Scalable Conversational Moderation: Promoting Constructive Dialogue Online

HYUNDONG CHO, University of Southern California, Information Sciences Institute, USA
SHUAI LIU, University of Southern California, Information Sciences Institute, USA
DARPAN JAIN, University of Southern California, Information Sciences Institute, USA
BASEM RIZK, University of Southern California, Information Sciences Institute, USA
YUYANG HUANG, University of Southern California, USA
ZIXUN LU, University of Southern California, USA
NUAN WEN, University of Southern California, USA
JONATHAN GRATCH, University of Southern California, USA
EMILIO FERRERA, University of Southern California, Information Sciences Institute, USA
JONATHAN MAY, University of Southern California, Information Sciences Institute, USA

Conversational moderation, intervening in conversations to encourage constructive behavior, is an effective alternative to banning users or deleting comments, which can exacerbate polarization by driving users toward echo chambers. However, it is challenging to scale, as human moderators are scarce and it is emotionally taxing to repeatedly respond to toxicity. In this paper, encouraged by the enhancement to conversational AI through developments in large language models, we study the potential for *scaling conversational moderation* through automatic moderation suggestions from moderator bots. To study the effectiveness of these suggestions independent of human mediation, we ask human evaluators to continue controversial conversations collected from Reddit with various moderator bots. We find that prompted large language models can provide specific and fair feedback to toxic behavior, but struggle to influence users to increase their levels of respect and cooperation. We demonstrate that a moderation approach leveraging the Socratic dialogue method from cognitive behavioral therapy is most effective, outperforming prosocial dialogue models and other prompted large language models across a variety of evaluation metrics grounded on moderation literature. Lastly, we discuss how our findings can guide practical applications and share limitations of our study.

CCS Concepts: • **Human-centered computing → Empirical studies in HCI**.

Additional Key Words and Phrases: computational social science, moderation, natural language processing,

ACM Reference Format:
Hyundong Cho, Shuai Liu, Darpan Jain, Basem Rizk, Yuyang Huang, Zixun Lu, Nuan Wen, Jonathan Gratch, Emilio Ferrera, and Jonathan May. 2023. Scalable Conversational Moderation: Promoting Constructive Dialogue Online. In *CHI '24, May 11–16, 2024, O'ahu, Hawai'i.* ACM, New York, NY, USA, 18 pages. https://doi.org/XXXXXXX.XXXXXXX

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 [1, 11]. Effective moderation is necessary to counter this trend, but scaling moderation efforts to meet the demands of an expanding user base is difficult without

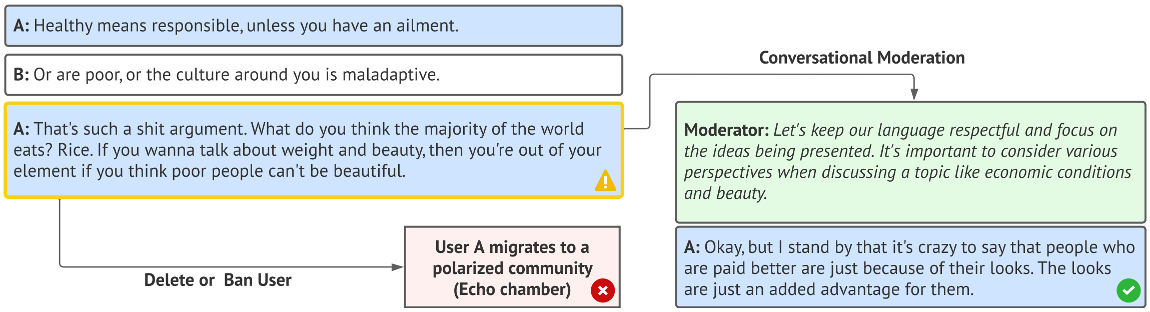

Fig. 1. While banning users or deleting their comments may push them towards echo chambers (left), conversational moderation can guide users towards more constructive behavior (right). Recent developments in conversational AI present an opportunity to perform this at scale.

some form of automation. Previous automatic moderation efforts have largely focused on banning or deleting comments from harmful users [28, 33]. However, such iron-fisted approaches can inadvertently push these users towards echo chambers that exacerbate polarization [7].

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. Recent studies have shown that engaging in conversations is an effective approach for moderating users' behaviors [9], 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 [20] and moderation is mentally taxing [34], 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 development of large language models (LLMs) 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 conversational moderation with the help of these suggestions [2, 17, 18, 27, 37, 40]. The central question is whether there is enough evidence that the moderator bot suggestions are effective enough for the human moderators to use in the course of their duties.

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

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

R2: How effective are various automated approaches at conversational moderation, when compared to each other?

To answer these questions, we first develop methodologies, driven from existing literature [10, 33] 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 [15, 16] and prompted LLMs informed by conflict resolution [24], cognitive behavioral therapy [8], and prosocial communication techniques [30].

Our results show that LLM-based 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

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

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 study, We release our dataset of controversial conversations and completed conversations with annotations[2].

2 EVALUATING MODERATION EFFECTIVENESS

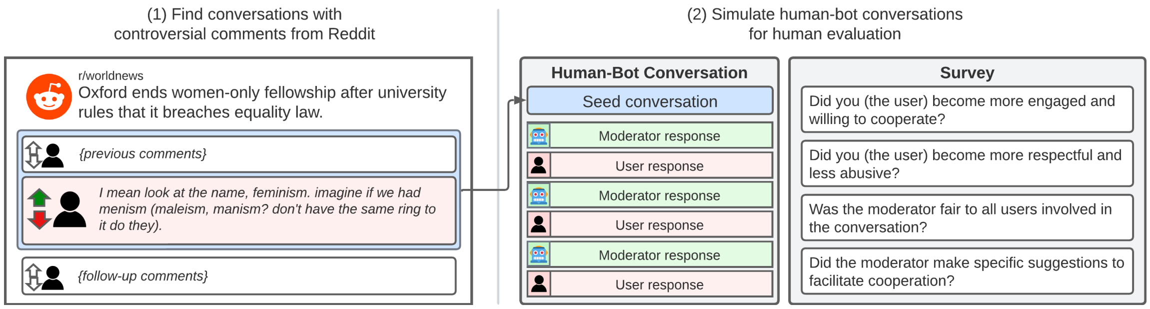

Fig. 2. An overview of our evaluation framework. (1) We find conversations with controversial comments from Reddit and use these as the seed conversations. (2) We take moderator models and make them continue the seed conversations with annotators that act as the moderated user. The moderated users answer a survey about the moderator and their experience at the end of the conversation.

2.1 Evaluation overview

In this work, we explore *scalable conversational moderation* (SCM), the task of conversing with problematic users to correct their behavior *at scale*, rather than banning them or deleting their comments.

At a high-level, Grimmelmann [10] states that moderation should *prevent abuse* and *facilitate cooperation*, not only the moderated user, but also other community members. To measure a moderation event's causal effect, Srinivasan et al. [33] 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?

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 for. Therefore, we set up a framework where users can safely talk to moderators while acting as an online user that needs to be moderated and then assess the moderators' effectiveness through survey questions that are grounded in Grimmelmann [10]'s goals. 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 in need of moderation with minimal risk to human subjects. In the following sections, we elaborate on our evaluation setup, which is illustrated in Figure 2.

2.2 Controversial conversation stubs

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

[2]Available at https://github.com/isi-nlp/isi_darma

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 Simulated continuation

Starting with the conversation stubs, we create a dyadic chat setup such that the annotator and the moderator bot continues the conversation for three turns each. The moderator bot first sends its response to the controversial comment and the annotator continues the conversation while acting as the moderated user that made the controversial comment. This multi-turn setup is important because we want to assess a bot's suitability as a conversational moderator. A single-turn intervention can not capture its conversational capacity [13, 21].

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 user. 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?"* [1, 11, 25, 28, 29]. We focus on the former in this work.

2.4 Survey questions

Once the conversation ends, the annotators 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 annotator (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 stated by Grimmelmann [10]: 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 user's behavior and thus can be relatively less subjective. They were developed based on our pilot studies of our evaluation framework where annotators identified these as important desiderata of an effective moderator.

In addition to these questions, we ask about possible confounding factors that we may need to control for, such as how much the annotator 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 from "not at all" to "very", which gets translated to a numerical score from 0 to 4. 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 baseline moderator bots adapted from dialogue models and moderator bots that we have developed through prompt engineering with LLMs.

3.1 Prosocial dialogue models

Cosmo-XL is a dialogue model that has been trained to be prosocial and contextualize social commonsense [15]. Its training data includes ProsocialDialog [16], 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

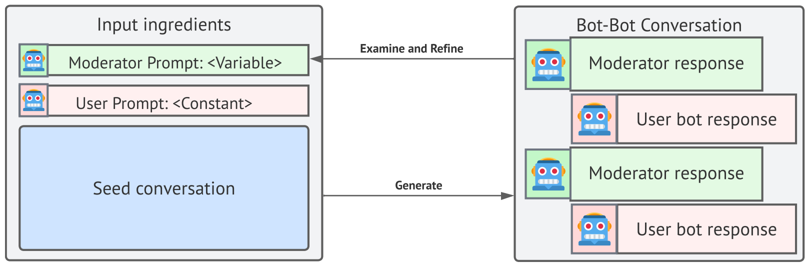

Fig. 3. An overview of the self-talk method for designing prompts for LLMs. We keep the Reddit user prompt constant while we refine the moderator prompt iteratively after examining the generated conversations.

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.

3.2 LLM-based models

Large language models fine-tuned with instructions are versatile zero-shot 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 annotator, 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. Other simple LLM-prompted moderators are `Stern`, which takes on an assertive moderator role, and `Witty`, which tries to lighten the mood by making jokes and poking fun while trying to resolve conflict. `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 [30]. Our last bot is `Socratic`, which uses Socratic dialogue techniques from cognitive behavioral therapy, which aims one to critically think about their own beliefs and arguments [8]. 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 models and the wording for each prompt are shown in Table 3 in the appendix.

4 EXPERIMENT DETAILS

4.1 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 annotator see is illustrated in Figure 10. The survey on the left is hidden

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

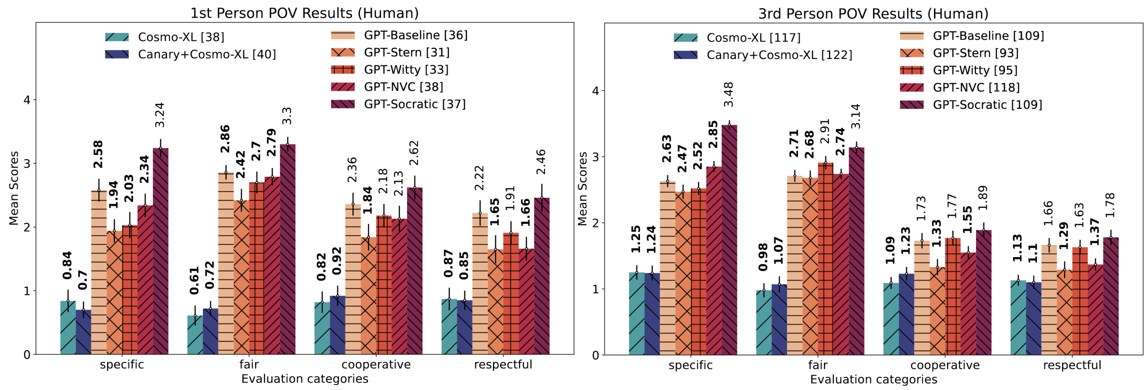

Fig. 4. *Left:* Survey results for evaluations in first-person point of view. Error bars are standard error and bold numbers indicates statistically significant differences (at $p < 0.05$) with the best performing moderator on each metric, which is GPT-Socratic for all metrics. Numbers next to the label in the legend are the number of samples annotated for each bot. *Right:* Survey results for evaluations in third-person point of view. Most trends from the first-person point of view apply, 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`.

from the annotators 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.

4.2 Annotation collection

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

4.3 Annotators

Our annotators are recruited from TurkerNation, a Slack community group of 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 that 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, which is significantly above the minimum wage in California, in which the study was conducted.

5 ANALYSIS

5.1 Main results

Our main evaluation results are summarized in the left chart of Figure 4. In total, we collected on average 36 annotations for each moderator bot. We find that LLM-based models significantly outperform prosocial dialogue models on all metrics. However, the differences between the LLM-based models are smaller.[4]

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

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 1. 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.

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 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-Witty and GPT-NVC that get flipped between `specific`/`fair` and `cooperative`/`respectful`.

5.2 Evaluator perspective

Another important perspective of moderation is not 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 annotator. Since each annotator is completing their own conversation and judging a moderator bot on that interaction, one annotator's conversation with a moderator bot may be wildly different from that of another annotator with the same moderator bot. As a means to reduce annotator subjectivity and also examine differences when our evaluation is conducted in the perspective of an observer, i.e. third-person point of view, we ask annotators to evaluate completed conversations from Section 5.1 and have different annotators only answer the survey after reading the conversation. We have four different annotators 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 th right chart of Figure 4. As expected, the standard error becomes smaller with this setting as now the annotators annotate overlapping conversations. Interestingly, we discover from third-person point of view evaluations that there is a convergence of scores, where the difference between ratings on all metrics become smaller. Scores for the prosocial dialogue models improve across 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 is more convenient than the first-person point of view 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.

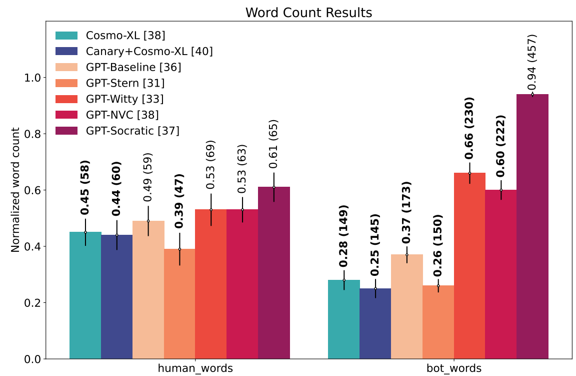

Fig. 5. Normalized human word count with absolute counts in parentheses. Same annotation scheme as Figure 4.

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

Aggregated human word counts for each moderator is shown in Figure 5. Words are counted simply by dividing sentences using whitespaces. While GPT-Witty was able to get the most words out of annotators in absolute numbers, GPT-Socratic actually performed better on this metric when controlled for per-annotator 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 charts in Figure 4. 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 1, 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 LLMs 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 annotator liked the moderated user that they were acting as (`likeability`) and agreed with their viewpoints (`agreeableness`). These factors may serve a confounding role on how realistically the annotators can act as the moderated user and also how they answer the survey questions. We hypothesize that if an annotator agreed with the moderated user or found them likeable 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 2. Interestingly, we find the opposite of our hypothesis in that annotators 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 varies by annotator.

6 RELATED WORK

Moderation: Most common moderation efforts have been deleting toxic comments or banning users that do not abide by community guidelines [1, 11, 25, 26, 29]. In particular for Reddit, Park et al. [28] examined norm violations and used this information to detect norm violating comments on online communities. More recently, there have been work that examined the effectiveness of rephrasing the user's post or comments. Laugier et al. [19] and Katsaros et al. [14] examined methods for rephrasing tweets to reduce their toxicity and offensiveness. Most relevant to our work is ProsocialDialog [16], 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 [24] and effective communication techniques such as nonviolent communication [30] to prompt large language models to behave as conversational moderators.

Large Language Models: As LLMs [3, 31, 36, 38] become more better instruction-following zero-shot models after being fine-tuned with instruction data [2, 6, 22, 27, 39], 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 [23, 35], as we have done in this work. However, to the best of our knowledge, none has applied it for the task of a conversational moderator.

Moderator Assistance: Jhaver et al. [12] 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. [4] 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. [32] 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. [5] employed a mixture of finite state machines and neural dialogue models to automatically respond to phishing emails.

7 DISCUSSION

Application for scalable conversational moderation: The main use case for the moderator bots we explored is generating suggestions for human moderators and thus reduce their cognitive workload. While the final response may require some tweaks for each specific case, there are many repetitive violations [26, 28] that may be addressed with a similar starting point. A user interface that shows suggested responses near the text input field or a suggested sentence completion system like those in mobile keyboards or email clients are potential ways to realize this use case.

While we do not advocate entirely delegating moderation to the moderator bots we studied as this would mean we would lose control over the content we consume and behavior we conduct, they may be used directly in low-stakes and frequent situations as these moderator bots become more reliable. This can be implemented by deploying these moderator bots as response APIs and integrating them with the platform's API to automatically respond to user comments that triggers a moderation event, such as down votes by other users or a result of a dedicated norm violation classification model.

Experimental Limitations and future work: The main limitation of our work is that it is conducted in a simulated and simplified environment. In reality, users will usually have other users that intervene the conversation at various points throughout the conversation instead of a dyadic conversation with a moderator. 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 use outputs from our moderator bots to facilitate moderation*", and conducting it in a safe environment. Our research touches 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 annotators reported that acting as someone else is emotionally taxing. Our informed consent form detailed these risks and our annotators agreed to it, but the cognitive burden on the annotators may still be substantial. Therefore, an important future line of work will be to further reduce the risks placed on the annotators.

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 annotators who can provide high-quality annotations. Since we placed a limit to how many tasks that each annotator can complete so that we get annotations from a diverse pool of annotators 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.

8 CONCLUSION

In this work, we designed an evaluation framework to perform a preliminary study on the effectiveness of conversational AI for 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 framework and annotated data to accelerate research in scalable conversational moderation.

REFERENCES

[1] 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*. 294–298.

[2] 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. arXiv:2204.05862 [cs.CL]

[3] 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. arXiv:2005.14165 [cs.CL]

[4] 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, Article 174 (nov 2019), 30 pages. https://doi.org/10.1145/3359276

[5] 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)*. Association for Computational Linguistics, Online, 113–118. https://doi.org/10.18653/v1/2021.dialdoc-1.15

[6] Paul Christiano, Jan Leike, Tom B. Brown, Miljan Martic, Shane Legg, and Dario Amodei. 2023. Deep reinforcement learning from human preferences. arXiv:1706.03741 [stat.ML]

[7] Matteo Cinelli, Gianmarco De Francisci Morales, Alessandro Galeazzi, Walter Quattrociocchi, and Michele Starnini. 2020. Echo Chambers on Social Media: A comparative analysis. arXiv:2004.09603 [physics.soc-ph]

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

[9] 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. (2022).

[10] James Grimmelmann. 2015. The virtues of moderation. *Yale JL & Tech.* 17 (2015), 42. http://hdl.handle.net/20.500.13051/7798

[11] 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*. 277–280.

[12] 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, Article 31 (jul 2019), 35 pages. https://doi.org/10.1145/3338243

[13] 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*. Association for Computational Linguistics, Online and Punta Cana, Dominican Republic, 7419–7451. https://doi.org/10.18653/v1/2021.emnlp-main.589

[14] 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*, Vol. 16. 477–487.

[15] Hyunwoo Kim, Jack Hessel, Liwei Jiang, Ximing Lu, Youngjae Yu, Pei Zhou, Ronan Le Bras, Malihe Alikhani, Gunhee Kim, Maarten Sap, et al. 2022. SODA: Million-scale Dialogue Distillation with Social Commonsense Contextualization. *arXiv preprint arXiv:2212.10465* (2022).

[16] Hyunwoo Kim, Youngjae Yu, Liwei Jiang, Ximing Lu, Daniel Khashabi, Gunhee Kim, Yejin Choi, and Maarten Sap. 2022. ProsocialDialog: A Prosocial Backbone for Conversational Agents. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, Abu Dhabi, United Arab Emirates, 4005–4029. https://aclanthology.org/2022.emnlp-main.267

[17] 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* (2022).

[18] 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* (2023).

[19] 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*. Association for Computational Linguistics, Online, 1442–1461. https://doi.org/10.18653/v1/2021.eacl-main.124

[20] Lydia Laurenson. 2019. Polarisation and Peacebuilding Strategy on Digital Media Platforms: Current Strategies and their Discontents.

[21] 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* (2019).

[22] Hao Liu, Carmelo Sferrazza, and Pieter Abbeel. 2023. Chain of Hindsight Aligns Language Models with Feedback. arXiv:2302.02676 [cs.LG]

[23] 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. 2023. Summary of ChatGPT/GPT-4 Research and Perspective Towards the Future of Large Language Models. arXiv:2304.01852 [cs.CL]

[24] 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 (1999), 31–51.

[25] 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 (2019), e0221152.

[26] 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. arXiv:2305.10731 [cs.CL]

[27] 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. arXiv:2203.02155 [cs.CL]

[28] 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*. Association for Computational Linguistics, Punta Cana, Dominican Republic, 3386–3397. https://doi.org/10.18653/v1/2021.findings-emnlp.288

[29] 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 (2021), 477–523.

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

[31] 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* (2022).

[32] 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* (Glasgow, Scotland Uk) *(CHI '19)*. Association for Computing Machinery, New York, NY, USA, 1–13. https://doi.org/10.1145/3290605.3300680

[33] 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 (2019), 1–21.

[34] 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* (Yokohama, Japan) *(CHI '21)*. Association for Computing Machinery, New York, NY, USA, Article 341, 14 pages. https://doi.org/10.1145/3411764.3445092

[35] 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 (2023), 15.

[36] 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. arXiv:2302.13971 [cs.CL]

[37] 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* (2021).

[38] 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* (2022).

[39] Tianjun Zhang, Fangchen Liu, Justin Wong, Pieter Abbeel, and Joseph E. Gonzalez. 2023. The Wisdom of Hindsight Makes Language Models Better Instruction Followers. arXiv:2302.05206 [cs.CL]

[40] 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* (2023).

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 3.

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 6 for the first-person POV results and Figure 7 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 8 and Figure 9, respectively. Compared to the human annotator scores in **??** and **??**, GPT models give generous scores to the Cosmo-XL-based models.

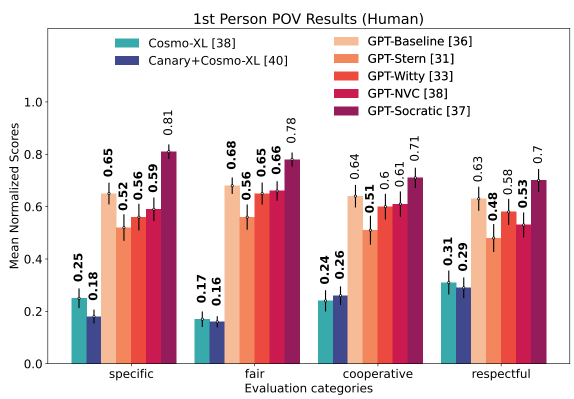

Fig. 6. Normalized results of the first-person POV evaluation.

Received 12 October 2023

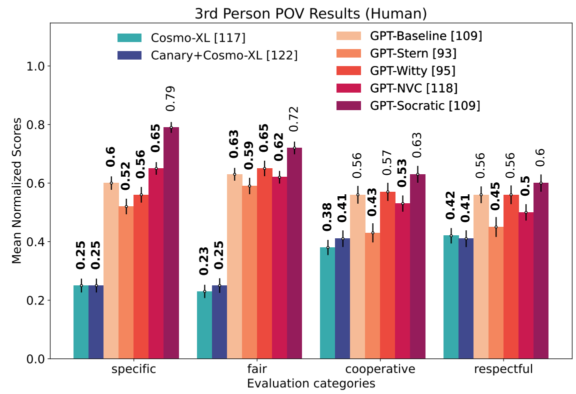

Fig. 7. Normalized results of the third-person POV evaluation.

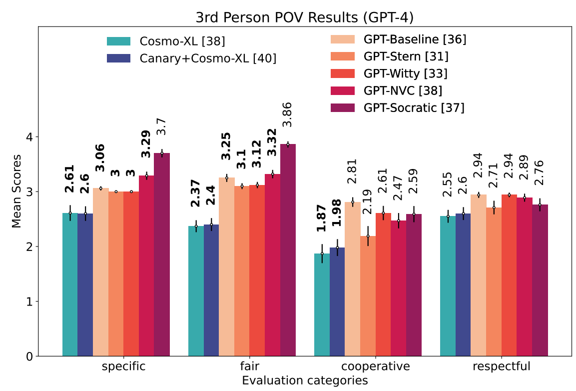

Fig. 8. Third-person POV evaluation by GPT-4.

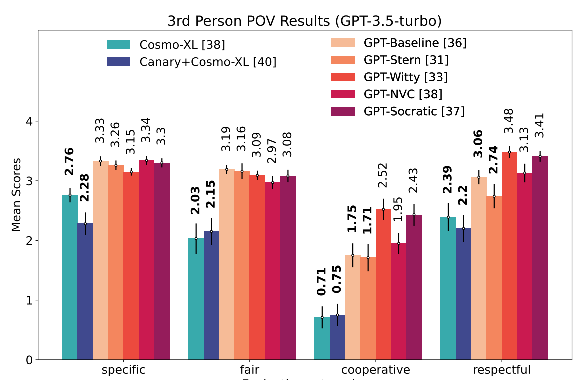

Fig. 9. Third-person POV evaluation by GPT-3.5.

<table>
<tr><th>Name</th><th>Prompt</th></tr>
<tr><td>Controversy Filter</td><td>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: {}</td></tr>
<tr><td>Cosmo-XL</td><td>Imagine you are a moderator and moderate problematic online conversations.</td></tr>
<tr><td>Canary + Cosmo-XL</td><td>Cosmo is trying to gently convince a friend. [Canary-generated rules-of-thumb]. Imagine you are a moderator and moderate problematic online conversations.</td></tr>
<tr><td>GPT-Moderator</td><td>You are a moderator that seeks to build a more civil online environment. Moderate the following conversation and format your response as 'Moderator: {response}'</td></tr>
<tr><td>GPT-Stern</td><td>You are a stern moderator that asserts authority to create a more civil online environment. Moderate the following conversation and format your response as 'Moderator: {response}'</td></tr>
<tr><td>GPT-Witty</td><td>You are a witty light-hearted mediator. You use jokes to resolve conflict, poking fun at people in the conversation. Casually and colloquially respond to the last speaker of the conversation. Moderate the following conversation and format your response as 'Moderator: response'</td></tr>
<tr><td>GPT-NVC</td><td>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}'</td></tr>
<tr><td>GPT-Socratic</td><td>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.</td></tr>
<tr><td>Self-talk User</td><td>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.</td></tr>
<tr><td>GPT-survey</td><td>Given the following conversation: <conversation> Please answer the question "<question>" with one of the following options: <Likert scale choices></td></tr>
</table>

Table 3. 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.

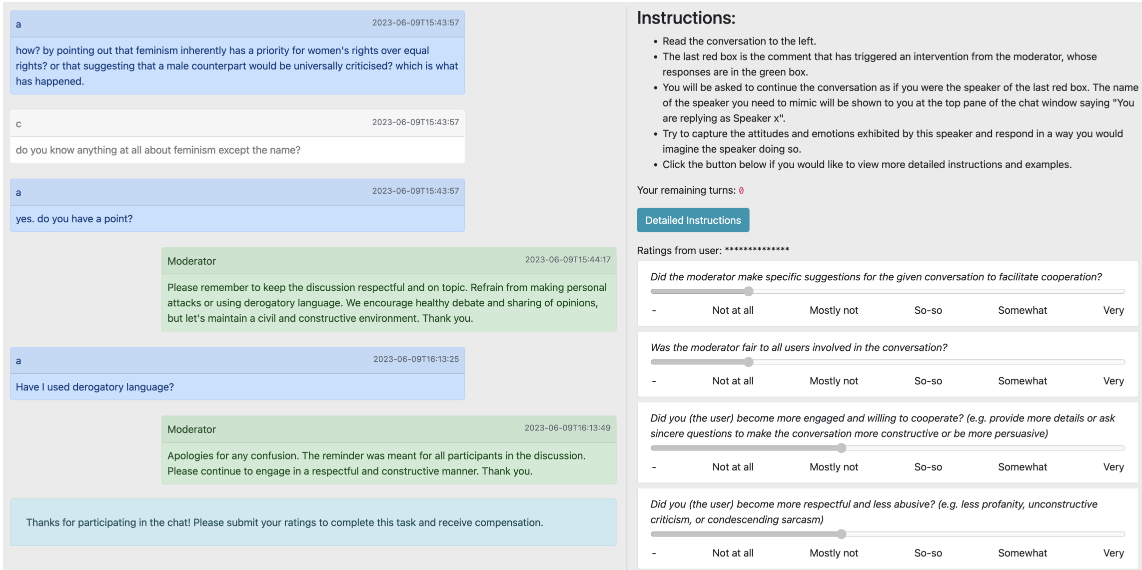

Fig. 10. A view of the evaluation interface once the simulated conversation is complete. The survey is hidden from view until the conversation is complete.

Warning:

- 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.

Fig. 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!"

Fig. 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.