PerspectiveMod: A Perspectivist Resource for Deliberative Moderation

Eva Maria Vecchi[1], **Neele Falk**[1], **Carlotta Quensel**[2],
Iman Jundi[1], and **Gabriella Lapesa**[3]
[1]Institute for Natural Language Processing, University of Stuttgart, Germany
[2]Institute of Artificial Intelligence - Leibniz University Hannover
[3]GESIS - Leibniz Institute for Social Sciences and Heinrich-Heine University of Düsseldorf
[1]first[-middle].last@ims.uni-stuttgart.de,
[2]c.quensel@ai.uni-hannover.de, [3]gabriella.lapesa@gesis.org

Abstract

Human moderators in online discussions face a heterogeneous range of tasks, which go beyond content moderation, or policing. They also support and improve discussion quality, which is challenging to model (and evaluate) in NLP due to its inherent subjectivity and the scarcity of annotated resources. We address this gap by introducing PerspectiveMod, a dataset of online comments annotated for the question: "Does this comment require moderation, and why?" Annotations were collected from both expert moderators and trained non-experts. PerspectiveMod is unique in its intentional variation across (a) the level of moderation experience embedded in the source data (professional vs. non-professional moderation environments), (b) the annotator profiles (experts vs. trained crowdworkers), and (c) the richness of each moderation judgment, both in terms on fine-grained comment properties (drawn from argumentation and deliberative theory) and in the representation of the individuality of the annotator (socio-demographics and attitudes towards the task). We advance understanding of the task's complexity by providing interpretation layers that account for its subjectivity. Our statistical analysis highlights the value of collecting annotator perspectives, including their experiences, attitudes, and views on AI, as a foundation for developing more context-aware and interpretively robust moderation tools.

1 Introduction

Online platforms that facilitate deliberative and argumentative discourse, such as civic participation portals and forums for persuasive dialogue, represent crucial arenas for collective reasoning and public voice. Effective moderation is essential to ensure that conversations remain constructive, inclusive, and aligned with platform or civic goals. While most NLP work on moderation has focused on detecting toxicity or harmful speech (Zampieri et al., 2019; Markov et al., 2023; Park et al., 2021; Hee et al., 2024; Kumar et al., 2024), such approaches fall short in deliberative contexts. Here, moderation must not only flag problematic behavior but also support participants in clarifying arguments, staying on topic, and engaging respectfully and productively (Innes, 2004; Trénel, 2009; Park et al., 2012; Deng et al., 2023). This is the type of moderation this paper focuses on, **deliberative moderation**, or facilitation, a form of intervention grounded in deliberative democratic theory (Habermas, 1981; Gutmann and Thompson, 2004).

Deliberative moderation is an **extremely complex task**, as it involves context-sensitive, value-driven decisions about when and how to intervene. A comment that breaks no rule may still derail the discussion or marginalize participants, while another that appears confrontational may, in context, advance meaningful deliberation (see e.g. Tbl. 1). These judgments are shaped by discourse goals, platform norms, and individual moderator perspectives. This **subjectivity** makes deliberative moderation especially difficult to model, and even harder to evaluate. There is no single "correct" intervention, and even trained experts may disagree on when action is needed or what form it should take. Yet existing annotated datasets overwhelmingly assume clearly defined classification tasks, leaving a gap in resources that reflect the ambiguity and normative nature of real-world moderation.

Growing calls to rethink annotation in NLP from a **perspectivist** lens (Fleisig et al., 2024) echo in this context. Rather than assuming a single ground truth, perspectivist approaches emphasize documenting the diversity of interpretations across annotators. Prior work has incorporated socio-demographic or attitudinal metadata to explain annotator variation (Frenda et al., 2023; Abercrombie et al., 2024), but these features often fall short of capturing the lived interpretive stances that shape language judgments (Hu and Collier,

Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing, pages 34175–34198
November 4-9, 2025 ©2025 Association for Computational Linguistics

2024). Here, we extend this line of work by designing structured interview-style questions that probe annotators' background, goals, and experience with online discussion and moderation. This enables a much richer representation of annotator perspective, grounding moderation judgments in both discourse context and the lived interpretive frameworks of those making them.

We introduce **`PerspectiveMod`**, a perspectivist dataset for studying moderation in deliberative and argumentation settings. We collect annotations from both expert and non-expert annotators on text sampled from three moderation ecosystems, ranging from professionally facilitated civic platforms to informal peer-led interventions on Reddit, capturing a wide range of moderation practices. The dataset includes 225 comments, each annotated 12 times (totaling 1,645 annotation hours), with labels covering need for moderation, type of intervention, and contributing discourse properties (e.g., clarity, constructiveness, tone). We also collect rich annotator metadata to examine how task interpretation, experience, and expectations shape moderation behavior. By documenting not just what decisions are made, but who is making them and why, we provide a crucial foundation for modeling moderation as a context-sensitive, subjective task. Indeed, our statistical analysis confirms that annotator attitudes, online discussion experience, and views on AI significantly shape moderation sensitivity, confirming recent trends in perspectivist NLP findings that socio-demographics alone are insufficient to explain complex annotation behavior in subjective tasks (Orlikowski et al., 2023).

Our contributions are three-fold: (i) `PerspectiveMod`, a novel dataset that includes annotations for moderation necessity, intervention type, discourse properties grounded in research on argument and deliberative quality (Falk and Lapesa, 2023), and annotator-level information such as expertise, background, and interpretive stance; (ii) an in-depth analysis of the subjective structure of moderation judgments, highlighting both content-driven and annotator-specific factors that complicate efforts to standardize or automate this task, such as goals, norms, and intervention thresholds. (iii) an interpretively rich, multi-perspective annotation framework which can be applied to other moderation domains and further contributes to the methodological research targeting the comparison of expert and non-expert annotations, and strategies to train non-experts (Lee et al., 2022).

Comment	**M**	**A**
Not exactly telecommuting but the city could work with employers to create staggard commute times. This way transportation is not overwhelmed during certain hours of the day. The traditional 9-5 is becoming obsolete, especially when taking into consideration the analytical service industries.	0	8%
The city should also implement a car2go like service and designate spots citywide (similar to Citi Bike). This could reduce ownership if the option is availble nearby and at your destination. Avaliable parking spaces could be reserved via app for a limited window.	0	33%
I'm not quoting. I'm expanding. You say it's fiscal. I say, "yeah, it's fiscal, but fiscal isn't something about numbers and money, it's about other people deciding what happens to the fruits of your breath and strenght".	1	50%
I have a feeling the vast majority of CMV (including myself) do not debate at international tournaments. I apologise for not being up to your standards.	0	75%
You're not trying to understand why it's an insult. You're stubbornly insisting that you didn't *intend* it as an insult, and thus it cannot possibly *be* an insult.	1	92%

Table 1: `PerspectiveMod` examples with sampling model (M) predictions (binary `to_moderate`, cf. Sect. 3.1) and overall annotator (A) judgments (rate `to_moderate` was positive).

2 Background & Related Work

Deliberative moderation in NLP NLP has so far focused on tasks of content moderation, such as detecting hate speech (Schmidt and Wiegand, 2017; Masud et al., 2024), offensiveness (Mostafazadeh Davani et al., 2024) or violation of community norms (Park et al., 2021) or promoting positive behavior through counter-speech (Bonaldi et al., 2023, 2024); hope speech (Chakravarthi and Muralidaran, 2021) or constructiveness (Kolhatkar and Taboada, 2017a,b). More concrete deliberative norms (including those that focus on positive behavior) are addressed by work in Argument Mining, in which different dimensions of argument quality (Habernal and Gurevych, 2016; Wachsmuth et al., 2017; Gurcke et al., 2021) or deliberative quality are annotated and automatically modeled, but their relationship to deliberative moderation is only rarely investigated.

A major bottleneck in studying deliberative moderation is the lack of high-quality data. As noted by Korre et al. (2025), very few datasets include human moderator interventions. While there is existing research on supervised classification of moderator interventions using moderated data (Falk et al., 2021; Falk and Lapesa, 2023), the signal remains noisy because the data is only incorporated into models post-hoc, and without insight into the consistency of moderator decisions or alternative plausible interventions. Although there is growing attention to both theoretical aspects of moderation

(Friess et al., 2020) and public attitudes toward AI-driven facilitation (Jungherr and Rauchfleisch, 2025), the decision-making process behind moderation and the annotations that guide modeling remain underexplored.

Another major bottleneck faced by existing research on deliberative moderation is the question of how to evaluate the quality of the interventions. In an ideal situation, NLP tools should be tested in real-world deliberation scenarios: in practice, this is in the majority of the cases not feasible because setting up hybrid deliberation is highly costly in terms of times and resources, not very welcome in online communities (and for good reasons, given the potential harms and biases AI necessarily brings). Related strands of work have explored practical applications of LLMs for constructive moderation, such as identifying grounding gaps in model behavior (Shaikh et al., 2024), reframing user comments to promote receptiveness (Kambhatla et al., 2024), and automatically evaluating contribution quality in deliberation platforms (Gelauff et al., 2024). While some work argues that LLM simulation is the solution (Tsirmpas et al., 2025), in this paper we take a human-centered approach and we collect fine-grained assessments of moderation decisions by both experts and trained non-experts, in hopes to encourage more comprehensive model implementations and evaluations.

Subjectivity & perspectivism in NLP Deciding whether to moderate a comment is a subjective task: this is why our work adopts a perspectivist approach. The perspectivist framework (Aroyo and Welty, 2015; Uma et al., 2021; Cabitza et al., 2023) promotes the idea that variation in annotation reflects the reality of language understanding, shaped by ambiguity, vagueness, and individual differences in background and experience, and that embracing this variation can lead to more robust and fair NLP models. As a result, recent work has explored perspectivism in modeling human label distributions (Lee et al., 2023), annotator-specific preferences (Bonaldi et al., 2023; Rodríguez-Barroso et al., 2024), and evaluation practices (Baan et al., 2022; Basile et al., 2021). Other studies focus on the design of un-aggregated datasets that include richer annotator metadata and analyze its influence on annotation patterns and model behavior (Vitsakis et al., 2024).

Within the broader perspectivist framework, data collection and application design in NLP require rethinking. Fleisig et al. (2024) outlines several challenges and concrete recommendations that remain underexplored in the field, above all, the need to capture richer information about annotators. While recent work has begun to incorporate socio-demographic and attitudinal metadata to account for annotator variation (Frenda et al., 2023; Abercrombie et al., 2024), such features have shown limited explanatory power on their own (Hu and Collier, 2024; Orlikowski et al., 2023). We address the limitations of relying solely on socio-demographics by designing interview questions that capture annotators' backgrounds and experience with online moderation, enabling a more nuanced understanding of their perspectives.

3 The PerspectiveMod Dataset

In this section, we walk the reader through the steps we followed to build `PerspectiveMod`, highlighting challenges and methodological considerations behind our design choices.

3.1 Data Sources & Instance Selection

`PerspectiveMod` draws from three datasets that reflect distinct moderation environments: (1) RegulationRoom (Park et al., 2012), a civic engagement platform with professional moderators; (2) `r/ChangeMyView` (CMV), a subreddit with semi-trained community moderators; and (3) UMOD, a CMV-based dataset capturing informal, user-driven moderation behavior (Falk et al., 2024). These sources represent a spectrum of moderation—from expert-led to grassroots interventions.

To systematically sample data for annotation, we used multi-task adapter models trained on moderation signals from each source. Following Falk and Lapesa (2023), the task for the adapter models was a binary prediction of whether a comment should (*positive*) or should not (*negative*) be moderated, based on an extended set of deliberative and argument quality features. These predictions were not taken as ground truth but used to select a varied set of candidate instances. Initial tests revealed wide performance variability across datasets (App. Tbl. 4; see App. D for details), and correlations with expert judgments were insignificant;[1] highlighting that deliberative and argument quality features alone are insufficient for robust modera-

[1] E.g., correlations between all expert annotator `to_moderate` judgments and adapter model prediction range from -0.08 to 0.08, see App. Tbl. 6.

tion prediction, and motivating the need for human annotation.

For annotation, we sampled 225 comments, maintaining a 2:1 ratio of predicted *need-to-moderate* to *not-need-to-moderate*. Each instance included the target comment (mean length: 130.5 tokens) and, when available, its topic (14.8 tokens) and preceding comment (95.3 tokens). Expert annotators were additionally shown model-generated visualizations of eight latent quality features influencing the sampling model's prediction. This visualization was provided for context but was not part of the annotation task itself.

3.2 Annotation Task

To capture the reasoning behind moderation decisions, we designed a multi-faceted annotation task (Table 2). We collected annotations from both expert moderators and non-expert crowdworkers, with different task designs tailored to each group. Across both groups, the annotation interface and instructions prioritized interpretive depth over annotation speed. This was an intentional design decision aimed at capturing the full complexity and contributions to subjectivity in moderation decisions. All materials, including guidelines and annotation forms, are publicly available (App. B).

Expert Annotations Experts were recruited from civic platforms (make.org), the CMV moderator team, and academic communities focused on argumentation and civic dialogue. All had verified professional or research experience in online moderation. Recruitment was time-intensive and logistically challenging, conducted over 9 months through direct outreach and academic networks.[2] Despite extensive efforts, our final set of experts remained small (i.e., 5), highlighting the scarcity of accessible moderation expertise and the difficulty of scaling such annotation efforts through traditional means.

Experts annotated all 225 instances[3] via a custom annotation interface.[4] Each instance included context (comment, topic, preceding comment), a visualization of the eight highest contributing properties to the model's decision, and tasks capturing: (1) moderation necessity; (2) priority/urgency; (3) appropriate moderation function(s) (Park et al., 2012); and (4) contributing quality properties (Falk and Lapesa, 2023). As part of an initial pilot study, they were also asked whether model-filtered comments might support real-life moderation workflows. We report this feedback separately (Sec. **??**), as it does not form part of the core annotation dataset. Experts received detailed guidelines throughout and were compensated €12.40/hour (avg. 10 hours).

Crowdsourced Annotations To complement the expert annotations and scale the dataset, we recruited non-expert annotators on the Prolific platform.[5] Participants met eligibility criteria (fluent English, secondary education, Anglophone residence) and were paid £10.50/hour, with a £5 bonus for elaboration in free-text responses. The 225 instances were split into 9 batches of 25, with each instance annotated by an average of 6.5 workers (after quality control).

Prolific workers viewed the same comment context but did not receive model visualizations. They were asked whether the comment (or thread) warranted moderation. They then rated 12 quality properties on a 5-point Likert scale and selected one or more appropriate moderation functions. In addition to the detailed guidelines, Prolific annotators received two structured tutorials introducing moderation for deliberative discourse and online platform, as well as describing the role of experienced moderators with examples.

Interview Questions To complement the annotations and better understand the perspective from which participants approached the task, we included introductory and concluding questionnaires capturing participants' backgrounds, experience with online moderation, and views on the goals and challenges of moderation. These responses revealed each annotator's implicit framework and self-assessed competence. Free-text answers were manually coded by two authors using an inductive approach: initial labels were generated independently, merged into a unified schema, and applied to the full dataset.[6]

3.3 Quantitative properties

We begin with a descriptive analysis to provide an initial perspective on annotation patterns. Inter-annotator agreement on the binary `to_moderate` label was low overall, including among domain experts (Fleiss' $\kappa = 0.055$; Krippendorff's $\alpha = 0.040$;

[2] See Appendix A.

[3] Participants were given the option to skip an instance if they were uncomfortable annotating it. Experts skipped 51 instances, crowdworkers skipped 19.

[4] Built with Streamlit and hosted on HuggingFace.

[5] https://www.prolific.com/

[6] See Appendix C for full coding details and examples.

	Annotation Layer	Feature	Details
Comment-level	Need for Moderation	`to_moderate` priority level	Binary Likert scale
	Moderation Function	Broadening Discussion, Improving Comment Quality, Content Correction, Keeping Discussion on Topic, Organizing Discussion, Policing, Resolving Site Use Issues, Social Functions, None, Other	Binary
	Properties (Prolific)	appropriateness, clarity, constructiveness, rationality, proposal, emotion, respect, reciprocity, storytelling, misinformation, moderation behavior	Likert scale
	Properties (Experts)	[above properties], common good, effectiveness, impact, overall quality, Q for justification, reasonableness, reference, Other	Binary, mult. choice
Annotator-level	Introductory Interview	personal moderation experience, profession, goals/objectives as moderator, what makes discussion good, what makes comment valuable vs. not to discussion	Pre-task
	Concluding Interview	ease of annotation task, usefulness of AI tools for moderation, moderation bottlenecks	Post-task
	Socio-demographics	fluent languages, highest education level completed, age, sex, ethnicity (simplified), country_birth, country_residence, nationality, language, student status, employment status	Via Prolific
	Annotation Time	overall time spent on annotation task	

Table 2: Overview of annotation layers and features collected in the annotation tasks.

see App. Tbl. 8). Agreement among crowdworkers varied widely by batch, with κ scores ranging from -0.095 to 0.221. These results underscore the subjective nature of moderation and suggest that decisions about whether to intervene differ both across and within annotator groups.

We also observed clear differences in overall moderation rates. Experts were generally more conservative, marking fewer than 30% of comments as needing intervention. In contrast, Prolific annotators showed much greater variability, with some labeling up to 90% of comments as requiring moderation (Fig. 1). This highlights not only group-level differences in moderation thresholds, but also substantial variation at the individual level.

When examining how annotators applied specific moderation functions (Fig. 2), crowdworkers most frequently selected functions like Improving Comment Quality, Policing, and Keeping on Topic. Expert annotators, however, displayed considerable variation in their functional preferences. For instance, one expert prioritized Keeping on Topic, while others emphasized functions like Broadening or Policing, reflecting individualized interpretations shaped by their moderation background and preferences. Broadly, Prolific annotators applied a wider range of functions more uniformly, whereas experts were more conservative and deliberate in their selections. This distinction reinforces the idea that moderation is not a one-size-fits-all task, but a context-sensitive practice grounded in experience and interpretive nuance. This highlights the importance of considering annotator expertise when evaluating moderation decisions.

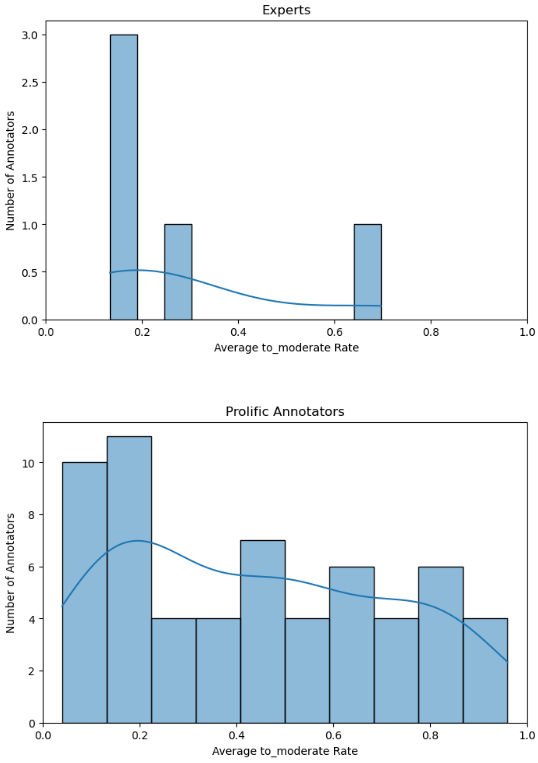


Figure 1: Distribution of `to_moderate` rates across annotators.

Finally, we explored how annotators used comment-level quality properties (Fig. 3). Prolific annotators most frequently flagged clarity, appropriateness, constructiveness, and rationality, while experts showed more heterogeneity, often emphasizing normative or deliberative properties such as common good, reciprocity, or impact. These property selections provide an early signal that experts and crowdworkers may approach moderation from different epistemic and normative standpoints. We pursue this hypothesis in more depth through statistical modeling in Sect. 4.

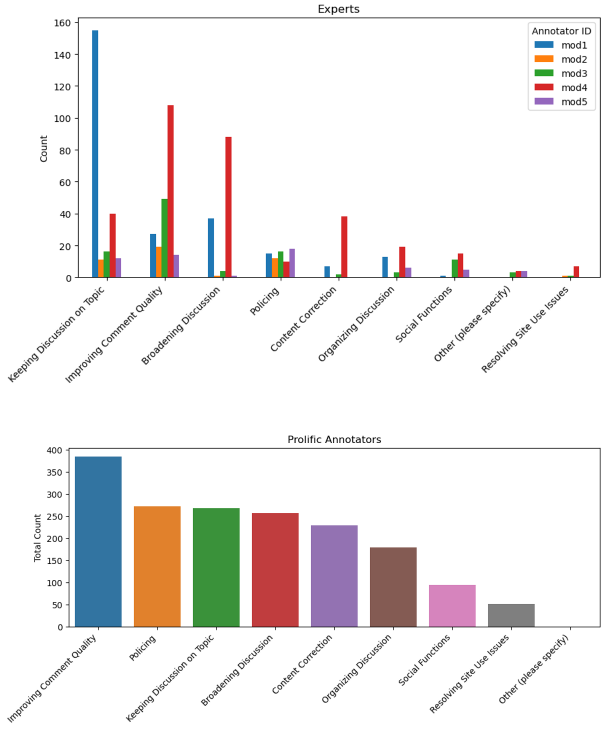


Figure 2: Distribution (experts) or usage (prolific) of moderation function annotations.

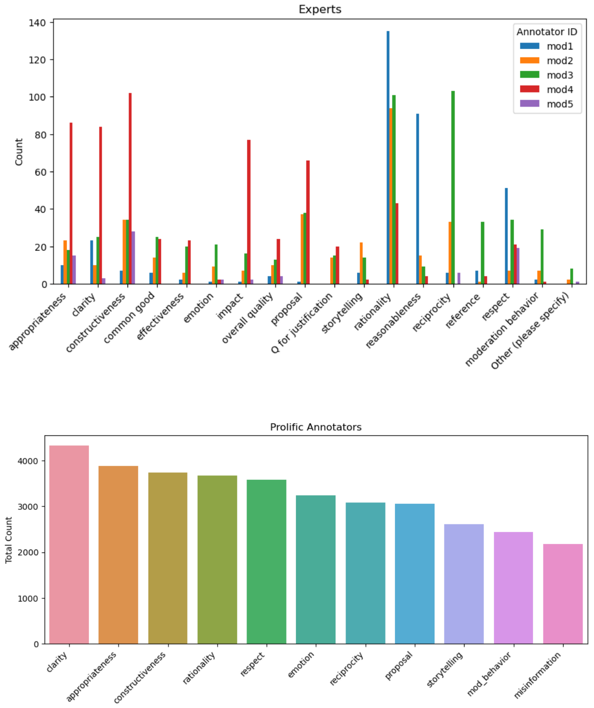


Figure 3: Distribution (experts) or usage (prolific) of comment property annotations.

4 Analysis of Annotator Perspectives

In this section, we employ generalized linear mixed models (GLMMs) to investigate how annotator-related variables (e.g., demographics, expertise, interview responses) and comment-level properties (e.g., deliberative quality dimensions) systematically and significantly influence annotation outcomes (i.e., increasing the likelihood of a comment being moderated).

Each data point in the regression model represents an individual annotation and includes the comment's rated properties and the annotator's background variables (Tbl. 2). We model moderation (`to_moderate`) as a binary outcome. Given the high subjectivity of the task, we introduce annotator ID and comment ID as random effects to account for individual biases and comment-specific variability.

This analysis helps us gain a better understanding of the subjectivity in the dataset and discover more general patterns that can inform the development of AI-assisted moderation: for example, do annotators of a certain level of expertise prioritize moderation of a different type of comment than lay people? Do annotators who frequently engage in online discussion have a different perspective on what to moderate than less active users?

4.1 Comment-related Properties

We first analyze which comment properties (aspects of deliberative/discursive quality) most influence moderation decisions and how this differs between crowdworkers and experts. For each group, we fit individual models using each property as a fixed effect – continuous for crowdworkers and categorical for experts.

Crowdworkers all properties significantly improved model fit ($R^2 > 2\%$) and are significantly better than a baseline model with no fixed effects[7]. Properties linked to harmful content, e.g., disrespect, inappropriateness, and misinformation, explained the most variance (14–21%), followed by core deliberative features like rationality, constructiveness, and proposals (10–13%). Properties that describe the style or type of argument such as emotion, storytelling, and clarity explained less (2–6%).

After identifying the impact of each property in isolation, we try to find the most explanatory model with several comment properties as predictors to

[7] Baseline model includes only random effects, we measure significance of improvement and a lower AIC value using the `model.comparison()` function of `dustinfife/flexplot`

compare their effect in combination and identify the most significant properties. A combined model identified misinformation, respect, and appropriateness as the strongest predictors, suggesting a focus on "negative moderation", i.e., intervening when civility is threatened. Emotional tone and low clarity also increased moderation likelihood, while comments that contain concrete suggestions or proposals for solutions decreased it (effect plot and relative importance of each predictor in App. Fig. 6, Tab. 7). The fixed effects (comment properties) of the best model explained 38% of the variance, indicating that the annotation task was designed well to explain the annotation behavior and that a diverse set of discursive properties guides crowdworker's moderation decisions. However, over half of the remaining variance (ICC = 0.55) is explained by random effects, indicating that it is important to investigate the characteristics of the annotators and how they systematically influence moderation decisions.

Experts For expert moderators, fewer properties were statistically significant or explanatory. Relevant factors included appropriateness, respect, constructiveness, presence of concrete proposals, emotions, reciprocity, and moderation behavior (each contributing 1–5%). The best-fitting model contained four quality dimensions as fixed effects (constructiveness, appropriateness, emotion, and reciprocity), explaining 13% of the variance, while 30% was attributable to annotator-specific differences. Although not directly comparable to the crowdworker setting (given differences in task design), expert moderators appeared to emphasize *positive moderation*, i.e., proactive, additive interventions aimed at enhancing discourse (e.g., fostering constructiveness, encouraging participation) (Strandberg et al., 2019; Mansbridge, 2010; Dillard, 2013; Kuhar et al., 2019; Boulianne et al., 2020). This stands in contrast to *negative moderation*, which generally focuses on removal or suppression of harmful content (Schmidt and Wiegand, 2017; Masud et al., 2024; Mostafazadeh Davani et al., 2024; Park et al., 2021; Gorwa et al., 2020). Constructiveness emerged as the strongest factor, suggesting that experts prioritize interventions that enrich discussion quality over those that police or restrict contributions (see effect plots and predictor importance in App. Fig. 8, Tab. 9).

4.2 Annotator-related Properties

In the following, we examine whether variation in crowdworker moderation behavior can be explained by annotator-related (i.e., demographics, expertise, task interpretations) or by interactions between these and comment properties (e.g., whether more experienced users are more sensitive to comment characteristics).

To start, we modeled each annotator-related variable individually, testing whether it significantly improved model fit over a baseline with only random effects (for both annotator ID and comment ID, as in the previous analysis). While some variables explained a small portion of variance, none of them produced a strong improvement, except for a weak trend showing that annotators who spent more time on the task were slightly more likely to moderate.

We then tested for interaction effects between annotator-specific variables and comment properties. Specifically, we asked whether adding each interaction significantly improved model fit, accounting for both complexity and predictive power of the regression model (lower AIC, significance of interaction). Below, we summarize key results from significant interactions.[8]

Active participants are more sensitive to appropriateness and moderation behavior. Annotators who actively engage in online discussions (reply and engage vs. users who only read comments) are more sensitive to appropriateness – more so than passive users for which this has less impact on the moderation choices (compare the orange with the green line in the interaction plot, Fig. 4). Active users also interpret moderation-like behavior as a positive signal, decreasing their likelihood to intervene, while passive users are less influenced by this feature (App. Fig. 10a).

A similar, even more divergent trend appears with annotator confidence and task difficulty. Annotators who were confident or found the annotation task easier were less likely to moderate comments showing moderation behavior, while those who struggled were more likely to flag them (App. Fig. 10b, 10c). This points to an interesting controversy: moderation-like behavior (users pointing out inaccuracies to others, drawing attention to potential rule violations) is seen by some as a constructive contribution to a good discussion environment, while others may view it as moral

[8] All effect plots and full model details are in Appendix E.

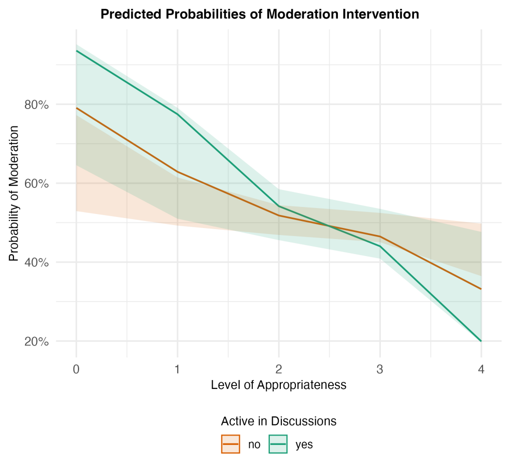


Figure 4: Predicted probabilities of moderation decisions by appropriateness level and online discussion experience, based on a mixed-effects model. The plot shows the interaction effect using marginalized fixed effects (estimated with the ggeffects package). Shaded areas indicate 95% confidence intervals.

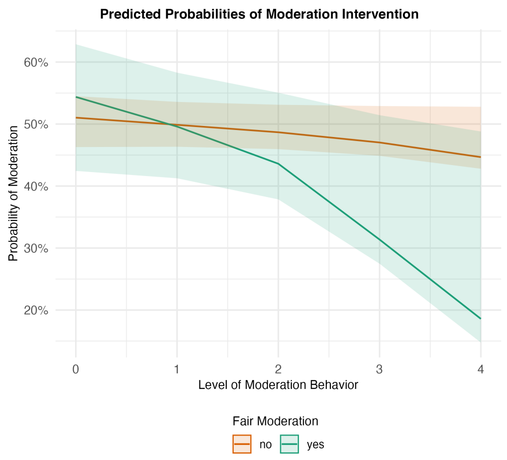


Figure 5: Predicted probabilities of moderation decisions by strength of moderation behavior and importance of fair moderation for a good discussion experience, based on a mixed-effects model. The plot shows the interaction effect using marginalized fixed effects, shaded areas indicate 95% confidence intervals.

policing, intrusive, and therefore worthy of moderation.

Values shape sensitivity to emotion, appropriateness, and clarity. When considering interactions between comment properties and annotators personal opinion about what constitutes a 'valuable' comment, we find different effects. Annotators who value diverse perspectives are more likely to moderate emotionally charged comments and less likely to moderate emotionally neutral ones (App. Fig. 10f). Annotators who pay more attention to a comment being supported with evidence or facts are tendentially more sensitive towards all comment properties (stronger negative or positive relationship) but mostly to appropriateness or clarity, clearly picking up on those signals when they decide for a moderation intervention (App. Fig. 11c-12b).

Sensitivity to misinformation is lower among experienced moderators and higher among student annotators. Although counterintuitive, we find that annotators with moderation experience are less likely to flag comments for misinformation, even at high levels (App. Fig. 10d). This suggests that they consider a broader set of contextual and constructive cues, while less experienced annotators rely more on clear negative signals. Moreover, we find that students are less likely to moderate at low levels of misinformation levels than non-students. As misinformation increases, both groups are more likely to moderate, but students show a steeper increase, indicating higher sensitivity (App. Fig. 10e).

Quality- or fairness-oriented annotators value moderation behavior. we find that for annotators who care a lot about the quality of comments and who believe that moderation decisions need to be clear, fair and transparent in order to achieve a good user experience are more sensitive to moderation behavior – when the comment signal this phenomenon, their probability to moderate is significantly lower than for annotators who do not focus on those aspects (see decreasing green line in Fig. 5, and interaction plots App. Fig. 13e, 13f).

Views on AI shape moderation sensitivity. Annotators who see AI as a tool for flagging harmful content are more sensitive to emotion and misinformation, viewing them as threats to civility. They tend to see a strong affective tone and unverified or false claims in comments as potential sources of harm that can undermine civility and the idea of the discussion being a 'safe space' (App. Fig. 12c, 12d). In contrast, those who view AI as supporting complex, context-aware moderation are less reactive to individual cues like appropriateness or rationality, instead considering a broader mix of signals (App. Fig. 12e, 12f). This finding is also supported by annotators opinions on bottlenecks a human moderator has to face: annotators who believe that interpretation of comments can be difficult and that finding a balance between freedom of speech and protecting users from harmful comments are more sensitive to comment properties, such as appropriateness or respect, again clearly picking up on negative cues to decide whether to intervene (App. Fig 13d).

In sum, while the variation in annotation of this phenomenon is very complex, we identify two main annotator profiles: those who rely on clear negative cues (misinformation, inappropriateness) and intervene accordingly, and those who weigh a broader mix of constructive and harmful traits, such as moderation-like behavior. While most show consistent patterns in how quality dimensions influence decisions, moderation behavior remains controversial: some highly value it, while others see it as a reason to intervene. Although many annotator-related factors (socio-demographics, background, and relationship to the task) have been accounted for in the analysis, we still observe a large amount of unexplained variance and substantial differences between annotators. This highlights the complexity of moderation as a task. Factors like broader context, reader fatigue, the topic at hand, the platform format, etc. might influence annotators' decisions.

5 Conclusion

This paper introduced `PerspectiveMod`, a perspectivist dataset for studying moderation in deliberative discourse. It goes beyond the limited existing resources by capturing diverse moderator expertise and enriching annotator profiles with task-specific experience and attitudes, addressing a crucial research gap.

Rather than pursuing scale alone, our focus is on capturing the variability and subjectivity inherent to constructive moderation. Moderation is not a uniform task; its norms and thresholds vary across users, communities, and platforms. `PerspectiveMod` reflects that variation and supports more context-aware approaches to modeling. Our findings reveal the deep subjectivity and complexity of moderation decisions, calling for future models that are fine-grained (sensitive to diverse discourse properties) and non-normative, serving as decision aids rather than replacements for human judgment.

6 Limitations

While `PerspectiveMod` offers a valuable resource for studying subjective and context-sensitive moderation, several limitations should be acknowledged. First, the dataset is relatively small in scale, which may constrain generalizability, particularly for training large-scale models. Our aim was depth and interpretive richness rather than breadth; however, future work could expand the data across more platforms or discourse settings.

Second, while we deliberately collected detailed background information from annotators, including their experience, values, and attitudes, there are likely unobserved factors that influence moderation decisions, such as mood, platform familiarity, or implicit biases. Capturing the full range of influences on human judgment remains a methodological challenge, and some degree of residual subjectivity is inevitable.

Third, although we included both expert and crowdworker perspectives, our annotator pool remains limited in geographic and cultural diversity. This may impact how certain norms (e.g., tone, politeness, argumentative style) are interpreted. More globally diverse annotations would strengthen claims about general moderation principles across communities.

Finally, we do not attempt to model moderation decisions in this work. Our focus is on laying the groundwork through annotation and analysis. While we highlight the obstacles and requirements for future modeling, designing systems that meaningfully engage with the perspectivist nature of moderation remains an open and complex challenge.

7 Ethics Statement

This research was conducted with careful attention to ethical considerations in data collection, annotation, and participant compensation. All annotators provided informed consent prior to participation. Crowdworkers were recruited via Prolific and compensated fairly according to local wage standards, with additional bonuses for thoughtful engagement. Expert annotators were compensated at a rate consistent with professional consulting.

To protect privacy, no personally identifying information was collected beyond general background categories (e.g., profession, experience). Free-text responses were anonymized prior to analysis and coding.

The datasets used in this work are drawn from public online platforms with moderation behavior already visible or implicit in platform interactions. Our annotation task focuses on moderation intent and quality dimensions, not on personal identity or protected categories.

We acknowledge that any research involving subjective annotation can reflect existing biases. Our approach explicitly embraces a perspectivist framework to document, rather than suppress, this variability. We aim to foster responsible future use of this dataset by making annotations and guidelines publicly available with appropriate caveats regarding subjectivity and context sensitivity.

Acknowledgements

We acknowledge funding by the Bundesministerium für Forschung, Technologie und Raumfahrt (BMFTR) through the project E-DELIB.

References

Gavin Abercrombie, Nikolas Vitsakis, Aiqi Jiang, and Ioannis Konstas. 2024. Revisiting annotation of online gender-based violence. In *Proceedings of the 3rd Workshop on Perspectivist Approaches to NLP (NLPerspectives) @ LREC-COLING 2024*, pages 31–41, Torino, Italia. ELRA and ICCL.

Lora Aroyo and Chris Welty. 2015. Truth is a lie: Crowd truth and the seven myths of human annotation. *AI Magazine*, 36(1):15–24.

Joris Baan, Wilker Aziz, Barbara Plank, and Raquel Fernandez. 2022. Stop measuring calibration when humans disagree. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 1892–1915, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

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.

Helena Bonaldi, Giuseppe Attanasio, Debora Nozza, and Marco Guerini. 2023. Weigh your own words: Improving hate speech counter narrative generation via attention regularization. In *Proceedings of the 1st Workshop on CounterSpeech for Online Abuse (CS4OA)*, pages 13–28, Prague, Czechia. Association for Computational Linguistics.

Helena Bonaldi, Yi-Ling Chung, Gavin Abercrombie, and Marco Guerini. 2024. NLP for counterspeech against hate: A survey and how-to guide. In *Findings of the Association for Computational Linguistics: NAACL 2024*, pages 3480–3499, Mexico City, Mexico. Association for Computational Linguistics.

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.

Federico Cabitza, Andrea Campagner, and Valerio Basile. 2023. Toward a perspectivist turn in ground truthing for predictive computing. *Proceedings of the AAAI Conference on Artificial Intelligence*, 37(6):6860–6868.

Bharathi Raja Chakravarthi and Vigneshwaran Muralidaran. 2021. Findings of the shared task on hope speech detection for equality, diversity, and inclusion. In *Proceedings of the First Workshop on Language Technology for Equality, Diversity and Inclusion*, pages 61–72, Kyiv. Association for Computational Linguistics.

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.

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

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.

Simona Frenda, Alessandro Pedrani, Valerio Basile, Soda Marem Lo, Alessandra Teresa Cignarella, Raffaella Panizzon, Cristina Marco, Bianca Scarlini, Viviana Patti, Cristina Bosco, and Davide Bernardi.

2023. EPIC: Multi-perspective annotation of a corpus of irony. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 13844–13857, Toronto, Canada. Association for Computational Linguistics.

Dennis Friess, Marc Ziegele, and Dominique Heinbach. 2020. Collective civic moderation for deliberation? Exploring the links between citizens' organized engagement in comment sections and the deliberative quality of online discussions. *Political Communication*, 38(5):624–646.

Lodewijk Gelauff, Mohak Goyal, Bhargav Dindukurthi, Ashish Goel, and Alice Siu. 2024. Estimating contribution quality in online deliberations using a large language model. In *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing*, volume 12, pages 65–74.

Robert Gorwa, Reuben Binns, and Christian Katzenbach. 2020. Algorithmic content moderation: Technical and political challenges in the automation of platform governance. *Big Data & Society*, 7(1):2053951719897945.

Timon Gurcke, Milad Alshomary, and Henning Wachsmuth. 2021. Assessing the sufficiency of arguments through conclusion generation. In *Proceedings of the 8th Workshop on Argument Mining*, pages 67–77, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Amy Gutmann and Dennis Thompson. 2004. *Why deliberative democracy?* Princeton University Press, Princeton, NJ.

Jürgen Habermas. 1981. *Theorie des kommunikativen Handelns: Handlungsrationalität und gesellschaftliche Rationalisierung*. Suhrkamp.

Ivan Habernal and Iryna Gurevych. 2016. What makes a convincing argument? Empirical analysis and detecting attributes of convincingness in web argumentation. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 1214–1223, Austin, Texas. Association for Computational Linguistics.

Ming Shan Hee, Shivam Sharma, Rui Cao, Palash Nandi, Preslav Nakov, Tanmoy Chakraborty, and Roy Ka-Wei Lee. 2024. Recent advances in online hate speech moderation: Multimodality and the role of large models. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 4407–4419, Miami, Florida, USA. Association for Computational Linguistics.

Tiancheng Hu and Nigel Collier. 2024. Quantifying the persona effect in LLM simulations. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 10289–10307, Bangkok, Thailand. Association for Computational Linguistics.

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

Andreas Jungherr and Adrian Rauchfleisch. 2025. Artificial intelligence in deliberation: The ai penalty and the emergence of a new deliberative divide. *Preprint*, arXiv:2503.07690.

Gauri Kambhatla, Matthew Lease, and Ashwin Rajadesingan. 2024. Promoting constructive deliberation: Reframing for receptiveness. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 5110–5132, Miami, Florida, USA. Association for Computational Linguistics.

Varada Kolhatkar and Maite Taboada. 2017a. Constructive language in news comments. In *Proceedings of the First Workshop on Abusive Language Online*, pages 11–17, Vancouver, BC, Canada. Association for Computational Linguistics.

Varada Kolhatkar and Maite Taboada. 2017b. Using New York Times picks to identify constructive comments. In *Proceedings of the 2017 EMNLP Workshop: Natural Language Processing meets Journalism*, pages 100–105, Copenhagen, Denmark. Association for Computational Linguistics.

Varada Kolhatkar, Hanhan Wu, Luca Cavasso, Emilie Francis, Kavan Shukla, and Maite Taboada. 2020. The sfu opinion and comments corpus: A corpus for the analysis of online news comments. *Corpus pragmatics*, 4:155–190.

Katerina Korre, Dimitris Tsirmpas, Nikos Gkoumas, Emma Cabalé, Dionysis Kontarinis, Danai Myrtzani, Theodoros Evgeniou, Ion Androutsopoulos, and John Pavlopoulos. 2025. Evaluation and facilitation of online discussions in the llm era: A survey. *Preprint*, arXiv:2503.01513.

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 Anees AbuHashem, and Zakir Durumeric. 2024. Watch your language: Investigating content moderation with large language models. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 18, pages 865–878.

Ji-Ung Lee, Jan-Christoph Klie, and Iryna Gurevych. 2022. Annotation curricula to implicitly train non-expert annotators. *Computational Linguistics*, 48(2):343–373.

Noah Lee, Na Min An, and James Thorne. 2023. Can large language models capture dissenting human voices? In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 4569–4585, Singapore. Association for Computational Linguistics.

Jane Mansbridge. 2010. Deliberative polling as the gold standard. *The good society*, 19(1):55–62.

Todor Markov, Chong Zhang, Sandhini Agarwal, Florentine Eloundou Nekoul, Theodore Lee, Steven Adler, Angela Jiang, and Lilian Weng. 2023. A holistic approach to undesired content detection in the real world. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pages 15009–15018.

Sarah Masud, Sahajpreet Singh, Viktor Hangya, Alexander Fraser, and Tanmoy Chakraborty. 2024. Hate personified: Investigating the role of LLMs in content moderation. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 15847–15863, Miami, Florida, USA. Association for Computational Linguistics.

Aida Mostafazadeh Davani, Mark Diaz, Dylan K Baker, and Vinodkumar Prabhakaran. 2024. D3CODE: Disentangling disagreements in data across cultures on offensiveness detection and evaluation. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 18511–18526, Miami, Florida, USA. Association for Computational Linguistics.

Matthias Orlikowski, Paul Röttger, Philipp Cimiano, and Dirk Hovy. 2023. The ecological fallacy in annotation: Modeling human label variation goes beyond sociodemographics. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 1017–1029, Toronto, Canada. Association for Computational Linguistics.

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

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*. ACM.

Nuria Rodríguez-Barroso, Eugenio Martínez Cámara, Jose Camacho Collados, M. Victoria Luzón, and Francisco Herrera. 2024. Federated learning for exploiting annotators' disagreements in natural language processing. *Transactions of the Association for Computational Linguistics*, 12:630–648.

Anna Schmidt and Michael Wiegand. 2017. A survey on hate speech detection using natural language processing. In *Proceedings of the Fifth International Workshop on Natural Language Processing for Social Media*, pages 1–10, Valencia, Spain. Association for Computational Linguistics.

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.

Kim Strandberg, Staffan Himmelroos, and Kimmo Grönlund. 2019. 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.

Matthias Trénel. 2009. Facilitation and inclusive deliberation. *Online deliberation: Design, research, and practice*, pages 253–257.

Dimitris Tsirmpas, Ion Androutsopoulos, and John Pavlopoulos. 2025. Scalable evaluation of online moderation strategies via synthetic simulations. *arXiv preprint arXiv:2503.16505*.

Alexandra N. Uma, Tommaso Fornaciari, Dirk Hovy, Silviu Paun, Barbara Plank, and Massimo Poesio. 2021. Learning from disagreement: A survey. *Journal of Artificial Intelligence Research*, 72:1385–1470.

Nikolas Vitsakis, Amit Parekh, and Ioannis Konstas. 2024. Voices in a crowd: Searching for clusters of unique perspectives. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 12517–12539, Miami, Florida, USA. Association for Computational Linguistics.

Henning Wachsmuth, Nona Naderi, Yufang Hou, Yonatan Bilu, Vinodkumar Prabhakaran, Tim Alberdingk Thijm, Graeme Hirst, and Benno Stein. 2017. Computational argumentation quality assessment in natural language. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*, pages 176–187, Valencia, Spain. Association for Computational Linguistics.

Marcos Zampieri, Shervin Malmasi, Preslav Nakov, Sara Rosenthal, Noura Farra, and Ritesh Kumar. 2019. Predicting the type and target of offensive posts in social media. *arXiv preprint arXiv:1902.09666*.

Timon Ziegenbein, Shahbaz Syed, Felix Lange, Martin Potthast, and Henning Wachsmuth. 2023. Modeling appropriate language in argumentation. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4344–4363, Toronto, Canada. Association for Computational Linguistics.

A Moderator recruitment efforts

To cast a big net, we contacted multiple different forums and administrators. For this procedure, we tailored a general message about the research group and the key points of the moderation study (goals, tasks, timeline, compensation) to the specifics of the respective forum. This tailoring included a level of formality, e.g., between self-moderated online forums and professionally moderated news comment sections, as well as the address and level of detail on the task, i.e., if the forum did not indicate that our email/message directly reached moderators, we shortened the task description, removed a direct link to the task and instead included a PDF for convenient forwarding of our request. An example email and forum message are shown in Tab. 3. We wrote to three main groups:

1. **News outlets** The websites of many newspapers or television stations include a discussion area below the articles or in a separate forum. As comments, especially regarding political news, can quickly become polarized, these areas usually come with detailed etiquette guidelines and are monitored by professional moderators. These moderators are either in-house experts or contracted from specialized moderation firms. To ensure the best possible success, after extensive research, we only contacted those news agencies whose websites contain a clear reference to an internal moderation team:
    - New York Times (`rd@nytimes.com`)
    - Washington Post (`comments@washpost.com`)
    - The Guardian (`moderation@theguardian.com`)
    - BBC (`Central-communities-team@bbc.co.uk`)
2. **Reddit forums** Reddit hosts a platform for specialized forums (*subreddits*) where communities form around specific topics or themes. A popular subreddit in the field of computational argumentation is *r/ChangeMyView*, which is dedicated to open debate. There are, however, multiple such forums that are dedicated to debate or discussion at large instead of a specific topic. As Reddit does not employ expert moderators, and the automated moderation (automod) is regex-based (i.e., only applies to regex-recognizable rule violations), most subreddits with more sophisticated guidelines for debate have community moderators. These are active members of the subreddit with administrative privileges (deleting and blocking comments or users), who volunteer to monitor the discussions for rule violations and reprimand users. A *direct message* addressed to a subreddit is seen by all current moderators. While searching for suitable forums that deal with debate and argumentation, we noticed a non-negligible overlap in the moderator base between related forums. We still decided to contact multiple subreddits to increase our chances of recruitment, which at first seemed promising, as multiple moderators asked clarifying questions about the definitions of moderation and argumentation we used and were interested in the annotation guidelines. However, we were only successful in recruiting one moderator through three Calls for Participation that were spaced out over two months.
    - r/ChangeMyView (three calls)
    - r/NeutralPolitics (two calls)
    - r/PoliticalDebate (one call)
    - r/PoliticalDiscussion (two calls)
    - r/WinMyArgument (two calls)
3. **Industry experts** Most deliberation studies are collaborations between researchers and deliberation platforms or regulators, which always employ continuous expert moderation. Thus, there is a potentially big network of researchers with connections to companies hosting participation platforms, community management professionals, or expert moderators themselves. We tried to use this network by contacting our own collaborators and colleagues in research (four people), as well as directly writing to participation agencies (three agencies) with our call for moderators. Of all possible avenues, this should be most successful, given our direct acquaintance with each contact and the established connection with moderation in any capacity, but we only gained one moderator from this approach.

	Call for Moderators
(1)	Hello [*Name*], I am writing on behalf of the [*group*] project, a group at [*affiliation*]. Our team is researching Natural Language Processing (NLP) with a focus on discussion moderation. We aim to develop AI tools to assist expert moderators, enabling them to better support online community discussions. We are currently conducting a research study to improve our approach and are seeking expert moderators, like those [*on the New York Times Community team*], to participate. For the study, we would ask moderators to review a set of comments our system predicts need or do not need moderation. Moderators would then answer questions about whether and why a comment might need moderation. This will help us assess the effectiveness of our tool. We estimate the task will take up to 8 hours. Participation and compensation options: • Moderators can register on prolific.com, a widely-used platform for studies, and be compensated at €12 per hour (preferred). • Moderators can participate directly and receive an Amazon voucher of equivalent value. If you believe your moderators might be interested, we would greatly appreciate it if you could share this request with them. Interested moderators can contact us directly at [*email*] to access the study. I am also available at any time to answer any further questions or concerns. Thank you for your time and assistance. Best regards, [*Name*]
(2)	**Annotation Task for Moderation Research - new PayPal payment option** Dear Moderators, We are researchers in AI and natural language processing (NLP) from the [*affiliation*][1], and we are working on computational approaches to make your moderation task easier and more efficient, allowing you to provide more support to the online community. We are looking for moderators to help with our (paid) annotation task! You might have already seen our Call a month ago, but we now have a new option to compensate you directly via PayPal. **The Task** All we need is for you to go through a set of (around 200) comments that we predict do (or do not) need moderator intervention in the discussion (you won't know which). You will then answer a few questions to see if having our tool would actually help your experience as a moderator. If you want to know more, you can have a closer look at the [guidelines](*url*). **Compensation** We predict up to 8 hours of work, and we will compensate 12€/hr (eq 13$/hr, eq 10.10£/hr). **How to participate** You have two options: • **Prolific** The annotation task and compensation can be carried out via prolific.com, a widely-used platform for annotation and survey studies. You can register and access the study on prolific via this [*access link*]. Compensation for participating in this study will be processed through Prolific. • **Direct Link** If you do not want to register at prolific, you can reach us here on Reddit or via email at [*email*] for a direct access link to the study. We can then compensate you directly on Paypal or with an Amazon voucher of equivalent value. If you're interested in participating or have any further questions, please don't hesitate to contact us directly. ~ [*Name*] [1] To learn more information about our research and our team, you can visit the [*affiliation*](*url*). I am also happy to answer any questions you may have.

Table 3: Examples of calls for moderators sent out as emails to newspaper moderation teams and hosts of deliberation platforms (1), or as direct messages to community moderation teams on online forums (2).

Therefore, we posted a last call for annotators directly to the forums of the *International Association of Facilitators*[9]. These facilitators received professional training in facilitation and mediation strategies, though the focus of their work is not necessarily on debate – instead, many members of the organization specialize in business applications, e.g., leading brainstorming sessions, mediating discussions on the future of a company or coaching teams and leadership in successful teamwork. This shifted focus does not, however, detract from the expertise of IAF members, two of which we could recruit for our study.

B Annotation Task Materials

All material used for the annotation tasks, both for expert moderators and for the crowdsourced annotations, are available here: https://github.com/emvecchi/PerspectiveMod. The set of materials includes: full annotation guidelines, the tutorials provided to the crowdworker annotators, the introductory and concluding interview questions, all expert moderator recruitment material, and a screenshot of each annotation form for a sample instance.

C Coding of interview answers

To assign the interview questions to high-level categories, two of the paper authors developed an annotation schema by following an inductive approach.

As a first step, for each of the four interview question employed in the statistical analysis, we draw an annotation sample from the annotator responses. We embedded the 62 answers with SBERT and carried out k-means clustering setting the number of desired clusters to 3. Note that this is not the final number of clusters, but just a sampling step (indeed, the final annotation schema contain more than 3 categories

[9] https://www.iaf-world.org/

per interview questions). For each cluster we then we sampled four items closer to the centroid and four random, to maximize diversity of the items employed to develop the coding scheme.

As a second step, each of the two annotators independently developed a set of categories by inspecting the sample of answers for each interview questions. They then compared their categorization, aligning labels and if necessary merging them, resulting in the final categorization.

As a final step, the two annotators applied the final categorization to each interview question.

In what follows, we list the annotation categories for each interview question, along with definitions and examples.

Interview questions required the annotators to put themselves into the moderator perspective.

Introductory question: what do you feel contributes for a good experience for the users/discussion?

- `respect/safe space`: emphasis on a respectful environment in which users feel comfortable in expressing their opinions: Example: "the users feel safe expressing their opinions without fear of personal attacks or hostility"

- `moderation fairness`: moderator actions do not create an advantage for some users: Example: "Treating everyone with equal privileges"

- `moderation clarity`: a good experience is ensured by community requirements clearly expressed in the guidelines, and by moderators who make their decisions understandable to the users. Example: "A good experience is shaped by clear, fair, and consistent moderation [...] (also labeled as `Moderation fairness`).

- `comment quality`: a good experience is connected to the overall quality of the discussion, and it is the role of the moderator to intervene to promote good quality comments and ensure, for example, that the discussion stays on topic. Examples: "Praising comments that share insightful views or highlight factual data"; "Easy understanding and not boring".

- `diversity`: a good experience in a discussion is defined as one that maximises inclusivity, in which all voices are heard. It is exactly the task of the moderator to ensure that this happens. Example: "That subjects are discussed with many people adding their opinion. Users feel safe and look forward to reading comments." (also labeled as `Respect/safe space`).

Introductory question: what makes a comment or contribution valuable?

- `topic relevance`: a valuable comment is one that is on topic. Example: "Comments or contributions which are clear and concise to the ongoing discussions. Keeping it relevant makes it so no one is confused about any being spoken about. For example, you wouldn't talk about cars in a discussion aimed at fruit salad recipes."

- `evidence support`: a valuable comment is one whose evidence is backed up by facts. Example: "If it shares factual data from research, linking resources or comments that driving deeper conversations."

- `adding value`: a valuable comment is one that triggers progress in the discourse, for example by uncovering a previously overlooked angle of the discussion, Example: "A comment or contribution is valuable when it enriches the discussion, fosters understanding , or helps achieve the goals of the conversation." "A valuable comment or contribution is something that gives both the user and the community value, for example an input that not only allows a user to contribute, but it adds to a topic and acts as a resource for current and future users, such as a comment that can be pinned."

- `inclusivity`: Example: "What makes a comment valuable is the consideration of multiple perspectives on an issue."

- `generic quality`: this category subsumes finer-grained aspects of quality. Example: "A comment is valuable when it has some meaning and when does not create any form of violence."

Concluding question 2: beyond the goal of this research and annotation task, what assistance do you feel computational tools (like AI) could provide to your task as a moderator?

- `attention/flagging`: AI can support the moderator by flagging comments that require an intervention. In doing so, it can help the moderators distribute their attention efficiently by indicating which comments need to be prioritized and avoiding that problematic comment slip off the moderator attention. Example: "AI can assist moderators by flagging comments that exhibit certain characteristics, such as offensive language, harmful rhetoric, or potential misinformation. AI tools could also prioritize comments based on severity or context."

- `complex tasks`: AI can support the moderator with more tasks beyond comment flagging, e.g., by providing discussion analytics, suggesting explanations for the decisions of the moderator. Example: "Often repeated sentences or words could be inserted or suggested as I type. [...]"; "Fact-checking would be especially beneficial"

- `bias (user harm)`: AI can support the moderator by assisting in the identification of disruptive content and avoid personal biases, thereby protecting forum users from harm Example: "Computational tools like AI could assist by offering a first line of support to identify comments that require closer attention, highlighting potential issues like hate speech, misinformation, or disruptive behavior.", "[...] AI could help ensure that content is not unfairly removed or flagged in situations where tone is misinterpreted."

- `safety (mod harm)` Example: "AI could provide content flagging and filtering, also reducing moderator burnout" (also labeled as `attention/flagging`)

- `Other`: general statements about AI (not related to the moderation task). Example: "I try not to use AI, like ever, honestly; I'm a graphic designer and in that field alone, it's a slap in the face how people are trying to shove it down your throat for any and all applications. I think AI has uses that would be beneficial, but not every field of study can, nor needs to try to benefit from it. It's a tool for a very niche skill, not a Swiss-army knife."; "The AI can't accurately relate human feelings but can be of use."

Concluding question: what do you think is the largest bottleneck moderators face in online discussions?

- `scaling up`: the challenge of scaling up to discussions with large crowds of participants. Example: "The largest bottleneck moderators face in online discussion is scaling effective moderation across large volumes of content while maintaining context-sensitive, fair, and timely decision making."

- `freedom/harm`: these comments refer to the challenge of identifying disrespectful and harmful comments; some additionally mention the challenge of finding a balance between protect users from harmful content and at the same time ensuring freedom of speech. Example: "What they face is maintaining a balance between ensuring free expression and preventing harm or disruption which involves Dealing with misinformation and disinformation and Burnout and emotional toll on moderators"

- `ambiguity/interpretation`: these comments refer to the challenge of correctly interpreting user comments, which may be ambiguous or simply lack context. Example: "It is having to understanding the nuances and contexts behind some discussions."

- `other`: additional finer-grained aspects that can make a discussion challenging to moderate, such as overuse of personal arguments and in general emotional burden to moderator. Example: "People getting into personal arguments instead of discussing the actual talking points "; " [...] and the motional toll [...]"

D Models Considered for Data Sampling

Below, we outline the feature transformers that were used in each multi-task fusion adapters that we considered for instance selection (Section 3.1). Additionally, we provide a performance overview of the trained adapter models on each of the three data sources (Fig. 5, as well as a performance comparison to the original implementation of (Falk and Lapesa, 2023) (Fig. 7).

Fusion Set	Adapter	Description	Pre-trained ST/Training Data
AQ	cogency	Acceptable and sufficient premises to draw a conclusion	falkne/cogency
	effectiveness	Persuasion, rhetorical, emotional appeal	falkne/effectiveness
	quality	General argument quality	falkne/ibm_rank
	overall	General argument quality	falkne/overall
	reasonableness	Contribution to resolution of issues, argument is accepted by universal audience	falkne/reasonableness
	[in]appropriateness	Misses commitment to discussion, uses toxic emotions, misses intelligibility, inappropriate	Appropriateness Corpus (Ziegenbein et al., 2023)
DQ	argumentative	Providing reasons and/or evidence in favor of or against a claim	falkne/argumentative
	cgood	Taking interests of the broader community or utilitarianism based values into account	falkne/cgood
	empathy	Speaker puts themselves in the perspective or emotional state of others	falkne/empathie
	interactivity	Respect towards other participants, reference to other participants arguments	falkne/interactivity
	justification	Rationality, providing reasons, reflection	falkne/justification
	narration	Personal experience, subjective description of an event or situation	falkne/narration
	posEmotion	Positive emotions are contained in the utterance	falkne/posEmotion
	negEmotion	Negative emotions are contained in the utterance	falkne/negEmotion
	proposal	A statement about what or how something is to be done	falkne/proposal
	QforJustification	Asks for the reasons for a statement or action	falkne/QforJustification
	story	Personal experience, subjective description of an event or situation	falkne/story
	reference	Participant refers to another discourse participant	falkne/reflexivity
	respect	Empathy or respect towards groups (e.g. immigrants)	falkne/respect
	socc_constructiveness	Crowd's annotation on constructiveness of an online news comments	SOCC (Kolhatkar et al., 2020)
	umod_constructiveness	High-quality comments, make a contribution to the conversation	UMOD (Falk et al., 2024)
MOD	rr_moderation	If comment has been moderated by the platform's expert moderators	RR (Park et al., 2012)
	cmv_moderation	If comment has been moderated by the platform's community moderators	`r/ChangeMyView`
	umod_moderation	If reply to comment was determined user moderation	UMOD (Falk et al., 2024)
ALLQ	{AQ} STs		
	{DQ} STs		
	clarity	Is it hard or easy to interpret the argument?	falkne/clarity
	impact	User likes / recommendations	falkne/impact

Table 4: Summary of which ST adapters are used in each fusion set tested with the path to the pre-trained model on HuggingFace or a reference to the dataset used for training.

Moderation Data	Fusion Set	$\mathbf{F1}_0$	$\mathbf{F1}_1$	Acc.
RR	AQ	0.51	0.47	0.49
	AQ$_{mod}$	0.66	0.45	0.58
	DQ	0.67	0.48	0.59
	DQ$_{mod}$	0.67	0.47	0.59
	mod$_{ST}$	0.45	0.44	0.45
	mod$_{fusion}$	**0.74**	**0.49**	**0.66**
	allQ	0.58	0.46	0.53
	allQ$_{mod}$	0.64	0.47	0.57
CMV	AQ	0.86	0.63	0.80
	AQ$_{mod}$	0.87	0.65	0.81
	DQ	**0.88**	**0.66**	**0.83**
	DQ$_{mod}$	0.88	0.65	0.82
	mod$_{ST}$	0.87	0.61	0.81
	mod$_{fusion}$	0.87	0.66	0.82
	allQ	0.87	0.64	0.81
	allQ$_{mod}$	**0.89**	**0.66**	**0.83**
UMOD	AQ	0.17	0.56	0.43
	AQ$_{mod}$	0.02	0.60	0.43
	DQ	**0.71**	**0.14**	**0.57**
	DQ$_{mod}$	0.00	0.61	0.44
	mod$_{ST}$	0.03	0.61	0.44
	mod$_{fusion}$	0.00	0.61	0.44
	allQ	0.05	0.56	0.40
	allQ$_{mod}$	0.72	0.00	0.56

Table 5: Performance on moderation prediction task for MT model, trained on each moderation dataset. Best fusion models results for each dataset in bold.

Fusion Set	**F1$_0$**	**F1$_1$**	**Accuracy**
AQ$_{original}$	0.778	0.478	0.702
DQ$_{original}$	0.758	0.554	0.706
AQ	0.766	0.468	0.680
AQ$_{mod}$	0.658	0.520	0.618
DQ	0.776	0.568	0.720
DQ$_{mod}$	0.768	0.566	0.716
mod	0.736	0.588	0.696
allQ	0.754	0.548	0.702
allQ$_{mod}$	0.758	0.582	0.706

Table 7: Comparison with Falk and Lapesa (2023) MT adapter models results on RR moderation for quality data. Average F1 scores and accuracy reported across the 5 splits used in original paper.

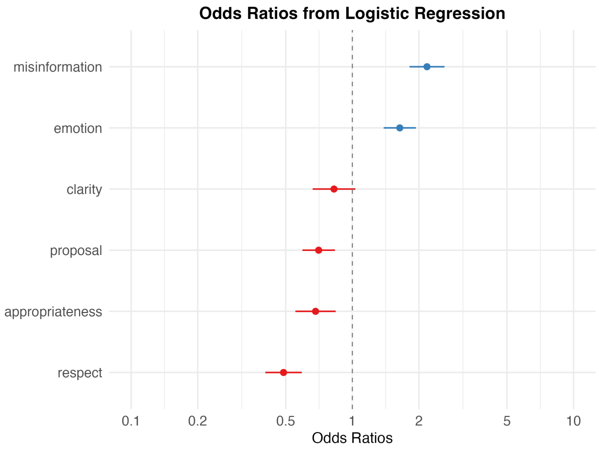


Figure 6: Forest plot showing estimated odds ratios and 95% confidence intervals for predictors the best model predicting when annotators (crowd) moderate (predictors are properties of deliberative quality). Odds ratios below 1 (red) suggest a negative association, so higher values here indicate a lower probability to moderate, while ratios above 1 (blue) indicate a positive relationship (higher values here increase probability to moderate).

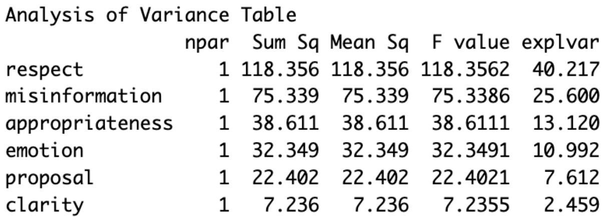
Annotator	**Correlation**	**P-value**
mod1	0.08	0.293
mod2	-0.05	0.501
mod3	-0.08	0.257
mod4	0.05	0.481
mod5	0.03	0.691

Table 6: Spearman correlation results between the 5 expert annotator `to_moderate` judgments and the binary `to_moderate` predictions of the trained adapter model on the full test set. *Sig.: ***p<0.001, **p<0.01, *p<0.05, ·p<0.1.*

E Supplementary Analysis Material

Batch	**Fleiss' κ**	**Krippendorff's α**
1	0.213	0.192
2	0.097	0.114
3	-0.028	-0.021
4	-0.027	0.002
5	0.067	0.073
6	-0.095	-0.059
7	0.039	0.045
8	0.074	0.043
9	0.221	0.195
Experts	0.055	0.040

Table 8: Inter-annotator agreement on the binary `to_moderate` label across Prolific batches and expert annotations.



```
Analysis of Variance Table
                  npar  Sum Sq Mean Sq  F value explvar
respect              1 118.356 118.356 118.3562  40.217
misinformation       1  75.339  75.339  75.3386  25.600
appropriateness      1  38.611  38.611  38.6111  13.120
emotion              1  32.349  32.349  32.3491  10.992
proposal             1  22.402  22.402  22.4021   7.612
clarity              1   7.236   7.236   7.2355   2.459
```

Figure 7: Effect sizes (relative importance) of each predictor in the best model predicting when annotators (crowd) moderate (predictors are properties of deliberative quality, coded as continuous predictors). Model is a mixed-effects model with comment properties as fixed effects and annotator and comment ID as random effects. Fixed predictors account for 38% of the variance, full model (including random effects) explains 72%. Intraclass correlation coefficient (ICC) for annotator ID is 0.54, 0.01 for comment ID.

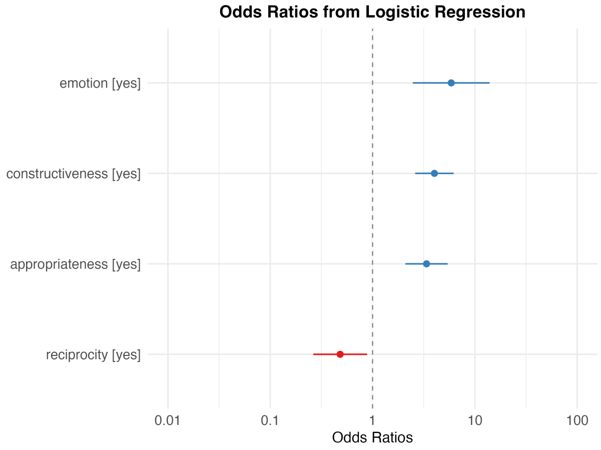
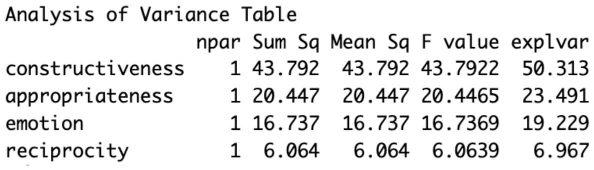
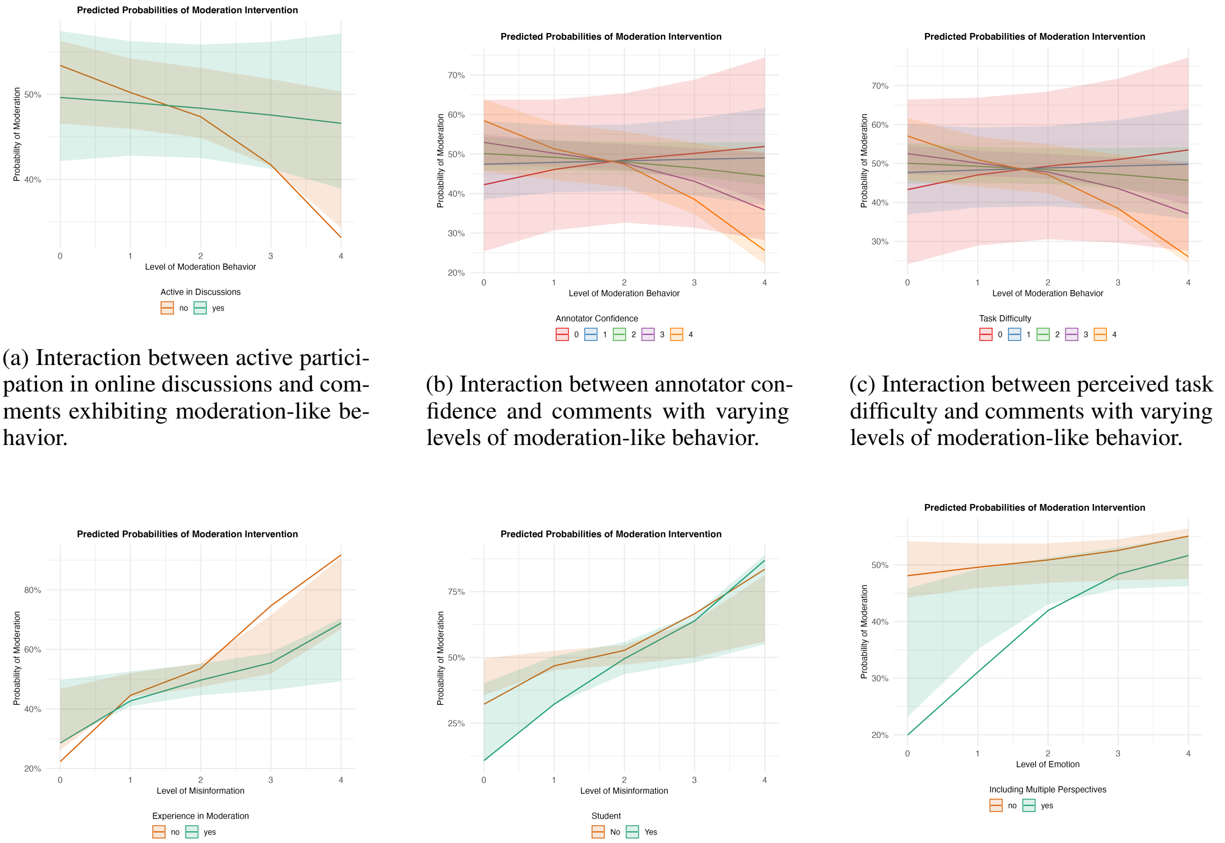


Figure 8: Forest plot showing estimated odds ratios and 95% confidence intervals for predictors the best model predicting when annotators (experts) moderate (predictors are relevant properties of deliberative quality, coded as categorical). Odds ratios below 1 (red) suggest a negative association, so when this property was relevant in moderation decision, the probability to moderate was lower, while ratios above 1 (blue) indicate a positive relationship (experts marking a property as relevant have a higher likelihood to moderate in this case).



```
Analysis of Variance Table
                 npar Sum Sq Mean Sq F value explvar
constructiveness    1 43.792  43.792 43.7922  50.313
appropriateness     1 20.447  20.447 20.4465  23.491
emotion             1 16.737  16.737 16.7369  19.229
reciprocity         1  6.064   6.064  6.0639   6.967
```

Figure 9: Effect sizes (relative importance) of each predictor in the best model predicting when annotators (experts) moderate (predictors are properties of deliberative quality, coded as categorical predictors). Model is a mixed-effects model with comment properties as fixed effects and annotator and comment ID as random effects. Fixed predictors account for 13% of the variance, full model (including random effects) explains 39%. Intraclass correlation coefficient (ICC) for annotator ID is 0.15, 0.15 for comment ID.



(a) Interaction between active participation in online discussions and comments exhibiting moderation-like behavior.

(b) Interaction between annotator confidence and comments with varying levels of moderation-like behavior.

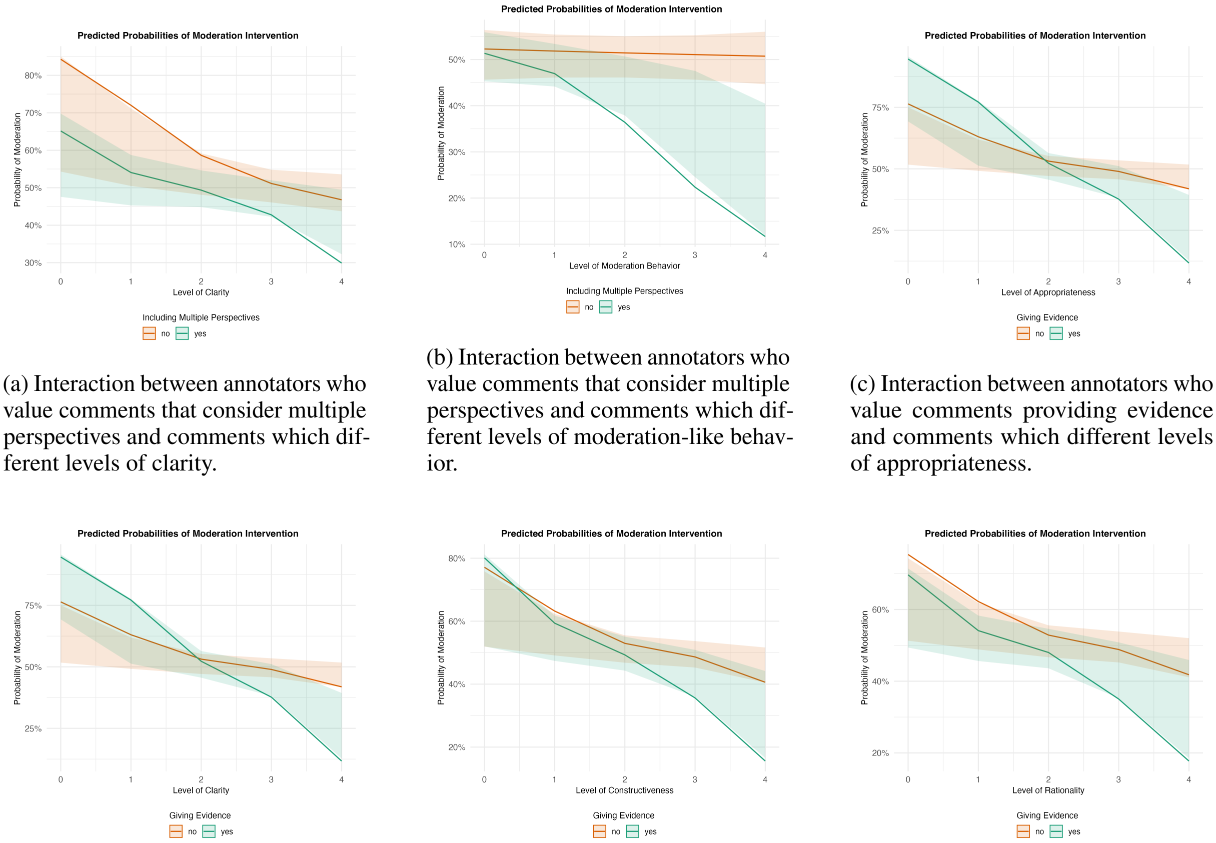
(c) Interaction between perceived task difficulty and comments with varying levels of moderation-like behavior.

(d) Interaction between prior moderation experience and comments with varying levels of misinformation.

(e) Interaction between student annotators and comments with varying levels of misinformation.

(f) Interaction between valuing inclusive perspective-taking and comments with varying levels of emotional content.

Figure 10: Marginalized fixed-effect predictions from mixed-effects models, illustrating interactions between annotator-specific characteristics and comment features (deliberative quality dimensions). Predictions were generated using the `ggeffects` package, averaging over random effects. Shaded areas represent 95% confidence intervals.



(a) Interaction between annotators who value comments that consider multiple perspectives and comments which different levels of clarity.

(b) Interaction between annotators who value comments that consider multiple perspectives and comments which different levels of moderation-like behavior.

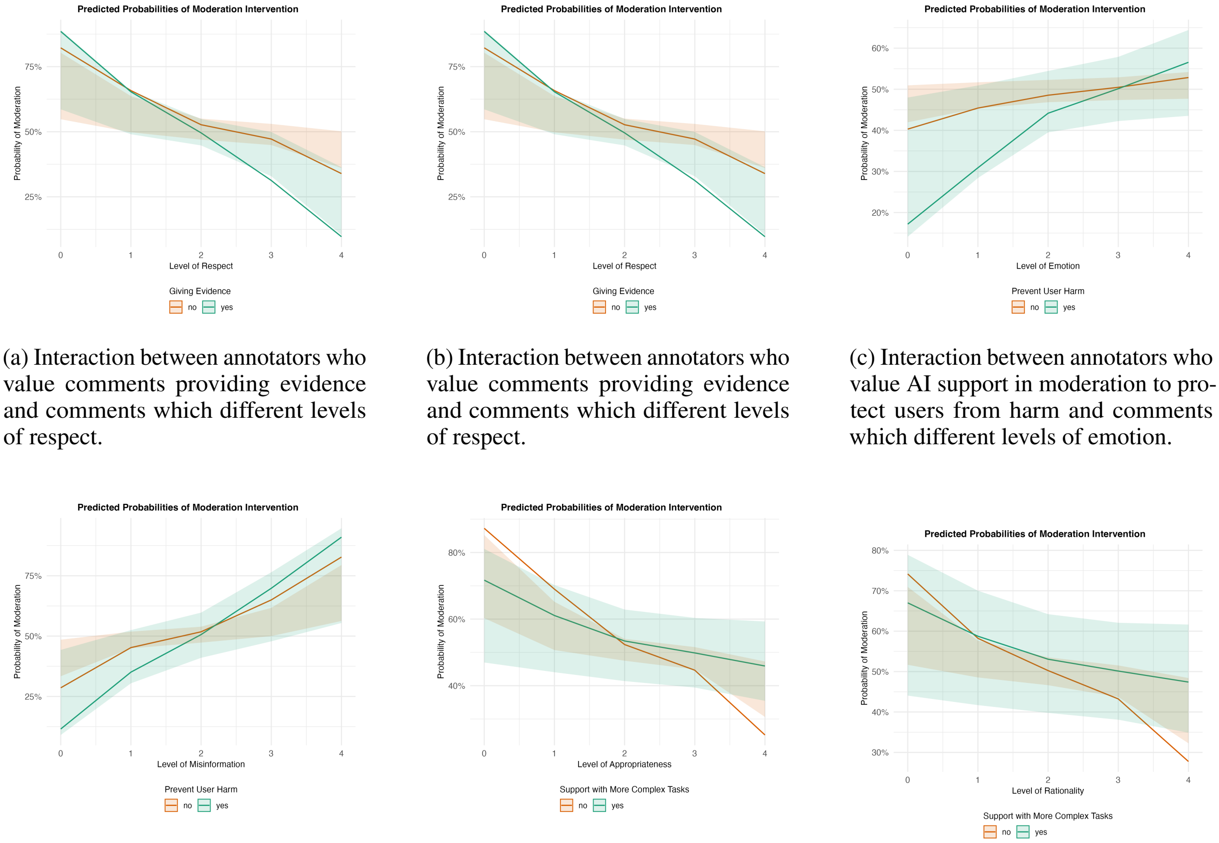
(c) Interaction between annotators who value comments providing evidence and comments which different levels of appropriateness.

(d) Interaction between annotators who value comments providing evidence and comments which different levels of clarity.

(e) Interaction between annotators who value comments providing evidence and comments which different levels of constructiveness.

(f) Interaction between annotators who value comments providing evidence and comments which different levels of rationality.

Figure 11: Marginalized fixed-effect predictions from different mixed-effects models, illustrating the interaction between different annotator-specific variables and comment properties (deliberative quality dimensions). Predictions were generated using the ggeffects package, averaging over random effects. Shaded areas represent 95% confidence intervals.



(a) Interaction between annotators who value comments providing evidence and comments which different levels of respect.

(b) Interaction between annotators who value comments providing evidence and comments which different levels of respect.

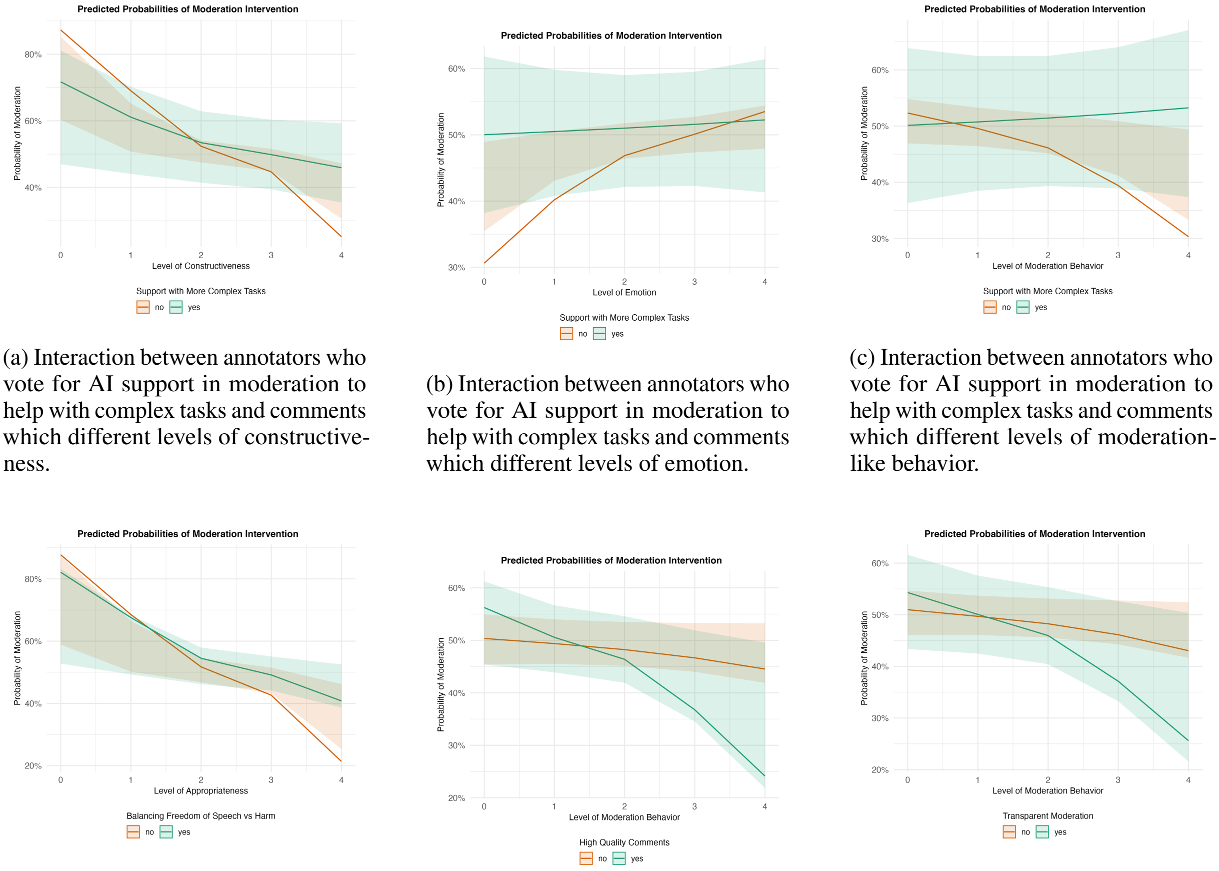
(c) Interaction between annotators who value AI support in moderation to protect users from harm and comments which different levels of emotion.

(d) Interaction between annotators who value AI support in moderation to protect users from harm and comments which different levels of misinformation.

(e) Interaction between annotators who vote for AI support in moderation to help with complex tasks and comments which different levels of appropriateness.

(f) Interaction between annotators who vote for AI support in moderation to help with complex tasks and comments which different levels of rationality.

Figure 12: Marginalized fixed-effect predictions from different mixed-effects models, illustrating the interaction between different annotator-specific variables and comment properties (deliberative quality dimensions). Predictions were generated using the ggeffects package, averaging over random effects. Shaded areas represent 95% confidence intervals.



(a) Interaction between annotators who vote for AI support in moderation to help with complex tasks and comments which different levels of constructiveness.

(b) Interaction between annotators who vote for AI support in moderation to help with complex tasks and comments which different levels of emotion.

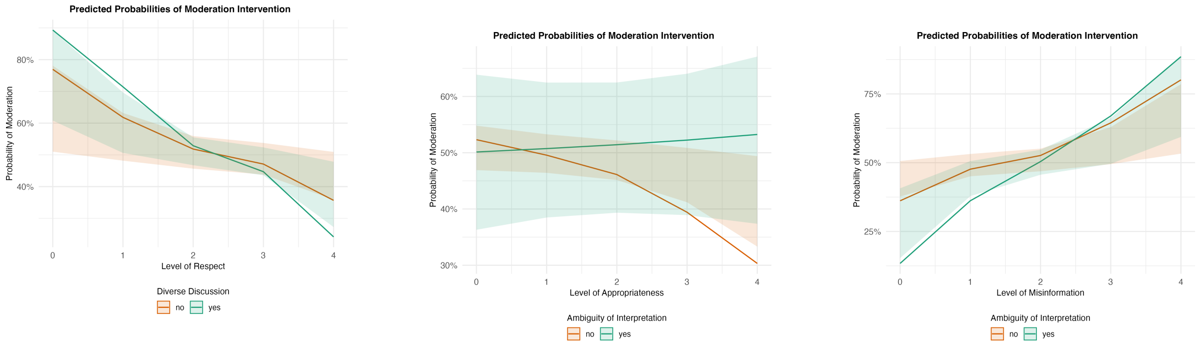
(c) Interaction between annotators who vote for AI support in moderation to help with complex tasks and comments which different levels of moderation-like behavior.

(d) Interaction between annotators who believe balancing freedom of speech and protecting users from harm is a bottleneck of human moderation and comments which different levels of appropriateness.

(e) Interaction between annotators believe that comments of high quality are essential for a good discussion experience and comments which different levels of moderation-like behavior.

(f) Interaction between annotators who believe transparency of regulations and moderation decisions are essential for a good discussion experience and comments which different levels of moderation-like behavior.

Figure 13: Marginalized fixed-effect predictions from different mixed-effects models, illustrating the interaction between different annotator-specific variables and comment properties (deliberative quality dimensions). Predictions were generated using the ggeffects package, averaging over random effects. Shaded areas represent 95% confidence intervals.



(a) Interaction between annotators believe that plurality of perspectives and a civility are essential for a good discussion experience and comments which different levels of appropriateness.

(b) Interaction between annotators who consider ambiguity in interpretation a bottleneck in human moderation and comments which different levels of appropriateness.

(c) Interaction between annotators who consider ambiguity in interpretation a bottleneck in human moderation and comments which different levels of misinformation.

Figure 14: Marginalized fixed-effect predictions from different mixed-effects models, illustrating the interaction between different annotator-specific variables and comment properties (deliberative quality dimensions). Predictions were generated using the ggeffects package, averaging over random effects. Shaded areas represent 95% confidence intervals.