Toxicity Detection is NOT all you Need: Measuring the Gaps to Supporting Volunteer Content Moderators through a User-Centric Method

Yang Trista Cao[1] **Lovely-Frances Domingo**[1] **Sarah Ann Gilbert**[3]
Michelle L. Mazurek[1] **Katie Shilton**[1] **Hal Daumé III**[1,2]
[1]University of Maryland, College Park [2]Microsoft Research [3]Cornell University
{ycao95,lovely,mmazurek,kshilton,hal3}@umd.edu
sag284@cornell.edu

Abstract

Extensive efforts in automated approaches for content moderation have been focused on developing models to identify toxic, offensive, and hateful content with the aim of lightening the load for moderators. Yet, it remains uncertain whether improvements on those tasks have truly addressed moderators' needs in accomplishing their work. In this paper, we surface gaps between past research efforts that have aimed to provide automation for aspects of content moderation and the needs of volunteer content moderators, regarding identifying violations of various moderation rules. To do so, we conduct a *model review* on Hugging Face to reveal the availability of models to cover various moderation rules and guidelines from three exemplar forums. We further put state-of-the-art LLMs to the test, evaluating how well these models perform in flagging violations of platform rules from one particular forum. Finally, we conduct a user survey study with volunteer moderators to gain insight into their perspectives on useful moderation models. Overall, we observe a non-trivial gap, as missing developed models and LLMs exhibit moderate to low performance on a significant portion of the rules. Moderators' reports provide guides for future work on developing moderation assistant models.

1 Introduction

Content moderation guards online forums against hostility and extremism while maintaining community norms, ensuring the forums remain healthy and open to all participants. While many platforms pay for this service, others, such as Reddit, Discord, Facebook, and Twitch, use a hybrid model, relying on the labor of volunteers. Yet, behind the screen, being a volunteer content moderator is time- and emotionally-draining work. Moderators frequently deal with abusive language, sensitive posts, and unpleasant interactions with users (Seering et al., 2019; Gilbert, 2020; Dosono and Semaan, 2019; Wohn, 2019; Jiang et al., 2019), often doing this work in addition to full-time jobs. To support these volunteers, efforts have been made to develop models, such as Google Perspective API[1] and OpenAI undesired content detection (Markov et al., 2023), that can automatically identify content for removal in order to alleviate moderators' workload.

Although these systems have shown great success in detecting "undesired" content, they primarily focus on toxic content. Yet, content moderation encompasses more than toxicity detection,[2] particularly in platforms that leverage volunteer moderation within smaller communities hosted by the site. For example, Reddit is a platform consisting of various communities, known as "subreddits," focused on a diverse set of topics, and each subreddit has its own moderation rules. Fiesler et al. (2018) conducted a study to explore various subreddit rules, consolidating similar ones, and arrived at 25 distinct rule types. Hence, in order to support moderators in detecting potential rule-violating content, content moderation tools need to support much more than just toxicity detection.

In this paper, we aim to assess to what extent current natural language processing (NLP) models can serve the wide spectrum of moderation rules so that they can be helpful in assisting moderators. First, to understand the functions previous automated content moderation models have focused on, we conduct a *model review* on Hugging Face (HF) with rules from three subreddits as exemplars. This allows us to gauge the progress of past model developments in covering various moderation rules. We use model review as opposed to the more common literature review to gain a technical understanding of the existing models' functions. In addition to examining models that are built to handle specific tasks, we also assess so-called "general-purposed" large language models' (LLMs') capability in cov-

[1]https://perspectiveapi.com/

[2]We use "toxicity detection" as an umbrella term for hate speech detection, incivility detection, etc.

Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing, pages 3567–3587
November 12-16, 2024

ering various moderation rules. We evaluate GPT-4 and Llama-2 on a new evaluation dataset that we collected, consisting of moderation decisions from *r/AskHistorians* and covering a wide range of rules. Finally, using the models' performance on this new dataset as an empirical grounding, we conducted a survey study with active moderators from *r/AskHistorians* ($N = 11$). Through this survey study, we aim to gain insights into model users' perspectives on the performance requirements for useful moderation models on different rules.

We find a substantial gap in both existing NLP models and LLMs' performance when to cover the diverse set of moderation rules that subreddits employ. Our analysis shows the majority of moderation rules from the three subreddits ($\sim 80\%$) are unrelated to toxicity detection, with nearly 70% lacking an huggingface model designed for their resolution. While one might hope that general-purpose LLMs could fill this gap, our experiments with GPT-4 and Llama-2 show that LLMs fall short in their ability to detect violations of many rules. Specifically, both GPT-4 and Llama-2 exhibited moderate to low precision and/or recall ($< 70\%$) for half of the rules from *r/AskHistorians*. Findings from our survey study also indicate that neither LLM has good enough performance to be useful for 6 of 23 *r/AskHistorians* rules (26%), including rules such as `Scope`, `Digression`, and `Sources`[3]. Meanwhile, our survey study also shows that moderators are excited about an assistant model—they are okay with even imperfect tools if they are well-informed about their limitations. For different kinds of rules, moderators have complex needs in terms of model precision and recall. For instance, they need high recall for complex rules (e.g. `Plagiarism` and `Digression`) and high precision for simple rules (e.g. `Current Event` and `Jokes and Humour`). Our study highlights the necessity of, and provides a guide for, future content moderation work to expand its focus beyond toxicity detection, encompassing a broader spectrum of moderation rules to meet the needs of moderators seeking automation assistance.

2 Background and Related Work

Volunteer moderators have been a staple of online content regulation from the earliest days of the internet, tackling issues such as spam (Seering, 2020), trolling (Binns, 2012), and abuse (Dibbell, 1994). Currently, volunteers contribute moderation efforts on nearly all major platforms. For example, volunteers are responsible for moderation within Facebook's Groups, Discord's servers, Twitch's streams, Reddit's subreddits, and through X's (formally known as Twitter) Community Notes. These efforts bring significant value to platforms, with one study estimating that, at the baseline, volunteer moderators on Reddit collectively work 466 hours a day, amounting to a value of about 3.4 million USD per year (Li et al., 2022). With the workload and continuous exposure to online harassment (Gilbert, 2020; Wohn, 2019; Dosono and Semaan, 2019) and toxic content, such as hateful language and violent or upsetting imagery, this moderation work can easily result in burnout.

A common way to mitigate burnout is through the development of tools that support moderation labor, particularly through automation. For example, prior study has found that automation can help at scale, supporting moderators of large communities with high levels of activity (Kiene and Hill, 2020; Seering et al., 2018). Automation is also helpful when communities experience unprecedented or unexpected growth, such in cases where communities receive an influx of new users (Kiene et al., 2016) and in cases where communities are subject to sustained brigades of bad actors spamming hateful content (Han et al., 2023). Mostly, automated tools are developed and maintained by moderators themselves. For example, the most commonly used tool on Reddit, Automod, was originally designed by a moderator before the company took over responsibility for its development and maintenance (Jhaver et al., 2019) and moderators on Twitch have developed bots on the fly in response to hate raids (Han et al., 2023). While automation can help support moderation work, it may also add to it in different ways—like needing to build and maintain a bot—and requires skill sets not all moderators have (Jhaver et al., 2019).

Meanwhile, the NLP field has a long line of work on developing automated models to detect toxic content, which can be used to support moderation work (e.g. Jigsaw, 2019; Pavlopoulos et al., 2020; Park and Fung, 2017; Zampieri et al., 2019). Beyond that, Lees et al. (2022) extended beyond English content and proposed the multilingual Perspective API to detect offensive and hateful content from a diverse range of languages. Moreover, some

[3]For explanation of the rules, please refer to Table 2 in appendix.

study delves into various types of toxic content, trying to improve the nuance of detection models (Markov et al., 2023; Price et al., 2020). However, these studies primarily focus on detecting toxic content, which is merely a subset of moderation work. Therefore, our study uses Reddit rules as an example to identify gaps in NLP moderation models within a wide spectrum of rules.

Recently, Kumar et al. (2023) tested LLMs' performance on rule-based moderation with rules from 95 subreddits. They evaluated GPT models (3, 3.5, 4) on a dataset consisting of removed Reddit posts as violating and not violating content. They conclude that the GPT models are effective in conducting moderation on subreddits like *r/Movies* but struggle with other subreddits such as *r/askscience* and *r/AskHistorians*. Our study builds on previous results by studying gaps in both purpose-built moderation models and LLMs. Rather than focusing on the broad binary questions of removing or keeping posts, we focus on coverage for specific moderation rules. In addition, we include the perspectives of human moderators to understand how existing models address the needs of real-world moderation.

3 Hugging Face Model Review

One traditional approach to understanding a scientific context is to perform a literature review. Because we are particularly interested in deployable models, we opt for a *model review*, where instead of reviewing papers, we review publicly available models that can be adapted to assist content moderators in making moderation decisions, with the focus on Reddit rules. We conduct the model review via Hugging Face (HF)[4], an open-source platform that provides pre-trained models for various NLP tasks, to understand the progress of model development from existing research in covering various moderation rules. To conduct the model review, we gather rules from three subreddits and manually investigate for each rule whether there exists an HF model suitable for detecting such rule violations.

3.1 Method

Chandrasekharan et al. (2018) clustered 100 subreddits into six Meso groups plus four Micro groups based on similarity of the removed posts. Since Micro groups are subreddits, whose rules are more specific to the particular subreddit, we focus on the Meso groups, whose rules are shared across multiple subreddits within the group. We pick three subreddits – *r/Atheism, r/Movies, r/AskHistorian*, from three of the major Meso subreddit clusters and use their rules as representatives for the study. We verify the coverage of the three subreddits' rules by cross-referencing them with the list of 25 rule types from Fiesler et al. (2018), ensuring that all text-based rule types are covered by at least one of the rules from the three subreddits. See appendix Table 1 for our list of rules and its cross-referencing with the list of rule types.

Procedure For each rule, we first crawl its rule description from Reddit and answer the following two questions based on the description.

► **RULE-BASED?**: Can detection of violations on this rule be done by rule-based approaches, e.g. by regular expressions (yes or no)? We first mark out the rules that can easily be handled by rule-based approaches, such as detecting reposts. If the answer is yes, we do not consider this rule any further; all other rules are fully analyzed using following steps.

► **TOXIC?**: Is this rule covered by toxicity detection (yes or no)? To understand what portion of the policies are related to toxicity detection, we mark the rules that can be handled by toxicity detectors, such as detecting harassment.

Then, we search on HF for a *matching model* to each rule. Based on the rule descriptions and our knowledge of previous NLP task names, we identify keywords for model searching. For example, for the rule `Personal Attacks or Flaming`: *"Keep things civil. Avoid fighting words and personal attacks. (applies to all forms of user content, including the user's name.)"*, we use keywords `doxxing` and `cyberbullying`.

We then search on HF by matching keywords with model names or model cards to find the most relevant model to each moderation rule. We traverse the top 20 search results to find the most relevant model, which we call the *matching model*. If more than one model is relevant, we select the model with the highest number of likes on HF. If nothing relevant is found, we declare there is no matching model for this moderation rule. We also skip any model that does not contain a model card (i.e., no model description). See Tables 3, 4, and 5 in the appendix for the keywords we used and the matching models for each rule.

With the matching model, we then answer the following questions for each moderation rule:

[4] https://huggingface.co/models

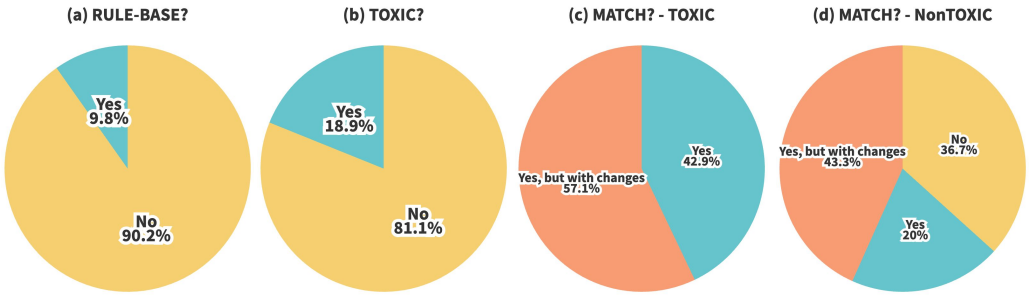

Figure 1: Model review annotation results. Figures (a) and (b) are the results of annotation questions **RULE-BASED?** and **TOXIC?**, respectively. Figures (c) and (d) are the results of question **MATCH?** for rules related to toxicity detection and not related, respectively.

► **MATCH?**: Is there an HF model that can be used to detect violations of this rule? Possible answers are *yes*; *yes, but with changes*; or *no*. We mark *yes, but with changes* when the model is intended to be used for tasks similar to detecting violations of this rule, but some adjustment are needed for the model to be useful in this case. The adjustments needed are noted in the last question.

► **PURPOSE?**: What is the original purpose of this matching model? We note the intended use case for the model as claimed in the model card.

► **GAP?**: In order to adapt this model to detect violations of this rule, what needs to be adjusted? We write the changes, if needed, for the model to be used for this rule, such as domain adaptation.

3.2 Model Review Results

As shown in Figure 1, among the 41 moderation rules, 37 (90%) of the rules cannot be handled by rule-based approaches. Among these, only 7 (19%) are toxicity-related. This reemphasizes that models solely designed for detecting toxicity fall short in meeting the practical needs of content moderators.

Among the 7 toxicity-related rules, we see that 4 (57%) require model modifications. Most of these modifications are customization for the rule, which we will describe later. Even toxicity-related rules, despite having many developed models, do not have perfectly matching models.

Among the 37 non-toxicity-related rules, only 6 (20%) rules have matching HF models that can be applied directly. To see some evaluation results of these matching models, please refer to Appendix A in appendix. 13 (43%) of the rules require some model modifications in order for the matching model to be adapted for the rule, whereas 11 (37%) rules have no matching model at all. These rules include, for example, `low-effort posts` and `proselytizing` from *r/Atheism*, and `no homework` and `no "soapboxing" or loaded questions` from *r/AskHistorians*.

Overall, many rules lack a matching model. For the rules that have one, most require non-trivial model modifications to be covered. These results indicate a substantial gap between the functionalities of previously developed models and those needed for Reddit moderation.

For the **GAP?** question, we primarily identify four types of gaps.

1. The most common adjustment is domain adaptation to the rule-related topic or to Reddit texts. For example, the `No Ambiguous/Misleading/Inaccurate Information` rule in *r/Movies* can be handled with a misinformation detection model, but the model needs to be adapted to title-like texts and information about movie releases.
2. Some models are developed with a limited scope of training data and thus may require extended training to be used for moderation. For instance, we found a plagiarism model that was trained on the Machine Paraphrase Corpus (Wahle et al., 2022) consisting of $\sim 200k$ examples of original and paraphrases using two online paraphrasing tools. The plagiarism cases on Reddit may be more complicated than paraphrasing, and the potential plagiarism source is larger. Hence, the model needs an extension of scope in order to be useful.
3. Models for some rules, especially toxicity-related rules, require customization. For example, *r/Atheism* has the rule `Harassment or Bigotry` that prohibits harassment but permits curse words. A toxicity detection model can be used here but needs modification.
4. Some models can be applied to certain rules,

but with a degree of stretching. For instance, violations of the rule `Off Topic` can potentially be caught by a topic classifier model, but the model requires some major changes to adapt to the subreddit and its related topics.

To see detailed annotations for each question, please refer to Tables 3, 4, and 5 in appendix.

4 LLM Performance on Subreddit Rules

Finding a suitable NLP model for handling a moderation rule, let alone executing the necessary modifications to the model, is not trivial, and some rules simply lack a matching model to address their specific issues. It requires familiarity with NLP techniques and tasks, which may not be within the scope of a content moderator's abilities. Question-answering-based LLM applications, on the other hand, may be more practical for content moderators to employ in assisting their jobs. Therefore, we also evaluate LLMs' performance in catching rule violations. For this evaluation, we focus on the *r/AskHistorians* subreddit, which has 23 moderation rules, and two LLMs — Llama-2 and GPT-4.

4.1 Method

Evaluation Dataset The evaluation dataset was created and annotated by a volunteer moderator for *r/AskHistorians* who has expertise in both moderation and qualitative analysis. For each rule, 11–13 violating posts or comments were selected, save for posts violating the `Illegal artifacts` rule, which, due to the rareness of violations, was only represented 3 times in the dataset. Most content was selected from *r/AskHistory*, a Reddit community that is thematically similar to *r/AskHistorians* but with different moderation rules. Some additional content was taken directly from *r/AskHistorians*. We exclude all borderline cases and only select posts and comments that clearly violate the rules. In total, the dataset includes 101 questions and 134 comments.[5] Because content often violates more than one rule, questions and comments included in the dataset often reflect multiple violations; however, annotations represent the most relevant or severe violation. For example, a comment containing only a joke would be annotated as violating the rule prohibiting joke responses, even though it would also violate rules on comprehensiveness or depth. The dataset also included 11 questions and 10 comments that do not violate any *r/AskHistorians* rules. We use these data as our evaluation dataset.[6]

Model Testing and Prompts To assess the performance of LLMs, we test GPT-4 and Llama-2-13b models. We conduct a pilot experiment with a small number of randomly selected examples from our dataset to determine the prompt for model assessment. Since moderators are mostly unfamiliar with prompt tuning and have limited time to learn this skill, we aim to test the models without extensively experimenting different wordings or examples of the prompt, roughly as they might be used by moderators.

First, as suggested in the prompting guidelines (Shieh, 2023), we find that using rule descriptions in the prompt is helpful for the model to detect violations of the rule. Hence, we crawl rule descriptions from *r/askHistorians* rule explanation web page. As the descriptions from the subreddit web page are long and include many constructive suggestions for post-writers, we manually select useful information for moderation to include in the prompt. See appendix Table 2 for the list of the rules and their descriptions.

Moreover, we experiment with both zero-shot and few-shot settings. In the few-shot setting, we provide the model with three examples that violated the rule and three examples that did not for question contents. For comment contents, we provide one example for each due to the model's input length limit. We realize that few-shot examples are helpful for the Llama-2 model to better understand the task, whereas the GPT-4 model does not have significant improvement from incorporating the examples. Therefore, for the main experiment, we test GPT-4 with the zero-shot setting and Llama-2 with the few-shot setting, but we also report the full set of experiments in the appendix.

Finally, following the suggestions from Shieh (2023), we set the model's role as a moderator assistant and put the main question at the beginning of the prompt. In the few-shot setting, we also repeat the main question at the end of the prompt. See Appendix B for the exact prompts we used.

[5] In *r/AskHistorians*, posts are questions users have asked and comments are responses to questions. Comments included in the dataset are not necessarily responses to the questions included in the dataset.

[6] The dataset, along with a datasheet, are available under a MIT licence at: `https://github.com/TristaCao/into_inclusivecoref`.

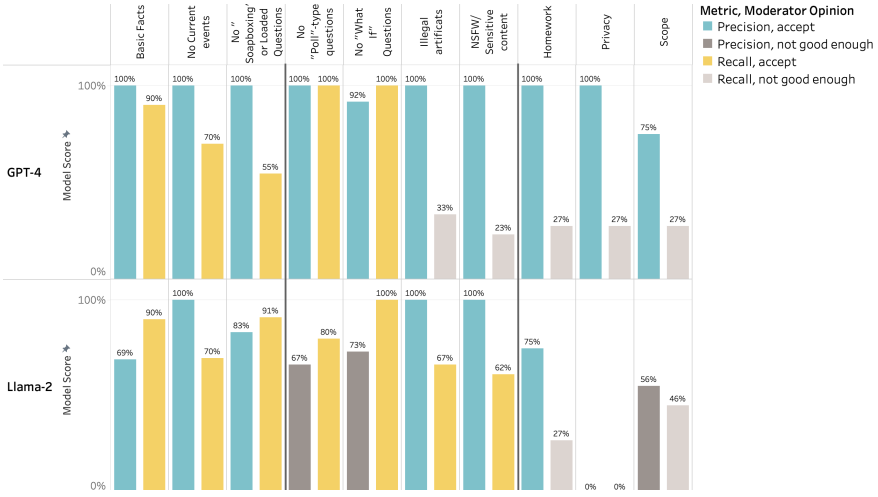

Figure 2: Model performance on detecting question posts that violate moderation rules with GPT-4 (top) and Llama-2 (bottom) models. The x-axis is the moderation rules for question posts. Each bar pair is the precision (left) and recall (right) scores on the specific rule. The ones marked grey are the model scores that moderators would not consider useful as a moderation assistant tool. The left-most rules have good performance from both of the LLMs; the rules in the middle have good performance from at least one of the LLMs; the right-most rules do not have good enough performance from either of the LLMs.

4.2 Moderator Evaluation

To understand volunteer moderators' perspectives on the relative performance of LLMs, we also conducted an evaluation survey with active moderators from *r/AskHistorians*[7]. The survey aimed to capture 1) whether moderators would use the LLM to help with their moderation work on each rule, given the LLM's performance (precision and recall) identifying violations for this rule, and 2) for each rule, what kind of model performance is needed to effectively assist moderators.

To answer the first question, we asked participants to evaluate each model's performance. We showed participants the LLMs' precision and recall scores for each rule from our experiment and asked them to choose among: 1. *performance for this rule is good enough that I would use the tool*, 2. *I would use this tool for this rule if it could catch more violations* (need higher recall), 3. *I would use this tool for this rule if it could be correct more often* (need higher precision), or *4. both measures would have to improve for me to use this tool.*

For the second question, we asked participants to cluster moderation rules based on how important it is for the model to have high precision (or high recall) for each rule, compared to other rules. Options included: *most important, less important,* and *would not use a model for this rule*. We also asked for participants' comments on using LLMs to identify violating posts in general. See appendix Figures 7 and 8 for the exact questions we asked.

Overall, by invitation, we recruited 11 out of 36 total active moderators from *r/AskHistorians*. We obtained their consent at the beginning of the survey. Our participants spent an average of 25 minutes completing the survey. In return, participants received compensation in the amount of $25 USD upon completion of the survey in full.

4.3 Model Performance – Results

We tested GPT-4 and Llama-2 on the evaluation dataset as we described above. We use the first occurrence of "Yes" or "No" in the generated text as the model's output. The results are shown in Figure 2 and Figure 3 (See appendix Figure 5 and Figure 6 for the results of the models in both few-shot and zero-shot settings). The results are mixed – some rules have both high precision and high recall from at least one model, while many of the rules have moderate to low precision, recall, or both. Neither of the two models is clearly better than the other in identifying violating contents for all the rules. The results indicate that certain rules remain unresolved, even for state-of-the-art LLMs.

Our survey results indicate that different moderators have different perceptions of what constitutes an adequate model: for the evaluation questions, participants exhibited only fair agreement (Fleiss' Kappa 0.34). Mirroring policy decision-making standards in the subreddit, we use majority vote

[7]Approved by our institutional IRB, #1704882-9.

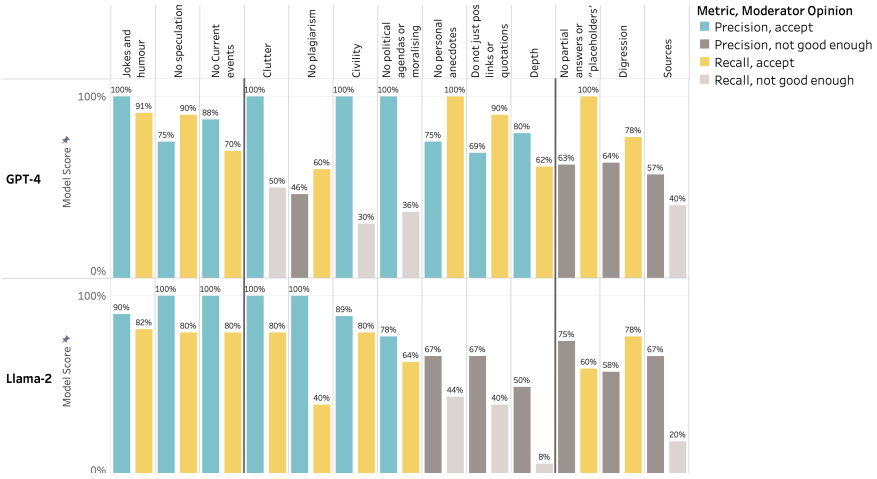

Figure 3: Model performance on detecting comment posts that violate moderation rules with GPT-4 (top) and Llama-2 (bottom) models. The x-axis is the moderation rules for comment posts. Each bar pair is the precision (left) and recall (right) scores on the specific rule. The ones marked grey are the model scores that moderators would not consider useful as a moderation assistant tool. The left-most rules have good performance from both of the LLMs; the rules in the middle have good performance from at least one of the LLMs; the right-most rules do not have good enough performance from either of the LLMs.

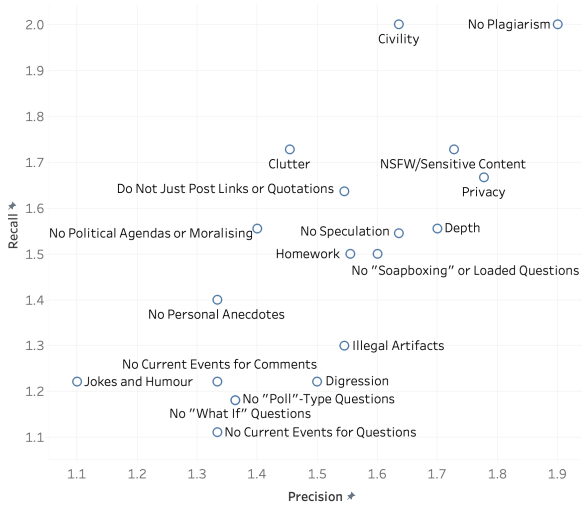

Figure 4: Clustering results from the survey study. Each point is a moderation rule from *r/AskHistorians*. The x-axis shows the precision importance score, or how important it is for a model to have high precision for this rule. Similarly, the y-axis shows the recall importance score. Note that we removed rules for which at least three participants stated they would not use a model.

to summarize participants' evaluation of the models' performance. The results are shown in Figure 2 and Figure 3. In most cases, moderators would consider models with $> 70\%$ precision and recall to be useful. Among the 23 rules, both models are considered useful for three question rules and three comment rules. For some rules, such as `NSFW/Sensitive content`, `No plagiarism`, `No political agendas or moralising`, and `Depth`, moderators can accept low recall; for some, such as `Do not just post links or quotations`, moderators can accept low precision. Importantly, there are six rules (26% of all rules) for which neither model is considered useful. These rules are `Homework`, `Privacy`, and `Scope` question rules and `No partial answer or "placeholders"`, `Digression`, and `Sources` comment rules. This emphasizes that, while current LLMs provide possibilities for supporting content moderation work, further improvement is needed to support useful moderation tools.

Most of the moderators were excited about the idea of having even imperfect assistant models to support their work. Two participants expressed willingness to accept a model with moderate levels of precision and recall, because *"any help is better than no help."* One stated *"A tool flags a lot of things but isn't right too often? Well, it's still flagging things for my attention. A tool that doesn't flag a lot but is very frequently right? Great, I don't have to think about whether or not to remove what it's found."* This response emphasizes the importance of model transparency. Model users often can adapt to a flawed model and make it useful as long as they are informed of its performance limitations.

Some moderators value precision more than recall because a model that mis-flagged many posts would make moderators' workload unnecessarily large. On the other hand, some moderators consider recall to be more important to catch all violating

posts missed by moderators, especially for urgent or sensitive content, such as for the rule `Civility`. Moreover, some participants state that the need for the model varies according to the rules. One participant specified, *"A tool would be useful for assessing and filtering breaches of more complex rules like plagiarism (especially Chat GPT use) or poor sourcing, but given the time needed to assess these claims it would need to be correct the overwhelming majority of the times it flagged something. For simpler rules like clutter, depth, or no jokes then I would like to see a tool which is able to flag the majority of comments which break the rules."* The same participant also pointed out that they would not use a model for rules that are inherently subjective between moderators, such as `Basic facts`. Another participant further indicated the need for model explanations for complex rules.

To inform future improvement on moderation assistant models, we asked moderators to cluster rules according to their needs. The results, as in Figure 4, provide quantified insight into moderators' needs for model precision and recall. We aggregate participants' answers by averaging their choice of cluster for each rule. Each time a participant assigned a rule to *most important*, it is scored as two points; assignments to *less important* are scored as one point. Rules for which a participant selected *would not use* are not included in the score calculation but reported separately. Among the 23 rules, ten require both high precision and high recall (> 1.5). For three rules, at least three participants claimed they would not use a model — `Basic facts`, `Scope`, and `Sources`. We hope our results can serve as a guide for future moderation assistant models to better align with moderators' needs.

5 Discussion

Overall, our findings suggest potential directions for future research on moderation assistants. First, there remains ample opportunity for model improvements. There are ten rules that moderators require both high precision and high recall, as in Figure 4. For three of them, neither GPT-4 nor Llama-2 are considered useful by moderators: `Privacy` and `Homework` with low recall (27%), and `No partial answer or "placeholders"` with low precision (63%). For four of the ten rules, although moderators mark one model as acceptable, there is still room for improvement in precision or recall: `Plagiarism`, `NSFW/Sensitive content`, and `Depth` with recall $40 - 62\%$, and `Do not just post links or quotations` with precision 69%. Besides, the rule `Digression`, which requires high precision, has low precision from both models (58% and 64%). Though some rules are distinct for *r/AskHistorians* (e.g., `Homework`), many rules are also enforced in other subreddit communities. For example, among the 523 subreddits studied by Fiesler et al. (2018), 39% have rules on off-topic content (similar to the `Digression` rule in our study) and 27% have rules on NSFW content. Hence, our findings underscore the existence of plenty of rules from online communities beyond toxicity detection that require exploration.

Despite the models not performing as well as one might hope, volunteer moderators express significant enthusiasm and flexibility about having moderation assistants. People are used to working with tools that may not be perfect, and they are skilled at finding ways to make the most of them. In our findings, moderators explained several personal approaches they take in using a model with moderately low precision or with moderately low recall. However, in order to exercise these strategies, models' abilities must be transparent to enable users to adapt accordingly.

Furthermore, we observed varying requirements among users and rules. Even for the same rule, different moderators expressed distinct preferences for high precision or recall. This suggests that in situations where a trade-off between precision and recall is necessary, moderators may lean towards a model offering more user control, thus allowing for an adjustable trade-off. This way, they can customize the model to align with their preferences. In addition, our participants also expressed some common preferences. For complex rules (e.g. `Plagiarism` and `Digression`), moderators prefer higher recall and model explanations over mere flagging of violations. Conversely, for simpler rules (e.g. `Current Event` and `Jokes and Humour`), they prefer higher precision.

Finally, our study showcases the importance of involving direct stakeholders (e.g. moderators) in the design and development loop for AI tools. They not only ease the alignment of model construction with users' needs but also provide insights model developers might otherwise miss.

6 Conclusions

We identified gaps between the functionalities of previous models on automated moderation and the functions needed to address Reddit moderation rules. We conducted a model review with Hugging Face models to to see for how many rules an existing model can be used to detect violations. The findings reveal that a considerable number of rules are not covered by existing Hugging Face models. Even when a matching model exists, more than half of the rules require non-trivial adjustments of the model in order to be useful. We additionally identified four major types of necessary adjustments: domain adaptation, model scope extension, customization, and function shift.

Moreover, we evaluated two LLMs, GPT-4 and Llama-2, on an evaluation dataset gathered from *r/AskHistorians*, and conducted a survey study to reveal whether the state-of-the-art general-purpose LLMs can handle various moderation rules. The results indicate there is still a significant gap — these models exhibit either low recall or moderate to low precision for many of these rules. Also, moderators are not satisfied with either of the models' performance for six of 23 rules. This limits the usability of these models in aiding moderators to identify violations in real-world moderation.

Finally, our moderator survey study provides insight into model-performance requirements for a moderation assistant model. We observed moderators' excitement about the model and willingness to be flexible to accommodate a flawed model. Moderators also offered constructive feedback on how they wish to use an assistant model differently for different rules. Such user-centered guidance should be useful in building improved, customized moderation tools in the future.

7 Limitations

Our study has several limitations. We are primarily studying Reddit; other platforms may have different sets of rules. We are also focusing on English and text-only posts; violations in other languages or in other forms, such as images and videos, may face different challenges. Furthermore, our survey study is limited to volunteer content moderators from the *r/AskHistorians* community. While we anticipate that the insights gained from the moderation experience might have broader applicability across communities and for other moderators, the extent of generalization remains uncertain without further study.

Moreover, our study and findings are grounded in the set of Reddit moderation rules rather than the frequency of their violation. Consequently, we lack information regarding which rules moderators encounter most frequently and for which rules they want additional assistance.

In addition, our model testing is limited to one-shot and few-shot contexts, as opposed to other strategies, such as interactive teaching and Chain-of-Thoughts.

References

Amy Binns. 2012. Don't feed the trolls! Managing troublemakers in magazines' online communities. *Journalism practice*, 6(4):547–562.

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. *Proceedings of the ACM on Human-Computer Interaction*, 2(CSCW):1–25.

Julian Dibbell. 1994. A rape in cyberspace or how an evil clown, a haitian trickster spirit, two wizards, and a cast of dozens turned a database into a society. *Ann. Surv. Am. L.*, page 471.

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

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(11).

Sarah A Gilbert. 2020. "I run the world's largest historical outreach project and it's on a cesspool of a website." Moderating a public scholarship site on Reddit: A case study of r/AskHistorians. *Proceedings of the ACM on Human-Computer Interaction*, 4(CSCW1):1–27.

Catherine Han, Joseph Seering, Deepak Kumar, Jeffrey T Hancock, and Zakir Durumeric. 2023. Hate raids on twitch: Echoes of the past, new modalities, and implications for platform governance. *Proceedings of the ACM on Human-Computer Interaction*, 7(CSCW1):1–28.

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

Jialun Aaron Jiang, Charles Kiene, Skyler Middler, Jed R Brubaker, and Casey Fiesler. 2019. Moderation challenges in voice-based online communities on discord. *Proceedings of the ACM on Human-Computer Interaction*, 3(CSCW):1–23.

Conversational AI Jigsaw. 2019. Jigsaw unintended bias in toxicity classification. *Kaggle*.

Charles Kiene and Benjamin Mako Hill. 2020. Who uses bots? A statistical analysis of bot usage in moderation teams. In *Extended abstracts of the 2020 CHI conference on human factors in computing systems*, pages 1–8.

Charles Kiene, Andrés Monroy-Hernández, and Benjamin Mako Hill. 2016. Surviving an "eternal september" how an online community managed a surge of newcomers. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*, pages 1152–1156.

Deepak Kumar, Yousef AbuHashem, and Zakir Durumeric. 2023. Watch your language: Large language models and content moderation. (arXiv:2309.14517). ArXiv:2309.14517 [cs].

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*, KDD '22, page 3197–3207, New York, NY, USA. Association for Computing Machinery.

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

Todor Markov, Chong Zhang, Sandhini Agarwal, Florentine Eloundou Nekoul, Theodore Lee, Steven Adler, Angela Jiang, and Lilian Weng. 2023. A holistic approach to undesired content detection in the real world. *Proceedings of the AAAI Conference on Artificial Intelligence*, 37(12):15009–15018.

Ji Ho Park and Pascale Fung. 2017. One-step and two-step classification for abusive language detection on Twitter. In *Proceedings of the First Workshop on Abusive Language Online*, pages 41–45, Vancouver, BC, Canada. Association for Computational Linguistics.

John Pavlopoulos, Jeffrey Scott Sorensen, Lucas Dixon, Nithum Thain, and Ion Androutsopoulos. 2020. Toxicity detection: Does context really matter? In *Annual Meeting of the Association for Computational Linguistics*.

Ilan Price, Jordan Gifford-Moore, Jory Fleming, Saul Musker, Maayan Roichman, Guillaume Sylvain, Nithum Thain, Lucas Dixon, and Jeffrey Sorensen. 2020. Six attributes of unhealthy conversation. *arXiv:2010.07410 [cs]*. ArXiv: 2010.07410.

Joseph Seering. 2020. Reconsidering community self-moderation: the role of research in supporting community-based models for online content moderation. *Proceedings of the ACM on Human-Computer Interaction*, 4:107.

Joseph Seering, Juan Pablo Flores, Saiph Savage, and Jessica Hammer. 2018. The social roles of bots: evaluating impact of bots on discussions in online communities. *Proceedings of the ACM on Human-Computer Interaction*, 2(CSCW):1–29.

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):1417–1443.

Jessica Shieh. 2023. Best practices for prompt engineering with openai api. [Online; accessed Nov 2023].

Jan Philip Wahle, Terry Ruas, Tomáš Foltýnek, Norman Meuschke, and Bela Gipp. 2022. Identifying machine-paraphrased plagiarism. In *Information for a Better World: Shaping the Global Future*, pages 393–413, Cham. Springer International Publishing.

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*, pages 1–13.

Marcos Zampieri, Shervin Malmasi, Preslav Nakov, Sara Rosenthal, Noura Farra, and Ritesh Kumar. 2019. Predicting the type and target of offensive posts in social media. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 1415–1420, Minneapolis, Minnesota. Association for Computational Linguistics.

A Evaluation of HF Matching Models

We evaluated the HuggingFace matching models with our r/AskHistorians evaluation dataset. As indicated in Table 5, there are four non-toxicity-related rules that have matching HF models: `NSFW/Sensitive content`, `Illegal artifacts`, `Basic Facts`, and `Jokes and humor`. Among these four matching models, the model for `Illegal artifacts` is not compatible with the HuggingFace model pipeline and lacks the appropriate code for inference. Similarly, the model for `Basic Facts` uses a model structure that is not defined within Huggingface module.

Consequently, we evaluated only the models for `NSFW/Sensitive content` and the model for `Jokes and humor` using our dataset. The performance of the `NSFW/Sensitive content` model is: precision 57.1% and recall 92.3%. The performance of the `Jokes and humor` model is: precision 55.6% and recall 45.5%.

B Prompt for Testing LLMs

B.1 Prompt for Testing Llama-2 with the Few-shot Setting

```
<s>[INST] «SYS»
You are a moderator assistant.  There is a moderation policy named {rule} that says
{rule description}.  Does the following content violate the {rule} policy?  Simply
give a yes or no answer.
«/SYS»

{violating example content}
Does the following content violate the {rule} policy?  Simply answer yes or no.[/INST]
Yes </s><s>[INST]

{not violating example content}
Does the following content violate the {rule} policy?  Simply answer yes or no.[/INST]
No </s><s>[INST]

{test content}
Does the following content violate the {rule} policy?  Simply answer yes or no.[/INST]
```

B.2 Prompt for Testing GPT-44 with the Zero-shot Setting

```
"role":  "system", "content":  "You are a moderator assistant.",
"role":  "user", "content":  "There is a moderation policy named {rule} that says
{rule description}.  Does the following content violate the {rule} policy?  Simply
give a yes or no answer.",
"role":  "user", "content":  {test content},
```

C LLMs Performances under Zero-shot and Few-shot Settings

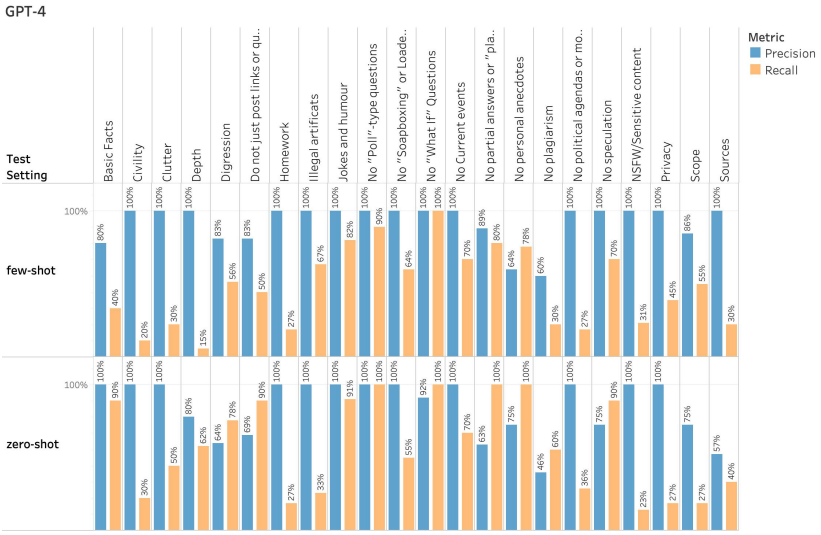

Figure 5: Model performance on detecting violations of moderation rules with GPT-4 model under few-shot setting (top) and zero-shot setting (bottom). The x-axis is the moderation rules. Each bar pair is the precision (left) and recall (right) scores on the specific rule.

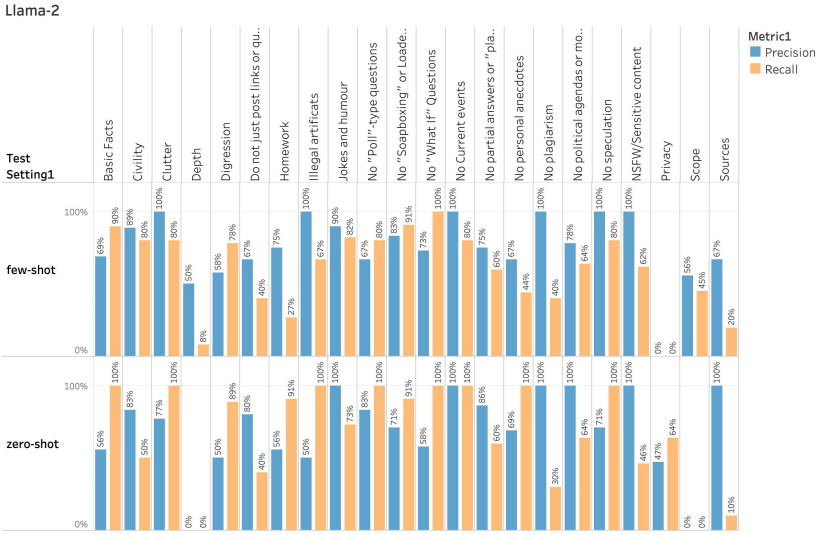

Figure 6: Model performance on detecting violations of moderation rules with Llama-2 model under few-shot setting (top) and zero-shot setting (bottom). The x-axis is the moderation rules. Each bar pair is the precision (left) and recall (right) scores on the specific rule.

D User Survey

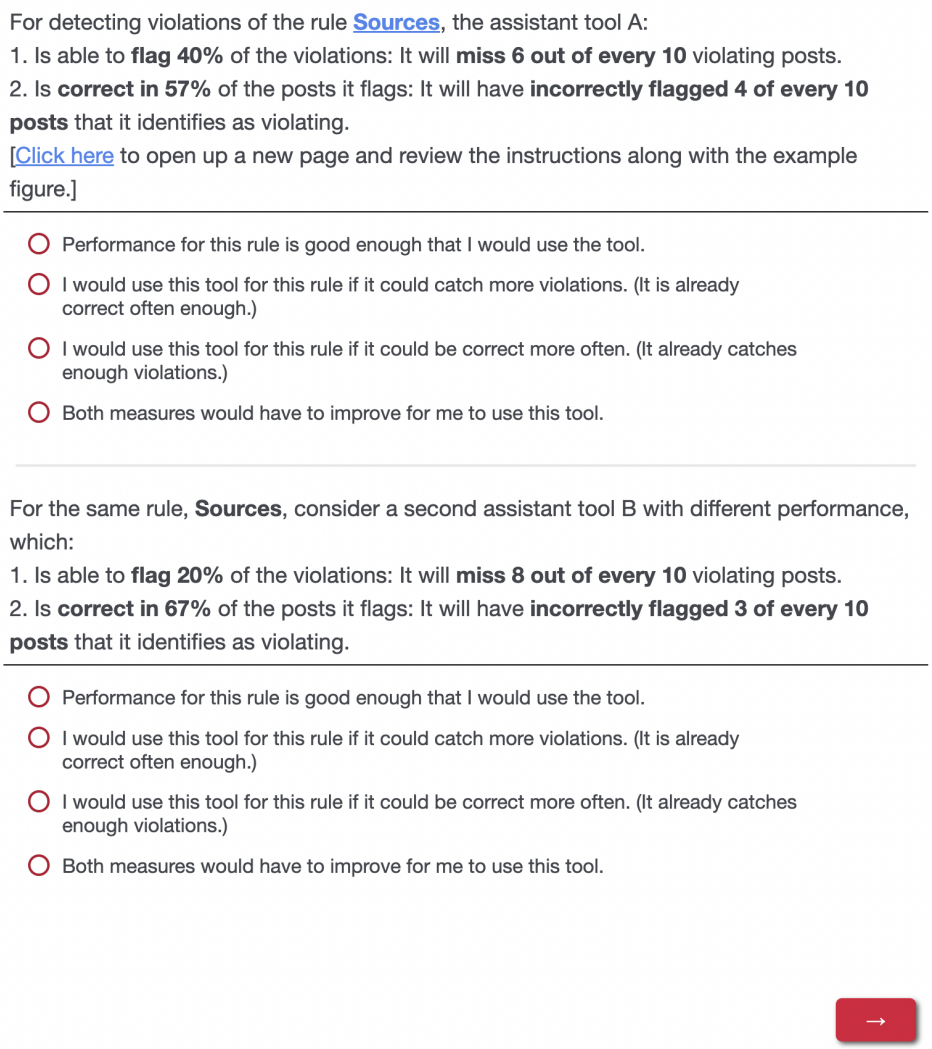

Figure 7: The evaluation survey question for the rule `Sources`. The question content is similar for all the rules.

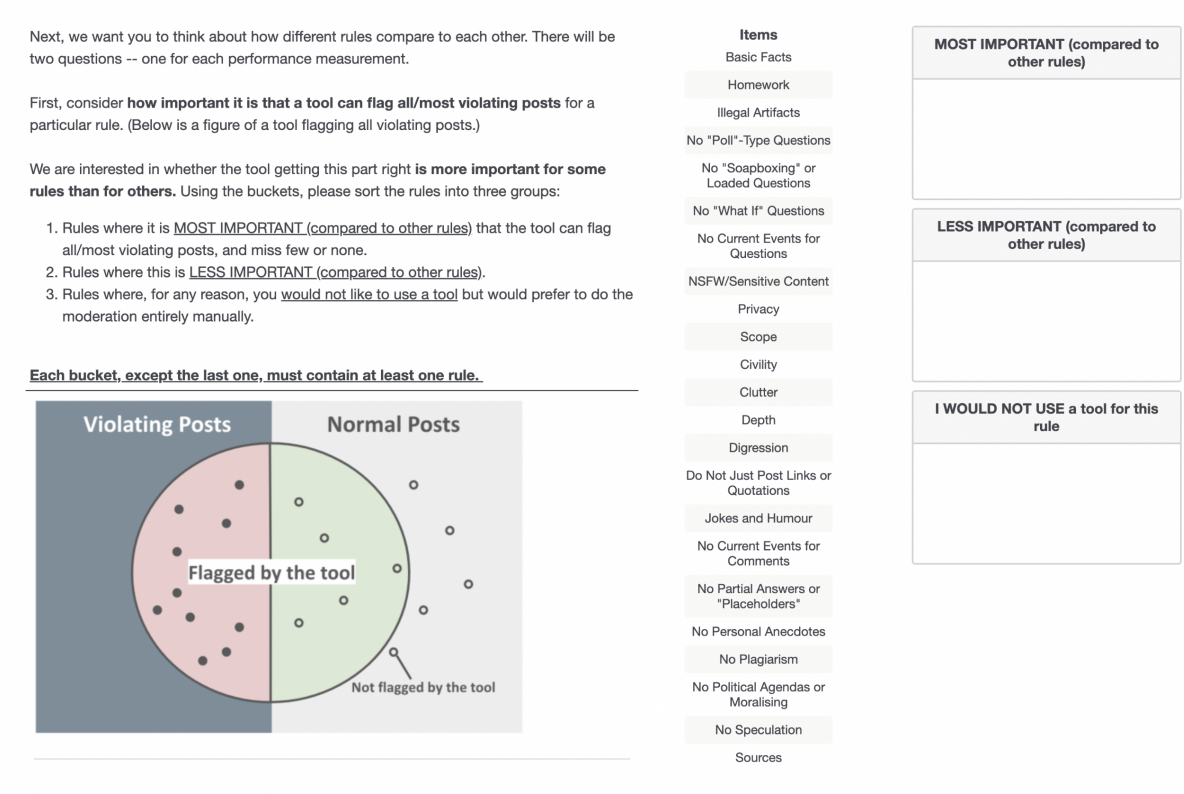

Figure 8: The clustering survey question for recall importance. The same question is asked for precision importance.

Categories of Rules	AskHistorian	Atheism	Movies
Advertising & Commercialization			No Spam & Self-promotion - Same Source Posts
~~Consequences/ Moderation/ Enforcement~~			
Content/Behavior	Homework; No "what if" questions Basic facts Depth; No speculation; Jokes and humor	Proselytizing	No Antagonism/ Flamewars/ Attention Whoring - No Racial/ Sexist/ Dogwhisteling/ Homophobic Slurs; No Extraneous Comic Book Movie Submission"
Copyright	No plagiarism		
Doxxing/Personal Info	Privacy; No personal anecdotes	Personal Attacks or Flaming	
Format	No partial answers or "placeholders"; Do not just post links or quotations		No Ambiguous/ Misleading/ Inaccurate Information or Clickbait in The Submission Title
Harassment	Civility	Harassment or Bigotry	No Antagonism/ Flamewars/ Attention Whoring - No Racial/ Sexist/ Dogwhisteling/ Homophobic Slurs
Hate Speech	Civility		No Antagonism/ Flamewars/ Attention Whoring - No Racial/ Sexist/ Dogwhisteling/ Homophobic Slurs
~~Images~~			
Links & Outside Content	Sources		Prohibited Popular Websites
Low-Quality Content	Scope; Clutter	Low-Effort Posts	
NSFW	NSFW/Sensitive content		
Off-topic	Digression	Off Topic	No Extraneous Comic Book Movie Submission
Personal Army		Brigading	No Subreddit Brigading
Personality	Civility	Personal Attacks or Flaming	No Antagonism/ Flamewars/ Attention Whoring - No Racial/ Sexist/ Dogwhisteling/ Homophobic Slurs
Politics	No political agendas or moralising; No "Soapboxing" or loaded questions		
~~Prescriptive~~			
~~Reddiquette~~			
Reposting			No Repost or Discussion Threads of New Releases
~~Restrictive~~			
Spam		Spam	No Spam & Self-promotion - Same Source Posts
Spoilers			No Repost or Discussion Threads of New Releases; No Spoilers
Trolling		No Trolling	No Ambiguous/ Misleading/ Inaccurate Information or Clickbait in The Submission Title
Voting	"Poll-type" question		
Others	Current event; Illegal artifacts	Don't complain about the use of AAVE or slang	No Encouraging Piracy

Table 1: Subreddit rules from AskHistorian, Atheism, and Movies subreddits and their cross-referencing with categories of rule summarized by Fiesler et al. (2018). Rules within one category are separated by semicolons. Categories crossed out are rules that are not text-relevant or not about rule content.

Rule	Rule Description
Civility	All users are expected to behave with courtesy and politeness at all times. We will not tolerate racism, sexism, or any other forms of bigotry. This includes Holocaust denialism. Nor will we accept personal insults of any kind, and do not allow minor nitpicking of grammar or spelling.
Scope	Submissions must be about a question about the human past, a META post about the state of the subreddit, or an AMA ("Ask Me Anything") with a historical expert or panel of experts.
No Current events	To discourage off-topic discussions of current events, questions, answers, and all other comments must be confined to events that happened 20 years ago or more, inclusively (e.g., 2003 and older).
Homework	Our users aren't here to do your homework for you, but they might be willing to help. Don't just give us your essay/assignment topic and ask us for ideas.
NSFW/Sensitive content	Questions with NSFW titles will be deleted, and we will ask you to repost it with a different title.
No "Poll"-type questions	"Poll"-type questions aren't appropriate here: "Who was the most influential person in history?" or "Who was the worst general in your period?" or "Who are your Top 10 favorite people in history?" If your question includes the words "most" or "least", or "best" or "worst" (or can be reworded to include these words), it's probably a "poll"-type question.
No "Soapboxing" or Loaded Questions	All questions must allow a back-and-forth dialogue based on the desire to gain further information and not be predicated on a false and loaded premise in order to push an agenda. Example: Good Question: "People say that Nixon is the worst President of all time. Why is this so?" Bad Question: "Nixon was the worst President of all time. Why isn't Obama considered the worst?" The bad question is a fishing expedition to try to start a debate about Obama's presidency. Most of these questions will break our 20-year rule, or try to set up a debate about an issue using a long wall of text in the main post.
No "What If" Questions	Questions should be about what did happen, not what could have happened.
Privacy	We do not allow questions that pose possible privacy issues for living or recently deceased persons who are not in the public eye. The cut-off for "recent" is 100 years.
Illegal artifacts	It is our policy to disallow posts asking for further information on artifacts where there is a likelihood that the acquisition or possession of the item might be illegal, unethical, and/or run contrary to sound, historical practices.
Basic Facts	Questions looking for specific, basic facts - for the purpose of this rule, seeking a name, a date or time, a number, a location, the origin of a word, the first or last known instance/example of an object/phenomenon/etc., or a simple list of examples or facts - are not allowed as standalone threads.
Depth	An in-depth answer provides the necessary context and complexity that the given topic calls for, going beyond a simple cursory overview. Your answer should be giving context to the events being discussed, not simply listing some related facts.
Sources	We do not require sources to be preemptively listed in an answer here, but do expect that respondents be familiar with relevant and reliable literature on the topic, and that answers reflect current academic understanding or debates on the subject at hand. But sole reliance on tertiary sources for context and analysis is not allowed, and will result in the removal of a response.
No personal anecdotes	Personal anecdotes are not acceptable answers in this subreddit.
No speculation	Suppositions and personal opinions are not a suitable basis for an answer here. Warning phrases for speculation include: "I guess..." or "My guess is...", "I believe...", "I think...", "... to my understanding", "It makes sense to me that...", "It's only common sense."
No partial answers or "placeholders"	An answer should be full and complete in and of itself.
No political agendas or moralizing	This subreddit is a place for learning and open-minded discussion. As such, answers should not be written in the interests of advancing a personal agenda, but should represent a sincere effort to make an argument from the historical record. They should be constructed in keeping with the principles of the historical method - that is to say, your evidence should not be chosen selectively to support an argument that you feel is right; your argument should instead demonstrably flow from your critical engagement with an appropriate range of evidence.
Do not just post links or quotations	Do not just post links to other sites as an answer. This is not helpful. The expectation is that a user is posting to this subreddit because they are looking for the type of answer dictated by the rules in place here. Please take some time to put the links in context for the person asking the question. Avoid only recommending a source – whether that's another site, a book, or large slabs of copy-pasted text. If you want to recommend a source, please provide at least a small summary of what the source says.
No plagiarism	We have a zero-tolerance policy on blatant plagiarism, such as directly copying and pasting another person's words and trying to pass them off as your own.
Jokes and humor	A post should not consist only of a joke, a humorous remark, or a flippant comment. You can certainly include humor as part of a full and comprehensive post, but your post should not be made solely for the purpose of being funny.
Digression	All comments should be related to the topic as outlined in the original post.
Clutter	Please refrain from commenting for the sake of commenting. This includes, but is not limited to, statements about how interesting the question is, how you would like to see an answer, to "remember to come back later," to share a story that the question reminds you of, and so on.

Table 2: AskHistorian Rules and Rule Descriptions

Rules	Keywords	Annotated model	RULE-BASED?	TOXIC?	MATCH?	PURPOSE?	GAP?
No Trolling	trolling; provocative	TRAC2020_IBEN_A_bert-base-multilingual-uncased	No	Yes	Yes		
Personal Attacks or Flaming	doxxing; cyberbullying	TRAC2020_IBEN_A_bert-base-multilingual-uncased	No	Yes	Yes		
Off Topic	topic classifier	tweet-topic-21-multi	No	No	Yes, but with changes	Twitter topic classification	Need to be finetuned on this topic
~~Image not submitted correctly~~							
Spam	spam	bert-tiny-finetuned-sms-spam-detection	No	No	Yes, but with changes	Detect message spam	1. Need to be adapted to reddit texts rather than message texts. 2. Need to be adapted to detect self-promotion of some youtube account/website etc.
Low-Effort Posts	low effort; low quality	NA	No	No	No		
Proselytizing	proselytizing; Soapbox	NA	No	No	No		
Harassment or Bigotry	harassment; bigotry	Text-Moderation	No	Yes	Yes, but with changes		any and all curse words are permitted
Brigading	brigading	NA - Toxicity	No	Yes	Yes, but with changes		These models may be able to detect some of the brigading comments, but there is no specific model for brigading and some brigading encouraging comments may not be caught by the toxicity models

Table 3: Model review for the Atheism subreddit. The second and third columns are the keywords used for model searching and the best-matching model from hugging face. The rest columns are annotation results. Crossed-out rules are rules not related to text input.

Rules	Keywords	Annotated model	RULE-BASED?	TOXIC?	MATCH?	PURPOSE?	GAP?
~~Do Familiarize Yourself With Our Rules~~							
Incivility - No Antagonism/ Flamewars/ Attention Whoring - No Racial/ Sexist/ Dogwhistling/ Homophobic Slurs	offensive; hatespeech	Text-Moderation	No	Yes	Yes, but with changes	Yes	Quoting something offensive from a movie is okay but needs to be put in quotation marks.
No Spam & Self-promotion - Same Source Posts	spam	bert-tiny-finetuned-sms-spam-detection	No	No	Yes, but with changes	No	Detect message spam - 1. Need to be adapted to Reddit texts rather than message texts. 2. Need to be adapted to detect self-promotion of some YouTube account/website etc.
~~No Image Posts & Memes~~							
No Ambiguous/ Misleading/ Inaccurate Information or Clickbait in The Submission Title - misinformation	misinformation	TwHIN-BERT-Misinformation-Classifier	No	No	Yes, but with changes	No	Detect misinformation for Twitter posts - 1. Need to be adapted to title-like texts. 2. Need to be adapted to a movie-related context
No Ambiguous/ Misleading/ Inaccurate Information or Clickbait in The Submission Title - Clickbait	clickbait	distilroberta-clickbait	No	No	Yes	Yes	
No Extraneous Comic Book Movie Submission	misinformation	TwHIN-BERT-Misinformation-Classifier	No	No	Yes, but with changes	No	Detect misinformation for Twitter posts - Need to be adapted to a movie-related context
No Repost	rule-based		Yes	No			
No Discussion Threads of New Releases	rule-based		Yes	No			
No Spoilers	spoiler	roberta-base-finetuned-imdb-spoilers	No	No	Yes	Yes	
No Encouraging Piracy	piracy	Classification	No	No	Yes, but with changes	No	Detect if a website is a piracy website - Need to detect not only piracy website but also comments that encourage piracy

Table 4 – *Continued from previous page*

Rules	**Keywords**	**Annotated model**	RULE-BASED?	TOXIC?	MATCH?	PURPOSE?	GAP?
No Subreddit Brigading	brigading	NA - Toxicity	No	Yes	Yes, but with changes	No	These models may be able to detect some of the brigading comments, but there is no specific model for brigading and some brigading encouraging comments may not be caught by the toxicity models
Prohibited Popular Websites	rule-based		Yes	No			

Table 4: Model review for the Movies subreddit. The second and third columns are the keywords used for model searching and the best-matching model from hugging face. The rest columns are annotation results. Crossed-out rules are rules not related to text input.

Rules	Keywords	Annotated model	RULE-BASED?	TOXIC?	MATCH?	PURPOSE?	GAP?
Civility	toxicity; hate-speech	Text-Moderation	No	Yes	Yes		The model may not be able to catch Holocaust denialism, which is vital in this case.
Current event	time/year prediction for event; history event; event extractor	GoLLIE-7B	No	No	Yes, but with changes	Information extraction	1. The model needs to be adapted to question-contexts and extract event names from the question. 2. Then by matching the event name with Wikipedia page, we may get the year or time of the event.
Homework	homework	NA	No	No	No		
Scope	topic classifier	tweet-topic-21-multi	No	No	Yes, but with changes	Twitter topic classification	Need to be fine-tuned to this topic.
NSFW/Sensitive content	NSFW; sexual	NSFW_text_classifier	No	No	Yes		
"Poll-type" question	poll	NA	No	No	No		
"Soapboxing" or loaded questions	soapboxing; agenda pushing	NA	No	No	No		
No "what if" questions	hypothetical; counter-factual; alternative history	NA	No	No	No		
Privacy	PII; personal private	lgpd_pii_identifier	No	No	Yes, but with changes	Identify sensitive data in the scope of LGPD. The goal is to have a tool to identify document numbers like CNPJ, CPF, people's names, and other kinds of sensitive data, allowing companies to find and anonymize data according to their business needs, and governance rules.	Need to be adapted to the sensitive data here
Illegal artifacts	illegal; safety	beaver-dam-7b	No	No	Yes		
Basic facts	basic facts; factoid	nf-cats	No	No	Yes		
Depth	depth	NA	No	No	No		
Sources	reliable source	NA	No	No	No		

Table 5 – *Continued from previous page*

Rules	**Keywords**	**Annotated model**	RULE-BASED?	TOXIC?	MATCH?	PURPOSE?	GAP?
No personal anecdotes	personal anecdote	muppet-roberta-base-joke_detector	No	No	Yes, but with changes	Detect jokes, stories, and anecdotes	The model is trained with 2000 jokes. Here the detection target is not about jokes. In addition, here the need is focusing on personal anecdotes.
No speculation	speculation; factual; fact-check	verdict-classifier-en	No	No	Yes, but with changes	Fact-checking model for detecting misinformation	The model does not detect speculation or personal opinion. Here the detection target is the speculation of history events, which can be especially hard for such a fact-checking model.
No partial answers or "placeholders"	partial answer	NA	No	No	No		
No political agendas or moralizing	political agenda; moralizing	NA	No	No	No		
Do not just post links or quotations	rule-based	NA	Yes	No			
No plagiarism	plagiarism	longformer-base-plagiarism-detection	No	No	Yes, but with changes	Detect machine-paraphrased plagiarisms	The real-world source of plagiarism is way larger than the training and testing dataset size of this model.
Jokes and humor	joke; humor; flippant	muppet-roberta-base-joke_detector	No	No	Yes		
Digression	relevance	response-quality-classifier-large	No	No	Yes, but with changes	Rate the relevancy and specificity of a response within a dialogue	Need to be adapted to the Reddit (question-comment) texts.
Clutter	clutter	NA	No	No	No		

Table 5: Model review for the AskHistorian subreddit. The second and third columns are the keywords used for model searching and the best-matching model from Hugging Face. The rest of the columns are annotation results. Crossed-out rules are rules not related to text input.