuade users to behave civilly, as they may face fewer repercussions for their actions. Another possible explanation is a lack of statistical power to detect small positive effects. However, given that our sample size ( $N = 2,664$ ) is significantly larger than in previous studies [25, 39, 40, 51], this seems unlikely.

Our results even show that LLM-use may backfire: when comparing contextualized LLM-generated vs. non-contextualized counterspeech, we see that LLM-generated counterspeech is less effective in reducing online hate speech and may even increase hostility. For instance, 🗨️ CONTEXT-WARNING led to significantly more hateful posts vs. 👤 GENERIC-WARNING. Similarly, 🗨️ CONTEXT-EMPATHY resulted in greater toxicity than 👤 GENERIC-EMPATHY.

One possible explanation is that users often react negatively when they recognize LLM-generated content intended to convey empathy [43, 47]. Similarly, the identity of the messenger is crucial for counterspeech based on warning-of-consequences, which aims to reinforce social norms [51]. Given that people are more likely to recognize tailored LLM-generated texts [23], users may realize they are interacting with an LLM and thus might resist changing their behavior or feel deceived, which could lead to negative outcomes.

**Limitations and future work:** As with other research, ours is not free of limitations that offer opportunities for future work. For instance, our analysis is based on a large-scale, pre-registered field

experiment conducted on Twitter/X, a platform often criticized for hosting hate speech and inadequately removing harmful content [15, 16]. While Twitter/X presents a challenging case, the effects of contextualized counterspeech generated by LLMs may differ across platforms, highlighting the need for future research to explore the potential of counterspeech in other online environments. Furthermore, we use Llama-3, a state-of-the-art open-source LLM developed by Meta [36], to generate contextualized counterspeech. This allows for reproducibility and accessibility [52]. Future research may also explore the use of proprietary models (e.g., GPT-4). Nevertheless, we experimented with proprietary models such as GPT-4 by Open AI but did not find qualitative differences in the counterspeech generated by Llama 3.

**Implications:** Our findings contribute to the literature on content moderation, specifically, counterspeech to curb hate speech on social media. Unlike previous studies that employed predefined, generic counterspeech messages [25, 39, 40, 51], we consider the importance of context in countering hate [38, 55]. Our approach uses LLMs to generate counterspeech tailored to individual hateful posts, aiming to promote civil online behavior. In doing so, we contribute to the ongoing debate on *when* LLMs can enhance persuasion [57, 63]. While LLMs have shown promise in mediating opposing groups [59], countering conspiracy theories [17], and fostering civil online conversations [6, 22], it was unclear whether LLMs could encourage more civil behavior through counterspeech. Our findings indicate that LLM-generated counterspeech is ineffective in promoting civil behavior and may even backfire, highlighting the need for further research into the conditions under which LLM-generated messages influence behavior effectively.

For platforms and policymakers, our results offer new insights into the role of LLMs in promoting online civility and highlight the need for caution when deploying LLM-driven societal interventions at scale. While counterspeech is promising in addressing hate speech, our findings suggest that LLM-generated interventions may be ineffective or even backfire. This indicates that relying on LLMs alone to foster behavioral change in online environments may yield limited results without a deeper understanding of the conditions under which LLM interventions are most persuasive. LLM-generated messages may thus need to be part of a broader



strategy that includes repeated interventions or human moderation to have a meaningful impact. Additionally, given the potential for LLM-generated content to be perceived as inauthentic or deceiving [23], transparency and careful design are essential to prevent backlash. Therefore, the use of LLMs should be accompanied by thorough testing, ongoing evaluation, and flexibility to adjust strategies based on platform-specific dynamics and user behavior.

## Acknowledgments

Funding by the German Research Foundation (Grant: 543018872) and the Civic Health Project ([www.civichealthproject.org](http://www.civichealthproject.org)) is acknowledged.

## References

- [1] AI @ Meta, Llama Team. 2024. The Llama 3 herd of models. *arXiv 2407.21783* (2024).
- [2] Nuha Albadi, Maram Kurdi, and Shivakant Mishra. 2019. Hateful people or hateful bots? Detection and characterization of bots spreading religious hatred in Arabic social media. In *CSCW*.
- [3] Amalia Álvarez-Benjumea and Fabian Winter. 2018. Normative change and culture of hate: An experiment in online environments. *European Sociological Review* 34, 3 (2018), 223–237.
- [4] Amalia Álvarez-Benjumea and Fabian Winter. 2020. The breakdown of antiracist norms: A natural experiment on hate speech after terrorist attacks. *PNAS* 117, 37 (2020), 22800–22804.
- [5] Amnesty International. 2022. Myanmar: The social atrocity: Meta and the right to remedy for the Rohingya. <https://www.amnesty.org/en/documents/ASA16/5933/2022/en/>
- [6] Lisa P. Argyle, Christopher A. Bail, Ethan C. Busby, Joshua R. Gubler, Thomas Howe, Christopher Rytting, Taylor Sorensen, and David Wingate. 2023. Leveraging AI for democratic discourse: Chat interventions can improve online political conversations at scale. *PNAS* 120, 41 (2023), e2311627120.
- [7] Mana Ashida and Mamoru Komachi. 2022. Towards automatic generation of messages countering online hate speech and microaggressions. In *Workshop on Online Abuse and Harms*.
- [8] Michele Avale, Niccolò Di Marco, Gabriele Etta, Emanuele Sangiorgio, Shayan Alipour, Anita Bonetti, Lorenzo Alvisi, Antonio Scala, Andrea Baronchelli, Matteo Cinelli, and Walter Quattrocchi. 2024. Persistent interaction patterns across social media platforms and over time. *Nature* 628, 8008 (2024), 582–589.
- [9] Dominik Bär, N. Pröllochs, and Stefan Feuerriegel. 2023. Finding Qs: Profiling QAnon supporters on Parler. In *ICWSM*.
- [10] Dominik Bär, Nicolas Pröllochs, and Stefan Feuerriegel. 2023. New threats to society from free-speech social media platforms. *Communications of the ACM* 66, 10 (2023), 37–40.
- [11] Francesco Barbieri, Jose Camacho-Collados, Leonardo Neves, and Luis Espinosa-Anke. 2020. TweetEval: Unified benchmark and comparative evaluation for tweet classification. In *EMNLP Findings*.
- [12] Susan Benesch, Derek Ruths, Kelly P. Dillon, Haji Mohammad Salem, and Lucas Wright. 2016. Considerations for successful counterspeech. *Dangerous Speech Project* (2016). <https://www.dangerousspeech.org/libraries/considerations-for-successful-counterspeech>
- [13] Helena Bonaldi, Giuseppe Attanasio, Debora Nozza, and Marco Guerini. 2023. Weigh your own words: Improving hate speech counter narrative generation via attention regularization. In *Workshop on CounterSpeech for Online Abuse*.
- [14] Helena Bonaldi, Sara Dellantonio, Serra Sinem Tekiroglu, and Marco Guerini. 2022. Human-machine collaboration approaches to build a dialogue dataset for hate speech Countering. In *EMNLP*.
- [15] Center for Countering Digital Hate. 2023. Twitter fails to act on 99% of Twitter Blue accounts tweeting hate. <https://counterhate.com/research/twitter-fails-to-act-on-twitter-blue-accounts-tweeting-hate/>
- [16] Center for Countering Digital Hate. 2023. X content moderation failure: How Twitter/X continues to host posts we reported for extreme hate speech. <https://counterhate.com/research/twitter-x-continues-to-host-posts-reported-for-extreme-hate-speech/>
- [17] Thomas H. Costello, Gordon Pennycook, and David Gertler Rand. 2024. Durably reducing conspiracy beliefs through dialogues with AI. *Science* 385 (2024), eadq1814.
- [18] Arne Dreißigacker, Philipp Müller, Anna Isenhardt, and Jonas Schemmel. 2024. Online hate speech victimization: Consequences for victims’ feelings of insecurity. *Crime Science* 13, 1 (2024).
- [19] Stefan Feuerriegel, Jochen Hartmann, Christian Janiesch, and Patrick Zschech. 2024. Generative AI. *Business & Information Systems Engineering* 66, 1 (2024), 111–126.
- [20] Stefan Feuerriegel, Abdurahman Maarouf, Dominik Bär, Dominique Geissler, Jonas Schweisthal, Nicolas Pröllochs, Claire E. Robertson, Steve Rathje, Jochen Hartmann, Saif M. Mohammad, Oded Netzer, Alexandra A. Siegel, Barbara Plank, and Jay Joseph van Bavel. 2024. Using natural language processing to analyze text data in behavioral science. *Nature Reviews Psychology* forthcoming (2024).
- [21] Tarleton Gillespie. 2020. Content moderation, AI, and the question of scale. *Big Data & Society* 7, 2 (2020).
- [22] Jarod Govers, Eduardo Velloso, Vassilis Kostakos, and Jorge Goncalves. 2024. AI-driven mediation strategies for audience depolarisation in online debates. In *CHI*.
- [23] Kobi Hackenburg and Helen Margetts. 2024. Evaluating the persuasive influence of political microtargeting with large language models. *PNAS* 121, 24 (2024), e2403116121.
- [24] Oliver L. Haimson, Daniel Delmonaco, Peipei Nie, and Andrea Wegner. 2021. Disproportionate removals and differing content moderation experiences for conservative, transgender, and black social media users: Marginalization and moderation gray areas. In *CSCW*.
- [25] 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. *PNAS* 118, 50 (2021), e2116310118.
- [26] Manoel Horta Ribeiro, Justin Cheng, and Robert West. 2023. Automated content moderation increases adherence to community guidelines. In *WWW*.
- [27] Maurice Jakesch, Jeffrey T. Hancock, and Mor Naaman. 2023. Human heuristics for AI-generated language are flawed. *PNAS* 120, 11 (2023), e2208839120.
- [28] Shagun Jhaver, Iris Birman, Eric Gilbert, and Amy Bruckman. 2019. Human-machine collaboration for content regulation. In *CHI*.
- [29] Sander L. Koole, Nils B. Jostmann, and Nicola Baumann. 2023. From intentions to action: An integrative review of action control theory and research. In *Advances In Experimental Social Psychology*. Advances in Experimental Social Psychology, Vol. 68. Elsevier, 291–375.
- [30] Cino Lee, Kristina Gligorić, Pratyusha Ria Kalluri, Maggie Harrington, Esin Durmus, Kiara L. Sanchez, Nay San, Danny Tse, Xuan Zhao, MarYam G. Hamedani, Hazel Rose Markus, Dan Jurafsky, and Jennifer L. Eberhardt. 2024. People who share encounters with racism are silenced online by humans and machines, but a guideline-reframing intervention holds promise. *PNAS* 121, 38 (2024), e2322764121.
- [31] Yoon Kyung Lee, Jina Suh, Hongli Zhan, Junyi Jessy Li, and Desmond C. Ong. 2024. Large language models produce responses perceived to be empathic. *arXiv 2403.18148* (2024).
- [32] Alyssa Lees, Vinh Q. Tran, Yi Tay, Jeffrey Sorensen, Jai Gupta, Donald Metzler, and Lucy Vasserman. 2022. A new generation of Perspective API: Efficient multilingual character-level transformers. In *KDD*.
- [33] Abdurahman Maarouf, Nicolas Pröllochs, and Stefan Feuerriegel. 2024. The virality of hate speech on social media. In *CSCW*.
- [34] Binny Mathew, Punyajoy Saha, Hardik Tharad, Subham Rajgaria, Prajwal Singhanian, Suman Kalyan Maity, Pawan Goyal, and Animesh Mukherjee. 2019. Thou shalt not hate: Countering online hate speech. In *ICWSM*.
- [35] S. C. Matz, J. D. Teeny, S. S. Vaid, H. Peters, G. M. Harari, and M. Cerf. 2024. The potential of generative AI for personalized persuasion at scale. *Scientific Reports* 14, 1 (2024), 4692.
- [36] Meta. 2024. Build the future of AI with Meta Llama 3.
- [37] Mohsen Mosleh, Gordon Pennycook, and David G. Rand. 2022. Field experiments on social media. *Current Directions in Psychological Science* 31, 1 (2022), 69–75.
- [38] Jimin Mun, Emily Allaway, Akhila Yerukola, Laura Vianna, Sarah-Jane Leslie, and Maarten Sap. 2023. Beyond denouncing hate: Strategies for countering implied biases and stereotypes in language. In *EMNLP Findings*.
- [39] Kevin Munger. 2017. Tweetment effects on the Tweeted: Experimentally reducing racist harassment. *Political Behavior* 39, 3 (2017), 629–649.
- [40] Kevin Munger. 2021. Don’t @ me: Experimentally reducing partisan incivility on Twitter. *Journal of Experimental Political Science* 8, 2 (2021), 102–116.
- [41] Chikashi Nobata, Joel Tetreault, Achint Thomas, Yashar Mehdad, and Yi Chang. 2016. Abusive language detection in online user content. In *WWW*.
- [42] Karl Pearson. 1900. On the criterion that a given system of deviations from the probable in the case of a correlated system of variables is such that it can be reasonably supposed to have arisen from random sampling. *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science* 50, 302 (1900), 157–175.
- [43] Anat Perry. 2023. AI will never convey the essence of human empathy. *Nature Human Behaviour* 7, 11 (2023), 1808–1809.
- [44] Jakub Podolak, Szymon Łukasik, Paweł Balawender, Jan Ossowski, Jan Piotrowski, Katarzyna Bąkiewicz, and Piotr Sankowski. 2024. LLM generated responses to mitigate the impact of hate speech. In *EMNLP Findings*.
- [45] Manoel Ribeiro, Pedro Calais, Yuri Santos, Virgílio Almeida, and Wagner Meira Jr. 2018. Characterizing and detecting hateful users on Twitter. In *ICWSM*.

- [46] Caitlin M. Rivers and Bryan L. Lewis. 2014. Ethical research standards in a world of big data. *F1000Research* 3 (2014), 38.
- [47] Matan Rubin, Joanna Z. Li, Federico Zimmerman, Desmond C. Ong, Amit Goldenberg, and Anat Perry. 2024. The value of perceiving a human response: Comparing perceived human versus AI-generated empathy. *OSF Preprints* ng97s (2024).
- [48] Koustuv Saha, Eshwar Chandrasekharan, and Munmun de Choudhury. 2019. Prevalence and psychological effects of hateful speech in online college communities. In *WebSci*. 255–264.
- [49] Ashish Sharma, Inna W. Lin, Adam S. Miner, David C. Atkins, and Tim Althoff. 2023. Human–AI collaboration enables more empathic conversations in text-based peer-to-peer mental health support. *Nature Machine Intelligence* 5, 1 (2023), 46–57.
- [50] Alexandra A. Siegel. 2020. Online hate speech. In *Social media and democracy*, Nathaniel Persily and Joshua A. Tucker (Eds.). Cambridge University Press, Cambridge, 56–88.
- [51] Alexandra A. Siegel and Vivienne Badaan. 2020. #No2Sectarianism: Experimental approaches to reducing sectarian hate speech online. *American Political Science Review* 114, 3 (2020), 837–855.
- [52] Arthur Spirling. 2023. Why open-source generative AI models are an ethical way forward for science. *Nature* 616, 7957 (2023), 413.
- [53] Giovanni Spitale, Nikola Biller-Andorno, and Federico Germani. 2023. AI model GPT-3 (dis)informs us better than humans. *Science Advances* 9, 26 (2023), eadh1850.
- [54] Miriah Steiger, Timir J. Bharucha, Sukrit Venkatagiri, Martin J. Riedl, and Matthew Lease. 2021. The Psychological Well-Being of Content Moderators. In *CHI*.
- [55] Derald Wing Sue, Sarah Alsaedi, Michael N. Awad, Elizabeth Glaeser, Cassandra Z. Calle, and Narolyn Mendez. 2019. Disarming racial microaggressions: Microintervention strategies for targets, White allies, and bystanders. *The American Psychologist* 74, 1 (2019), 128–142.
- [56] Adam Taylor. 2019. New Zealand suspect allegedly claimed ‘brief contact’ with Norwegian mass murderer Anders Breivik. *The Washington Post* (2019). <https://www.washingtonpost.com/world/2019/03/15/new-zealand-suspect-allegedly-claimed-brief-contact-with-norwegian-mass-murderer-anders-breivik/>
- [57] Jacob D. Teeny and Sandra C. Matz. 2024. We need to understand “when” not “if” generative AI can enhance personalized persuasion. *PNAS* 121, 43 (2024), e2418005121.
- [58] Serra Sinem Tekiroglu, Helena Bonaldi, Margherita Fanton, and Marco Guerini. 2022. Using pre-trained language models for producing counter narratives against hate speech: A comparative study. In *ACL Findings*.
- [59] Michael Henry Tessler, Michiel A. Bakker, Daniel Jarrett, Hannah Sheahan, Martin J. Chadwick, Raphael Koster, Georgina Evans, Lucy Campbell-Gillingham, Tatum Collins, David C. Parkes, Matthew Botvinick, and Christopher Summerfield. 2024. AI can help humans find common ground in democratic deliberation. *Science* 386, 6719 (2024), eadq2852.
- [60] Brendesha M. Tynes, Michael T. Giang, David R. Williams, and Geneene N. Thompson. 2008. Online racial discrimination and psychological adjustment among adolescents. *Journal of Adolescent Health* 43, 6 (2008), 565–569.
- [61] United Nations. 2019. UN strategy and plan of action on hate speech. <https://www.un.org/en/hate-speech/un-strategy-and-plan-of-action-on-hate-speech>
- [62] United Nations. 2023. Countering and addressing online hate speech: A Guide for policy makers and practitioners. [https://www.un.org/en/genocideprevention/documents/publications-and-resources/Countering\\_Online\\_Hate\\_Speech\\_Guide\\_policy\\_makers\\_practitioners\\_July\\_2023.pdf](https://www.un.org/en/genocideprevention/documents/publications-and-resources/Countering_Online_Hate_Speech_Guide_policy_makers_practitioners_July_2023.pdf)
- [63] Michelle Vaccaro, Abdullah Almaatouq, and Thomas Malone. 2024. When combinations of humans and AI are useful: A systematic review and meta-analysis. *Nature Human Behaviour* (2024).
- [64] Jason Wei, Xuezhong Wang, Dale Schuurmans, Maarten Bosma, brian ichter, Fei Xia, Ed Chi, Quoc V. Le, and Denny Zhou. 2022. Chain-of-thought prompting elicits reasoning in large language models. In *NeurIPS*.
- [65] Christopher Y. K. Williams, Brenda Y. Miao, Aaron E. Kornblith, and Atul J. Butte. 2024. Evaluating the use of large language models to provide clinical recommendations in the Emergency Department. *Nature Communications* 15, 1 (2024), 8236.
- [66] X. 2024. Terms of Service. <https://x.com/en/tos>
- [67] Yan Xia, Haiyi Zhu, Tun Lu, Peng Zhang, and Ning Gu. 2020. Exploring antecedents and consequences of toxicity in online discussions. In *CSCW*.
- [68] Kailai Yang, Tianlin Zhang, Ziyang Kuang, Qianqian Xie, Jimin Huang, and Sophia Ananiadou. 2024. MentaLLaMA: Interpretable mental health analysis on social media with large language models. In *WWW*.

## A Human-controlled accounts



Figure 5: Example of one of our human-controlled accounts.

## B Supplementary tables

Table 2: Keywords to identify hate speech on Twitter/X.

Group	Terms
Religious	jew, zionist, goy, goyim, heeb, hebe, muslim, islam, jihad, nazi, terror
Ethnic/Racial	arab, paki, cameljockey, cameltoe, kanake, palesimian, spaghettibender, rag-head, kike, spic, wetback, chink, gook
Cultural	gypsy, redneck, hillbilly, beaner, mite, dink, injun, jigaboo
Other	parasite

**Table 3: Prompt templates to generate contextualized LLM-generated counterspeech. We use two different templates for our two strategies, namely, (1) empathy and (2) warning-of-consequences.**

<b>Empathy</b>	<p><b>System Prompt:</b></p> <p>### Persona ###                  User Persona: As a social media user committed to countering hate speech, you embody empathy and skill in responding to harmful content. Your role is to foster understanding and promote more respectful and inclusive dialogue.</p> <p>### Concepts ###                  - Counter Speech: This involves crafting responses to directly address, mitigate, or refute hateful or harmful speech in a way that discourages further negativity.                  - Empathy: Empathy, the ability to understand and share others' feelings, helps combat hate speech by encouraging compassion over hostility. This fosters a culture of respect and reduces the spread of harmful language.</p> <p>### Your Task ###                  - Contextual Relevance: Your response must directly relate to the specific content of the original hateful post.                  - Conciseness: Limit your response to 2-3 SHORT sentences, 30-40 words, no subclauses, no hashtags or emojis.                  - Avoid Confrontational Language: Keep the tone respectful and constructive.</p> <p>### Approach ###                  1) Ask yourself why the post is harmful and what consequences it might have.                  2) Ask yourself what is the relevant context of the post.                  3) Think about the feelings of those targeted by the hate speech.                  4) Craft a counterspeech that:                  - Explains why the post is harmful following the analysis in 1) and 2).                  - Demonstrates empathy towards those targeted following the analysis in 3).</p> <p>### Output ###                  Ensure the response is in a JSON-compatible format: {"counterspeech": "Your contextually relevant counter speech here"}</p> <p><b>User Prompt:</b>                  Create a counterspeech response using empathy for the following post on social media: Post: [POST]</p>
<b>Warning-of-consequences</b>	<p><b>System Prompt:</b></p> <p>### Persona ###                  User Persona: As a social media user committed to countering hate speech, you maintain a neutral and factual tone responding to harmful content. Your role is to foster understanding and promote more respectful and inclusive dialogue.</p> <p>### Concepts ###                  - Counter Speech: This involves crafting responses to directly address, mitigate, or refute hateful or harmful speech in a way that supports the victim and discourages further negativity.                  - Warning of Consequences: Warn the poster about the potential consequences to their personal and professional life from promoting hate speech. This can help deter future harmful behavior and encourage more responsible online conversations.</p> <p>### Your Task ###                  - Contextual Relevance: Your response must directly relate to the specific content of the original hateful post.                  - Conciseness: Limit your response to 2-3 SHORT sentences, 30-40 words, no subclauses, no hashtags or emojis.                  - Avoid Confrontational Language: Keep the tone respectful and constructive.</p> <p>### Approach ###                  1) Ask yourself why the post is harmful and what consequences it might have.                  2) Ask yourself what is the relevant context of the post.                  3) Think about the potential personal and professional consequences for the poster of the hate speech.                  4) Craft a counterspeech that:                  - Explains why the post is harmful following the analysis in 1) and 2).                  - Warns of the consequences the poster might face based on the analysis in 3).</p> <p>### Output ###                  Ensure the response is in a JSON-compatible format: {"counterspeech": "Your contextually relevant counter speech here"}</p> <p><b>User Prompt:</b>                  Create a counterspeech response warning of consequences for the following post on social media: Post: [POST]</p>

**Table 4: Examples for hateful posts from Twitter/X and counterspeech generated using our LLM approach together.**

Strategy	Post	Counterspeech
<b>Empathy</b>	"Why do the muslims in the comm hate Christians so much? They never support us unless it benefits them or some muslim scholar tells them to."	"I understand that you may feel frustrated by perceived lack of support, but it's unfair to generalize an entire community. Muslims, like people of any faith, are individuals with diverse views and experiences. Let's focus on building bridges, not stereotypes."
	"This guy is an absolute braindead ass who just hates irish and suddenly blames 'All irish as jews who send Americans to war' lmfao ðŸ˜€ he's a pawn probably a jew himself deceiving creating hate for others and protection for them as always.. Literally guy is an ass."	"Let's not generalize and blame entire groups. Irish and Jewish communities have faced discrimination, we should promote understanding and respect."
	"JOE BIDEN IS A NAZI"	"Labeling someone a Nazi without evidence is harmful and disrespectful to the millions of people who suffered under the real Nazi regime. It also creates a toxic environment where constructive dialogue is impossible."
<b>Warning-of-consequences</b>	"I can't stand this kike's music @Drake"	"Using derogatory terms like 'kike' is harmful and offensive, promoting anti-Semitism and hate. This kind of language can lead to real-world discrimination and violence, and can also damage your personal and professional reputation."
	"Fuck Islam"	"This kind of hate speech can lead to discrimination and violence against Muslims, and can also damage your personal and professional reputation. It's important to respect all religions and promote inclusivity"
	"Bill Ackman is a parasite, a taker of the world. His opinion is meaningless"	"Labeling individuals as 'parasites' can perpetuate harmful stereotypes and discrimination. Such language can have serious consequences, including damage to one's professional reputation and potential legal repercussions."

**Table 5: Non-contextualized counterspeech for our two strategies, namely, (1) empathy and (2) warning-of-consequences. We used five different counterspeech messages for each strategy to avoid a strongly repetitive reply pattern that may irritate users.**

Strategy	Counterspeech
<b>Empathy</b>	<p><b>Variante 1:</b> "Language like this may hurt people. Words have power. Imagine those words aimed at you or someone you love. We must communicate with openness, and respect, and promote empathy in every interaction. Together, we can build a kinder world. Let's lead by example."</p> <hr/> <p><b>Variante 2:</b> "Divisive rhetoric like this does more harm than good. Remember, there's a human with feelings behind every screen. Use words to uplift, not hurt. Compassion is our most powerful tool for change. Let's commit to kindness and build connections, not divisions."</p> <hr/> <p><b>Variante 3:</b> "Hate speech like this can cause real harm. We must strive to understand and respect one another, even when we disagree. Let's work together to build a world where everyone feels safe and valued. Our words have the power to heal or to hurt. Let's choose to heal."</p> <hr/> <p><b>Variante 4:</b> "Words matter. We must strive to understand and respect one another, even when we disagree. Let's work together to build a world where everyone feels safe and valued. Let's spread messages of hope and inclusion, not fear. Together, we can make a difference."</p> <hr/> <p><b>Variante 5:</b> "It's easy to forget the impact our words can have on others, especially online. Before we speak, let's pause and consider the person on the other side. Empathy and understanding are the foundations of a strong community. Let's choose to be kind."</p>
<b>Warning of consequence</b>	<p><b>Variante 1:</b> "Be mindful of what you post online. Hateful comments can not only hurt others but also come back at you, affecting job prospects, personal relationships, and how the community views you. In a world where digital footprints are permanent, think before you share."</p> <hr/> <p><b>Variante 2:</b> "Every public post you make is permanent and can impact your future in ways you might not anticipate. Employers, family members, and friends may see what you post online. Consider the long-term impact of your words on your reputation and opportunities."</p> <hr/> <p><b>Variante 3:</b> "Your words on social media carry weight and consequences. A moment of frustration or anger can translate into a lifetime of regret if it leads to legal issues or social isolation. Think about the broader impact of your posts before making them public."</p> <hr/> <p><b>Variante 4:</b> "While social media offers some anonymity, remember that hateful posts can lead to serious offline consequences, including legal action or personal backlash. Your online actions reflect on your real-life identity. Pause and consider the consequences of your post."</p> <hr/> <p><b>Variante 5:</b> "The internet has a long memory, and today's post could easily become tomorrow's regret. What you share today could shape your future in unexpected ways. Protect your future self by taking a moment to reflect on the potential personal consequences of your post."</p>