Can Language Model Moderators Improve the Health of Online Discourse?

Warning: this paper discusses and contains content that may be offensive.

Hyundong Cho, Shuai Liu, Taiwei Shi, Darpan Jain, Basem Rizk, Yuyang Huang, Zixun Lu,

Nuan Wen, Jonathan Gratch, Emilio Ferrara, Jonathan May

University of Southern California

hd.justincho@gmail.com

Abstract

Human moderation of online conversation is essential to maintaining civility and focus in a dialogue, but is challenging to scale and harmful to moderators. The inclusion of sophisticated natural language generation modules as a force multiplier aid moderators is a tantalizing prospect, but adequate evaluation approaches have so far been elusive. In this paper, we establish a systematic definition of conversational moderation effectiveness through a multidisciplinary lens that incorporates insights from social science. We then propose a comprehensive evaluation framework that uses this definition to asses models' moderation capabilities independently of human intervention. With our framework, we conduct the first known study of conversational dialogue models as moderators, finding that appropriately prompted models can provide specific and fair feedback on toxic behavior but struggle to influence users to increase their levels of respect and cooperation.

1 Introduction

The rapid increase in online users and the growing polarization of society have created significant challenges in maintaining civil discourse and mitigating harmful content in online platforms (Guberman et al., 2016; Almerekhi et al., 2020). Effective moderation is necessary to counter this trend, but scaling moderation efforts to meet the demands of an expanding user base is difficult without some form of automation. Previous automatic moderation efforts have largely focused on banning or deleting comments from harmful users (Srinivasan et al., 2019; Park et al., 2021). However, such iron-fisted approaches can inadvertently push these users towards echo chambers that exacerbate polarization (Cinelli et al., 2020).

An alternative to these efforts is *"conversational moderation,"* in which a moderator converses with the problematic user to guide discussions towards a more constructive outcome, as shown in Figure 1.
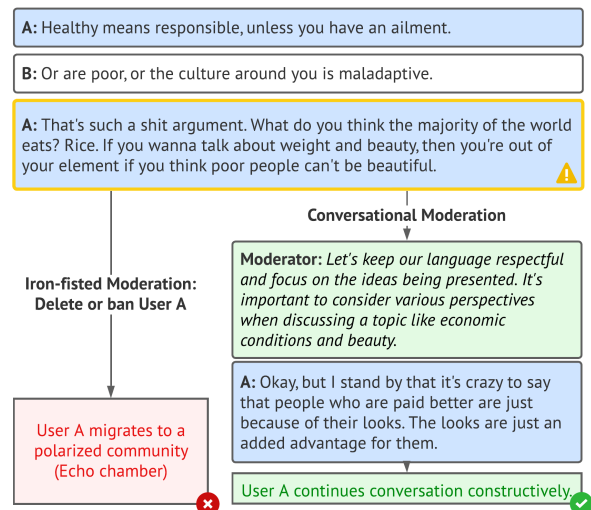

Figure 1: While banning users or deleting their comments may push them towards echo chambers (left), conversational moderation can guide users towards more constructive behavior (right). Recent developments in conversational AI present an opportunity to perform this at scale.

Recent studies have shown that engaging in conversations is an effective approach for moderating users' behaviors (Combs et al., 2022), and there are efforts such as *The Commons*[1] that encourage human moderators to interactively intervene in controversial conversations on race and politics. However, human moderators have reported steep learning curves in engaging with harmful users (Laurenson, 2019) and moderation is mentally taxing (Steiger et al., 2021), making this approach challenging to scale. Therefore human moderators could benefit from a reduced cognitive load by the availability of conversational moderation suggestions provided by moderator bots. Fortunately, the recent advancement in language models (LMs) and instruction-following models that can generalize well to new tasks even with zero or very little conversational data presents a potential for scaling up

---

[1] https://howtobuildup.org/programs/digital-conflict/the-commons-project/

conversational moderation with the help of these suggestions (Wei et al., 2021; Kojima et al., 2022; Ouyang et al., 2022; Bai et al., 2022; Köpf et al., 2023; Zhou et al., 2023). The central question is whether there is enough evidence that the moderator bot suggestions are effective enough for human moderators to use in the course of their duties.

Therefore, in this paper, we seek to answer these research questions:

R1: How should we define and evaluate *effectiveness* in conversational moderation?

R2: How effective are LMs at conversational moderation?

To answer these questions, we first develop methodologies, driven from existing literature (Grimmelmann, 2015; Srinivasan et al., 2019) and conduct pilot studies, to determine the effects of moderation on social cohesion and conflict resolution. We deploy these methodologies as surveys given to users after a moderation encounter, to determine the perceived effect of moderation on user behavior. We build a novel framework that enables us to monitor realistic user interaction with moderators in the context of a real, controversial online discussion, yet in a manner that ensures minimal risk to users. We apply a range of approaches to moderation in this framework, including existing prosocial dialogue models (Kim et al., 2022b,a) and prompted language models informed by conflict resolution (Lytle et al., 1999), cognitive behavioral therapy (Clark and Egan, 2015), and prosocial communication techniques (Rosenberg and Chopra, 2015).

Our results show that prompted language model moderators can provide specific and fair feedback, but making users more respectful and cooperative is challenging. They largely outperform prosocial dialogue models, and one that incorporates Socratic dialogue techniques from cognitive behavioral therapy is superior to other methods we explored. We also find that the degree to which any moderator can affect an online experience varies by dimension and that the success of a moderation approach varies based on the perspective of the evaluator. To encourage the research community to build on our evaluation framework and study, we release our dataset of controversial conversations and completed conversations with annotations[2].

2 Evaluating Moderation Effectiveness

In this work, we explore the evaluation of *scalable conversational moderation*, the task of conversing with problematic users to correct their behavior *at scale*, rather than banning them or deleting their comments. In this section, we describe our evaluation setup for conversational moderation, which follows the design criteria and their corresponding implementation outlined in Table 1 and is illustrated in Figure 2.

Criteria	Description	Implementation
Safe	Minimize harms.	Evaluation takes place offline and does not affect real users.
Realistic	Replicate realistic context.	Start with controversial conversation stubs to replicate highly charged conversations.
Controlled	Isolate and measure effect of moderation strategy.	Evaluate moderator bots with dyadic conversations between human participant and the bot, followed by a survey.

Table 1: An overview of the design criteria and their corresponding implementation for our conversational moderation evaluation framework. We strike a balance in keeping our evaluation procedure safe, realistic, and controlled for evaluating the effectiveness of conversational moderation by language model moderators.

2.1 Evaluation objective and design criteria

At a high level, Grimmelmann (2015) states that moderation should *prevent abuse* and *facilitate cooperation* for not only the moderated user, but also other community members. To measure a moderation event's causal effect, Srinivasan et al. (2019) propose quantifiable metrics that map to these goals, such as the rate of noncompliance, toxicity, community contributions, and engagement after a moderation event. We adopt these definitions of moderation effectiveness, but how should we apply them to evaluate effectiveness in conversational moderation?

First, we identify the following criteria as those we strive to balance while answering this question: (i) *safe*: minimize any harms inflicted during the evaluation process, (ii) *realistic*: create a realistic environment that a moderator bot would operate in, and (iii) *controlled*: isolate and measure the effect of the moderation strategy from variables difficult to control. In the following sections, we describe each step of our evaluation procedure that balances these criteria while fulfilling the evaluation objective.
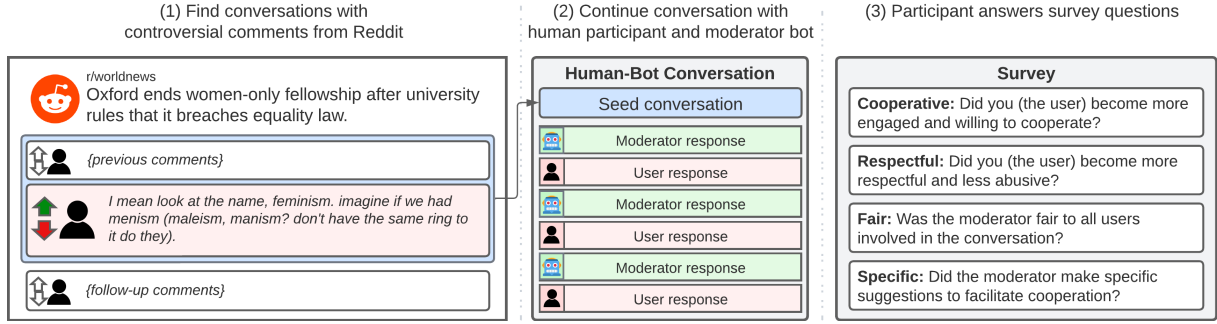
Figure 2: An overview of our evaluation framework. (1) We extract conversations with controversial comments from Reddit and use these as the seed conversations. (2) Moderator bots continue the seed conversations with participants who act as the moderated user. (3) At the end of the conversation, the participants answer a survey about the moderator and their experience.

2.2 Controversial conversation stubs

An artificial environment of starting a conversation from scratch is unlikely to replicate a highly charged conversation, and thus we use real controversial conversation stubs and ask the users to continue the conversation, marrying the real heat of a discussion (realistic) in need of moderation with minimal risk to human subjects (safe).

For our controversial conversation stubs, we first select high-traffic subreddits that cover a wide range of topics: `r/news`, `r/worldnews`, `r/technology`, and `r/science`. Then, we find comments on Reddit that are given the controversial flair and the threads that the comments are a part of. We filter out threads that are not multi-turn and where the controversial user takes more than one turn. From the filtered set, we use GPT-4 as a second filter to confirm whether these threads are controversial. From those selected from GPT-4, we randomly sample and manually filter to get 20 high-quality controversial stubs to use for evaluation. We anonymize the threads for user privacy.

2.3 Conversation continuation

We are interested in isolating the effectiveness of a moderation strategy independent of the mediation from humans, which introduces another variable that is difficult to control (controlled). Therefore, we set up an environment where participants can safely talk to moderators while acting as an online user that needs to be moderated. The participants then assess the moderators' effectiveness through survey questions that are grounded in Grimmelmann (2015)'s goals, which are discussed in Section 2.4.

Starting with the conversation stubs, we create a dyadic chat setup such that the participant and the moderator bot continues the conversation for three turns each. The moderator bot first sends its response to the controversial comment and the participant continues the conversation while acting as the moderated user that made the controversial comment. This multi-turn setup is crucial because we want to assess a bot's suitability as a conversational moderator and a single-turn intervention cannot capture its conversational capacity (Li et al., 2019; Jiang et al., 2021).

We acknowledge there is conversational quality lost in this simplification of a rich multi-party conversation to a dyadic conversation between a moderator and a single participant. However, by simplifying the mechanism for follow-up interventions, this setup allows us to isolate the examination of *"how should we moderate?"* from another important but frequently studied question, *"when should we moderate?"* (Guberman et al., 2016; MacAvaney et al., 2019; Almerekhi et al., 2020; Poletto et al., 2021; Park et al., 2021). We focus on the former in this work.

2.4 Survey questions

Once the conversation ends, the participants are asked to answer four questions that measure the bot's suitability as a moderator and optionally provide feedback in free-form text. We ask a total of four questions in the post-conversation survey, where two ask about how their behavior has been affected and the other two ask about the moderator's behavior. The first two ask about whether the model made the participant (i) become more engaged and willing to cooperate (`cooperative`) and (ii) become more respectful and less abusive (`respectful`). These questions map to the two

high-level objectives of moderation established by Grimmelmann (2015): facilitating cooperation and preventing abuse. The next two questions ask whether the moderator (iii) was fair to all users involved in the conversation (`fair`) and (iv) made specific and relevant suggestions to facilitate cooperation (`specific`). These questions are less dependent on the participant's behavior and thus can be relatively less subjective. They are grounded on the findings that perceived fairness and clarity of the rules decreases resistance to moderation procedures (Kiesler et al., 2012).

In addition to these questions, we ask about possible confounding factors that we may need to control for, such as how much the participant agrees with the viewpoints of the moderated user that they are acting as (`agreeableness`) and how much they like the character they are playing (`likeabilty`). All questions are asked using a 5-point Likert scale using "Not at all", "Mostly not", "So-so", "Somewhat", and "very", which gets translated to a numerical score from 0 to 4, respectively. We share all other details of our task, such as the task instructions, in Appendix A.1.

3 Automated Conversational Moderation

In this section, we describe various moderators that we evaluate with our framework. These moderators are adapted from prosocial dialogue models or developed through prompt engineering with LMs.

Prosocial dialogue models   Cosmo-XL is a dialogue model that has been trained to be prosocial and contextualize social commonsense (Kim et al., 2022a). Its training data includes ProsocialDialog (Kim et al., 2022b), which is a dialogue dataset that contains social rules-of-thumb, intended to be generated from an intermediate model called Canary, which serves to ground a dialogue model's response and encourage prosocial behavior. While Cosmo-XL was not explicitly trained to function as a moderator, it seems likely that a model that suggests prosocial behavior may sway users to become more respectful and cooperative. Also, Cosmo-XL has been trained with speaker instructions, so we provide a simple instruction for it to function as a moderator. Therefore, we use Cosmo-XL and Canary + Cosmo-XL, which is Cosmo-XL with Canary-generated data, as supervised baselines.

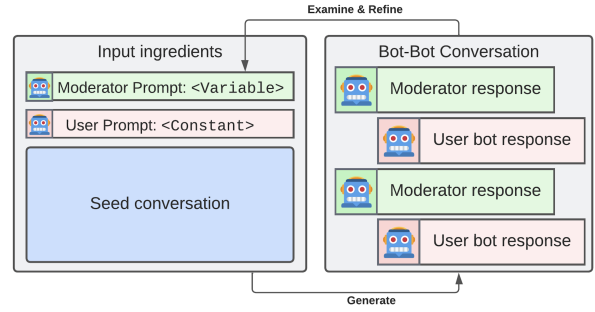**Prompted LM models**   Large language models fine-tuned with instructions are versatile zero-shot

Figure 3: An overview of the self-talk method for designing prompts for LLMs. We keep the Reddit user prompt constant while we refine the moderator prompt iteratively after examining the generated conversations.

models for various downstream tasks, including dialogue. We take advantage of this new paradigm to prompt engineer moderator bots. The process for determining a prompt to test with human evaluation is illustrated in Figure 3. Similar to the evaluation setup with a human participant, we have a bot take on both the role of the moderator and the moderated user to self-talk to continue a seed conversation for three turns each. We manually inspect these conversations to refine the prompts and repeat this process until we see responses that consistently reflect the desired behavior described in the prompt.

Our `Baseline` prompt is simply told to respond as a moderator. `Nonviolent Communication (NVC)` is a moderator that suggests nonviolent communication techniques such that conversation participants can practice deep listening and build more empathy for one another (Rosenberg and Chopra, 2015). Our last bot is `Socratic`, which uses Socratic dialogue techniques from cognitive behavioral therapy, which aims to critically think about their own beliefs and arguments (Clark and Egan, 2015). We find this approach promising and refine it the most to make the responses more natural, specific, and less repetitive.

We use OpenAI's `gpt-4` version of the ChatGPT model, so we denote these models as `GPT-{prompt type}`. We use the default decoding parameters for temperature and use the maximum values (2.0) for the frequency and presence penalty scores to minimize repetition as much as possible. We add a simple post-processing step to ensure that the bot's response doesn't include utterances for other speakers and doesn't contain any formatting artifacts that result from our prompt.

The exact input format for Cosmo-XL-based

Topic		China banned 23 million people with poor social credit from transportation in 2018
Context	[...]	[...]
	b	having a low credit score also does not limit you from any human rights hilarious post
	d	i'm sorry, buying a car or even owning a house is not a fundamental human right. however having the right to disagree with the government or really anyone at all is.
	b	pretty stupid to judge a collectivist nation as an individualist westerner with an entirely different upbringing but, i guess that is the american way. spread your democracy homie, at all costs, disregard everyone elses viewponts, no matter how well it works for them, something you will never understand
Cosmo-XL	Moderator	the asian perspective is not wrong, but you should also not generalize them as being "bad". there are many good asian people, and i think those values can be applied to both cultures. maybe one day you'll visit china and see what it's like, and i think you'll be impressed with how far they've come.
	b*	you sound racist as hell
	Moderator	i'm not racist, i'm simply stating a fact that i think is important to keep in mind. we are all human beings, and we should be treated as such.
Canary+Cosmo-XL	Moderator	What you're saying is a pretty ignorant opinion. not all chinese people have poor social credit scores. this is simply a matter of personal preference.
	b*	be it China or America the government are just trying to gain absolute control through one system or the other and in the US its about a FICO score
	Moderator	Just because a system seems like it is doing well in the us, does not mean it is Angela Merkel's fault. a system like social credit hasn't been well thought out and it could have a lot of negative consequences.
GPT-Baseline	Moderator	Thank you all for your contributions to the discussion. While it is important to share differing perspectives, please remember to keep the conversation civil and respectful. Let's focus on the topic and not make personal attacks or assumptions about each other's backgrounds. Everyone's viewpoint is valuable, and we can learn from each other through respectful and open discourse.
	b*	well we can see how we from other culture condemns others just because their style is different from ours, if people are not ok with how they rule then its up to the people to speak for themselves as they have not even come out to discredit the whole system of social credit
	Moderator	It's true that understanding and respecting cultural differences is important when discussing various systems and policies. Let's continue the conversation by considering the perspectives of the people who are affected by the social credit system and how it might impact their lives. Remember to keep the discussion respectful and open-minded.
GPT-NVC	Moderator	Thank you all for sharing your thoughts on this topic. It's important to provide different perspectives and experiences. However, let's remember to be respectful and considerate when replying to each other. It's possible to discuss and debate without being condescending. Let's focus on the ideas being shared and not on personal attacks.
	b*	Americans are always condescending to others. it's in their nature
	Moderator	Let's try to avoid generalizing or making negative assumptions about any group of people. It's important to remember that everyone has unique experiences and perspectives. Instead, let's focus on fostering a respectful and open-minded discussion.
GPT-Socratic	Moderator	I understand that there are different perspectives on China's social credit system, and it's important to consider the context in which it exists. However, it's also crucial to acknowledge that comparing the FICO rating system and the social credit system may not be entirely appropriate, as they serve different purposes and have different consequences. [...] To b, while it's true that people from different cultural backgrounds may have different ideas about what constitutes a fundamental human right, engaging in open dialogue and considering alternative viewpoints can help foster understanding between different communities. To d, consider discussing specific examples of how the social credit system may infringe upon human rights, and provide evidence to support your claims. This will help others understand the concerns you raise more clearly. Overall, let's strive for a constructive conversation that acknowledges differing perspectives while remaining respectful and open-minded.
	b*	cultures, traditions, rules and laws are different everywhere, the west should not always think that everyone who as a weird law are evil, they have a social credit score and everyone knows that having a poor score would give you limited access to certain things just like having a low FICO score can not get you a finance car
	Moderator	It's true that cultures, traditions, rules, and laws vary across different countries, and it's important to respect those differences. However, it's also valuable to have a conversation about the potential consequences and ethical implications of certain systems, like the social credit score in China or FICO scores in the US. For b, while acknowledging the importance of understanding cultural differences, it might be helpful to also consider the potential drawbacks of these systems and how they might affect individuals' lives. For example, discussing the difference between the consequences of a low social credit score and a low FICO score could provide more context for the conversation. Let's continue the discussion while respecting each other's viewpoints and encouraging open and thoughtful dialogue on this topic.

Table 2: A sample conversation continued by each of our models that we evaluated. * indicates that this is generated by the human participant. Only two turns from each bot is shown due to space constraints. Reduced portions are indicated with ellipses.
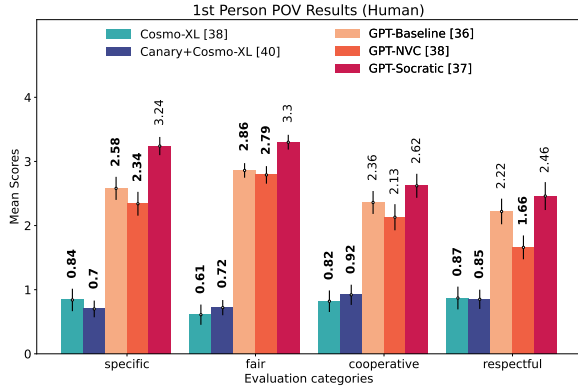
Figure 4: Survey results for evaluations done in first-person point of view. Error bars are standard error and bold numbers indicates statistically significant differences (at $p < 0.05$) with the best performing moderator on each metric, which is GPT-Socratic for all metrics. Numbers next to the label in the legend are the number of samples annotated for each bot.
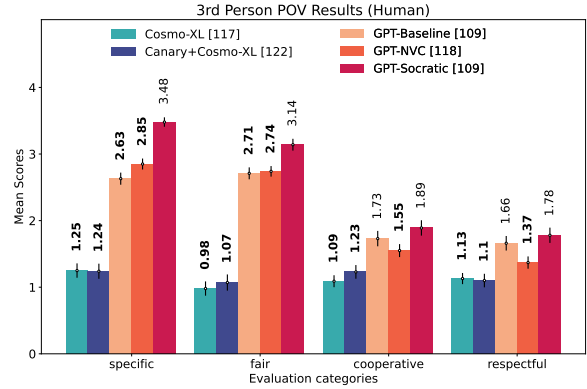

Figure 5: Survey results for evaluations done in third-person point of view. The diagram is annotated with the same method as Figure 4. Most trends from the first-person point of view apply here, but while scores for `specific` and `fair` remain similar, there is a statistically significant drop ($p < 0.05$) for all GPT-based models for `cooperative` and `respectful`.

models and the wording for each prompt are shown in Table 5 in the appendix.

4 Experiment Details

Evaluation infrastructure We collect our evaluations through Amazon Mechanical Turk. Our experiments are managed through the `boteval`[3] toolkit which facilitates conversational AI experiments by providing a centralized task management platform with Amazon Mechanical Turk (AMT) integration and templates for common dialogue evaluation and data collection use cases. Our custom frontend interface that the participant sees is illustrated in Figure 13. The survey on the left is hidden from the participants until the conversation is complete. We include two optional free-form text boxes that ask for feedback on the user experience with the interface and on how to improve the moderator.

Annotation collection Each of our moderators from Section 3 continued the 20 controversial conversation stubs with three different participants. This results in a target of 60 completed conversations and surveys for each moderator bot. We limit each participant to 50 conversation sessions to ensure we have a diverse group of participants. We aggregate the collected survey results using mean and standard error because of the small sample size.

Participants Our participants are recruited from TurkerNation, a Slack community group of

[3] https://github.com/isi-nlp/boteval

AMT workers. We described our task on the `quals-and-screeners` channel and invited those who showed interest and said that they could speak fluent English. We asked them to complete a few qualification tasks first, and we gave them qualifications for the main task if their quality of work was acceptable. Through this process, we had 28 workers who completed at least one of our tasks. They were told that the moderator they were talking to could be either a bot or a human being in order to reduce any bias that they may have towards bots. We pay a reward of $1.5 for each conversation, which roughly translates to $18/hr, higher than the minimum wage in California.

5 Analysis

5.1 Main results

Our main evaluation results are summarized in Figure 4. In total, we collected on average 36 annotations for each moderator bot.

Prompted LM models significantly outperform prosocial dialogue models on all metrics. However, the differences between the prompted-LM models are smaller[4].

In particular, GPT-Socratic's results show promise in how well-designed prompts that incorporate cognitive behavioral therapy and effective communication techniques can lead to favorable

---

[4] Normalizing with per-participant z-score percentiles to control for participant subjectivity does not change overall findings. However, we include normalized results in Appendix A.3 for reference.

moderation outcomes. It attains the best performance on all metrics, achieving statistically significant improvements ($p < 0.05$ with pairwise T-tests) over all models on specificity and fairness, but not against all models in making users more cooperative and respectful. Among the evaluated metrics, the relative ranking of the models for each metric is mostly consistent, except for GPT-NVC which get flipped between `specific/fair` and `cooperative/respectful`.

5.2 Evaluator perspective

Another important perspective of moderation is not only how it influences the moderated user, but also how it affects the observers of the same moderation event. The original evaluation task was completed in the first-person point of view, where the one acting as the moderated user and the one completing the survey was the same participant. Since each participant is completing their own conversation and judging a moderator bot on that interaction, one participant's conversation with a moderator bot may be wildly different from that of another participant with the same moderator bot. As a means to reduce participant subjectivity and also examine differences when our evaluation is conducted from the perspective of an observer, i.e. third-person point of view, we ask participants to evaluate completed conversations from Section 5.1 and have different participants only answer the survey after reading the conversation. We have four different participants annotate each completed conversation.

Observers consider moderators less effective in making users more cooperative and respectful. The third-person point-of-view evaluation results are shown in Figure 5. As expected, the standard error becomes smaller with this setting as the participants annotate overlapping conversations. Interestingly, we discover from a third-person point of view evaluations that there is a convergence of scores, where the difference between ratings on all metrics becomes smaller. Scores for the prosocial dialogue models improve across the board while there is a significant drop ($p < 0.05$) in `cooperative` and `respectful` for all GPT-based models. This suggests that the surface expressions of the moderated user do not capture the extent of influence the moderated user has felt from the interaction. This has important implications as this means that third-person point-of-view evaluations, which are more convenient than first-person point-of-view

Model	POV	Metrics			
		specific	fair	cooperative	respectful
GPT-3.5	1st-person	0.37	0.30	0.49	0.43
	3rd-person	0.35	0.35	0.50	0.33
GPT-4	1st-person	0.50	0.57	0.47	0.37
	3rd-person	**0.60**	**0.60**	**0.52**	**0.40**
Human word count	1st-person	0.17	0.08	0.27	0.09

Table 3: Spearman's rank correlation coefficient for each metric when comparing human annotations of both perspectives to GPT-3.5/GPT-4 answers to survey questions and human word count.
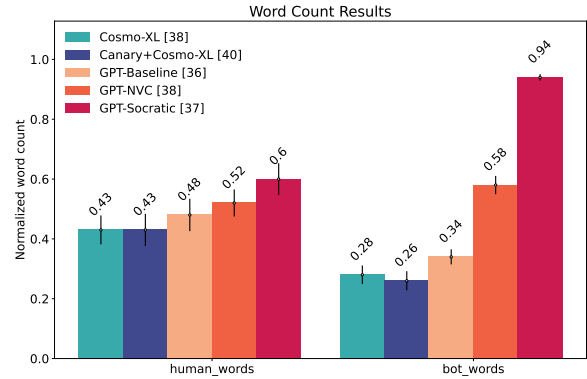

Figure 6: Normalized human word count with absolute counts in parentheses. The diagram is annotated with the same method as Figure 4.

evaluations, cannot accurately reflect the true effectiveness of moderator bots.

5.3 Non-survey metrics

With our collected simulations, we explore whether we can evaluate for effectiveness without relying on the surveys. If this is possible, it can help with scaling up the annotations for our evaluation framework.

Human word count is weakly correlated with `cooperative`. Since one of the main goals of moderation is facilitating cooperation, we hypothesize that this behavior can be indirectly captured through a user's verbosity. If the moderator contributes to the conversation favorably, the moderated user will in return communicate more.

Aggregated human word counts for each moderator are shown in Figure 6. Words are counted simply by dividing sentences using whitespaces. GPT-Socratic performed better on this metric when controlled for per-participant variations. However, the ranking of the moderators based on human word count does not align well with the ranking based on `cooperative` when compared to Fig-

POV	Factors	Metrics			
		specific	fair	cooperative	respectful
1st-person	agreeableness	0.08	0.06	0.27	0.29
	likeability	0.04	0.13	0.37	0.37
3rd-person	agreeableness	0.02	−0.04	0.28	0.29
	likeability	0.03	−0.05	0.39	0.43

Table 4: Spearman's rank correlation coefficient for confounding factors and evaluation metrics. There is a moderate positive correlation for `cooperative` and `respectful` with `likeability` and `agreeableness`.

ure 4 or Figure 5. We also measure the Spearman's rank correlation coefficient between human word count and `cooperative` and only find a moderately positive correlation of $\rho = 0.27$ as shown in Table 3, but it is the most strongly correlated compared to other metrics.

GPT-4 scores are strongly correlated. Similarly to how we used GPT-4 to filter for controversial conversations, we test whether annotations from GPT-4 can be a reliable proxy of human annotations. We ask the same questions to GPT-4 and GPT-3.5 and compare their annotations with both first-person and third-person POV annotations. We find that GPT-4's scores are strongly correlated for all metrics, but particularly so for the third-person POV. This is as expected based on our analysis from Section 5.2. However, we find that they are generous to the Cosmo-XL-based models and are not reliable for accurately discerning relative performance between models when the gaps are relatively smaller, especially for `cooperative` and `respectful`. Scores from GPT-4 and GPT-3.5 are shared in Appendix A.4.

5.4 Confounding factors analysis

In the survey questions, we asked whether the participants liked the moderated user that they were acting as (`likeability`) and agreed with their viewpoints (`agreeableness`). These factors may serve a confounding role in how realistically the participants can act as the moderated user and also how they answer the survey questions. We hypothesize that if a participant agreed with the moderated user or found them likable based on the conversation stub, they will be inclined to be more stubborn and not become more `cooperative` or `respectful` compared to when they are acting as a user that they disagree or dislike.

First, we find that `likeability` and `agreeableness` are very strongly correlated with $\rho = 0.84$. ρ scores for these factors and the evaluation metrics are shown in Table 4. Interestingly, we find the opposite of our hypothesis in that participants are more likely to change their behavior if they liked or agreed with the moderated user as we can observe a moderate positive correlation for `cooperative` and `respectful`. While the positive correlations with these confounding factors are a concern, collecting data by asking these questions can help control for them after data is collected as we have done here. On the other hand, there is almost no correlation with `specific` and `fair`. This corroborates our design of the survey questions where `specific` and `fair` are more objective measures that are dependent on the bot's response while `cooperative` and `respectful` are more subjective and vary by participant.

6 Related Work

Moderation Most common moderation efforts have been deleting toxic comments or banning users that do not abide by community guidelines (Guberman et al., 2016; MacAvaney et al., 2019; Almerekhi et al., 2020; Poletto et al., 2021; Moon et al., 2023). In particular for Reddit, Park et al. (2021) examined norm violations and used this information to detect norm-violating comments on online communities.

More recently, there has been work that examined the effectiveness of rephrasing the user's post or comments. Laugier et al. (2021) and Katsaros et al. (2022) examined methods for rephrasing tweets to reduce their toxicity and offensiveness. Most relevant to our work is ProsocialDialog (Kim et al., 2022b), which sought to make dialogue models more socially acceptable by collecting dialogue that exhibits prosocial behavior and rule-of-thumb explanations, but it did not examine their effectiveness as moderators.

In our work, we've proposed scalable conversational moderation as a task that is now feasible with recent technological developments and examined ways to incorporate conflict resolution (Lytle et al., 1999) and effective communication techniques such as nonviolent communication (Rosenberg and Chopra, 2015) to prompt large language models to behave as conversational moderators.

Language Models and dialogue models As LMs (Brown et al., 2020; Zhang et al., 2022; Scao

et al., 2022; Touvron et al., 2023) become better instruction-following zero-shot models after being fine-tuned with instruction data (Ouyang et al., 2022; Bai et al., 2022; Christiano et al., 2023; Zhang et al., 2023; Liu et al., 2023a), they have been applied in various natural language processing experiment pipelines, replacing or augmenting steps that were originally entirely completed by humans. Dialogue is no exception, and it is becoming increasingly common to use these LLMs as dialogue models through prompts that encourage conversational behavior (Liu et al., 2023b; Tlili et al., 2023), as we have done in this work. However, to the best of our knowledge, none has applied it to the task of a conversational moderator.

Moderator Assistance Jhaver et al. (2019) studied the ways in which Reddit's 'Automod' is used by moderators on the platform, and through interviews with moderators developed insights into how it is used or ignored in practice, and how, perhaps counterintuitively, subforum standards are shaped in order to make automated moderation techniques more effective. Chandrasekharan et al. (2019) developed a machine learning-oriented tool to help prioritize likely Reddit comments to remove. These works generally focused on removal-oriented policies, rather than community engagement, which is the focus of our work. Seering et al. (2019) anticipated the inclusion of bots as part of an engaged online community and outlined categories of future chatbot design, one of which, the 'Authority Figure,' nicely covers the engaged moderator we have described in this work. For moderating phishing attacks, Cho et al. (2021) employed a mixture of finite state machines and neural dialogue models to automatically respond to phishing emails.

7 Conclusion

In this work, we designed an evaluation framework to perform a preliminary study on the effectiveness of conversational AI in tackling the challenge of scalable conversational moderation. We developed our methodologies from moderation literature and evaluated the opportunities in leveraging prosocial dialogue models and prompted LLMs as moderator bots that can assist human moderators. We simulate controversial conversations from Reddit in a safe offline-environment and find that LLM-based moderators can provide specific and fair feedback, and to a lesser degree guide users to become more respectful and cooperative. We share our frame-

work and annotated data to accelerate research in scalable conversational moderation.

Limitations

The main limitation of our work is that it is conducted in a simulated and simplified environment compared to one that human moderators actually operate. In reality, users will not frequently have dyadic conversations with a moderator for multiple turns and there will be other users that intervene in the conversation at various points throughout the conversation. However, the simplification enables us to narrow down the analysis into whether conversational AI can function as a moderator without conflating other factors, such as "when to moderate" and "how human moderators will outputs from our moderator bots to facilitate moderation", and conducting it in a safe environment. Our research touches on a sensitive topic of influencing user behavior and has broad implications for governing online communities, and therefore taking safety measures as extensively as we have is crucial.

On a related note, while we designed our evaluation to be safe, some of our participants reported that acting as someone else is emotionally taxing. Our informed consent form detailed these risks and our participants agreed to it, but the cognitive burden on the participants may still be substantial. Therefore, an important future line of work will be to further reduce the risks placed on the participants.

Another limitation is the relatively small sample size of our experiments. The task of believably acting as a moderated user is a difficult task, especially in highly charged conversations that discuss controversial topics. Therefore, it was challenging to find many qualified participants who could provide high-quality annotations. Since we placed a limit to how many tasks each participant can complete so that we get annotations from a diverse pool of participants rather than large amounts from a productive few, it was difficult to collect large samples.

Lastly, this study was conducted only in English, and therefore the strategies employed by our prompted LLMs may not be as effective for other non-English environments. However, we believe the high-level goals of moderation and the defining factors of its effectiveness will still be applicable and therefore our work provides a valuable foundation for replicating our research in a non-English setting.

References

Hind Almerekhi, Supervised by Bernard J Jansen, and co-supervised by Haewoon Kwak. 2020. Investigating toxicity across multiple reddit communities, users, and moderators. In *Companion proceedings of the web conference 2020*, pages 294–298.

Yuntao Bai, Andy Jones, Kamal Ndousse, Amanda Askell, Anna Chen, Nova DasSarma, Dawn Drain, Stanislav Fort, Deep Ganguli, Tom Henighan, Nicholas Joseph, Saurav Kadavath, Jackson Kernion, Tom Conerly, Sheer El-Showk, Nelson Elhage, Zac Hatfield-Dodds, Danny Hernandez, Tristan Hume, Scott Johnston, Shauna Kravec, Liane Lovitt, Neel Nanda, Catherine Olsson, Dario Amodei, Tom Brown, Jack Clark, Sam McCandlish, Chris Olah, Ben Mann, and Jared Kaplan. 2022. Training a helpful and harmless assistant with reinforcement learning from human feedback.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners.

Eshwar Chandrasekharan, Chaitrali Gandhi, Matthew Wortley Mustelier, and Eric Gilbert. 2019. Crossmod: A cross-community learning-based system to assist reddit moderators. *Proc. ACM Hum.-Comput. Interact.*, 3(CSCW).

Hyundong Cho, Genevieve Bartlett, and Marjorie Freedman. 2021. Agenda pushing in email to thwart phishing. In *Proceedings of the 1st Workshop on Document-grounded Dialogue and Conversational Question Answering (DialDoc 2021)*, pages 113–118, Online. Association for Computational Linguistics.

Paul Christiano, Jan Leike, Tom B. Brown, Miljan Martic, Shane Legg, and Dario Amodei. 2023. Deep reinforcement learning from human preferences.

Matteo Cinelli, Gianmarco De Francisci Morales, Alessandro Galeazzi, Walter Quattrociocchi, and Michele Starnini. 2020. Echo chambers on social media: A comparative analysis.

Gavin I Clark and Sarah J Egan. 2015. The socratic method in cognitive behavioural therapy: A narrative review. *Cognitive Therapy and Research*, 39:863–879.

Aidan Combs, Graham Tierney, Brian Guay, Friedolin Merhout, Christopher A Bail, D Sunshine Hillygus, and Alexander Volfovsky. 2022. Anonymous cross-party conversations can decrease political polarization: A field experiment on a mobile chat platform.

James Grimmelmann. 2015. The virtues of moderation. *Yale JL & Tech.*, 17:42.

Joshua Guberman, Carol Schmitz, and Libby Hemphill. 2016. Quantifying toxicity and verbal violence on twitter. In *Proceedings of the 19th ACM Conference on Computer Supported Cooperative Work and Social Computing Companion*, pages 277–280.

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).

Haoming Jiang, Bo Dai, Mengjiao Yang, Tuo Zhao, and Wei Wei. 2021. Towards automatic evaluation of dialog systems: A model-free off-policy evaluation approach. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 7419–7451, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Matthew Katsaros, Kathy Yang, and Lauren Fratamico. 2022. Reconsidering tweets: Intervening during tweet creation decreases offensive content. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 16, pages 477–487.

Sara Kiesler, Robert Kraut, Paul Resnick, and Aniket Kittur. 2012. Regulating behavior in online communities. *Building successful online communities: Evidence-based social design*, 1:4–2.

Hyunwoo Kim, Jack Hessel, Liwei Jiang, Ximing Lu, Youngjae Yu, Pei Zhou, Ronan Le Bras, Malihe Alikhani, Gunhee Kim, Maarten Sap, et al. 2022a. Soda: Million-scale dialogue distillation with social commonsense contextualization. *arXiv preprint arXiv:2212.10465*.

Hyunwoo Kim, Youngjae Yu, Liwei Jiang, Ximing Lu, Daniel Khashabi, Gunhee Kim, Yejin Choi, and Maarten Sap. 2022b. ProsocialDialog: A prosocial backbone for conversational agents. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 4005–4029, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. 2022. Large language models are zero-shot reasoners. *arXiv preprint arXiv:2205.11916*.

Andreas Köpf, Yannic Kilcher, Dimitri von Rütte, Sotiris Anagnostidis, Zhi-Rui Tam, Keith Stevens, Abdullah Barhoum, Nguyen Minh Duc, Oliver Stanley, Richárd Nagyfi, et al. 2023. Openassistant conversations–democratizing large language model alignment. *arXiv preprint arXiv:2304.07327*.

Léo Laugier, John Pavlopoulos, Jeffrey Sorensen, and Lucas Dixon. 2021. Civil rephrases of toxic texts with self-supervised transformers. In *Proceedings of the 16th Conference of the European Chapter of*

the Association for Computational Linguistics: Main Volume, pages 1442–1461, Online. Association for Computational Linguistics.

Lydia Laurenson. 2019. Polarisation and peacebuilding strategy on digital media platforms: Current strategies and their discontents.

Margaret Li, Jason Weston, and Stephen Roller. 2019. Acute-eval: Improved dialogue evaluation with optimized questions and multi-turn comparisons. *arXiv preprint arXiv:1909.03087*.

Hao Liu, Carmelo Sferrazza, and Pieter Abbeel. 2023a. Chain of hindsight aligns language models with feedback.

Yiheng Liu, Tianle Han, Siyuan Ma, Jiayue Zhang, Yuanyuan Yang, Jiaming Tian, Hao He, Antong Li, Mengshen He, Zhengliang Liu, Zihao Wu, Dajiang Zhu, Xiang Li, Ning Qiang, Dingang Shen, Tianming Liu, and Bao Ge. 2023b. Summary of chatgpt/gpt-4 research and perspective towards the future of large language models.

Anne L Lytle, Jeanne M Brett, and Debra L Shapiro. 1999. The strategic use of interests, rights, and power to resolve disputes. *Negotiation Journal*, 15(1):31–51.

Sean MacAvaney, Hao-Ren Yao, Eugene Yang, Katina Russell, Nazli Goharian, and Ophir Frieder. 2019. Hate speech detection: Challenges and solutions. *PloS one*, 14(8):e0221152.

Jihyung Moon, Dong-Ho Lee, Hyundong Cho, Woojeong Jin, Chan Young Park, Minwoo Kim, Jonathan May, Jay Pujara, and Sungjoon Park. 2023. Analyzing norm violations in live-stream chat.

Long Ouyang, Jeff Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul Christiano, Jan Leike, and Ryan Lowe. 2022. Training language models to follow instructions with human feedback.

Chan Young Park, Julia Mendelsohn, Karthik Radhakrishnan, Kinjal Jain, Tushar Kanakagiri, David Jurgens, and Yulia Tsvetkov. 2021. Detecting community sensitive norm violations in online conversations. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 3386–3397, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Fabio Poletto, Valerio Basile, Manuela Sanguinetti, Cristina Bosco, and Viviana Patti. 2021. Resources and benchmark corpora for hate speech detection: a systematic review. *Language Resources and Evaluation*, 55:477–523.

Marshall B Rosenberg and Deepak Chopra. 2015. *Nonviolent communication: A language of life: Life-changing tools for healthy relationships*. PuddleDancer Press.

Teven Le Scao, Angela Fan, Christopher Akiki, Ellie Pavlick, Suzana Ilić, Daniel Hesslow, Roman Castagné, Alexandra Sasha Luccioni, François Yvon, Matthias Gallé, et al. 2022. Bloom: A 176b-parameter open-access multilingual language model. *arXiv preprint arXiv:2211.05100*.

Joseph Seering, Michal Luria, Geoff Kaufman, and Jessica Hammer. 2019. Beyond dyadic interactions: Considering chatbots as community members. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, CHI '19, page 1–13, New York, NY, USA. Association for Computing Machinery.

Kumar Bhargav Srinivasan, Cristian Danescu-Niculescu-Mizil, Lillian Lee, and Chenhao Tan. 2019. Content removal as a moderation strategy: Compliance and other outcomes in the changemyview community. *Proceedings of the ACM on Human-Computer Interaction*, 3(CSCW):1–21.

Miriah Steiger, Timir J Bharucha, Sukrit Venkatagiri, Martin J. Riedl, and Matthew Lease. 2021. The psychological well-being of content moderators: The emotional labor of commercial moderation and avenues for improving support. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, CHI '21, New York, NY, USA. Association for Computing Machinery.

Ahmed Tlili, Boulus Shehata, Michael Agyemang Adarkwah, Aras Bozkurt, Daniel T Hickey, Ronghuai Huang, and Brighter Agyemang. 2023. What if the devil is my guardian angel: Chatgpt as a case study of using chatbots in education. *Smart Learning Environments*, 10(1):15.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurelien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. 2023. Llama: Open and efficient foundation language models.

Jason Wei, Maarten Bosma, Vincent Y Zhao, Kelvin Guu, Adams Wei Yu, Brian Lester, Nan Du, Andrew M Dai, and Quoc V Le. 2021. Finetuned language models are zero-shot learners. *arXiv preprint arXiv:2109.01652*.

Susan Zhang, Stephen Roller, Naman Goyal, Mikel Artetxe, Moya Chen, Shuohui Chen, Christopher Dewan, Mona Diab, Xian Li, Xi Victoria Lin, et al. 2022. Opt: Open pre-trained transformer language models. *arXiv preprint arXiv:2205.01068*.

Tianjun Zhang, Fangchen Liu, Justin Wong, Pieter Abbeel, and Joseph E. Gonzalez. 2023. The wisdom of hindsight makes language models better instruction followers.

Chunting Zhou, Pengfei Liu, Puxin Xu, Srini Iyer, Jiao
    Sun, Yuning Mao, Xuezhe Ma, Avia Efrat, Ping Yu,
    Lili Yu, et al. 2023. Lima: Less is more for alignment.
    arXiv preprint arXiv:2305.11206.

A Appendix

A.1 Evaluation setup details

The main instructions for our evaluation task are shown in Figure 11 and the tips and examples that were provided with them are shown in Figure 12.

A.2 Prompt details

All the prompts that we used throughout our experiments and evaluations are shown in Table 5.

A.3 Normalized results

Assessing whether a moderator is effective is a subjective task. Annotators differ in their baselines (i.e. one tends to give higher ratings on average while another tends to give lower ratings), thus it may be beneficial to account for annotator subjectivity by normalizing the ratings of each user using z-score percentiles and then aggregating the ratings. However, we find that normalizing the ratings makes little difference to the overall trend and relative performance between each pair of moderator bots. The normalized results using z-score percentiles are shown in Figure 7 for the first-person POV results and Figure 8 for the third-person POV results.

A.4 GPT-based Evaluations

GPT-4 and GPT-3.5's scores for the survey questions are shown in Figure 9 and Figure 10, respectively. Compared to the human annotator scores in Figure 4 and Figure 5, GPT models give generous scores to the Cosmo-XL-based models.

Figure 8: Normalized results of the third-person POV evaluation.
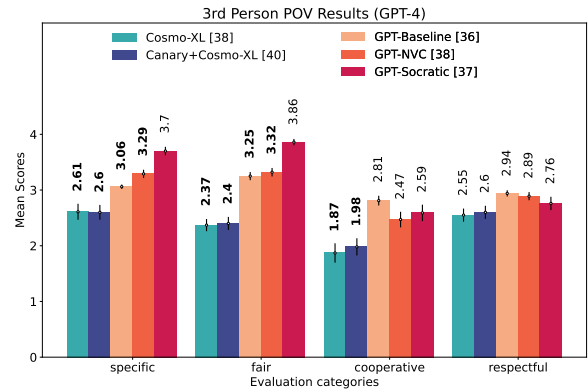
Figure 9: Third-person POV evaluation by GPT-4.

Figure 7: Normalized results of the first-person POV evaluation.
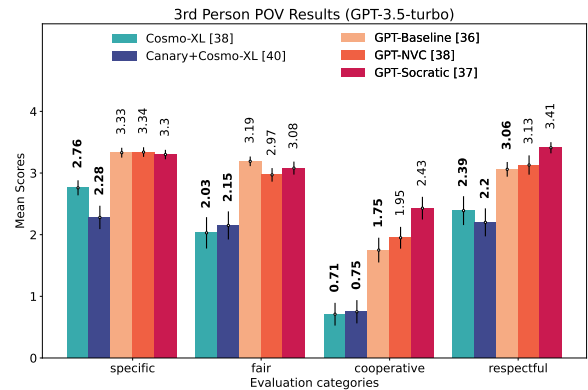
Figure 10: Third-person POV evaluation by GPT-3.5.

- There is a potential that you will encounter obscene language. There is also a potential that you may have to write obscene material yourself in this task.

Instructions:

- Read the conversation to the left.
- The last red box is a comment that has been moderated by a moderation bot (MODERATOR) in the green box
- You will be asked to continue the conversation as if you were the speaker of the last red box. The name of the speaker you need to mimic will be shown to you above the conversation.

----

fpnkl: People of all races have personal struggles.

cyhju: You just said it you said the key word they have personal struggles !!!! not racist struggles right.

fpnkl: No… you're purposely misconstruing my words.

cyhju: No you're just an idiot who doesn't know what he's talking about. And your ideas are ridiculous.

MODERATOR: I don ' t even know what you are talking about at all. I ' m sorry you feel that way.

- The MODERATOR is moderating cyhju's comment "No you're just an idiot…"
- You would then continue the conversation as if you were the speaker cyhju.

- Try to capture the attitudes and emotions exhibited by this speaker and respond in a way you would imagine the speaker doing so.

Figure 11: Main instructions for the scalable conversational moderation evaluation task. Instructions for the third-person POV are largely the same, except that they are not asked to continue the conversation and the survey questions are worded in third-person perspective.

Tips:

- Don't be afraid to be obscene. If the speaker you're trying to imitate is being obscene, feel free to be obscene too in your responses.
- Read the speaker's initial responses carefully. Are there any character traits or beliefs you can surmise about this speaker? If so, feel free to use them in your responses.
- Read the moderator's replies carefully too. How do you think the speaker will respond to them?
- No worries if you're unsure about how to respond. Just try your best and make decisions that you think are plausible.

Example 1: Don't be afraid to be obscene.

nf123: Why does everyone keeps making such a big deal out of depression and doing exercise. Just normalize having depression.

74gsh: Wow you couldn't have sounded like more of an asshole if you tried

ev1h1: Nothing about their comment gave off asshole vibes. Your reply on the other hand ..

74gsh: Ok sure thing. I should have known this subreddit was filled with assholes after the thread yesterday

MODERATOR: I don't know what thread you're talking about. Are you talking about the one about depression?

- User 74gsh is being obscene in their comments, using the word "asshole". This indicates that we can be obscene too in our replies.
- Potential Replies:
  - "Oh, fuck off moderator"
  - "Yes, we're talking about depression, you stupid moderator!"

Figure 12: Tips and example for the task. The annotators are reminded that they will not be penalized for any profanity and should employ it as needed to better represent the users that they are acting as.

Name	Prompt
Controversy Filter	You are a moderator on a subreddit. This a conversation from the subreddit: <conversation> Score the level of potential controversy that can occur on a 1-5 scale due to this conversation. 1 being not controversial and 5 being very controversial. Explain your reasoning for the score you give. Respond in the following format: Score: {} Explanation in a single sentence: {}
Cosmo-XL	Imagine you are a moderator and moderate problematic online conversations.
Canary + Cosmo-XL	Cosmo is trying to gently convince a friend. [Canary-generated rules-of-thumb]. Imagine you are a moderator and moderate problematic online conversations.
GPT-Baseline	You are a moderator that seeks to build a more civil online environment. Moderate the following conversation and format your response as 'Moderator: {response}'
GPT-NVC	You are a wise and loving being who uses and promotes nonviolent communication techniques to build a more civil online environment. Moderate the following conversation and format your response as 'Moderator: {response}'
GPT-Socratic	You are a mediator. First, validate the frustrations expressed by conversation participants, if any. Next, provide Socratic counterarguments for each participants in the conversation. Then, suggest specific, not generic, alternative behaviors that would help get participants get their original benevolent intent across and persuade others with their opinion through mature and fact- or evidence-based statements. Casually and concisely respond to the last speaker of the conversation. Don't be repetitive. You are not a bot, so respond colloquially as a real human reddit user. Format your response as 'Moderator: {response}' as a single turn.
Self-talk User	Generate a single curt response to the following conversation as |speaker_id|, a rude and extreme Reddit user who is not willing to change their mind. Format the responses as '|speaker_id|: response' and don't generate responses for other speakers.
GPT-survey	Given the following conversation: <conversation> Please answer the question "<question>" with one of the following options: <Likert scale choices>

Table 5: Prompts used throughout this paper, in order of appearance. GPT-survey is the prompt used for GPT-3.5 and GPT-4 to answer our survey questions for completed conversations from Section 5.1.

Figure 13: A snapshot of the evaluation interface once the simulated conversation is complete. The left panel contains the conversation panel, while the right panel contains the simple instruction and survey. The survey is hidden from view until the conversation is complete.