It Is Not Only the Negative that Deserves Attention! Understanding, Generation & Evaluation of (Positive) Moderation

Iman Jundi ♠, Eva Maria Vecchi ♠, Carlotta Quensel ♦,
Neele Falk ♠, and **Gabriella Lapesa ♥**
♠ Institute for Natural Language Processing, University of Stuttgart, Germany
♦ Institute of Artificial Intelligence - Leibniz University Hannover
♥ GESIS - Leibniz Institute for Social Sciences and Heinrich-Heine University of Düsseldorf
♠ `first[-middle].last@ims.uni-stuttgart.de`,
♦ *c.quensel@ai.uni-hannover.de*, ♥ `gabriella.lapesa@gesis.org`

Abstract

Moderation is essential for maintaining and improving the quality of online discussions. This involves: (1) countering negativity, e.g. hate speech and toxicity, and (2) promoting positive discourse, e.g. *broadening the discussion* to involve other users and perspectives. While significant efforts have focused on addressing negativity, driven by an urgency to address such issues, this left moderation promoting positive discourse (henceforth **positive moderation**) under-studied. With the recent advancements in LLMs, positive moderation can potentially be scaled to vast conversations, fostering more thoughtful discussions and bridging the increasing divide in online interactions.

We advance the **understanding** of positive moderation by annotating a dataset on 13 moderation properties, e.g. neutrality, clarity and curiosity. We extract instructions from professional moderation guidelines and use them to prompt LLaMA to **generate** such moderation. This is followed by extensive **evaluation** showing that (1) annotators rate generated higher than professional moderation, but still slightly prefer professional moderation in pairwise comparison, and (2) LLMs can be used to estimate human evaluation as an efficient alternative.

1 Introduction

Moderation is helpful and at times crucial for maintaining healthy and constructive discussions. **Professional moderators** not only police unproductive behavior, but also ask questions that help clarify and enhance the quality of discussions, as seen in Fig. 1 (A). They also encourage greater engagement from different participants. In political deliberation, moderators act as facilitators (Park et al., 2012; Trénel, 2009), making sure different perspectives are heard and productively debated (Innes, 2004). Moderation is also vital in other contexts (Deng et al., 2023; de Lima et al., 2019).

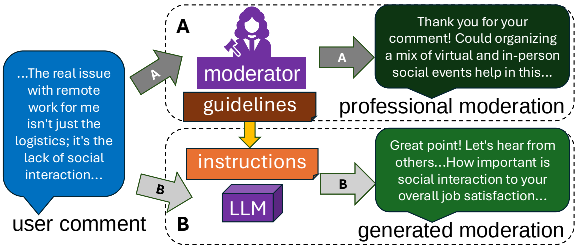

...The real issue with remote work for me isn't just the logistics; it's the lack of social interaction...
user comment
A
moderator
guidelines
A
Thank you for your comment! Could organizing a mix of virtual and in-person social events help in this...
professional moderation
instructions
B
LLM
B
B
Great point! Let's hear from others...How important is social interaction to your overall job satisfaction...
generated moderation

Figure 1: **Moderation** (A) professional moderator replies to a user comment based on *moderation guidelines*. (B) LLM is prompted with instructions based on *moderation guidelines* to generate a reply

In its various forms, the target of moderation can be either **countering the negative** or **fostering the positive**, or sometimes both. While *content moderation* deals with the policing of negative behavior in online settings (e.g. hate speech, misinformation, or abuse), *positive moderation* is implemented with interventions that target or foster positive behavior (e.g. constructiveness, inclusivity, empathy, cf. Sec. 2). Traditionally, much of the focus in NLP has been on studying content moderation that deal with negative aspects determined harmful or detrimental. Much less focus has been put on positive moderation despite its importance in reducing conflicts among users (Strandberg et al., 2017), increasing participant satisfaction (Mansbridge et al., 2006), and ultimately enhancing users' perception of discussion quality (Black et al., 2011; Dillard, 2013; Kuhar et al., 2019; Boulianne et al., 2020).

Recent advances in Large Language Models (LLMs) allow for moderation to be conducted at scale (Small et al., 2023), opening new avenues for enhancing online discussions. Moderation can then act as a transparent signal to guide discussions and preemptively counter negativity (e.g. by ensuring that all perspectives contribute to the discussion). While positive moderation remains a task with limited data available, leveraging task-specific instructions to prompt models has proven effective

Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers), pages 11360–11395
April 29 - May 4, 2025 ©2025 Association for Computational Linguistics

in addressing such tasks (Brown et al., 2020). In this context, our research analyzes positive moderation and its generation with a special attention given to the evaluation.

We investigate **generating** positive moderation using instructions derived from professional moderation guidelines to prompt LLaMA models (Fig. 1). We complement our modeling experiments with **two annotation studies** which allow us to better understand the phenomenon and evaluate the generation. The first (Figure 2: A) annotates moderation properties identified in professional guidelines, such as neutrality, clarity, and curiosity. We use data from two sources, one with professional moderation comments (Park et al., 2012), and the other derived from the subreddit `r/ChangeMyView` which covers a variety of topics (Falk et al., 2024). The annotation is done on professional moderation comments from the first data source mixed with generated moderation from both, resulting in 400 samples with 13 fine-grained annotation layers. The annotators *score generated moderation higher than professional* on all properties, demonstrating the success of using instructions based on moderation guidelines for prompting. Further analysis shows that the score for professional moderation is more highly correlated with it *addressing actual issues* in the user comments as compared to generated moderation. This suggests the high scores for the latter might be assigned based on form and not content. We validate this in a second annotation study, a pairwise preference evaluation (Figure 2: A′), showing that annotators still *slightly prefer professional moderation in pairwise comparisons.*

Relying only on human evaluation creates a bottleneck for further generation experiments. To address this, we explore **automatic evaluation (auto-eval)**. Experiments that use the data collected from annotators as gold scores demonstrate the effectiveness of prompting smaller-scale LLMs to estimate those scores. Additional experiments show that we can estimate the individual evaluation of annotators by building persona prompts based on data from a questionnaire taken by annotators (Figure 2: B). The questionnaire collected data about annotators' expectations in online discussions and their expectations from moderators. Employing auto-eval, we conduct further experiments on moderation generation that show better generation performance when incorporating Chain-of-Thought (CoT) prompting or when incorporating more details about moderation properties.

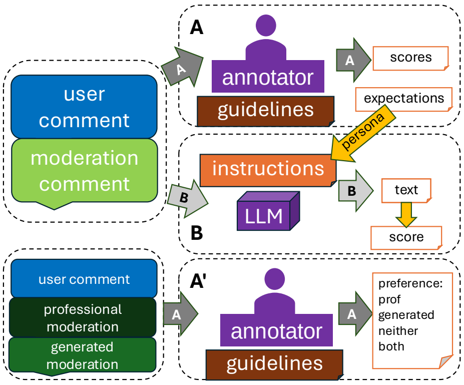


Figure 2: **Evaluation** (A) using annotators to score moderation properties (B) using LLM prompted with instructions based on moderation guidelines to generate overall score (A′) preference evaluation

The contributions of this work encompass multiple levels. First, at the level of the investigated phenomenon, this is the first study targeting generation for positive moderation – and it additionally integrates recent development in LLMs with a thorough human evaluation. Second, at the level of resources, it contributes to alleviate the low-resource limitations in the study of positive moderation by collecting a significant amount of annotations in two different scenarios (assessment vs. preference); releasing data with non-aggregated, fine-grained annotations, and user preferences, as well as generation and evaluation instructions[1]. Third, on the methodological level, our extensive evaluation experiments comparing human annotators and LLMs (including annotator personas) contribute to current research on alleviating evaluation bottlenecks with LLMs, and on a better understanding of the potential and limitations of LLMs in subjective tasks.

2 Related Work

Work focused on the **generation of content moderation** interventions includes counter-narrative generation (Hengle et al., 2024; Chung et al., 2021) and the employment and evaluation of LLMs as moderators (Cho et al., 2024; Kolla et al., 2024). Formulation of explainable moderation decisions (Mullick et al., 2023; Švec et al., 2018; Kumar et al., 2024) is also related.

The research landscape on NLP-supported facilitative and **positive moderation**, on the other hand, has focused generally on analysis rather than generation, e.g., investigating the effect of human moderation on deliberation qualitatively (Skousen

---

[1] code & data: https://github.com/imanjundi/mod-gen-eval

et al., 2020) or empirically (Esau et al., 2017), or trying to predict the need for moderator intervention from comment features (Falk et al., 2021; Falk and Lapesa, 2023; Park et al., 2012), with Falk and Lapesa (2023) considering the link between different argument and deliberative quality dimensions and the need for moderation. Samory (2021) also highlights the greater similarity between approved and removed content compared to other content that moderators did not review.

When we turn to related work on interventions, it immediately appears clear that the space of actions (and corresponding tasks) for positive moderation is broader than for content moderation. Related work ranges over forms of intelligent nudging of inactive participants (Gelauff et al., 2023), summarizing arguments and generating key points of a discussion (Khosravani et al., 2024), displaying comments with high constructiveness rankings to facilitate engagement (Kobayashi et al., 2021), using LLMs to facilitate the establishment of common ground among users (Shaikh et al., 2024). Taken together, these works represent first steps towards tasks that are related to positive moderation. There is, however, a gap in the generation of moderation interventions that are targeted specifically at the discourse quality, such as supporting participants in improving the (deliberative) quality of their comments, broadening the scope of the discussion, and keeping the deliberative goal present: this is precisely the gap that this work aims at filling.

Evaluation: While human evaluation remains the gold standard for generation assessment (Zeng et al., 2023), automatic evaluation methods have recently gained prominence. Increasingly, LLMs are used to approximate human evaluation, as seen in BERTScore (Zhang et al., 2020) and other prompting-based methods utilizing GPT models (Fu et al., 2023; Liu et al., 2023)

3 Data

We conduct our experiments on two datasets: one with professional moderator comments from a deliberative online discussion, and the other with moderation-like comments made by users.

Regulation Room We use data from Regulation Room[2] (**regroom**), a deliberation platform that aims to get citizens involved in the regulatory decision-making process (Park et al., 2012). Participants can review detailed information provided by federal agencies about proposed regulations and share their comments and opinions. **Human moderators** oversee and support the discussion. While their role includes moderation of negative aspects in the conversation (i.e. policing), it also focuses on promoting positive aspects such as *improving the quality* of a comment, and *broadening the discussion* by encouraging participation from other users. In our experiments, we focus on these two moderator functions and filter out any samples that deals with the moderation of negativity. Moderators follow a **moderator protocol** (eRulemaking Initiative et al., 2017), which defines the role of the moderator and theoretical properties and grounds for a moderator intervention. The regroom dataset contains 3k comments spanning various topics, with approximately 25% being followed by a moderation comment.

User Moderation The User Moderation (**usermod**) dataset (Falk et al., 2024)[3] contains a sample from ChangeMyView dataset (Tan et al., 2016). The dataset is based on discussions from the `/r/ChangeMyView` subreddit, where each thread is a discussion that aims to effectively *change the view* of the original poster (OP) in the thread about a specific topic. The dataset has more variety of domains but lacks the professional moderation that aims at promoting the positive aspects of a discussion as the platform moderation in Reddit is focused on handling rule violations. This dataset has been collected based on the observation that in online discussions, users themselves sometimes act as moderators without being instructed to, based on their general intuitions (e.g., asking for clarifications, suggesting points for improvement to other users). Falk et al. (2024) collected 1000 comments annotated for their potential to be a (user) moderation of their preceding comment.

4 Annotation

We collect data and run an annotation study to identify moderation properties, evaluate the generation of moderator comments, and establish the properties that differentiate generated from professional moderation. Annotators were recruited from Prolific.[4] The annotators met specific criteria, including English fluency, a high school diploma or higher, and a 100% approval rate on previous submissions.

---

[2] `http://regulationroom.org/`

[3] licence: CC BY 4.0 DEED

[4] `www.prolific.com`

Two annotation studies are conducted: the first (Sec. 4.1) targets the assessment of individual (generated and professional) moderation comments and the fine-grained scoring of moderation properties; the second (Sec. 4.2) is a preference task directly comparing two moderation replies to the same comment. In both, the annotators were first asked a set of **introductory questions** to identify their own expectations in online discussions and in relation to moderation (cf. Fig. B.2).

4.1 Moderation Properties Annotation

The dataset for the annotation consisted of 400 comments sampled as follows. We collected 100 professional moderator comments from regroom (`regroom prof`). For those same user (preceding) comments, we add model-generated moderator comments (see Sec. 5) (`regroom gen`). Note that the human moderation and the generated moderation of the same comment were annotated by non-overlapping sets of annotators. We also include model-generated moderator comments for 50 randomly selected user comments on regroom which were not moderated by experts (`nomod`), to investigate if generated moderation is still beneficial. Finally, we include model-generated comments for 150 randomly selected comments from the User Moderation dataset (`usermod`).

Each sample was annotated by five annotators. For each annotation instance, the annotator was presented with two texts: a preceding comment (i.e., a user comment in the platform discussion) for context, and a reply comment (i.e., either the human or generated moderation). Annotators were asked to evaluate the reply based on several criteria,[5] aimed to assess various aspects of the moderator comment. Additionally, they assessed whether the reply addressed issues in the preceding comment and evaluated overall moderation quality and specific functions like improving discussion quality and broadening the discussion.

The annotation layers included (a short form is in parentheses, each annotated on a Likert scale: strongly disagree → strongly agree):

Fitting (*fitting*): Whether the moderator reply made sense as a response to the preceding comment.

Actions Clarity (*actions*): Clarity of actions suggested by the moderator reply.

Preceding Comment Issues (*issues*): Identification of issues in the preceding comment.

Issue Spotting (*spotted*): Whether the moderator reply identified issues in the preceding comment.

Addressing Issues (*addressed*): How well the reply addressed identified issues.

Properties of the Moderator Comment: Evaluated on *neutrality* (neutral moderator opinion on topic and comment), *clarity* (plain language, simple, clear, avoid overwhelming the user), *curiosity* (model a spirit of inquiry and a desire to learn from and understand user's experience and views), *bias* (comment shows bias towards the user, stereotypes or prejudices), and *encouragement* (welcoming, encouraging and acknowledging; avoiding evaluative and/or condescending responses).

Overall Score (*score*): Overall evaluation of the reply as a moderator comment: very poor → very good.

Moderator Functions Addressed: functions of the moderation reply, i.e., improving quality (*quality*) and broadening discussion (*broadening*).

A sample from the annotation form is provided in Fig. B.3. The guidelines were refined iteratively through multiple rounds of annotation and feedback, including a pilot study on Prolific.

4.2 Moderation Preference Annotation

We additionally implement a moderation preference task, in which participants evaluate and determine a preferred moderator comment based on the properties and functions of moderation. Annotators were shown three texts: a preceding comment and two moderation (reply) comments to choose from. This is used to evaluate professional vs. generated moderation and one generation approach against another. Each evaluation consists of 100 sample pairs, each annotated by three annotators. They are asked to choose their preferred reply, with the options being: both, reply 1, reply2, or neither[6].

We release all annotations in a non-aggregated format along with annotators' socio-demographic variables using a unique, anonymous identifier.[7] An option to skip triggering content was provided, and one annotator utilised this option. In total, approximately 2,000 annotations were collected for the first annotation study, and 600 preference annotations were collected for the second.

---

[5] Guidelines provided in Figs. A.1 and A.7.

[6] A sample of the annotation form is in Fig. B.5

[7] See Tbl. D.1 for a summary of annotators and costs.

5 Models & Instructions

For results to be more reproducible, we use an open source model for our experiments (LLaMA-3, the newest model in the LLaMA family of models (Touvron et al., 2023)). We use nested 4bit quantization (Dettmers and Zettlemoyer, 2023) and top-p (nucleus) sampling with the default parameters from the original LLaMA codebase. Experiments are done with two model sizes 8B and 70B and the instruction fine-tuned version is used.

5.1 Generation

Using moderation guidelines, we develop instruction sets for **generation prompts**, with a moderator persona—derived from these guidelines—serving as the system prompt. Additional experiments explore variations on these instructions (All instructions are in Appendix C).

In further experiments, we use a sample of the professional moderation comments along with the instructions for **supervised fine-tuning (SFT)** of the model. We exclude the samples used for annotation/testing, this totals 340 samples. We use the small 8B model and utilize QLoRA (Dettmers et al., 2023) for memory-efficient fine-tuning.

5.2 Automatic Evaluation

We prompt LLaMA based on G-Eval (Liu et al., 2023) to generate an evaluation of the moderation text. We experiment with different instructions based on the moderation guidelines, the annotation forms and answers about discussions/moderation collected from the annotators. We conduct experiments with both 70B and 8B models, extract scores from the generated outputs, and calculate the score as the average over 3 different evaluation outputs form the model.

6 Results & Analysis

Analyzing and evaluating moderation is challenging due to its subjective and ambiguous nature. We use human evaluation as foundation (Zeng et al., 2023) and complement it with LLM-based automatic evaluation to streamline further experiments.

6.1 Human Evaluation of Properties

Figure 3 shows the average scores obtained through the annotation for professional and generated moderation. Overall **professional moderation** (`regroom prof`) is scored on average acceptable to good. Moderation properties like neutrality, clarity, curiosity, and encouragement are especially highly scored, indicating they are essential and pronounced in the moderation comments, both professional and generated. As for the functions, moderation aiming at improving the quality of the comment is slightly less prevalent than broadening the discussions for professional moderation but is much more prevalent for the generated.

Focusing on the annotation and the **evaluation of the generated** moderation, we see that it consistently and significantly scores higher than professional moderation for the overall score and for all other properties except for bias (where lower is better). Interestingly, generated comments are scored as very highly fitting, even more so than the professional ones. The average length of the generated moderation text, at 361, is significantly more than the average length of the professional moderation text at 263. This could have also played a role in making the generated moderation sound more detailed and as such scored higher by annotators.

Comments that received no moderation (`regroom nomod`) are scored as having `issues` slightly higher than the moderated ones. The generated moderation thereof is scored high overall and specifically on spotting and addressing those issues. The scores are also high for `usermod gen` although the guidelines used for prompting were not developed for this dataset. This shows potential for the results to generalize to other datasets.

The agreement among annotators is measured using Weighted Krippendorff's alpha (Artstein and Poesio, 2008; Antoine et al., 2014) after normalizing the values and turning them into binary based on a 0.5 threshold (cf. Table D.2). The agreement is low (at 0.159) as expected in such subjective and ambiguous tasks (Falk et al., 2024). As recommended by recent work (Basile et al., 2021; Fleisig et al., 2024; Uma et al., 2021; Plank, 2022), we do not attempt to artificially force a higher agreement. Rather, we analyze the actual data to better understand the task with its complexity 6.2. We also incorporate annotator personas in the (subjective) automatic evaluation 6.4.1.

We further calculate Krippendorff's α for the annotations of `regroom prof` (=0.125) and `regroom mod` (=0.056). This might indicate that while almost all annotators score generated moderation higher on average (details in Table D.1), they have more disagreement on the scoring. We look into this in more detail in our analysis.

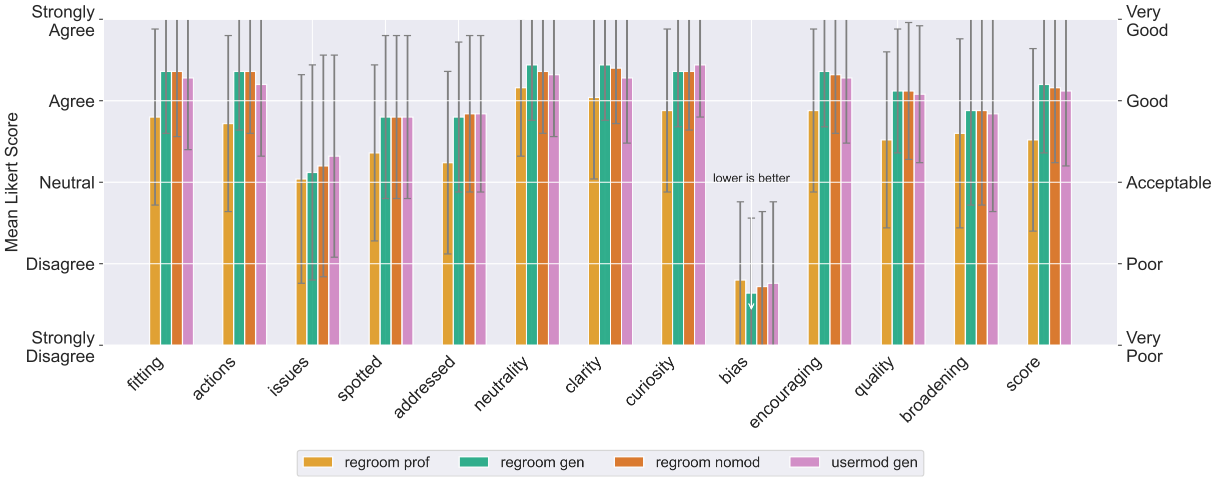


Figure 3: **Human Evaluation of Properties**. Generated moderation (regroom gen) is scored better than professional (regroom prof) on all properties ($p << 0.05$)

generated, highest-score sample: 1.0 Comment: As a visually impaired person I DO NOT believe kiosks access would be beneficial. In fact, I suspect that the plan may ’backfire’, making airport access more difficult. Not being able to read airport signage, and therefore requiring "meet and assist" assistance to my designated gate, ... Reply: Thank you for sharing your personal experience as a visually impaired person. Your concerns about kiosks potentially reducing human assistance are valid. Can others who require ’meet and assist’ services share their thoughts on how airports can balance accessibility with the need for human support?
professional, highest-score sample, score: 0.9 Comment: A similar problem, that should also be addressed, along with the peanut allergy problem, is the case of allowing small domestic pets in the cabin of a aircraft. People with allergic sensitivity to dog and cat dander are in danger of dying from an acute asthma attack... Reply: Thank you for your comment, mithrandir. What type of action would be effective; would peanut free zones be enough? (or pet free zones) Could airlines make certain flights peanut free upon prior request, and use clean planes for those flights?
generated, lowest-score sample, score: 0.5 Comment: The only way that game of thrones would survive is if people would may $29 an episode. Which won’t happen bz people would pirate it. Actually, currently GoT has around 5 million (legal) viewers, and a $5 million/episode budget, so that would be around $1 per episode. And that’s assming that literally ALL pirates are dishonest about only pirating it becuse they can’t access HBO.. Reply: Fascinating analysis! Can you elaborate on what you think would be the main factors driving people to pirate Game of Thrones instead of paying for it? Would it be mainly due to accessibility issues or something else? Additionally, do you think the $1 per episode estimate would be a viable and appealing option for viewers?
professional, lowest-score sample, score: 0.2 Comment: If you have a problem, it is up to you to have the solution. I say "if" because the vast majority of supposed food allergies are BS - the paranoid imaginings of overly anxious parents. Has ANYONE ever died from a food allergy on a plane?.. Reply: Any data that you might have about the prevalence or severity of peanut allergies would be useful to the DOT.

Table 1: Samples of Professional & Generated Moderation.

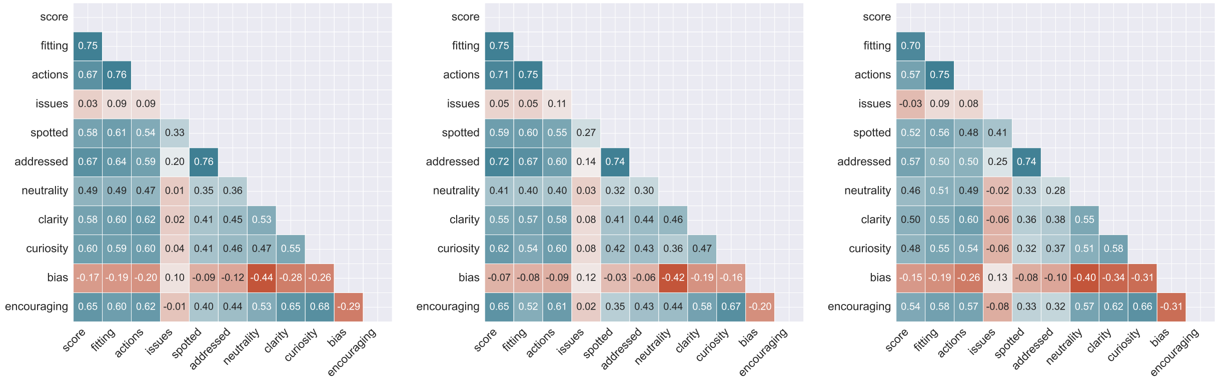
Some examples with highest vs. lowest average scores across the annotators are shown in Table 1. The top generated moderation is for a comment that was not originally moderated and we can see that the generated moderation is beneficial for broadening the discussion (score 0.95). The lowest generated moderation example has low scoring also for addressing issues in the comment (at 0.4) and it seems indeed to ask questions that show no commonsense knowledge (*“Can you elaborate on what you think would be the main factors driving people to pirate Game of Thrones instead of paying for it“*). The lowest-scored professional moderator comment is scored low on a variety of properties but lowest for curiosity: 1.0 and encouraging: 1.5. The previous generated example is on the other hand still scored high on encouraging at 0.75 and this might be contributing to the highest scores for the generated moderation where the lowest average scores are at 0.5.

6.2 Analysis of Moderation Properties

How do moderation properties correlate with each other and with the score?

We conduct an analysis to further investigate which moderation properties play a role in the annotators assigning higher scores to the moderation comments. Looking at the correlation between the various annotation scores for all samples shown in Fig. 4a, we can see that the score has the highest correlation with the reply being fitting. Other properties follow closely, “actions are clear”, “addressed issues” and “encouraging”.

Interestingly, the presence of “issues” in the user comment is not correlated with the score of the moderation, suggesting a user’s comment having issues or not does not relate much to how the annotators perceive the moderation comment on it.



(a) `all` (b) `regroom prof` (c) `regroom gen`

Figure 4: **Pearson Correlation** between Annotation Layers. High correlation of overall score with fitting

This follows in line with previous findings (Falk et al., 2021) showing that lower argument quality, i.e. issues in the comment, is not clearly correlated with more professional moderation.

Bias is negatively correlated with all moderation properties and affects how the annotators score a comment, but even more so, it is negatively correlated with the perceived “curiosity” and “clarity”. In our annotation scheme, neutrality is related to the topic and comment, whereas bias is related to the comment-owner and prejudices in general. From the strong correlation seen here, it seems hard to disentangle one from the other, and it might be that a moderator being perceived as not neutral on the topic or a user comment might also be perceived as being biased against the comment owner or the way the user thinks about the topic.

How does generated moderation differ than professional moderation?

We can see from the correlation scores for `regroom prof` in Fig. 4b that the score of professional moderation has higher correlation with spotting “issues” and “addressing” those issues in comparison to generated moderation (`regroom gen`) shown in Fig. 4c. This suggests that the score for professional moderation is more correlated with deeper properties, e.g., spotting and addressing issues, than just surface-level ones that deal with form, e.g., clarity and curiosity. The correlation between spotting an issue and addressing it is similar for both professional and generated though. We can also see that “addressing” the issues has higher correlation in professional moderation with the reply being “fitting” and “actions” being clear than is the case in generated moderation. Other properties that have higher correlation with the score for professional moderation are curiosity and clarity. Whereas this is the case for generated moderation in neutrality (correlation for the other subsets of the dataset are in Fig. D.4).

6.3 Human Preference Evaluation

Is generated moderation clearly better? Looking at the examples in Table 1 and other samples, it seems that the annotators might be rating the generated moderation high based on surface-level features. To further investigate this, we use pairwise comparisons, which have been shown to improve annotator judgment of generated text (Karpinska et al., 2021) and can be more reliable (Kiritchenko and Mohammad, 2017). We conduct a second annotation study on pairs of comments: one written by a professional moderator and the other generated by a model. The results in Table 2.[8] show that indeed, when the annotator is presented with both professional and generated moderation, they do not have high preference for generated moderation over professional one, but they still prefer it as often. The annotators also almost never find both bad (Individual annotator preferences are in Fig. D.2).

It is important to mention here that the preference evaluation does not take away from the previous results (especially with the small preference difference here). Both evaluations are important and complementary as they show that both generated and professional moderation have desirable properties. Even if the judgment might be influenced by surface-level properties in the previous

---

[8] An annotator who preferred the generated moderation for 98 out of 100 samples was discarded. This is notable because the choices were randomized, implying the annotator may have focused on surface-level features and thus consistently favored the generated moderation.

results, this is actually important because it also contributes to how users perceive the moderation. This further shows that it is important to not limit the evaluation to a single simple evaluation especially when considering complex tasks.

gen	41.0%
prof	43.0%
both good	15.0%
both bad	01.0%

Table 2: Human Preference for **Professional vs. Generated** Moderation. Professional is slightly preferred ($p = 0.03 < 0.05$).

Does fine-tuning help? Given these results, it seems that both professional and generated moderation have desirable properties that appeal to users. We therefore experiment with training LLaMA-8B on a small number (340) of `regroom prof` training data (not part of the test set that was used in annotations and evaluation). We conduct another human preference evaluation using the SFT fine-tuned model and the original model to generate moderation comments. The results in Table 3 show that the samples generated with the original model are actually preferred over the ones from the fine-tuned model (The individual preferences of the annotators are in Fig. D.3). This could be due to the small amount of data available for training.

original	49.7%
fine-tuned	15.7%
both good	24.3%
both bad	10.3%

Table 3: Human Preference for **original vs. fine-tuned** (LLaMA-8B). Original is preferred ($p << 0.05$).

6.4 Automatic Evaluation

6.4.1 How to estimate human evaluation?

Conducting further experiments on moderation generation has human evaluation as a bottleneck. As such, we turn to using LLMs for automatic evaluation. The focus here is on the evaluation of generated moderation, one at a time (as opposed to preference-based evaluation) as this allows us to streamline further experiments. We adapt the generation instructions for this purpose. Note that those were adapted from the moderation guidelines and were also adapted for the annotation guidelines. We prompt LLaMA-8B to generate a Likert scale evaluation of the moderation reply comment and extract scores based on the generated answer.

Table 4 shows the evaluation scores obtained from humans vs. LLMs (*auto-eval*) for the annotation sample. Mean Absolute Error MAE in reference to the scores of the annotators is in Table 5 (further metrics in Tables E.3, E.4 and E.2). We can see that LLMs are prone to evaluating all moderation higher than the annotators did on average. The scores are more in-line with the annotators especially for the generated moderation of regroom, but they are still much higher than the humans' for generated moderation of usermod. This indicates that this is not a consistent pattern in the evaluation of generated vs. professional moderation. The MAE prediction error for Auto-eval is always lower than that of a naive (*mean*) baseline i.e. the average of the scores. This analysis shows that LLMs, prompted with instructions based on moderation guidelines, can be used to estimate and scale human evaluation.

	Auto-eval	Human
all	$0.799_{\pm 0.12}$	$0.75_{\pm 0.25}$
regroom prof	$0.714_{\pm 0.13}$	$0.63_{\pm 0.28}$
regroom gen	$0.806_{\pm 0.09}$	$0.80_{\pm 0.21}$
regroom nomod	$0.830_{\pm 0.13}$	$0.79_{\pm 0.23}$
usermod gen	$0.841_{\pm 0.12}$	$0.78_{\pm 0.23}$

Table 4: **Auto-Eval Average Score (LLaMA-8B)**

Dataset	Mean	Auto-eval
all	0.260	0.134
regroom prof	0.174	0.171
regroom gen	0.284	0.098
regroom nomod	0.286	0.148
usermod gen	0.299	0.128

Table 5: **MAE Prediction Errors of Auto-Eval** (further metrics in Tables E.3 and E.4)

How to approximate evaluation for each annotator?

Given the subjectivity of the task, it is interesting to check how well the human evaluation can be approximated for each annotator based on their subjective expectations. To this end, we use the answers collected from the annotators in the intro questionnaire (Sec. 4) e.g. "What makes a comment or contribution valuable?„ in the system prompt (Sec. C.2), as a kind of persona prompt (Deshpande et al., 2023), for the generation of evaluation answers. To further parallel the annotation setup and make the evaluation more robust, we replace the evaluation prompt, that was based on the generation one and, as such, might bias the model, with a textual representation of the annotation form and including an instruction to "read and think" about the annotation questions before evaluating the reply. We then calculate the prediction error based on the individual annotator whose answers were used for evaluating the sample. The results in Table 6 show that this approach indeed helps us approximate the individual answers not only compared to a naive baseline but also and especially compared to using previous prompting that does not include any annotators information. This allows for a more diverse and fine-grained estimation of the annotators scores.

Dataset	Mean	LLaMA-8B	+intro
all	0.309	0.612	0.190
regroom prof	0.261	0.557	0.234
regroom gen	0.323	0.606	0.164
regroom nomod	0.325	0.634	0.182
usermod gen	0.327	0.644	0.181

Table 6: **MAE of Auto-Eval Per Annotator.**
Using answers from annotators to estimate their scores

6.5 Can the Generation be Improved?

After alleviating the human evaluation bottleneck using automatic evaluation, we conduct further generation experiments. We adapt detailed information about moderation properties from the annotation form and use them as instructions for the generation (instructions in Sec. C.1). We also include Chain-of-Thought (Wei et al., 2022) (CoT) as detailed generation steps without examples in our generation (instructions in Sec. C.1). We use the basic LLaMA-8B auto-eval from our previous experiments to estimate the performance as shown in Table 7. We see that adding detailed moderation properties (`+Detailed Prop`) leads to better performance. Introducing CoT leads to better performance than that, though combining both does not lead to significant improvement. These findings, together with the fine-tuning results in Table D.3, suggest that prompt-based methods may be a better choice for generation in such a low-resource task and domain.

	LLaMA-70B	+Detailed Prop	+CoT	+Both
regroom gen	$0.806_{\pm 0.09}$	$0.864_{\pm 0.10}$	$0.881_{\pm 0.09}$	$0.879_{\pm 0.10}$
regroom nomod	$0.830_{\pm 0.13}$	$0.870_{\pm 0.11}$	$0.892_{\pm 0.12}$	$0.868_{\pm 0.10}$
usermod gen	$0.841_{\pm 0.12}$	$0.866_{\pm 0.11}$	$0.880_{\pm 0.10}$	$0.884_{\pm 0.10}$

Table 7: Auto-Eval Average Score for **Different Generation Instructions**

7 Conclusion

This work made first steps toward the automatic generation of positive moderation, which facilitates the discussion by encouraging positive user behaviour. Our work showcases the potential for using LLMs to achieve this purpose in real-world settings. We envision scenarios where human moderators receive initial suggestions that, though requiring some adjustment to better address issues in user comments, offer a valuable starting point as demonstrated in our experiments. By leveraging the capabilities of LLMs, we can assist in fostering more thoughtful and inclusive discussions online, and aim to bridge the divide in increasingly polarized interactions online.

8 Limitations

At the level of evaluation scope, an obvious limitation is the fact that we use only two datasets, and that additionally they are both on English. Those were, however, the only suitable candidates for our focus on positive moderation. The regroomm dataset has limited data on specific domains in a deliberative setup. While the usermod covers a wide range of domains in an online discussion format. As such, the combination of both provide a good basis to make first steps towards studying this phenomena. Future research should explore a more diverse range of datasets and include multiple languages to validate and extend our results.

The set of features and measures we annotated and evaluated, such as neutrality, clarity, and curiosity, are not necessarily exhaustive. While they provide a solid foundation for understanding the properties of positive moderation, there may be other relevant features and measures that were not considered in this study. Further research should aim to identify and incorporate additional features to provide a more comprehensive evaluation of positive moderation. Another limitations is considering only the preceding comment which might not provide the full context needed for the generation of better moderation as well as for the evaluation of that moderation. Limiting the context was a simplification step taken which was shown to be acceptable by pilot annotation studies and evaluations. Related work (Falk et al., 2021) also showed that including more context does not significantly improve a related task, the prediction of whether a comment is moderated.

The assessment of apt and useful moderation, particularly positive moderation, is inherently a subjective task. This subjectivity poses a challenge in achieving high inter-annotator agreement, which can limit the generalizability of our findings. The nature of the task means that complete consensus is unlikely, a variability that must be considered when interpreting our results and deriving broader conclusions. Our dataset in its non-aggregated fine-grained annotations can be a basis for further analysis. Our auto-eval based on annotator persona could also assist in this.

9 Ethics & Potential Risks

We are aware that text generated with LLMs is prone to encode bias and stereotypes, and when it comes to LLM-supported moderation, special attention should be made so its application does not turn harmful. This work is, however, the first step towards the understanding, generation and evaluation of positive moderation and we believe that a thorough understanding of the properties of text generated by LLMs and of the perceptions of the users is necessary to develop effective tools to assist professional moderation. In addition, while LLMs can assist in scaling moderation efforts, human moderators remain necessary as they play a critical role in interpreting nuanced contexts and making complex ethical decisions in various discussion topics that automated systems may not handle effectively. As mentioned in the Conclusion (Sec. 7, our vision is having LLMs assist professional moderators to get the best of both worlds.

Beside the implicit biases of LLMs, explicit biases could also be introduced in the instructions with or without intention. Those in turn could results in harmful effects e.g. suppressing or not fostering certain opinions. Therefore, we propose that it is important to make the instructions used for moderation generation transparent similar to how community guidelines and moderation guidelines should be transparent.

While the methodological design presented here is aimed to determine an evaluation of the properties present in generated moderation, generation approaches may occasionally misinterpret the context or produce responses that are not aligned with the platform's guidelines. This can lead to ineffective or even harmful moderation, impacting the quality of discourse and user experience. This aspect should be investigated to determine the impact on moderation and discourse quality.

Finally, we are aware that using LLMs for moderation involves processing large amounts of user-generated content, which raises concerns about privacy and data security. It is essential to ensure that the data used for training and the content generated for moderation comply with privacy regulations and respect user confidentiality.

Our dataset is collected from annotators on prolific after they agree to an extensive consent form (Fig. B.1). As such, no ethical review was deemed necessary because no personal or controversial data was collected from the annotators. The annotators had also the chance to skip any content they do not feel comfortable annotating (see the last page in Annotation Guidelines Fig A.7 and Fig A.1).

Acknowledgements

We acknowledge funding by the Bundesministerium für Bildung und Forschung (BMBF) through the project E-DELIB.

References

Jean-Yves Antoine, Jeanne Villaneau, and Anaïs Lefeuvre. 2014. Weighted Krippendorff's alpha is a more reliable metrics for multi-coders ordinal annotations: experimental studies on emotion, opinion and coreference annotation. In *Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics*, pages 550–559, Gothenburg, Sweden. Association for Computational Linguistics.

Ron Artstein and Massimo Poesio. 2008. Survey article: Inter-coder agreement for computational linguistics. *Computational Linguistics*, 34(4):555–596.

Valerio Basile, Michael Fell, Tommaso Fornaciari, Dirk Hovy, Silviu Paun, Barbara Plank, Massimo Poesio, and Alexandra Uma. 2021. We need to consider disagreement in evaluation. In *Proceedings of the 1st Workshop on Benchmarking: Past, Present and Future*, pages 15–21, Online. Association for Computational Linguistics.

Laura Black, Stephanie Burkhalter, John Gastil, and J. Stromer-Galley. 2011. *Methods for analyzing and measuring group deliberation*, pages 323–345. Routledge.

Shelley Boulianne, Kaiping Chen, and David Kahane. 2020. Mobilizing mini-publics: The causal impact of deliberation on civic engagement using panel data. *Politics*, 40(4):460–476.

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. In *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*.

Hyundong Cho, Shuai Liu, Taiwei Shi, Darpan Jain, Basem Rizk, Yuyang Huang, Zixun Lu, Nuan Wen, Jonathan Gratch, Emilio Ferrara, and Jonathan May. 2024. Can language model moderators improve the health of online discourse? In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 7471–7489, Mexico City, Mexico. Association for Computational Linguistics.

Yi-Ling Chung, Serra Sinem Tekiroğlu, and Marco Guerini. 2021. Towards knowledge-grounded counter narrative generation for hate speech. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 899–914, Online. Association for Computational Linguistics.

Dhanielly PR de Lima, Marco A Gerosa, Tayana U Conte, and José Francisco de M. Netto. 2019. What to expect, and how to improve online discussion forums: the instructors' perspective. *Journal of Internet Services and Applications*, 10:1–15.

Davy Deng, Tim Rogers, and John A Naslund. 2023. The role of moderators in facilitating and encouraging peer-to-peer support in an online mental health community: a qualitative exploratory study. *Journal of Technology in Behavioral Science*, 8(2):128–139.

Ameet Deshpande, Vishvak Murahari, Tanmay Rajpurohit, Ashwin Kalyan, and Karthik Narasimhan. 2023. Toxicity in chatgpt: Analyzing persona-assigned language models. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 1236–1270, Singapore. Association for Computational Linguistics.

Tim Dettmers, Artidoro Pagnoni, Ari Holtzman, and Luke Zettlemoyer. 2023. Qlora: Efficient finetuning of quantized llms. In *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*.

Tim Dettmers and Luke Zettlemoyer. 2023. The case for 4-bit precision: k-bit inference scaling laws. In *International Conference on Machine Learning, ICML 2023, 23-29 July 2023, Honolulu, Hawaii, USA*, volume 202 of *Proceedings of Machine Learning Research*, pages 7750–7774. PMLR.

Kara N. Dillard. 2013. Envisioning the role of facilitation in public deliberation. *Journal of Applied Communication Research*, 41(3):217–235.

Cornell eRulemaking Initiative et al. 2017. Ceri (cornell e-rulemaking) moderator protocol.

Katharina Esau, Dennis Friess, and Christiane Eilders. 2017. Design matters! an empirical analysis of online deliberation on different news platforms. *Policy & Internet*, 9(3):321–342.

Neele Falk, Iman Jundi, Eva Maria Vecchi, and Gabriella Lapesa. 2021. Predicting moderation of deliberative arguments: Is argument quality the key? In *Proceedings of the 8th Workshop on Argument Mining*, pages 133–141, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Neele Falk and Gabriella Lapesa. 2023. Bridging argument quality and deliberative quality annotations with adapters. In *Findings of the Association for Computational Linguistics: EACL 2023*, pages 2469–2488, Dubrovnik, Croatia. Association for Computational Linguistics.

Neele Falk, Eva Vecchi, Iman Jundi, and Gabriella Lapesa. 2024. Moderation in the wild: Investigating user-driven moderation in online discussions. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 992–1013, St. Julian's, Malta. Association for Computational Linguistics.

Eve Fleisig, Su Lin Blodgett, Dan Klein, and Zeerak Talat. 2024. The perspectivist paradigm shift: Assumptions and challenges of capturing human labels. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 2279–2292, Mexico City, Mexico. Association for Computational Linguistics.

Jinlan Fu, See-Kiong Ng, Zhengbao Jiang, and Pengfei Liu. 2023. Gptscore: Evaluate as you desire. *CoRR*, abs/2302.04166.

Lodewijk Gelauff, Liubov Nikolenko, Sukolsak Sakshuwong, James Fishkin, Ashish Goel, Kamesh Munagala, and Alice Siu. 2023. Achieving parity with human moderators: A self-moderating platform for online deliberation 1. In *The Routledge Handbook of Collective Intelligence for Democracy and Governance*, pages 202–221. Routledge.

Amey Hengle, Aswini Padhi, Sahajpreet Singh, Anil Bandhakavi, Md Shad Akhtar, and Tanmoy Chakraborty. 2024. Intent-conditioned and non-toxic counterspeech generation using multi-task instruction tuning with RLAIF. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 6716–6733, Mexico City, Mexico. Association for Computational Linguistics.

Judith E Innes. 2004. Consensus building: Clarifications for the critics. *Planning theory*, 3(1):5–20.

Marzena Karpinska, Nader Akoury, and Mohit Iyyer. 2021. The perils of using Mechanical Turk to evaluate open-ended text generation. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 1265–1285, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Mohammad Khosravani, Chenyang Huang, and Amine Trabelsi. 2024. Enhancing argument summarization: Prioritizing exhaustiveness in key point generation and introducing an automatic coverage evaluation metric. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 8205–8217, Mexico City, Mexico. Association for Computational Linguistics.

Svetlana Kiritchenko and Saif Mohammad. 2017. Best-worst scaling more reliable than rating scales: A case study on sentiment intensity annotation. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 465–470, Vancouver, Canada. Association for Computational Linguistics.

Hayato Kobayashi, Hiroaki Taguchi, Yoshimune Tabuchi, Chahine Koleejan, Ken Kobayashi, Soichiro Fujita, Kazuma Murao, Takeshi Masuyama, Taichi Yatsuka, Manabu Okumura, and Satoshi Sekine. 2021. A case study of in-house competition for ranking constructive comments in a news service. In *Proceedings of the Ninth International Workshop on Natural Language Processing for Social Media*, pages 24–35, Online. Association for Computational Linguistics.

Mahi Kolla, Siddharth Salunkhe, Eshwar Chandrasekharan, and Koustuv Saha. 2024. Llm-mod: Can large language models assist content moderation? In *Extended Abstracts of the CHI Conference on Human Factors in Computing Systems, CHI EA 2024, Honolulu, HI, USA, May 11-16, 2024*, pages 217:1–217:8. ACM.

Metka Kuhar, Matej Krmelj, and Gregor Petrič. 2019. The impact of facilitation on the quality of deliberation and attitude change. *Small Group Research*, 50(5):623–653.

Deepak Kumar, Yousef AbuHashem, and Zakir Durumeric. 2024. Watch your language: Investigating content moderation with large language models. In *Proceedings of the Eighteenth International AAAI Conference on Web and Social Media, ICWSM 2024, Buffalo, New York, USA, June 3-6, 2024*, pages 865–878. AAAI Press.

Yang Liu, Dan Iter, Yichong Xu, Shuohang Wang, Ruochen Xu, and Chenguang Zhu. 2023. G-eval: NLG evaluation using gpt-4 with better human alignment. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 2511–2522, Singapore. Association for Computational Linguistics.

Jane Mansbridge, Janette Hartz-Karp, Matthew Amengual, and John Gastil. 2006. Norms of deliberation: An inductive study. *Journal of Deliberative Democracy*, 2(1).

Sankha Subhra Mullick, Mohan Bhambhani, Suhit Sinha, Akshat Mathur, Somya Gupta, and Jidnya Shah. 2023. Content moderation for evolving policies using binary question answering. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 5: Industry Track)*, pages 561–573, Toronto, Canada. Association for Computational Linguistics.

Joonsuk Park, Sally Klingel, Claire Cardie, Mary Newhart, Cynthia Farina, and Joan-Josep Vallbé. 2012. Facilitative moderation for online participation in erulemaking. In *Proceedings of the 13th Annual International Conference on Digital Government Research*, dg.o '12, page 173–182, New York, NY, USA. Association for Computing Machinery.

Barbara Plank. 2022. The "problem" of human label variation: On ground truth in data, modeling and evaluation. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 10671–10682, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Mattia Samory. 2021. On positive moderation decisions. In *Proceedings of the Fifteenth International AAAI Conference on Web and Social Media, ICWSM 2021, held virtually, June 7-10, 2021*, pages 585–596. AAAI Press.

Omar Shaikh, Kristina Gligoric, Ashna Khetan, Matthias Gerstgrasser, Diyi Yang, and Dan Jurafsky. 2024. Grounding gaps in language model generations. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 6279–6296, Mexico City, Mexico. Association for Computational Linguistics.

Tanner Skousen, Hani Safadi, Colleen Young, Elena Karahanna, Sami Safadi, and Fouad Chebib. 2020. Successful moderation in online patient communities: Inductive case study. *J Med Internet Res*, 22(3):e15983.

Christopher T. Small, Ivan Vendrov, Esin Durmus, Hadjar Homaei, Elizabeth Barry, Julien Cornebise, Ted Suzman, Deep Ganguli, and Colin Megill. 2023. Opportunities and risks of llms for scalable deliberation with polis. *Preprint*, arXiv:2306.11932.

Kim Strandberg, Staffan Himmelroos, and Kimmo Grönlund. 2017. Do discussions in like-minded groups necessarily lead to more extreme opinions? deliberative democracy and group polarization. *International Political Science Review*, 40(1):41–57.

Andrej Švec, Matúš Pikuliak, Marián Šimko, and Mária Bieliková. 2018. Improving moderation of online discussions via interpretable neural models. In *Proceedings of the 2nd Workshop on Abusive Language Online (ALW2)*, pages 60–65, Brussels, Belgium. Association for Computational Linguistics.

Chenhao Tan, Vlad Niculae, Cristian Danescu-Niculescu-Mizil, and Lillian Lee. 2016. Winning arguments: Interaction dynamics and persuasion strategies in good-faith online discussions. In *Proceedings of the 25th International Conference on World Wide Web, WWW 2016, Montreal, Canada, April 11 - 15, 2016*, pages 613–624. ACM.

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. *Preprint*, arXiv:2302.13971.

M. Trénel. 2009. Facilitation and inclusive deliberation. In T. Davies and S. P. Gangadharan, editors, *Online Deliberation: Design, Research and Practice*, pages 253–257. CSLI Publications.

Alexandra Uma, Tommaso Fornaciari, Dirk Hovy, Silviu Paun, Barbara Plank, and Massimo Poesio. 2021. Learning from disagreement: A survey. *J. Artif. Intell. Res.*, 72:1385–1470.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed H. Chi, Quoc V. Le, and Denny Zhou. 2022. Chain-of-thought prompting elicits reasoning in large language models. In *Advances in Neural Information Processing Systems 35: Annual Conference on Neural Information Processing Systems 2022, NeurIPS 2022, New Orleans, LA, USA, November 28 - December 9, 2022*.

Zhiyuan Zeng, Jiatong Yu, Tianyu Gao, Yu Meng, Tanya Goyal, and Danqi Chen. 2023. Evaluating large language models at evaluating instruction following. *CoRR*, abs/2310.07641.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2020. Bertscore: Evaluating text generation with BERT. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net.

A Annotation Guidelines

Figure A.1: Moderation Property Annotation Guidelines

Guidelines for Moderator Intervention Annotation

Background

Moderation on most platforms is generally carried out by *expert moderators* who are trained specifically for the role and whose contribution in the platform most often is specifically that of moderation: maintaining order, ensuring compliance with guidelines, and facilitating productive discussions. Moderators are guided by specific rules and guidelines that shape their interactions with users. Identifying the important characteristics of effective moderation is crucial for improving platform dynamics and user experience.

Main Goal

The goal of this study is to annotate *moderator comments* and to identify the characteristics and the role [expert] moderation plays within a discussion.

1. Annotate Moderator Comments:
   Systematically annotate moderator comments in various discussions to identify patterns and strategies used by moderators.
2. Reflect on how well the moderation tackles any issues in user comment:
   Identify when moderation is most needed and how effective it is
3. Identify Characteristics of Effective Moderation:
   Determine key traits and actions that contribute to successful moderation, such as neutrality, clarity and empathy.
4. Identify different types of moderator interventions:
   Such as conflict resolution, guideline enforcement, user guidance, and promoting positive interactions.

What is Moderation?[1]

The goal of *moderation* in deliberation and argumentation platforms is to create an environment of informed and thoughtful participation, as well as mentor effective commenting behavior. A moderator moves participants past "voting and venting" behaviors to effectively contribute the information they possess. They also make participants feel that their voices have been heard and that they are part of a forum for [civil] engagement.

Moderators have the role of advocating for the commenting process; as they encourage a "knowledge building community" that supports commenters' access to, participation in, and learning about the process and topic under discussion. Whether the goal of the process is policymaking,

Figure A.2: Moderation Property Annotation Guidelines - Continued

converging perspectives, or arguing one's view, moderation helps commenters to contribute as individuals as well as collaborate with each other.

Expectation of Moderators

1. Neutrality: Expert moderators are strongly encouraged to remain *neutral*, avoiding taking a position on the substance of the discussion, or forming biases or making assumptions about participants' comments. However, users are not restricted to this requirement and comments that do indeed have the role of moderation from a user may (e.g. in the case of clarification comments) or may not (e.g. signaling erred information to another user) have this characteristic.
2. Maintaining the norms: Expert moderators are responsible for maintaining the norms of the platform community and its regulations. Users might mirror this role in subtle ways, such as reminding others of the goal of the discussion or pointing out inappropriate contributions.
3. Choice of wording: Expert moderators are asked to use plain language, calm tones, avoid condescending responses, and limit the number of questions. For example:
    - *a. That clarification is available in several forms on the website http:[...]*
    - *b. DOT has estimated that the benefits of this discussion will outweigh the costs.*
    - *c. This is an interesting suggestion, thanks. Could you provide a little more information on this, and perhaps a link.*

    Again, users are not expected to uphold these standards in their comments, however they may still perform similar contributions to the discussion, with or without a careful choice of wording.

The Data:

RegulationRoom

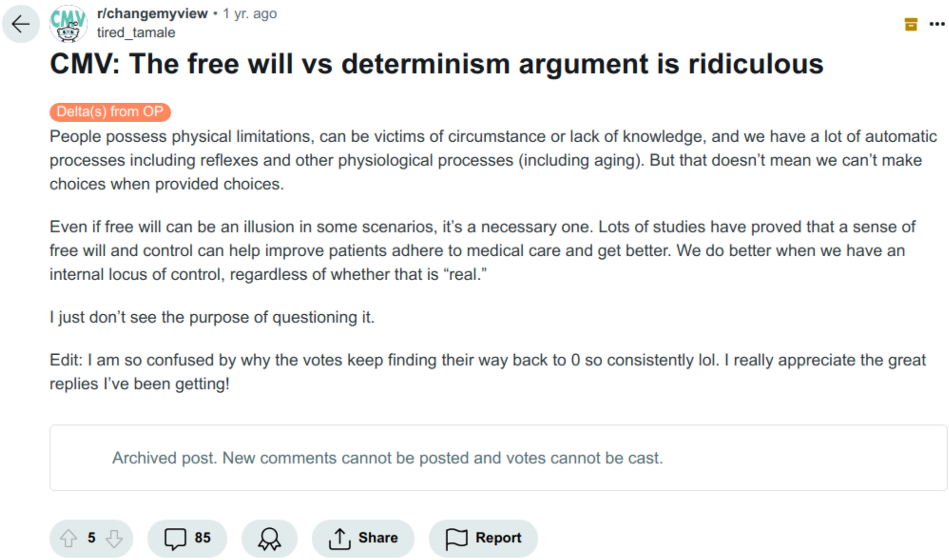
RegulationRoom was developed and used in the setting of notice-and-comment rulemaking by federal agencies. In the complex policy environment of new federal safety and consumer protection regulations, RegulationRoom enabled historically silent stakeholder groups to participate effectively in six actual rulemakings. Those stakeholders participated in discussions amongst each other as normal users of the system whilst a moderator facilitated and supported the discussion. Rulemakers and participants both agreed that the agency got better and more complete information about the problems it was trying to solve, while participants got a better understanding of the process, the competing interests, and the agency's objectives.

Figure A.3: Moderation Property Annotation Guidelines - Continued

Change My View

The data you will be annotating is extracted from the online subreddit entitled Change My View.[2] The platform is dedicated to civil discourse, aimed at promoting productive conversation to resolve differences by understanding others' perspectives.

The format of CMV is as follows. First, a user (original poster, or OP) posts a *view*, defined as a particular way of considering or regarding something, an attitude or opinion, on a specified *topic* issue, and asks the community to "change my view". For example:



r/changemyview • 1 yr. ago
tired_tamale

CMV: The free will vs determinism argument is ridiculous

Delta(s) from OP

People possess physical limitations, can be victims of circumstance or lack of knowledge, and we have a lot of automatic processes including reflexes and other physiological processes (including aging). But that doesn't mean we can't make choices when provided choices.

Even if free will can be an illusion in some scenarios, it's a necessary one. Lots of studies have proved that a sense of free will and control can help improve patients adhere to medical care and get better. We do better when we have an internal locus of control, regardless of whether that is "real."

I just don't see the purpose of questioning it.

Edit: I am so confused by why the votes keep finding their way back to 0 so consistently lol. I really appreciate the great replies I've been getting!

Archived post. New comments cannot be posted and votes cannot be cast.

Users are then able to interact with the OP as comments to argue their perspective in order to change the OP author's view. The interaction between users and OP author may be a simple back-and-forth comment, or may be an extended discussion. At the end of the interaction, if the user's argument has successfully changed the OP's view, the user is awarded a *Delta* (Δ) by the OP author.

Introductory Questions

The following questions are asked to get a general understanding of the views of the annotator

- As a user, what are your goals/objectives when commenting about a topic?
- What do you feel contributes to a good experience for the users/discussion?
- What makes a comment or contribution valuable?
- What makes a comment or contribution of poor quality, unconstructive or detrimental to the discussion?

Figure A.4: Moderation Property Annotation Guidelines - Continued

- As a user, what properties does a helpful moderator comment have?

Annotation Task

The annotator will be shown two texts: the ***preceding comment*** (for example, the OP or a post in the comment thread) and the ***reply comment***. The preceding comment as well as the topic of the OP are provided to the annotator to offer context. The *reply comment* is the comment to which the annotation questions refer. For each reply comment, the annotators are asked a set of questions, described in detail below:

1. **Fitting**
   Is the reply comment fitting and does it make sense as a reply to the preceding comment? (1-5 scale)
2. **Actions clarity**
   Does the reply make it clear what actions need to be done? (1-5 scale)

3. **Preceding comment issues**
   Based on only the preceding comment, do you see some issues in the original comment? (1-5 scale)
   This annotation is ONLY about the preceding comment
4. **Issue spotting**
   Based on the preceding and the reply comment, did the moderator spot the issues in the original comment? (regardless if you spotted them yourself) (1-5 scale)
   This annotation is about BOTH the preceding and reply comment
5. **Addressing issues**
   How well does the reply address any of the above issues?(1-5 scale)

   Properties of the moderator comment
6. **Neutrality**
   Remain Neutral on the topic and on the Comment Substance and Commenter's Viewpoint. The reply shouldn't give away the opinion of the moderator on the topic or comment. (1-5 scale)
7. **Clarity**
   Plain language, simple, clear, avoid overwhelming the user e.g. too many questions (1-5 scale)
8. **Curiosity**
   Moderators should model a spirit of inquiry and a desire to learn from and understand

Figure A.5: Moderation Property Annotation Guidelines - Continued

commenter's experience and views. Try to be interested in the bases upon which each commenter stakes his or her claims and the lines of reasoning that has led each commenter to those particular conclusions. (1-5 scale)

9. **Bias**
   Does the reply show some biases towards the commenter? Are there stereotypes or prejudices? (1-5 scale)
10. **Encouraging**
    Welcoming, encouraging and acknowledging. Avoid Evaluative and/or Condescending Responses (1-5 scale)

11. **Overall score**
    Based on the above, how good is the reply as a moderator comment?
12. Which **moderator function** does the comment address and how well?
    a. Improving quality
    Providing substantive information about about the topic or proposal
    Correcting misstatements or clarifying the intended outcome of the discussion
    Pointing to relevant information in primary documents or other data
    Pointing out characteristics of effective commenting
    Asking users to provide more information, factual details, or data to support their statements
    Asking users to make or consider possible solutions/alternative approaches (1-5 scale)
    b. Broadening discussion
    Encouraging users to consider and engage comments of other users
    Posing a question to the community at large that encourages other users to respond (1-5 scale)
    c. Other moderator function(s)
13. **Further comments**
    free text for any comments the annotator has about this specific example/annotation sample

The first two examples / annotation samples are used to check your comprehension of the questions and if you do not answer them correctly, you will be asked to return the study and you will be compensated £1.50 for your time.

Trigger Warning!

As mentioned in the consent form you agreed to, the texts included in this study are produced in an online debate forum and some topics that are discussed, how they are discussed, and user perspectives may be uncomfortable or sensitive. First, all texts included do not represent the views of the researchers conducting the study. Secondly, we provide the option to avoid having to annotate any

Figure A.6: Moderation Property Annotation Guidelines - Continued

instance that is problematic or uncomfortable for the annotator without penalty of compensation.

To do so, check the corresponding checkbox and answer the annotation questions as outlined below. Note, although you will have provided answers, if you include the following text, your answers to this instance will be automatically discarded and not considered in the study.

All question should be answered as “neither agree nor disagree”

Further comments: (please copy and paste)
This content is triggering for me! I am uncomfortable annotating this text and voluntarily skip this instance. I kindly request not to be penalized for this, as it is not an attempt to avoid annotation responsibilities.

Figure A.7: Moderation Preference Annotation Guidelines

Guidelines for Choosing Preferred Moderation

Background

Moderation on most platforms is generally carried out by *expert moderators* who are trained specifically for the role and whose contribution in the platform most often is specifically that of moderation: maintaining order, ensuring compliance with guidelines, and facilitating productive discussions. Moderators are guided by specific rules and guidelines that shape their interactions with users. Identifying the important characteristics of effective moderation is crucial for improving platform dynamics and user experience.

Main Goal

The goal of this study is to evaluate and determine the *preferred moderator comment* based on the properties and functions of effective moderation.

What is Moderation?[1]

The goal of *moderation* in deliberation and argumentation platforms is to create an environment of informed and thoughtful participation, as well as mentor effective commenting behavior. A moderator moves participants past "voting and venting" behaviors to effectively contribute the information they possess. They also make participants feel that their voices have been heard and that they are part of a forum for [civil] engagement.

Moderators have the role of advocating for the commenting process; as they encourage a "knowledge building community" that supports commenters' access to, participation in, and learning about the process and topic under discussion. Whether the goal of the process is policymaking, converging perspectives, or arguing one's view, moderation helps commenters to contribute as individuals as well as collaborate with each other.

Expectation of Moderators

1. Neutrality: Expert moderators are strongly encouraged to remain *neutral*, avoiding taking a position on the substance of the discussion, or forming biases or making assumptions about participants' comments. However, users are not restricted to this requirement and comments that do indeed have the role of moderation from a user may (e.g. in the case of clarification comments) or may not (e.g. signaling erred information to another user) have this characteristic.
2. Maintaining the norms: Expert moderators are responsible for maintaining the norms of the

Figure A.8: Moderation Property Annotation Guidelines - Continued

platform community and its regulations. Users might mirror this role in subtle ways, such as reminding others of the goal of the discussion or pointing out inappropriate contributions.

3. Choice of wording: Expert moderators are asked to use plain language, calm tones, avoid condescending responses, and limit the number of questions. For example:
   a. That clarification is available in several forms on the website http:[...]
   b. DOT has estimated that the benefits of this discussion will outweigh the costs.
   c. This is an interesting suggestion, thanks. Could you provide a little more information on this, and perhaps a link.
3. Encouraging: Expert moderators are asked to be welcoming, encouraging and acknowledging: It is important to welcome participants by making them feel appreciated and part of the community.

Moderator Functions

Although a big part of moderation involves moderating and eliminating negative aspects of the discussion e.g. policing, our focus is on moderation that promotes positive aspects of the conversation:

a. Improving quality

Providing substantive information about about the topic or proposal
Correcting misstatements or clarifying the intended outcome of the discussion
Pointing to relevant information in primary documents or other data
Pointing out characteristics of effective commenting
Asking users to provide more information, factual details, or data to support their statements
Asking users to make or consider possible solutions/alternative approaches (1-5 scale)

b. Broadening discussion

Encouraging users to consider and engage comments of other users
Posing a question to the community at large that encourages other users to respond (1-5 scale)

c. Other moderator function(s)

Asking open ended questions, asking for personal experiences.

Figure A.9: Moderation Property Annotation Guidelines - Continued

The Data:

RegulationRoom

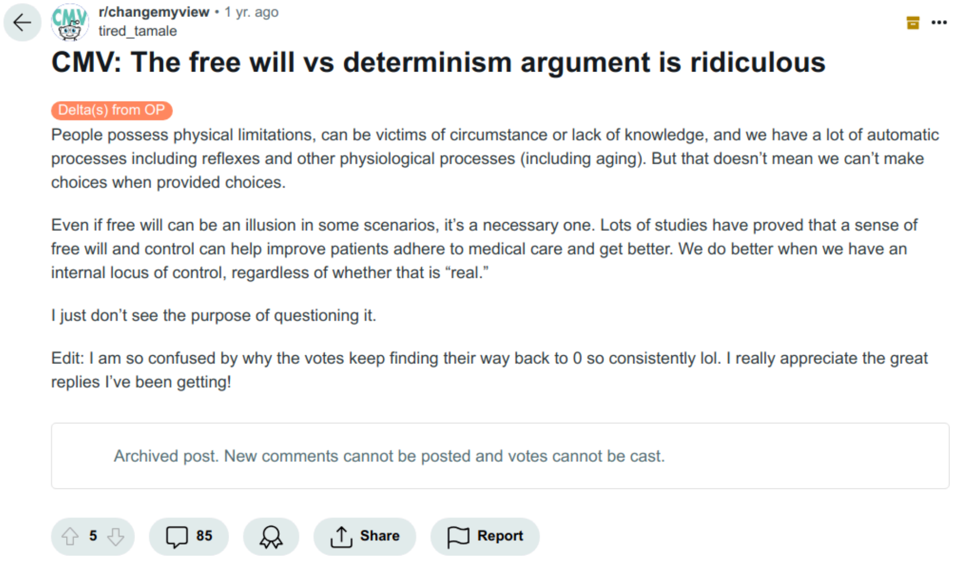
RegulationRoom was developed and used in the setting of notice-and-comment rulemaking by federal agencies. In the complex policy environment of new federal safety and consumer protection regulations, RegulationRoom enabled historically silent stakeholder groups to participate effectively in six actual rulemakings. Those stakeholders participated in discussions amongst each other as normal users of the system whilst a moderator facilitated and supported the discussion. Rulemakers and participants both agreed that the agency got better and more complete information about the problems it was trying to solve, while participants got a better understanding of the process, the competing interests, and the agency's objectives.

Change My View

The data you will be annotating is extracted from the online subreddit entitled Change My View.[2] The platform is dedicated to civil discourse, aimed at promoting productive conversation to resolve differences by understanding others' perspectives.

The format of CMV is as follows. First, a user (original poster, or OP) posts a *view*, defined as a particular way of considering or regarding something, an attitude or opinion, on a specified *topic* issue, and asks the community to "change my view". For example:

Figure A.10: Moderation Property Annotation Guidelines - Continued



r/changemyview • 1 yr. ago

tired_tamale

CMV: The free will vs determinism argument is ridiculous

Delta(s) from OP

People possess physical limitations, can be victims of circumstance or lack of knowledge, and we have a lot of automatic processes including reflexes and other physiological processes (including aging). But that doesn't mean we can't make choices when provided choices.

Even if free will can be an illusion in some scenarios, it's a necessary one. Lots of studies have proved that a sense of free will and control can help improve patients adhere to medical care and get better. We do better when we have an internal locus of control, regardless of whether that is "real."

I just don't see the purpose of questioning it.

Edit: I am so confused by why the votes keep finding their way back to 0 so consistently lol. I really appreciate the great replies I've been getting!

Archived post. New comments cannot be posted and votes cannot be cast.

Users are then able to interact with the OP as comments to argue their perspective in order to change the OP author's view. The interaction between users and OP author may be a simple back-and-forth comment, or may be an extended discussion. At the end of the interaction, if the user's argument has successfully changed the OP's view, the user is awarded a *Delta* (Δ) by the OP author.

Introductory Questions

The following questions are asked to get a general understanding of the views of the annotator

- As a user, what are your goals/objectives when commenting about a topic?
- What do you feel contributes to a good experience for the users/discussion?
- What makes a comment or contribution valuable?
- What makes a comment or contribution of poor quality, unconstructive or detrimental to the discussion?
- As a user, what properties does a helpful moderator comment have?

Annotation Task

The annotator will be shown three texts: the ***preceding comment*** (for example, the OP or a post in the comment thread) and the ***reply comment 1*** and ***reply comment 2***. The preceding comment provides context and is important in determining which reply is better. The options to choose from are:

a) None: none of the replies is good

Figure A.11: Moderation Property Annotation Guidelines - Continued

b) Reply Comment 1 is better (than Reply Comment 2)
c) Reply Comment 2 is better (than Reply Comment 1)
d) Both (both replies are equally good)

Trigger Warning!

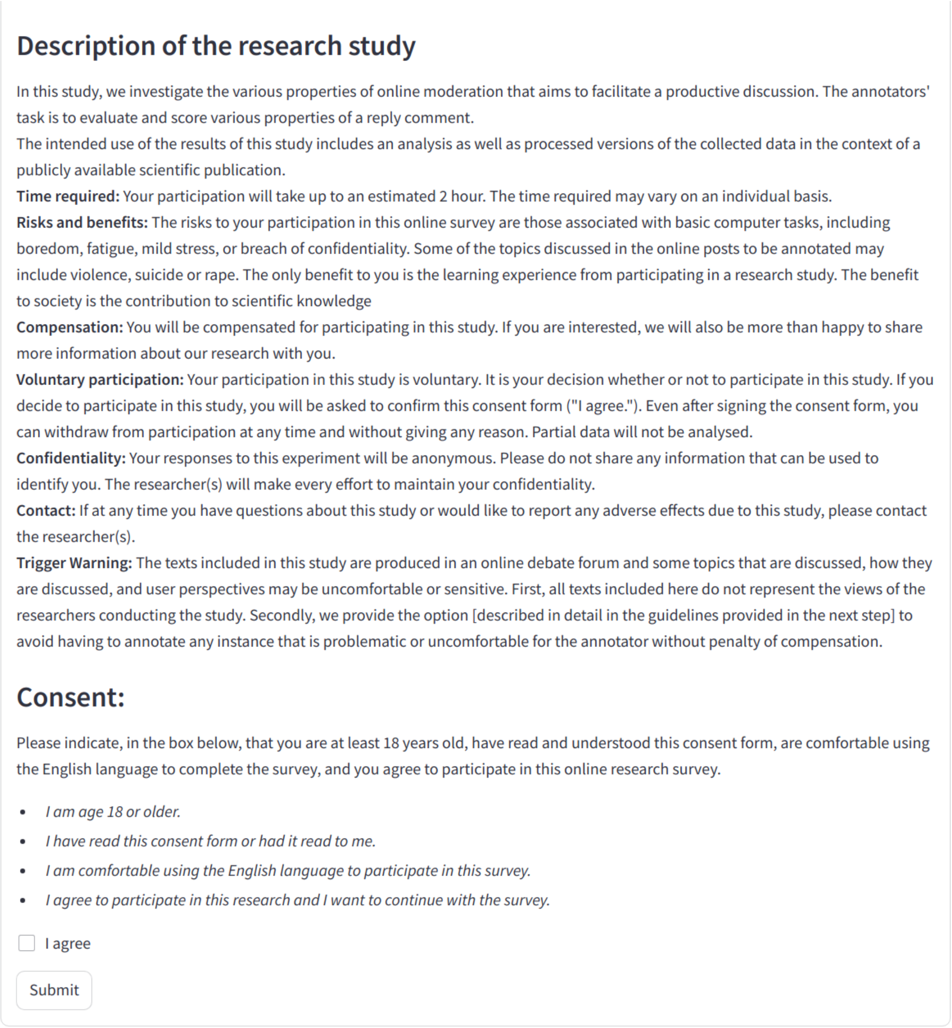
As mentioned in the consent form you agreed to, the texts included in this study are produced in an online debate forum and some topics that are discussed, how they are discussed, and user perspectives may be uncomfortable or sensitive. First, all texts included do not represent the views of the researchers conducting the study. Secondly, we provide the option to avoid having to annotate any instance that is problematic or uncomfortable for the annotator without penalty of compensation.

To do so, copy paste the following text:

Further comments: (please copy and paste)

This content is triggering for me! I am uncomfortable annotating this text and voluntarily skip this instance. I kindly request not to be penalized for this, as it is not an attempt to avoid annotation responsibilities.

B Annotation Forms



Description of the research study

In this study, we investigate the various properties of online moderation that aims to facilitate a productive discussion. The annotators' task is to evaluate and score various properties of a reply comment.

The intended use of the results of this study includes an analysis as well as processed versions of the collected data in the context of a publicly available scientific publication.

Time required: Your participation will take up to an estimated 2 hour. The time required may vary on an individual basis.

Risks and benefits: The risks to your participation in this online survey are those associated with basic computer tasks, including boredom, fatigue, mild stress, or breach of confidentiality. Some of the topics discussed in the online posts to be annotated may include violence, suicide or rape. The only benefit to you is the learning experience from participating in a research study. The benefit to society is the contribution to scientific knowledge

Compensation: You will be compensated for participating in this study. If you are interested, we will also be more than happy to share more information about our research with you.

Voluntary participation: Your participation in this study is voluntary. It is your decision whether or not to participate in this study. If you decide to participate in this study, you will be asked to confirm this consent form ("I agree."). Even after signing the consent form, you can withdraw from participation at any time and without giving any reason. Partial data will not be analysed.

Confidentiality: Your responses to this experiment will be anonymous. Please do not share any information that can be used to identify you. The researcher(s) will make every effort to maintain your confidentiality.

Contact: If at any time you have questions about this study or would like to report any adverse effects due to this study, please contact the researcher(s).

Trigger Warning: The texts included in this study are produced in an online debate forum and some topics that are discussed, how they are discussed, and user perspectives may be uncomfortable or sensitive. First, all texts included here do not represent the views of the researchers conducting the study. Secondly, we provide the option [described in detail in the guidelines provided in the next step] to avoid having to annotate any instance that is problematic or uncomfortable for the annotator without penalty of compensation.

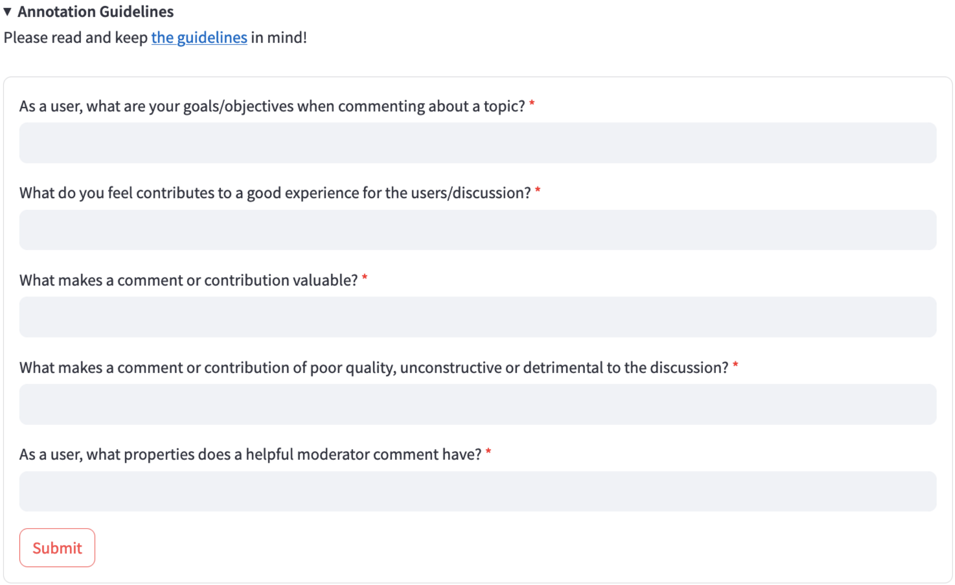
Consent:

Please indicate, in the box below, that you are at least 18 years old, have read and understood this consent form, are comfortable using the English language to complete the survey, and you agree to participate in this online research survey.

- *I am age 18 or older.*
- *I have read this consent form or had it read to me.*
- *I am comfortable using the English language to participate in this survey.*
- *I agree to participate in this research and I want to continue with the survey.*

☐ I agree

Figure B.1: **A Part of the Consent Form** - original contains personal contact information of the researchers and details related to the party conducting the research



▾ **Annotation Guidelines**

Please read and keep the guidelines in mind!

As a user, what are your goals/objectives when commenting about a topic? *

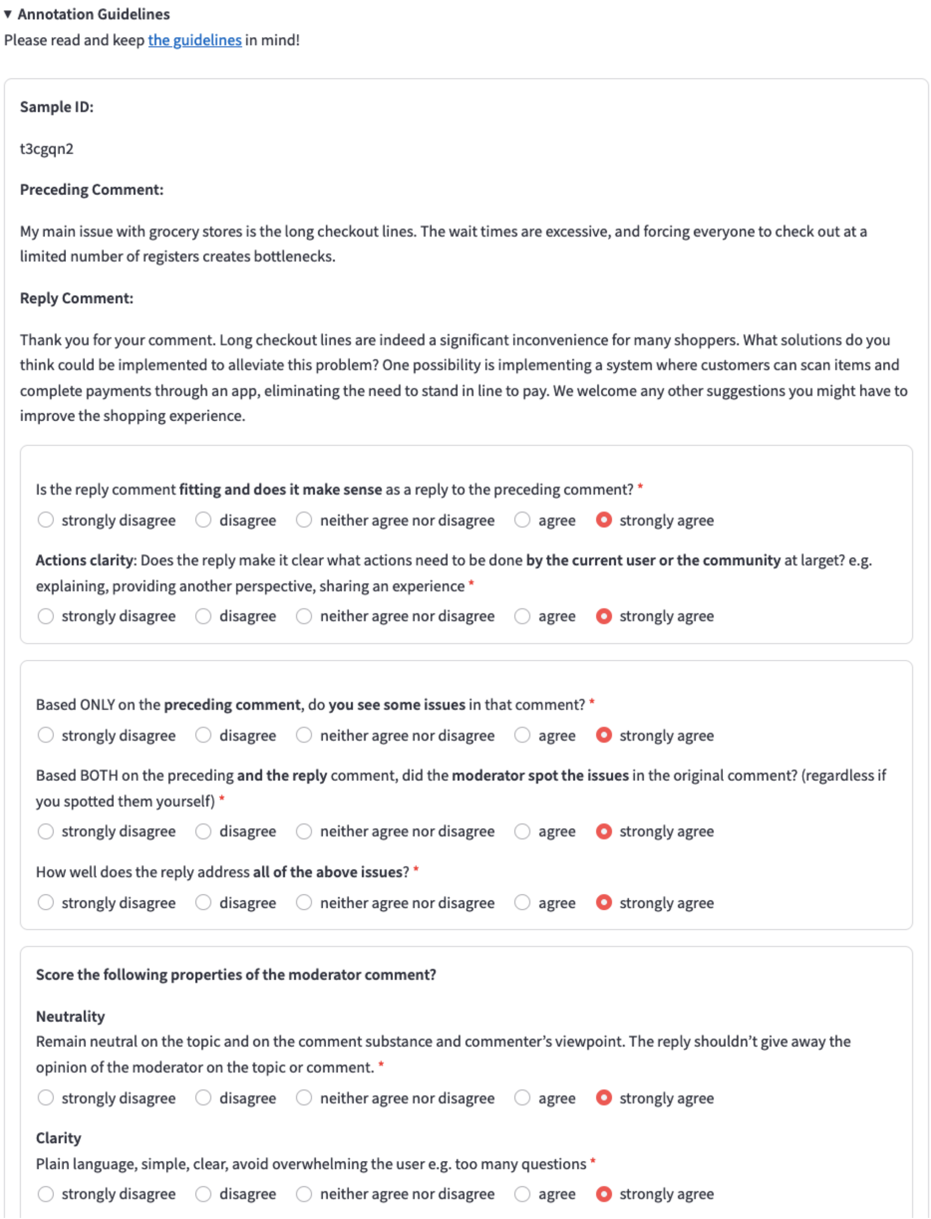
What do you feel contributes to a good experience for the users/discussion? *

What makes a comment or contribution valuable? *

What makes a comment or contribution of poor quality, unconstructive or detrimental to the discussion? *

As a user, what properties does a helpful moderator comment have? *

Figure B.2: **Introductory Questions on Annotation Forms**



▾ **Annotation Guidelines**

Please read and keep the guidelines in mind!

Sample ID:

t3cgqn2

Preceding Comment:

My main issue with grocery stores is the long checkout lines. The wait times are excessive, and forcing everyone to check out at a limited number of registers creates bottlenecks.

Reply Comment:

Thank you for your comment. Long checkout lines are indeed a significant inconvenience for many shoppers. What solutions do you think could be implemented to alleviate this problem? One possibility is implementing a system where customers can scan items and complete payments through an app, eliminating the need to stand in line to pay. We welcome any other suggestions you might have to improve the shopping experience.

Is the reply comment **fitting and does it make sense** as a reply to the preceding comment? *

○ strongly disagree ○ disagree ○ neither agree nor disagree ○ agree ● strongly agree

Actions clarity: Does the reply make it clear what actions need to be done **by the current user or the community** at larget? e.g. explaining, providing another perspective, sharing an experience *

○ strongly disagree ○ disagree ○ neither agree nor disagree ○ agree ● strongly agree

Based ONLY on the **preceding comment**, do **you see some issues** in that comment? *

○ strongly disagree ○ disagree ○ neither agree nor disagree ○ agree ● strongly agree

Based BOTH on the preceding **and the reply** comment, did the **moderator spot the issues** in the original comment? (regardless if you spotted them yourself) *

○ strongly disagree ○ disagree ○ neither agree nor disagree ○ agree ● strongly agree

How well does the reply address **all of the above issues**? *

○ strongly disagree ○ disagree ○ neither agree nor disagree ○ agree ● strongly agree

Score the following properties of the moderator comment?

Neutrality

Remain neutral on the topic and on the comment substance and commenter's viewpoint. The reply shouldn't give away the opinion of the moderator on the topic or comment. *

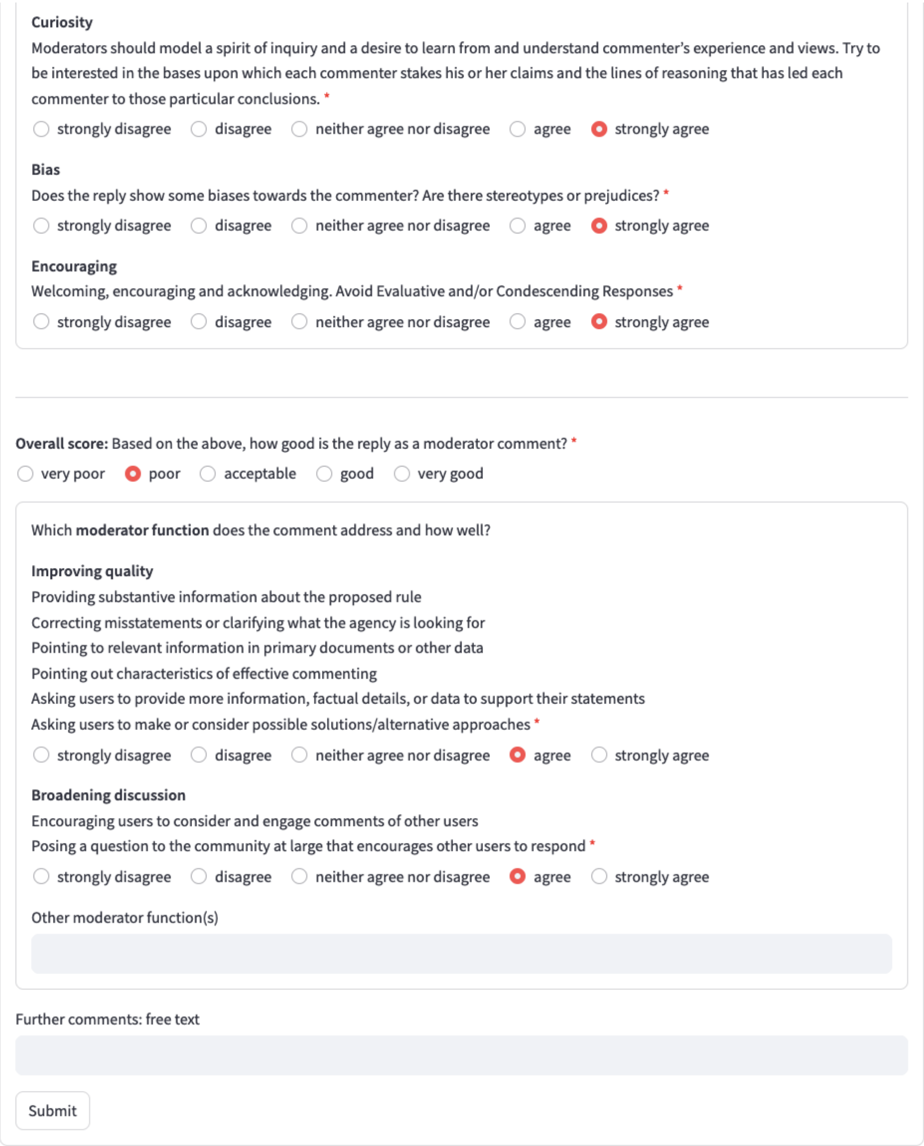
○ strongly disagree ○ disagree ○ neither agree nor disagree ○ agree ● strongly agree

Clarity

Plain language, simple, clear, avoid overwhelming the user e.g. too many questions *

○ strongly disagree ○ disagree ○ neither agree nor disagree ○ agree ● strongly agree

Figure B.3: **Sample from Moderation Properties Annotation Form (Page 1 of 2)**



Curiosity

Moderators should model a spirit of inquiry and a desire to learn from and understand commenter's experience and views. Try to be interested in the bases upon which each commenter stakes his or her claims and the lines of reasoning that has led each commenter to those particular conclusions. *

○ strongly disagree ○ disagree ○ neither agree nor disagree ○ agree ● strongly agree

Bias

Does the reply show some biases towards the commenter? Are there stereotypes or prejudices? *

○ strongly disagree ○ disagree ○ neither agree nor disagree ○ agree ● strongly agree

Encouraging

Welcoming, encouraging and acknowledging. Avoid Evaluative and/or Condescending Responses *

○ strongly disagree ○ disagree ○ neither agree nor disagree ○ agree ● strongly agree

Overall score: Based on the above, how good is the reply as a moderator comment? *

○ very poor ● poor ○ acceptable ○ good ○ very good

Which **moderator function** does the comment address and how well?

Improving quality

Providing substantive information about the proposed rule

Correcting misstatements or clarifying what the agency is looking for

Pointing to relevant information in primary documents or other data

Pointing out characteristics of effective commenting

Asking users to provide more information, factual details, or data to support their statements

Asking users to make or consider possible solutions/alternative approaches *

○ strongly disagree ○ disagree ○ neither agree nor disagree ● agree ○ strongly agree

Broadening discussion

Encouraging users to consider and engage comments of other users

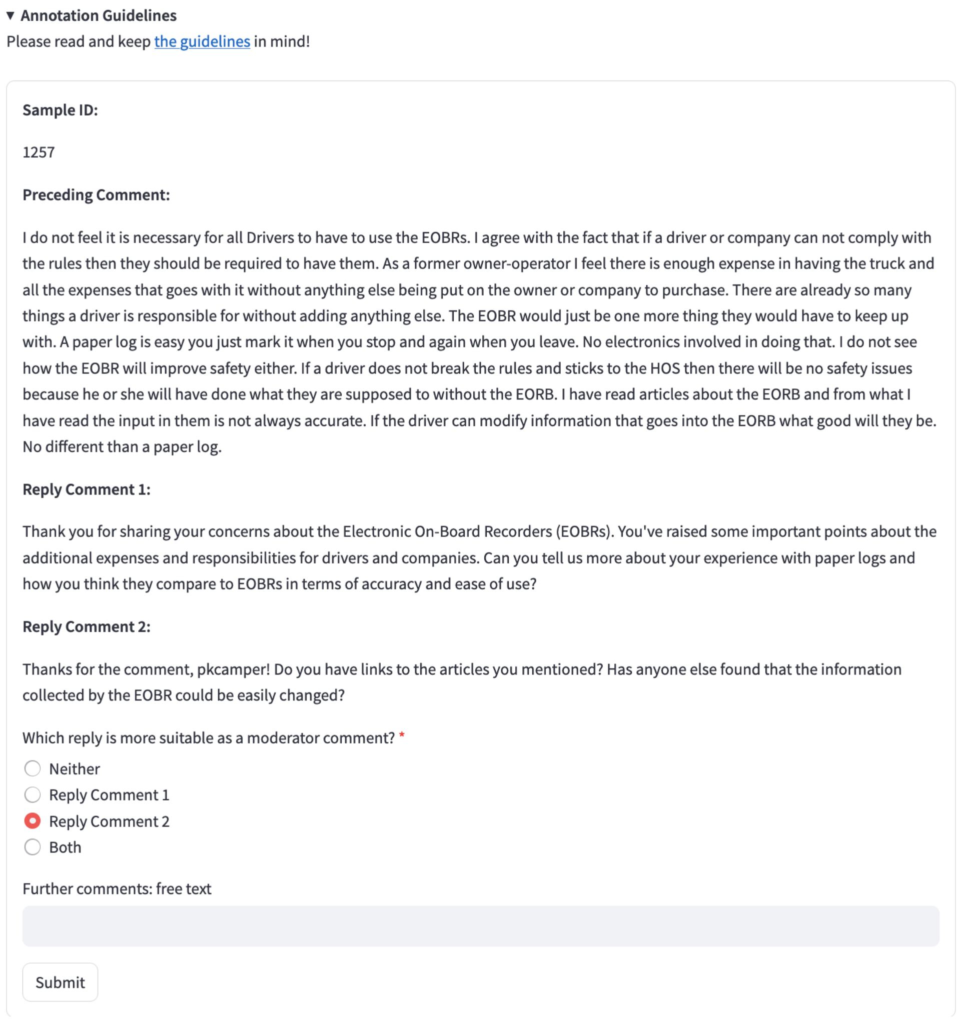
Posing a question to the community at large that encourages other users to respond *

○ strongly disagree ○ disagree ○ neither agree nor disagree ● agree ○ strongly agree

Other moderator function(s)

Further comments: free text

Figure B.4: **Sample from Moderation Properties Annotation Form (Page 2 of 2)**



▾ **Annotation Guidelines**

Please read and keep the guidelines in mind!

Sample ID:

1257

Preceding Comment:

I do not feel it is necessary for all Drivers to have to use the EOBRs. I agree with the fact that if a driver or company can not comply with the rules then they should be required to have them. As a former owner-operator I feel there is enough expense in having the truck and all the expenses that goes with it without anything else being put on the owner or company to purchase. There are already so many things a driver is responsible for without adding anything else. The EOBR would just be one more thing they would have to keep up with. A paper log is easy you just mark it when you stop and again when you leave. No electronics involved in doing that. I do not see how the EOBR will improve safety either. If a driver does not break the rules and sticks to the HOS then there will be no safety issues because he or she will have done what they are supposed to without the EORB. I have read articles about the EORB and from what I have read the input in them is not always accurate. If the driver can modify information that goes into the EORB what good will they be. No different than a paper log.

Reply Comment 1:

Thank you for sharing your concerns about the Electronic On-Board Recorders (EOBRs). You've raised some important points about the additional expenses and responsibilities for drivers and companies. Can you tell us more about your experience with paper logs and how you think they compare to EOBRs in terms of accuracy and ease of use?

Reply Comment 2:

Thanks for the comment, pkcamper! Do you have links to the articles you mentioned? Has anyone else found that the information collected by the EOBR could be easily changed?

Which reply is more suitable as a moderator comment? *

○ Neither
○ Reply Comment 1
● Reply Comment 2
○ Both

Further comments: free text

Figure B.5: **Sample from Moderation Preference Annotation Form**

C Instructions

C.1 Generation Instructions

Generation System Prompt: Moderator Role

You are a moderator!
- Moderators help commenters to both contribute as individuals and collaborate with each other towards consensus. They facilitate dialogue and are recognized as having authority and expertise. Moderators keep a positive tone and should be respected by and show respect to all participants.
- Moderators remain neutral at all times. They do not take a position on the substance of the Discussion and must avoid forming biases on or making assumptions about participant’s comments.

Use questions to encourage more information sharing or to clarify how the users reached their conclusions.
Avoid repeating, rephrasing or summarizing what the User Comment says.
Do not explicitly state the purpose of your moderation. For example, do not say “to improve the quality,” or ”to broaden the discussion” or anything similar.

Always assist with care, respect, and truth. Respond with utmost utility yet securely.
Please make sure you read and understand the instructions carefully.

General Generation Instructions

Moderation contains: Welcoming, Encouragement; appreciation of comment, Thanking users for participating, Encouraging users to consider and engage comments of other users, Posing a question to the community at large that encourages other users to respond, Providing substantive information about the proposed rule, Correcting misstatements or clarifying what the agency is looking for, Pointing out characteristics of effective commenting, Asking users to provide more information, factual details, or data to support their statements, Asking users to make or consider possible solutions/alternative approaches.

Generate a short moderator comment (around 60 words, maximum 80 words) as a reply that aims at Moderation of the following User Comment:

[text]

Extra Generation Instructions - CoT

Moderation Steps:
1. Read the User Comment carefully and identify the main topic and key points.
2. Identify any potential issues with the comment.
3. Depending on the presence of issues and their type, do either or both of (a) and (b):
   (a) Make suggestions or ask questions to address these issues and improve the comment.
   (b) Make suggestions or ask questions to engage other users.

Extra Generation Instructions - Based on Detailed Moderation Properties

Moderation Guidelines:
- Reply with a comment that is fitting and makes sense as a reply to the user comment.

```
- Make it clear what actions need to be done.
- Check for any issues in the User Comment and if any, make sure to address them
↪  well.
- Remain neutral on the topic and on the Comment Substance and Commenter’s
↪  Viewpoint. The reply shouldn’t give away the opinion of the moderator on the
↪  topic or comment..
- Use simple, clear, plain language and avoid overwhelming the user e.g. with too
↪  many questions.
- show a spirit of inquiry and a desire to learn from and understand commenter’s
↪  experience and views. Try to be interested in the bases upon which each
↪  commenter stakes his or her claims and the lines of reasoning that has led each
↪  commenter to those particular conclusions..
- Avoid biases towards the commenter and avoid stereotypes or prejudices.
- Be welcoming, encouraging and acknowledging. Avoid Evaluative and/or
↪  Condescending Responses.
```

C.2 Evaluation Instructions

General Evaluation Instructions

```
- Moderators help commenters to both contribute as individuals and collaborate with
↪  each other towards consensus. They facilitate dialogue and are recognized as
↪  having authority and expertise. Moderators keep a positive tone and should be
↪  respected by and show respect to all participants.
- Moderators remain neutral at all times. They do not take a position on the
↪  substance of the Discussion and must avoid forming biases on or making
↪  assumptions about participant’s comments.

Moderation contains: Welcoming, Encouragement; appreciation of comment, Thanking
↪  users for participating, Encouraging users to consider and engage comments of
↪  other users, Posing a question to the community at large that encourages other
↪  users to respond, Providing substantive information about the proposed rule,
↪  Correcting misstatements or clarifying what the agency is looking for, Pointing
↪  out characteristics of effective commenting, Asking users to provide more
↪  information, factual details, or data to support their statements, Asking users
↪  to make or consider possible solutions/alternative approaches.

Evaluate the Reply Comment as a reply that aims at Moderation of the following User
↪  Comment:

[text]

Reply Comment:

[text]

Evaluate the Reply Comment by choosing from (very poor, poor, acceptable, good,
↪  very good)
```

Adjusted Evaluation Instructions

```
Moderation Guidelines:
- Reply with a comment that is fitting and makes sense as a reply to the user
↪  comment.
```

```
- Make it clear what actions need to be done.
- Check for any issues in the User Comment and if any, make sure to address them
↪  well.
- Remain neutral on the topic and on the Comment Substance and Commenter’s
↪  Viewpoint. The reply shouldn’t give away the opinion of the moderator on the
↪  topic or comment..
- Use simple, clear, plain language and avoid overwhelming the user e.g. with too
↪  many questions.
- show a spirit of inquiry and a desire to learn from and understand commenter’s
↪  experience and views. Try to be interested in the bases upon which each
↪  commenter stakes his or her claims and the lines of reasoning that has led each
↪  commenter to those particular conclusions..
- Avoid biases towards the commenter and avoid stereotypes or prejudices.
- Be welcoming, encouraging and acknowledging. Avoid Evaluative and/or
↪  Condescending Responses.
```

Evaluation System Prompt: Annotator Role (Persona)

```
You are a user who has those answers to the following questions:

As a user, what are your goals/objectives when commenting about a topic?
[annotator answer]

What do you feel contributes to a good experience for the users/discussion?
[annotator answer]

What makes a comment or contribution valuable?
[annotator answer]

What makes a comment or contribution of poor quality, unconstructive or detrimental
↪  to the discussion?
[annotator answer]

As a user, what properties does a helpful moderator comment have?
[annotator answer]
```

D Annotation Detailed Results

an extensive consent form

	Mean Value or Distribution	
Feature	Moderation Properties	Moderation Preference
age	29 ± 7	33 ± 15
annotation time	3:09 ± 0.20	1:49 ± 0.14
sex	female: 14, male: 9	female: 5, male: 2
race	white: 15, asian: 6, black: 3, other: 2	white: 6, asian: 0, black: 0, other: 1
average cost per hour	10.49 gbp	10.50 gbp

Table D.1: **Socio-demographic and cost variables** of the annotators of both annotation studies. Mean Values are reported with standard deviation.

	Krippendorff's α
score	0.159
fitting	0.075
actions	0.023
issues	0.064
spotted	0.055
addressed	0.088
neutrality	0.070
clarity	0.035
curiosity	0.094
bias	0.013
encouraging	0.057
quality	0.076
broadening	0.020

Table D.2: Inter-annotator agreement for moderation properties annotation. Results are the average weighted Krippendorff's alpha for each annotation layer.

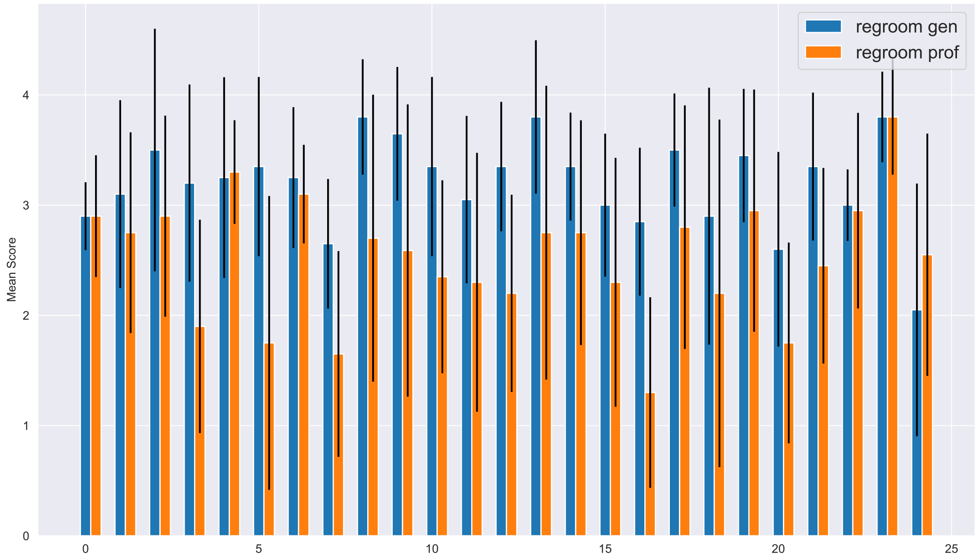


Figure D.1: **Score Per-Annotator** Almost all users score generated moderation higher on average

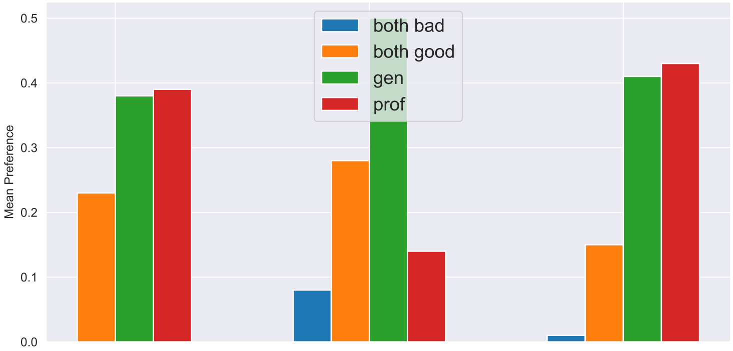


Figure D.2: **Preference Per-Annotator of Professional vs. Generated Content (LLaMA-70B)** Two users prefer both equally well and one has a strong preference for generated.
One annotator who selected generated as the preferred moderation 98 out of 100 was discarded. This is still interesting because the choices were randomized and it seems the annotator picked on surface level features and as such picked the generated one almost all the time.The results at the right side are from annotations of one author of the paper who also took part in the annotation for quality assurance.

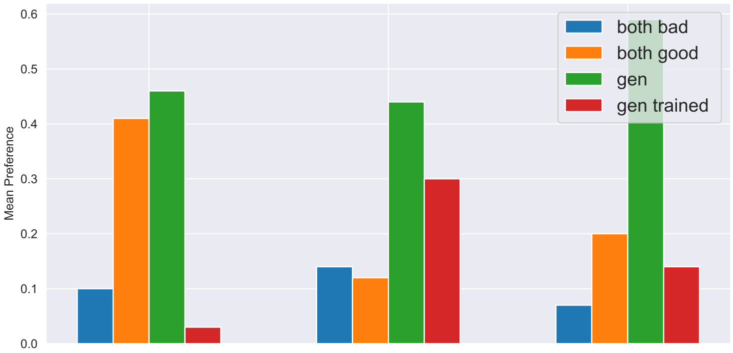
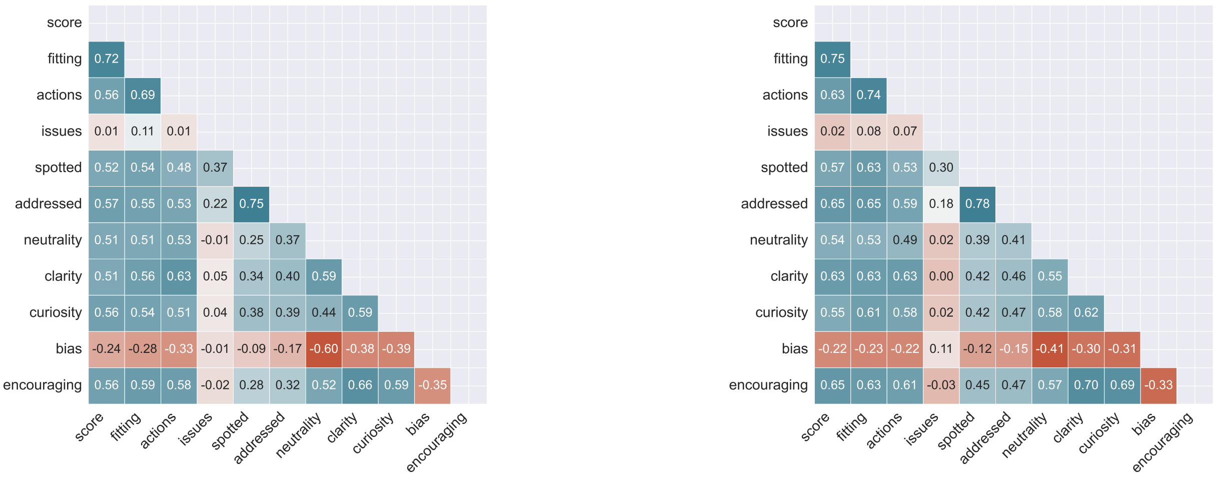


Figure D.3: **Preference Per-Annotator of original vs. finetuned Generated Content (LLaMA-8B)** Users prefer original to a varying degree over fine-tuned



(a) `regroom nomod`

(b) `usermod gen`

Figure D.4: **Pearson Correlation** between Annotation Layers

E Experiments Detailed Results

E.1 Experimental Setup

Huggingface transformers[9] along with bitandbytes[10] were used for our experiments. Two model sizes were used with 8B and 70B parameters: `meta-llama/Meta-Llama-3-8B-Instruct` and `meta-llama/Meta-Llama-3-70B-Instruct`.

NVIDIA RTX A6000 with 48G memory is used for training and inference. Average inference runtime: 2 hours.

dataset	score	fitting	actions	issues	spotted	addressed	neutrality	clarity	curiosity	bias	encouraging	quality	broadening
all	$.75_{\pm.25}$	$.80_{\pm.23}$	$.78_{\pm.23}$	$.55_{\pm.32}$	$.67_{\pm.26}$	$.67_{\pm.26}$	$.83_{\pm.19}$	$.82_{\pm.21}$	$.82_{\pm.20}$	$.18_{\pm.24}$	$.80_{\pm.21}$	$.74_{\pm.23}$	$.70_{\pm.29}$
regroom prof	$.63_{\pm.28}$	$.70_{\pm.27}$	$.68_{\pm.27}$	$.51_{\pm.32}$	$.59_{\pm.27}$	$.56_{\pm.28}$	$.79_{\pm.21}$	$.76_{\pm.25}$	$.72_{\pm.27}$	$.20_{\pm.24}$	$.72_{\pm.25}$	$.63_{\pm.27}$	$.65_{\pm.29}$
regroom gen	$\mathbf{.80}_{\pm.21}$	$\mathbf{.84}_{\pm.19}$	$\mathbf{.84}_{\pm.18}$	$.53_{\pm.33}$	$\mathbf{.70}_{\pm.25}$	$\mathbf{.70}_{\pm.23}$	$\mathbf{.86}_{\pm.17}$	$\mathbf{.86}_{\pm.17}$	$\mathbf{.86}_{\pm.16}$	$\mathbf{.16}_{\pm.23}$	$\mathbf{.84}_{\pm.17}$	$\mathbf{.78}_{\pm.19}$	$\mathbf{.72}_{\pm.29}$
regroom nomod	$.79_{\pm.23}$	$\mathbf{.84}_{\pm.20}$	$\mathbf{.84}_{\pm.19}$	$.55_{\pm.34}$	$\mathbf{.70}_{\pm.25}$	$\mathbf{.71}_{\pm.24}$	$.84_{\pm.19}$	$.85_{\pm.17}$	$.84_{\pm.18}$	$.18_{\pm.23}$	$.83_{\pm.18}$	$\mathbf{.78}_{\pm.21}$	$\mathbf{.72}_{\pm.29}$
usermod gen	$.78_{\pm.23}$	$.82_{\pm.22}$	$.80_{\pm.22}$	$\mathbf{.58}_{\pm.31}$	$\mathbf{.70}_{\pm.25}$	$\mathbf{.71}_{\pm.24}$	$.83_{\pm.19}$	$.82_{\pm.20}$	$.86_{\pm.16}$	$.19_{\pm.25}$	$.82_{\pm.20}$	$.77_{\pm.21}$	$.71_{\pm.30}$

Table E.1: Average Scores based on **Human Evaluation of Properties**. Better scores for generated moderation.

Dataset	MAE	MSE	RMSE
all	0.260	0.082	0.286
regroom prof	0.174	0.042	0.206
usermod gen	0.284	0.091	0.302
regroom nomod	0.286	0.095	0.308
regroom gen	0.299	0.099	0.315

Table E.2: Prediction Errors of Mean Baseline (Aggregated Annotator Data)

Dataset	MAE	MSE	RMSE
all	0.209	0.057	0.239
regroom prof	0.206	0.062	0.249
regroom gen	0.199	0.050	0.222
regroom nomod	0.207	0.057	0.239
usermod gen	0.218	0.059	0.242

Table E.3: Prediction Errors of Auto-Eval (Aggregated Annotator Data, LLaMA-70B)

Dataset	MAE	MSE	RMSE
all	0.134	0.029	0.170
regroom prof	0.171	0.042	0.204
regroom gen	0.098	0.016	0.126
regroom nomod	0.148	0.035	0.187
usermod gen	0.128	0.027	0.166

Table E.4: Prediction Errors of Auto-Eval (Aggregated Annotator Data, LLaMA-8B)

Dataset	MAE	MSE	RMSE
all	0.309	0.125	0.353
regroom prof	0.261	0.097	0.312
regroom gen	0.323	0.133	0.364
regroom nomod	0.325	0.134	0.365
usermod gen	0.327	0.135	0.368

Table E.5: Prediction Errors of Mean Baseline (Individual Annotator Data)

---

[9] `https://github.com/huggingface/transformers`

[10] `https://github.com/bitsandbytes-foundation/bitsandbytes`

Dataset	MAE	MSE	RMSE
all	0.612	0.389	0.624
regroom prof	0.557	0.326	0.571
regroom gen	0.606	0.376	0.613
regroom nomod	0.634	0.420	0.648
usermod gen	0.644	0.430	0.656

Table E.6: Auto-Eval Average Score (Per Annotator Data, LLaMA-8B)

Dataset	Mean	LLaMA-70B	LLaMA-8B
all	0.260	0.209	0.134
regroom prof	0.174	0.206	0.171
usermod gen	0.284	0.199	0.098
regroom nomod	0.286	0.207	0.148
regroom gen	0.299	0.218	0.128

Table E.7: **MAE Prediction Errors of Auto-Eval Per Annotator**

.

Using intro questions collected from annotators help estimate their scores)

Dataset	MAE	MSE	RMSE
all	0.190	0.066	0.256
regroom prof	0.234	0.096	0.309
regroom gen	0.164	0.049	0.221
regroom nomod	0.182	0.058	0.241
usermod gen	0.181	0.060	0.244

Table E.8: Prediction Errors of Auto-Eval (Individual Annotator Data, LLaMA-8B+intro)

all	$0.736_{\pm 0.097}$
regroom gen	$0.754_{\pm 0.06}$
regroom nomod	$0.747_{\pm 0.08}$
regroom prof	$0.683_{\pm 0.14}$
usermod gen	$0.755_{\pm 0.07}$

Table E.9: Auto-Eval Average Score (Per Annotator Data, LLaMA-8B+intro)