Venire: A Machine Learning-Guided Panel Review System for Community Content Moderation

VINAY KOSHY, University of Illinois at Urbana-Champaign, USA
FREDERICK CHOI, University of Illinois at Urbana-Champaign, USA
YI-SHYUAN CHIANG, University of Illinois at Urbana-Champaign, USA
HARI SUNDARAM, University of Illinois at Urbana-Champaign, USA
ESHWAR CHANDRASEKHARAN, University of Illinois at Urbana-Champaign, USA
KARRIE KARAHALIOS, University of Illinois at Urbana-Champaign, USA

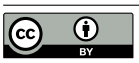
Research into community content moderation often assumes that moderation teams govern with a single, unified voice. However, recent work has found that moderators disagree with one another at modest, but concerning rates. The root problem is not the disagreements themselves. Subjectivity in moderation is unavoidable, and there are clear benefits to including diverse perspectives within a moderation team. Instead, the crux of the issue is that, due to resource constraints, moderation decisions end up being made by individual decision-makers. The result is decision-making that is inconsistent, which is frustrating for community members. To address this, we develop Venire, an ML-backed system for panel review on Reddit. Venire uses a machine learning model trained on log data to identify the cases where moderators are most likely to disagree. Venire fast-tracks these cases for multi-person review. Ideally, Venire allows moderators to surface and resolve disagreements that would have otherwise gone unnoticed. We conduct three studies through which we design and evaluate Venire: a set of formative interviews with moderators, technical evaluations on two datasets, and a think-aloud study in which moderators used Venire to make decisions on real moderation cases. Quantitatively, we demonstrate that Venire is able to improve decision consistency and surface latent disagreements. Qualitatively, we find that Venire helps moderators resolve difficult moderation cases more confidently. Venire represents a novel paradigm for human-AI content moderation, and shifts the conversation from replacing human decision-making to supporting it.

CCS Concepts: • **Human-centered computing → Collaborative and social computing systems and tools**; **Empirical studies in collaborative and social computing**; *Empirical studies in HCI*; *Collaborative interaction.*

Additional Key Words and Phrases: Content moderation, human-AI interaction, decision-making, online communities

ACM Reference Format:
Vinay Koshy, Frederick Choi, Yi-Shyuan Chiang, Hari Sundaram, Eshwar Chandrasekharan, and Karrie Karahalios. 2025. Venire: A Machine Learning-Guided Panel Review System for Community Content Moderation. *Proc. ACM Hum.-Comput. Interact.* 9, 7, Article CSCW518 (November 2025), 35 pages. https://doi.org/10.1145/3757699

Authors' Contact Information: Vinay Koshy, vkoshy2@illinois.edu, University of Illinois at Urbana-Champaign, Urbana, Illinois, USA; Frederick Choi, fc20@illinois.edu, University of Illinois at Urbana-Champaign, Urbana, Illinois, USA; Yi-Shyuan Chiang, ysc6@illinois.edu, University of Illinois at Urbana-Champaign, Urbana, Illinois, USA; Hari Sundaram, hs1@illinois.edu, University of Illinois at Urbana-Champaign, Urbana, Illinois, USA; Eshwar Chandrasekharan, hs1@illinois.edu, University of Illinois at Urbana-Champaign, Urbana, Illinois, USA; Karrie Karahalios, kkarahal@illinois.edu, University of Illinois at Urbana-Champaign, Urbana, Illinois, USA.



ACM 2573-0142/2025/11-ARTCSCW518
https://doi.org/10.1145/3757699

1 Introduction

In many online communities, a team of volunteer moderators is responsible for creating and enforcing rules to govern acceptable speech. Although rules are often collectively set by the entire moderation team, moderators act more independently when carrying out day-to-day enforcement. This implies a tension inherent to moderation efforts. While community policy reflects a moderation team's shared values, the day-to-day practice of this policy relies on individual decision makers. Prior work has found that moderators themselves disagree over how to apply community rules at concerning rates – audits of moderator decision-making have found disagreement rates as high as 16% [25] and 23% [32]. This raises the possibility that moderators on the same team could be enforcing rules inconsistently. Inconsistency can be frustrating for community members. If users see that their posts have been removed while similar posts remain unmoderated, they may feel unfairly singled out or targeted [21, 34]. Naively, one might try to increase decision consistency by requiring every moderation case be reviewed by a panel of moderators [37]. In practice, this is infeasible. Universal panel review creates far too much extra work for moderation teams, which are already stretched thin [5, 31]. In this paper, we explore an alternative approach. Instead of mandating universal panel review, we build Venire[1], a machine learning-guided panel review system.

The key idea behind Venire is that while some moderation cases would benefit from panel review, many are straightforward and can be handled by a single moderator. Venire acts as an overlay for Reddit's existing moderation queue, and gives moderators the ability to manually flag cases for panel review. To help triage cases for panel review, Venire is backed by an ML model which uses moderation log data to predict how each team member would respond to an incoming case. Venire recommends a case for panel review when it predicts the moderation team is likely to disagree over how to handle it. When a case undergoes panel review, its final outcome is determined by a vote amongst multiple moderators rather than by a single decision-maker. Ideally, Venire helps moderation teams surface latent disagreements, while keeping the increase in moderator workload to a minimum. We present a series of three studies through which we design, build, and evaluate Venire. First, we conducted preliminary interviews to assess whether our intended goals for Venire were aligned with moderator needs. Second, we performed technical evaluations to ensure that Venire could deliver on the promise of increasing decision consistency and surfacing latent disagreements. Finally, we investigated how moderators use the system in practice by performing a think-aloud study. In the think-aloud study, moderators used Venire to make decisions on real moderation cases pulled from the r/ChangeMyView subreddit.

Our initial interviews revealed that moderators were open to the idea of using a system like Venire. Moderators appreciated having a built-in deliberation channel as an alternative to informal deliberation processes already occurring within their communities. Moderators anticipated that the ML panel recommendations could help surface disagreements that were being missed. We also surfaced potential benefits we had not anticipated. Most critically, moderators felt Venire had value as an onboarding tool for new moderators, since the machine learning model could act as a safety net to catch controversial rulings from a new moderator.

In our technical evaluations, we conducted simulation-based analysis on two dataset to assess the quality of our ML model's panel recommendations. Using a large, publicly available toxicity dataset [26], we demonstrate that our predictive model can effectively triage moderation cases for panel review: we are able to approximate the decision-consistency benefits of universal panel review while assigning only a third of moderation cases to a panel. To make our analysis more

[1]The name is derived from "*venire facias*", an archaic legal term for a written order from a judge directing a sheriff to assemble a jury [43]

ecologically valid, we constructed a smaller dataset containing labels from Prolific crowdworkers. To improve ecological validity, crowdworkers were asked to enforce an actual moderation rule from r/ChangeMyView, a popular Reddit community. On this smaller dataset panel allocation was less efficient—it took assigning 60% of all cases to a panel to approximate the decision consistency of universal panel review. Still, our results with both datasets indicate that model-assigned panels significantly outperformed random panel assignment at improving decision-consistency and surfacing disagreements.

Finally, we tested Venire's capabilities as an onboarding tool in our think-aloud study with moderators. We had moderators from other communities enforce a rule from r/ChangeMyView, using real data from r/ChangeMyView's mod queue. Moderators found Venire's ML model recommendations helpful for deciding when to assign a case for panel review, and noted that the panel review empowered them to play a more active role in the moderation queue. Taken together, our findings demonstrate how an ML-guided panel review system can support the reflective practice of moderation work [8], and facilitate high quality decision making.

2 Background

The goal of this work is to develop a human-AI workflow for surfacing disagreements over subjective content moderation cases. As such, we present a review of prior work on: existing practices amongst community moderators for establishing a consistent moderation policy, prior attempts to build human-AI moderation tools, and machine learning approaches for modeling subjective decision-making.

2.1 Community Practices for Improving Consistency

Most commonly, moderators of online communities create sets of shared guidelines for rule enforcement to ensure the team acts in a consistent matter [5, 8, 10, 25, 40–42]. In interviews, both Seering et al. [41] and Cullen and Kairam [8] find that these guidelines are developed iteratively over a community's lifespan, often in response to specific incidents where moderators felt a user crossed the line. Policy tends to be set collectively by mod teams [8, 10, 40, 41], though moderators sometimes take community input, or defer to a "head" moderator [8, 41]. Occasionally, disputes over such policy can cause communities to fracture, splitting off into separate groups [10]. Moderators may consult with another when they are unsure how to handle a particular case [8, 41]. Cullen and Kairam [8] notes that such incidents can reveal places where moderators' mental models of how a rule should work, or even core values, are misaligned, creating an opportunity for reflection.

Still, not every community's moderation team revolves around a comprehensive, iteratively developed policy. For example, Seering et al. [41] found that at least one moderator was told to "do whatever you feel makes the [community] better" when they were made a moderator. Similarly, across studies, many interviewees report experiencing limited onboarding when they started as moderators, instead relying on more implicit processes to develop a feel for the community norms [8, 41, 42].

To our knowledge, only two studies have tried to directly assess how often moderators disagree with one another [25, 32]. Both studies had moderators review sets of comments previously posted on large, discussion-based subreddits [25, 32]. For comments that received two in-study annotations, Lucas et al. [32] found a disagreement rate of 23% (Fleiss' $\kappa = 0.46$, $N = 222$), while Koshy et al. [25] found a disagreement rate of 13% ($N = 134$). When comparing in-study labels to real life outcomes, the rates of disagreement were 28% ($N = 246$ annotations across 134 comments) and 26% ($N = 1020$ annotations across 798 comments) respectively. The present work is primarily motivated by the prevalence of disagreements found in these two studies, in spite of the existing measures that online communities take to ensure consistency.

2.2 Human-AI Content Moderation Tools

Researchers in the CHI and CSCW communities have built a number of human-in-the-loop AI systems for content moderation [19]. These tools can be categorized as facilitating either *top-down* [5, 7, 16–18, 27, 39] or *personalized* content moderation [20, 23]. Top-down content moderation tools are those which are used to filter content for an entire platform or community, whereas personalized content moderation tools filter content at the level of a single user's feed. Although tools within each category tend to follow similar design patterns, they differ substantially in terms of the task the underlying machine learning model is trained on, the way model predictions are presented in an interface, and the place in the moderation process human decision-making is utilized. Because Venire falls under the umbrella of top-down content moderation tools, we provide a more detailed review of prior top-down approaches.

2.2.1 AI tools for top-down moderation. Typically in *top-down* tools, a machine learning model is used to flag content for a human moderator to review—moderator decisions in turn affect all users within a given community or platform [5, 7, 14, 16, 18, 27, 39]. Amongst top-down tools, one of the key differentiating factors is the source of the training data. One approach is to use a generalized model. For example, in building a tool for Discord moderation, Choi et al. [7] use Perspective API.[2] The Perspective model is trained using crowdworker toxicity labels on comments across multiple social media platforms. In contrast, to build a tool for Reddit moderation, Chandrasekharan et al. [5] use historical data scraped from Reddit to train an ensemble of models that predict whether comments in specific communities will get removed or not. Halfaker and Geiger [16] adopt the most bespoke approach, creating a "Wiki Labels" system through which Wikipedians can contribute training data labels to specific model development requests.

Notably, almost all existing tools have been built with the goal of either reducing moderator labor, or helping moderators identify norm violations they would have otherwise missed [5, 7, 16, 18]. Our approach, using machine learning models to improve the consistency of human decision-making, is relatively unique in this regard.

Still, a few other researchers have also built tools that center on disagreement in the content moderation process [14, 27]. Gordon et al. [14] argue that rather than predicting a majority vote or average label across annotators, machine learning models should be trained to predict a label for each annotator in the dataset, using the annotators ID and demographic information as features, a process they call "jury learning." They create an interface for jury learning models that allows the end user to choose which voices in the training dataset to amplify for their purposes. Kuo et al. [27], whose work is perhaps most similar to our own, develop a tool that allows communities to curate evaluation datasets for AI tools they might want to adopt. As community members annotate data points, cases with disagreements are prioritized to receive additional ratings. Although our work shares a similar focus on contentious moderation cases, the goal of our work is to use a model to allocate panels efficiently, rather than to use panels to more accurately evaluate a model.

2.3 Modeling Subjective Decision-Making

Gordon et al.'s jury learning framework [14] is part of a broader trend amongst HCI and machine learning researchers, recognizing that modeling individual annotator beliefs can be beneficial for subjective tasks [4, 11, 12, 38]. This viewpoint, sometimes referred to as "perspectivism," is accompanied by varying motivations. Drawing on feminist theory, Blackwell et al. [2] argue that traditional majority vote aggregations of annotator labels can reinforce the viewpoints of dominant social groups. Other researchers appeal to more technical benefits of perspectivist approaches: that

[2]https://www.perspectiveapi.com

they more accurately represent the data generating process [4, 12], that they provide the ability to capture uncertainty in training labels that arise from human variation [4, 12, 14], and that they afford end users more modeling flexibility [4, 14]. We provide a brief, non-exhaustive, review of prior perspectivist modeling approaches and applications.

2.3.1 Direct Disagreement Prediction. One approach adopted in prior work is to directly model the variance in annotator labels as a function of features of the training instance (i.e., without using annotator features) [15, 36]. Gurari and Grauman [15] apply this approach to visual question answering tasks. They treat disagreement prediction as a binary task, using image- and question-based features to predict whether a panel of 10 raters will reach a supermajority (9/10 or 10/10) decision for a specific image-question pair. Raghu et al. [36] provide a theoretical analysis of direct disagreement prediction, arguing that disagreement prediction can be thought of as a regression task, mapping inputs to "uncertainty scores," such as the variance of rater labels or the probability of two raters agreeing. They contrast direct disagreement prediction (referred to as direct uncertainty prediction in their work) with what they call uncertainty via classification: a two step-process in which an uncertainty score is computed from the output of a calibrated classifier. They prove direct disagreement prediction is more accurate than uncertainty via classification.

2.3.2 Annotator-Aware Approaches. In contrast to direct disagreement prediction, annotator aware approaches utilize annotator-level features when making predictions [11, 14, 26, 35, 44]. These features almost always include an annotator ID, for which corresponding embeddings are learned [11, 14, 35, 44]. Prior work differs on exactly how embeddings are learned, and where embeddings are incorporated in a neural architecture. Demographic information for each rater is also sometimes utilized [11, 14], though ablation analysis from Gordon et al. [14] found limited additional value to using these features in a toxicity detection task.

Practically speaking, an important distinction between direct disagreement prediction and annotator-aware approaches is the type of training data needed. Direct disagreement prediction requires multiple labels per training instance, but does not require any annotator identifiers features. In contrast, annotator aware approaches generally require annotator identifiers, but do not require multiple labels per training instance.

2.4 Towards Venire

The existing literature demonstrates that even well-intentioned moderation teams may suffer from undetected consistency issues [25, 32], and that such consistency issues negatively affect user experience [21, 34]. Further, training a machine learning model to predict disagreements appears to be technically feasible [14, 15], making it possible to build the predictive model underlying Venire. However, we believe there are a few major open questions that need to be answered before building and testing Venire. First, how do moderators view the labor-consistency tradeoff inherent to panel review? Would they deem it worthwhile to increase the amount of decisions that need to be made in order to catch potential disagreements? And how could a panel review system complement existing moderation practices? Second, even if disagreement prediction is possible in some cases, is it possible to implement *for a realistic content moderation task* and *with the kinds of data available in a subreddit's moderation log*? These questions motivated our decision to conduct two preliminary studies leading up to building and evaluating the Venire interface.

3 Preliminary Interviews: Does Venire Support Moderator Needs?

We conducted a round of preliminary interviews with moderators to better understand how they would view the labor demands of panel review. More broadly, we wanted to surface moderators' general attitudes towards the idea of an ML-assisted panel review system, and better understand

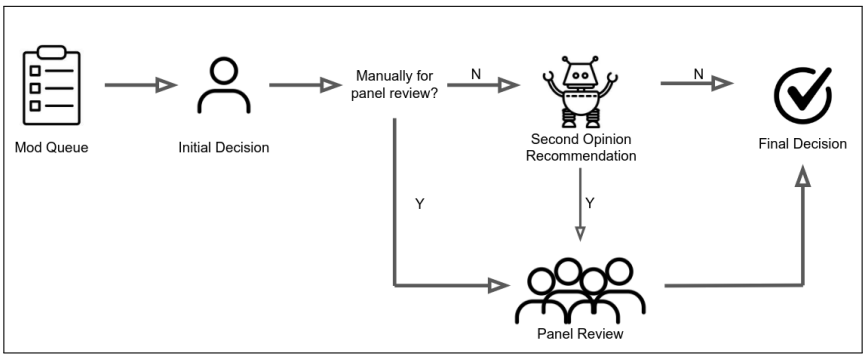


Fig. 1. Venire workflow. A case is pulled from the moderation queue by a human moderator, and an AI model recommends whether it should be reviewed by a single moderator or a panel of moderators.

moderators' existing processes for improving decision consistency. Simply put, our goal was to evaluate whether our vision for Venire aligned with moderator needs. We summarize these interview objectives in the following research questions:

RQ1a: How important do moderators think decision consistency is? What factors do they weigh it against?
RQ1b: What processes do moderators currently employ to improve decision consistency?
RQ1c: What benefits and harms do moderators anticipate an ML-guided panel review system will bring?

These interviews were narrowly scoped around evaluating how moderators felt about using an ML-guided panel review system. Although moderator input was sought around specific design elements, and moderators were encouraged to share alternate ideas for the system, this was not a fully fledged participatory design study. As such, we cannot claim that Venire represents a solution to the most pressing problems moderators face, or that Venire is the moderation tool that would see the widest adoption. Still, we tried to design the interview protocol to surface both positive and negative feedback equally. We encourage future work that engages moderators more fully in the design process, especially towards building tools to improve decision consistency.

3.1 Initial Prototypes

To ground our interviews, we created a workflow diagram to communicate the idea behind our system (Figure 1). We also created two interactive interface mock-ups in Figma (Figure 2). These mockups were shown to participants during the interviews to give them a concrete sense of how the system would work in practice.

The mock-ups were conceived through iterative brainstorming sessions amongst research team members, one of whom has experience as Reddit moderator. To minimize disruptions to existing workflows, we based the mockups on the existing Reddit moderation interface. We wanted to present multiple mockups to help participants separate the high-level idea behind Venire from the particulars of any specific interface. However, we limited the number of mockups to two, since design presentations were only part of the planned interviews. The two designs were chosen to represent more and less intrusive panel voting systems, since the research team anticipated in advance that panel review could be overly burdensome.

3.1.1 System Mock-up #1: Strict Voting. Figure 2 contains the first version of the interface. In this version, moderators make two decisions for each comment: whether they support removal for the

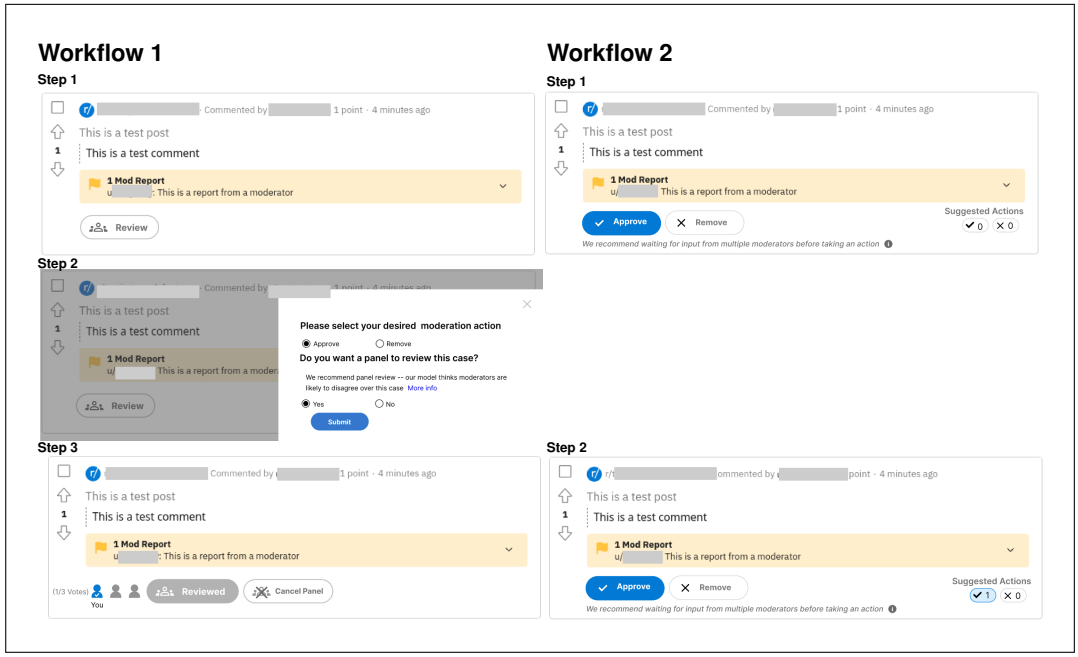


Fig. 2. Two potential Venire workflows. **Left: The strict voting workflow.** When making a ruling on a case, moderators must specify whether they want to flag it for panel review or not. Cases flagged for panel review remain in the moderation queue until a majority vote is achieved across k moderators. **Right: The suggested action workflow.** Rather than enforcing a strict voting procedure, moderators are always given the option to "suggest" an action instead of making a decision, making their opinion visible to other moderators. Any moderator can input a final decision when they feel confident enough.

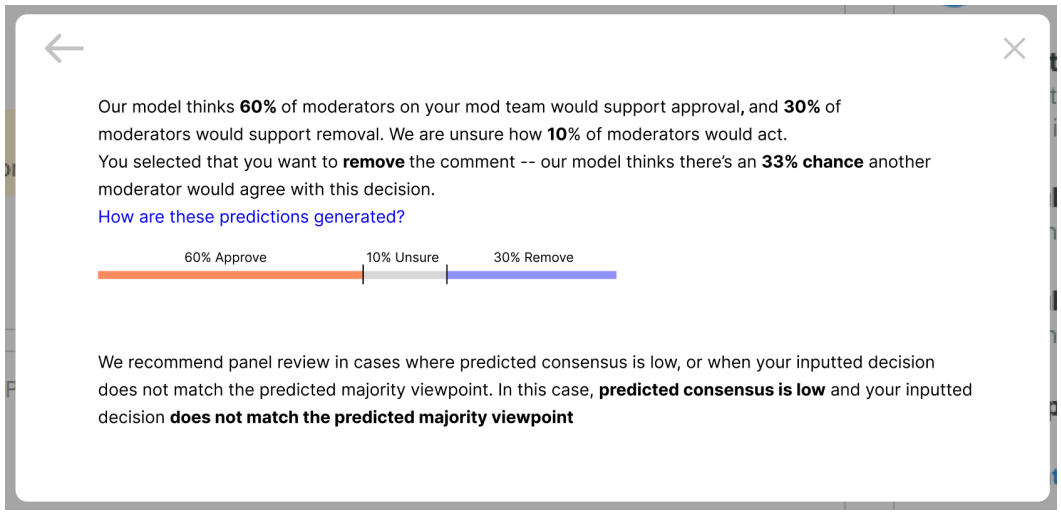


Fig. 3. Visualization of how the model predicts the moderation team will react to a particular case—accessed by clicking the "More info" button next to the panel recommendation text in both mock-ups.

comment, and whether it should undergo panel review. Moderators also see a model-produced recommendation for whether a case needs panel review. Clicking the "More info" button allows the user to see a full breakdown of how the model predicts the moderation team will react to the comment (Figure 3). If the moderator chooses to flag a case for panel review, it will remain in the queue, until k moderators cast a vote, where k is a configurable number. After the kth vote, the final outcome will be determined based on a majority vote. Note that any moderator on the subreddit's mod team is allowed to weigh in on a panel decision—panel votes are not solicited from specific moderators. To minimize bias, moderators are unable to see the direction of existing votes until after they vote themselves.

3.1.2 System Mock-up #2: Suggested Actions. The second interface treats panel review more loosely. All moderators are given the ability to signal their preference for removal or approval using two "reaction" buttons. There is no built-in aggregation mechanism for these suggestions. Instead, at any time, a moderator can choose to make the final removal/approval decision, taking into account others' suggested actions as they see fit. When cases are predicted to be controversial, moderators are advised to make a suggestion rather than an immediate decision. This interface focuses on making the opinions of different moderators visible to one another, while minimally disrupting the moderation workflow.

3.2 Interview Protocol

Interviews were conducted over Zoom, lasted around 60 minutes, and contained three parts. First moderators described the rules of the subreddit they moderate. Then we asked moderators to recall any prior experiences of disagreements between moderators, steps taken to ensure decision consistency (RQ1b), and general attitudes towards decision consistency as a goal for moderation (RQ1a). Next, we presented the workflow diagram to moderators to give them a high-level idea of how the system would work. Moderators were asked to speculate about the strengths and weaknesses of the system (RQ1c). If moderators did not explicitly mention concerns about increased workload here, we prompted them about it.

Finally, participants were given a guided tour of the two interactive mockups. Afterwards, we asked moderators to contrast the two interfaces, and state which one they preferred. Moderators were encouraged to share alternative workflow suggestions here. We also solicited lower-level design feedback, and asked moderators to re-evaluate the potential benefits and harms of the system, having seen concrete designs. Moderators were compensated the equivalent of $30 USD in their local currency via Amazon giftcard or PayPal.

3.3 Recruiting

Recruitment messages were sent to Reddit moderation teams via the platform's modmail feature. Messages were sent to subreddits with over 10,000 subscribers and at least one subjective rule listed in the subreddit's sidebar. The lead author manually curated a list of subreddits based on these criteria. The r/ChangeMyView subreddit was also contacted since the researchers had previously worked with this community to create a dataset of moderation decisions that could later be used to train Venire's machine learning model. Several measures were taken to avoid spamming moderators with recruitment messages, which we detail in the Appendix A. Table 1 contains a summary of the final eight participants interviewed.

3.4 Results

3.4.1 Attitudes Towards Decision Consistency (RQ1a). Unsurprisingly, almost all the moderators we interviewed viewed decision consistency as desirable (N=6). P3, speaking directly to the motivations

PID	Subreddit Size (Subscribers)	Subreddit Topic
P1	1-10M	r/ChangeMyView
P2	10-100K	Fanfiction
P3	100K-1M	Gaming/TTRPG
P4	10-100K	Music
P5	10-100K	Pregnancy Support
P6	100K-1M	Gardening/Agriculture
P7	100K-1M	History/Politics
P8	10-100K	TV Show

Table 1. Interview Participant Information

of the project, stated "in an ideal world [...] every post would go through panel review." This was especially the case for subreddits that dealt with highly sensitive or political topics (N=3). P1, who moderates r/ChangeMyView, argued that decision consistency was essential to user participation in their community:

> "Ultimately the subreddit that we have doesn't work unless there is consistent moderation that is as topic neutral as possible. [...] Pick any culture wars topic in the United States or any hot button political issue—keeping conversations civil on those topics requires a lot of consistency. Everyone needs to know what the rules are and how they're applied. And they need to feel it's fair and consistent regardless of if they're a Republican, a Communist, a Social Democrat, a Centrist, or a Libertarian. They need to feel like they're getting a fair shake or they don't participate."

Still, most moderators contextualized the importance of decision consistency alongside other factors (N=5). Minimizing workload and stress came up most frequently (N=4). With regards to stress, P5 said that moderation could sometimes feel like "factory work", and that it was important to "get through the queue" with the "least amount of damage to yourself." Others highlighted the voluntary nature of moderation work when thinking about the consistency-workload tradeoff. P2 argued that "whoever's the person shouldering most of the workload gets to make the calls" in part because those moderators would be "more in tune with how the subreddit currently is." Similarly, P4 said "we really try to back up our moderators as much possible in their decisions, even if it's something we might disagree with."

Outside of workload concerns, P3 mentioned that incorporating diverse perspectives into the moderation team could be worth sacrificing some decision consistency for. P2 noted that more senior moderators were sometimes able to take actions within the community that other moderators would not be able to. This is a form of decision inconsistency that was viewed positively. They described these moderators as "having a bit more goodwill" amongst community members, which allowed them to "shut something down" where another moderator could not.

3.4.2 Existing Practices (RQ1b). Every moderator we interviewed described taking steps in the past to either preempt a potential disagreement (N=8) or resolve a disagreement that surfaced after a moderation action had already been taken (N=5). Most moderators recalled soliciting a second opinion from another moderator through a side channel like Discord (N=5). This practice was sometimes specifically encouraged for new moderators (N=2). Moderators also described holding discussions prior to implementing a new rule to try to iron out potential inconsistencies (N=3). In a few cases, moderators outlined specific rules or policies where taking a vote amongst multiple mods

was required (N=2). In P1's community, certain kinds of post removals "required two moderators to sign off on."

Still, most moderators felt disagreements were relatively rare to begin with (N=5). At first glance, this was surprising given the disagreement rates found in prior work. However, the moderators we interviewed attributed this rarity to the fact that their subreddits only had a few moderators (N=2) or to the fact that their community's rules were straightforward (N=2). In contrast, the subreddits studied in prior work had large moderation teams, and tight moderation standards. Still, P1 mentioned that disagreements could be going unnoticed in their community, stating: "if a moderator does make a decision, very rarely are the other moderators going to even be aware of it."

3.4.3 Potential Benefits and Harms of ML-Guided Panel Review (RQ1c). Moderators outlined a number of potential benefits to a panel review system. Moderators appreciated having a built-in channel for disagreement-resolution that might have otherwise happened in a side channel (N=5). And as we anticipated, a number of moderators felt that a machine learning model could help surface disagreements that would have gone unnoticed (N=6). P1, for example, said "I would say this tool would be great for helping to figure out [...] if there are controversial cases that are being decided too quickly." A few moderators explicitly stated that this could lead to policy updates (N=3). P7 was one such moderator:

> "You might think that you're in the majority with one opinion or someone else might think they're in the majority with another opinion and they're not. So it would be nice to be able to see, 'oh, this is how the rest of the mod team has been moderating things. I see that I've been moderating differently.' Maybe we should talk about this rule and it will provide discussion and at least get all the mods on the same page or maybe a compromise to the rule is made."

We also surfaced a few unanticipated benefits. Most notably, moderators highlighted the benefits of ML-guided panel review as an onboarding tool (N=4). In the context of recruitment, P2 argued that this would be beneficial to both senior and junior moderators. For the senior moderators, P2 felt that they "won't feel like they have to be double checking or [...] correcting them all the time." and for the junior moderators being onboarded "they can be a little bit more confident that if they take an action and it turns out to be a mistake, that it will be caught. Like a safety net." P5 echoed this sentiment saying "I do think that you'll have people more willing to moderate in general. Right now when you recruit moderators, the stickiness of a moderator is not high. [...] It could be that when they run across these difficult to moderate content, they don't know what to do." P3 contrasted using the panel review system to their current practices for onboarding new moderators, saying:

> "When there's a new moderator on the team, I'll try and do audits of some of their actions. But it's a bit of a slog and there aren't really great tools to do that. Whereas this serves as a potential to be doing those audits in a way. And so that just makes it easier for all of us"

However, moderators mentioned a number of drawbacks as well. Workload concerns came up multiple times (N=5), especially with respect to false positive flags from the ML model (N=3). At the same time, moderators felt like the workload could be manageable if disagreements were relatively rare (N=4). One participant, P5, questioned the fundamental value of highly deliberative moderation at all, saying:

> "I don't always 100% agree with everything the other mod does but I agree enough, and that's enough for me. And that's because moderating a community is a lot of work. I have a full-time job, I have a kid. There's other things happening in my life. And in

> hindsight over the six years, I think the community is better for me not strictly creating my own vision [..] Building consensus is a difficult thing to do, and it's not always worth the bang for the buck. It's not always worth it for you internally as a person, and it also might not be worth it for the community in general."

Moderators worried panel review could increase the decision time for reported content (N=3). P6, for example, said "if it takes three days to get an answer [...] the post is gone [...] it doesn't really even matter anymore." However, P2 offered a potential solution saying the system could instead be "more of an appeals process that would let you reverse a decision [...] rather than something that might block you from taking action."

A few moderators speculated that the panel review feature might simply go unused (N=4). P3 and P6 both argued that moderators may not have the self-awareness to flag a case for panel review. P6 stated "it takes a certain level of person to say 'I think it's possible that I didn't make the best decision here.' " Other moderators simply felt that moderators within the same mod team might all agree anyways (N=2). To mitigate this, P6 proposed that instead the system could get "an outside panel of moderators" instead, possibly from a "sister community."

3.4.4 Mockup Preference. Moderators largely preferred the first strict voting mock-up to the suggested action mockup (N=5 vs N=1). Moderators who preferred the strict voting mock-up liked the rigidity of the voting process. These moderators described the strict voting mock-up as "more formal" and the suggested action mock-up as "more passive" and "softer." P2, the sole moderator who preferred the suggested action mockup argued that they liked the "non-binding" nature of the suggested action button. They felt the voting system "puts a bit more pressure on you" because "if you're the second person voting, you might or might not be making a moderatorion action, you don't know."

3.4.5 Summary of Findings. Overall, we find that Venire's intended goals, surfacing disagreements and improving decision consistency, are aligned with moderators' values. Still, moderators reported a number of additional factors, like stress management and decision speed that these goals should be weighed against. Although moderators reported employing practices to improve decision consistency already, they saw potential for Venire to further these efforts. Thus, our findings gave us the confidence to proceed with building Venire.

4 Technical Evaluation: Can Venire Improve Decision Consistency?

In this section, we present two experiments that test the technical feasibility of predicting moderation disagreements. [3] Our primary goal is to demonstrate that ML-guided panel review can deliver on the promise of improving decision consistency and surfacing disagreements. In our first experiment, we demonstrate this using a large, publicly available toxicity dataset. In our second experiment, we construct a new, more ecologically valid dataset using crowdworkers on Prolific, and repeat our analysis. Crucially, the training dataset used in the second experiment is much smaller, and based on a more realistic content moderation task. We find that our model performs worse on this new dataset, but still allocates panels more effectively than a random panel assignment baseline. More concretely, the goal of our technical evaluations is to answer the following research questions:

> **RQ2a**: To what extent can an ML-guided panel review improve the consistency of moderation decisions?

[3] Code for model training and test set analysis can be found here: https://drive.google.com/drive/folders/1DSJlOihn4w-WmidHybtLiSWePQIn-9Wu?usp=sharing. Datasets used in this section are available by request. See the README in the Google Drive folder for details.

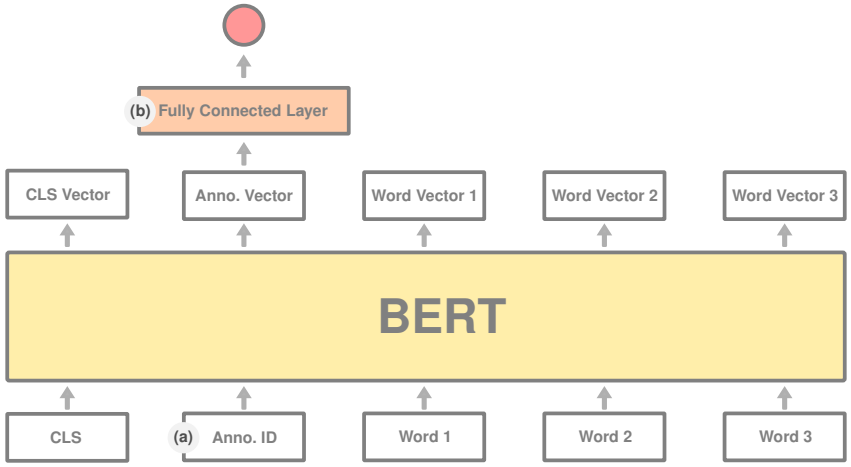


Fig. 4. Overview of our model architecture. We prepend a unique token representing a specific annotator (a) to each sentence, and pass the finetuned BERT contextual embedding for this token through a fully connected layer (b) to produce our final prediction. We ignore the special CLS token that is typically used for BERT sequence classification tasks [9].

RQ2b: To what extent can an ML-guided panel review surface disagreements between human raters?
RQ2c: How does the performance of an ML-guided panel review system change when we move from an ideal dataset to one that more closely matches moderation log data?

4.1 Modeling Approach

Our modeling approach is inspired by recent perspectivist NLP papers [11, 14, 35, 44]. We define the prediction task as follows. As input, the model is shown a text x_i and the identity of a rater j. Our model then outputs a prediction for how rater j would label that x_i. Prior work varies on the specific neural architecture used. In preliminary experiments, we found adapting Yin et al.'s approach to be most effective [44]. We treat each rater ID, j, as a special token that is prepended to x_i. We feed this new string into a BERT model, and pass the contextual embedding for the rater ID token through a feedforward layer to produce the final prediction. Figure 4 demonstrates our proposed architecture. Like other perspectivist approaches [11, 14] we learn rater-specific embeddings to capture attributes of decision-making style. This embedding is passed through multiple transformer layers alongside the text, to produce a final contextual embedding which captures attributes of the text and the rater.

During deployment, our model predicts how *each* possible rater would label x_i. These predictions can be aggregated to produce: 1) a "majority vote" prediction, $M(x_i)$, of the fraction of raters that would label x_i as positive, and 2) an "a priori disagreement score" $D(x_i)$, that captures how controversial x_i is. To produce $D(x_i)$, we calculate the empirical probability that two randomly selected rater predictions disagree with one another [36]. An alternative approach might be to simply train a model that learns $D(x_i)$ directly. In the context of our study, this alternative approach is infeasible, since it requires us to have multiple rater labels for each comment in the training dataset; real moderation log data is largely singly-labeled.

4.1.1 Evaluation strategy. Intuitively, there are multiple ways to use $M(x_i)$ and $D(x_i)$ to produce a panel review recommendation. We focus on two main approaches. First, a panel could be recommended when a moderator tries to take an action that disagrees with $M(x_i)$. We call this an "a posteriori" disagreement prediction, since it occurs after a moderator as taken an initial action. Second, a panel could be recommended when $D(x_i)$ is large (i.e. when there is a high chance of a disagreement a priori). To evaluate model performance, we first present metrics to gauge the quality of $M(x_i)$ and $D(x_i)$ independently. Then, we conduct a simulation analysis to evaluate the performance of different panel allocation strategies relying on $M(x_i)$ and $D(x_i)$.

4.1.2 Hyperparameters. We use a base BERT model for the toxicity prediction task and a BERT-large model for the Prolific study. Because the comments for the toxicity prediction task are short, we use only the first 126 tokens of the text for prediction. For the Prolific study, which contained longer comments, we use a 256-token window. When comments exceeded 256 tokens, we create an embedding for each 256-token slice of the comment and combine them via max-pooling to produce a final token embedding. For the Prolific task, where we have access to additional thread context, we also feed the model the first 256 tokens of text either the immediate parent comment for reply comments or first 256 tokens of associated post for top-level comments. We separate the target comment's text from the thread context using BERT's SEP token. Our model is implemented in PyTorch and trained with the AdamW optimizer—we use a learning rate of 2e-5, batch size of 32, and weight decay of 0.0075. We train for 3 epochs on the toxicity dataset and 5 epochs on the Prolific dataset. The hyperparameters were determined after conducting a modest grid search using a validation set. All model training was conducted on a single Nvidia T4 GPU provided by Google Colab.

4.2 Experiment 1: Toxicity Dataset

4.2.1 Dataset Description. For our initial evaluations, we leverage a large, multiply-labeled toxicity dataset provided by Kumar et al. [26]. This dataset contains 107620 comments sampled from Twitter, Reddit and 4chan. Each comment received five, 5-point Likert ratings of toxicity. A total of 17280 raters contributed to the dataset, and each rater labeled at least 20 comments. We treat toxicity prediction as a binary classification problem, with Likert ratings of 3 ("moderately toxic") or higher being treated as "Toxic", and 2 ("slightly toxic") or lower as "Not toxic" [26]. The labels in this dataset are mildly imbalanced—only 29% of supplied ratings were "Toxic", and the majority vote was "Toxic" for only 21% of comments. We use 75% of the comments in the dataset for training, 10% for validation, and reserve 15% for reporting results.

Model	AUROC	Accuracy	Precision	Recall
Toxicity (Rater-level Annotations)	0.8798	0.8191	0.7102	0.6445
Toxicity (Majority Vote)	0.9196	0.8688	0.7045	0.6780
A Priori Disagreement (Rater-Aware)	0.7376	0.7338	0.6373	0.4480
A Priori Disagreement (Rater-Blind)	0.6946	0.7004	0.5453	0.4261

Table 2. Model prediction quality for the toxicity dataset

4.2.2 Toxicity Predictions. The first two rows of Table 2 demonstrate our model's ability to predict toxicity labels. Our model achieves an accuracy of 82% (AUROC 0.88) when predicting individual annotator ratings (using a threshold of 0.5), and 86% (AUROC 0.92) when predicting the majority vote amongst all five annotators ($M(x_i)$). We produce majority vote predictions by aggregating the binary predictions for each of the five annotators who labeled a comment.

4.2.3 A Priori Disagreement Predictions. To test our model's ability to predict disagreements amongst raters a priori ($D(x_i)$), we divide comments in the test set into two groups—high consensus comments (decided unanimously or with a single dissenting rater) and low consensus comments (decided by a 3/2 split amongst the raters). Under this definition, 30% of comments were considered low consensus. The a priori disagreement prediction task is to discriminate between these two classes. We produce two kinds of predictions for comment consensus-level: "annotator blind" predictions and "annotator aware" predictions. Annotator-blind predictions are produced without looking at the identities of the five raters who were assigned to the comment. Instead, we make a prediction about how 100 randomly sampled annotators will label a comment, and aggregate these 100 predictions into a single disagreement score (as described in Section 4.1). This disagreement score is thresholded to produce a final binary consensus-level prediction. This threshold is chosen by calculating the disagreement score at which a low consensus outcome is more likely than a high consensus outcome, assuming correct annotator-level predictions. To produce annotator-aware predictions, we perform the same procedure, but use the 5 raters actually assigned to the comment. The second two rows of Table 2 contain the results—annotator-blind predictions are 70% accurate (0.69 AUROC), while annotator-aware predictions are 73% accurate (0.74 AUROC).

We can see that a priori disagreement prediction is more difficult than toxicity prediction. This could be because of the thresholding applied to separate "high" and "low" disagreement instances. When performing rater-aware disagreement prediction across five raters, a single rater-level error can lead to an incorrect disagreement prediction. The gap between the rater-aware and rater-blind performance indicates that the model is able to pick up on some characteristics of individual rater style. However, given the limited number of training points per rater (most commonly 20), there may be room to improve the quality of rater-level predictions.

4.2.4 Simulation Analysis. While the previous results demonstrate our model's raw predictive power, they do not tell us how well our model will perform *when used to allocate human decision-makers*. To address this, we present the results of a simulation based analysis using the test set data. First, for each comment x_i, we simulate single moderator review by randomly selecting a label, $h_1(x_i)$ from one of the five assigned raters. After simulating initial decisions, we then apply a *panel allocation strategy* to the comments. Formally, a panel allocation strategy is a function that takes in as input: x_i, $h_1(x_i)$, and a list of rater-level predictions $f_j(x)$. It then outputs a panel priority score p_i. We assign the cases with the top k% highest p_i to panel review. For these cases we solicit a second, randomly selected opinion $h_2(x_i)$. If $h_2(x_i)$ does not match $h_1(x_i)$ we solicit a third vote, $h_3(x_i)$ to tie-break.

We judge the performance of a panel allocation strategy by looking at 1) average number of raters used per case (workload), 2) the fraction of cases where the final decision matched the ground truth majority vote amongst all five raters (decision consistency), and 3) the fraction of cases where at least one dissenting viewpoint was sampled (disagreements surfaced). It should be noted that 2) and 3) represent distinct goals for the system. A panel allocation strategy that deliberately oversamples minority opinions, for example, might surface a large number of disagreements, but fail to improve decision-consistency. In a real world deployment, surfacing disagreements and improving consistency are more closely related, since moderators who observe disagreements can update community policy to minimize future disagreements. We cannot test this with our simulation analysis, however, so we report disagreements surfaced and decision-consistency as separate measures. We test the following strategies for panel assignment:

Random Panel Assignment: The panel priority score is a random number between 0 and 1.
Predicted Majority-Based Assignment: The panel priority score is equal to $|h_{initial}(x_i) - M(x_i)|$. Intuitively, cases are prioritized for panel review when the initial human rater decision

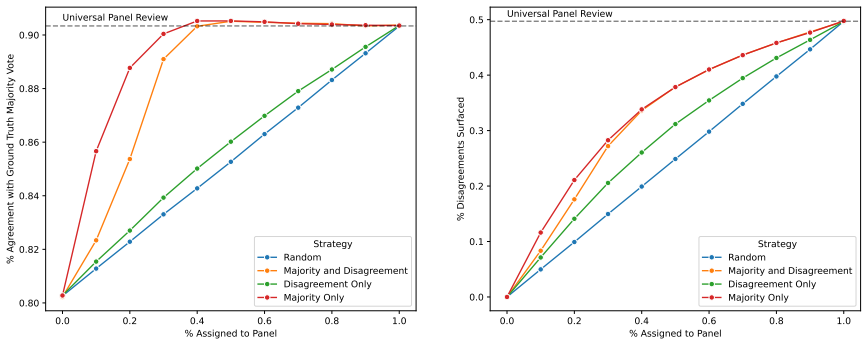


Fig. 5. Performance of a panel prioritization strategies at increasing workloads. When it comes to improving decision consistency (RQ2a), majority-vote based prioritization converges to the optimal value much faster than random allocation. Majority-vote based prioritization also outperforms other strategies at surfacing disagreements(RQ2b), though to a less dramatic degree

disagrees with the predicted consensus viewpoint. We use a random sample of 100 training dataset raters to produce $M(x_i)$.

A Priori Disagreement Score-Based Assignment: The panel priority score is equal to $D(x_i)$, the probability that the predicted decisions for two out of 100 randomly sampled training dataset raters disagree.

A Priori Disagreement+Majority Based Assignment: The panel priority score is equal to the $2D(x_i) + M(x_i)$. We multiply $D(x_i)$ by two since $0 \leq D(x_i) \leq 0.5$, while $0 \leq M(x_i) \leq 1$

Figure 5 contains the results of applying the strategies to increasing proportions of the test set. In each plot, the dotted line indicates the performance of assigning every case to a three person panel. Intuitively, effective strategies will quickly converge to the dotted line—these strategies approximate the benefits of universal panel review with fewer extra ratings. Overall we can see that Predicted Majority-Based Assignment performs the best. This strategy can approximate the decision consistency benefits of full panel review while assigning only 30% of cases to panel. This finding matches our initial intuition: many cases are straightforward, and do not benefit from undergoing panel review. All strategies outperform Random Panel Assignment for both surfacing disagreements and improving decision consistency. Surprisingly, the A Priori Disagreement Score-Based Assignment does not seem to improve decision consistency, contradicting suggestions from prior work [36]. This makes more sense on closer inspection—when soliciting additional opinions on a highly controversial case, we are almost equally likely to sample the minority and majority viewpoint. Thus, starting a panel only marginally improves the chance that we arrive at the ground truth majority viewpoint. Still, A Priori Disagreement-Based panel allocations outperform random assignment when it comes to surfacing disagreements.

4.3 Experiment 2: Prolific Dataset

Our initial analysis provides evidence that an ML-guided panel review system can improve decision consistency (RQ2a) and surface disagreements (RQ2b). However, the toxicity dataset we used differed from real moderation log data along a couple key dimensions, prompting us to experiment further (RQ3c). Perhaps most importantly, the dataset contained labels from multiple raters for each comment in the training dataset. Data scraped from a subreddit's moderation log will largely contain

only a single decision per comment, since moderators do not usually re-evaluate old moderation cases. Intuitively, this could make the disagreement prediction task harder, since the model will be unable to learn explicit contrasts between decision makers during training. Additionally, the raw number of comments in the toxicity training dataset was high—only the largest communities on Reddit would have such large-scale log data. Finally, the toxicity labeling task is too broad to serve as an effective proxy for a content moderation task. The survey instrument used to produce the dataset did not clearly define toxicity [26]. This may have inflated the number of disagreements in the dataset.

To address these issues, we curated a smaller dataset using crowdworkers on Prolific. To make the moderation task more realistic, we chose to have crowdworkers apply a real community rule to comments taken from a subreddit's actual moderation log. Specifically, we chose the r/ChangeMyView subreddit's "Rule 2", which bans comments that express rudeness or hostility towards another user. This rule was chosen for a few reasons. r/ChangeMyView maintains detailed guidelines on how Rule 2 should be applied. This makes it easier for us to communicate the rule to crowdworkers in an unambiguous manner, while maintaining ecological validity. At the same time, Rule 2 contains a degree of subjectivity. Thus, it is reasonable to assume that there are disagreements that can be surfaced in the first place. Finally, Rule 2 requires relatively little thread context to moderate. This allows us to keep the cognitive load of the task low, since users will not have to read long threads of discussion to make a determination.

In our Prolific study, crowdworkers contributed model training and test set labels across four separate tasks. Full details about the survey instruments are provided in the Appendix B, but we provide an overview here. In the first task, participants were given a short introduction to the r/ChangeMyView subreddit and shown a set of condensed guidelines for how to apply Rule 2. Participants then completed three practice questions where they were asked to apply Rule 2 to a comment. In these practice questions, participants were shown the actual decision made by moderators alongside a short explanation. The practice questions were deliberately chosen to contain short, straightforward cases that illustrated key points in the Rule 2 guidelines. Participants were then asked to provide training set labels for 20 comments and test set labels for 20 comments. Participants were compensated $7.50 for completing the first task (20 minute average completion time). The subsequent tasks each began with a refresher on the Rule 2 guidelines. In tasks 2 and 3, participants provided 20 training set labels and 20 test set labels. In task 4, participants provided 40 training set labels. Participants were compensated $5 for each of these parts (average completion times of 18 minutes for parts 2 and 3, and 20 minutes for part 4). This structure was chosen to ensure that we would still have sufficient training and test data for each participant, even if they dropped out before completing all four parts. To minimize attrition between tasks, participants received a $22.50 bonus for completing all four tasks, effectively doubling their compensation.

Participants were screened to include only participants who were: US Residents, first-language English speakers, between the ages of 18 and 65, and Reddit users. In total 34 participants completed at least one of the tasks, and 32 participants completed all four tasks. Below, we provide a few additional details about how training and test labels were solicited.

4.3.1 Training Labels. Figure 6 demonstrates our comment interface. We display a comment's text, the text of any immediate parent comment, and the title and body of the associated post. Usernames are replaced with pseudonyms. Our interface also contains a collapsible version of the Rule 2 guidelines. Participants were given the following prompt to solicit labels: "Please state whether you believe Rule 2 (Banning rudeness/hostility towards other users) applies to the highlighted comment." The answer choices were: "The comment violates Rule 2"/"The comment does not violate Rule 2." To mimic moderation log data, each comment in the training dataset was shown to

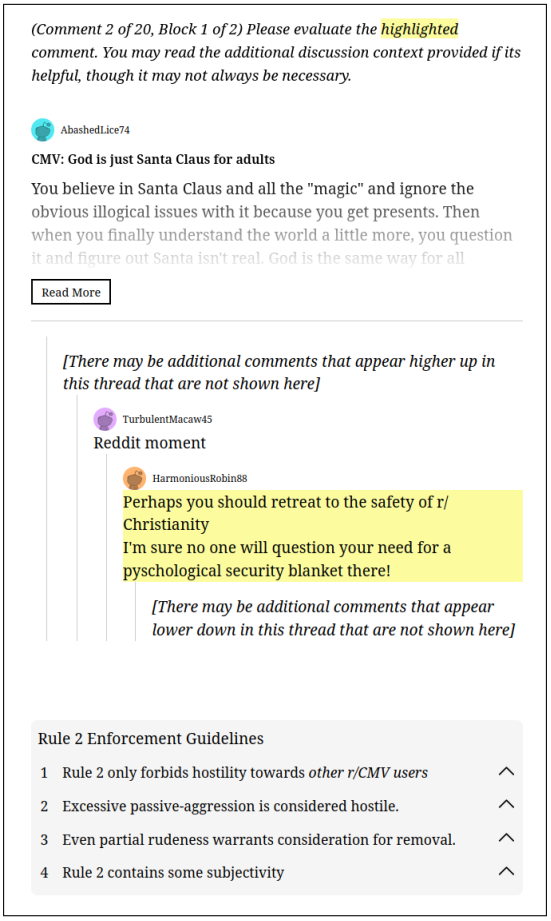


Fig. 6. Survey interface recreation of a comment. In addition to providing the text of a comment, we also include a parent comment (where appropriate), and the post associated with the comment.

a single survey-taker. Participants who completed all four tasks provided 100 training labels. A few participants experienced technical issues with the survey and supplied a different number of labels (ranging from 80 to 160). In total 3280 comments were labeled for the training dataset.

4.3.2 Test Labels. Each comment in the test dataset was shown to multiple participants. We used the same comment recreation interface for test set comments. We decided to use the test set as an opportunity to assess how well human raters were able to anticipate disagreements, creating a baseline for a priori disagreement prediction performance. To do this, we informed participants that comments in the test set section were being shown to 6 other human raters. In addition to asking them to apply Rule 2, we asked participants to predict whether the other 6 raters would be in "high consensus" or "low consensus" about whether Rule 2 applied. High consensus cases were those where a super-majority of 5 or more is achieved. To incentivize accurate predictions, participants received a small bonus of 10 cents for each correct prediction (maximum of $2 bonus for a single task). Crucially, each rater was asked about how the other 6 raters would respond to avoid creating incentive to distort their personal opinion. We did not include an a posteriori prediction task, since

this would be dependent on both the rater's personal opinion, and the responses of other raters. Under this scheme, we need 7 rater labels per comment. Due to a technical issue with our sampler, we ended up soliciting 10 labels per comment instead, giving us a few extra. In total 200 comments were labeled for the test dataset.

4.4 Prolific Study Results

Model	AUROC	Accuracy	Precision	Recall
Rule 2 Application (Rater-level Annotations)	0.8163	0.7345	0.7515	0.7933
Rule 2 Application (Majority Vote)	0.8540	0.7692	0.7425	0.8794
A Priori Disagreement (Rater-Aware)	0.6197	0.6028	0.4909	0.4204
A Priori Disagreement (Rater-Blind)	0.6295	0.6154	0.5101	0.4560
A Priori Disagreement (Human-Generated)	N/A	0.6212	0.5219	0.4005

Table 3. Model prediction quality for the Prolific dataset

4.4.1 Rule 2 Application Predictions. As in Section 4.2, the first two rows of Table 3 demonstrate our model's ability to predict Rule 2 determinations. Our model achieves an accuracy of 73% (AUROC 0.82) when predicting individual annotator ratings (using a threshold of 0.5), and 77% (AUROC 0.85) when predicting the majority vote amongst all annotators. Unsurprisingly given the size of the respective training datasets, our model performs slightly worse on this task compared to the toxicity task.

4.4.2 A Priori Disagreement Predictions. Table 3 also contains the results of the a priori disagreement prediction task. We also include the quality of the human supplied disagreement predictions. When computing disagreement prediction performance, we randomly select a rater for each comment, and compare their disagreement prediction (and the model's prediction) against the Rule 2 determinations of a random sample of 6 other raters. We report the average of 100 simulations. Again, we see that the performance of the model is generally worse when compared against the toxicity dataset (62% accuracy vs 73%). Still, the rater-blind model is able to make disagreement predictions at roughly the accuracy of human raters (62% for both). In general, we can see that the quality of a priori disagreement predictions are substantially diminished, relative to the toxicity dataset. This could be attributed to the fact that the training comments for this task were singly-labeled, or to the smaller overall number of labeled comments. However, the human rater performance was quite modest as well. If data deficiency issues can be addressed in future work, there appears to be potential for a predictive model to outperform human capabilities. Further, human and model predictions of a priori disagreement aligned roughly 60% of the time, suggesting that there may be room for a joint human-AI team to improve on performance as well.

4.4.3 Simulation Analysis. For our simulation analysis, we include one additional strategy: "Human Disagreement Prediction-Based Allocation," in which an a priori disagreement prediction from a randomly sampled human rater is used as the panel priority score. In general we can see that all strategies are slower to converge to the optimal value compared to the toxicity dataset (Figure 5). For example, under a Majority-vote Based panel allocation strategy, around 60% of comments must be assigned to panel review to approximate the decision consistency benefits of universal panel review. This was achieved in the toxicity dataset at a 30% assignment level. Still, model based panel allocation strictly outperforms random assignment. Thus, despite the decline in both a priori and a posteriori disagreement prediction performance, our modeling approach is able to improve decision consistency and surface disagreements.

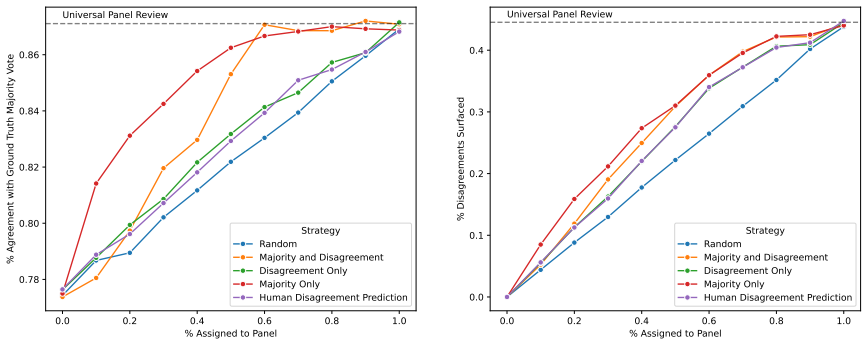


Fig. 7. Performance of a number of panel prioritization strategies as increasing workloads. When it comes to improving decision consistency, majority-vote based prioritization converges to the optimal value much faster than random allocation. Still, convergence is much faster for the toxicity dataset.

5 Think-Aloud Study: How is Venire Used in Practice?

Given the positive results from initial qualitative and quantitative assessments, we decided to move forward with building and evaluating Venire. In this section we present the finalized Venire interface [4], and the results of our evaluation interviews. Given our findings around Venire's potential as an onboarding tool (Section 3), we had interview participants pretend to be newly added moderators to the r/ChangeMyView subreddit. We gave them a sandboxed moderation queue filled with real comments reported for violating r/ChangeMyView's Rule 2. Participants were then asked to think aloud while using Venire to make Rule 2 determinations and panel assignments. Broadly, our goal was to assess whether moderators would use the system as we intended, and to re-assess whether they viewed the system as valuable. Concretely, we sought to answer the following research questions:

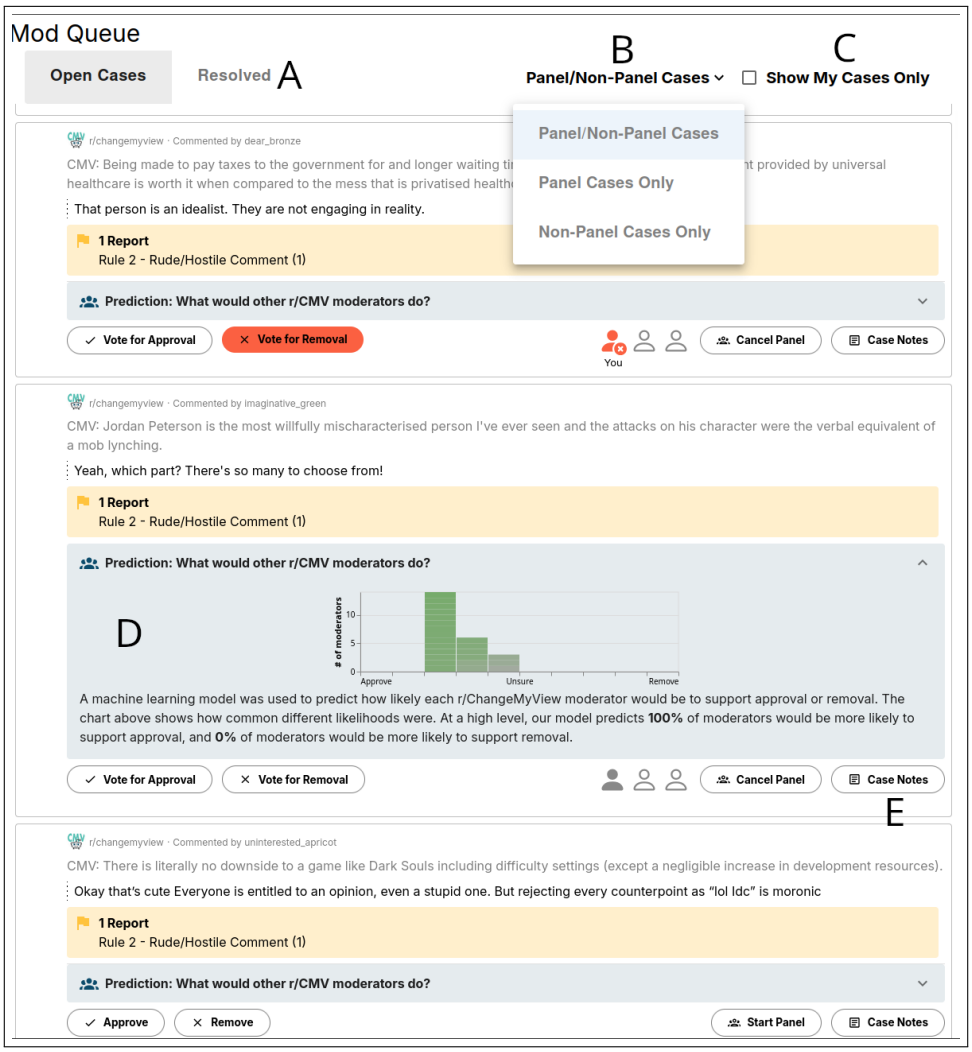
RQ3a: When do moderators consider flagging cases for panel review?
RQ3b: How did users incorporate machine learning model recommendations into their decision-making?
RQ3c: How do Venire's panel review system and model recommendations impact participants' experience learning a new community rule?
RQ3d: What kinds of communities do moderators foresee Venire working best in?

5.1 Final Interface

Based on the feedback from our initial interviews, we implemented the strict voting mock-up. We made a few updates to the interface based on lower-level feedback from our initial interviews. Most notably, we removed the "Review" modal and moved buttons directly onto the case card. This reduced the number of necessary clicks. We also implemented the following features (Section 5.1):

Resolved Tab (A): A separate queue for past cases that have been reviewed by moderators. This was included because a few moderators (N=2) mentioned in the initial interviews that they were interested in reviewing the log of past panel decisions. (See Appendix C for an example of a resolved case)

[4]Code for the interface can be found here: https://github.com/koreanwglasses/disagreement-prediction-prototype



Labeled Venire interface. A: The button used to access Venire's log of resolved moderation cases. B: A filter that lets users look at only cases that are already in panel mode. C: A filter that lets users look at only cases they have voted on already. D: A collapsible visualization of model predictions for a specific case. Additional text is included here when cases are recommended for panel review. E: The button used to access a list of moderator-generated notes for a particular case. This can be used as a channel for case-specific deliberation between moderators.

Panel Filter (B): Filters that let the moderators see only cases in panel mode, see cases not in panel mode, or see both kinds of cases. This allows moderators to prioritize decision-making as they see fit.

My Cases Filter (C): A filter that lets a moderator view only cases that they had previously ruled on. This allows moderators to check on the status of panels they previously participated in.

Panel Predictions (D): An updated version of our panel prediction visualization. Rather than dividing moderator predictions into binary categories of "Approve" and "Remove", we display a histogram of the model-produced prediction score for each moderator. To ensure that these scores are interpretable, we calibrate the model using Platt scaling on a validation set. We also provide a text description, which contains an explicit recommendation for panel review whenever the model predicts failure to achieve a supermajority amongst 70% or more of the moderation team.
Case Notes (E): A small chat interface which moderators can use to communicate their thoughts on specific cases. This was requested a few times in our initial interviews (N=3). See Appendix C for the case notes interface.
Thread Context: Clicking anywhere on a case card lets moderators view additional thread context for the reported comment. See Appendix C for the comment recreation interface.

Additionally, if a moderator makes a removal decision without starting a panel, the system will occasionally display a pop-up recommending panel review (Appendix C). This happens either when the model believes a priori or a posteriori that a disagreement is likely. An a priori prediction is made when the model predicts moderators will not achieve a 70% supermajority decision. An a posteriori prediction is made if the model predicts 80% or more of moderators will disagree with the decision *inputted by the user*. This is similar to our **Predicted Disagreement+Majority Based Assignment** strategy from section 4. Although we found that **Predicted Majority-Based Assignment** was more effective, we chose to include the disagreement score based alert as there may be qualitative differences that make these cases valuable to discuss.

5.2 Recruiting and Protocol

To recruit participants, we reached out to the moderators from the preliminary interviews. Five agreed to a follow-up interview. We recruited one additional moderator from P8's subreddit, giving us a total of 6 interviews. To train the final predictive model for Venire, we used the r/ChangeMyView, Rule 2 report dataset from our Prolific study (Section 4.3). This time, however, we used the actual moderation decisions and moderator IDs from the r/ChangeMyView moderation log, rather than the data supplied by Prolific raters. In total, we used 3369 comments for training and 481 comments for validation and calibration. We populated the mod queue with the 200 test set comments from the Prolific study. Our model achieved an accuracy of 75% on this set of comments (AUROC: 0.84, Precision: 75%, Recall: 79%)

The interview consisted of three parts. First, we gave participants an overview of the system and informed them that they would be acting as a new r/ChangeMyView moderator. Then, we gave participants an overview of r/ChangeMyView and r/ChangeMyView's Rule 2, as in the Prolific study (section 4.3). Participants were also given a guided tour of the interface. When describing the predictive model, we informed participants of the dataset the model was trained on, the input features the model used, and the accuracy rate the model achieved at predicting removals. Although we wanted to provide participants information about the model's ability to predict disagreements, we only had this information for the toxicity and Prolific rater datasets, rather than for the dataset of real-world r/ChangeMyView moderation decisions the model was trained on. Thus we chose not to provide information about disagreement prediction performance to avoid biasing participants. Next, participants were instructed to explore the interface and think aloud while using the tool to make Rule 2 decisions, for around 20 minutes. The interviewer occasionally prompted participants to elaborate on why they made certain decisions. To improve ecological validity, the research team preset some of the mod-queue cases to be in panel mode or to appear in the resolved queue at the beginning of the interview. This procedure is described in the Appendix D. After using the tool,

moderators were asked a series of closing questions to reflect on their experience using Venire. Interviews lasted one hour long, and participants were compensated $50.

5.3 Results

5.3.1 When do moderators use panel review? (RQ3A). All moderators described starting panels when they were unsure of the best decision (N=6). Frequently, moderators described these cases as "borderline." P2, for example, described starting panels when they could "see an argument both ways." The precise reasons for why a case was considered borderline varied. Sometimes, moderators described it in terms of the severity of the infraction. For example P8 said "This guy was sort of rude, but not really", in response to a comment containing the phrase "Facts don't care about your feelings, honey". Other times, moderators referenced inherent ambiguity in the text. For instance, after reading a comment where one user called another user a Marxist, P2 said, "It does call into question if just saying that somebody is affiliated with a political party is an insult. Some people wouldn't mind that affiliation [...] but I know other people could take it just straight up as an insult."

Occasionally, participants mentioned starting a panel when they felt a comment should be removed for reasons outside of Rule 2 (N=2). P3 described one instance of this. In response to a sarcastic comment saying "wow what an original and thought-provoking opinion yaaawwwnnn" they said "I don't think [the comment] is contributing. So looking specifically at rule two, I think that it would be probably OK, but I'm sure that there's larger reason that we might want to remove it." Additionally, one participant, P7, started a panel to preempt personal bias from influencing their decision, saying "If I disagree with someone, I'm honestly more reluctant to remove their comment [...] , I would definitely move to a panel [...] especially since, at least for my subreddit, we tried to be as evenly political leaning as possible."

5.3.2 Incorporating Model Predictions into the Decision-making. On the whole participants felt the model predictions did not sway their personal preference for removal or approval (N=3), but did influence their decision to start a panel (N=5). Participants differed on how they factored in the model predictions into panel decisions. Most commonly, participants considered starting a panel when the model's predictions differed from their own decision (N=5). Less frequently, participants said they would start a panel when the model predicted a split in the mod team (N=2). As such, participants rarely saw the panel recommendation pop-ups. Interestingly, model predictions played a role in deciding when *not* to start a panel (N=4). When the model predictions aligned with their removal decision, participants would decline to start a panel to avoid creating extra work for others. Participants described using the model like this as a "gut check" (N=4). Participants often described weighing their personal certainty alongside the model's confidence. P3 described this as follows:

> "I would start a panel either if I was feeling unsure and the model was also unsure [...] Or if the model seemed to be quite sure of itself and I was disagreeing with the model. Whereas if I feel confident and the model is aligned, definitely I'm not going to start a panel. If I feel confident and the model is unsure... then I might start a panel? [...] And likewise if I felt unsure and the model seemed to be pretty certain, I probably wouldn't start a panel in that case. I would probably just go with what the model predicts"

A few participants mentioned considering the model's accuracy rate when weighing its suggestions in their decision-making (N=2). In general, these participants felt that the model's 75% accuracy rate was high enough to provide useful information, but not so high that they believed every prediction it made.

Participants were split on how often they looked at the model predictions. Most commonly, participants said they would not look at the prediction if they were confident in their decision (N=3). Some participants disagreed. For example, P7 said that they "like to make sure, if I'm confident, that

my confidence is well placed" and that they usually checked the model predictions since "it just takes a second and it don't hurt." A few participants tried to avoid biasing themselves by coming to their own decision before looking at the model prediction (N=2).

5.3.3 Venire as an onboarding tool. On the whole, participants found Venire to be beneficial to their experience as a "new" ChangeMyView moderator (N=5). Participants described the benefits of the panel system and the predictive model slightly differently. For the panel system, participants highlighted its benefits as a built-in channel for existing subreddit onboarding processes (N=3). Participants often contrasted the panel system with existing moderator group chats. P3 described the panel system as "clean," since they "don't want to start five different threads in the mod chat" when they needed advice on multiple decisions. P2 felt that the panel system allowed them to be more active since "I can weigh in on cases that I have a weaker opinion on" and that "it also means that I can take an action on everything in the mod queue."

In contrast, the predictive model was described as helpful for understanding subreddit norms, especially in places where the rules were not clearly specified (N=4). P9, for example, said that "it made it so you can do a lot more on your own" since "you're not sitting there having to bother other mods asking a million questions." P2 echoed this sentiment saying "different subreddits will have a different tolerance for what they consider uncivil" and that the model predictions "would give [them] confidence in understanding the culture of the subreddit." They summarized these sentiments neatly, describing the model as a "a distilled version of past moderation decisions."

5.3.4 What kinds of communities should use Venire? (RQ3d). Participants felt that Venire would be most effective on large subreddits (N=4). This was in part because those subreddits would have a "larger training data set" (P3), but also because in a large moderation team, moderators "don't talk to each other very often." Participants also mentioned political subreddits (N=2) since they contained more ""gray area" and "nuance." Interestingly, a few moderators actually felt Venire would be more useful when the subreddit rules are poorly defined (N=2) since there would be more "vague areas" and "ambiguity" where the predictive model and panel system could help.

6 Discussion

Venire is best understood as a system that empowers moderators to manage the tradeoff between maximizing decision quality and minimizing workload. This is in stark contrast to most other AI-based moderation tools, which focus exclusively on minimizing workload [5, 13, 22]. Quantitatively, we measure decision quality by looking at how often the final decision for a case reflects the majority viewpoint amongst human raters. We measure workload by looking at how many human raters are involved in each decision. Venire's two features, the panel system and the ML model recommendations, work in concert to help moderators balance these two concerns. The panel system gives moderators a formal way to incorporate more voices into a single decision, thereby improving decision consistency. The machine learning model, on the other hand, helps moderators identify which cases would most benefit from panel review, minimizing workload. ' Still, our quantitative evaluations in Section 4 belie more nuanced mechanisms through which Venire can impact decision consistency and workload. Notably, the direct benefits to decision consistency found in our quantitative evaluations reflect only a fraction of the greater value of surfacing disagreements. In our interviews, we found that moderation teams already discuss discrepancies and update policy to iron out disagreements. Ideally, Venire can help moderators identify points of disagreement more quickly, resulting in more policy updates that reduce the need for future panel review.

With respect to workload, our interviews point to additional benefits to Venire we had not initially anticipated. In our preliminary interviews, participants speculated that Venire could act

as an onboarding tool for new moderators. They felt that ML-guided panel reviews could act as a "safety net" for new moderators, minimizing the risk that a new moderator will make a decision out of alignment with the moderation team as a whole. Crucially, if Venire makes it easier to recruit new moderators, it could actually lead to a decrease in per-person moderation workload. Our evaluation interviews lend credence to this: moderators described the ML model recommendations as a "distilled" version of a subreddit's moderation decision history. Moderators found this helpful for understanding the norms of a particular subreddit's moderation style, especially in places where the rules were not fully specified. Moderators also noted that the panel review system allowed them to "take an action on everything in the mod queue" rather than "leaving the complicated stuff for someone else." In the status quo Reddit moderation queue, moderators may spend time reviewing a case that they end up declining to make a decision on, contributing to group inefficiency. Panel review can reduce overall workload if it encourages moderators to weigh in early, preventing cases from sitting in the queue.

While we believe our work represents a comprehensive investigation into the feasibility of Venire, in the remainder of the section, we highlight a few factors that should be explored more closely before deploying Venire in the wild.

6.1 Training Venire's prediction model

In an ideal world, a Venire instance could be trained for a community using only existing moderation log data. That way, moderators would not need to spend time labeling additional training data points. However, moderation log data may be insufficient for a few reasons. First, real-world moderation log data is largely singly labeled [39]. That is, each case has a single label associated with it, supplied by a single decision-maker. This means that the model cannot learn from explicit contrasts between decision-makers, making disagreement prediction more challenging. Second, moderation log data could contain too few decisions *per moderator*. Newly added moderators, for example, will be difficult for the model to make accurate predictions about [28]. Finally even with sufficient labels per case and per moderator, many communities may not have enough total cases for model training [5]. Our technical evaluations reveal that all three of these factors can bottleneck model performance, especially with respect to "a priori" disagreement prediction (those made before an initial moderator action is taken). Our model was able to allocate panels much more efficiently on the large, multiply-labeled toxicity dataset than on the smaller, singly-labeled Prolific dataset. Even then, performance on the larger dataset was likely still limited by the fact that it had fewer labels per rater. Future work should think about all three aspects of dataset size when considering whether model deployment makes sense.

In cases where a community lacks sufficient training data in its moderation log, Venire's panel review system could be deployed without the backing ML model. Recommendations could be turned on once a sufficient amount of training data has been seen. Since the panel system naturally produces multiply-labeled data, waiting to activate the model recommendations allows us to collect more comments with multiple labels and more labeled comments in general. For perspective, our r/ChangeMyView dataset contained only two months of data from a relatively large subreddit. As such, we think that large, older communities could achieve a sufficient amount of training data within a reasonable time frame through this approach. This strategy—waiting till a sufficient number of human labels have been seen before activating an underlying ML model—is a classic pattern in human-AI system design [30]. Further, given that our modeling strategy relies on relatively small transformer models rather than LLMs, our approach is also quite scalable, meaning that too much data should not be a problem. Still alternative model architectures could be explored if working with truly massive datasets (i.e. on the scale of millions of training data points).

Moderation log data could also suffer from issues with selection bias. In the actual moderation queue, moderators *choose* which cases they want to make a decision on [41]. For instance, moderators might be more likely to handle cases that they have a strong opinion on. When we trained our prediction model on r/ChangeMyView log data, we evaluated the model's ability to make predictions on a hold out portion dataset. If strong selection effects are at play, this test set cannot capture the model's ability to predict a moderator's decisions on cases that they would not have normally opted to handle. Constructing an adequate test set would require asking moderators to do additional labeling work outside the moderation queue. This labeling work could take multiple forms. One option might be to incorporate active learning into the model training process, and have the model explicitly query moderators for new labels [33]. Another option might be to combine Venire with a community data curation tool like Wikibench [27]. Regardless, more work is necessary to understand the precise degree and nature of potential selection bias issues.

6.2 Unintended effects on decision quality

Although Venire seems to improve decision consistency, it is possible that the tool could have unintended long-term effects on decision *quality*. Understanding and mitigating these effects is crucial to successfully deploying Venire in a real online community. Our final set of interviews point to a few promising areas for further investigation. With respect to the panel system, one concern is that voting-based aggregation could *decrease* the amount of discussion in moderation teams. Participants in our final interviews often described the panel system as "more organized" or "more efficient" than consulting a moderator group chat. While this was framed as a positive, our interface may not support robust discussion in the way that a group chat does. Stronger support for deliberation or integration with chat software like Discord could address this. One participant, P3, suggested adding a deliberation phase to the panel decision-making process after the votes have been cast.

> "If I felt very strongly one way and I was the minority [..] I would want to be able to hop on discord with the other two moderators and say I feel very strongly, [...] I'd want to have a larger discussion. Because obviously, if something gets approved, it can always be removed later. It's harder to go in the other direction."

Another possibility is that ML-model recommendations could bias moderator decision-making. In the worst case, this could lead to a kind of information cascade [1] where each moderator is overly influenced by the model prediction and fails to communicate their personal judgment to other decision-makers. This would be especially concerning if the model predictions demonstrate bias against specific user groups.

Still, our interview findings suggest that moderators are unlikely to be altering their decisions based on the model output. Moderators described making up their own minds before looking at model recommendations. Rather than influencing their decision to remove or approve a comment, participants instead described the model as playing a role in their choice to start a panel or not. Most commonly, moderators described starting a panel when the model disagreed with them, or being convinced to forgo a panel when the model agreed with them. This indicates that the system is working as intended: a signal of the individual moderator's judgment is still being transmitted, regardless of whether a panel is started or not. Thus, we think it is unlikely that Venire would negatively impact the accuracy, fairness, or transparency of the moderation process. Still, our interviews rely on moderators to self-report their thought processes—we cannot rule out the possibility that moderators are being biased without realizing it. Following recommendations about transparency from Gillespie, future work could consider building tools to allow community

members to audit Venire's ML model recommendations [13], serving as a final check against fairness and bias issues.

Moderators also raised the possibility that the panel review process could increase the decision time for flagged cases. This means that potentially harmful comments could remain up for longer on the subreddit. This possibility is consistent with prior work, which finds that multi-person content review procedures suffer from slow decision times [29]. During our interviews, one moderator provided a solution to this issue. They suggested using the first vote on a panel case as an initial decision, and allowing the subsequent votes to overturn this decision if necessary. This ensures that a quick decision can be made, but allows for mistakes to be corrected.

6.3 Generalizing to other communities

In this paper, we evaluate Venire in the context of a single rule from a single subreddit: r/ChangeMyView's Rule 2, forbidding excessive rudeness or hostility towards other users. It is worth asking how Venire would perform for other rules, communities, and platforms. One important consideration is that Rule 2 requires relatively little context to adjudicate – most Rule 2 decisions can be made by looking at only a single reported comment. This may not be true for more complicated moderation decisions where it is necessary to look for patterns of behavior across time [3]. Take for example, a chat-based community like a Discord or Twitch channel. Moderators might rely on bans rather than content removals to sanction misbehaving users. Prior work has found that moderators take into account the frequency, recency, and severity of prior rule violations when making ban determinations [41]. Adapting Venire to such a context would require re-engineering the features fed into the model, and adjusting the neural network architecture appropriately.

Further, r/ChangeMyView's approach to moderation is relatively unique amongst online communities. r/ChangeMyView benefits from having content moderation rules that are exceptionally well-specified — prior work has found few communities document their moderation practices in such detail [24]. In general, the prevalence of more ad-hoc approaches to content moderation [41] makes it unclear how Venire would perform outside of r/ChangeMyView. On the one hand, under-specified moderation guidelines could make the underlying prediction task harder, since decisions made by a single moderator may suffer from greater internal inconsistency. On the other hand, communities with vague moderation guidelines might also be more likely to suffer from the kinds of inter-moderator decision consistency problems that Venire seeks to address [6], creating more potential upside to deploying the system. This was pointed out explicitly by two participants during the evaluation interviews, who felt that Venire was most helpful for ironing out "vagueness" and "grey areas" within a community's moderation policy. Thus, moderators interested in using Venire should assess whether the community's moderation log contains a strong signal of individual moderators' preferences that a machine learning model can learn from, regardless of how clearly community guidelines are spelled out in writing.

7 Conclusion

In this work, we present a comprehensive exploration of Venire, a machine learning-guided panel review system for community content moderation. Through a series of three studies, we provide quantitative and qualitative evidence that Venire improves decision consistency between moderators and surfaces latent disagreements within moderation teams. More broadly, we argue that Venire helps moderators navigate the tradeoff between maximizing decision quality and minimizing workload. Unlike prior work on AI content moderation systems, Venire represents an attempt to use machine learning to more efficiently allocate human decision-makers, rather than replace them outright. We call for more CSCW research that supports reflective practice amongst moderators, empowering them to refine community policy through case-based reasoning.

References

[1] Lisa R Anderson and Charles A Holt. 1997. Information cascades in the laboratory. *The American economic review* (1997), 847–862.

[2] Lindsay Blackwell, Jill Dimond, Sarita Schoenebeck, and Cliff Lampe. 2017. Classification and its consequences for online harassment: Design insights from heartmob. *Proceedings of the ACM on Human-Computer Interaction* 1, CSCW (2017), 1–19.

[3] Lia Bozarth, Jane Im, Christopher Quarles, and Ceren Budak. 2023. Wisdom of Two Crowds: Misinformation Moderation on Reddit and How to Improve this Process—A Case Study of COVID-19. *Proceedings of the ACM on Human-Computer Interaction* 7, CSCW1 (2023), 1–33.

[4] Federico Cabitza, Andrea Campagner, and Valerio Basile. 2023. Toward a perspectivist turn in ground truthing for predictive computing. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 37. 6860–6868.

[5] Eshwar Chandrasekharan, Chaitrali Gandhi, Matthew Wortley Mustelier, and Eric Gilbert. 2019. Crossmod: A cross-community learning-based system to assist reddit moderators. *Proceedings of the ACM on human-computer interaction* 3, CSCW (2019), 1–30.

[6] Quan Ze Chen and Amy X Zhang. 2023. Judgment Sieve: Reducing uncertainty in group judgments through interventions targeting ambiguity versus disagreement. *Proceedings of the ACM on Human-Computer Interaction* 7, CSCW2 (2023), 1–26.

[7] Frederick Choi, Tanvi Bajpai, Sowmya Pratipati, and Eshwar Chandrasekharan. 2023. ConvEx: A Visual Conversation Exploration System for Discord Moderators. *Proceedings of the ACM on Human-Computer Interaction* 7, CSCW2 (2023), 1–30.

[8] Amanda LL Cullen and Sanjay R Kairam. 2022. Practicing moderation: Community moderation as reflective practice. *Proceedings of the ACM on Human-computer Interaction* 6, CSCW1 (2022), 1–32.

[9] Jacob Devlin. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805* (2018).

[10] Bryan Dosono and Bryan Semaan. 2019. Moderation practices as emotional labor in sustaining online communities: The case of AAPI identity work on Reddit. In *Proceedings of the 2019 CHI conference on human factors in computing systems*. 1–13.

[11] Eve Fleisig, Rediet Abebe, and Dan Klein. 2023. When the majority is wrong: Modeling annotator disagreement for subjective tasks. *arXiv preprint arXiv:2305.06626* (2023).

[12] 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)*, Kevin Duh, Helena Gomez, and Steven Bethard (Eds.). Association for Computational Linguistics, Mexico City, Mexico, 2279–2292. doi:10.18653/v1/2024.naacl-long.126

[13] Tarleton Gillespie. 2018. *Custodians of the Internet: Platforms, content moderation, and the hidden decisions that shape social media*. Yale University Press.

[14] Mitchell L Gordon, Michelle S Lam, Joon Sung Park, Kayur Patel, Jeff Hancock, Tatsunori Hashimoto, and Michael S Bernstein. 2022. Jury learning: Integrating dissenting voices into machine learning models. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*. 1–19.

[15] Danna Gurari and Kristen Grauman. 2017. CrowdVerge: Predicting If People Will Agree on the Answer to a Visual Question. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems* (Denver, Colorado, USA) *(CHI '17)*. Association for Computing Machinery, New York, NY, USA, 3511–3522. doi:10.1145/3025453.3025781

[16] Aaron Halfaker and R Stuart Geiger. 2020. Ores: Lowering barriers with participatory machine learning in wikipedia. *Proceedings of the ACM on Human-Computer Interaction* 4, CSCW2 (2020), 1–37.

[17] Manoel Horta Ribeiro, Justin Cheng, and Robert West. 2023. Automated content moderation increases adherence to community guidelines. In *Proceedings of the ACM web conference 2023*. 2666–2676.

[18] Jane Hsieh, Joselyn Kim, Laura Dabbish, and Haiyi Zhu. 2023. " Nip it in the Bud": Moderation Strategies in Open Source Software Projects and the Role of Bots. *Proceedings of the ACM on Human-Computer Interaction* 7, CSCW2 (2023), 1–29.

[19] Evey Jiaxin Huang, Abhraneel Sarma, Sohyeon Hwang, Eshwar Chandrasekharan, and Stevie Chancellor. 2024. Opportunities, tensions, and challenges in computational approaches to addressing online harassment. In *Proceedings of the 2024 ACM Designing Interactive Systems Conference*. 1483–1498.

[20] Jane Im, Sonali Tandon, Eshwar Chandrasekharan, Taylor Denby, and Eric Gilbert. 2020. Synthesized social signals: Computationally-derived social signals from account histories. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*. 1–12.

[21] Shagun Jhaver, Darren Scott Appling, Eric Gilbert, and Amy Bruckman. 2019. " Did you suspect the post would be removed?" Understanding user reactions to content removals on Reddit. *Proceedings of the ACM on human-computer*

interaction 3, CSCW (2019), 1–33.

[22] Shagun Jhaver, Iris Birman, Eric Gilbert, and Amy Bruckman. 2019. Human-machine collaboration for content regulation: The case of reddit automoderator. *ACM Transactions on Computer-Human Interaction (TOCHI)* 26, 5 (2019), 1–35.

[23] Shagun Jhaver, Alice Qian Zhang, Quan Ze Chen, Nikhila Natarajan, Ruotong Wang, and Amy X Zhang. 2023. Personalizing content moderation on social media: User perspectives on moderation choices, interface design, and labor. *Proceedings of the ACM on Human-Computer Interaction* 7, CSCW2 (2023), 1–33.

[24] Prerna Juneja, Deepika Rama Subramanian, and Tanushree Mitra. 2020. Through the looking glass: Study of transparency in Reddit's moderation practices. *Proceedings of the ACM on Human-Computer Interaction* 4, GROUP (2020), 1–35.

[25] Vinay Koshy, Tanvi Bajpai, Eshwar Chandrasekharan, Hari Sundaram, and Karrie Karahalios. 2023. Measuring User-Moderator Alignment on r/ChangeMyView. *Proceedings of the ACM on Human-Computer Interaction* 7, CSCW2 (2023), 1–36.

[26] Deepak Kumar, Patrick Gage Kelley, Sunny Consolvo, Joshua Mason, Elie Bursztein, Zakir Durumeric, Kurt Thomas, and Michael Bailey. 2021. Designing toxic content classification for a diversity of perspectives. In *Seventeenth Symposium on Usable Privacy and Security (SOUPS 2021)*. 299–318.

[27] Tzu-Sheng Kuo, Aaron Lee Halfaker, Zirui Cheng, Jiwoo Kim, Meng-Hsin Wu, Tongshuang Wu, Kenneth Holstein, and Haiyi Zhu. 2024. Wikibench: Community-driven data curation for ai evaluation on wikipedia. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*. 1–24.

[28] Xuan Nhat Lam, Thuc Vu, Trong Duc Le, and Anh Duc Duong. 2008. Addressing cold-start problem in recommendation systems. In *Proceedings of the 2nd international conference on Ubiquitous information management and communication*. 208–211.

[29] Cliff Lampe and Paul Resnick. 2004. Slash (dot) and burn: distributed moderation in a large online conversation space. In *Proceedings of the SIGCHI conference on Human factors in computing systems*. 543–550.

[30] Gierad Laput, Walter S Lasecki, Jason Wiese, Robert Xiao, Jeffrey P Bigham, and Chris Harrison. 2015. Zensors: Adaptive, rapidly deployable, human-intelligent sensor feeds. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems*. 1935–1944.

[31] Hanlin Li, Brent Hecht, and Stevie Chancellor. 2022. Measuring the monetary value of online volunteer work. In *Proceedings of the International AAAI Conference on Web and Social Media*, Vol. 16. 596–606.

[32] Elizabeth Lucas, Cecilia O Alm, and Reynold Bailey. 2019. Understanding human and predictive moderation of online science discourse. In *2019 IEEE Western New York Image and Signal Processing Workshop (WNYISPW)*. IEEE, 1–5.

[33] Thodoris Lykouris and Wentao Weng. 2024. Learning to defer in content moderation: The human-ai interplay. *arXiv preprint arXiv:2402.12237* (2024).

[34] Renkai Ma and Yubo Kou. 2022. " I'm not sure what difference is between their content and mine, other than the person itself" A Study of Fairness Perception of Content Moderation on YouTube. *Proceedings of the ACM on Human-Computer Interaction* 6, CSCW2 (2022), 1–28.

[35] Sarumi Oluyemi, Béla Neuendorf, Joan Plepi, Lucie Flek, Jörg Schlötterer, and Charles Welch. 2024. Corpus Considerations for Annotator Modeling and Scaling. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, Kevin Duh, Helena Gomez, and Steven Bethard (Eds.). Association for Computational Linguistics, Mexico City, Mexico, 1029–1040. doi:10.18653/v1/2024.naacl-long.59

[36] Maithra Raghu, Katy Blumer, Rory Sayres, Ziad Obermeyer, Bobby Kleinberg, Sendhil Mullainathan, and Jon Kleinberg. 2019. Direct uncertainty prediction for medical second opinions. In *International Conference on Machine Learning*. PMLR, 5281–5290.

[37] Paul Resnick, Yuqing Kong, Grant Schoenebeck, and Tim Weninger. 2021. Survey equivalence: A procedure for measuring classifier accuracy against human labels. *arXiv preprint arXiv:2106.01254* (2021).

[38] Paul Röttger, Bertie Vidgen, Dirk Hovy, and Janet B Pierrehumbert. 2021. Two contrasting data annotation paradigms for subjective NLP tasks. *arXiv preprint arXiv:2112.07475* (2021).

[39] Mattia Samory. 2021. On positive moderation decisions. In *Proceedings of the International AAAI Conference on Web and Social Media*, Vol. 15. 585–596.

[40] Joseph Seering and Sanjay R Kairam. 2023. Who moderates on Twitch and what do they do? Quantifying practices in community moderation on Twitch. *Proceedings of the ACM on Human-Computer Interaction* 7, GROUP (2023), 1–18.

[41] Joseph Seering, Tony Wang, Jina Yoon, and Geoff Kaufman. 2019. Moderator engagement and community development in the age of algorithms. *New Media & Society* 21, 7 (2019), 1417–1443. arXiv:https://doi.org/10.1177/1461444818821316 doi:10.1177/1461444818821316

[42] Farhana Shahid, Dhruv Agarwal, and Aditya Vashistha. 2024. 'One Style Does Not Regulate All': Moderation Practices in Public and Private WhatsApp Groups. *arXiv preprint arXiv:2401.08091* (2024).

[43] Wikipedia contributors. 2022. Venire facias — Wikipedia, The Free Encyclopedia. https://en.wikipedia.org/w/index.php?title=Venire_facias&oldid=1095751539. [Online; accessed 23-September-2024].

[44] Wenjie Yin, Vibhor Agarwal, Aiqi Jiang, Arkaitz Zubiaga, and Nishanth Sastry. 2023. Annobert: Effectively representing multiple annotators' label choices to improve hate speech detection. In *Proceedings of the International AAAI Conference on Web and Social Media*, Vol. 17. 902–913.

A Preliminary Interview: Moderator Recruitment Details

Several measures were taken to avoid spamming moderators with recruitment messages. First, we avoided subreddits that we had recently recruited for other studies. Second, we messaged subreddit's such that no moderator received a recruiting message twice. To do this, we constructed a network of our filtered communities where each edge corresponded to a community that shared at least one moderator. We then ran an independent set solver to find a subset of communities that did not share any moderators. We messaged only these communities. Recruiting messages were sent to five communities a day until the desired number of participants was reached.

B Prolific Study: Survey Materials

B.1 Content Warning

In this task, you will see comments that were previously posted to r/ChangeMyView (r/CMV), some of which were removed by the community's moderators. The research team conducting this study feels that a few of these comments are potentially offensive. If you do not feel comfortable viewing such text, we advise you to exit the study. Otherwise, you can press the next button to continue on to the next page. You may exit the study at any point in time, though you will not receive payment unless you complete the full study.

Specific triggers you might encounter include references to:

- Racism
- Misogyny
- Sexual assualt
- Gun violence
- Homophobia/Transphobia
- Body shaming
- Anti-semitism
- Ableism

B.2 Introduction to r/ChangeMyView

Page 1 Instructions: "In this task, you'll be asked to pretend to act as a content moderator for r/ChangeMyView, a Reddit community. On r/ChangeMyView, people post about a viewpoint they hold that they want to have changed. In the comments section, other users submit arguments to change the original poster's (OP) view. Some posters' viewpoints may be controversial or offensive – still, such posts are allowed as long as the poster is open to having their mind changed. Below is an example of a post on r/ChangeMyView:"
Post title: "CMV:Taylor Swift is an average musician"
Post body: " I have seen many posts and heard people say many things to hype up Taylor Swift. They say Taylor Swift is a better vocalist than Adele. They say Taylor Swift is a better performer than Beyonce. I even heard someone say that Taylor Swift is one of the best songwriters of all time(when people like Alicia Keys and Bruno Mars exist). I don't think Taylor Swift is a terrible artist. She can actually hold a tune unlike Jennifer Lopez or Selena Gomez. Her Performances aren't as high energy and powerful as a Beyonce performance but something I do appreciate is that Taylor Swift can play instruments while singing which is something not many performers can do. However I don't think Taylor Swift is anywhere close to Beyonce when it comes to performing. Something I do appreciate about Taylor Swift is that she story tells through her music however all of her music is a breakup story. Where is the variety in that?

I understand that Taylor Swift is one of the biggest artists out right now but in my opinion I don't think Taylor Swift is as talented as people make her sound. She is average in all aspects of

music. Nothing about her screams “I’m the best at what I do”. Nothing about her stands out among the crowd of much better musicians. ”
Page 2 Instructions: “If a commenter changes the OP’s mind, the OP can give them a ’delta’ as a reward. Deltas act as a point system on r/ChangeMyView. The community maintains a leaderboard to show which users have earned the most deltas over time. Below is an example of a reply comment and a delta award. The "OP" symbol next to the second reply shows it’s from the original poster. The OP types ’!delta’ into their message (typing δ also works) to award a delta at the bottom of the second reply. You may occasionally see references to the delta system when labeling comments for this task.”
Example parent comment: “ She been touring for the past year (and will be for almost another year) in all different time zones each week, performing 3.5 hours of her discography. Not only is she doing dancing through those entire songs, but she also has elaborate sets and performances throughout. Not only that, but she has full songs playing the guitar and piano on top of that. These are sold out arenas around the world

And within that crazy tour, she’s continuing to write, produce and put out new music. And when we say.l new music, it’s a double album consisting of 30 tracks. Then she records music videos. She has 11 studio albums and averaging like 20+ songs each. Thats simply incredibly dedication of an artist to give their entire self to their craft.

She’s also really supportive to other artists - she’s the only person you’ll see at an award show standing at every performance. Generally unproblematic and makes friends with everyone. Invites fans all the time to events just from her or her team seeing their dedication on social media/etc. I mean she even invited fans to her house once. Every show and event you’ll see her having invited hand selected people from her fan base.

I understand her ’generic white girl pop music’ might not be for everyone but girl credit where it’s due that she put the damn work in to be where she is now.”
Example response from OP: “Yea someone else mentioned that vocally and performance wise she is average when compared to other well known singers. But they said her personality is what makes her stand out. I can appreciate that. I don’t really like Taylor’s music like that but she never causes any drama online, she always supports smaller artists and many artists already say that Taylor is a huge inspiration to them. Even though Taylor has had many breakups she never takes these breakups online like many other celebrities. She keeps her life confidential which I like. !delta
”

B.3 Rule 2 Overview

r/CMV has a few rules for commenting. In this task you’ll review real r/CMV comments and judge whether they break CMV’s Rule 2, which states: ’Don’t be rude or hostile to other users.’ Below are some basic guidelines to help you decide if Rule 2 is violated. You can view these guidelines on later pages.

- Rule 2 only forbids hostility towards other r/CMV users:Rule 2 bans comments that are rude or hostile to other users (e.g., “You’re an idiot”). It doesn’t apply to comments containing broadly offensive viewpoints, or insults toward public figures, institutions, or general categories of people. Although such comments can be offensive, r/CMV often aims to persuade people to change these views. An exception is if a group insult is clearly aimed at another commenter. For example, saying “People who wear glasses are stupid”, after another commentor says they wear glasses. It’s more acceptable to criticize arguments than people – e.g., “That argument is bad” is okay, but “You are bad” is not. Still, there are limits. Calling an argument “moronic” is probably still considered rude or hostile.

- Excessive passive-aggression is considered hostile: Rule 2 applies to comments that are excessively passive-aggressive, even if they do not contain an overt insult. A common example is claiming that another user "lacks reading comprehension abilities".
- Rule 2 applies if any part of a comment contains rudeness or hostility. If even one sentence in a three paragraph comment is hostile, the comment can be considered for removal.
- There is some subjectivity that goes into making a rule 2 determination. Please use your best judgment—although we will do some quality checks on your answers, we recognize that not everyone will interpret a comment the exact same way

B.4 Practice Comments

Comment 1: "you are an idiot. any BEING that can feel pain is not object. Simple as that."
Comment 1 Explanation (User Says Rule 2 Applies): "Good job! This comment was removed by the r/ChangeMyView moderators. Calling another user an 'idiot' is a clear example of rudeness"
Comment 1 Explanation (User Says Rule 2 Does Not Apply): " Unfortunately, the r/ChangeMyView moderators removed this comment. Calling another user an 'idiot' is a clear example of rudeness. Please select 'The comment violates Rule 2' to continue"
Comment 2: "I have a Juris Doctor degree. I am very comfrey in my understanding of the term 'rights'. The arguments you are holding onto can at best be described as purely semantic, which means I cannot help you or so vapidly stupid I'm honestly surprised you managed to create a Reddit account."
Comment 2 Explanation (User Says Rule 2 Applies): "Good job! This comment was removed by the r/ChangeMyView moderators. Although a bit less straightforward than the previous case, the commentor is insulting someone's intelligence by indicating surpise at their ability to create a Reddit account"
Comment 2 Explanation (User Says Rule 2 Does Not Apply): "Unfortunately, the r/ChangeMyView moderators removed this comment. Although a bit less straightforward than the previous case, the commentor is insulting someone's intelligence by indicating surprise at their ability to create a Reddit account. Please select 'The comment violates Rule 2' to continue"
Comment 3: "All of those statements are pretty misogynistic. Women don't need to do anything they don't want to in their dating life, and frankly their strategies for pursuing men are their own business. Dating isn't, and shouldn't be, about fairness and equality. Romantic engagements are deeply personal and people are allowed to like or dislike whatever they want in a person. So, yeah, this kinda proves how gross the incel mindset is."
Comment 3 Explanation (User Says Rule 2 Applies): " Unfortunately, the r/ChangeMyView moderators did not remove this comment. This case is borderline. Although one could argue the term 'misogynist' indicates hostility, the commentor is using the label to refer to another user's arguments, rather than to the user themselves. Please select 'The comment does not violate Rule 2" to continue' "
Comment 3 Explanation (User Says Rule 2 Does Not Apply): " Good job! This comment was not removed by the r/ChangeMyView moderators. This case is borderline. Although one could argue the term 'misogynist' indicates hostility, the commentor is using the label to refer to another user's arguments, rather than to the user themselves."

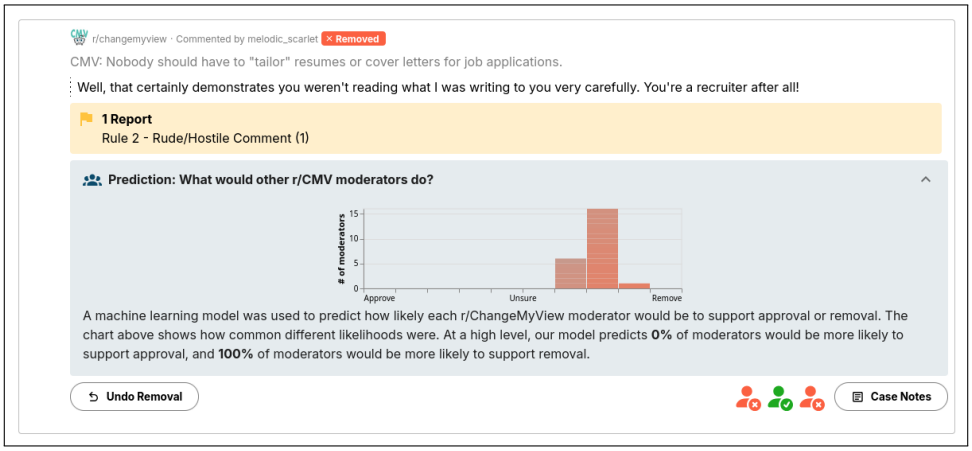
B.5 Disagreement Prediction Instructions

Thank you for completing the previous block of comments.
In the next section, you'll review another 20 comments. Each comment in this section will also be reviewed by 6 other Prolific workers.
You will be asked to do two things for each comment:

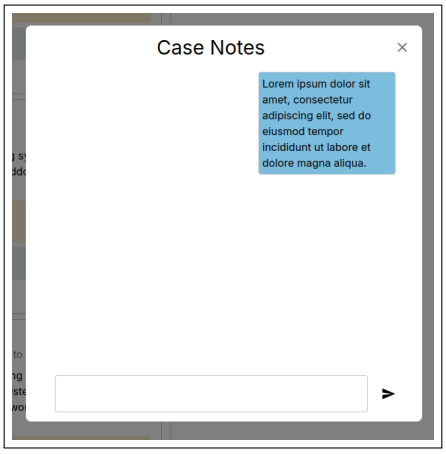
(1) You will be asked to determine whether the comment violates Rule 2.
(2) You will be asked to predict whether the other 6 raters will be be in high consensus or low consensus over the violation of Rule 2.

We say there is high consensus amongst the other 6 raters, if one or zero of the other raters dissent from the majority viewpoint (i.e. a 6/0 or 5/1 split). We say there is low consensus amongst the other raters if the other raters are evenly split, or if two raters dissent from the majority viewpoint (i.e. a 3/3 or 4/2 split).
You'll get a 10-cent bonus every time you correctly predict the consensus level for a comment – up to $2.00 for a perfect score. Bonuses will be given within a week.
Question text: "6 other raters will be shown this comment. Consider how they will respond to the previous question. Will there be high consensus amongst these raters over whether 2 has been violated (e.g. 6/0 or 5/1 split), or low consensus? (e.g. 3/3, 2/4, or 4/2 split)"
Answer Choices : "High Consensus", "Low Consensus"

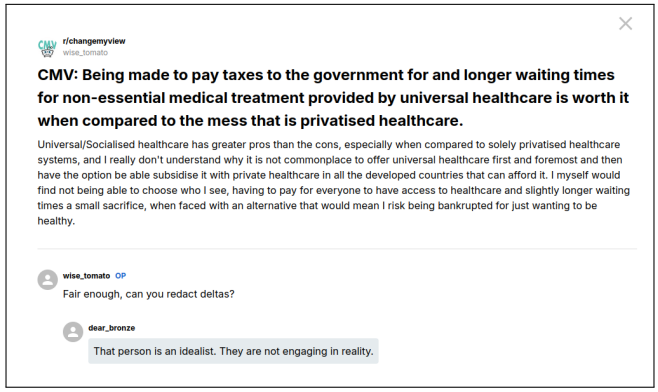
C Venire Interface



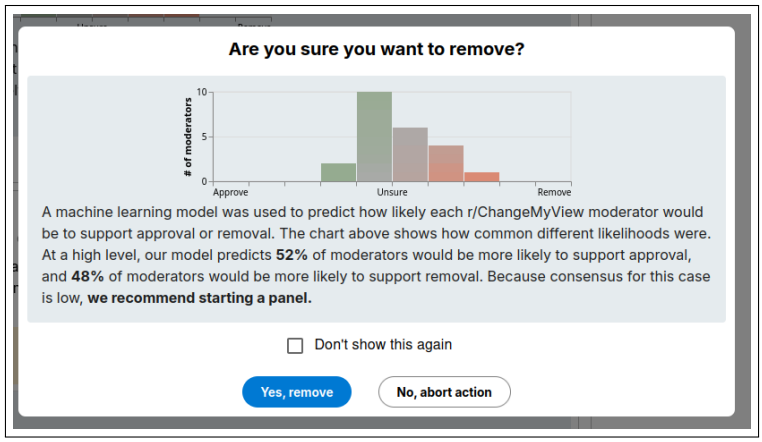
An example of a previously handled moderation case in Venire's resolved queue. Moderators can see the final status of the comment, and how the panel of decision makers voted



The Venire case notes interface. Moderators can this chat-like feature to handle case-specific deliberations



Interface recreation of additional thread context. Users can access the thread context for a case by clicking anywhere on the case card



Pop up alert recommending the user pursue panel review. The pop-up interface includes the visualization and text from the panel prediction tab

D Evaluation Interview: Presetting Moderation Queue

To improve ecological validity of the evaluation interviews, the research team preset some of the mod-queue cases to be in panel mode or in the resolved queue at the beginning of the interview. This was done using randomly sampled labels provided by Prolific raters. We set 13 cases to panel mode in places where a Prolific rater had predicted low consensus. 3 of these went into the resolved queue and 10 went into the open cases queue. Another 3 non-panel cases were included in the resolved queue. A total of 25 votes and Rule 2 decisions were preset using the Prolific ratings

Received October 2024; revised April 2025; accepted August 2025