Proceedings of the Twentieth International AAAI Conference on Web and Social Media (ICWSM 2026)

Asking For It: Question-Answering for Predicting Rule Infractions in Online Content Moderation

Mattia Samory[1], **Diana Pamfile**[1], **Andrew To**[2], **Shruti Phadke**[2]
[1]Sapienza University of Rome
[2]Drexel University
mattia.samory@uniroma1.it, pamfile.1943337@studenti.uniroma1.it, dt686@drexel.edu, sp3945@drexel.edu

Abstract

Online communities rely on a mix of platform policies and community-authored rules to define acceptable behavior and maintain order. However, these rules vary widely across communities, evolve over time, and are enforced inconsistently—posing challenges for transparency, governance, and automation. In this paper, we model the relationship between rules and their enforcement at scale, introducing ModQ, a novel question-answering framework for rule-sensitive content moderation. Unlike prior classification or generation-based approaches, ModQ conditions on the full set of community rules at inference time and identifies which rule best applies to a given comment. We implement two model variants—extractive and multiple-choice QA—and train them on large-scale datasets from Reddit and Lemmy, the latter of which we construct from publicly available moderation logs and rule descriptions. Both models outperform state-of-the-art baselines in identifying moderation-relevant rule violations, while remaining lightweight and interpretable. Notably, ModQ models generalize effectively to unseen communities and rules, supporting low-resource moderation settings and dynamic governance environments.

Introduction

Content moderation on platforms like Reddit and Lemmy is often delegated to the community members, enabling decentralized governance structures where local norms, values, and enforcement thresholds reflect the collective will of each community in addition to platform-wide policies. This approach to moderation empowers communities with localized, context-specific decision-making where community members enforce their own standards of acceptable behavior, often through a set of community-specific rules.

Local rules defined by communities are not static; they evolve dynamically in response to shifts in user demographics, emergent behaviors, platform-wide developments, and external sociopolitical contexts. Prior research has shown that such rules reflect the distinctive social, cultural, and organizational logics of the communities in which they are embedded, leading to significant variation even within a single platform ecosystem (Fiesler et al. 2018; Reddy and Chandrasekharan 2023; Frey et al. 2022). Moderation practices are thus highly contextual, temporally sensitive, and shaped by localized governance priorities.

Despite growing recognition of the sociotechnical complexity of moderation, most existing automated approaches treat rule enforcement as a binary classification task: should a given piece of content be removed or retained? These models primarily focus on the content itself—whether it constitutes, for instance, harassment or hate speech—rather than its alignment with community-specific rules. While recent studies have demonstrated that including rule-level information improves the predictive accuracy of moderation systems (Park et al. 2021; Xin et al. 2024; He, May, and Lerman 2024), these approaches are typically limited in scope. They assess content compliance one rule or category at a time and often require extensive computational resources, constraining their scalability and utility in real-world, volunteer-driven community settings.

To address these limitations, we introduce a novel formulation of community rule enforcement as an information extraction task. Leveraging a comprehensive, longitudinal dataset of moderation actions from Lemmy—a federated platform where rules and enforcement decisions are publicly logged— and existing Reddit moderation datasets, we model the relationship between user comments and the complete set of community rules. Rather than mapping content to a fixed label or rule category, our approach identifies the specific rule violated, conditioned on the full textual set of rules defined by the community at that point in time.

Specifically, we propose two novel information extraction models that consider moderation as a question-answering problem (Figure 1)—ModQ-Extract, an extractive question answering (Q&A) model, and ModQ-Select, a multiple-choice Q&A model. ModQ-Extract treats rule enforcement as a span prediction problem. Given a user comment and the full set of community rules as context, the ModQ-Extract is trained to extract the specific span corresponding to the rule that justifies moderation. ModQ-Select, by contrast, frames rule identification as a multiple-choice Q&A task. For each comment, the model scores its alignment with each of the community's rules and selects the most appropriate rule.

We evaluate our proposed models against several strong baselines, including NormVio—a community-sensitive classifier for rule categories (Park et al. 2021)—and CPL-NoViD (He, May, and Lerman 2024), a state-of-the-art

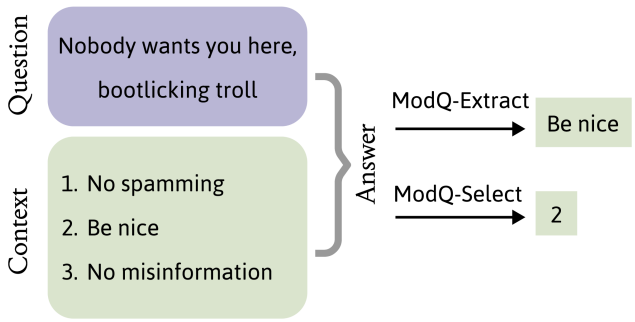

Figure 1: ModQ-Extract and ModQ-Select models presented in this paper for modeling content moderation as a question and answer task

prompting-based model that generates per-rule moderation decisions. Both of our proposed models, ModQ-Extract and ModQ-Select, demonstrate strong performance in identifying rule violations across Reddit (Table 4) and Lemmy (Table 3) datasets. ModQ-Select consistently outperforms all baselines across rule categories and moderation tasks, while ModQ-Extract performs comparably to state-of-the-art methods despite addressing the more granular task of identifying the exact rule rather than a coarse rule category. Crucially, our models generalize better than existing approaches to unseen communities (Figure 3) and previously unencountered rules (Figure 4). This is a key advantage in federated or fast-evolving platforms like Lemmy, where new communities and rules are continually introduced. While performance predictably declines in these out-of-domain settings, both ModQ variants maintain a clear edge over prior models.

Overall, we offer a new formulation of rule-sensitive content moderation as a question-answering task, advancing both methodological and practical goals in the field of computational moderation. The key contributions are as follows:

- We introduce ModQ, a modeling approach that recasts moderation as either extractive or multiple-choice question answering. Unlike prior classification or generation-based methods, our framework is designed to be interpretable, computationally efficient, and capable of handling an open set of community-specific rules.
- By outputting rule-level predictions, the proposed models can be embedded in flagging systems, rule-hinting tools, or rationale-assistance features for moderators. This transparency enhances trust, reduces ambiguity in enforcement, and supports both moderators and end users.
- Our QA formulation can be used to simulate rule revisions, test alignment between stated and enacted policies, and better understand the institutional logics of community governance.

This work contributes to the ICWSM community, bridging NLP and social computing to address real-world challenges in community-specific moderation, transparency, and platform governance at scale.

Related Work

Rules and Moderation Enforcement in Online Communities

Given the growing recognition of its impact on users' lives—from democratic processes to economic opportunities—the governance of online communities is increasingly under scrutiny. While legal and platform-wide regulatory frameworks establish parameters for user behavior (Katzenbach et al. 2023; Vlist et al. 2022), the fine-grained norms are often articulated through community-specific rules, which vary significantly even within the same online platform (Fiesler et al. 2018). These rules are not static; rather, they evolve in response to the changing demographics, prevalent behaviors, and emergent issues within a community, reflecting its unique characteristics and temporal development (Reddy and Chandrasekharan 2023; Frey et al. 2022).

This time-varying and community-specific nature of rules poses a significant challenge for automated content moderation. Research has highlighted the importance of community context, including shared and idiosyncratic norms (Chandrasekharan et al. 2018); conversation context, where subtle cues can indicate violations (Hardaker 2013; Samory and Peserico 2017); user context, leveraging past behavior; interpretation context, addressing mismatches between moderators and users (Koshy et al. 2023; Munzert et al. 2025); and the difficulties of deliberation context in nuanced cases (Koshy et al. 2025). The inability of current automated systems to adapt to an increasingly diverse set of community-specific rules and to account for their evolving nature remains a significant limitation. This work directly addresses the gap by modeling the relationship between content and contextually-defined, community-specific rules.

Automated Content Moderation

The increasing need for scalable solutions to online content moderation has bolstered the adoption of automated moderation systems that commonly employ rule-based filters (Jhaver et al. 2019), keyword detection (Horta Ribeiro et al. 2025), and sophisticated machine learning techniques to identify and flag potentially violating content (Chandrasekharan et al. 2019). Natural Language Processing research contributed substantially to the advancement of the technology for detecting harmful language, such as hate speech and sexism (Yu et al. 2024; Samory et al. 2021). The maturity of this field is evident in the commodification of automated moderation as a service (e.g., Jigsaw's Perspective APIs, OpenAI's Moderation API) (Parker and Ruths 2023). However, a notable disconnection between NLP research and moderation practice is that the former often relies on definitions of undesirable content—and their operationalizations in terms of datasets and tasks—which are often misaligned with the norms and requirements of online communities (Cao et al. 2024). Thus, a growing body of research in the intersection of NLP and Social Computing focuses on developing resources and models that may directly support moderators (Koshy et al. 2025; Chandrasekharan et al. 2019). Most closely related to this paper, recent work aims

to model the enforcement of community-specific rules (Park et al. 2021; Xin et al. 2024; He, May, and Lerman 2024).

The prevailing formulation of rule enforcement as an NLP task has been as a classification problem. In this paradigm, community rules are mapped to a predefined set of classes, and classifiers are trained to predict these classes for new content (e.g., (Park et al. 2021; Xin et al. 2024)). One limitation of this approach is that rules are often mapped to coarse-grained taxonomies, despite the presence of thousands of distinct rules within platforms such as Reddit. A second, more recently explored formulation involves language generation. Here, given a comment and contextual information, a model is tasked with generating the specific rule that applies (e.g., (He, May, and Lerman 2024; Wang et al. 2025; Zhan et al. 2025)). This approach, leveraging sequence-to-sequence models or prompting large language models, offers the potential to model context-dependent moderation, as the same comment appearing in communities with different rules could theoretically be processed differently. However, it also lacks inherent explainability and incurs significant computational demands. The present work introduces a novel formulation of rule enforcement as an information extraction task, sharing the advantage of language generation in that it allows us to model different rules for different communities, while relying on much more computationally efficient models.

Automated Governance beyond Sanctioning

While automation offers scalability in managing vast quantities of user-generated content(Jhaver et al. 2019; Wright 2022), it poses significant challenges (Gorwa, Binns, and Katzenbach 2020). In particular, the lack of transparency in automated moderation raises concerns about over-censorship or the suppression of legitimate speech (e.g., (Thach et al. 2024)). Recent thrusts in social computing research identify compounded opportunities in developing interpretable automated moderation systems. Beyond the automation of sanctioning (e.g., content removal, user suspension), transparent automated moderation may support governance more broadly (Eslami et al. 2024; Park et al. 2021). This includes leveraging computational tools to provide communities with insights into their norms, facilitate more informed rule-making processes, and support deliberative processes among community members and moderators (e.g., (Schneider et al. 2021; Bajpai and Chandrasekharan 2024; Kuo et al. 2025)). The present work adds to this body of work toward developing transparent and interpretable moderation systems, to ultimately empower communities to take a more active and data-driven role in shaping their own online environments, moving beyond a purely reactive model of content policing (e.g., (Filippi and Schneider 2021)).

Data Collection and Curation

We evaluate our approach using moderation data from two multi-community, peer-moderated platforms: Reddit and Lemmy. Tables 1 and 2 present summary statistics for the corresponding datasets.

Reddit

We leverage the Reddit moderation dataset introduced by (Park et al. 2021), which consists of 20K conversation threads in which the final comment was removed by a moderator. These moderated conversations were identified by locating moderator comments that explicitly cited a rule number or quoted rule text in a public response to the removed content. Park et al. retrieved the entire preceding conversation thread—including the original post, all parent comments, and the removed comment itself—using the Pushshift API. Additionally, they included a control set of 32,000 unmoderated conversations, selected based on temporal proximity and matched to the same original post as their moderated counterparts. Each unmoderated conversation was associated with the rule applied in the corresponding moderated case. Finally, the rules were categorized according to the taxonomy developed by Fiesler et al., using BERT-based classifiers fine-tuned for each rule category.

While providing a valuable resource for modeling moderation, the dataset exhibits inherent limitations stemming from its collection methodology. By focusing exclusively on cases where moderators publicly explained their reasoning by referencing a rule, the dataset is limited to a specific subset of moderation actions — those deemed by moderators to warrant such public justification, which may introduce a bias towards certain types of rule violations or moderation styles. In particular, the dataset only includes cases of the application of a select few rules in each community. We address this limitation by constructing a new dataset based on a full record of moderation actions and community rules.

To support our QA-based models, we augmented this dataset by identifying all instances of rules associated with each community across the original dataset. We compiled these rule mentions to construct a rule set for every community, which serves as the contextual input to our models. While the original dataset only includes a limited subset of applied rules, this reconstruction enables us to model rule selection in a more realistic setting where a more comprehensive set of community rules is available at inference time.

Lemmy

Lemmy, a popular platform in the Fediverse, presents itself as an open-source, non-commercial alternative to Reddit. Lemmy consists of thousands of federated instances that host volunteer-moderated communities. We used Lemmy's public API to build a dataset of safe and removed comments. To discover communities on Lemmy, we gathered available lists of instances from multiple portals to the network (fedidb, fediverse observer, awesomelemmy, lemmyverse.net) and recursively snowballed through federated instances.

Moderated comments Unlike most proprietary platforms, Lemmy provides direct access to moderation logs (modlogs) through its public and free API.[1] Modlogs are a complete record of moderator actions, including content removals and rule enforcement justifications (Samory 2021;

[1] An example public modlog endpoint for Lemmy is https://lemmy.world/api/v3/modlog

	Modlogs	Communities	Avg. no. rules per comment	Safe comments	Removed comments	Top 5 communities
Lemmy	40,534	413	6	21,518	19,016	worldnews, memes, technology, news, politics
Reddit	50,732	2,264	3	31,119	19,613	Coronavirus, AmItheAsshole, classicwow, CanadaPolitics, Games

Table 1: Lemmy and Reddit data description

Li, Hecht, and Chancellor 2022; Juneja, Subramanian, and Mitra 2020) as displayed in Figure 2(a).

We collected moderator-removed comments from modlogs in communities local to each instance to avoid duplication. To maintain label accuracy, we excluded comments that had been mass-removed or later reinstated by moderators. We further restricted our analysis to comments that retained their original text, were in English, and included rule enforcement justifications provided by moderators.

Community rules The free-text self-descriptions of communities on Lemmy serve as the conventional location for publishing community rules. We accessed community descriptions directly from modlogs at the time of each moderation action. This approach allowed us to retrieve the precise rule set in effect when the comment was removed, accounting for temporal changes in community governance. We used GPT-4o to automate the structured extraction of rules from these descriptions (Figure 2(b)). To validate the accuracy of this extraction process, three authors manually evaluated 100 randomly sampled rule sets, finding a high accuracy rate of 86.9

Following the approach used for Reddit, we categorized rules using classifiers fine-tuned on data from (Fiesler et al. 2018), mapping them to the coarser-grained categories used by (Park et al. 2021; He, May, and Lerman 2024). Rules that could not be mapped were grouped under the "Other" category. It is important to note that the ModQ models proposed in this paper rely solely on the comment text and the extracted rule texts (Figure 2(d)); rule categories are not used during modeling. Rather, they are included solely to enable clearer comparisons with the baseline models.

Matching rules with removal reason To identify the specific rule violated by a comment in the modlog, we begin by analyzing the *reason* provided by moderators. In many cases, moderators explicitly cite a rule number (e.g., "Rule 6") that corresponds directly to a rule in the community description—these references are captured using regular expressions. However, moderators may also provide a free-text, descriptive rationale. To align such rationales with the correct rule, we tokenize both the moderator-provided reason and all extracted community rules using the lightweight `all-MiniLM` model.[2] We then compute cosine similarity between the tokenized embeddings (Figure 2(c)). If the similarity between a reason and a rule exceeds a threshold of 0.85, we treat that rule as the match. While this threshold is heuristically chosen, manual inspection of 50 reason–rule mappings verified via similarity revealed no mismatches. In the final dataset, 80% of moderator reasons were matched using rule numbers, and the remaining 20% through similarity-based matching.

Safe comments To provide a counterbalance to the moderated comments, we also collected a random sample of publicly accessible comments that were not removed by moderators. These are encoded under the "Safe" class. To ensure consistency, we stratified this sampling by community, limiting it to the same set of communities that contributed moderated comments, and applied the same filtering criteria used for removed content.

Preparing QA Data

For each dataset, we retain three key components: the comment text, the full set of community rules (each with an associated rule number), and the specific rule that was violated. We use the violated rule text to derive the model supervision signal. In ModQ-Extract, the violated rule is treated as a span within the concatenated rule set, allowing the model to learn to extract the relevant substring (Figure 1). In ModQ-Select, the violated rule is instead represented by its corresponding rule number, enabling the model to perform rule selection from a fixed set of candidates (Figure 1).

Models

ModQ

We formulate rule prediction as a question-answering (Q&A) task. Similar to language generation approaches, our method provides the model with a user comment and a contextual input consisting of all applicable community rules. However, instead of asking the model to generate a rule as free-form text, we train it to select the rule that best matches the enforcement decision, either by extracting it as a text span or selecting it from a predefined list.

This reframing allows us to leverage pretrained Q&A architectures, which are highly effective at answer extraction, while avoiding the verbosity and instability often associated with generative outputs. In addition to being more interpretable, this approach offers practical benefits: by including the full rule set as part of the model's context at inference time, we enable the system to operate across a variety of communities and rule configurations without retraining. Crucially, our models do not rely on computationally expensive large language models (LLMs), making them suitable

[2]https://huggingface.co/sentence-transformers/all-MiniLM-L6-v2

	Incivil	Hate	Spam	Content	Doxx	Format	Harass	Meta	Off-tpc	Troll	Other	Safe
Lemmy	8,917	4,739	2,264	6,572	618	2,217	4,202	15,304	1,703	4,412	511	21,518
Reddit	9,574	815	2,574	1,655	244	1,735	2,254	643	1,560	843	-	31,119

Table 2: Rule category distribution in Lemmy and Reddit datasets

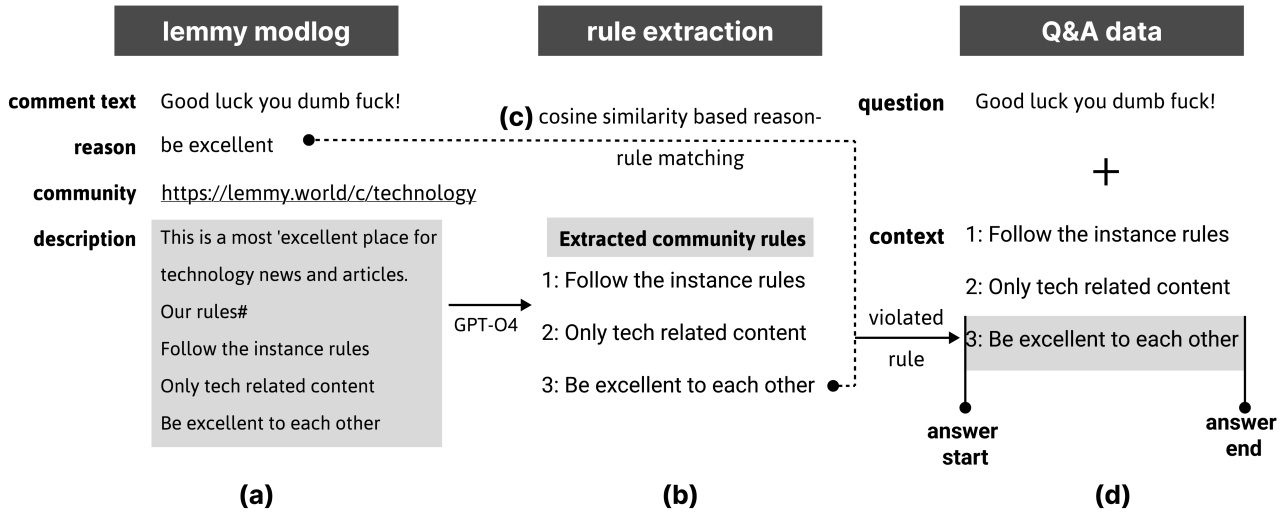

Figure 2: **Lemmy data preparation process:** Figure illustrating various stages in the data preparation phase for Lemmy modlogs. (a) displays a typical Lemmy modlog queried from Lemmy's public API. We use GPT-O4 mini to extract structured community rules from the community description (b) and match the removal *reason* provided by moderators with one of the extracted rules (c). We then prepare Q&A data for the BERT model in the form of *question*, *context*, and *start and end* of the answer as displayed in (d).

for deployment in real-world moderation workflows, including those with limited technical or computational resources.

We implement two model variants aligned with common Q&A paradigms: an extractive model and a multiple-choice model. Both variants fine-tune the same pretrained transformer backbone—`conversational-bert-base-cased`.[3]

ModQ-Extract

Model specifications ModQ-Extract is an extractive Q&A model that identifies the span of text within the rule set that corresponds to the rule applied to a given comment. The model is trained to predict the start and end positions of the relevant span using a cross-entropy loss function. To improve stability during optimization, we apply gradient accumulation and an exponential moving average.

The model input is a pair *(community, QA)*, where *QA* includes the comment and a concatenated list of all rules in the associated community. The rule list serves as the context for extraction. To help the model learn the structural boundaries between rules, we format each rule as `'{rule_n}. {rule_text}'` and wrap both the comment and the rule context with custom tokens.

[3]https://huggingface.co/DeepPavlov/bert-base-cased-conversational

At inference time, the model predicts the most likely start and end positions of a span within the concatenated rule context. We then identify which rule contains that span: if the predicted span falls entirely within one rule, that rule is selected; if it overlaps multiple rules, we assign the prediction to the rule that is most fully covered by the span. This design enables fine-grained token-level reasoning while retaining interpretability at the rule level.

Data augmentation To encourage generalization and reduce overfitting, we apply three data augmentation strategies during training:

- Rule shuffling: Randomizing the order of rules in the context to prevent the model from memorizing positional patterns.
- Number permutation: Randomly renumbering rule tokens to prevent the model from overfitting to specific rule indices.
- Rule exclusion: Removing one random rule from the context, to prevent the model from memorizing rule sets and enabling it to generalize to their variation. This may include removing the correct rule, forcing the model to learn from counterfactual inputs.

ModQ-Select

Model specifications Unlike extractive Q&A, multiple-choice Q&A trains a model to select the correct answer from a set of candidate options. ModQ-Select follows this paradigm by evaluating a user comment against each rule in the community and predicting the likelihood that each rule applies. Similar to the approach in (He, May, and Lerman 2024), the model scores each comment–rule pair for applicability. However, unlike (He, May, and Lerman 2024), our model evaluates all candidate rules in a single batch, allowing for efficient and scalable inference over full rule sets.

Formally, the model receives an input sequence of the form `'{content} [SEP] {community}'`, along with the full set of rule texts and the index of the rule enforced by the moderator. During training, the model constructs a separate input for each comment–rule pair and assigns a binary label indicating whether the rule is correct. At inference time, it computes a score for each rule, then selects the rule with the highest predicted score as the final output.

State-of-the-Art

We compare our models against two state-of-the-art baselines: (1) NormVio, a classifier developed for the NormVio dataset that identifies the category of the violated rule (e.g., "spam"), and (2) CPL-NoViD, a generative, prompt-based approach that extends NormVio by predicting the specific rule being enforced (e.g., "No excessive reposting") (Park et al. 2021; He, May, and Lerman 2024).

NormVio Park et al. model community-sensitive norm violations using transformer-based classifiers that incorporate community-level context. This work serves as a strong baseline for capturing rule-sensitive moderation behavior, particularly on heterogeneous, multi-community platforms like Reddit and Lemmy. We use NormVio's best-performing configuration, which takes both the comment text and the community name as input to contextualize rule application. (Notably, more computationally demanding configurations that included additional metadata did not improve performance.) NormVio consists of nine binary classifiers—one per rule category. Unlike our approach, it does not predict the specific rule violated, but rather the broader rule category. To compare NormVio with other models, we run all category-specific classifiers for each test comment and collect the predicted rule categories. If none of the classifiers are triggered, the comment is labeled as "safe."

CPL-NoViD He, May, and Lerman introduce CPL-NoViD, a prompt-based learning model that incorporates contextual signals such as user history and conversational context to improve accuracy. It represents the state of the art among generative approaches. The model is finetuned using a prompt template of the form:

"In the [subreddit] subreddit, there is a rule: [rule]. A comment was posted: [comment]. Does the comment violate the subreddit rule? [MASK]"

For consistency, we use the zero-shot configuration of CPL-NoViD, which excludes prior conversational context. To ensure fair comparison with other models, we evaluate CPL-NoViD on each rule in the community associated with a given test comment, similar to our setup for NormVio.[4] It has to be noted that the CPL-NoViD approach relies on benefiting from the conversational context. Hence, the comparison provided in this paper largely remains applicable only to the reduced form of CPL-NoViD prompt without the conversational context that matches our setup.

Baselines

Random predict This baseline predicts rule applications according to their frequency distribution in the training data. Because the number of rules per comment varies and the rule distribution is highly skewed, this serves as a reference point for comparing model performance against random assignment.

Naive Bayes We implement a Complement Naive Bayes classifier using TF-IDF vectorization of the comment text. This provides a lightweight and interpretable baseline grounded in traditional text classification techniques.

We train all baseline models to predict the rule number applied to each comment, treating "rule 0" as the class corresponding to safe (i.e., unmoderated) content. One model is trained per community to account for differences in rule sets and moderation patterns across communities. All models' predictions are then aggregated before evaluation.

Experimental Set-Up

Data splits To rigorously evaluate model performance, we employed a multi-stage data splitting strategy. We created train, development, and test sets through random sampling, respectively 80%, 10%, and 10% of the data, stratifying the proportions of moderated comments in each split.

Prior to creating these splits, we held out data from two sources to simulate real-world generalization scenarios. First, we excluded all data from a random sample of $N = 20$ communities to form a *leave-N-communities-out* test set. Second, we excluded $N = 20$ randomly selected rules—defined as unique (rule, community) pairs—from the remaining dataset to construct a *leave-N-rules-out* test set.

These two held-out test sets present challenging out-of-distribution conditions, simulating practical settings where new communities are created or where new rules are added.

Training We train both ModQ-Extract and ModQ-Select models for 5 epochs, selecting the checkpoint with the highest macro F1 score on the development set for final evaluation. Both models are trained using a learning rate of 1×10^{-5} and a weight decay of 0.001. Preliminary experiments with alternative learning rates and weight decay values yielded comparable results. For state-of-the-art baselines, we retain the original training configurations provided

[4]The evaluation script provided by the authors tests whether a given comment violates a specific, pre-identified rule but does not compare that rule to others in the same community. That is, if a comment was removed due to "No personal attacks," the model is only tested on that rule—not on whether "No hate speech" might also apply. This limits interpretability at the rule category level and reduces comparability across candidate rules.

	Incivil	Hate	Spam	Content	Doxx	Format	Harass	Meta	Off-tpc	Troll	Other	Safe
Random	0.20	0.22	0.10	0.33	0.22	0.13	0.28	0.40	0.10	0.23	0.14	0.52
Naive bayes	0.80	0.84	0.75	0.82	0.93	0.70	0.85	0.79	0.68	0.87	0.71	0.76
NormVio	0.86	0.89	**0.85**	0.87	0.93	0.75	**0.91**	0.87	0.76	**0.89**	-	0.84
CPL-NoViD	**0.87**	0.87	0.82	0.80	0.92	0.73	0.87	0.79	**0.78**	0.88	0.75	0.80
ModQ-Extract	0.85	0.88	0.82	0.86	0.96	0.73	0.88	0.86	0.77	0.88	0.71	0.85
ModQ-Select	**0.87**	**0.90**	0.84	**0.89**	**0.97**	**0.76**	**0.91**	**0.88**	**0.78**	**0.89**	**0.77**	**0.87**

Table 3: **Lemmy ModQ macro F1 results:** Table providing comparative macro F1 results on the Lemmy dataset aggregated across categories between different models. Our proposed models—ModQ variants—consistently outperform other models across most categories.

	Incivil	Hate	Spam	Content	Doxx	Format	Harass	Meta	Off-tpc	Troll	Safe
Random	0.49	0.49	0.49	0.49	0.49	0.49	0.49	0.49	0.49	0.49	0.28
Naive bayes	0.44	0.49	0.48	0.49	0.49	0.49	0.48	0.49	0.49	0.49	0.28
NormVio	0.85	0.76	0.81	0.74	0.66	0.80	0.85	0.81	0.74	0.72	0.85
CPL-NoViD	0.72	0.79	0.51	0.82	**0.91**	0.54	0.85	0.54	0.78	0.80	0.53
ModQ-Extract	0.85	**0.82**	0.85	0.77	0.81	0.82	0.87	**0.83**	**0.81**	**0.81**	0.83
ModQ-Select	**0.86**	0.81	**0.86**	**0.79**	0.76	**0.84**	**0.89**	**0.83**	0.80	0.80	**0.86**

Table 4: **Reddit ModQ macro F1 results:** Table providing comparative macro F1 results on the Reddit dataset aggregated across categories between different models. Our proposed models—ModQ variants—consistently outperform other models across most categories.

by the respective authors. All models were trained on an NVIDIA L40 GPU, with average per-epoch training time between 18 and 35 minutes depending on the model. The ModQ-Select model was also trained in parallel on RTX 3060 with 1 hour per epoch time.

Evaluation We report performance over rule categories, which summarize how well models perform on common classes of rule transgressions. We associate each comment with the ground-truth and predicted categories of the rules it was moderated for, adding the "safe" category to comments that either were safe in the ground-truth, or for which models did not predict any other category. To enable comparison with the SOTA, we treat each category as a separate binary classification task and macro-average performance between the two classes (category, not-category). We also provide binary classification metrics distinguishing whether a comment violates *any* moderation rule or is "safe", which allows clearer inspection of the model's safety detection.

Results

Next, we describe the results of model evaluation. We start by discussing overall performance — Tables 4 and 3 summarize key metrics of model performance on the Reddit and Lemmy datasets, respectively. We break down performance according to categories of transgressions and the binary task of determining whether a comment should be moderated. We then unpack how performance translates to challenges in real-world settings. We study to what extent models generalize to new rules and communities. Finally, we discuss which rule categories are often conflated by the model, assessing to what extent erroneous predictions could mislead moderators. Note that the performance breakdown by categories is only performed to offer a direct comparison with other models.

Performance on Different Rule Categories

Both ModQ variants demonstrate strong and competitive performance relative to state-of-the-art (SOTA) models on the Reddit and Lemmy datasets. Across nearly all rule categories, ModQ-Select consistently outperforms all other models, including its extractive counterpart. ModQ-Extract also achieves performance on par with or just slightly below the best-performing baselines, despite tackling the more challenging task of identifying the specific rule text rather than a broad category.

On Lemmy, ModQ-Select reaches the highest F1 scores in nearly every category, including difficult-to-detect ones like content (0.89), harassment (0.91), and doxxing (0.97). ModQ-Extract closely follows, maintaining F1 scores comparable to NormVio and CPL-NoViD, despite using a single model rather than multiple binary classifiers. This underscores the strength of the extractive QA formulation in retaining interpretability while matching the state-of-the-art.

On Reddit, ModQ-Select again outperforms other models across most categories. For instance, it surpasses CPL-NoViD in incivility, spam, and harassment, and matches or exceeds NormVio in safe and trolling categories. Notably, both ModQ variants outperform CPL-NoViD on "safe" classification—a critical measure for avoiding false positives in moderation. Thus, both ModQ variants provide strong performance on the NormVio dataset and achieve granular, per-rule predictions akin to CPL-NoViD. Notably, ModQ-Select

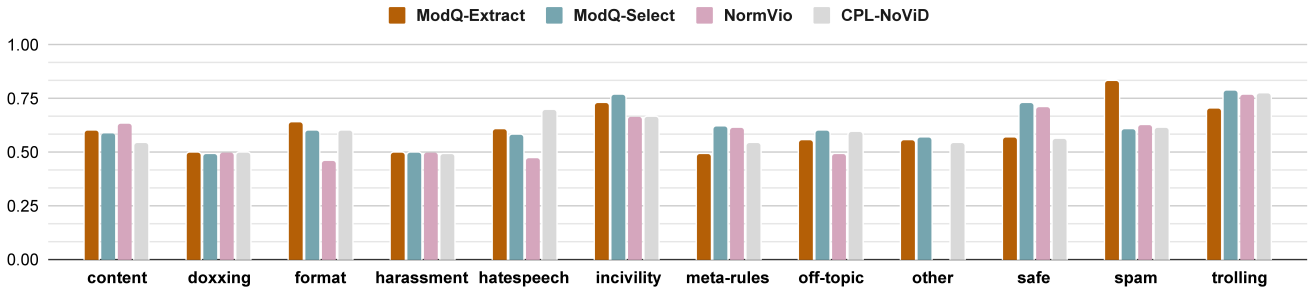

Figure 3: **Lemmy data *leave-N-communities-out*:** Figure illustrating macro F1 results for *leave-N-communities-out* test set.

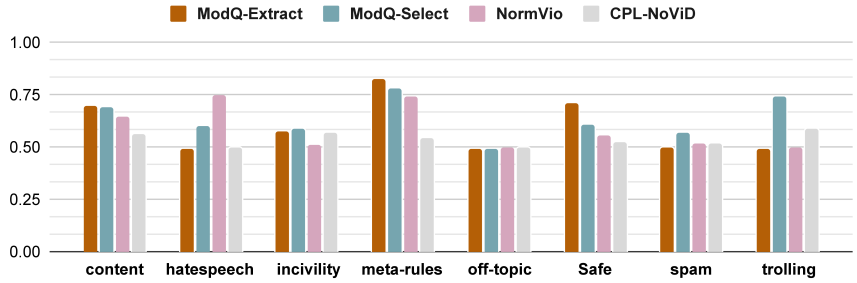

Figure 4: **Lemmy data *leave-N-rules-out*:** Figure illustrating macro F1 results for *leave-N-rules-out* test set.

	Safe (macro F1)	Not Safe (macro F1)
NormVio	0.85	0.83
ModQ-Extract	0.86	0.84
ModQ-Select	**0.87**	**0.86**

Table 5: Binary classification macro F1 for the Lemmy dataset. ModQ-Select outperforms all other models on the binary task.

achieves this strong performance with lower computational demands than both ModQ-Extract and CPL-NoViD.

Overall, these results demonstrate that our proposed ModQ models offer a favorable balance of accuracy, interpretability, and efficiency—achieving strong alignment with moderator decisions while delivering more fine-grained rule-level outputs than prior work.

Binary Rule Classification Task

Both ModQ variants outperform the SOTA on the binary task of predicting whether a comment should be approved or removed according to moderation guidelines. For the NormVio task, we considered a comment to be 'Safe' if it was not flagged by any of the classifiers. CPL-NoViD setup is inherently incompatible with binary safe-not safe prediction task (He, May, and Lerman 2024). Hence, we do not include CPL-NoViD results to avoid unfair comparison.

Generalization to Unseen Rules and Communities

Generalization to unseen communities and rules remains a significant challenge across all models. As expected, performance decreases when models are evaluated on out-of-distribution settings—either due to new community contexts (leave-N-communities-out) in Figure 3 or novel rule sets (leave-N-rules-out) in Figure 4. These conditions reflect real-world scenarios where new communities emerge or moderation policies evolve, demanding robustness beyond static, in-domain training.

Despite this performance drop, both ModQ-Extract and ModQ-Select consistently outperform strong baselines, including NormVio and CPL-NoViD, across nearly all categories in these challenging settings. In the leave-N-communities-out test (Figure 3), ModQ-Select leads across categories, particularly in incivility, off-topic, meta-rules, safe, and trolling, suggesting its strength in transferring moderation behavior across different community norms. ModQ-Extract performs competitively as well, with only marginally lower performance than ModQ-Select, and often exceeding other models in format and spam.

In the leave-N-rules-out setting (Figure 4), where models must generalize to previously unseen moderation policies, the gap between our models and the baselines becomes even more evident. Both ModQ variants maintain stronger macro F1 scores across categories like content, incivility, meta-rules, safe, trolling, and spam. This suggests that ModQ's QA-based formulation effectively leverages the textual structure of rules, enabling it to handle semantic variation in rule wording better than other models.

Taken together, these results demonstrate that ModQ

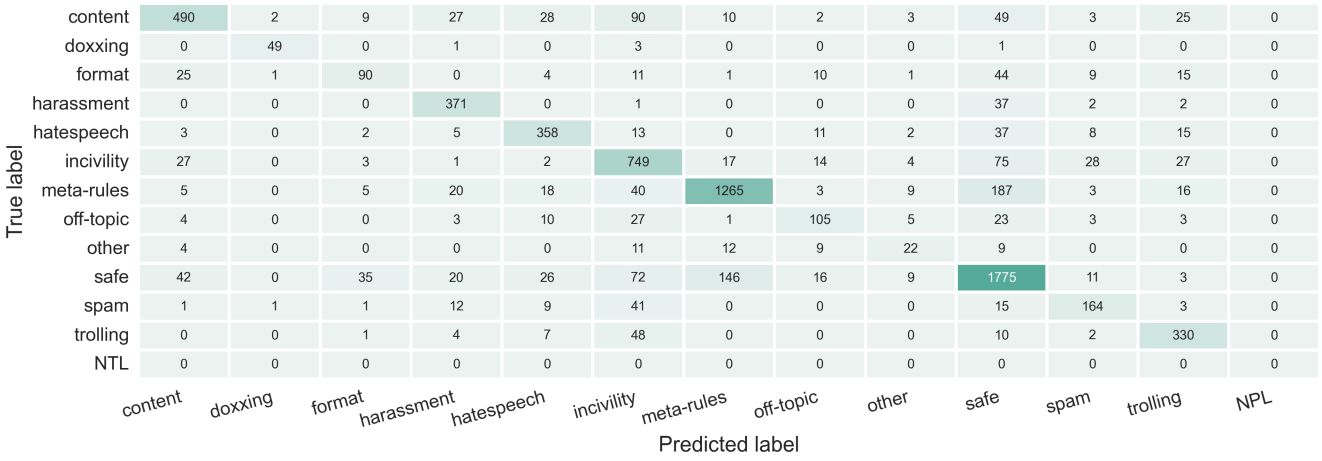

Figure 5: Confusion matrix produced by ModQ-Select between true (y axis) and predicted (x axis) labels

models not only match or exceed state-of-the-art performance in-distribution but also offer superior robustness in out-of-distribution tests. This capability is particularly promising for bootstrapping moderation systems in new or low-resource communities, where labeled training data is scarce or non-existent. By conditioning on the rule set at inference time, rather than relying on fixed taxonomies or extensive fine-tuning, ModQ supports adaptive, scalable moderation that aligns with evolving community standards.

Confusion between Rule Categories

To better understand model behavior and failure modes, we analyze which rule categories are most frequently confused with one another. For brevity, we focus on the best-performing model—ModQ-Select—and visualize its predictions in the confusion matrix shown in Figure 5.[5]

A common pattern is the overprediction of the "safe" label, particularly when the true category belongs to meta-rules, incivility, content, or format. These categories often require contextual or interpretive nuance, and moderation boundaries may be blurry. Importantly, this overprediction skews the model toward caution—minimizing false positives at the cost of some false negatives—making it less likely to flag borderline cases, which may be desirable in human-in-the-loop moderation settings.

Beyond "safe" misclassifications, the model's errors are generally semantically reasonable. For instance format is often predicted as content as the two categories are often entangled by governing post structure and informational quality (e.g., "Mark spoilers" [content], "Tag all posts" [format]. Furthermore, meta-rules, which often include behavioral expectations not tied to a single post, are mistaken for incivility, likely due to their conceptual breadth and lack of precise linguistic markers.

Some categories remain particularly difficult due to abstractness or low support. For example, meta-rules, off-topic, and other violations often lack consistent textual cues, making them harder to learn. Conversely, doxxing—despite being rare—is detected with high reliability, likely due to its more distinctive linguistic signatures and narrow scope.

Overall, this analysis suggests that ModQ-Select's errors tend to arise in borderline or ambiguous cases, particularly where rule boundaries are inherently soft or where training data is sparse. This reinforces the importance of using such models as decision support tools rather than fully autonomous moderators, particularly in nuanced domains.

Discussion and Implications

Before concluding, we discuss the implications of our work for content moderation practices.

Moderation Support

One of the key contributions of this work lies in its potential to enhance the design of moderation support tools.[6] By producing granular, per-rule predictions, our models improve both the interpretability and transparency of automated systems—two features that current machine learning-based moderation tools often lack. Rather than reducing decisions to opaque binary outcomes, the ModQ framework provides actionable outputs tied directly to community-authored rules. This capability unlocks several practical applications. First, it can power flagging systems that not only identify potentially violating content but also specify the rule it may contravene. This helps moderators make faster and more informed decisions, particularly in high-volume environments where consistency is challenging. Further, ModQ's outputs can be used to automatically draft rationale templates for enforcement actions. Rather than requiring moderators to manually justify each action—a known

[5]We report a Multi-Label Confusion Matrix, as each rule is associated with one or more categories (Heydarian, Doyle, and Samavi 2022).

[6]To foster future research on moderation, and to enable moderators to make use of ours, we release ModQ's code at https://github.com/hide-ous/modq

source of cognitive and emotional labor—automated rationales grounded in the rule set can streamline communication with users, while also promoting consistency and accountability. In community contexts where moderation actions are public, this can foster greater trust and reduce perceptions of arbitrariness. Finally, by surfacing rule-based predictions proactively, ModQ can support user-facing nudging tools. For example, when a user begins composing a comment, the model could flag likely rule violations and suggest revisions before submission—much like spell-check or grammar-correction tools. Such preemptive interventions may reduce downstream conflict, lighten the load on moderators, and improve the overall health of online communities.

Low-Resource Moderation

Our approach holds particular promise for addressing the challenges of content moderation in low-resource settings. Notably, ModQ's low computational demands compared to the SOTA make it an accessible solution for computationally-versed community moderators. New online communities, often lacking the extensive moderation logs required to train conventional supervised models, can leverage Q&A models pre-trained on other communities to bootstrap their moderation efforts. Furthermore, ModQ's adaptability to different rule sets at inference time implies that communities with limited computational resources can benefit from sharing models or accessing third-party-hosted ones, which they can access on demand.

Governance Insights

Beyond applications in content moderation, modeling the relationship between rules and their enforcement at scale enables quantifying and exploring crucial aspects of governance in online communities. The Q&A formulation allows us to treat community rules as a structured knowledge base, which can then be queried to study their role in community health. For example, the model can be used to test hypothetical moderation scenarios under different rule sets, enabling communities to proactively evaluate the impact of proposed rule changes. Moreover, model outputs can be analyzed to assess the alignment between formal rules and actual moderation decisions. Discrepancies between predicted rule matches and historical moderation behavior can reveal enforcement gaps, inconsistent application of norms, or implicit biases in moderator judgments. These insights may inform community audits, deliberative reform processes, or broader questions of legitimacy and transparency in digital governance. ModQ provides not just a tool for rule enforcement, but a lens through which to study the institutional logic of rule-governed systems online.

Limitations and Ethical Considerations

Several limitations in our data and methodology warrant discussion. First, the replication of NormVio and CPL-NoViD, and the testing of ModQ on the Reddit dataset are conditioned by missing data, due to the restriction of Reddit's API access. Especially, we could not retrieve all subreddit rules to make a comparable setup as for Lemmy; we could also not rehydrate the text of the rules that are truncated in the original dataset. More broadly, although we followed best practices and evaluated each step, the data processing pipeline includes several inference steps (e.g., for rule extraction, rule matching, and rule categorization), each of which carries inherent risks of inaccuracy.

We also acknowledge the potential for bias in our dataset, starting from its limitation to English-language communities. While our approach shows promising performance, it is important to emphasize that content moderation is a broader and more complex task, involving different roles and expertise, and every transgression and normative act requires contextualization. Notably, a definitive "ground truth" about what should be moderated and whether a rule applies may always exist—even expert human moderator teams disagree on whether a comment violates a rule. Therefore, the models we propose are intended as support to scale up the effort of human moderators, for example, as a flagging system, rather than as an autonomous decision-making tool.

Finally, we recognize that moderation data is sensitive. In particular, the moderator logs we collected from Lemmy carry the risk of re-victimization of the targets of moderated comments, brigading against those whose content was moderated, as well as potential retaliation against moderators. Although this data is publicly accessible through Lemmy API, we have decided not to share it to limit these potential harms.

Conclusions

This paper introduced Q&A as a novel formulation of automated content moderation in online communities, and introduced ModQ: two model variants that identify whether a comment infringes community rules. The proposed models correspond to extractive and multiple-choice Q&A tasks. The models, although computationally lightweight, outperform the state of the art in emulating approve/remove moderation decisions as well as in detecting content across a range of common infraction categories. This framework applies out-of-the-box to rule sets that may vary from community to community or over time.

References

Bajpai, T.; and Chandrasekharan, E. 2024. Towards a Better Modqueue: Designing for Diversity Across Moderator Objectives and Workflows. *arXiv preprint arXiv:2409.16840*.

Cao, Y.; Domingo, L.-F.; Gilbert, S.; Mazurek, M.; Shilton, K.; and Iii, H. D. 2024. Toxicity Detection is NOT all you Need: Measuring the Gaps to Supporting Volunteer Content Moderators through a User-Centric Method. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, 3567–3587.

Chandrasekharan, E.; Gandhi, C.; Mustelier, M. W.; and Gilbert, E. 2019. Crossmod: A Cross-Community Learning-based System to Assist Reddit Moderators. *Proceedings of the ACM on Human-Computer Interaction*, 3(CSCW): 1–30.

Chandrasekharan, E.; Samory, M.; Jhaver, S.; Charvat, H.; Bruckman, A.; Lampe, C.; Eisenstein, J.; and Gilbert, E.

2018. The Internet's Hidden Rules: An Empirical Study of Reddit Norm Violations at Micro, Meso, and Macro Scales. *Proceedings of the ACM on Human-Computer Interaction*, CSCW.

Eslami, M.; Gilbert, E.; Schoenebeck, S.; Baumer, E. P.; Chandrasekharan, E.; De Mooy, M.; Karahalios, K.; Karger, D.; Cottom, T. M.; Monroy-Hernández, A.; et al. 2024. The Future of Research on Social Technologies: CCC Workshop Visioning Report. *arXiv preprint arXiv:2404.10897*.

Fiesler, C.; Jiang, J.; McCann, J.; Frye, K.; and Brubaker, J. 2018. Reddit rules! characterizing an ecosystem of governance. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 12.

Filippi, P. D.; and Schneider, N. 2021. Editorial: Peer Governance in Online Communities. *Frontiers in Human Dynamics*, 3: 771586.

Frey, S.; Zhong, Q.; Bulat, B.; Weisman, W. D.; Liu, C.; Fujimoto, S.; Wang, H.; and Schweik, C. M. 2022. Governing Online Goods: Maturity and Formalization in Minecraft, Reddit, and World of Warcraft Communities. *Proceedings of the ACM on Human-Computer Interaction*, 6(CSCW2): 1–23.

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

Hardaker, C. 2013. "Uh. . . . not to be nitpicky,,,,,but. . . the past tense of drag is dragged, not drug.": An overview of trolling strategies. *Journal of Language Aggression and Conflict*, 1: 58–86.

He, Z.; May, J.; and Lerman, K. 2024. Cpl-novid: Context-aware prompt-based learning for norm violation detection in online communities. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 18, 569–582.

Heydarian, M.; Doyle, T. E.; and Samavi, R. 2022. MLCM: Multi-Label Confusion Matrix. *IEEE Access*, 10: 19083–19095.

Horta Ribeiro, M.; West, R.; Lewis, R.; and Kairam, S. 2025. Post guidance for online communities. *Proceedings of the ACM on Human-Computer Interaction*, 9(2): 1–26.

Jhaver, S.; Birman, I.; Gilbert, E.; and Bruckman, A. 2019. Human-Machine Collaboration for Content Regulation. *ACM Transactions on Computer-Human Interaction (TOCHI)*, 26(5): 1–35.

Juneja, P.; Subramanian, D. R.; and Mitra, T. 2020. Through the looking glass: Study of transparency in Reddit's moderation practices. *Proceedings of the ACM on Human-Computer Interaction*, 4.

Katzenbach, C.; Kopps, A.; Magalhães, J. C.; Redeker, D.; Sühr, T.; and Wunderlich, L. 2023. The Platform Governance Archive v1: A longitudinal dataset to study the governance of communication and interactions by platforms and the historical evolution of platform policies (Data Paper).

Koshy, V.; Bajpai, T.; Chandrasekharan, E.; Sundaram, H.; and Karahalios, K. 2023. Measuring User-Moderator Alignment on r/ChangeMyView. *Proceedings of the ACM on Human-Computer Interaction*, 7(CSCW2): 1–36.

Koshy, V.; Choi, F.; Chiang, Y.-S.; Sundaram, H.; Chandrasekharan, E.; and Karahalios, K. 2025. Venire: A Machine Learning-Guided Panel Review System for Community Content Moderation. *Proceedings of the ACM on Human-Computer Interaction*, 9(7): 1–35.

Kuo, T.-S.; Chen, Q. Z.; Zhang, A. X.; Hsieh, J.; Zhu, H.; and Holstein, K. 2025. PolicyCraft: Supporting Collaborative and Participatory Policy Design through Case-Grounded Deliberation. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*, 1–24.

Li, H.; Hecht, B.; and Chancellor, S. 2022. Measuring the Monetary Value of Online Volunteer Work. *Proceedings of the International AAAI Conference on Web and Social Media*, 16: 596–606.

Munzert, S.; Traunmüller, R.; Barberá, P.; Guess, A.; and Yang, J. 2025. Citizen preferences for online hate speech regulation. *PNAS Nexus*, pgaf032.

Park, C. Y.; Mendelsohn, J.; Radhakrishnan, K.; Jain, K.; Kanakagiri, T.; Jurgens, D.; and Tsvetkov, Y. 2021. Detecting Community Sensitive Norm Violations in Online Conversations. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, 3386–3397.

Parker, S.; and Ruths, D. 2023. Is hate speech detection the solution the world wants? *Proceedings of the National Academy of Sciences*, 120(10): e2209384120.

Reddy, H.; and Chandrasekharan, E. 2023. Evolution of Rules in Reddit Communities. *Computer Supported Cooperative Work and Social Computing*, 278–282.

Samory, M. 2021. On positive moderation decisions. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 15, 585–596.

Samory, M.; and Peserico, E. 2017. Sizing up the troll: A quantitative characterization of moderator-identified trolling in an online forum. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*, 6943–6947.

Samory, M.; Sen, I.; Kohne, J.; Flöck, F.; and Wagner, C. 2021. "Call me sexist, but...": Revisiting sexism detection using psychological scales and adversarial samples. In *Proceedings of the international AAAI conference on web and social media*, volume 15, 573–584.

Schneider, N.; De Filippi, P.; Frey, S.; Tan, J. Z.; and Zhang, A. X. 2021. Modular politics: Toward a governance layer for online communities. *Proceedings of the ACM on Human-Computer Interaction*, 5(CSCW1): 1–26.

Thach, H.; Mayworm, S.; Delmonaco, D.; and Haimson, O. 2024. (In) visible moderation: A digital ethnography of marginalized users and content moderation on Twitch and Reddit. *New Media & Society*, 26(7): 4034–4055.

Vlist, F. N. v. d.; Helmond, A.; Burkhardt, M.; and Seitz, T. 2022. API Governance: The Case of Facebook's Evolution. *Social Media + Society*, 8(2): 20563051221086228.

Wang, L.; Yurechko, K.; Dani, P.; Chen, Q. Z.; and Zhang, A. X. 2025. End User Authoring of Personalized Content Classifiers: Comparing Example Labeling, Rule Writing, and LLM Prompting. *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*, 1–21.

Wright, L. 2022. Automated platform governance through visibility and scale: On the transformational power of automoderator. *Social Media+ Society*, 8(1): 20563051221077020.

Xin, W.; Wang, K.; Fu, Z.; and Zhou, L. 2024. Let community rules be reflected in online content moderation. *arXiv preprint arXiv:2408.12035*.

Yu, Z.; Sen, I.; Assenmacher, D.; Samory, M.; Fröhling, L.; Dahn, C.; Nozza, D.; and Wagner, C. 2024. The unseen targets of hate: A systematic review of hateful communication datasets. *Social Science Computer Review*, 08944393241258771.

Zhan, X.; Goyal, A.; Chen, Y.; Chandrasekharan, E.; and Saha, K. 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)*, 8774–8790.

Paper Checklist

1. For most authors...
   (a) Would answering this research question advance science without violating social contracts, such as violating privacy norms, perpetuating unfair profiling, exacerbating the socio-economic divide, or implying disrespect to societies or cultures? Yes
   (b) Do your main claims in the abstract and introduction accurately reflect the paper's contributions and scope? Yes
   (c) Do you clarify how the proposed methodological approach is appropriate for the claims made? Yes
   (d) Do you clarify what are possible artifacts in the data used, given population-specific distributions? Yes
   (e) Did you describe the limitations of your work? Yes
   (f) Did you discuss any potential negative societal impacts of your work? Yes
   (g) Did you discuss any potential misuse of your work? Yes
   (h) Did you describe steps taken to prevent or mitigate potential negative outcomes of the research, such as data and model documentation, data anonymization, responsible release, access control, and the reproducibility of findings? Yes
   (i) Have you read the ethics review guidelines and ensured that your paper conforms to them? Yes
2. Additionally, if your study involves hypotheses testing...
   (a) Did you clearly state the assumptions underlying all theoretical results? NA
   (b) Have you provided justifications for all theoretical results? NA
   (c) Did you discuss competing hypotheses or theories that might challenge or complement your theoretical results? NA
   (d) Have you considered alternative mechanisms or explanations that might account for the same outcomes observed in your study? NA
   (e) Did you address potential biases or limitations in your theoretical framework? NA
   (f) Have you related your theoretical results to the existing literature in social science? NA
   (g) Did you discuss the implications of your theoretical results for policy, practice, or further research in the social science domain? NA
3. Additionally, if you are including theoretical proofs...
   (a) Did you state the full set of assumptions of all theoretical results? NA
   (b) Did you include complete proofs of all theoretical results? NA
4. Additionally, if you ran machine learning experiments...
   (a) Did you include the code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL)? Yes
   (b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? Yes
   (c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? No, because we observed very small variability over multiple runs of different splits
   (d) Did you include the total amount of compute and the type of resources used (e.g., type of GPUs, internal cluster, or cloud provider)? Yes
   (e) Do you justify how the proposed evaluation is sufficient and appropriate to the claims made? Yes
   (f) Do you discuss what is "the cost" of misclassification and fault (in)tolerance? Yes
5. Additionally, if you are using existing assets (e.g., code, data, models) or curating/releasing new assets, **without compromising anonymity**...
   (a) If your work uses existing assets, did you cite the creators? Yes
   (b) Did you mention the license of the assets? NA
   (c) Did you include any new assets in the supplemental material or as a URL? NA
   (d) Did you discuss whether and how consent was obtained from people whose data you're using/curating? Lemmy data collected is publicly available through Lemmy API.
   (e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? Yes, and we are not releasing this dataset for ethical considerations
   (f) If you are curating or releasing new datasets, did you discuss how you intend to make your datasets FAIR (see FORCE11 (2020))? We are not releasing this dataset for ethical considerations

(g) If you are curating or releasing new datasets, did you create a Datasheet for the Dataset (see Gebru et al. (2021))? NA

6. Additionally, if you used crowdsourcing or conducted research with human subjects, **without compromising anonymity**...

(a) Did you include the full text of instructions given to participants and screenshots? NA

(b) Did you describe any potential participant risks, with mentions of Institutional Review Board (IRB) approvals? NA

(c) Did you include the estimated hourly wage paid to participants and the total amount spent on participant compensation? NA

(d) Did you discuss how data is stored, shared, and deidentified? NA