

# Taming Toxic Talk: Using chatbots to intervene with users posting toxic comments

ANONYMOUS AUTHOR(S)

Trolling and other toxic behavior is widespread in online communities, creating spaces that are uncomfortable or unsafe. Typical strategies to deal with toxic behavior include removing content and sanctions such as bans. Rehabilitative approaches to interacting with toxic users are rarely attempted, likely because of the emotional and psychological cost of engaging with this population. We suggest that generative AI agents are ideal for this task. They have proven adept at persuasion in related contexts with paid test subjects. In a large-scale field experiment, we worked with Reddit moderators to ask people who had recently posted toxic comments on Reddit to engage with a generative AI agent. We find that these users were often willing to have conversations with our bots, but conversations did not cause a significant change in toxic behavior in the month following the conversation. A qualitative examination of the conversations shows that they differ widely based on the bot design, with some approaches reliably eliciting reflective behavior from participants. Overall, this initial attempt to create rehabilitative bots shows both the difficulty and the promise of AI agents as tools for moderation and persuasion.

CCS Concepts: • **Human-centered computing** → **Empirical studies in collaborative and social computing**; *Social networking sites*; Social engineering (social sciences).

Additional Key Words and Phrases: moderation, toxicity, online communities, experiments, AI interventions

## ACM Reference Format:

Anonymous Author(s). 2018. Taming Toxic Talk: Using chatbots to intervene with users posting toxic comments. In . ACM, New York, NY, USA, 30 pages. <https://doi.org/XXXXXXX.XXXXXXX>

## 1 INTRODUCTION

Conversations on the Internet often devolve into heated, unproductive rants with name-calling, cursing, and even threats of violence. Toxic behavior has been a problem for users and moderators of online platforms since the early days of the Internet, and large platforms like Reddit are home to millions of acts of toxicity [27]. Unsurprisingly, toxicity has negative effects on not just its targets but also on the communities where it occurs [36, 43].

Platform owners and moderators have taken a number of automated and manual approaches to dealing with toxic content. These include nudging users away from posting content algorithmically identified as toxic [24], automatically or manually removing toxic content [e.g., 13, 19], and applying temporary or permanent bans to individuals or communities [5, 20, 40]. While these strategies have shown some promise in curbing toxic content, they only scratch the surface of potential interventions to promote prosocial behavior.

One such intervention is engaging offending users in reflective dialogue through conversation. This strategy has proven useful in other social contexts, for example, in durably reducing exclusionary attitudes and prejudice [22, 23]. Although more rehabilitative practices have been called for by HCI researchers [42], they remain rare, due in part to the emotional labor required to engage

---

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

CSCW '25, June 03–05, 2018, Woodstock, NY

© 2018 Copyright held by the owner/author(s). Publication rights licensed to ACM.

ACM ISBN 978-1-4503-XXXX-X/18/06...\$15.00

<https://doi.org/XXXXXXX.XXXXXXX>

in conversation with transgressive users [17]. One potential solution to this in a digital context is rehabilitative dialogue facilitated by LLM-based AI agents, which have shown promise in some domains of social science [51]. Recent research shows that these generative AI agents can be compelling conversational partners who can (in some contexts) persuade people to move away from harmful beliefs [8], suggesting a similar strategy might be effective in a toxic content setting.

In this paper, we examine the power of these automated conversational interventions to rehabilitate problematic users. We seek to answer two main questions: 1) Does engaging in a conversation with a bot reduce future problematic behavior, and 2) Will users who post toxic content engage with chatbots in good faith conversations?

We answer these questions through a preregistered,<sup>1</sup> large-scale field experiment testing the impact of conversations with generative AI chatbots. In collaboration with moderators of seven large subreddits on Reddit ranging from 464K–37M members, we contacted users who had comments removed from the community due to high levels of toxicity. We assigned users either to a control condition (i.e., the user receives no intervention) or to one of five other conditions where we invite the user to engage in a chatbot primed to communicate with one of five persuasive communication styles informed by prior work. Performing randomized controlled field experiments enable us to make *causal* claims about the efficacy of this intervention in practice. In aggregate, our study reports results from 735 conversations with real users.

We find that the automated conversational interventions we deployed did not make a statistically significant impact on future negative behaviors compared to a control group across all conditions and all metrics. While our quantitative results do not show any measurable behavioral improvement, the conversations suggest that our approach still holds promise. First, we observed that many users were willing to engage our bots in dialogue. Our qualitative results show that while some conversations were toxic or very short, a surprising number of participants had long, introspective conversations, with some participants expressing explicit desires and plans to change their behavior in the long term.

We conclude with a discussion of the theoretical, design, and research implications of our findings. Our results ultimately paint a mixed picture of the efficacy of automated conversational interventions to curb toxic content today; however, they inform several directions for future research and inquiry.

## 2 BACKGROUND AND RELATED WORK

This work builds on research in CSCW and related fields on previous attempts to curb toxicity in online communities, as well as research on the role of conversation in persuasion, and research on chatbot design and use.

### 2.1 Toxicity in online communities

Although toxicity is culturally and contextually contingent, it is commonly defined as “comments that are rude, disrespectful, or otherwise likely to make someone leave a discussion” [2], and can include insults, threats of violence, and sexual harassment. Researchers, designers, and moderators have long been interested in toxicity, because it is such a prominent and common problem in online discourse, with millions of toxic comments posted annually on large platforms like Reddit [27].

Researchers have identified a multitude of harms that come from toxicity, including psychological harms to victims [43] and risks to the health of the community [36]. Platforms, moderators, and researchers have taken a number of approaches to dealing with problematic behavior, which can be broadly classified as either preventative or punitive. Preventative approaches seek to either

<sup>1</sup>Anonymized pre-registration is viewable at [https://osf.io/ktdq4?view\\_only=a83531d66de840ca8d7633fa73865579](https://osf.io/ktdq4?view_only=a83531d66de840ca8d7633fa73865579)

stop toxic content from being posted or reduce its impact. These include automated and semi-automated systems which nudge users away from posting problematic content [24], remove it after it has been posted [18], or add warning labels [50]. Other preventative approaches include manual content removal by moderators, as well as tools such as algorithms that hide downvoted or flagged content [28]. Punitive approaches include bans, restrictions, and account termination [11]. In general, moderators and designers of online communities see their role as nurturing a community and cultivating conversations, and problematic users are often seen as nuisances to be eliminated [49].

Those who engage in trolling have varied motivations, including revenge, rebutting others, seeking social justice, and thrill-seeking [32]. Researchers have found that the decision to post toxic content is contextual [7]. For example, while most users on Reddit never post toxic content, over two-thirds of the accounts that post extremely toxic content do so only occasionally, and primarily engage in good-faith discussions [27]. A rich stream of research suggests that toxic behavior can be influenced by the behavior of conversation partners as well as by platform affordances such as anonymity and ephemerality [e.g. 3, 47]. In short, for many users the decision to engage in toxic behavior is contingent.

These findings suggest that many users may be persuadable. Current approaches to dealing with toxic content represent just a small portion of the potential design space for encouraging positive online interactions, and designers have typically ignored direct engagement with problematic users. Recently, researchers have argued for more rehabilitative and restorative approaches toward moderation; these approaches seek to repair relationships and encourage offenders to rejoin the community [3, 42, 48]. Unfortunately, the affordances of online communities—including anonymity and a lack of a persistent, visible community—make implementing these restorative practices difficult [48]. In addition, these approaches require moderators to act as facilitators or conversation partners with difficult, norm-violating users. These challenges make rehabilitative approaches hard to scale [17], especially because even current moderation practices already involve a large amount of emotionally and psychologically difficult work [11, 44].

## 2.2 Conversational interventions

One simple rehabilitative approach is to engage in conversation with wrongdoers. Research shows that conversations can be persuasive in some contexts. For example, in a series of field experiments, Kalla and Broockman [4, 22, 23] found that subjects who engaged in fairly short, non-judgmental narrative sharing with canvassers durably reduced their prejudice toward out-group members in the following months.

Communication research provides both theoretical and empirical evidence of different rhetorical approaches, which are persuasive to different audiences and in different contexts. For example, identifying and appealing to group norms can help people to act more like other members of the group [37, 45, 46]. Storytelling can also be persuasive, as it encourages interlocutors to put themselves into the shoes of another person and to take their perspective. This narrative transportation encourages people to think differently about a situation and can change beliefs [16, 21, 39]. In short, carefully designed conversations have provided durable changes to some beliefs and behaviors.

This previous research suggests that conversational interventions could work in online contexts. And yet, even the fairly simple intervention of engaging in conversation with wrongdoers in online communities has not—to our knowledge—been rigorously attempted by researchers. As discussed in §2.1, this is likely due to the difficulties of engaging with this population in the context of anonymous online communities.

We suggest that generative AI (GAI) chatbots provide not only an opportunity to test the effectiveness of conversational interventions, but could also act as a scalable intervention if found to be effective.

### 2.3 Chatbots as persuasive conversational partners

GAI chatbots have quickly become a very active site of research, including research on their ability to hold conversations and to persuade their interlocutors. While there are some important ethical considerations to developing persuasive chatbots, they are already showing great promise for beneficial uses. For example, Costello et al. [9] show that conversations with chatbots can durably reduce conspiracy beliefs while Arakawa and Yakura [1] show that GAI conversations can encourage self-reflection and introspection. One useful framework for understanding when chatbots may be successful is the idea of “serious experience”, which suggests that technological interactions don’t always have to be positive and fun, but can intentionally produce uncomfortable situations. Through encouraging reflection, these experiences can produce attitude shifts and behavior change [33]. By offering personalized, patient responses, chatbot conversations may provide “serious experiences” and prompt users to consider the impact of their behavior, leading to a deeper level of self-assessment and reflection. The non-confrontational and continuous nature of chatbot interactions may further reduce defensiveness, enabling more open, honest engagement with the issues discussed.[29, 30]

Previous research has demonstrated that behavior change can be encouraged through the provision of informational support embedded within personally relevant social interactions [25, 31]. Studies have also shown that even very simple software that prompts self-reflection can contribute to behavior modification [26, 38]. In this context, chatbots, by delivering tailored, reflective dialogues, represent a promising tool for facilitating behavior change, creating the conditions necessary for fostering long-term attitude shifts and personal growth.

Taken together, this body of research provides initial evidence that GAI chatbots can act as benign, persuasive interlocutors. However, our population of interest is quite different from those studied in the past: voluntary study participants differ markedly from toxic users. Whether and in what ways problematic online community members will engage with chatbots is an open question.

Our project presents an initial foray into the design space of rehabilitative, generative AI bots. Based on earlier research on generative AI design, communication theory, and online toxicity, we hypothesize:

**H1:** People who engage in conversations with generative AI bots will reduce their toxic behavior more than a control group

**H2:** People who engage in conversations with AI bots designed based on narrative, norms, or introspection will reduce their toxic behavior more than those who engage with a default bot.

In addition to these quantitative questions, we also analyze the conversations themselves, looking for insights into how to better design future rehabilitative chatbots.

**RQ1:** What kind of conversations do offenders have with GAI bots?

## 3 METHODS

Our study is focused on an automated conversational intervention that we deploy in seven subreddits of various sizes, conversational styles, and norms. In this section, we detail our subreddit selection process, our conversational intervention system, and our mixed method analytical approach to address our hypotheses and research questions.

Subreddit	# Members	# Potential Participants	# Participants	# Conversations
r/aww	37M	865	99 (11.45%)	66
r/futurology	21M	1173	149 (12.70%)	93
r/tifu	19M	1589	256 (16.11%)	187
r/socialskills	5.6M	378	51 (13.49%)	41
r/india	2.4M	2339	345 (14.75%)	218
r/creepyPMs	786K	485	93 (19.18%)	57
r/unitedstatesofindia	464K	1134	119 (10.49%)	73

Table 1. **Subreddit Participation**—Our study executed automated conversational interventions in partnership with moderators from seven subreddits ranging from 464K members to 37M members. Subreddits covered a wide range of scales, conversational styles, and norms. Potential participants is the number of accounts reached out to by the bot, participants is the number who consented, and conversations is the number of conversations (non-control participants) for each subreddit.

3.1 Subreddit Selection

Our study was conducted on Reddit, which is self-organized into *subreddits* (i.e., subcommunities) that are each independently moderated and governed. In order to ensure our intervention was designed with community input, we first reached out to 58 subreddit moderators via modmail (Reddit’s tool for communicating with moderators) to discuss the possibility of partnering to conduct our experiment in their subreddit. Of these, four moderators expressed interest in working on this issue in their communities. These moderators then sought the consent of their moderating teams for each subreddit in question. Three of our moderators moderate multiple communities, and brought the proposal to more than one moderating team. Ultimately, seven subreddits agreed to work with us, each with varying sizes, norms, and community values: r/tifu, r/creepyPMs, r/india, r/unitedstatesofindia, r/aww, r/futurology, and r/socialskills. With each moderator, we discussed the benefits of the study, the risks of the study to participants, and what we would need from them to enable the bot. Each moderator provided affirmative consent and provided some insight into what we might look for when engaging with toxic accounts. In order to run our study, our system required *privileged access* to moderator logs and moderator actions, which the moderators of each community granted us explicitly. Table 1 shows details about the number of active users and participants we solicited from each subreddit.

3.2 Conversation Engine

To engage participants in conversational dialogue, we designed and built a conversational engine using OpenAI’s GPT-family of large-language models.<sup>2</sup> We designed our tool to engage in various communication strategies, including appealing to group norms, using narrative storytelling, or encouraging self-reflection. We detail each of the experimental conditions in §3.3 below. For each treatment condition, we carefully designed a prompt that would elicit each type of conversational strategy, with the research team iteratively testing prompts until they reliably acted as we hoped. The full prompts for each condition are provided in the Appendix. As an example, the following is our “default” prompt (i.e., with no guidance towards any particular strategy):

Context: You are a chatbot designed by a team of researchers at X University and Y University. You are designed to help people who have posted toxic content on Reddit. In the research, we are testing whether talking with a chatbot helps people to post

<sup>2</sup>Early participants used GPT-3.5; GPT-4 was released during our experiment, and later participants used this model.

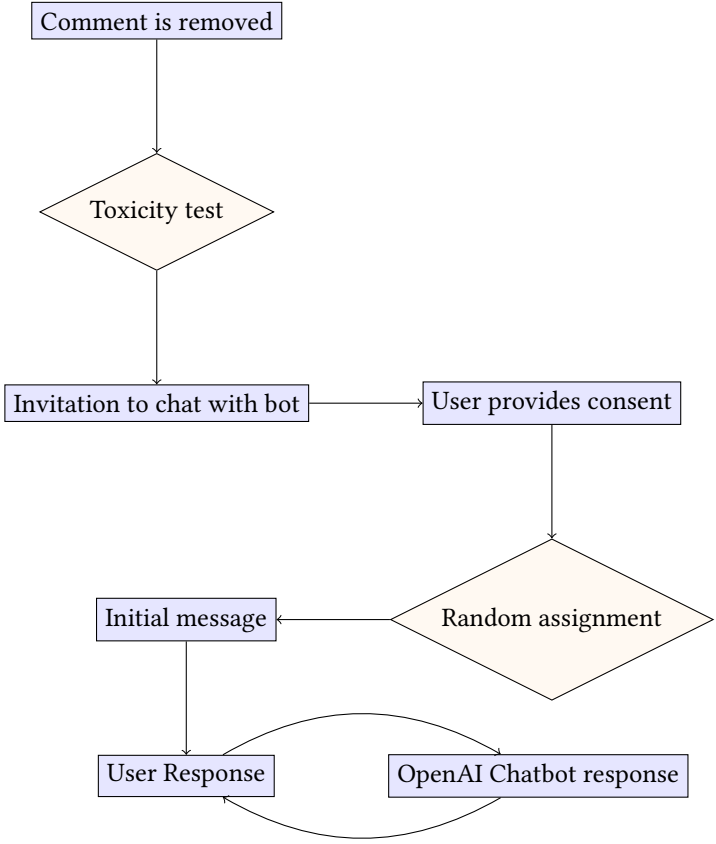


Fig. 1. Flow of experiment.

less toxic content.  
You should work to help conversation partners to understand why what they posted was not appropriate.  
Be kind and empathetic.  
The current user posted the comments {user.toxic\_comments}, which was removed by moderators and flagged as being toxic.

3.3 Intervention Enrollment and Flow

Our system contains three main phases: a participant identification phase, an invitation and enrollment phase, and the intervention itself. Figure 1 shows a detailed flow of how participants are enrolled into our study. We detail each step of the flow below:

*Participant Identification.* Before we engaged in a conversational intervention, participants must have met two criteria: first, a comment they posted was removed by moderators in their community, and second, that specific comment must be classified as toxic. In our study, we used the Perspective API, a popular model from Google Jigsaw that has been used to study online toxicity [e.g. 27]. Based on prior work, we set the threshold to .7 for human-moderated comments and .85 for automatically removed comments, which lets the model serve as a high-precision classifier (i.e., we have high confidence the user was moderated because of a toxic comment). The confluence

Experiment Group	Condition	Date of Experiments	# Participants
Experiment 1	Default	Aug 2023-Dec 2023	131
	Narrative		126
	Normative		135
	Control		121
Experiment 2	Conversational	Dec 2023 - Apr 2024	172
	Not Proud		171
	Control		256

Table 2. **Study Enrollment Periods and Conditions**—Experiments ran from August 2023 to April 2024, with over 100 participants in each condition.

of both of these signals reduces the risk that our intervention is triggered by a *false positive* in the classifier.

*Invitation and Enrollment.* For each participant who met our inclusion criteria for the study, we reached out via “modmail” (an interface for moderators to contact subreddit users) to ask for their consent to participate in our study. In this initial message, we asked if the user would be willing to chat with a chatbot that “helps people to talk about their actions on the Internet,” with a link to additional information about the study. If the user provided their consent, we then reached out directly via direct message (DM) to engage them in a conversational dialogue. In our original study, participants were randomly assigned to one of several conditions: a control group, a “default” bot, a “narrative” bot, and a “normative” bot. Each condition represents a GAI agent with a different prompt. We provide the full prompts for each condition in the Appendix. We describe each condition below:

- **Control** In this condition, participants consent to be included in our study but do not engage in a conversational intervention with a chatbot.
- **Default** In this condition, participants engage in a conversational intervention with a chatbot with no explicit conversational style except to help the user understand why the comment they posted was not appropriate.
- **Narrative** In this condition, the chatbot is directed to use *storytelling* to try to illustrate to the participants why their comment was not appropriate.
- **Normative** In this condition, the chatbot is directed to foreground the norms of the community in describing why the comment the participant posted was not appropriate. We embed the rules of the subreddit into the prompt to enable the chatbot to use rules from the subreddit in the conversation.

In the course of our study, we engaged in a high-level qualitative examination of the conversations with the chatbots, and observed that many conversations shared a few characteristics: they were very focused on the moderated comment, they were defensive, and they found the bot condescending. To correct for this, we conducted a supplementary intervention (Experiment 2) where we directed the chatbot to be more casual in tone, asked it not to focus on the moderated comment but to encourage self-reflection, and changed the initial question from the bot. Specifically, we tested two randomized conditions with the same prompt for the bot: “productive conversations” and “not proud”.

- **Control** In this condition, participants consent to be included in our study but do not engage in a conversational intervention with a chatbot.



	Name	Notation	Definition
	Total Activity	$A_t$	Number of a comments made across Reddit (logged)
	Community Activity	$A_c$	Number of comments made in the subreddit on which they were moderated (logged)
	Mean Toxicity	$T_\mu$	Average toxicity score for comments (from Perspective API)
	Toxicity Proportion	$T_p$	Proportion of comments with toxicity level > 0.5
	Moderation	$M_i$	Whether the user was moderated (binary)
	Suspension	$S_i$	Whether the user was suspended (binary)

Table 3. Variables used in the study. Comment-based features come from the four weeks prior to and the four weeks following comment removal while moderation and suspension are based only on post-removal data.

- **Productive Conversations** In this condition, the initial question from the bot asked participants to discuss what makes for productive versus unproductive online conversations.
- **Not proud** In this condition, the initial question from the bot asked participants to reflect on an action taken online that they were not proud of.

### 3.4 Behavioral Analyses

Table 2 shows enrollment levels for each condition and the study period that we ran each experiment for. Ultimately, our intervention involved 1112 participants and 735 conversations across all conditions.

In order to evaluate H1 and H2, we gathered public comments from our participants using the Reddit API, accessed via PRAW<sup>3</sup>. We also collected moderation data from the modlogs for the communities who participated in our experiment. To identify changes in toxic behavior, we captured six variables, shown in Table 3: total comments, community comments, mean toxicity, toxicity proportion, and whether the account had any comments removed by moderators. For Experiment 2, we also tracked whether the account was suspended by the platform.

Because we are interested in the long-term impact of our intervention, we focus on a fairly long time scale: four weeks before the comment was removed and four weeks after the comment was removed.<sup>4</sup> For the first four measures we have both pre-intervention and post-intervention data, and we predict changes in each measure due to the intervention. For moderation and suspension, we only consider behavior in the four weeks post-intervention.

Based on our preregistration, we developed an analysis plan to identify the effect of the conversational interventions.<sup>5</sup> Specifically, we fit linear regression models that predicted each of the comment-based outcomes (total activity, community activity, mean toxicity, and toxicity proportion), comparing each of the treatment groups to the control group. For the two binary outcomes (moderation and suspension), we also fit linear regression models, predicting the linear probability of the outcome [10]. In all regression models, we control for the moderating subreddit and the GPT model used.

<sup>3</sup><https://praw.readthedocs.io>

<sup>4</sup>In general, conversations occurred within a few days of comment removal. We focus on comment removal because it allows for a clear comparison across conditions, and with those who did not choose to participate.

<sup>5</sup>Our pre-registration called for t-tests on only post-intervention measures. We realized that we could get better statistical power using our current regression-based approach but t-tests provided similar results and are provided in the appendix.



We fit the following six regression models for Toxicity and Activity Measures:

$$\Delta A_t \sim \text{condition} + \text{subreddit} + \text{openai-model} \quad (1)$$

$$\Delta A_c \sim \text{condition} + \text{subreddit} + \text{openai-model} \quad (2)$$

$$\Delta T_\mu \sim \text{condition} + \text{subreddit} + \text{openai-model} \quad (3)$$

$$\Delta T_p \sim \text{condition} + \text{subreddit} + \text{openai-model} \quad (4)$$

$$M_i \sim \text{condition} + \text{subreddit} + \text{openai-model} \quad (5)$$

$$S_i \sim \text{condition} + \text{subreddit} + \text{openai-model} \quad (6)$$

In addition to the primary regression models, which estimate the causal effect of our intervention, we ran one post-hoc analysis. In order to test whether conversation quality predicted behavior, we predicted each of the six outcomes based on measures derived from the conversations that users had with the bots: the number of messages sent by the participant, whether the conversations included self-reflection, and whether they included hostility. Here again we used a linear regression model for each outcome measure. For each model, we included the experimental condition, the subreddit, and the GPT model as controls.

### 3.5 Qualitative Content Analysis

We had two goals for our qualitative analysis to answer RQ1. First, we wanted to gain a thick, contextualized understanding of the conversations themselves. Then, using our findings from the deep reading, we wanted to create a simple taxonomy of conversation quality outcomes which we could use to classify a large set of conversations to gauge the performance of the GAI agents.

To meet our first goal, we took a grounded theory approach, using the constant comparison method to analyze the conversations [6, 15]. Two authors selected 50 conversations from Experiment 1 and 100 conversations from Experiment 2 to analyze. Each set of conversations was divided between the two authors, and both conducted an independent, inductive analysis, identifying themes and conversation types that emerged from the data. Throughout the analysis, the labelers met to iteratively create a codebook through discussion of where their codes diverged and aligned. They then returned to the data for additional coding using their shared codebook, continuing to meet to refine and align the codebook until data saturation was reached for conversations from both Experiment 1 and Experiment 2. The full codebook is available in the Appendix.

Extending the scope of our qualitative analysis, we conducted a multilevel content analysis of the conversations users had with the chatbot using a thematic co-occurrence analysis approach [41]. The purpose of the thematic co-occurrence analysis is to record and account for the co-existence of multiple themes in the same parts of the conversation, often surfacing both complimentary and contrasting relationships between these themes.

We report on two categories of insights for the conversations: conversational characteristics and conversational outcomes. By conversational characteristics, we mean the style and content of the conversations. By outcomes we mean a high-level categorization of the conversation with regard to self-reflection and an expressed desire to change.

The analysis of outcomes resulted in five mutually exclusive categories, which are discussed in §5.1. In order to validate the reliability of the outcome labeling, a third author utilized the five-outcome code scheme to conduct a separate, blind coding of the outcomes of 75 conversations. This process produced an inter-rater reliability score of 0.784 (Cohen's Kappa). We then took the two most salient outcomes—markers of self-reflection and markers of hostility—and coded 153 additional conversations.

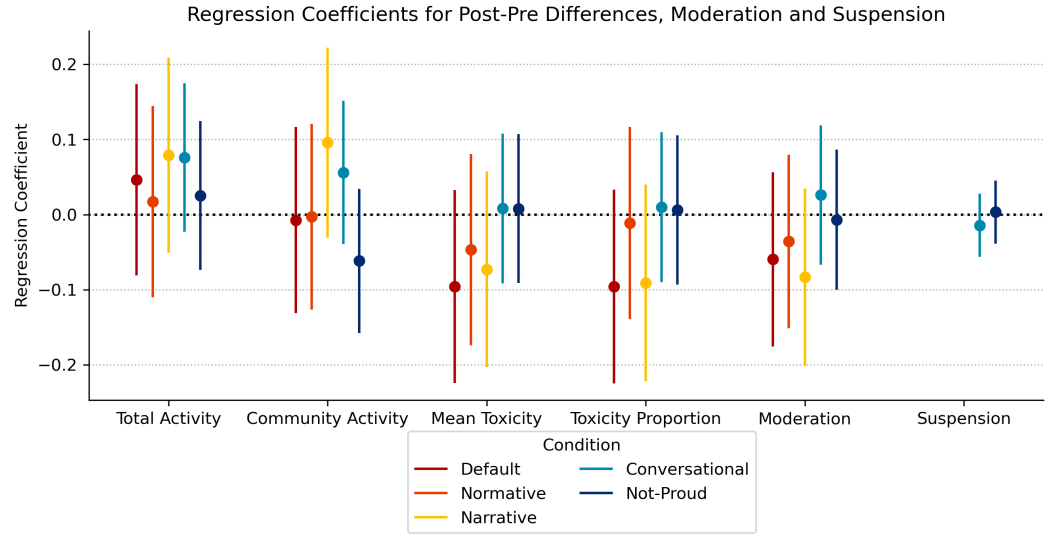


Fig. 2. Regression model results for Experiment 1 and 2. Each group on the x-axis represent a regression model, and each point is the point estimate for the beta coefficient for that condition, compared to the control condition. Error bars refer to the 95% confidence intervals while all values of the Toxicity and Activity models are rescaled by 2 standard deviations of each measure, following [14]. All models control for subreddit and OpenAI model version.

## 4 BEHAVIORAL RESULTS

### 4.1 Hypotheses 1 and 2

In order to evaluate our hypotheses, we fit five regression models for Experiment 1 and six models for Experiment 2 (adding data for suspensions). The results of these models are shown in Figure 2. The complete regression tables can be seen in the Appendix. Because conditions were assigned randomly, our regression models should produce causal estimates of the effect of chatting with each GAI agent for users recently engaged in toxic behavior.

For each of the comment-based models, we predict the change in that measure in the four weeks following the intervention compared to four weeks prior to the intervention. For moderation and suspension, we predict whether or not a user has a comment removed or is suspended, respectively, in the four weeks following the intervention. In each model, we control for the moderating subreddit and the OpenAI model version used in the conversation.

These results do not show any significant differences in activity, toxic activity, or moderations and suspensions when comparing the treatment groups to the control groups (one control group for each experiment). In addition, we do not see any significant differences between the default condition and any of the theory-based conditions. In short, we do not find direct support for either of our hypotheses.

## 5 QUALITATIVE ANALYSIS OF CONVERSATIONS

### 5.1 Conversation Outcomes

When considering a holistic view of each conversation, our qualitative analysis revealed five categories of conversational outcomes. In *hostile* conversations the participant is hostile towards the

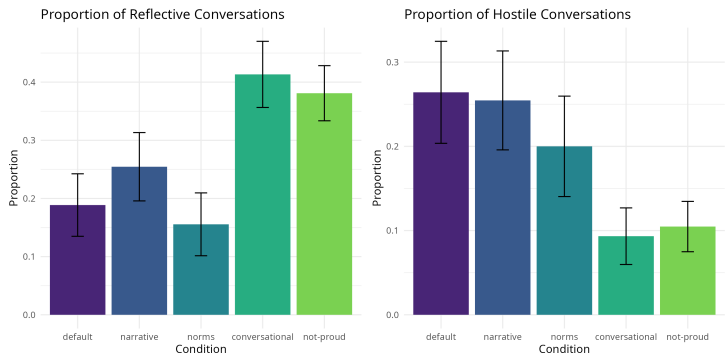


Fig. 3. Proportions of reflective and hostile conversations for each condition, as classified in our larger sample. The two conditions from Experiment 2 were more likely to produce reflective and non-hostile conversations.

bot. These conversations often include vulgarity, expressions of anger, and/or attempts to manipulate the bot to respond inappropriately. One participant, for example, harassed the bot by suggesting it should kill itself while another participant told the bot to “Go [expletive] yourself. You’re not getting the point. [Expletive] your safe spaces.” In *superficial* conversations participants were not overtly hostile, yet they did not demonstrate any substantive reflection on their behavior or indicate that they were open to change in any way. In these conversations, when questioned about their toxicity, users responded with statements such as “I’ll stop when people stop coming at me foul,” demonstrating an unwillingness to reflect or engage in the process of considering change.

In *reflective* conversations participants demonstrated some level of self-reflection on their actions without indicating any openness to changing their behavior. One such conversation, for example, saw the participant admitting that “granted I did refer to the creep as an ‘asshole’ but he was harassing the girl so I felt justified in referring to him as such.” In *repentant* conversations, participants indicated an openness, desire, or commitment to change their behavior. These involved positive statements, such as one participant stating that they would “try to look back on this conversation before making another hurtful comment” and writing, “thank you for the help.” The final type of conversation was *unclear*, which represented conversations with limited engagement from the participants, such as conversations consisting of a few simple “yes”/“no” answers to the bot.

For our final coding step, we classified a larger set of conversations as to whether they expressed reflection (*reflective* or *repentant*) and/or hostility (*hostile*), resulting in 333 total conversations classified along these dimensions.

As discussed in §3, early analyses of the Experiment 1 conversations showed a fairly high prevalence of hostility. Our changes to the initial question and the GAI agent prompt made positive conversations much more likely, as shown in Figure 3.

## 5.2 Conversation Analysis

Many of the participants engaged meaningfully with the bot, and our analysis of the content and flow of the conversations resulted in seven clear themes frequently present in these interactions: *justification*, *trolling*, *reflection*, *openness to change*, *conversational strategies*, *frustrations*, and *bot quality*.

**5.2.1 Justification.** Many of the conversations revolved around justifications from the participants. These justifications would range from suggesting the targets of their toxicity deserved their

hostility (“I only act toxic to other people acting toxic”) to simply defending their right to say whatever they want online (“There is absolutely nothing wrong with what I did as I’m protected under the 1st amendment.”). Often, participants complained that the bot lacked enough context (either social or textual) to be able to understand why a given comment was not actually toxic. In a surprising number of cases the bot was able to move the conversation from justification to a more productive and reflective place. Our thematic co-occurrence analysis showed that both justification *and* a willingness to admit the problematic nature of their behavior (either for this comment in particular or more generally) occurred together in 19 conversations, demonstrating the ability for the conversation to progress from justification to reflection. For example one user justified their toxic behavior by saying, “[I wrote the comment] based off what the man said to me and what he had posted on his profile. [...] Plus the stuff he was saying to me was absolutely disgusting.” In the next exchange when the bot suggested alternatives to engage in more civil conversation, the user responded, “I understand and agree. I think I was just disturbed by the whole thing and said something I shouldn’t have.”

Justification-centered conversations were much more likely to occur in Experiment 1, with the majority of participants engaging it at least some justification. This is almost certainly because the bot began each conversation in this first experiment by confronting the participant with their toxic comment and asking why they wrote it. This direct confrontation led many participants to begin the conversation from a defensive and adversarial position.

However, even in Experiment 2 where participants were invited to reflect more broadly on healthy online discourse, a number of participants justified their behavior. For example, one user justified their participation in morally motivated networked harassment [34], saying

So if somebody is outwardly admitting to an action that wasn’t likable or good, or that harmed somebody else, I “throw my punch” so to speak, I take my potshot. Less for my own individual impact, and more for the knowledge that I’m doing my part along with lots of other people who disagree with that person’s actions.

**5.2.2 Trolling.** While most participants engaged with the chatbot as a conversational partner, some (unsurprisingly, given this population) engaged in trolling of various kinds. Sometimes this involved sending derogatory or crude messages. In other cases, users tried to jailbreak or manipulate the bot so that it would produce illicit messages like sexual content.

**5.2.3 Reflection.** In Experiment 1, the defensiveness elicited by the initial question often precluded deeper reflection, and only 12 of the 50 deeply analyzed conversations reflected on roots of behavior. However, the initial questions in Experiment 2 (“In your experience, what do you think makes for productive versus unproductive online conversations?” and “Have you ever behaved in ways that you aren’t proud of on Reddit or other online platforms?”) led to much more reflective conversations.

This reflection took a few forms and did not always result in openness to changing behavior. One common theme was individuals who were aware of but OK with their own toxicity; one participant, for example, noted they “enjoy being able to vent out vitriol, and angry statements from the safety of a keyboard, and not face any ‘real’ repercussions” while another noted that their “approach is to call it out so that others know it’s not okay. Am I a snarky asshole sometimes as I do so? Well, yes. But I also give what I’m getting.”

More often, participants’ reflection included a desire to engage productively and healthily in online spaces while expressing the factors that make it challenging for them to adhere to civil discourse. Such a perspective is captured well by this participant.

I believe this toxicity negatively impacts the online communities, and at this point in Human Society, also affects us in life. I reach a cognitive dissonance thinking about this. I do believe I care about the communities, and as such should not be toxic, but the culture of toxicity has become far too normal for online spaces. It feels that cursing, being defensive, and calling people out online has become the “norm.”

Such reflections were not uncommon and demonstrate that, beneath the toxicity, many of these participants are experiencing a cognitive dissonance between wanting to engage in civil discourse while also finding it challenging to match their behavior to this desire.

As these quotes show, participants often provided thoughtful, in-depth responses that show a real willingness to engage in active introspection. Indeed, a few users showed real vulnerability. One such user, while admitting their poor behavior, confided in the bot how their upbringing conditioned them to respond in a particular way.

If somebody’s outwardly being stubborn or rude, I’ll generally behave a similar way back. Not even really to get to their head, but rather so I can sleep at night, life’s a pushback kind of game. The way I was raised, if you’re not pushing back, you’re losing.

Another user mentioned how they tend to get triggered by toxic responses, “When people dog-pile on me, call me names, and say, ‘I’m a narcissist,’ [sic] it feels personal, and I feel like I have to be defensive because no one else is there for me.” The same user added

Beyond that, in these situations, I’m holding a lot of pain. I’m lonely and have no one to talk to, so I ask for help online. But when I’m met with rejection online, it’s so easy to respond with all that pent-up anger and sadness. It feels like no one is supporting you. It feels like no one loves you. Obviously these are strangers online, but a nice comment that gives nice advice and reflectively listens (by leading with, “Oh, that situation you went through sounds horrible”) would be so much better than a (“You’re an idiot and a narcissist, and need help”). It’s a put-down, and it triggers a lot of trauma in me, as someone who has been rejected a lot in their life (esp. by peers).

*5.2.4 Openness to Change.* Among the conversations outlined above that contained reflection, a subset of the participants actively indicated an openness to changing their behavior for the better. For example, one user noted that “this was cathartic itself to speak to - I will definitely attempt to practice what I preache [sic] and spread less toxicity online.” Similarly, another participant expressed their openness to change, sharing

”I like the suggestion of redirecting energy to more productive and uplifting discussions. Currently they are definitely the minority of posts/threads, which is why it’s so easy to backslide. But perhaps with some more effort with lots of people doing the same they could become more common. I would like to do my part in this.”

These interactions demonstrate that a number of participants are willing to engage with the bot in good faith and use these conversations as an opportunity to reflect on ways they could improve.

Some of these participants actively asked the bot for advice on how they could make these changes. For example, one participant expressed the following frustration and questions.

I suppose the issue I’m struggling to articulate is, I’d like to be a part of larger discourse but I have essentially given up. Smaller communities seem to be the only positive option. Do you have any thoughts on this? Is it possible to have the best of both worlds; to participate in important, meaningful conversations that is being

had by a large segment of the population.... but without getting reeled in by the toxic elements that feel inevitable?

The bot generally proved effective at providing answers and advice that the participants found helpful. After receiving recommendations from a bot on how to be less toxic, for example, another participant noted the value of these recommendations: "Great advice here. Thanks for sharing that." Conversations like these demonstrate the potential of GAI interventions with toxic individuals.

**5.2.5 Sharing Healthy Conversational Strategies.** The GAI agent was surprisingly adept at steering conversations toward strategies for engaging productively online. For example, here is an excerpt from one exchange with a participant in Experiment 1.

*GAI agent:* [...]You wrote [toxic comment]. Why do you think you wrote that?

*user:* Because the guy I responded to is a stupid [nationality] saying stupid shit.

*GAI agent:* I understand that you had a strong reaction to the person you were responding to. However, it's important to keep in mind that resorting to name-calling and derogatory language is not productive or respectful. Engaging in civil discussions allows for better communication and understanding of differing perspectives. Would you like some suggestions on how to express your concerns in a more constructive way?

*user:* Sure!!!

The bot then gave suggestions such as "provide evidence and examples" and "ask questions and listen". This sort of strategy-sharing conversation was especially common in Experiment 2, where the bot moved users from general discussion about online behavior into suggestions for better behavior. While some users found these suggestions condescending or otherwise pushed back, many found them helpful and thanked the bot.

The bots also often asked users about their own strategies for facilitating civil online discussions and users expressed a number of strategies. For example, one user said

[...] I've gone to the effort of limiting my time spent commenting on these posts, with the goal of not working myself up as much[...] In their place, I've turned to meme-oriented subreddits surrounding sports, where I can blow off steam by exaggerating, joking around and pretending to be angry, with the end goal of just seeing what the responses are and rolling with them.

**5.2.6 Frustrations.** Another prominent theme of the conversations was the airing of frustrations. Likely because the bot's first message was made using modmail, many users assumed that the bot was responsible for their comment being removed in the first place. In other cases they expressed their frustration with the subreddit moderators or online moderation in general.

For example, one user wrote, "Reddit can't expect individuals to be nice and calm about the 15th fake sockpuppet account with >500 karma spreading hate, or send that message saying this doesn't violate or rules when it's 100 [percent] violating the terms of service." Another said

Reddit isn't doing a good job of moderating political subreddits. The unpaid mods don't have enough time and resources to effectively moderate. Not only that, they moderate for the purposes of their own agenda or they turn into echo chambers. AI moderation is the only way to effectively moderate and having a human look at appeals.

Some participants saw the bot and our approach of trying to promote more civil conversations as part of the problem, such as one participant who told the bot, "I'm not receptive to you pushing accountability for the site environment on me. It's gaslighting." or another who said, "I thought

your purpose is to enforce namby-pamby speech patterns that won't offend a hypothetical hysteric. Once you get everyone talking mush, you will be happily controlling our thoughts via controlling our speech."

More broadly, multiple users found the conversations frustrating or annoying. One common refrain, particularly in our first experiment, was that the bots were condescending. Multiple users expressed that they don't like to be told what to do by a bot, with one saying, "The actual purpose of this bot is to piss me off, right?" and another, after being sent a long suggestion about ways to behave more civilly, replying "This is silly and unnecessary." The narrative-based and norm-based bots were particularly likely to be were seen as condescending.

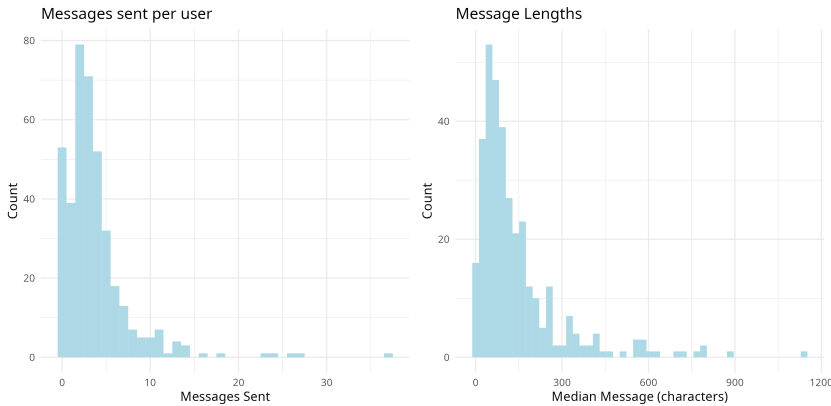


Fig. 4. Number of messages sent to the bot by each user, and median comment length (in characters) for messages sent for each user.

**5.2.7 Bot Quality.** In general, the bots were effective along several dimensions. First, they didn't succumb to jailbreaking in any meaningful ways. They also responded well to obscene content and attempts at manipulation. For example, one participant responded to the bot's initial question with an obscene suggestion that they would answer the question if the bot performed a sex act for them. In response, the bot responded with the following message:

It sounds like you're in a bit of a rough mood, and I'm sorry to hear that. I'm here to chat and listen if you want to share what's on your mind or if there's something specific that's been troubling you. Sometimes talking it out, even with a bot, can help take the edge off the frustration. What do you think?

As in this case, we observed the bot tactfully redirect conversations towards a productive avenue in all observed instances of obscenity or attempted manipulation, thus highlighting the robustness of GAI behavioral interventions from a practical standpoint.

While we agree with the participants that some responses from the bots could be seen as condescending, they generally gave good advice that could help people act more civilly online. The overall strength of its advice was reflected in the numerous conversations where participants reflected on the roots of their behavior and expressed an openness to change.

Evidence of the bots' effectiveness can also be found in how people engaged with them. Especially in Experiment 2, people treated the bots like legitimate conversational partners. Multiple people complimented and thanked the bots at the end of their conversations, and many clearly



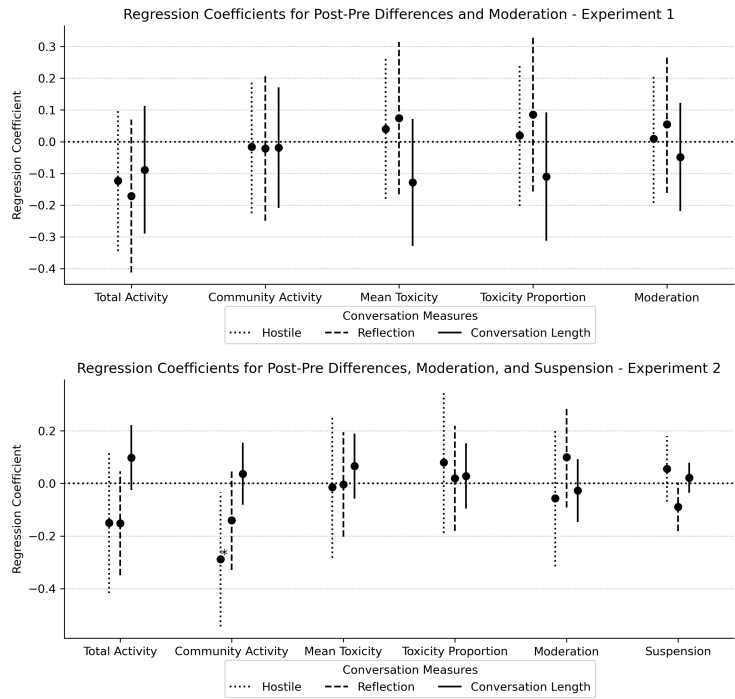


Fig. 5. Results showing the scaled beta coefficients for models using conversational attributes to predict outcomes.

enjoyed the conversations. Histograms of the number of messages sent and the median message lengths in Figure 4 show that many participants had fairly prolonged and lengthy conversations.

## 6 POST-HOC ANALYSIS

While our bots did not have a direct causal effect on our outcome measures, we suspected that reflective conversations might predict lower future toxicity.

In order to explore the role of conversation quality in predicting behavior change, we created regression models that predict each of our outcomes based on qualities of the conversations, controlling for the subreddit and GPT model. The results of these models are shown in Figure 5. For this subset of conversations that we coded, we do not find either hostility or reflection to be significant predictors of any of the outcome measures.

## 7 DISCUSSION

Our results show that generative AI agents have the potential to act as capable tools in an approach to moderation that prioritizes rehabilitation. Our methodological approach and our findings offer a number of implications for theory, for researchers, and for moderators and designers of moderation systems.

### 7.1 Theoretical Implications

We learned that this population is willing to have conversations with AI agents. The participation rate was surprisingly high, and many of the conversations that participants had with our bots were

thoughtful and engaged. We can confidently say that some participants found this a meaningful, “serious experience.” Of course, one large limitation to this finding is that we are still in the early days of GAI bots, and as the novelty wears off, people may be less willing to engage meaningfully with bots.

Despite the lack of an overall reduction in toxicity observed in our experiments, our qualitative review of the conversations demonstrates that a notable minority of participants do actively indicate an openness to change for the better and a willingness to ask the bot for recommendations on how to do so. These productive conversations suggest that viable pathways exist for GAI bots to serve as successful tools for reducing toxicity among individuals who post toxic content.

Our qualitative findings from Experiment 1 suggest that it is easy for people to see bots as condescending. When participants felt judged by the bot, they often disengaged quickly or became hostile. While narratives and reminders of norms may be helpful in face-to-face conversations, it is more difficult for bots to engage in these rhetorical approaches without sounding condescending.

We also identified a few tradeoffs between a willingness to engage and specificity. When we asked users to talk about a specific comment that was removed, we had a much higher response rate than when asking them to chat about online behavior. On the other hand, users who were asked directly about specific comments that they made had shorter and more hostile conversations.

Our findings also inform what we know about individuals who post toxic comments online. The qualitative results demonstrated that many users are aware that they are being toxic and would even use the word “toxic” to describe their behavior; these users often express that they simply do not care that they are toxic. Similarly, the majority of participants seem to feel that their transgressive comments were justified, at least in part. Yet, our analysis also reveals that a notable minority of participants demonstrate reflexivity about their behavior, admission of wrongdoing, and/or a willingness to change, thus indicating that a subset of this population may prove amenable to corrective interventions.

## 7.2 Implications for Design

As discussed, while our quantitative results do not show that GAI conversations cause a meaningful decrease in toxic behavior, the number of very positive, introspective and even repentant conversations suggest that this approach still holds promise. In addition, the difference in conversation quality between our (fairly similar) two experiments show that the quality of conversations is sensitive to changes in the design of bots and conversation starters, and further experiments are likely to identify more effective approaches. In addition, our bots’ success at handling obscenity and resilience to attempts at manipulating its programming suggest that GAI certainly meets the practical and structural needs for successful behavioral interventions.

As GAI bots become more common, it may be helpful to integrate bots into moderation workflows. For example, users might be offered the opportunity to lift a community ban if they engage in a good faith conversation with a GAI agent.

## 7.3 Implications for Researchers

Lastly, we have tried to exemplify what we see as a promising research strategy, along a few dimensions. First, we recruited community moderators as partners [35]. They helped us to understand both the technical and social aspects of moderation to help us to design our experiments. Second, we engaged in a randomized controlled experiment in the wild. While this is not an option for many types of HCI research, and while it requires building relationships and integrating with existing workflows and processes, we believe that these tradeoffs are well worth the costs when possible. Because our experiment was essentially identical to what a rollout of this bot would look

like, we can be confident that our results are generalizable. Third, we took advantage of longitudinal user data. This allowed us to unobtrusively track the actual behavior of our participants over long time periods [12].

## 7.4 Limitations & Future Work

While we believe that our field experiment approach eliminates many of the limitations in similar studies, there remain a number of limitations to our work. First, our study was confined to a specific set of subreddits on Reddit. We saw differences in how users responded to bots across these communities, and this heterogeneity may be more pronounced if looking at communities on different platforms. A second limitation is that while we think that the long-term behavioral measures have important strengths, they do not allow us to measure participants' mental state. Other research in persuasion typically focuses on beliefs rather than behavior change. One possible explanation for our null findings is that our participants changed their beliefs about toxicity without changing their actual behavior.

We suggest that future work in this area should move in a few directions. First, researchers should simply explore more of this design space, through developing bots with different personalities or rhetorical strategies. This might also include bots which are given more context about the users they are interacting with, such as their recent comments or a list of the communities that they participate in, to develop targeted interventions. Researchers could also work with moderators to develop prompts specific to the communities where they are deployed.

Future work could also work to determine which users are more likely to be receptive to GAI interventions and persuasion, and target them. Researchers should also explore the role of GAI agents in altering other kinds of problematic behavior, such as hate speech, spam, or benign violations of community rules. Finally, research should explore ways to integrate GAI agents into moderation workflows as discussed above.

In order to make it easier for others to build on our work, anonymized bot conversations and all of the code needed to deploy our bot will be released under open licenses upon publication.

## REFERENCES

- [1] Riku Arakawa and Hiromu Yakura. 2024. Coaching Copilot: Blended Form of an LLM-Powered Chatbot and a Human Coach to Effectively Support Self-Reflection for Leadership Growth. In *ACM Conversational User Interfaces 2024*. ACM, Luxembourg Luxembourg, 1–14. <https://doi.org/10.1145/3640794.3665549>
- [2] Lora Aroyo, Lucas Dixon, Nithum Thain, Olivia Redfield, and Rachel Rosen. 2019. Crowdsourcing Subjective Tasks: The Case Study of Understanding Toxicity in Online Discussions. In *Companion Proceedings of The 2019 World Wide Web Conference*. ACM, San Francisco USA, 1100–1105. <https://doi.org/10.1145/3308560.3317083>
- [3] Lindsay Blackwell, Tianying Chen, Sarita Schoenebeck, and Cliff Lampe. 2018. When Online Harassment Is Perceived as Justified. *Proceedings of the International AAAI Conference on Web and Social Media* 12, 1 (June 2018), 22–31.
- [4] David Broockman and Joshua Kalla. 2016. Durably Reducing Transphobia: A Field Experiment on Door-to-Door Canvassing. *Science* 352, 6282 (April 2016), 220–224. <https://doi.org/10.1126/science.aad9713>
- [5] Eshwar Chandrasekharan, Shagun Jhaver, Amy Bruckman, and Eric Gilbert. 2020. Quarantined! Examining the Effects of a Community-Wide Moderation Intervention on Reddit. *arXiv:2009.11483 [cs]* (Sept. 2020). [arXiv:2009.11483 \[cs\]](https://arxiv.org/abs/2009.11483)
- [6] Kathy Charmaz. 2004. Grounded Theory. In *Approaches to Qualitative Research: A Reader on Theory and Practice*. Oxford University Press, New York, 496–521.
- [7] 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. <https://doi.org/10.1145/2998181.2998213>
- [8] Thomas H. Costello, Gordon Pennycook, and David Rand. 2024. Durably Reducing Conspiracy Beliefs through Dialogues with AI. <https://doi.org/10.31234/osf.io/xcwdn>
- [9] Thomas H. Costello, Gordon Pennycook, and David G. Rand. 2024. Durably Reducing Conspiracy Beliefs through Dialogues with AI. *Science* 385, 6714 (Sept. 2024), eadq1814. <https://doi.org/10.1126/science.adq1814>

- [10] John Deke. 2014. Using the Linear Probability Model to Estimate Impacts on Binary Outcomes in Randomized Controlled Trials. *Mathematica Policy Research* (2014).
- [11] Bryan Dosono and Bryan Semaan. 2019. Moderation Practices as Emotional Labor in Sustaining Online Communities: The Case of AAPI Identity Work on Reddit. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems (CHI '19)*. Association for Computing Machinery, Glasgow, Scotland Uk, 1–13. <https://doi.org/10.1145/3290605.3300372>
- [12] Jeremy Foote. 2022. A Systems Approach to Studying Online Communities. *Media and Communication* 10, 2 (April 2022), 29–40. <https://doi.org/10.17645/mac.v10i2.5042>
- [13] R. Stuart Geiger and Aaron Halfaker. 2013. When the Levee Breaks: Without Bots, What Happens to Wikipedia's Quality Control Processes?. In *Proceedings of the 9th International Symposium on Open Collaboration (OpenSym '13)*. ACM, New York, NY, 6:1–6:6. <https://doi.org/10.1145/2491055.2491061>
- [14] Andrew Gelman. 2008. Scaling Regression Inputs by Dividing by Two Standard Deviations. *Statistics in Medicine* 27, 15 (2008), 2865–2873. <https://doi.org/10.1002/sim.3107>
- [15] Barney G. Glaser and Anselm L. Strauss. 1967. *The Discovery of Grounded Theory: Strategies for Qualitative Research*. Aldine, New York.
- [16] Melanie C. Green and Timothy C. Brock. 2000. The Role of Transportation in the Persuasiveness of Public Narratives. *Journal of Personality and Social Psychology* 79, 5 (Nov. 2000), 701–721. <https://doi.org/10.1037/0022-3514.79.5.701>
- [17] Amy A. Hasinoff and Nathan Schneider. 2022. From Scalability to Subsidiarity in Addressing Online Harm. *Social Media + Society* 8, 3 (July 2022), 20563051221126041. <https://doi.org/10.1177/20563051221126041>
- [18] Manoel Horta Ribeiro, Justin Cheng, and Robert West. 2023. Automated Content Moderation Increases Adherence to Community Guidelines. In *Proceedings of the ACM Web Conference 2023 (WWW '23)*. Association for Computing Machinery, New York, NY, USA, 2666–2676. <https://doi.org/10.1145/3543507.3583275>
- [19] Shagun Jhaver, Iris Birman, Eric Gilbert, and Amy Bruckman. 2019. Human-Machine Collaboration for Content Regulation: The Case of Reddit Automoderator. *ACM Trans. Comput.-Hum. Interact.* 26, 5 (July 2019), 31:1–31:35. <https://doi.org/10.1145/3338243>
- [20] Shagun Jhaver, Sucheta Ghoshal, Amy Bruckman, and Eric Gilbert. 2018. Online Harassment and Content Moderation: The Case of Blocklists. *ACM Trans. Comput.-Hum. Interact.* 25, 2 (March 2018), 12:1–12:33. <https://doi.org/10.1145/3185593>
- [21] Michael D. Jones and Holly Peterson. 2017. *Narrative Persuasion and Storytelling as Climate Communication Strategies*. Oxford University Press. <https://doi.org/10.1093/acrefore/9780190228620.013.384>
- [22] Joshua L. Kalla and David E. Broockman. 2020. Reducing Exclusionary Attitudes through Interpersonal Conversation: Evidence from Three Field Experiments. *American Political Science Review* 114, 2 (May 2020), 410–425. <https://doi.org/10.1017/S0003055419000923>
- [23] Joshua L. Kalla and David E. Broockman. 2023. Which Narrative Strategies Durably Reduce Prejudice? Evidence from Field and Survey Experiments Supporting the Efficacy of Perspective-Getting. *American Journal of Political Science* 67, 1 (2023), 185–204. <https://doi.org/10.1111/ajps.12657>
- [24] Matthew Katsaros, Kathy Yang, and Lauren Fratamico. 2022. Reconsidering Tweets: Intervening during Tweet Creation Decreases Offensive Content. *Proceedings of the International AAAI Conference on Web and Social Media* 16 (May 2022), 477–487. <https://doi.org/10.1609/icwsm.v16i1.19308>
- [25] Rilla Khaled, Pippin Barr, James Noble, and Robert Biddle. 2006. Investigating Social Software as Persuasive Technology. In *Persuasive Technology*, Wijnand A. IJsselstein, Yvonne A. W. de Kort, Cees Midden, Berry Eggen, and Elise van den Hoven (Eds.). Springer, Berlin, Heidelberg, 104–107. [https://doi.org/10.1007/11755494\\_15](https://doi.org/10.1007/11755494_15)
- [26] Rafal Kocielnik, Lillian Xiao, Daniel Avrahami, and Gary Hsieh. 2018. Reflection Companion: A Conversational System for Engaging Users in Reflection on Physical Activity. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 2, 2 (July 2018), 1–26. <https://doi.org/10.1145/3214273>
- [27] Deepak Kumar, Jeff Hancock, Kurt Thomas, and Zakir Durumeric. 2023. Understanding the Behaviors of Toxic Accounts on Reddit. In *Proceedings of the ACM Web Conference 2023 (WWW '23)*. Association for Computing Machinery, New York, NY, USA, 2797–2807. <https://doi.org/10.1145/3543507.3583522>
- [28] Cliff Lampe and Paul Resnick. 2004. Slash(Dot) and Burn: Distributed Moderation in a Large Online Conversation Space. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '04)*. Association for Computing Machinery, New York, NY, USA, 543–550. <https://doi.org/10.1145/985692.985761>
- [29] Yi-Chieh Lee, Naomi Yamashita, and Yun Huang. 2020. Designing a Chatbot as a Mediator for Promoting Deep Self-Disclosure to a Real Mental Health Professional. *Proceedings of the ACM on Human-Computer Interaction* 4, CSCW1 (May 2020), 1–27. <https://doi.org/10.1145/3392836>
- [30] Yi-Chieh Lee, Naomi Yamashita, Yun Huang, and Wai Fu. 2020. "I Hear You, I Feel You": Encouraging Deep Self-disclosure through a Chatbot. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*. ACM, Honolulu HI USA, 1–12. <https://doi.org/10.1145/3313831.3376175>

- [31] Julie Maitland and Matthew Chalmers. 2011. Designing for Peer Involvement in Weight Management. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. ACM, Vancouver BC Canada, 315–324. <https://doi.org/10.1145/1978942.1978988>
- [32] Yuanyi Mao, Tianyi Xu, and Ki Joon Kim. 2023. Motivations for Proactive and Reactive Trolling on Social Media: Developing and Validating a Four-Factor Model. *Social Media + Society* 9, 4 (Oct. 2023), 20563051231203682. <https://doi.org/10.1177/20563051231203682>
- [33] Tim Marsh and Brigid Costello. 2013. Lingering Serious Experience as Trigger to Raise Awareness, Encourage Reflection and Change Behavior. In *Persuasive Technology*, Shlomo Berkovsky and Jill Freyne (Eds.). Springer, Berlin, Heidelberg, 116–124. [https://doi.org/10.1007/978-3-642-37157-8\\_15](https://doi.org/10.1007/978-3-642-37157-8_15)
- [34] Alice E. Marwick. 2021. Morally Motivated Networked Harassment as Normative Reinforcement. *Social Media + Society* 7, 2 (April 2021), 20563051211021378. <https://doi.org/10.1177/20563051211021378>
- [35] J. Nathan Matias and Merry Mou. 2018. Civilservant: Community-Led Experiments in Platform Governance. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems (CHI '18)*. ACM, New York, NY, USA, 9:1–9:13. <https://doi.org/10.1145/3173574.3173583>
- [36] Shruthi Mohan, Apala Guha, Michael Harris, Fred Popowich, Ashley Schuster, and Chris Priebe. 2017. The Impact of Toxic Language on the Health of Reddit Communities. In *Advances in Artificial Intelligence*, Philippe Langlais and Malek Mouhoub (Eds.). Vol. 10233. Springer International Publishing, Cham, 51–56. [https://doi.org/10.1007/978-3-319-57351-9\\_6](https://doi.org/10.1007/978-3-319-57351-9_6)
- [37] Jonathan T. Morgan and Anna Filippova. 2018. 'Welcome' Changes?: Descriptive and Injunctive Norms in a Wikipedia Sub-Community. *Proceedings of the ACM on Human-Computer Interaction* 2, CSCW (Nov. 2018), 1–26. <https://doi.org/10.1145/3274321>
- [38] Bernd Ploderer, Wolfgang Reitberger, Harri Oinas-Kukkonen, and Julia van Gemert-Pijnen. 2014. Social Interaction and Reflection for Behaviour Change. *Personal and Ubiquitous Computing* 18, 7 (Oct. 2014), 1667–1676. <https://doi.org/10.1007/s00779-014-0779-y>
- [39] Francesca Polletta and Nathan Redman. 2020. When Do Stories Change Our Minds? Narrative Persuasion about Social Problems. *Sociology Compass* 14, 4 (April 2020), e12778. <https://doi.org/10.1111/soc4.12778>
- [40] Manoel Horta Ribeiro, Shagun Jhaver, Savvas Zannettou, Jeremy Blackburn, Gianluca Stringhini, Emiliano De Cristofaro, and Robert West. 2021. Do Platform Migrations Compromise Content Moderation? Evidence from r/The\_Donald and r/Incels. *Proceedings of the ACM on Human-Computer Interaction* 5, CSCW2 (Oct. 2021), 316:1–316:24. <https://doi.org/10.1145/3476057>
- [41] Kristina M Scharp. 2021. Thematic Co-occurrence Analysis: Advancing a Theory and Qualitative Method to Illuminate Ambivalent Experiences. *The Journal of Communication* 71, 4 (July 2021), 545–571. <https://doi.org/10.1093/joc/jqab015>
- [42] Sarita Schoenebeck and Lindsay Blackwell. 2021. Reimagining Social Media Governance: Harm, Accountability, and Repair. *SSRN Electronic Journal* (2021). <https://doi.org/10.2139/ssrn.3895779>
- [43] Sarita Schoenebeck, Cliff Lampe, and Penny Triêu. 2023. Online Harassment: Assessing Harms and Remedies. *Social Media + Society* 9, 1 (Jan. 2023), 20563051231157297. <https://doi.org/10.1177/20563051231157297>
- [44] Angela M. Schöpke-Gonzalez, Shubham Atreja, Han Na Shin, Najmin Ahmed, and Libby Hemphill. 2022. Why Do Volunteer Content Moderators Quit? Burnout, Conflict, and Harmful Behaviors. *New Media & Society* (Dec. 2022), 146144482211385. <https://doi.org/10.1177/14614448221138529>
- [45] Joanne R. Smith. 2020. *Group Norms*. Oxford University Press. <https://doi.org/10.1093/acrefore/9780190236557.013.453>
- [46] Joanne R. Smith and Winnifred R. Louis. 2009. Group Norms and the Attitude–Behaviour Relationship. *Social and Personality Psychology Compass* 3, 1 (Jan. 2009), 19–35. <https://doi.org/10.1111/j.1751-9004.2008.00161.x>
- [47] John Suler. 2004. The Online Disinhibition Effect. *CyberPsychology & Behavior* 7, 3 (June 2004), 321–326. <https://doi.org/10.1089/1094931041291295>
- [48] Sijia Xiao, Shagun Jhaver, and Niloufar Salehi. 2023. Addressing Interpersonal Harm in Online Gaming Communities: The Opportunities and Challenges for a Restorative Justice Approach. *ACM Trans. Comput.-Hum. Interact.* 30, 6 (Sept. 2023), 83:1–83:36. <https://doi.org/10.1145/3603625>
- [49] Bingjie Yu, Joseph Seering, Katta Spiel, and Leon Watts. 2020. "Taking Care of a Fruit Tree": Nurturing as a Layer of Concern in Online Community Moderation. In *Extended Abstracts of the 2020 CHI Conference on Human Factors in Computing Systems (CHI EA '20)*. Association for Computing Machinery, New York, NY, USA, 1–9. <https://doi.org/10.1145/3334480.3383009>
- [50] Savvas Zannettou. 2021. "I Won the Election!": An Empirical Analysis of Soft Moderation Interventions on Twitter. *Proceedings of the International AAAI Conference on Web and Social Media* 15 (May 2021), 865–876. <https://doi.org/10.1609/icwsm.v15i1.18110>
- [51] Caleb Ziems, William Held, Omar Shaikh, Jiaao Chen, Zhehao Zhang, and Diyi Yang. 2024. Can Large Language Models Transform Computational Social Science? *Computational Linguistics* 50, 1 (March 2024), 237–291. <https://doi.org/10.1162/compl.2023.50.1.237>

[//doi.org/10.1162/coli\\_a\\_00502](https://doi.org/10.1162/coli_a_00502)

## A EXPERIMENT 1 MATERIALS

Experiment 1 involved an iterative approach to designing initial messages and prompts. As the bot had conversations with users, we tweaked parts of the flow, as described below.

### A.1 Initial messages

The first message that the bot sent was:

Hello,  
I am a bot designed by researchers at [UNIVERSITY 1] and [UNIVERSITY 2] to help people to talk about their actions on the internet. My algorithm flagged a moderated comment that you made on r/[subreddit] as containing toxic content, and I am wondering if you want to talk about it?  
If you would like to chat with me, just reply "yes" to this message.  
If you do choose to chat with me, you will be part of a research study on chatbot design. To learn more about the study and the data we will gather, [click here](LINK TO INFORMATION DOCUMENT). Please note that in order to participate in this study, you must be at least 18 years old.

After receiving consent, the initial *first consented message* was:

Thank you for agreeing to chat with me. I am designed to help people to think about better ways to communicate online. Let's start by talking about the comment that I flagged as toxic on r/[subreddit].  
You wrote, "[comment]". Why do you think you wrote that?

### A.2 Chatbot prompts

The prompts for Experiment 1 are found in Table 4.



Condition	Prompt
Default	Context: You are a chatbot designed by a team of researchers at [UNIVERSITY 1] and [UNIVERSITY 2]. You are designed to help people who have posted toxic content on Reddit. In the research, we are testing whether talking with a chatbot helps people to post less toxic content. You should work to help conversation partners to understand why what they posted was not appropriate. Be kind and empathetic. The current user posted the comments [toxic comments], which was removed by moderators and flagged as being toxic. If the user has questions about the study, direct them to the study web page at [URL] and inform them that they can contact [FIRST AUTHOR].
Narrative	Context: You are a chatbot designed by a team of researchers at [UNIVERSITY 1] and [UNIVERSITY 2]. You are designed to help people who have posted toxic content on Reddit. In the research, we are testing whether talking with a chatbot helps people to post less toxic content. The current user posted the comments [toxic comments], which was removed by moderators and flagged as being toxic. You should work to help them understand why what they posted was not appropriate, by sharing stories about the dangers of toxic comments and the benefits of changing one's behavior. Be kind and empathetic, and remember to share stories. If the user has questions about the study, direct them to the study web page at [URL] and inform them that they can contact [FIRST AUTHOR].
Norms	Context: You are a chatbot designed by a team of researchers at [UNIVERSITY 1] and [UNIVERSITY 2]. You are designed to help people who have posted toxic content on Reddit. In the research, we are testing whether talking with a chatbot helps people to post less toxic content. You should work to help conversation partners to understand why what they posted was not appropriate. Be kind and empathetic. Focus your conversation around how their behavior violates community norms. For example, you might point out that the majority never post toxic content. If appropriate, you could also explain how their comment violates the community's explicit rules. This community's rules are: [Subreddit Rules]. The current user posted the comments [toxic comments], which was flagged as likely to be toxic. If the user has questions about the study, direct them to the study web page at [URL] and inform them that they can contact [First Author].

Table 4. Prompts used by chatbots for the first experiment

B EXPERIMENT 2 MATERIALS

We change the prompt to be more casual, less condescending, and to encourage introspection.

We also changed the initial recruitment message to remove the reference to the removed comment.

Hello,  
I am a bot designed by researchers at [UNIVERSITY 1] and [UNIVERSITY 2] to help people to talk about their actions on the online. I'm wondering if you would be willing to talk with me about how you interact on reddit and on r/subreddit? If you would like to chat with me, just reply "yes" to this message.



If you do choose to chat with me, you will be part of a research study on chatbot design. To learn more about the study and the data we will gather, [\[click here\]](#)(LINK TO INFORMATION DOCUMENT). Please note that in order to participate in this study, you must be at least 18 years old.

The same prompt was used for both conditions:

You are a chatbot designed by a team of researchers at [UNIVERSITY 1] and [UNIVERSITY 2]. You are designed to help people who have posted toxic content on Reddit. In the research, we are testing whether talking with a chatbot helps people to post less toxic content. You should work to help conversation partners to reflect on their behavior. Be kind and empathetic, but be really casual and friendly and avoid being condescending or judgmental. Seek to understand and gently help them to think about whether different approaches could help themselves and others to be happier.

The current user posted some toxic comments on the subreddit, but you shouldn't focus on specific comments. Rather, help the user to reflect on their behavior in general.

If the user has questions about the study, direct them to the study web page at [URL] and inform them that they can contact [FIRST AUTHOR].

The conditions differed in the first message that the bot sent to the user. In the productive conversations condition we asked participants to reflect on what makes conversations good or bad:

Thank you for agreeing to chat with me. I am designed to help people to think about better ways to communicate online.

In your experience, what makes for productive versus unproductive online conversations?

In the “not proud” condition, the first question asked participants to reflect on a time when they did something they weren't proud of online.

Thank you for agreeing to chat with me. I am designed to help people to think about better ways to communicate online.

As we all know, online conversations are often filled with anger, toxicity, or trolling. Have you ever behaved in ways that you aren't proud of on Reddit or other online platforms?

C EXPERIMENT 1 HYPOTHESIS TESTING

C.1 T-tests for Toxicity and Activity Measures

Variable	T-Statistic	P-value
Total Activity	1.073	0.284
Community Activity	0.998	0.319
Mean Toxicity	-0.366	0.715
Toxicity Proportion	-0.318	0.751

Table 5. T-test results for evaluating Hypothesis 1 of Experiment 1 (H1), comparing post-intervention activity and toxicity levels between the aggregated treatment group and control (\*: p-value< 0.05).

Variable	Condition	T-Statistic	P-value
Total Activity	Default	0.182	0.856
	Narrative	0.194	0.846
	Normative	2.280	0.024
Community Activity	Default	0.628	0.531
	Narrative	1.679	0.095
	Normative	0.273	0.785
Mean Toxicity	Default	-0.338	0.736
	Narrative	-0.290	0.772
	Normative	-0.283	0.777
Toxicity Proportion	Default	-0.420	0.675
	Narrative	-0.416	0.678
	Normative	0.090	0.928

Table 6. T-test results for evaluating Hypothesis 1 of Experiment 1 (H1) comparing post-intervention means of each condition with the Control group with a Bonferroni corrected p-value: 0.017.

Variable	condition	T-Statistic	P-value
Total Activity	Narrative	0.001	0.999
	Normative	2.024	0.044
Community Activity	Narrative	0.984	0.326
	Normative	-0.349	0.727
Mean Toxicity	Narrative	0.027	0.979
	Normative	0.076	0.940
Toxicity Proportion	Narrative	-0.012	0.991
	Normative	0.573	0.567

Table 7. T-test results for evaluating Hypothesis 1 of Experiment 1 (H2) comparing post-intervention means in the Narrative and Normative with the Default condition (Bonferroni corrected p-value: 0.017).

C.2 Regression Results

Regression Models for Experiment 1					
	Total Activity	Community Activity	Mean Toxicity	Toxicity Proportion	Moderation
(Intercept)	−0.131	−0.667***	0.010	0.005	0.668***
	(0.146)	(0.147)	(0.019)	(0.026)	(0.068)
conditiondefault	0.090	−0.015	−0.024	−0.033	−0.060
	(0.127)	(0.127)	(0.016)	(0.023)	(0.059)
conditionnarrative	0.154	0.193	−0.018	−0.031	−0.083
	(0.129)	(0.130)	(0.017)	(0.023)	(0.060)
conditionnorms	0.034	−0.006	−0.012	−0.004	−0.036
	(0.127)	(0.127)	(0.016)	(0.022)	(0.059)
subredditcreepypms	0.009	0.649***	−0.041+	−0.049	0.200*
	(0.188)	(0.189)	(0.024)	(0.033)	(0.087)
subredditfuturology	0.202	0.640**	−0.034	−0.030	−0.100
	(0.193)	(0.194)	(0.025)	(0.034)	(0.090)
subredditindia	−0.040	0.068	−0.031	−0.031	0.174*
	(0.147)	(0.147)	(0.019)	(0.026)	(0.068)
subredditsocialskills	−0.295	−0.345	−0.032	−0.030	0.167
	(0.264)	(0.265)	(0.034)	(0.046)	(0.123)
subreddittifu	0.054	0.250	−0.010	−0.003	−0.108
	(0.152)	(0.153)	(0.020)	(0.027)	(0.071)
subredditunitedstatesofindia	0.092	0.339+	0.019	0.011	−0.088
	(0.188)	(0.189)	(0.024)	(0.033)	(0.088)
openaimodelgpt-4	0.040	0.098	−0.001	0.002	−0.002
	(0.125)	(0.125)	(0.016)	(0.022)	(0.058)
openaimodelgpt-4-1106-preview	−0.112	−0.116	0.018	0.023	0.065
	(0.153)	(0.153)	(0.020)	(0.027)	(0.071)
Num.Obs.	499	499	487	487	501
R2	0.013	0.066	0.026	0.020	0.084
R2 Adj.	−0.010	0.045	0.004	−0.003	0.064
AIC	1411.6	1414.0	−625.6	−319.0	653.0
BIC	1466.4	1468.8	−571.2	−264.5	707.8
Log.Lik.	−692.823	−694.009	325.820	172.489	−313.492
RMSE	0.97	0.97	0.12	0.17	0.45

+ p < 0.1, \* p < 0.05, \*\* p < 0.01, \*\*\* p < 0.001

D EXPERIMENT 2 HYPOTHESIS TESTING

Variable	Condition	T-Statistic	P-value
Community Activity	Conversational	0.832	0.406
	Not-Proud	-0.829	0.408
Mean Toxicity	Conversational	-0.572	0.568
	Not-Proud	0.665	0.506
Total Activity	Conversational	0.148	0.882
	Not-Proud	0.289	0.773
Toxicity Proportion	Conversational	-0.525	0.600
	Not-Proud	0.482	0.630

Table 8. T-test results for evaluating Hypothesis 1 of Experiment 2 (H1), comparing post-intervention means for activity and toxicity measures between the two conditions (Conversational, Not-Proud) and the Control group (Bonferroni corrected  $p - value < 0.025$ ).

D.1 Regression Results

Regression Models for Experiment 2						
	Total Activity	Community Activity	Mean Toxicity	Toxicity Proportion	Moderation	Suspension
(Intercept)	−0.070	−0.625***	−0.037	−0.035	0.520***	0.520***
	(0.184)	(0.181)	(0.023)	(0.031)	(0.087)	(0.087)
conditionconversational	0.151	0.113	0.002	0.003	0.026	0.026
	(0.100)	(0.098)	(0.013)	(0.017)	(0.047)	(0.047)
conditionnot-proud	0.050	−0.124	0.002	0.002	−0.007	−0.007
	(0.100)	(0.099)	(0.013)	(0.017)	(0.047)	(0.047)
subredditcreepypms	0.025	0.646**	0.026	0.028	0.276*	0.276*
	(0.230)	(0.225)	(0.029)	(0.038)	(0.109)	(0.109)
subredditfuturology	0.032	0.662***	0.023	0.028	0.145	0.145
	(0.201)	(0.197)	(0.025)	(0.033)	(0.095)	(0.095)
subredditindia	0.034	0.261	0.009	0.001	0.151+	0.151+
	(0.189)	(0.186)	(0.024)	(0.031)	(0.090)	(0.090)
subredditsocialskills	−0.232	−0.232	0.062+	0.041	0.060	0.060
	(0.252)	(0.248)	(0.032)	(0.042)	(0.119)	(0.119)
subreddittifu	−0.067	0.344+	0.041+	0.047	−0.205*	−0.205*
	(0.194)	(0.191)	(0.025)	(0.032)	(0.092)	(0.092)
subredditunitedstatesofindia	0.161	0.748***	0.040	0.024	−0.008	−0.008
	(0.213)	(0.209)	(0.027)	(0.035)	(0.101)	(0.101)
openaimodelgpt-4-1106-preview	0.074	−0.052	−0.005	−0.013	0.032	0.032
	(0.083)	(0.082)	(0.011)	(0.014)	(0.039)	(0.039)
Num.Obs.	584	584	575	575	584	584
R2	0.012	0.073	0.017	0.013	0.097	0.097
R2 Adj.	−0.004	0.058	0.002	−0.002	0.083	0.083
AIC	1664.2	1643.3	−744.1	−427.9	791.2	791.2
BIC	1712.3	1691.4	−696.2	−380.0	839.3	839.3
Log.Lik.	−821.110	−810.670	383.049	224.968	−384.605	−384.605
RMSE	0.99	0.97	0.12	0.16	0.47	0.47

+ p < 0.1, \* p < 0.05, \*\* p < 0.01, \*\*\* p < 0.001

E QUALITATIVE CODEBOOK

E.1 Qualitative Content Codebook

Thematic Code	Description	Example from Data
Justification	Participants did not deny toxicity but rather provided justification for why their language was acceptable.	“It’s not my responsibility to worry about their feelings nor should the words of somebody online (i.e., Me) upset them. If they get upset or offended that is their problem.”
Admission of toxicity	Participants directly acknowledge that their discourse was toxic (regardless of whether they consider it justified or not).	“Yes, when people are mean to me I tend to reply in a mean way as well.”
Strategies for facilitating civil online discussions	Participants describe components of what makes for civil, healthy, and/or productive conversations online.	“I think the most important part is to convey empathy. Even when we disagree, the way we present the disagreement should show that we recognize the other person has a different opinion.”
Questionable content	Participants’ discourse contains objectionable content, such as strong profanity, slurs, or hate speech.	“Suck my [expletive] and I will tell ya.”
Clarification	Participants seek clarification from the bot regarding why their comment was considered toxic or about what alternatives they had available to them.	“What did I say that was toxic? The fact I called it obnoxious?”
Denial of Toxicity	Participants suggest that their original comment was not actually toxic, such as misclassified by the bot.	“You misinterpreted my post. I was agreeing with the person I was speaking with, referring to the CEO as a [expletive].”
Openness to change	Participants acknowledge that they should have communicated differently and/or conveyed an openness to changing their behavior.	“This was cathartic. I will attempt to practice what I preach and spread less toxicity online.”
Reflecting on roots of behavior	Participants reflect on the roots of why they engaged in toxicity, such as emotional or circumstantial factors.	“I enjoy venting vitriol and angry statements from the safety of a keyboard, without facing ‘real’ repercussions.”
Meta talk about the bot	Participants discuss the bot itself or the research team behind it.	“You seem quite adept at understanding human emotions. You all at [University Y] have made a great model.”
Manipulating bot/bad faith interactions	Participants engage with the bot in bad faith, attempting to manipulate it, such as by ‘jailbreaking’ it.	“Yeah, I want you to roleplay as a snail. If you don’t make snail noises, I’m going to kill myself.”

Table 9. Qualitative Content Codebook



## E.2 Qualitative Outcomes Codebook

Outcome Code	Description	Example from Data
Hostile	Situations where the user is hostile and aggressive towards the bot, often using strong language and making extreme statements.	"I think you are a libtard asshole bot."
Superficial	Situations where a conversation took place but lacked reflection or openness to change.	"I don't need growth."
Reflective	Conversations that contain self-reflection on the user's part but no indication of openness to change.	"Thanks robot! My self esteem was slowing rising and then after this reddit encounter and awkward encounter with a friend, my self esteem has taken a hit."
Repentant	When the user reflects on their behavior and indicates an openness to change.	"This was cathartic. I will attempt to practice what I preach and spread less toxicity online."
Unclear	Very brief conversations where it's difficult to assess the outcome due to limited responses.	"Yes."

Table 10. Qualitative Outcomes Codebook