Needling Through the Threads: A Visualization Tool for Navigating Threaded Online Discussions

Yijun Liu
University of Illinois
Urbana-Champaign
Urbana, IL, USA
yijun6@illinois.edu

Frederick Choi
University of Illinois
Urbana-Champaign
Urbana, IL, USA
fc20@illinois.edu

Eshwar Chandrasekharan
University of Illinois
Urbana-Champaign
Urbana, IL, USA
eshwar@illinois.edu

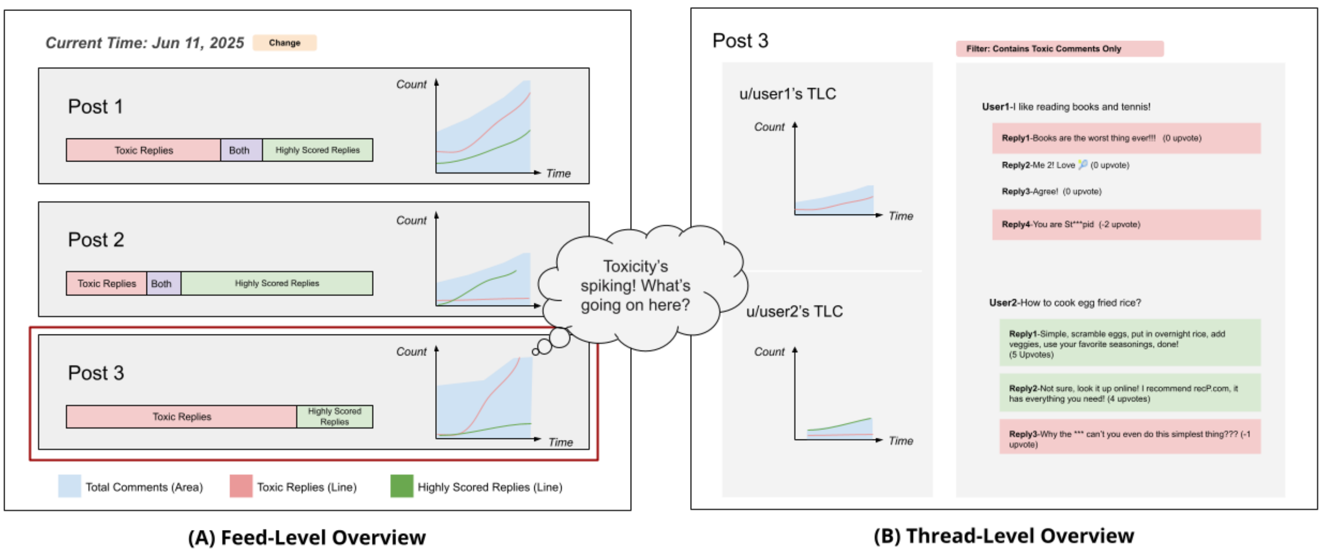

Figure 1: Needle's Multi-Level Visual Analytics Approach. (A) The feed-level overview shows temporal visualizations and toxicity patterns across all subreddit posts, with visual anomalies (like Post 3's toxicity spike) drawing immediate attention. (B) The thread-level overview supports drilling down into specific threads with Top-Level Comment (TLC) visualizations as well as highlighting and filtering capabilities to quickly identify problematic content requiring moderator attention.

Abstract

Navigating large-scale online discussions is difficult due to their rapid pace and high volume of content. Platforms like Reddit employ "threads" to visually organize parallel discussions, but deep nesting obscures conversation flow. For moderators, this fragmentation compounds the difficulty of following evolving conversations and maintaining context across threads, which limits timely and effective moderation. In this paper, we present Needle, an interactive system that applies visual analytics to summarize key conversational metrics: activity, toxicity, and voting trends over time. Needle provides both high-level overviews and detailed breakdowns of threads, enabling moderators to identify priority areas without reading through entire nested conversations. Through a user study with ten Reddit moderators, we find that Needle provides a practical solution to maintain contextual understanding when navigating threaded discussions. Based on these findings, we propose design guidelines for future visualization-based tools that shape how people consume, interpret, and make sense of large-scale online discussions.

CCS Concepts

• **Human-centered computing → Collaborative and social computing systems and tools**.

Keywords

Visual Analytics, Conversational Trajectories, Summarization, Moderation, Conversational Metrics

CHI '26, Barcelona, Spain

ACM ISBN 979-8-4007-2278-3/2026/04
https://doi.org/10.1145/3772318.3790365

ACM Reference Format:
Yijun Liu, Frederick Choi, and Eshwar Chandrasekharan. 2026. Needling Through the Threads: A Visualization Tool for Navigating Threaded Online Discussions. In *Proceedings of the 2026 CHI Conference on Human Factors in Computing Systems (CHI '26), April 13–17, 2026, Barcelona, Spain.* ACM, New York, NY, USA, 18 pages. https://doi.org/10.1145/3772318.3790365

1 Introduction

Online communities and their members are frequently exposed to *undesirable* content (e.g., toxicity and harassment) which can reduce user perceptions of quality and safety [1, 20, 78], and sometimes even lead to negative psychological effects [59, 72]. To combat online misbehavior, content moderation is the primary mechanism employed by social media platforms [32, 82, 85]. Most online communities employ the services of human moderators (either paid or volunteer, centralized or distributed) to regulate user-generated content [25, 71]. Prior work has highlighted the numerous challenges faced by human moderation approaches [25, 71, 77, 86], including the substantial impact of constant exposure to toxic content on the mental health of moderators [46, 86].

To alleviate some of these challenges, prior work in HCI has developed computational tools and approaches to augment the process of manual review by moderators [41, 45, 65, 81] and support human-in-the-loop approaches to content moderation [12, 16, 40]. Despite the increasing adoption of human-AI systems for content moderation—particularly in light of recent advances in NLP and the emergence of large language models (LLMs)—automated approaches continue to face limitations, including poor understanding of varied community norms [12], lack of human precision to fully address complex cases [52], and diminished transparency in moderation processes that obscures how and why decisions are made [28].

1.1 Complexities in Navigating Threaded Online Discussions

Large-scale online discussions further exacerbate moderation challenges due to their volume and the rapid pace of content generation [9–11, 46]. The scale and complexity of online discourse make it difficult to track conversations and maintain contextual understanding of the surrounding discussions, hindering timely and effective moderation [26, 44, 63]. For instance, a recent study highlighted how Discord moderators struggle to monitor multiple discussions that take place in parallel on different channels, track how individual conversations flow or evolve, all while maintaining contextual understanding of the surrounding discussion [16].

Platforms like Reddit employ "threads" to visually organize or structure parallel discussions. Threads can occur at different levels, from posts to comments, as shown in Figure 2 (see Threaded Discussion A), fragmenting context across multiple layers. Threaded structures offer a degree of organization by allowing discussions to branch logically, which can help users follow discussion paths with greater ease [7, 91]. However, as discussions become increasingly nested (i.e., "threaded"), maintaining context and identifying key points across multiple layers becomes challenging [34], as illustrated in Figure 2 (see Threaded Discussion B). Threaded structures fragment conversations across multiple layers, making it difficult to maintain local context in large-scale discussions, especially when topics have drifted [2, 67]. Prior work has shown that greater structural complexity reduces overall comprehensibility and readability, and that threaded formats can make it harder for readers to locate the most recent comments, creating additional navigation burdens for moderators who must track evolving conversations across multiple branches [2, 36]. For moderators, this fragmentation compounds the difficulty of tracking evolving conversations across branches and identifying problematic contents, showing a need for systems that help moderators navigate complex threaded discussions that does not require reading through entire nested conversations [16].

HCI researchers have explored tools to enhance sense-making, summarization, and visualization of large-scale online discussions in different contexts [91]. Specifically, text summarization/synthesis and visualization techniques have shown promise in making sense of complex discussions forums and Wikis [93] and conversations on Slack [92]. In the context of content moderation, previous systems have demonstrated that visual cues can improve user awareness when interacting with *stranger* accounts on Twitter [42], in addition to highlighting how visual summaries can guide moderation attention to problematic conversations on Discord [16]. We extend this line of work on using visualizations to assist content moderation in the context of single-thread conversations (or sequential chat) and develop a visualization tool to navigate a new format of online discourse—*threaded* discussions on Reddit. Our system provides high-level syntheses of key metrics to guide moderators' attention through visualizations, while also overlaying visual cues directly on the conversations to improve contextual understanding and comprehension of deeply nested conversations.

1.2 Summary of Contributions

In this paper, we present *Needle*, an interactive system that provides high-level synthesis via visual analytics and aims to support the navigation of threaded discussions on social platforms like Reddit. *Needle* can visualize key comment metrics including activity, toxicity, and number of upvotes or downvotes, as well as temporal trends of how these metrics change over time. *Needle* is designed to support moderators by providing intuitive and interactive line graph summaries of recent commenting activity and toxicity levels across different threads within a given post (or Reddit submission). Upon clicking a post, moderators can delve into a detailed visual breakdown of comment activity in each Top Level Comment (TLC) as well as the comments themselves.

Through a user study with Reddit moderators ($N = 10$), we find that *Needle* can help moderators easily navigate threaded discussions in their communities, monitor key conversational metrics, and quickly identify areas of concern within threads. The use of visualizations in *Needle* enables moderators to maintain a clear overview of parallel conversations within threads, making it easier to manage complex threaded discussions and prioritize content that requires immediate moderator attention. *Needle*'s design provides a practical solution that helps moderators maintain contextual understanding of surrounding comments in the conversation when reviewing comments within threaded discussions, without increasing cognitive load and immediate exposure to toxic content.

The contributions of this paper are three-fold. First, we make a systems contribution by developing a novel sociotechnical system to help Reddit moderators navigate threaded online discussions. Unlike prior work that primarily targeted linear conversations such as Slack or Discord, our system addresses the unique fragmentation and complexity of nested threads. *Needle* is one of the first systems to combine interactive visual analytics with *human-in-the-loop* moderation for threaded online conversations.

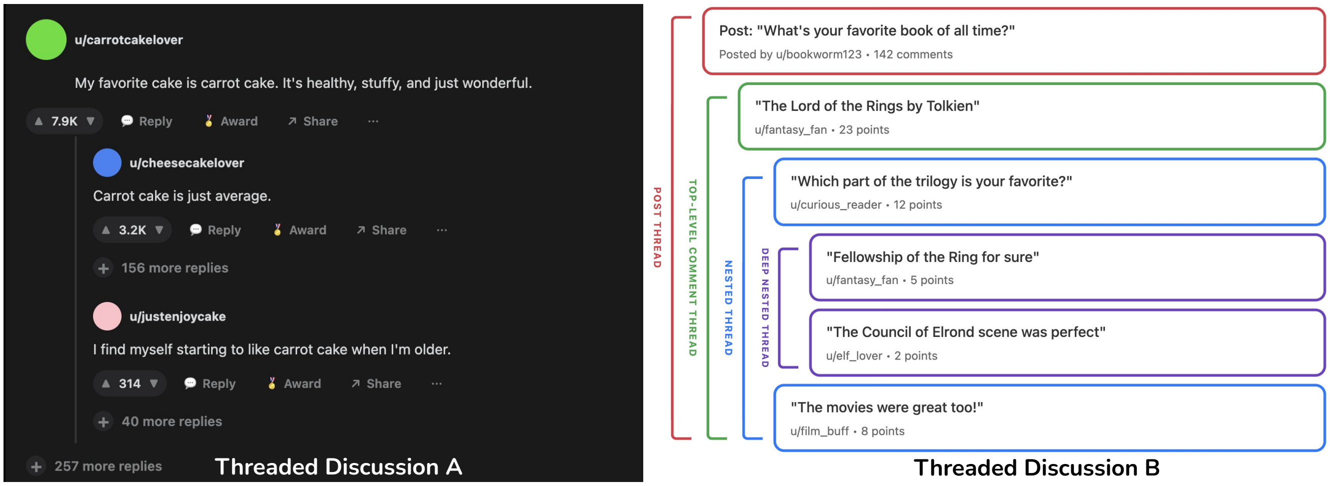

Figure 2: Threaded Discussion A is a mock discussion designed to reflect the scale and comment volume of a real large-scale Reddit thread. Threaded Discussion B illustrates the "threads" that occur at multiple levels within online discussions using synthetic data.

Second, we examine how well design guidelines synthesized from prior HCI work on content moderation, visualization, and sociotechnical systems translate into practice. We develop *Needle* based on these design guidelines and evaluate both the guidelines and resulting design features through a user study with ten moderators. In addition to validating the usability and practical viability of *Needle*, we present insights into which features proved particularly useful and how *Needle* can be improved. Finally, we present additional guidelines and considerations for designers of future tools to navigate and summarize online discussions using visualizations, especially when threaded conversations are involved. Our broader goal is to inform the design of visualization-based tools that can guide how individuals process, understand, and make sense of large-scale online discussions.

2 Related Work

In this section, we review prior literature to summarize key challenges in human moderation approaches and situate our work in the context of recent advances in computational approaches to content moderation and related visual interfaces for navigating large-scale online discussions.

2.1 Challenges in Human Moderation

Online harassment, toxicity, and other forms of harmful content present serious risks to user well-being and community health [41, 64], making content moderation a critical function for digital platforms [32]. As highlighted by prior research, human moderation remains the primary approach for managing harmful content across these platforms [25, 82, 85]. But the increasing amounts of manual labor performed by human moderators has the potential to result in severe cognitive and emotional strain [71]. Frequent exposure to toxic content and manually reviewing large volumes of user-generated content has been linked to detrimental impact on moderators, including negative outcomes like moderator burnout, stress, and "secondary trauma" [46, 80, 86]. These challenges are especially pronounced in volunteer-driven communities [50] like Reddit [61], Discord [46], and Twitch [8], where only a few moderators need to perform thousands of actions every day [35, 86].

The scale of user-generated content presents an additional challenge [26, 44, 63], with major platforms such as TikTok receiving millions of new posts daily [43]. This volume renders purely manual review impossible and has led to an increasing reliance on computational approaches to assist human moderation.

2.2 Computational Approaches and Tools to Assist Content Moderators

To scale moderation efforts, platforms and researchers have increasingly developed automated tools and algorithmic approaches to detect norm violations and assist moderator decision-making [28, 41, 65, 75]. These computational approaches generally fall into three categories: text-based classifiers and filters [44, 76], pattern and behavior analysis [15, 22, 73, 74, 84, 88, 89], and the use of large and small language models [30, 52, 90]. Text-based classifiers and filters are used to identify formatting violations and filter content based on lists of predefined keywords [44, 76]. For example, Reddit employs a built-in rule-based tool called AutoModerator (AutoMod).[1] Structural approaches analyze engagement patterns. Prior work indicates that toxic conversations and patterns of how misinformation cascades exhibit distinct structural patterns [74, 84]. Prior work demonstrates that identifying deviations from predicted engagement levels [88] and analyzing the topology of discussion threads [89] can help characterize and understand anomalies in conversations, in addition to pure text-based metrics. Furthermore, these structures capture situational dynamics, such as the influence of "social catalysts" (e.g., users who ignite conversations) [73] or trolling behaviors due to surrounding discussions [15].

[1]https://www.reddit.com/r/reddit.com/wiki/automoderator/

Yet these systems remain limited by specific trade-offs. Structural signals often require a conversation to mature before patterns emerge [74], limiting their utility for immediate, real-time detection. Text-based filters often fail in their ability to interpret nuance, handle linguistic subtleties like sarcasm or coded language, and understand community-specific norms [14, 28, 75]. For example, fully automated filters may mislabel heated but productive debate as toxic, or overlook veiled harassment masked in polite phrasing [94]. As a result, automated moderation systems continue to struggle with context-dependent nuance [13] and cultural variations [28]—even with recent approaches that rely on large language models [52].

These limitations have prompted calls for human-in-the-loop systems that blend computational efficiency with human discernment [12, 16, 37, 40]. Such systems do not aim to replace human moderators but to enhance their capabilities by surfacing high-risk content for review, visualizing user activity patterns, or highlighting contextual signals [16, 75]. For example, ConvEx [16] utilizes visualizations and AI to guide moderator attention to toxic messages on Discord, and Venire [53] helps Reddit moderation teams identify reports where moderators are likely to have disagreements. Importantly, research shows that moderators value tools that increase transparency and interpretability, and facilitate informed and community-aligned decision-making [14].

Extending this line of work, we aim to develop a system that supports human-in-the-loop moderation and incorporates computational techniques to assist, rather than replace, human judgment.

2.3 Visual Interfaces for Navigating Large-Scale Conversations

As platforms like Reddit employ threaded discussions to structure and organize content, these threaded structures introduce complexity as conversations branch into multiple pathways and foster increased social reciprocity between users [2, 91]. This branching architecture, although beneficial for organization, generates long, intricate conversation networks that impose substantial cognitive demands on moderators attempting to maintain contextual understanding across discussion threads [2, 48]. Consequently, moderators face increasing difficulty in synthesizing the necessary context for informed decision-making as conversations expand in both breadth and depth.

Researchers in HCI have explored visualization methods to address these challenges in making sense and navigating through large online discussions. Early visualization efforts like Thread Arcs provided compact graphical representations of both chronological sequences and hierarchical relationships, improving users' ability to grasp thread structures quickly [49]. More recent visualization approaches have emphasized multi-scale and interactive timelines. For instance, DisVis employed timeline and post activity graphs combined with detailed summary statistics, allowing users to rapidly identify periods of intense activity or topic shifts [66].

Similarly, analytic systems like ConVis [38] and ConvEx [16] incorporated linguistic analysis into visual interfaces, helping users quickly pinpoint emerging trends and contentious issues within expansive discussions. AudienceView leverages NLP and LLMs to analyze and visualize themes, sentiment, and topic clusters from hundreds of thousands of YouTube comments to support journalistic sense-making [6]. Additionally, Liao et al. [62] uses both concept maps and comment tree maps to help novice learners navigate large video collections through structured content and social discussion themes. Overall, these visual interfaces have consistently demonstrated their ability to reduce cognitive load and the potential to enhance moderator's experience by clearly presenting complex conversational structures and temporal dynamics.

In this paper, we examine how visualizations can be used to easily navigate and summarize online discussions involving threaded conversations, specifically focusing on Reddit moderators.

3 Design Goals

Next, we present the three primary design goals that guided the development of our new visual analytics system to enhance moderation on Reddit, a threaded discussion platform. These design goals were derived from a comprehensive review of prior literature on content moderation and the specific needs of moderators working with large-scale online discussions on platforms like Reddit.

3.1 Supporting Multi-level Summarizations through Visual Aids

Research on content moderation consistently shows that moderators struggle to maintain awareness across multiple discussions simultaneously [46, 86]. The sheer volume and complexity of online discussions often overwhelm a moderator's capacity to efficiently identify areas requiring intervention, leading to delayed interventions on problematic content and moderator burnout [86]. Previous moderation tools have typically focused on content flagging and reporting mechanisms, which provide limited contextual information. Visualization approaches like those explored by Choi et al. [16] have shown promise in reducing cognitive load through visual summarization, but these lack the multi-level perspective needed for complex threaded discussions.

Information visualization theory suggests that effective overviews combined with details-on-demand can significantly improve situation awareness for complex monitoring tasks [19]. We apply this in the context of online moderation, supporting moderator awareness by surfacing trends in recent activity through multi-level visual summaries.

Design Goal 1 (G1): *Needle* should provide multi-level visual summaries (subreddit and thread levels) and surface trends in recent activity to support efficient monitoring and improve moderator awareness.

3.2 Directing Attention and Providing Actionable Insights

With limited time and cognitive resources, moderators benefit from tools that help prioritize their attention and provide contextual information for decision-making [82]. However, existing approaches fail to provide sufficient support for moderators. While community reports and flagging systems provide the primary mechanism for helping moderators prioritize their attention, they rely on community members to identify and report problematic content [21] and flagged content appears in a separate interface (e.g., Reddit's Mod Queue). This approach can miss subtler issues or novel forms of

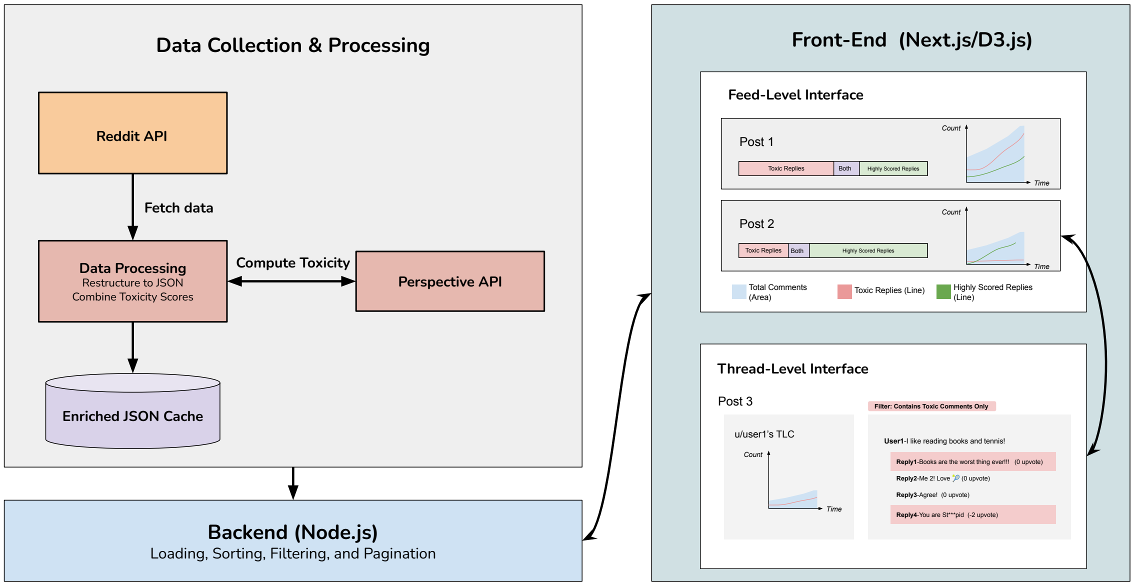

Figure 3: *Needle*'s System Architecture. Data from Reddit is processed and augmented with toxicity scores, stored as a JSON cache, and served by a Node.js backend to a Next.js/D3.js frontend; front-end is powered by two pages: Feed-level and Thread-level interfaces.

policy violations that community members do not recognize or report, as well as showing flagged content in an isolated queue may limit moderators' understanding of the conversations trends [4]. Previous approaches have also explored automated content detection using machine learning [28], but these often lack contextual awareness and have high false positive rates when used alone.

However, research on attention guidance in visual analytics suggests that combining algorithmic signals with community indicators and temporal patterns can more effectively guide moderator attention [16]. We follow this approach to support informed decision-making by directing moderator attention to key discussion and comments.

Design Goal 2 (G2): *Needle* should direct moderators' attention to key discussions and comments using metrics and visual cues that support informed decision-making.

3.3 Adapting to Diverse Community Needs

Online communities vary widely in their norms, sizes, and content policies [80]. One-size-fits-all approaches to moderation fail because what constitutes problematic content in one community may be acceptable or even encouraged in another [44]. However, many current moderation systems typically offered limited customization options, forcing communities to adapt to tools rather than adapting tools to communities [44].

Moderators need tools that can adapt to these differences without requiring technical expertise to configure them. Additionally, discussion sizes range from small exchanges to massive threads with thousands of comments, requiring scalable approaches to visualization and analysis. Research on end-user customization suggests that providing moderators with accessible controls can improve adoption and usages [16, 32, 51]. Visualization research also indicates that adaptive displays that respond to data characteristics can better maintain usability across varying scales of information [79]. We thus aim to provide customization options that allow moderators to adapt the tool to the needs and characteristics of their specific communities.

Design Goal 3 (G3): *Needle* should adapt to different community needs and discussion sizes through customizable settings.

4 *Needle*: A Visualization Tool to Navigate Threaded Discussions

In this paper, we present *Needle*, a sociotechnical system designed to help Reddit moderators navigate, interpret, and act on complex threaded discussions.[2] *Needle*'s development includes a data processing pipeline and two interfaces: a feed-level overview (Figure 4) that summarizes activity and toxicity across posts, and a thread-level drill-down (Figure 5) that supports detailed inspection of individual threads. In this section, we describe the system architecture, interfaces, and visualizations, as well as show how specific features within them support each specific design goal (Section 3): multi-level summarization (G1), attention guidance (G2), and community customization (G3).

4.1 System Architecture

Needle is implemented as a client-side web application using Next.js for the front end, D3.js for interactive visualization, and Node.js for the server-side API layer that serves preprocessed data. In this

[2]Code is avaliable in the following repository: https://github.com/scuba-illinois/Needle

section, we will walk through the details of implementing *Needle*. The complete system architecture is shown in Figure 3.

4.1.1 Data Collection & Processing. We retrieved posts and comments from the Reddit API. For each post and comment, we extracted metadata including `id`, `author`, `text`, `timestamp`, and `score`. Score, obtained directly from the Reddit API, is roughly the net total of upvotes minus downvotes displayed next to the voting arrows. It is the main mechanism for users to affect comments' visibility in Reddit [39]. Comments were recursively expanded to preserve the threaded structure. This structure allows the frontend to preserve post → top-level comment → reply hierarchies in the threaded nature.

In addition to data from the Reddit API, we also leveraged the Perspective API [60] to assess toxicity levels in comments—specifically, we used the TOXICITY model.[3] This tool provides a continuous score indicating the likelihood that a comment may be perceived as toxic, ranged from 0 to 1, as 1 being the most toxic. The Perspective API is widely used in HCI and NLP research [16, 69, 87, 95] and serves as a validated baseline for toxicity detection. It provides a standardized and interpretable score, supports direct comparability with prior moderation studies, and avoids the computational and ethical challenges of training custom models that solely rely on removed content.

4.1.2 Backend. The backend is implemented in Node.js and acts as a lightweight server layer between the static JSON data cache and the frontend. Its primary responsibilities are to load the cached dataset, apply client-side requests for sorting, filtering, and pagination, and return results efficiently.

Unlike systems that recompute toxicity or restructure comments on demand, our backend is deliberately minimal: it only serves preprocessed data to ensure low latency, reproducibility, and robustness under study conditions. This design allows us to control for variability in the Reddit API (e.g., rate limits, changing comment states) while still supporting interactive exploration on the frontend.

4.1.3 Frontend. The frontend is implemented in Next.js with D3.js for interactive visualizations. Its role is to render preprocessed data served by the backend. The next section provides a walkthrough of the front-end interfaces and features.

4.2 Interfaces and Features

Needle provides two levels of interfaces aligned with Reddit's existing structure: a feed-level overview (Figure 4) and a thread-level drill-down (Figure 5). Below we introduce these two interfaces and our design choices, as well as alignment to our design goals.

Metric Selection Rationale. In *Needle*, we use toxicity, score, and activity as the key metrics to visualize and guide moderators' attention. Toxicity captures hostile or harassing language (e.g., insults and threats), which is strongly correlated with removals and community sanctions [5, 18, 83]. We prioritized toxicity specifically because it addresses the challenge of scale: in high-volume discussions where exhaustive reading is impossible, toxicity scores help surface problematic content that community reports often miss. Score (upvotes minus downvotes, voting trends) reflects community sentiment on whether a comment is norm-aligned or not, and provides an indirect signal of removal likelihood [31, 33, 58, 83]. Activity measures overall thread engagement (e.g., the total number of comments comes in), providing essential context for interpreting the relative frequency of toxic or highly scored posts within each discussion.

These metrics serve as interpretative signals rather than absolute determinations. Recognizing that acceptable thresholds for toxicity can vary significantly across communities, *Needle* relies on the moderator's judgment for final adjudication. To facilitate this calibration, the interface includes histograms that visualize the distribution of these scores (Fig. 4B), allowing moderators to quickly establish a baseline and discern what constitutes an "unusual" or "noteworthy" deviation in their specific community context. To contextualize these metrics, we include Table 4 in Appendix A that maps specific score and toxicity values to paraphrased example comments from our user studies.

Feed-level interface. The feed-level page (Figure 4) provides an overview of the subreddit and each posts, similar to the main subreddit page in Reddit. It summarizes subreddit-wide activity and toxicity at both global and post levels. Global histograms visualize overall distributions of toxicity and scores, while post-level bar chats and temporal activity charts show patterns of toxicity and score within individual posts.

Moderators may begin with the control panel (Figure 4-A), adjusting thresholds and zoom levels to see how distributions shift across the subreddit; thresholds instantly propagate to post-level views. Hover tooltips clarify metrics like "toxicity" and "score." Posts are then browsed (Figure 4-D), by default sorted by recency, but sortable by other criteria, including highest score, highest toxicity, and highest activity. These posts' visualizations can guide attention, for example, a post with a sudden toxic spike becomes more salient for further inspection. Once familiar with *Needle*, moderators can leverage additional controls, such as customizing the time range (Figure 4-a3) or toggling specific metrics on and off to refine their view (Figure 4-a4).

Thread-level interface. The thread-level interface (Figure 5) enables detailed examination of individual discussions' comments. This interface uses the temporal visualization approach from the feed-level view, displaying activity patterns within each top-level comment (TLC) thread. Beyond the standard customization controls in the previous page, this page incorporates new functionality including comment highlighting and filtering capabilities, and integrated moderation actions including content approval or deletion to mirror the workflow of Reddit's existing Mod Queue interface.

Landing on the page, moderators first see the full set of comments (Figure 5-G) as they occupied the largest area on screen. Toxicity highlights are applied by default (Figure 5-g1). Moderators can then refine the view with filters (Figure 5-e2) such as showing only toxic or high-scoring comments, or change highlight rules (Figure 5-e1). Comment visualizations (Figure 5-F) provide temporal overviews of each TLC (Figure 5-f4), while hover effects (Figure 5-f2) and "Jump to Comment" (Figure 5-f3) link the visualizations directly to their

[3]The Perspective TOXICITY model's is a machine-learning classifier trained on online comments. Its architecture consists of multilingual BERT-based models distilled into language-specific convolutional neural networks [27].

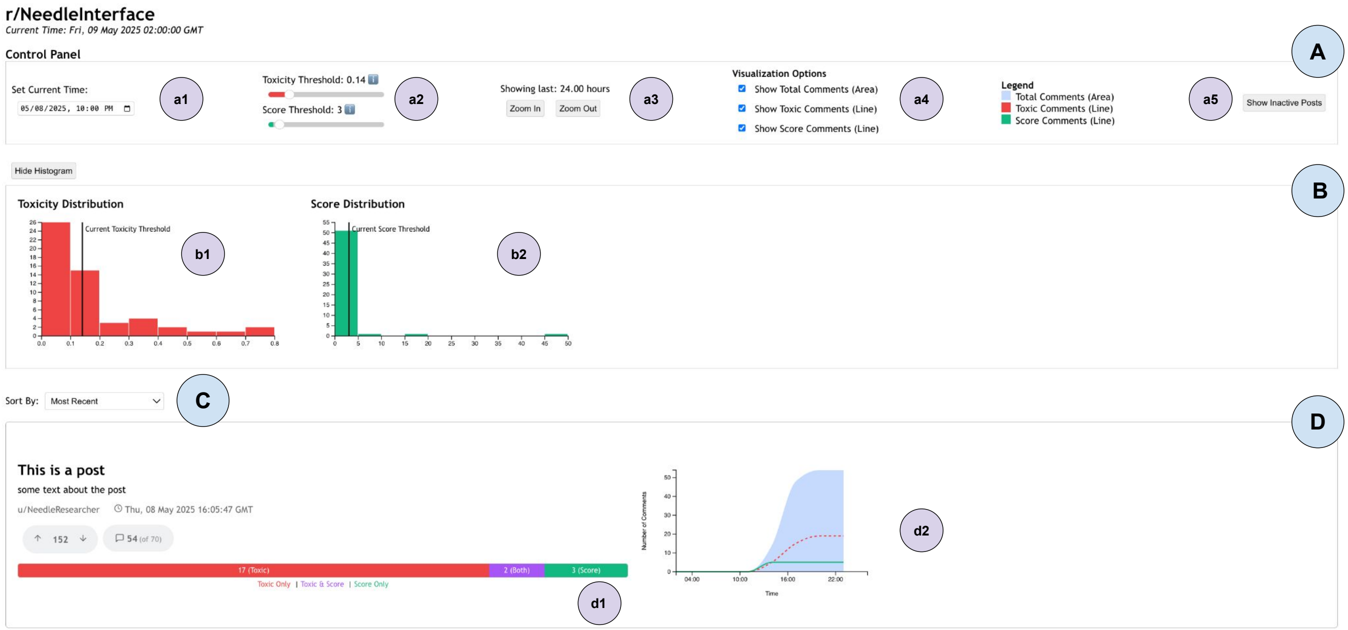

Figure 4: The feed-level interface of *Needle*, which presents subreddit-wide comment activity and toxicity. (A) Control panel for setting time (a1), toxicity and score threshold (a2), zoom level for time range (a3), line toggles with legend (a4), and toggle show inactive posts (a5). (B) Global histograms for toxicity (b1) and score distributions (b2). (C) Sorting dropdown to organize posts by criteria. (D) Post summaries, with a segmented comment breakdown bar (d1), and a temporal activity chart (d2).

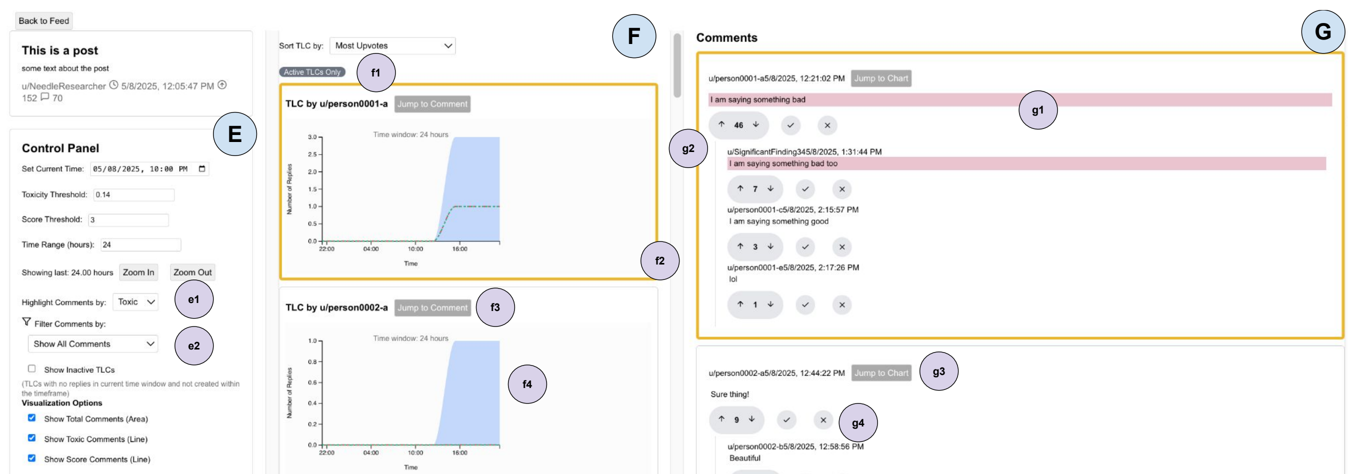

Figure 5: The thread-level interface of *Needle*, which provides detailed visualization of individual discussion threads. (E) Control panel with the same customization options as the feed-level interface, adding two new features of applying highlights to comments (e1) and filter comments (e2). (F) Comment visualization features including: active TLCs only indicator (f1), hover effects to indicate linkage between visualizations to comments (f2), jump to Comment button for quick navigation (f3), and temporal activity chart for the selected TLC (f4). (G) Actual comments: toxic comment highlighting (g1), hover effects (g2), navigation controls to move between charts and comments (g3), and moderation buttons (g4).

source comments. Finally, moderation buttons (Figure 5-g4) allow immediate actions such as removing content.

Summary of Features. Table 1 summarizes the key features of *Needle*, grouped according to our three design goals: multi-level summarizations (G1), attention direction and actionable insights (G2), and community customization (G3). Features are listed with

Feature	Location	Description	Use Cases
Multi-level Summarizations (G1)			
Toxicity & score histograms	Figure 4-B	Visualizes the distribution of comment toxicity and score levels across all posts	Assess overall subreddit health and establish appropriate filtering thresholds
Post temporal visualizations	Figure 4-d2	Charts how discussion activity evolves within individual posts over time	Analyze post engagement patterns and identify toxicity, score, or activity peaks
TLC temporal visualizations	Figure 5-f4	Charts how discussion activity evolves within individual top-level comment threads over time	Monitor how specific conversation threads develop and escalate
Attention Direction and Actionable Insights (G2)			
Sort options	Figure 4-C	Reorders posts and top-level comments by various criteria	Prioritize content based on toxicity levels, activity, recency, or scores
Segmented breakdown bars	Figure 4-d1	Color-coded bars that visually segment comments by toxicity and score numbers	Identify problematic or highly-rated content within posts
Inactive content toggle	Figure 4-a5	Shows or hides posts and TLCs with minimal recent activity	Reduce interface clutter by focusing on active discussions
Comment highlighting	Figure 5-e1	Visually emphasizes comments that exceed toxicity or score thresholds	Direct moderator attention to comments requiring immediate review
Comment filtering	Figure 5-e2	Displays only comments meeting specific toxicity or score thresholds	Concentrate moderation efforts on the most critical content
Community Customization (G3)			
Threshold configuration	Figure 4-a2	Allows adjustment of toxicity and score sensitivity levels	Adapt parameters to match different community standards
Time point navigation	Figure 4-a1	Enables quick jumps to specific timestamps in the activity feed	Review historical activity during specific time periods
Temporal zoom control	Figure 4-a3	Adjusts the time window scope from 8 minutes to 24 hours	Explore community activities across different temporal scales

Table 1: Feature overview of *Needle* including interface locations, descriptions, and use cases.

their location, function, and representative use cases. This overview illustrates how visual summaries, attention cues, and configurable settings work together to support effective and efficient moderation.

4.3 Visualizations and Visual Cues

While the table above inventories *Needle*'s features, here we explain the rationale for our visualization choices for the four types of visualizations we used: temporal visualizations, comment highlighting, segmented breakdown bars, and histograms. Our goal was not simply to display raw data, but to transform the three core metrics—score, toxicity, and activity—into representations that moderators could quickly interpret and act upon.

Temporal visualizations (Figure 4-d2 and Figure 5-f4) are the key visualizations in *Needle* representing multi-level summarization (G1). They were included because activity and pacing are key signals of escalation: bursts of comments or sudden spikes in toxicity often mark moments when intervention is most needed. Raw timestamps are difficult to parse in threaded discussions, but lines and areas in charts allow moderators to perceive these dynamics at a glance.

Our approach to temporal visualization requires a clear definition of recency within *Needle*. Rather than adopting a single fixed threshold, we define recency as a flexible, moderator-controlled window of activity anchored to a chosen point in time. The "current time" setting (Figure 4-a1) allows moderators to anchor their analysis at any moment. This anchor pairs with adjustable time range (Figure 4-a3) that scale from minutes to hours, allowing communities with different pacing to calibrate what counts as "recent" and to focus on that time frame.

To balance overview with detail, *Needle* visualizes discussions at the feed and Top-Level Comment (TLC) levels. We intentionally do not create separate visualizations for nested replies beneath each TLC, as doing so would introduce excessive visual complexity and cognitive overhead. While the visualization aggregates branches, the comment panel preserves the full hierarchy (Figure 4G). To bridge these views, comment highlighting (Figure 4-g1, that highlights toxicity, score, or both) pinpoints problematic content at any depth, enabling moderators to navigate complex threads without parsing multiple visual layers. Highlighting also draws the eye directly to outliers. We see these design choices as essential to reducing cognitive load and supporting efficient moderation at scale.

Segmented breakdown bars (Figure 4-d1) is another important visualization added to to create "at-a-glance" profiles of posts. These bars compress all comments into a simple signature, making it easy to notice when a post contains unusually high proportions of toxic or highly scored contributions. This supports triaging: moderators can scan many posts quickly without reading individual comments.

In addition to the two key visualizations, we also used histograms (Figure 4-b1 & b2) to provide necessary context for interpreting toxicity and score as a whole for a subreddit. Instead of providing post-level or thread-level supports, histograms provide a global view of how the subreddit is reacting to the toxicity and score metrics.

Participant	Subreddit Size (approx.)	Years Moderating	Subreddit Type
P1	2300k	5	Academic
P2	400k	7	Sports
P3	700k	2	TV shows
P4	400k	9	Games
P5	600k	13	Games
P6	800k	8	Games
P7	2300k	9	Anime
P8	5900k	8	DIY
P9	4400k	3	Sports
P10	600k	1	Anime

Table 2: Overview of study participants. Subreddit sizes rounded to nearest 100k; years moderating rounded to nearest year.

A toxicity score of 0.7 may be problematic in one community but normal in another; subreddit-wide distributions help moderators calibrate thresholds to their own norms. Without this comparative view, thresholds risk being arbitrary or biased by isolated cases.

5 User Study

Next, we conduct a user study to examine how well the design guidelines synthesized from prior literature translate into practice. Specifically, we ran an online user study with moderators from various subreddits by simulating a realistic workflow using *Needle*. This allows to examine how *Needle* could support Reddit moderators navigating threaded discussions and how it fits into their existing workflows. In addition to evaluating the usability and practical viability of both the guidelines and resulting design features in *Needle*, we identify specific features and design elements that are most and least helpful to moderators.

5.1 Participants

We recruited 10 moderators who are current Reddit moderators located in the United States, varied in community sizes and topics. On average, moderators had 6.5 years of experience (SD = 3.72), and the mean subreddit community size was 1.84M members (SD = 1.92M). Moderators were recruited via mod mail from large, active subreddits (top 1500 popular subreddits[4]), and they were invited to fill out an interest form, including their moderation experience and communities they were moderating for. Participants were required to be active moderators on Reddit. We confirmed their eligibility in the interest form by verifying their moderator status via their provided username.

5.2 Procedure

During the screening survey, we asked moderators to indicate which subreddit(s) they currently moderate and a specific subreddit that they would like to test *Needle* with. For the study sessions, the researcher collected the subreddit's data spanning two days, including posts either created or replied to within that period, to use as the database for the user study via the data collection and processing procedures outlined in Section 4.1.1.

In the study, the researcher first introduced *Needle*'s features. Next, moderators engaged in two primary tasks using *Needle* to simulate their content moderation workflow (§5.3), followed by a semi-structured interview that asked them to comment on the features and design of *Needle*, as well as how *Needle* can be integrated into their current workflow. Moderators were encouraged to think aloud while completing the tasks, verbalizing their decisions, questions, and reactions when they were interacting with *Needle*.

All user study sessions were conducted via Zoom in English. Each session lasted approximately one hour, and each participant was compensated with a $20 Amazon gift card. Each session was audio-recorded, transcribed, and anonymized. We also log their interaction data with *Needle*.

Ethical Considerations. To minimize disruption to online communities, we simulated the moderation environment using scraped data collected a day prior to each user study session. This approach ensured participants worked with recent, familiar content while avoiding direct intervention in live communities. Participant recruitment was conducted through modmail, and we carefully excluded communities whose rules explicitly prohibited unsolicited messages or research recruitment. We obtained oral informed consent from all participants and implemented measures to protect anonymity by avoiding collecting real names or personally identifiable information. All potentially identifying information were removed. All research procedures were reviewed and approved by our Institutional Review Board.

5.3 Tasks

To mimic a typical moderation process and to encourage the exploration of *Needle*'s features, we designed the following tasks for moderators. During the study, moderators were encouraged to think aloud and describe how each feature influenced their decision-making process. If moderators did not explore certain features, researchers would reintroduce the features to ensure comprehensive evaluation.

5.3.1 Locating Threads. In the first task, moderators were asked to locate discussion threads of interest within their subreddit. Using *Needle*'s overview features, they explored the feed to identify posts that might warrant their attention.

5.3.2 Moderating Comments. In the second task, moderators clicked into one of the identified posts to review its recent comment threads. For each comment, they could decide whether to remove it if it clearly violated subreddit rules, or report it if it was ambiguous.

5.3.3 Interview. Following the two tasks, each participant took part in a semi-structured interview. The interview explored their overall impressions of *Needle*, how it compared to their existing

[4]Retrieved from Reddit API reddit.subreddits.popular

moderation workflow, and how specific features supported or hindered their decision-making. Participants also provided feedback on desired improvements and reflected on *Needle*'s potential to reduce cognitive load during moderation.

5.4 Analysis

We analyzed the interview data using thematic analysis. Given our focus on system evaluation, we deductively mapped participant feedback against our three initial Design Goals (Section 3) to validate *Needle*'s objectives. Simultaneously, we remained open to inductive themes from the data outside of these pre-defined categories. During the coding process, two researchers first collaboratively coded three transcripts, discussing their interpretations and identifying preliminary themes, such as challenges moderators face, their needs, and how system features could address those needs. These discussions produced an initial, flexible set of codes and themes. One researcher then coded the remaining transcripts, continuing to consult with the second researcher when uncertainties arose.

6 Results

From our study, we found that *Needle* successfully transformed design guidelines into practices and can be integrated to support moderation workflows in several key ways. All participants agreed that *Needle* has the potential to aid in moderation tasks or in their own moderation processes. Reactions were predominantly positive, though some concerns emerged regarding the toxicity detection mismatches with community-specific norms. Despite these considerations, five of ten moderators had outlined ways that *Needle* would complement their existing toolkit effectively (Section 6.5).

We organize our findings to reflect both *Needle*'s intended design objectives and the broader sociotechnical context of moderation. We begin with moderators' overall reactions to *Needle*'s interface and features (Section 6.1), which provides important context for understanding how they engaged with specific design elements. We then examine how *Needle* addressed each of its three design goals (Sections 6.2–6.4), evaluating whether *Needle* achieved its intended objectives. Finally, we present themes that emerged inductively regarding *Needle*'s practical viability in real-world contexts: how moderators envision integrating *Needle* into their existing workflows (Section 6.5) and remaining challenges that need to be addressed for practical deployment (Section 6.6).

6.1 Moderator Impressions of Interfaces and Features in *Needle*

Overall, moderators expressed that the interface is quick and easy to use. In the feed-level interface, eight out of ten moderators expressed that they found the features involving toxicity values to be useful. They adjusted the toxicity threshold (Figure 4-a2) more often than the score threshold, with an average 270.8 adjustments to the toxicity threshold per user study session compared to 35.6 for the score threshold. The histograms (Figure 4-B) were found to be generally helpful for making sense of the toxicity threshold as well as for understanding the high-level community health. However, one moderator did report that the histogram was difficult to understand and that they were not familiar with this type of visualization. Seven out of ten moderators found the time setting and time zooming (zoom in/out timespan) not necessary, as they mostly moderate relatively frequently on the subreddits and do not need to adjust time range.

Participants also found the sorting effective, and many would sort by the highest toxicity comments (Figure 4-C). In terms of each post and the accompanying visualizations, moderators found it easy to understand. For most moderators (7 out of 10), the temporal graphs (Figure 4-d2) were useful for highlighting anomalies in threads and understanding trends; however, many of them (5 out of 10) also note that they used the bar (Figure 4-d1) to make comparisons between posts, therefore guiding their attention to the most toxic posts.

At the feed-level, moderators generally began with the system's default toxicity threshold (0.2) and sorted by the highest-toxicity comments (6 out of 10), as three of them mentioned sorting is a common functionality they used in Reddit's moderation tool (Mod Queue). To gain an early sense of community health and to orient themselves to the range of toxicity values, they often glanced at the visualizations for the first few surfaced posts. From there, moderators drew on the visual summaries in ways that fit their own routines: the histogram helped them interpret the overall distribution of toxicity across the subreddit, while the temporal activity charts could show toxicity spikes in problematic behavior occurred as they adjusted the threshold. Because their first task was to identify a thread to examine in depth, all moderators adjusted the threshold or sort until the charts revealed a post containing at least some toxic comments.

In the thread-level interface, moderators generally found the interface easy to navigate without confusion. The highlighting and filtering features (Figure 5-e1&e2) were the most appreciated features identified by moderators, with eight out of ten reporting them as useful. Moderators also expressed that the the approve and delete buttons were also intuitive and helpful as they showed their moderation actions on the posts. However, due to the low number of replies most top-level comments receive, the visualizations for TLCs (Figure 5-F) were not as helpful as the visualizations for threads, as 6 out of 10 moderators noted, and the "jumping to" feature (Figure 5-f3&g3) was not often used.

At the thread level, moderators were typically drawn directly to the highlighted comments or filtered to view only toxic comments. They then began examining these comments more closely, adjusting the toxicity threshold as needed once they could see how the toxicity values mapped onto the actual content.

6.2 Design Goal 1: Transforming Textual Content into Intuitive Visual Patterns

We found that our multi-level visual summaries were largely effective in surfacing patterns in recent activity and supporting efficient moderator awareness.

Seven out of ten moderators reported that thread-level visualizations were effective, with three moderators specifically highlighting the effectiveness of visualizing individual TLCs. Moderators frequently expressed that our visualization approach offered superior functionality compared to existing Reddit moderation tools and has the potential to fit right into their preexisting workflows. One moderator specifically stated how *Needle*'s feed-level overview aligns with their typical practices:

> *"This is interesting, like, [visualizations] for threads is much more useful than what we have...in Mod Queue or toolbox, you can't see that[discussions/replies] at all. So I've gone into the thread to go make sure I get everything that's in there."* (P1)

The visualization features that most effectively supported awareness and detection of problematic content were the thread-level toxicity graphs and the temporal activity visualizations. These tools transformed moderation from sequential text review to pattern recognition, allowing moderators to quickly identify anomalies without manually reviewing each comment one by one. As P2 noted:

> *"If I see this graph and it starts not looking like this[toxicity line stays low], then I should do something about it. I think there's value in that, and it makes it very easy for me."*

This capability proved especially valuable for identifying contentious discussions:

> *"If I could just look at that visualization of what's happening on the subreddit, then that could help me choose which threads to look at and see if there's too much fighting going on."* (P7)

These visual patterns may substantially alter traditional moderation workflows. Rather than processing comments one-by-one, moderators adopted a more strategic approach: first identifying potential hotspots through our visualizations, then investigating those specific areas. This triage-based workflow can support more efficient attention allocation by directing moderator attention to areas of greatest concern. Additionally, the ability to visually scan for patterns enabled more proactive moderation, with moderators identifying potentially problematic threads before they accumulated multiple reports.

Within individual TLC visualization, moderators reported limited utility due to low amount of activity within the replies of individual TLCs. As P1 noted:

> *"I think the visualizations work pretty well for what their intended purpose is like, even with just the small number of comments you scraped like they made sense... the visualizations work. We just don't give it enough to visualize."*

This suggests that while the visualizations function as intended even with modest data, visualization effectiveness scales with content volume, being most valuable at the thread level where patterns become more meaningful.

6.3 Design Goal 2: Directing Attention Through Metrics and Visual Cues

We found that *Needle*'s implementation of toxicity metrics and highlighting features was successful in supporting informed decision making for moderators. These features effectively addressed the moderators' need for attention-directing mechanisms that could help them navigate large volumes of content efficiently.

6.3.1 Enhanced Decision Support Through Toxicity Metrics. Moderators found *Needle*'s toxicity metrics were notably more effective than Reddit's native tools. As P6 remarked:

> *"[Needle] seems to work a little bit more intelligently than the tools that Reddit has initially...this is the sort of tool that can help you find that[unreported contents] for any sort of subreddits."*

Specifically, these toxicity metrics helped moderators pinpoint problematic content that might otherwise go unnoticed, with P1 noting that it was more helpful than standard Reddit tools since the Mod Queue sometimes missed toxic comments that were unreported.

The value of toxicity metrics was particularly evident for high-volume subreddits where manual review becomes impractical. P5 explained the challenge of scale in very large subreddits where moderators need to perform thousands of actions per day just to keep pace with incoming content. While our study did not measure task completion time quantitatively, four moderators mentioned that in these environments, toxicity detection significantly reduced cognitive load to identify toxicity. Three moderators expressed interest in customizable toxicity thresholds that would allow them to receive alerts when unusual patterns emerge. P9 suggested they would set high thresholds to identify threads receiving an abnormal number of toxic comments, allowing for more targeted intervention.

Another common usage of toxicity metrics was as a reference point to aid in their decision-making process. For example, P10 said that:

> *"[Toxicity is like a] second opinion on stuff if I don't know I should remove it or not."*

6.3.2 Improving Content Processing Through Visual Highlighting. The highlighting feature proved particularly effective at directing moderator attention, with eight out of ten moderators acknowledging and appreciating this function primarily for its efficiency benefits. P1 noted that highlighting would help them process threaded sections much faster. And as P5 noted, quick processing is a fundamental requirement for effective moderation. The color-coded system provided immediate visual cues that guide moderator attention, with P3 explaining that their eye naturally gets drawn to content highlighted in red, which guided their attention similarly to how their existing filters flag content.

Visual highlighting also appeared to improve moderation thoroughness, with three moderators mentioning how it helped them catch problematic content they might otherwise have missed when reviewing comments across their entire subreddit. By combining toxicity metrics with visual highlights, moderators were able to implement a more strategic, targeted approach to moderation.

Although our highlight function identifies comments based on multiple metrics—including high scores, toxicity levels, and combined metrics (toxicity and score)—six moderators specifically noted that score metric fall outside their typical concerns in moderation process. Therefore, score is not a useful metric for them. Thus, the combination of toxicity and visual highlighting emerged as particularly effective. Together, they were described by moderators as reducing cognitive effort and enabling more systematic navigation of large comment volumes.

6.4 Design Goal 3: Adapting to Community Needs Through Customizable Settings

We found that our implementation of adjustable thresholds gave moderators meaningful ways to customize *Needle* to their needs. Our findings support the importance of this flexibility, particularly in how moderators interpret and apply toxicity thresholds across diverse community contexts.

Moderators actively experimented with toxicity thresholds to calibrate the tool to their community's specific communication norms. They demonstrated a process of exploratory calibration to find appropriate settings. P4 described needing to adjust thresholds repeatedly to determine the right sensitivity level for their community. By testing content against different thresholds, moderators gained additional context for judgment calls. Moderators also expressed interest in understanding the underlying mechanics of toxicity detection to better calibrate thresholds. For example, P1 noted curiosity about what specific text patterns triggered higher toxicity values.

6.5 Integration with Existing Moderation Workflows

We investigated how *Needle* can integrate with existing moderation practices as it is essential for evaluating its practical utility. Moderators reported using a variety of native Reddit tools including AutoMod, Mod Queue, and user history cards, while operating across both mobile and desktop platforms. Three moderators also mentioned utilizing external communication tools such as Discord and Slack for moderator discussions and coordination. Therefore, due to the diversity in moderation tools, there is no "one-size-fits-all" solution that *Needle* can fit into every tool.

We found that the visualization graphs could integrate into existing workflows in meaningful ways by helping moderators identify emerging problematic topics and add relevant keywords from these discussions to their filters. For example, P2 explained that:

> *"If we see a big spike in that graph, we're just like, okay, what is the topic that's being discussed? Let's go ahead and add, you know, the words related to this topic to our auto moderator."*

This demonstrates how *Needle*'s visualizations can directly enhance existing AutoMod configurations by helping moderators proactively identify conversation trends and trajectories that require additional monitoring.

Toxicity metrics were also reported to be useful to integrate into existing workflows. Five moderators mentioned performing moderation activities directly from the Reddit feed, using sorting functions to prioritize content review. This behavior aligns well with *Needle*'s feed-level interface and enhanced sorting capabilities (with toxicity metrics), suggesting another way to integrate the tool into existing workflows.

6.6 Open Challenges and Unaddressed Moderator Needs

While *Needle* demonstrated effectiveness in addressing workflow efficiency challenges, our interviews revealed a persistent challenge that requires further development: the nuanced understanding of community-specific language and norms. Six moderators consistently emphasized how community-dependent language complicates automated toxicity detection across different communities. This highlights the need for more community- and context-sensitive models to detect norm violations [30, 47].

Moderators across gaming, sports, and fandom subreddits highlighted how domain-specific language often triggers false positives in toxicity detection. P5 (from a game's subreddit) noted that "toxic" phrases can be praise in gaming culture, while P2 (from a sport's subreddit) pointed to emotional sports talk as a source of misclassification. P3 (from a TV show's subreddit) and P6 (from a game's subreddit) emphasized the role of show-specific jargon and fictional context in defining toxicity. While *Needle*'s adjustable thresholds help, moderators like P9 stressed the need for a model that understands niche vocabularies and cultural nuance, as well as specific removal patterns in the community.

7 Discussion

In this paper, we presented a multi-tiered visual analytics approach in *Needle*, translated design principles from prior HCI research into practice, and demonstrated how *Needle* can enhance moderation workflows beyond Reddit's current tools. Through a user evaluation study with 10 Reddit moderators, we found that *Needle*'s design and features effectively addressed three core challenges in content moderation: (1) detecting anomalies via visual patterns in text, (2) surfacing problematic content through toxicity metrics and visual cues, and (3) accommodating diverse community norms through customizable thresholds.

Needle provides aggregated metrics at the subreddit level for comprehensive community health assessment, visualizes toxicity and activity patterns at the thread level to reveal conversation health without requiring exhaustive reading, and incorporates temporal visualizations to help moderators detect emerging issues. This layered approach helped transform overwhelming textual content into intuitive, scannable visualization patterns that can reduced cognitive load on Reddit moderators (G1).

We also introduced attention-directing mechanisms in *Needle* that combined multiple data sources to identify potentially problematic content (G2). By integrating toxicity analysis, community signals like voting patterns, temporal anomalies that may indicate coordinated behavior, and customizable highlighting, *Needle* helped moderators focus on areas requiring intervention. By synthesizing these signals, *Needle* supported more informed decision-making while preserving human judgment in the moderation process.

Finally, we developed a customizable moderation system that can be adapted to diverse community needs and contexts (G3). *Needle* incorporates adjustable thresholds for different metrics, visualization scaling that accommodates various discussion sizes, and customizable panels that allows moderators to tailor their view based on community-specific priorities. This flexibility ensures that the tool can serve communities with different norms, sizes, and moderation approaches without requiring technical expertise to configure.

Next, we discuss the implications of our findings, design considerations to inform future tools, and next steps for *Needle*'s development.

7.1 Tools to Navigate Threaded Discussions and Guide Moderator Attention

Needle's success highlights the effectiveness of visualization-based approaches in transforming moderation workflows from sequential text review to pattern recognition. By converting textual content into visual patterns, moderators can quickly identify anomalies and potential issues without manually examining each comment. This approach fundamentally changes how moderators navigate large volumes of content—moving from linear processing to a more strategic workflow that improves both efficiency and effectiveness, as well as helping distance moderators from immediate trauma [46, 80, 86].

While prior systems like ConvEx [16] or ConVis [38] support sense-making in linear or topic-based formats, they do not address the "branching" complexity of threaded discussions [2]. In *Needle*, particularly effective was the thread-level toxicity visualization that allowed moderators to identify problematic discussions at a glance. As P2 noted, these visualizations make it "very easy" to determine when intervention is necessary. This capability proves especially valuable for high-volume subreddits where manual review becomes impractical, addressing what P5 described as the "scale challenge" where moderators need to perform thousands of actions daily.

Additionally, *Needle* supports a shift from individual to community-level situational awareness. Unlike Reddit's current ModQueue, where comments are presented as isolated infractions stripped of conversational context [4], *Needle*'s visualizations serve as a signal directing moderator attention to specific, escalating threads. Furthermore, customizable thresholds allow moderators to explicitly calibrate these signals to their specific community norms [14, 28, 75], empowering them to intervene early in toxic cascades [14].

To illustrate the practical implications of *Needle*, Table 3 compares the workflows in *Needle* with native moderation tools including AutoMod and Mod Queue in Reddit. Reddit's ModQueue is the primary interface used by moderators, which shows a linear list of comments that are identified through AutoMod rules or flagged by users pending reviews from the moderators [4].

7.2 Design Implications

We identify several key implications for designing effective tools specifically suited for community-driven moderation:

7.2.1 Customizability and Adaptability to Community Norms. A critical theme that emerged in our study is the need to account for community-specific norms. Unlike commercial moderation that enforces uniform global policies [80], volunteer moderation often involves adapting to community rules and cultures [12]. This variability creates challenges for standardized toxicity detection systems, which often fail to account for contextual nuances in community discourse. Future moderation tools should incorporate community-sensitive models that can learn and adapt to specific subreddit cultures, rules, and communication patterns.

The need for customization goes beyond adjustable toxicity thresholds to fit into different moderation philosophies. Some communities prefer tools that support positive reinforcement in addition to content removal. Four moderators, for instance, described practices such as awarding flairs, identifying potential future moderators, and extracting high-quality comments to include into community wikis. This aligns with recent HCI work exploring the role of positive reinforcement in moderation [54] and highlights the need for future research into systems that facilitate such positive moderation actions.

7.2.2 Multi-level Agency and Overview. *Needle*'s implementation of both feed-level and thread-level interfaces demonstrates the importance of providing multiple perspectives on community content. Moderators need the ability to transition between high-level pattern recognition and detailed comment review. This multi-level approach supports different moderation strategies—from proactive identification of emerging issues [35, 75] to targeted intervention in specific discussions [16].

Effective moderation tools should provide clear visualization pathways between these levels, allowing moderators to quickly drill down from community-wide patterns to specific problematic content. This capability supports informed decision-making by providing both context and detail when needed.

7.2.3 Human-in-the-Loop Approaches. While automation can significantly improve moderation efficiency, our study reinforces the critical role of human judgment in content moderation. Our findings suggest that moderators do not view algorithmic signals as absolute truth but rather as one of several inputs to be triangulated with their own domain expertise.

Moderators valued *Needle* not as a replacement for their decision-making but as a tool that enhances their capabilities through such as "second opinion" support (P10) and strategic attention direction. Specifically, when content sits in a gray area, these "second opinions" (e.g., toxicity scores) help moderators to preform a sanity-check their own interpretations and gain confidence in borderline decisions. Additionally, by design, Needle is intended to assist moderators by drawing attention to potentially problematic areas rather than taking automated actions on their behalf. The toxicity metrics support strategic attention direction by filtering large volumes of content and highlighting areas that warrant closer human review. Three moderators also explicitly mentioned that all their moderation decisions are purely human-based to capture community norms and adherence to past moderation standards.

This finding suggests that moderation tools should be designed with a human-in-the-loop philosophy, where technology augments rather than replaces moderator judgment [12, 16, 37, 40]. The ideal approach combines algorithmic efficiency with human discernment—using technology to filter and highlight potential issues while preserving moderator agency in final decision-making. This human oversight is particularly important in the generative AI era, where automated tools and LLMs still struggle with nuance and false positives [16, 52]. Human moderators remain necessary to correct algorithmic errors, making visualization tools like *Needle* essential for scaling this human review process.

7.2.4 Enabling Proactive Moderation. A significant limitation of traditional moderation tools is their predominantly reactive nature, requiring moderators to respond to issues after they emerge. *Needle*'s visualization capabilities demonstrate potential for shifting toward more proactive moderation approaches. By visualizing

Design Goal (DG)	Native Tools	*Needle*'s Workflow	Conceptual Shift
DG1: Multi-level Visual Summarizations	Linear comment lists in Mod Queue strip conversational context across threaded branches. Moderators must manually click into threads to understand discussion flow and reply hierarchies [4].	Feed-level temporal visualizations and thread-level toxicity graphs enable pattern recognition at a glance. Segmented breakdown bars compress entire discussions into scannable signatures showing toxic/highly-scored content proportions.	Moderators transitioned from sequential text review to strategic triage, identifying hotspots before reading individual comments.
DG2: Directing Attention & Providing Actionable Insights	Moderators rely on community reports and keyword-based AutoMod filters [44]. Flagged content appears isolated as individual comments in Mod Queue without surrounding conversational context, potentially missing unreported violations.	Visualizes toxicity and voting signals in situ within the threaded conversations. Uses color-coded highlights to guide attention to problematic clusters without removing them from the surrounding dialogue. Customizable sorting prioritizes high-toxicity threads.	Moderators shifted moderation from reactive flagging to proactive conversation health monitoring.
DG3: Adapting to Diverse Community Needs	Static AutoMod keyword rules require manual configuration and apply uniformly once set [44]. While keywords can be tailored to communities, they lack nuance for context-dependent toxicity (e.g., whether gaming slang is toxic depends on conversational context, not just the words themselves) [13].	Adjustable toxicity and score thresholds with real-time histogram feedback showing community-wide distributions. Moderators calibrate sensitivity through exploratory threshold testing to match their subreddit's communication culture.	Moderators actively experimented with thresholds to accommodate domain-specific language compared to a rigid list of keywords.

Table 3: Conceptual comparison of moderation workflows supported by native Reddit tools and *Needle*, categorized by our design goals.

temporal patterns and discussion trends, moderators can monitor conversational trajectories and identify potentially problematic conversations earlier in their development [14, 56, 94]—before they escalate to require intervention.

7.3 Next steps for *Needle*

Based on our findings, we identify several promising directions for future development of *Needle*:

7.3.1 Developing Community-Specific Toxicity Models. Moderators consistently noted the limitations of generalized toxicity detection. Future work should develop subreddit-specific models trained on community norms, prior moderation actions, and unique community vocabulary. Additionally, incorporating contextual signals, such as discussion topics and user reputation, could help the system better interpret niche language and improve detection accuracy.

Designers and developers should collect details on the specifics about community rules and practices, as they vary in forms [23, 70]. For example, one moderator explicitly mentioned that they used word count to remove comments, as their subreddit required sophisticated answers citing credible sources; another moderator expressed that an image-based models would be useful to detect toxicity, as a lot of toxicity came from images in their community.

7.3.2 User-Level Information Integration. Five moderators emphasized the value of incorporating user history and behavior patterns into moderation decisions. Future versions of *Needle* could include user-level signals such as prior comment quality or moderation history directly in the interface, helping moderators determine whether a comment is part of a broader behavioral pattern. These signals could be further visualized or calculated to similar metrics or summaries currently used for comment- and thread-level insights.

7.3.3 Positive Moderation Features. While *Needle* currently focuses on identifying problematic content, moderators expressed interest in tools that would also help them identify and promote positive reinforcement [17, 57]. Future development could expand *Needle*'s capabilities to include features for recognizing desirable behavior [29], highlighting quality comments [68], and implementing positive reinforcement through awards, flairs, highlights, or adding to the subreddit's wiki or weekly digest to incentivize participation and motivate contributions [24, 54, 55].

7.3.4 Expanding Needle to Assist Reddit Users. Beyond moderation, *Needle* could also support everyday users in navigating and making sense of large-scale discussions. For example, visual summaries like activity bursts and toxicity spikes can help users obtain overviews of discussion threads. Similarly, surfacing highly active threads can guide users to the most relevant or interesting parts of a discussion. These features may improve engagement and promote more informed participation among social media users. Broadly, further research is needed to examine the influence of both textual and visual summaries on how users consume, interpret, and make sense of large-scale online discussions.

7.4 Limitations

The study presents several limitations.

First, the evaluation was conducted in a simulated environment with pre-scraped data rather than during live moderation. This choice minimized disruption to real communities and allowed us to control for variability in Reddit's API, but it also meant that we were unable to capture the pressures and dynamics of real-time moderation.

Second, *Needle* was evaluated as a standalone tool rather than integrated with moderators' existing workflows. We chose this approach to isolate and examine its design features, and we asked participants to reflect on how it might fit alongside their current tools. However, this setup leaves open questions about exactly how *Needle* would be adopted and used in practice when combined with tools such as AutoMod and ModQueue. We envision a practical path to adoption as a browser-based extension. This approach would allow visualizations to function as an additional layer over Reddit's current interface.

Third, our evaluation did not include an ablation study to isolate the individual contributions of specific system components (e.g., temporal visualizations or toxicity metrics). We chose this approach because these components are tightly coupled and intended to be used together to support moderation workflows. As a result, we do not attribute observed workflow changes to individual components in isolation. As we envision *Needle* working alongside existing moderation ecosystems, future longitudinal studies could examine how these features perform in combination or isolation with other moderation tools.

Fourth, regarding scalability and data retrieval, *Needle* currently relies on a pre-fetched JSON cache to ensure low latency for the user study. In deployment, however, exclusively using real-time Reddit API calls would face rate limits and network delays. A more scalable approach would be to register target subreddits, pre-fetch historical conversations, and run continuous background scraping, enabling a hybrid of caching and incremental loading that balances responsiveness with data completeness.

Lastly, we focused exclusively on Reddit to examine threaded discussions. While threaded discussions also appear on other online platforms (e.g., X or other forums with similar affordances and moderation ecosystems [3]), we selected Reddit as a representative testbed given its large scale and deeply layered thread structures. Further research is needed to examine how visualizations can support users in navigating threaded discussions on other platforms.

8 Conclusion

Our system, *Needle*, demonstrates how visualization-based approaches can transform content moderation workflows by converting overwhelming textual discussions into interpretable visual patterns. Our user study with 10 Reddit moderators revealed that the system's multi-level summarization capabilities, toxicity metrics, and customizable interfaces effectively reduce cognitive load while enabling proactive identification of problematic content. Although challenges still remain in automated toxicity detection across diverse community contexts, *Needle*'s human-in-the-loop design successfully augments moderator capabilities without replacing human judgment. Our system can be readily extended to incorporate additional conversational metrics and community-specific signals. We hope that our findings will inform future approaches to creating visualization-driven tools that enhance human moderation capabilities across diverse online communities.

Acknowledgments

The authors would like to thank the participants for their time and contributions and the reviewers for their thoughtful feedback. We also thank members of the Social Computing Lab (SCUBA) at the University of Illinois for their feedback and input on this work.

References

[1] Ana Aleksandric, Sayak Saha Roy, Hanani Pankaj, Gabriela Mustata Wilson, and Shirin Nilizadeh. 2024. Users' Behavioral and Emotional Response to Toxicity in Twitter Conversations. *Proceedings of the International AAAI Conference on Web and Social Media* 18, 1 (May 2024), 29–42. doi:10.1609/icwsm.v18i1.31295

[2] Pablo Aragón, Vicenç Gómez, and Andreaks Kaltenbrunner. 2017. To Thread or Not to Thread: The Impact of Conversation Threading on Online Discussion. *Proceedings of the International AAAI Conference on Web and Social Media* 11, 1 (May 2017), 12–21. doi:10.1609/icwsm.v11i1.14880

[3] Tanvi Bajpai, Drshika Asher, Anwesa Goswami, and Eshwar Chandrasekharan. 2022. Harmonizing the Cacophony with MIC: An Affordance-aware Framework for Platform Moderation. *Proc. ACM Hum.-Comput. Interact.* 6, CSCW2, Article 385 (Nov. 2022), 22 pages. doi:10.1145/3555110

[4] Tanvi Bajpai and Eshwar Chandrasekharan. 2026. "Think about it like you're a firefighter": Understanding How Reddit Moderators Use the Modqueue. In *Proceedings of the 2026 CHI Conference on Human Factors in Computing Systems (CHI '26)*.

[5] Katy Blumer and Jon Kleinberg. 2025. Tracking patterns in toxicity and antisocial behavior over user lifetimes on large social media platforms. *Scientific Reports* 15 (July 2025), 25369. doi:10.1038/s41598-025-07086-3

[6] William Brannon, Doug Beeferman, Hang Jiang, Andrew Heyward, and Deb Roy. 2024. AudienceView: AI-Assisted Interpretation of Audience Feedback in Journalism. In *Companion Publication of the 2024 Conference on Computer-Supported Cooperative Work and Social Computing* (San Jose, Costa Rica) *(CSCW Companion '24)*. Association for Computing Machinery, New York, NY, USA, 65–68. doi:10.1145/3678884.3681821

[7] Ceren Budak, R. Kelly Garrett, Paul Resnick, and Julia Kamin. 2017. Threading is Sticky: How Threaded Conversations Promote Comment System User Retention. *Proceedings of the ACM on Human-Computer Interaction* 1, CSCW (Dec. 2017), 1–20. doi:10.1145/3134662

[8] Jie Cai and Donghee Yvette Wohn. 2021. After Violation But Before Sanction: Understanding Volunteer Moderators' Profiling Processes Toward Violators in Live Streaming Communities. *Proc. ACM Hum.-Comput. Interact.* 5, CSCW2, Article 410 (Oct. 2021), 25 pages. doi:10.1145/3479554

[9] Jackie Chan, Aditi Atreyasa, Stevie Chancellor, and Eshwar Chandrasekharan. 2022. Community Resilience: Quantifying the Disruptive Effects of Sudden Spikes in Activity within Online Communities. In *Extended Abstracts of the 2022 CHI Conference on Human Factors in Computing Systems* (New Orleans, LA, USA) *(CHI EA '22)*. Association for Computing Machinery, New York, NY, USA, Article 339, 8 pages. doi:10.1145/3491101.3519813

[10] Jackie Chan, Fred Choi, Koustuv Saha, and Eshwar Chandrasekharan. 2026. Examining Algorithmic Curation on Social Media: An Empirical Audit of Reddit's r/popular Feed. *Proceedings of the International AAAI Conference on Web and Social Media* (2026).

[11] Jackie Chan, Charlotte Lambert, Frederick Choi, Stevie Chancellor, and Eshwar Chandrasekharan. 2024. Understanding Community Resilience: Quantifying the Effects of Sudden Popularity via Algorithmic Curation. *Proceedings of the International AAAI Conference on Web and Social Media* 18 (05 2024), 227–240. doi:10.1609/icwsm.v18i1.31310

[12] 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 (Nov. 2019), 1–30. doi:10.1145/3359276

[13] Eshwar Chandrasekharan, Mattia Samory, Shagun Jhaver, Hunter Charvat, Amy Bruckman, Cliff Lampe, Jacob Eisenstein, and Eric Gilbert. 2018. The Internet's Hidden Rules: An Empirical Study of Reddit Norm Violations at Micro, Meso,

and Macro Scales. *Proc. ACM Hum.-Comput. Interact.* 2, CSCW, Article 32 (Nov. 2018), 25 pages. doi:10.1145/3274301

[14] Jonathan P. Chang, Charlotte Schluger, and Cristian Danescu-Niculescu-Mizil. 2022. Thread With Caution: Proactively Helping Users Assess and Deescalate Tension in Their Online Discussions. *Proc. ACM Hum.-Comput. Interact.* 6, CSCW2, Article 545 (Nov. 2022), 37 pages. doi:10.1145/3555603

[15] Justin Cheng, Michael Bernstein, Cristian Danescu-Niculescu-Mizil, and Jure Leskovec. 2017. Anyone Can Become a Troll: Causes of Trolling Behavior in Online Discussions. In *Proceedings of the 2017 ACM Conference on Computer Supported Cooperative Work and Social Computing*. ACM, Portland Oregon USA, 1217–1230. doi:10.1145/2998181.2998213

[16] Frederick Choi, Tanvi Bajpai, Sowmya Pratipati, and Eshwar Chandrasekharan. 2023. ConvEx: A Visual Conversation Exploration System for Discord Moderators. *Proc. ACM Hum.-Comput. Interact.* 7, CSCW2, Article 262 (Oct. 2023), 30 pages. doi:10.1145/3610053

[17] Frederick Choi, Charlotte Lambert, Vinay Koshy, Sowmya Pratipati, Tue Do, and Eshwar Chandrasekharan. 2025. Creator Hearts: Investigating the Impact Positive Signals from YouTube Creators in Shaping Comment Section Behavior. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems (CHI '25)*. Association for Computing Machinery, New York, NY, USA, Article 1212, 18 pages. doi:10.1145/3706598.3713521

[18] Lorenzo Cima, Benedetta Tessa, Amaury Trujillo, Stefano Cresci, and Marco Avvenuti. 2025. Investigating the heterogeneous effects of a massive content moderation intervention via Difference-in-Differences. *Online Social Networks and Media* 48 (2025), 100320. doi:10.1016/j.osnem.2025.100320

[19] B. Craft and P. Cairns. 2005. Beyond guidelines: what can we learn from the visual information seeking mantra?. In *Ninth International Conference on Information Visualisation (IV'05)*. 110–118. doi:10.1109/IV.2005.28

[20] Naomi Craker and Evita March. 2016. The dark side of Facebook®: The Dark Tetrad, negative social potency, and trolling behaviours. *Personality and Individual Differences* 102 (2016), 79–84. doi:10.1016/j.paid.2016.06.043

[21] Kate Crawford and Tarleton Gillespie. 2016. What is a flag for? Social media reporting tools and the vocabulary of complaint. *New Media & Society* 18, 3 (March 2016), 410–428. doi:10.1177/1461444814543163

[22] Nicholas Diakopoulos and Mor Naaman. 2011. Towards quality discourse in online news comments. In *Proceedings of the ACM 2011 conference on Computer supported cooperative work*. ACM, Hangzhou China, 133–142. doi:10.1145/1958824.1958844

[23] Casey Fiesler, Jialun Jiang, Joshua McCann, Kyle Frye, and Jed Brubaker. 2018. Reddit Rules! Characterizing an Ecosystem of Governance. *Proceedings of the International AAAI Conference on Web and Social Media* 12, 1 (Jun. 2018). doi:10.1609/icwsm.v12i1.15033

[24] Sarah Gilbert. 2017. Motivations for Participating in Online Initiatives: Exploring Contributory Behaviour Across Initiative Types. In *Companion of the 2017 ACM Conference on Computer Supported Cooperative Work and Social Computing* (Portland, Oregon, USA) *(CSCW '17 Companion)*. Association for Computing Machinery, New York, NY, USA, 65–68. doi:10.1145/3022198.3024941

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

[26] April Glaser. 2018. Want a Terrible Job? Facebook and Google May Be Hiring. https://slate.com/technology/2018/01/facebook-and-google-are-building-an-army-of-content-moderators-for-2018.html

[27] Google. 2026. Perspective API Model Cards. https://developers.perspectiveapi.com/s/about-the-api-model-cards?language=en_US.

[28] Robert Gorwa, Reuben Binns, and Christian Katzenbach. 2020. Algorithmic content moderation: Technical and political challenges in the automation of platform governance. *Big Data & Society* 7, 1 (2020), 2053951719897945. doi:10.1177/2053951719897945

[29] Agam Goyal, Charlotte Lambert, Yoshee Jain, and Eshwar Chandrasekharan. 2026. Uncovering the Internet's Hidden Values: An Empirical Study of Desirable Behavior Using Highly-Upvoted Content on Reddit. *Proceedings of the International AAAI Conference on Web and Social Media* (2026).

[30] Agam Goyal, Xianyang Zhan, Yilun Chen, Koustuv Saha, and Eshwar Chandrasekharan. 2025. MoMoE: Mixture of Moderation Experts Framework for AI-Assisted Online Governance. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*.

[31] Timothy Graham and Aleesha Rodriguez. 2021. The Sociomateriality of Rating and Ranking Devices on Social Media: A Case Study of Reddit's Voting Practices. *Social Media + Society* 7, 3 (July 2021), 20563051211047667. doi:10.1177/20563051211047667 Publisher: SAGE Publications Ltd.

[32] James Grimmelmann. 2015. The virtues of moderation. *Yale JL & Tech.* 17 (2015), 42.

[33] Cheng Guo and Kelly Caine. 2025. Throwaway Accounts and Moderation on Reddit. In *Proceedings 2025 Symposium on Usable Security*. Internet Society, San Diego, CA, USA. doi:10.14722/usec.2025.23031

[34] Omkar Gurjar, Agam Goyal, and Eshwar Chandrasekharan. 2025. ArgCMV: An Argument Summarization Benchmark for the LLM-era. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*.

[35] Hussam Habib, Maaz Bin Musa, Fareed Zaffar, and Rishab Nithyanand. 2019. To Act or React: Investigating Proactive Strategies For Online Community Moderation. arXiv:1906.11932 [cs.SI] https://arxiv.org/abs/1906.11932

[36] Rafik Hadfi and Takayuki Ito. 2024. Structural complexity predicts consensus readability in online discussions. *Social Network Analysis and Mining* 14, 1 (March 2024), 51. doi:10.1007/s13278-024-01212-1

[37] Aaron Halfaker and R. Stuart Geiger. 2020. ORES: Lowering Barriers with Participatory Machine Learning in Wikipedia. *Proc. ACM Hum.-Comput. Interact.* 4, CSCW2, Article 148 (Oct. 2020), 37 pages. doi:10.1145/3415219

[38] E. Hoque and G. Carenini. 2014. ConVis: A Visual Text Analytic System for Exploring Blog Conversations. *Computer Graphics Forum* 33, 3 (2014), 221–230. doi:10.1111/cgf.12378

[39] Benjamin D. Horne, Sibel Adali, and Sujoy Sikdar. 2017. Identifying the Social Signals That Drive Online Discussions: A Case Study of Reddit Communities. In *2017 26th International Conference on Computer Communication and Networks (ICCCN)*. 1–9. doi:10.1109/ICCCN.2017.8038388

[40] 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. *Proc. ACM Hum.-Comput. Interact.* 7, CSCW2, Article 301 (Oct. 2023), 29 pages. doi:10.1145/3610092

[41] 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* (Copenhagen, Denmark) *(DIS '24)*. Association for Computing Machinery, New York, NY, USA, 1483–1498. doi:10.1145/3643834.3661623

[42] 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* (Honolulu, HI, USA) *(CHI '20)*. Association for Computing Machinery, New York, NY, USA, 1–12. doi:10.1145/3313831.3376383

[43] David Inserra. 2024. *A Guide to Content Moderation for Policymakers*. Policy Analysis 974. Cato Institute, Washington, DC. https://www.cato.org/policy-analysis/guide-content-moderation-policymakers

[44] Shagun Jhaver, Iris Birman, Eric Gilbert, and Amy Bruckman. 2019. Human-Machine Collaboration for Content Regulation: The Case of Reddit Automoderator. *ACM Trans. Comput.-Hum. Interact.* 26, 5, Article 31 (July 2019), 35 pages. doi:10.1145/3338243

[45] Shagun Jhaver, Amy Bruckman, and Eric Gilbert. 2019. Does Transparency in Moderation Really Matter? User Behavior After Content Removal Explanations on Reddit. *Proc. ACM Hum.-Comput. Interact.* 3, CSCW, Article 150 (Nov. 2019), 27 pages. doi:10.1145/3359252

[46] Jialun Aaron Jiang, Charles Kiene, Skyler Middler, Jed R. Brubaker, and Casey Fiesler. 2019. Moderation Challenges in Voice-based Online Communities on Discord. *Proc. ACM Hum.-Comput. Interact.* 3, CSCW, Article 55 (Nov. 2019), 23 pages. doi:10.1145/3359157

[47] David Jurgens, Libby Hemphill, and Eshwar Chandrasekharan. 2019. A Just and Comprehensive Strategy for Using NLP to Address Online Abuse. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, Anna Korhonen, David Traum, and Lluís Màrquez (Eds.). Association for Computational Linguistics, Florence, Italy, 3658–3666. https://aclanthology.org/P19-1357/

[48] Jeon-Hyung Kang and Jihie Kim. 2011. Analyzing Answers in Threaded Discussions Using a Role-Based Information Network. In *2011 IEEE Third International Conference on Privacy, Security, Risk and Trust and 2011 IEEE Third International Conference on Social Computing*. 111–117. doi:10.1109/PASSAT/SocialCom.2011.107

[49] B. Kerr. 2003. Thread Arcs: an email thread visualization. In *IEEE Symposium on Information Visualization 2003 (IEEE Cat. No.03TH8714)*. 211–218. doi:10.1109/INFVIS.2003.1249028

[50] Charles Kiene, Kate Grandprey-Shores, Eshwar Chandrasekharan, Shagun Jhaver, Jialun "Aaron" Jiang, Brianna Dym, Joseph Seering, Sarah Gilbert, Kat Lo, Donghee Yvette Wohn, and Bryan Dosono. 2019. Volunteer Work: Mapping the Future of Moderation Research. In *Companion Publication of the 2019 Conference on Computer Supported Cooperative Work and Social Computing* (Austin, TX, USA) *(CSCW '19 Companion)*. Association for Computing Machinery, New York, NY, USA, 492–497. doi:10.1145/3311957.3359443

[51] Charles Kiene, Jialun Aaron Jiang, and Benjamin Mako Hill. 2019. Technological Frames and User Innovation: Exploring Technological Change in Community Moderation Teams. *Proceedings of the ACM on Human-Computer Interaction* 3, CSCW (Nov. 2019), 1–23. doi:10.1145/3359146

[52] Mahi Kolla, Siddharth Salunkhe, Eshwar Chandrasekharan, and Koustuv Saha. 2024. LLM-Mod: Can Large Language Models Assist Content Moderation?. In *Extended Abstracts of the CHI Conference on Human Factors in Computing Systems* (Honolulu, HI, USA) *(CHI EA '24)*. Association for Computing Machinery, New York, NY, USA, Article 217, 8 pages. doi:10.1145/3613905.3650828

[53] 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 (Oct. 2025), 35 pages. doi:10.1145/3757699

[54] Charlotte Lambert, Frederick Choi, and Eshwar Chandrasekharan. 2024. "Positive reinforcement helps breed positive behavior": Moderator Perspectives on Encouraging Desirable Behavior. *Proc. ACM Hum.-Comput. Interact.* 8, CSCW2, Article 390 (Nov. 2024), 33 pages. doi:10.1145/3686929

[55] Charlotte Lambert, Agam Goyal, Eunice Mok, and Eshwar Chandrasekharan. 2025. Mind Your Ps and Qs: Supporting Positive Reinforcement in Moderation Through a Positive Queue. arXiv:2509.18437 [cs.HC] https://arxiv.org/abs/2509.18437

[56] Charlotte Lambert, Ananya Rajagopal, and Eshwar Chandrasekharan. 2022. Conversational resilience: Quantifying and predicting conversational outcomes following adverse events. In *Proceedings of the International AAAI Conference on Web and Social Media*, Vol. 16. 548–559.

[57] Charlotte Lambert, Koustuv Saha, and Eshwar Chandrasekharan. 2025. Does Positive Reinforcement Work?: A Quasi-Experimental Study of the Effects of Positive Feedback on Reddit. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems (CHI '25)*. Association for Computing Machinery, New York, NY, USA, Article 1213, 16 pages. doi:10.1145/3706598.3713830

[58] 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*. ACM, Vienna Austria, 543–550. doi:10.1145/985692.985761

[59] Na Yeon Lee and Ahran Park. 2024. How online harassment affects Korean journalists? The effects of online harassment on the journalists' psychological problems and their intention to leave the profession. *Journalism* 25, 4 (2024), 900–920. doi:10.1177/14648849231166511

[60] Alyssa Lees, Vinh Q. Tran, Yi Tay, Jeffrey Sorensen, Jai Gupta, Donald Metzler, and Lucy Vasserman. 2022. A New Generation of Perspective API: Efficient Multilingual Character-level Transformers. In *Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*. ACM, Washington DC USA, 3197–3207. doi:10.1145/3534678.3539147

[61] Hanlin Li, Brent Hecht, and Stevie Chancellor. 2022. All That's Happening behind the Scenes: Putting the Spotlight on Volunteer Moderator Labor in Reddit. *Proceedings of the International AAAI Conference on Web and Social Media* 16, 1 (May 2022), 584–595. doi:10.1609/icwsm.v16i1.19317

[62] Jingxian Liao, Mrinalini Singh, Yi-Ting Hung, Wen-Chieh Lin, and Hao-Chuan Wang. 2023. ConceptCombo: Assisting Online Video Access with Concept Mapping and Social Commenting Visualizations. In *Companion Publication of the 2023 Conference on Computer Supported Cooperative Work and Social Computing* (Minneapolis, MN, USA) *(CSCW '23 Companion)*. Association for Computing Machinery, New York, NY, USA, 362–364. doi:10.1145/3584931.3607499

[63] Alexis C. Madrigal. 2018. Inside Facebook's Fast-Growing Content-Moderation Effort. https://www.theatlantic.com/technology/archive/2018/02/what-facebook-told-insiders-about-how-it-moderates-posts/552632/

[64] J. Nathan Matias. 2019. Preventing harassment and increasing group participation through social norms in 2,190 online science discussions. *Proceedings of the National Academy of Sciences* 116, 20 (2019), 9785–9789. doi:10.1073/pnas.1813486116

[65] J. Nathan Matias and Merry Mou. 2018. CivilServant: Community-Led Experiments in Platform Governance. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems* (Montreal QC, Canada) *(CHI '18)*. Association for Computing Machinery, New York, NY, USA, 1–13. doi:10.1145/3173574.3173583

[66] Drashko Nakikj and Lena Mamykina. 2017. DisVis: Visualizing Discussion Threads in Online Health Communities. In *AMIA Annual Symposium Proceedings*, Vol. 2016. American Medical Informatics Association, 944–953.

[67] Albert Park, Andrea L. Hartzler, Jina Huh, Gary Hsieh, David W. McDonald, and Wanda Pratt. 2016. "How Did We Get Here?": Topic Drift in Online Health Discussions. *Journal of Medical Internet Research* 18, 11 (2016), e284. doi:10.2196/jmir.6297

[68] Deokgun Park, Simranjit Sachar, Nicholas Diakopoulos, and Niklas Elmqvist. 2016. Supporting Comment Moderators in Identifying High Quality Online News Comments. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems* (San Jose, California, USA) *(CHI '16)*. Association for Computing Machinery, New York, NY, USA, 1114–1125. doi:10.1145/2858036.2858389

[69] Vyoma Raman, Eve Fleisig, and Dan Klein. 2023. Centering the Margins: Outlier-Based Identification of Harmed Populations in Toxicity Detection. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, Houda Bouamor, Juan Pino, and Kalika Bali (Eds.). Association for Computational Linguistics, Singapore, 9316–9329. doi:10.18653/v1/2023.emnlp-main.579

[70] Harita Reddy and Eshwar Chandrasekharan. 2023. Evolution of Rules in Reddit Communities. In *Companion Publication of the 2023 Conference on Computer Supported Cooperative Work and Social Computing* (Minneapolis, MN, USA) *(CSCW '23 Companion)*. Association for Computing Machinery, New York, NY, USA, 278–282. doi:10.1145/3584931.3606973

[71] Sarah T Roberts. 2019. *Behind the Screen: Content Moderation in the Shadows of Social Media*. Yale University Press.

[72] Koustuv Saha, Eshwar Chandrasekharan, and Munmun De Choudhury. 2019. Prevalence and Psychological Effects of Hateful Speech in Online College Communities. In *Proceedings of the 10th ACM Conference on Web Science* (Boston, Massachusetts, USA) *(WebSci '19)*. Association for Computing Machinery, New York, NY, USA, 255–264. doi:10.1145/3292522.3326032

[73] Martin Saveski, Farshad Kooti, Sylvia Morelli Vitousek, Carlos Diuk, Bryce Bartlett, and Lada A. Adamic. 2021. Social Catalysts: Characterizing People Who Spark Conversations Among Others. *Proceedings of the ACM on Human-Computer Interaction* 5, CSCW2 (Oct. 2021), 1–20. doi:10.1145/3476023

[74] Martin Saveski, Brandon Roy, and Deb Roy. 2021. The Structure of Toxic Conversations on Twitter. In *Proceedings of the Web Conference 2021*. ACM, Ljubljana Slovenia, 1086–1097. doi:10.1145/3442381.3449861

[75] Charlotte Schluger, Jonathan P. Chang, Cristian Danescu-Niculescu-Mizil, and Karen Levy. 2022. Proactive Moderation of Online Discussions: Existing Practices and the Potential for Algorithmic Support. *Proc. ACM Hum.-Comput. Interact.* 6, CSCW2, Article 370 (Nov. 2022), 27 pages. doi:10.1145/3555095

[76] Joseph Seering, Robert Kraut, and Laura Dabbish. 2017. Shaping Pro and Anti-Social Behavior on Twitch Through Moderation and Example-Setting. In *Proceedings of the 2017 ACM Conference on Computer Supported Cooperative Work and Social Computing* (Portland, Oregon, USA) *(CSCW '17)*. Association for Computing Machinery, New York, NY, USA, 111–125. doi:10.1145/2998181.2998277

[77] 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. doi:10.1177/1461444818821316

[78] Amit Sheth, Valerie L. Shalin, and Ugur Kursuncu. 2022. Defining and detecting toxicity on social media: context and knowledge are key. *Neurocomputing* 490 (2022), 312–318. doi:10.1016/j.neucom.2021.11.095

[79] B. Shneiderman. 1996. The eyes have it: a task by data type taxonomy for information visualizations. In *Proceedings 1996 IEEE Symposium on Visual Languages*. 336–343. doi:10.1109/VL.1996.545307

[80] Miriah Steiger, Timir J Bharucha, Sukrit Venkatagiri, Martin J. Riedl, and Matthew Lease. 2021. The Psychological Well-Being of Content Moderators: The Emotional Labor of Commercial Moderation and Avenues for Improving Support. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems* (Yokohama, Japan) *(CHI '21)*. Association for Computing Machinery, New York, NY, USA, Article 341, 14 pages. doi:10.1145/3411764.3445092

[81] TL Taylor. 2018. Regulating the networked broadcasting frontier. In *Watch me play: Twitch and the rise of game live streaming*. Princeton University Press, Chapter 5.

[82] Sahaj Vaidya, Jie Cai, Soumyadeep Basu, Azadeh Naderi, Donghee Yvette Wohn, and Aritra Dasgupta. 2021. Conceptualizing Visual Analytic Interventions for Content Moderation. In *2021 IEEE Visualization Conference (VIS)*. IEEE, 191–195. doi:10.1109/VIS49827.2021.9623288

[83] Chris Vargo, Gina Masullo, and Tobias Hopp. 2024. Deciding to Delete Posts on Reddit: What Factors Influence Content Removal. In *Proceedings of the 57th Hawaii International Conference on System Sciences*. ScholarSpace, University of Hawai'i at Manoa, Hawaii, 2593–2602. doi:10.24251/HICSS.2024.314

[84] Michela Del Vicario, Alessandro Bessi, Fabiana Zollo, Fabio Petroni, Antonio Scala, Guido Caldarelli, H. Eugene Stanley, and Walter Quattrociocchi. 2016. The spreading of misinformation online. *Proceedings of the National Academy of Sciences* 113, 3 (2016), 554–559. doi:10.1073/pnas.1517441113

[85] David Wadden, Tal August, Qisheng Li, and Tim Althoff. 2021. The Effect of Moderation on Online Mental Health Conversations. *Proceedings of the International AAAI Conference on Web and Social Media* 15, 1 (May 2021), 751–763. doi:10.1609/icwsm.v15i1.18100

[86] Donghee Yvette Wohn. 2019. Volunteer Moderators in Twitch Micro Communities: How They Get Involved, the Roles They Play, and the Emotional Labor They Experience. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems* (Glasgow, Scotland Uk) *(CHI '19)*. Association for Computing Machinery, New York, NY, USA, 1–13. doi:10.1145/3290605.3300390

[87] Zachary Yang, Nicolas Grenon-Godbout, and Reihaneh Rabbany. 2023. Towards Detecting Contextual Real-Time Toxicity for In-Game Chat. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, Houda Bouamor, Juan Pino, and Kalika Bali (Eds.). Association for Computational Linguistics, Singapore, 9894–9906. doi:10.18653/v1/2023.findings-emnlp.663

[88] Yulin Yu, Houming Chen, Daniel Romero, and Paramveer S. Dhillon. 2025. Signals in the Noise: Decoding Unexpected Engagement Patterns on Twitter. *Proceedings of the ACM on Human-Computer Interaction* 9, 7 (Oct. 2025), 1–33. doi:10.1145/3757657

[89] Yulin Yu, Julie Jiang, and Paramveer S. Dhillon. 2024. Characterizing the Structure of Online Conversations Across Reddit. *Proceedings of the ACM on Human-Computer Interaction* 8, CSCW2 (Nov. 2024), 1–23. doi:10.1145/3686913

[90] Xianyang Zhan, Agam Goyal, Yilun Chen, Eshwar Chandrasekharan, and Koustuv Saha. 2025. SLM-Mod: Small Language Models Surpass LLMs at Content Moderation. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*. Association for Computational Linguistics, Albuquerque, New Mexico, 8774–8790. doi:10.18653/v1/2025.naacl-long.441

[91] Amy X. Zhang, Michael S. Bernstein, David R. Karger, and Mark S. Ackerman. 2024. Form-From: A Design Space of Social Media Systems. *Proc. ACM Hum.-Comput. Interact.* 8, CSCW1, Article 167 (April 2024), 47 pages. doi:10.1145/3641006

[92] Amy X. Zhang and Justin Cranshaw. 2018. Making Sense of Group Chat through Collaborative Tagging and Summarization. *Proc. ACM Hum.-Comput. Interact.* 2, CSCW, Article 196 (Nov. 2018), 27 pages. doi:10.1145/3274465

[93] Amy X. Zhang, Lea Verou, and David Karger. 2017. Wikum: Bridging Discussion Forums and Wikis Using Recursive Summarization. In *Proceedings of the 2017 ACM Conference on Computer Supported Cooperative Work and Social Computing* (Portland, Oregon, USA) *(CSCW '17)*. Association for Computing Machinery, New York, NY, USA, 2082–2096. doi:10.1145/2998181.2998235

[94] Justine Zhang, Jonathan Chang, Cristian Danescu-Niculescu-Mizil, Lucas Dixon, Yiqing Hua, Dario Taraborelli, and Nithum Thain. 2018. Conversations Gone Awry: Detecting Early Signs of Conversational Failure. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, Iryna Gurevych and Yusuke Miyao (Eds.). Association for Computational Linguistics, Melbourne, Australia, 1350–1361. doi:10.18653/v1/P18-1125

[95] Jiangrui Zheng, Xueqing Liu, Mirazul Haque, Xing Qian, Guanqun Yang, and Wei Yang. 2024. HateModerate: Testing Hate Speech Detectors against Content Moderation Policies. In *Findings of the Association for Computational Linguistics: NAACL 2024*, Kevin Duh, Helena Gomez, and Steven Bethard (Eds.). Association for Computational Linguistics, Mexico City, Mexico, 2691–2710. doi:10.18653/v1/2024.findings-naacl.172

A Toxicity and Score Values for Selected Comments

Table 4 presents a selection of comments from our user studies alongside their toxicity and score values. To protect Reddit user privacy and prevent the amplification of harmful content, all comment text has been paraphrased to preserve the original intent and tone while ensuring they are no longer searchable. To maintain the validity of these examples, we re-evaluated the paraphrased text using the Perspective API (consistent with our main study) to ensure they retained similar toxicity levels (±0.2).

Paraphrased Comment Text	Toxicity	Score
The saw is a capacitance-based sensor, not a metal detector. You can think of it like a touch screen. Can usually saw through a small nail without the safety mechanism activating.	0.01	31
What's your issue? Are you busy being a Jack ass to learn how to respond to a simple question?	0.66	1
Missed times to say Oregon turns into charred duck everytime they get close to a natty	0.39	-10

Table 4: Paraphrased comments displaying their Toxicity (0.0–1.0) and Score.