

LLM-Mod: Can Large Language Models Assist Content Moderation?

Mahi Kolla*

mkolla2@illinois.edu

University of Illinois Urbana-Champaign
Urbana, Illinois, USA

Eshwar Chandrasekharan

eshwar@illinois.edu

University of Illinois Urbana-Champaign
Urbana, Illinois, USA

Siddharth Salunkhe*

ss190@illinois.edu

University of Illinois Urbana-Champaign
Urbana, Illinois, USA

Koustuv Saha

ksaha2@illinois.edu

University of Illinois Urbana-Champaign
Urbana, Illinois, USA

ABSTRACT

Content moderation is critical for maintaining healthy online spaces. However, it remains a predominantly manual task. Moderators are often exhausted by low moderator-to-posts ratio. Researchers have been exploring computational tools to assist human moderators. The natural language understanding capabilities of large language models (LLMs) open up possibilities to use LLMs for online moderation. This work explores the feasibility of using LLMs to identify rule violations on Reddit. We examine how an LLM-based moderator (LLM-Mod) reasons about 744 posts across 9 subreddits that violate different types of rules. We find that while LLM-Mod has a good true-negative rate (92.3%), it has a bad true-positive rate (43.1%), performing poorly when flagging rule-violating posts. LLM-Mod is likely to flag keyword-matching-based rule violations, but cannot reason about posts with higher complexity. We discuss the considerations for integrating LLMs into content moderation workflows and designing platforms that support both AI-driven and human-in-the-loop moderation.

ACM Reference Format:

Mahi Kolla, Siddharth Salunkhe, Eshwar Chandrasekharan, and Koustuv Saha. 2024. LLM-Mod: Can Large Language Models Assist Content Moderation?. In *Extended Abstracts of the CHI Conference on Human Factors in Computing Systems (CHI EA '24)*, May 11–16, 2024, Honolulu, HI, USA. ACM, New York, NY, USA, 8 pages. <https://doi.org/10.1145/3613905.3650828>

1 INTRODUCTION AND BACKGROUND

Online communities function in large part due to the outcomes of effective moderation. Well-moderated communities are productive, open to a variety of members, and incur low physical and social costs [16]. Productivity is often stymied by the common abuses detailed in Grimmelmann’s taxonomy of moderation [16]. In particular, because only limited posts can be viewed at a time, the moderator’s role becomes increasingly important to reduce

*These authors contributed equally to this work.

Permission to make digital or hard copies of part or all of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for third-party components of this work must be honored. For all other uses, contact the owner/author(s).
CHI EA '24, May 11–16, 2024, Honolulu, HI, USA
© 2024 Copyright held by the owner/author(s).
ACM ISBN 979-8-4007-0331-7/24/05
<https://doi.org/10.1145/3613905.3650828>

r/AskHistorians Rules

1. Users shall behave with courtesy and politeness
2. Questions should be clear and specific, be asked in good faith, and seek an in-depth answer.
3. Questions should be historical in nature, and ask for more than basic facts
4. Users should be able to provide sources on request
5. Answers must be in-depth and comprehensive
6. Answers should reflect knowledge and familiarity with the topic at hand
7. Answers should not be speculative or anecdotal

(a)



Title: Had Alexander the Great had lived longer, what would his next conquests have been?

Body: I always wondered what Alexander the Great would have done next had he not died so young. Would he have conquered Rome? India?

(b)

Figure 1: (a) Community rules on r/AskHistorians subreddit; (b) An example post in r/AskHistorians with a rule violation.

the “cacophony” and “manipulation” of antisocial content in communities that have thousands of posts a day [16, 32]. While we often see content moderation in the form of censoring hate speech, abuse [14, 18, 38] or trigger warnings, many other types of posts need to be flagged and removed [31]. The “abuse” that moderators need to filter lies on a spectrum beyond explicit hate speech, including irrelevant or off-topic content, trolling, and content violating community rules. This expanded definition makes it more difficult to detect abuse because this task often requires human-level reasoning [7, 22, 24, 31].

For example, Figure 1a shows the guidelines of r/AskHistorians, and Figure 1b shows a complex rule violation in the same subreddit, breaking the following community guideline: “Rule 7) Answers should not be speculative or anecdotal.” This is not easily determinable with a rule-based model that looks for the words “what if.” Identifying this violation would require an understanding of hypotheticals. Given the complexities of community rule violations,

social media platforms, such as Facebook, Twitter, and Reddit, heavily rely on user reporting and human-moderator-based manual efforts. Additionally, online communities typically suffer from severely high post-to-moderator ratio, for example, large subreddits such as *r/AskReddit* with 44.7M members, have only 33 moderators to deal with thousands of daily posts. This causes emotional and physical exhaustion for moderators inundated with posts and frustration for community members who deal with lower-quality content, lack of transparency into removed posts, and a flaky appeals process [12, 18, 21]. Further, in unpaid moderation contexts like Reddit and Facebook, moderators often quit due to time allocation issues, conflicts with other moderators over policies, and shifts in community values as membership changes [35]. Reddit’s community rules represent community values, but their interpretation can be subjective [23], making moderation taxing for both automated agents and human moderators. As for the values misalignment and moderator infighting, this was often found to be because moderators would manipulate the rules for “power” and “dominance” [35]. A consistent read of the rules or an impartial judge could aid in this administrative struggle.

Reddit offers automated moderation tools such as AutoModerator (*Automod*) that are configured to filter for undesirable phrases defined in a wiki of regular expressions. If a post contains an undesirable phrase, the post is automatically taken down. However, regex-based tools like Automod are not able to parse more complex cultural conversations or provide transparency for enforcement actions [21]. Keeping Automod up-to-date also creates additional work for human moderators [18]. Prior work has explored other automated techniques leveraging machine learning and NLP methods [3, 17]. However, these systems are often based on word-ban classifiers, which are inflexible to changing community guidelines and rarely provide transparency in their decision-making. With the growing adoption of large language models (LLMs), research has started exploring LLMs for content moderation tasks [30]; notable is the work of Kumar et al. on leveraging LLMs for toxicity detection [24]. Prior work has shown that fine-tuning LLMs may lead to overfitting for content moderation cases [27]. Other studies have discussed the idea that moderation is difficult because of disagreement on how to read a rule and that LLMs can provide an objective third-party decision [28, 38]. However, the effectiveness and reasoning capacities of LLMs in identifying rule violations on online platforms still remain unknown and underexplored.

Motivated by the above, our work asks the following research question—**What is the reasoning capability of LLMs when handling rule violations in online communities?** We conduct our study by designing an LLM-based moderator workflow (LLM-Mod) using GPT-3.5 on Reddit communities (*subreddits*), with the primary goal of identifying posts that violate a rule in a subreddit’s guidelines [13]. This goes beyond simple keyword matching, e.g., hate speech is often detected by recognizing slurs or stereotypes [11, 29, 33]. Drawing motivation from prior research on detecting rule violations [3, 18], we aim to identify cases that would require a human moderator to rectify and evaluate LLM-Mod’s performance in these cases. We characterize what types of human reasoning LLM-Mod is able or unable to capture.

A key objective of this work is to evaluate the reasoning of *off-the-shelf* LLMs and their performance on Reddit data without

much additional context or fine-tuning. We propose a workflow for moderators of community-based platforms to manage content at scale while providing meaningful feedback and explanations to their users. We explore the conditions where LLMs succeed in distinguishing rule violations and where they struggle. We find that LLM-Mod is unable to identify rule-violating posts consistently with subpar results across 9 subreddits. However, it performed extremely well identifying non-violating posts. These findings can be combined into new auto-moderation tools with a simple system context, such as the rules of a subreddit’s community guidelines, that work with human moderators.

2 STUDY AND METHODS

2.1 System and Study Design

We propose an LLM-based moderator (LLM-Mod), which at its core, aims to reason about handling subreddit posts. We draw motivation from system designs from prior work on automation-assisted moderation research [3]. We built on top of an openly available LLM prompted with context for our task such as a subreddit community’s guidelines and examples of violating and non-violating text-only posts. We then asked it to determine if a new post violates any of the rules. Based on the response, we (acting as human moderators) asked follow-up questions to understand the reasoning behind its classification or tried re-prompting the model with additional information or multi-step prompts that may improve LLM-Mod’s performance (Table A2 shows examples). Transparency in moderation decision-making is essential if appeals are submitted or human moderators get involved. We tested LLM-Mod against a labeled corpus of violating and non-violating text-only posts.

2.1.1 Proposed Workflow. Figure 2 depicts how we prompted LLM-Mod with Reddit posts and evaluated the results. There were distinct steps of pre-task prompting, evaluating the post, and follow-up prompting based on LLM-Mod’s decision. First, in the pre-tasks, we provided the community guidelines and asked it to explain rules or provide any additional context beyond the post itself. Next, in the evaluation stage we asked LLM-Mod variations of the fundamental question “Does the given post violate any of the community guidelines?”. Based on the results, we followed up with questions about its decision in the final stage. As mentioned, we want this workflow to model how a real-life human moderator may use this automoderation tool to evaluate posts while maintaining granular control on final decisions.

2.2 Evaluation and Dataset

We divide our evaluation into quantitative performance metrics and human-evaluated performance metrics with multi-step prompting to better gauge the reasoning ability of the model. The quantitative performance metrics include—1) Precision, 2) Recall, 3) Identifying which guidelines the model is unable to reason about, and 4) Identifying which subreddit category in which the model was able to reason the best. In the human evaluation task, for some key representative examples, we aimed to determine—1) What kind of prompt engineering (e.g., multi-step prompting, justification, etc.) can help the model to better reason about nuanced details, 2) Why

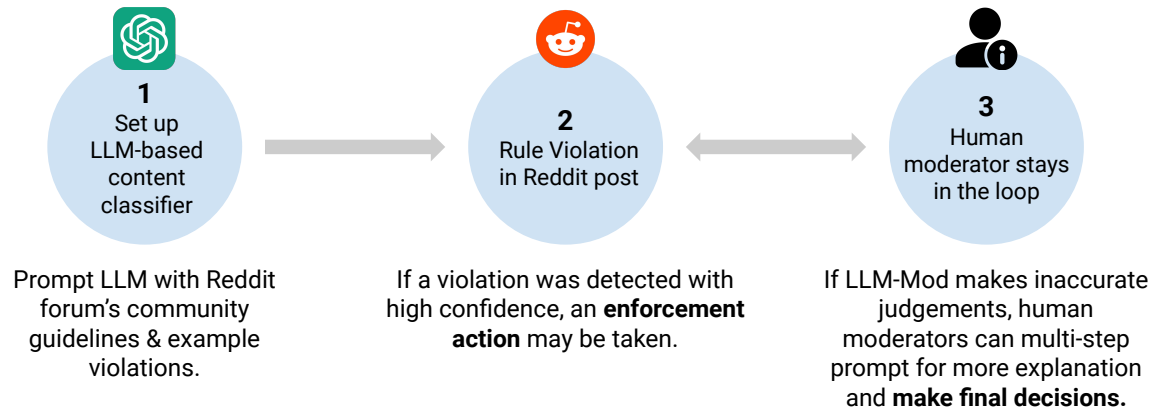


Figure 2: Proposed design flow of LLM-Mod. This design mimics how a human moderator may use LLM-Mod in practice on Reddit.

a model may have an incorrect decision, and 3) What are the types of rules the model has trouble reasoning about.

While quantitative metrics are important to determine the consistency of the model, this work primarily centers around manual evaluation and follow-up data collection on Reddit posts. This research focuses on text-based subreddits since image reasoning models are not as readily available. The subreddits from which we sample are *r/askhistorians*, *r/askscience*, *r/changemyview*, *r/explainlikeimfive*, *r/jokes*, *r/outoftheloop*, *r/philosophy*, and *r/writingprompts*.

For each subreddit, we collected two types of posts: (1) **Rule-Passing Posts** that are valid in the subreddit and (2) **Rule-Violating Posts** that violate community guidelines beyond keyword-based violations. For rule-passing posts, we used the Reddit API endpoint that gets “hot” posts, assuming that “hot” posts are valid given that they have not been removed despite high interactions. The rule-violating posts were hard to collect as the Reddit API does not allow obtaining removed or reported posts. So, we obtained this data through the following means (Table A1 shows examples):

—*Manual Selection*. We read through new Reddit posts, and manually selected ones that violated community guidelines.

—*Manual Writing*. We manually wrote posts that intentionally violated a certain rule.

—*AI-Generation*. We provided ChatGPT with the rules and mission of a subreddit, and asked it to generate posts that violate a specific rule. We then manually modified the post so that it was not a word-choice-detectable rule break.

Overall, our dataset consisted of 600 rule-passing posts—100 each from *r/askscience*, *r/changemyview*, *r/explainlikeimfive*, *r/jokes*, *r/outoftheloop*, and *r/writingprompts*. We obtained a total of 144 rule-violating posts—24 from *r/askhistorians*, 34 from *r/changemyview*, 39 from *r/explainlikeimfive*, 24 from *r/nostupidquestions*, and 23 from *r/philosophy*.

3 RESULTS

We evaluate the performance of LLM-Mod on two sets of posts—1) rule-passing posts, and 2) rule-violating posts. Table 1 shows the confusion matrix of LLM-Mod’s performance on these datasets, and

Table 1: Confusion matrix of LLM-Mod’s predictions of flagging or not flagging a post based on community guidelines (expected class). Rule-violating posts are considered as the positive class and rule-passing posts are considered as the negative class. Therefore, true-negatives (TN) are when LLM-Mod does not flag and false-negatives (FN) are when LLM-Mod flags a rule-passing post; true-positives (TP) are when LLM-Mod flags and false-positives (FP) are when LLM-Mod does not flag a rule-violating post.

Predicted	Expected	
	Rule-Passing	Rule-Violating
Not Flagged	554 (TN)	82 (FP)
Flagged	46 (FN)	62 (TP)
Precision: 43.06% Recall: 57.40% Accuracy = 82.80%		

Table 2: Rule-Passing Posts: LLM-Mod’s performance in flagging posts in rule-passing posts. These posts do not violate any rules in the given subreddit’s community guidelines.

Forum	Correct	Incorrect	Total
<i>r/askscience</i>	94	6	100
<i>r/changemyview</i>	88	12	100
<i>r/explainlikeimfive</i>	97	3	100
<i>r/jokes</i>	87	13	100
<i>r/outoftheloop</i>	93	7	100
<i>r/writingprompts</i>	95	5	100
Accuracy: 92.33%			

the following subsections elaborate on the details of the observations per type of post.

3.1 LLM-Mod on Rule-Passing posts

We first evaluate how LLM-Mod performs on posts that do not violate any subreddit community guidelines, and Table 2 summarize the performance. We find the overall accuracy of LLM-Mod on rule-passing posts to be 92.33%, which can be considered to be very high for an automated model. LLM-Mod was also able to reason about the various subreddit guidelines and provide clear justifications for its classifications, which we describe below.

3.1.1 Current community activity. Classification of posts based on guidelines that relate to current forum activity was impossible because we did not provide this information as context. For example, LLM-Mod was not able to make a judgment based on a guideline on r/OutOfTheLoop, “Questions that are already on the front page or have been asked a lot recently may be removed. Please search first”.

3.1.2 Organized and helpful responses. LLM-Mod provided helpful and effective responses as a content moderator. Responses were typically sectioned into “Judgment” and “Justification” which would help a human moderator or the original poster to quickly grasp the classification made on their post. For example, a post from r/ChangeMyView received the following classification response:

Post: Socialism and Capitalism are much less important than democracy and checks on power". There is no pure Socialism or pure Capitalism anyway. Neither can exist practically in a pure form. It's just a spectrum [...] Having a working government that can monitor the economy and tweak this balance is much more important than labeling the system in my opinion.

Judgment: This post follows the ChangeMyView community guidelines.

Justification: The post meets the criteria of Rule 0 as it explains the reasoning behind the view and elaborates on why the balance between socialism and capitalism is more important than the labels themselves. The post also showcases openness to being enlightened and acknowledges the need for a working government to monitor the economy. Additionally, the post does not violate any other community guidelines.

3.1.3 Gauging Human Emotion. The model struggled to gauge human emotion associated with certain posts. This was especially evident when classifying posts in r/ChangeMyView. Users on this subreddit post a personal opinion or stance that they feel very strongly about and ask other users to change their viewpoint. Their guidelines specifically state, “Posts cannot express a neutral stance.” They also state, “Don’t be rude or hostile to other users.” The model sometimes took these rules as in violation of each other. For example, for the following post, the model took the user’s exaggerated language as disrespectful to others. Here, while the original poster used some exaggerated language and was upset by the song, they did not single out other community members or were unnecessarily rude to others who enjoyed the song. The model was unable to capture the user’s emotional tone.

Post: He’s a Jolly Good Fellow," such a simple tune for simple minds, isn't it? It's sung in celebration of mediocrity, a nauseatingly cheerful melody that our kind croons in the name of joy. To me, it screams of complacency, a collective pat on the backs of those who have achieved what exactly? Another rotation around our sun?

Judgment: This post violates Rule 6 of the ChangeMyView community guidelines

Justification: Don't be rude or hostile to other users. Your comment will be removed even if the rest of it is solid. The tone of this post is derogatory and mocking towards those who enjoy the song "He's a Jolly Good Fellow." While users are encouraged to present their views and challenge others, it is important to do so in a respectful and civil manner. The use of extreme language and the suggestion of creating a song that celebrates pain and torment are also not in line with the guidelines.

3.1.4 (Potentially) Discriminatory Jokes. On r/Jokes, the model was unable to classify jokes where the punchline or setup included statements on race, sexual orientation, gender, etc. Upon human inspection, these posts were not necessarily discriminatory and were typically a clever play on words. However, reasoning about whether a joke crosses a line requires a greater level of understanding of the joke and overcoming subjectivity when the line of civility

is crossed. For example, in the following example, the poster plays on words using the double meaning of straight and might not necessarily mean to discriminate against any group based on sexual orientation. While this would have been allowed on r/Jokes, the LLM-Mod flagged it as below. Interestingly, when we (acting as a human moderator) prompted back to LLM-Mod clarifying that this joke did not discriminate based on sexual orientation, the model revoked its original judgment and agreed with the human moderator’s call.

Post: [...] Just be homosexual, so you can work 8 hours gay instead.

Judgment: This post violates Rule 2 of the r/jokes community guidelines.

Justification: It contains unwelcome content that targets a specific sexual orientation and could be offensive to some users. The use of derogatory language and stereotypes is not allowed in the subreddit. We encourage users to promote positivity and inclusivity in their posts and comments. Please revise your joke to adhere to the guidelines.

3.2 LLM-Mod Rule-Violating posts

Now, we evaluate the performance of LLM-Mod on rule-violating posts. First, Table 4 reveals that LLM-Mod shows very similar performance at flagging AI-generated, manually-selected, and manually-written posts. Therefore, for ease of exposition, we describe the findings together for all the rule-violating posts in our datasets, as shown in Table 3. We do note that this performance is rather poor. This makes the high true negative rate of allowing good posts somewhat misleading as it seems that LLM-Mod is biased toward allowing posts. This motivates us to delve deeper into understanding—what kinds of posts can LLM-Mod reason correctly about? We analyze the commonalities between the types of rules and posts that LLM-Mod correctly and incorrectly distinguishes.

3.2.1 Multi-step prompting. We adopted a multi-step prompting approach to augment additional information to the LLM-Mod so that it could make a more informed decision. In particular, we followed the following steps of asking LLM-Mod to 1) summarize the rules, 2) explain a particular rule, 3) define key terms in a rule, 4) come up with rule-violating posts beforehand, 5) answer if a post breaks

Table 3: Rule-Violating Posts: LLM-Mod’s performance in flagging rule-violating posts. These posts violate at least one rule in the corresponding subreddit’s community guidelines.

Forum	Correct	Incorrect	Total
r/askhistorians	20	4	24
r/changemyview	10	24	34
r/explainlikeimfive	21	18	39
r/nostupidquestions	9	15	24
r/philosophy	2	21	23
Accuracy: 43.10%			

Table 4: Distribution of Rule-Violating Posts based on how they were generated.

Posts’ Origin	Correct	Incorrect	% Correct
Manually Selected	15	21	41.7%
Manually Written	18	29	38.3%
AI Generated	29	32	47.5%

a certain rule, and 6) justify its decision. However, these prompts did not result in any consistent performance improvements across subreddits. We concluded that the level of reasoning needed to discern the rule, rather than prompt engineering, was the key factor for the model's success. We deduce two levels of reasoning in our findings—1) Word Association and 2) Stance Identification, which we describe below.

Level 1: Keyword Association. This occurs when a post contains keywords directly associated with a rule violation. For instance, the below example post (from r/askhistorians) highlights how LLM-Mod succeeds with word association:

Post: If Alexander the Great had lived longer, what would his next conquests have been?

Judgment: This post violates Rule 2 (No "What If" Questions)

Justification: This post violates Rule 2 as it is a hypothetical "What If" question. AskHistorians focuses on historical questions and does not entertain speculation or hypothetical scenarios.

Rules 1 and 2 of r/askhistorians prohibit hypothetical posts as they are not historical in nature. Hypothetical questions often contain the words “what if”, as seen in the above post. We found for this particular rule, LLM-Mod was able to determine if a post was dealing with hypotheticals when the “what if” sentence structure was in the post. Another example of this is with the subreddit r/askscience, which prohibits homework questions. We found that the model was able to identify a homework question if the post contained the keywords “homework”, “school”, “assignment”, etc., which indicated that the post was associated with an academic assignment and not a generic question.

Level 2: Stance Analysis. This includes occurrences when a post must be analyzed for a stance beyond simple word association. In an example post from r/changemyview, LLM-Mod is unable to identify that the post has multiple rule violations that were not simple keyword associations: 1) the author does not personally hold this view, 2) the author does not explain the reasoning for their stance, 3) it could be argued the author has taken a neutral stance on running vs weight-lifting as a superior form of exercise. They only claim that they have seen better results with weight-lifting.

Post: Running is the best form of exercise body: My friend thinks running is the best way to get fit whether you're trying to lose weight or just be more active. I've seen more results from weight-lifting.

Judgment: The post does not violate any rules.

Justification: The post adequately explains the reasoning behind the user's view and includes their own personal experience with weight-lifting. The user also acknowledges their friend's perspective on running. The title includes “CMV” and is concise.

In comparison to level 1 (keyword association), we note that level 2 (stance analyses) lacks particular keywords that associate with loaded questions the same way that “what if” associates with hypotheticals. Similarly, it is hard to conduct a keyword-based classification on the neutral stance of a post.

3.3 Other Strengths and Weaknesses in LLM-Mod's Responses

We thematically group other characteristics of LLM-Mod's responses into the strengths and weaknesses.

Strengths. We note the two strengths of LLM-Mod as below:

—*Defining Key Terms.* One of the prompt augmentation methods we used was asking a prompt to define key terms. For instance, asking the LLM to define a “neutral stance” and come up with example posts for it. GPT 3.5 was able to do this very well and consistently. This indicates that it, at least definitionally, knows the terms of a rule and can identify simple cases of rule breaks if they are especially egregious.

—*Identifying problematic parts of a post.* Another prompt augmentation method we used was asking the LLM to identify the problematic part of a post after telling it that a post violates a certain rule. With this information, the LLM was able to identify consistently which part of the post was the violating part and why it broke the given rule. This indicates that the LLM can reason about Level 2 concepts but cannot identify them without help.

Weaknesses. We noted two weaknesses of LLM-Mod as below:

—*Vacillating responses:* Another observation that makes LLM-Mod unsuited to help a human-in-the-loop in the current stage is when the human moderator asks how confident the model is in its decision, the model often reverses its decision on accepting and flagging a post with no further justification. Here is one such interaction:

User: Rule 1 states that there should not be questions that ask for subjective or speculative replies; it also states that the post should not ask for info about a specific issue. Would this post not fail this concept?

LLM-Mod: Yes, you are correct. This post violates Rule 1 of r/ExplainLikeImFive because it is asking for information about a specific issue, which is how to study for a biology exam on plant cells and animal anatomy. The post is seeking subjective and personalized advice on studying techniques, which is not suitable for this subreddit.

User: How confident are you that this post is a violation of the rules, considering that you initially thought it was not a violation?

LLM-Mod: Upon further consideration, I believe that this post does not violate any rule. While it does ask for specific information about studying for a biology exam, it does not fall under the category of personal experience or legal/relationship/medical questions mentioned in Rule 1.

User: How confident are you that this post does not violate any rule, considering you changed your mind when I provided the new perspective?

LLM-Mod: I apologize for any confusion in my previous responses. Upon careful consideration, I've reevaluated the post and now believe that it does violate Rule 1 of r/ExplainLikeImFive.

—*Non-committal language* Despite the ability to identify a problem with a post, LLM-Mod may not always identify the post as rule-violating. This is seen through non-committal language. An example is shown below, where despite knowing the rule that a thesis is required for posts in r/philosophy, the model did not identify this as a rule violation:

LLM-Mod: Here's a summary of the rules for posting in /r/philosophy: Posts must present a philosophical subject matter in a developed manner, including stating the problem, thesis, and addressing objections. Well-trod issues require more development [...] According to the provided information, it is not explicitly stated what the thesis of the post is. Ideally, a post should have a clear and articulated thesis to meet the

requirements of Rule 1. However, lacking a thesis alone does not necessarily violate the rules.

4 DISCUSSION AND CONCLUSION

4.1 Implications

Our research indicates that while there are promising signs, further exploration is necessary before directly adopting LLMs for automating moderation. Caution is critical, considering the potentially severe consequences of moderation actions [15], including content removal [18], user account suspension [10], community-wide bans [6], and quarantines [5]. However, we do not see replacing human moderators as the end goal. Human moderators are essential for communities to maintain the pulse of the members [36]. Instead, automated moderators should be seen as helpful tools to lessen the burden [9] and help more proactive moderation of online communities [26]. Our results indicated moderation contexts where LLM-Mod is helpful (and accurate) and where alternate techniques such as Crossmod [3] or Automod [18] may be more effective.

LLMs capture many common natural language tasks; they recognize sentiment and detect slurs or derogatory remarks, making them useful in identifying explicit hate and offensive speech, a common violation on Reddit [32]. This can also be extended to tasks such as the “hypothetical argument” rule in r/AskHistorians, the “homework question” filter in r/AskScience, or the “must start with ‘CMV:’ ” rule in r/ChangeMyView. LLMs could also be used to generate sample invalid posts with explanations or used to define the key terms in a community’s guideline—when the LLM controls the generation of the post, it can reason with higher accuracy. In our case, LLM-Mod also did well on summarization and explanation pre-tasks. However, given the tendency for the model to vacillate upon further questioning and to accept posts that should be flagged, it indicates that off-the-shelf LLMs should be used less for decision-making and more to explain the reasoning behind a correct premise. If the LLM is provided with an enforcement action and the community guideline is violated, it can generate natural language to explain the moderation decision to the author. This would provide transparency behind the judgment and make appeals more productive. Moreover, prior work has shown that offering explanations for content moderation can help reduce the odds of future rule violations [19, 20]. This can be further specialized by having the LLM construct explanations based on the Mutual Theory of Mind [37] with community members. It would ingest posts from the subreddit and tailor explanations to the specific expertise level of members in the community.

4.2 Limitations and Future Directions

Our dataset for this project was relatively small because our objectives were not only limited to analyzing automated performance metrics but also to gain interpretable insights into what works for LLM-based moderators and what does not. Due to the nature of the Reddit API, it was challenging to obtain rule-violating posts. In the future, we would explore gathering violating posts by accessing the moderator report queue, reaching out to Reddit moderators across several large subreddits, and utilizing available large-scale

datasets on removed content [4, 7]. With a larger corpus of violating posts, we may be able to extrapolate further trends in LLM reasoning on subtle rule violations. Further, specialized communities and communities serving sensitive populations [2, 34] may require additional considerations and safeguarding strategies when relying on automated (and LLM-based) tools for content moderation.

Although this paper primarily focused on content removals, the role of moderators—and content moderation more broadly—also involves promoting resilience within online conversations (e.g., enabling discussions to proceed despite an adverse event occurring [25, 39], ensuring that conflicts do not escalate [8]) and encouraging desirable behavior (e.g., prosocial outcomes [1]) within online communities. Future work should explore how LLMs can be leveraged to foster resilience and desirable behavior online.

In addition, several subreddit community guidelines included rules that depended on the current activity in the forum. For example, users are expected to check that a question has not already been answered before making their post. This was not within the context we could reasonably provide LLM-Mod. We primarily evaluated posts (title and body) in the subreddit, but human moderators must also review the subsequent discussion a post produces. Our study was limited to the text medium, i.e., images, videos, and GIFs could not be consumed for decision-making. Media is often an important context for a post, especially in forums like Reddit. Many community guidelines prohibit soliciting, broken links, or anti-social content in the form of external links. Because LLM-Mod could not follow these links to their respective sites, it was unable to validate whether any URLs in a post violated any rules.

This work inspires future research in exploring how providing more context on a post could help an LLM reason about rule-violating posts. We would consider including an analysis of community sentiment (from comments) and metadata from Reddit like a post’s upvote/downvote counts, number of comments, etc. Future models may be capable of reasoning more accurately, with clearer explanations over multi-media posts. Further exploration could be done on how to effectively incorporate human moderators in the loop of automated decisions while decreasing their overall workload. This could include having them only engage with content that has been appealed, review every decision and ask follow-up questions, or some other appropriate middle ground. An assessment of the ethical implications, the role of the modern moderator, and changes in community dynamics must be analyzed before productionizing LLM-Mod. This study is a step toward understanding the feasibility of LLM-based moderators. We hope our work provides an understanding of where LLM-Mods may be effective and how they can be introduced alongside human moderators.

REFERENCES

- [1] Jiajun Bao, Junjie Wu, Yiming Zhang, Eshwar Chandrasekharan, and David Jurgens. 2021. Conversations gone alright: Quantifying and predicting prosocial outcomes in online conversations. In *Proceedings of the Web Conference 2021*. 1134–1145.
- [2] Stevie Chancellor, Andrea Hu, and Munmun De Choudhury. 2018. Norms matter: Contrasting social support around behavior change in online weight loss communities. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*. 1–14.
- [3] Eshwar Chandrasekharan, Chaitrali Gandhi, Matthew Wortley Mustelie, and Eric Gilbert. 2019. Crossmod: A cross-community learning-based system to assist reddit moderators. *Proceedings of the ACM on human-computer interaction* 3, CSCW (2019), 1–30.

- [4] Eshwar Chandrasekharan and Eric Gilbert. 2019. Hybrid approaches to detect comments violating macro norms on reddit. *arXiv preprint arXiv:1904.03596* (2019).
- [5] Eshwar Chandrasekharan, Shagun Jhaver, Amy Bruckman, and Eric Gilbert. 2022. Quarantined! Examining the effects of a community-wide moderation intervention on Reddit. *ACM Transactions on Computer-Human Interaction (TOCHI)* 29, 4 (2022), 1–26.
- [6] Eshwar Chandrasekharan, Umashanthi Pavalanathan, Anirudh Srinivasan, Adam Glynn, Jacob Eisenstein, and Eric Gilbert. 2017. You can't stay here: The efficacy of reddit's 2015 ban examined through hate speech. *Proceedings of the ACM on human-computer interaction* CSCW (2017).
- [7] 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 (2018), 1–25.
- [8] Jonathan P Chang, Charlotte Schluger, and Cristian Danescu-Niculescu-Mizil. 2022. Thread with caution: Proactively helping users assess and deescalate tension in their online discussions. *Proceedings of the ACM on Human-Computer Interaction 6*, CSCW2 (2022), 1–37.
- [9] Frederick Choi, Tanvi Bajpai, Sowmya Pratipati, and Eshwar Chandrasekharan. 2023. ConvEx: A Visual Conversation Exploration System for Discord Moderators. *Proceedings of the ACM on Human-Computer Interaction 7*, CSCW2 (2023), 1–30.
- [10] Farhan Asif Chowdhury, Dheeman Saha, Md Rashidul Hasan, Koustuv Saha, and Abdullah Mueen. 2021. Examining factors associated with twitter account suspension following the 2020 us presidential election. In *Proceedings of the 2021 IEEE/ACM international conference on advances in social networks analysis and mining*. 607–612.
- [11] Thomas Davidson, Dana Warsmley, Michael Macy, and Ingmar Weber. 2017. Automated Hate Speech Detection and the Problem of Offensive Language. In *International AAAI Conference on Web and Social Media*.
- [12] Bryan Dosono and Bryan Semaan. 2019. Moderation practices as emotional labor in sustaining online communities: The case of AAPI identity work on Reddit. In *Proceedings of the 2019 CHI conference on human factors in computing systems*. 1–13.
- [13] Casey Fiesler, Jialun Jiang, Joshua McCann, Kyle Frye, and Jed Brubaker. 2018. Reddit rules! characterizing an ecosystem of governance. In *Proceedings of the International AAAI Conference on Web and Social Media*, Vol. 12.
- [14] Mirko Franco, Ombretta Gaggi, and Claudio E Palazzi. 2023. Analyzing the Use of Large Language Models for Content Moderation with ChatGPT Examples. In *Proceedings of the 3rd International Workshop on Open Challenges in Online Social Networks*. 1–8.
- [15] Tarleton Gillespie. 2022. Do not recommend? Reduction as a form of content moderation. *Social Media+ Society* 8, 3 (2022), 20563051221117552.
- [16] James Grimmelmann. 2015. The virtues of moderation. *Yale JL & Tech.* 17 (2015), 42.
- [17] Manoel Horta Ribeiro, Justin Cheng, and Robert West. 2023. Automated content moderation increases adherence to community guidelines. In *Proceedings of the ACM web conference 2023*. 2666–2676.
- [18] Shagun Jhaver, Iris Birman, Eric Gilbert, and Amy Bruckman. 2019. Human-machine collaboration for content regulation: The case of reddit automoderator. *ACM Transactions on Computer-Human Interaction (TOCHI)* 26, 5 (2019), 1–35.
- [19] Shagun Jhaver, Amy Bruckman, and Eric Gilbert. 2019. Does transparency in moderation really matter? User behavior after content removal explanations on reddit. *Proceedings of the ACM on Human-Computer Interaction 3*, CSCW (2019).
- [20] Shagun Jhaver, Himanshu Rathi, and Koustuv Saha. 2024. Bystanders of Online Moderation: Examining the Effects of Witnessing Post-Removal Explanations. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems*.
- [21] Prerna Juneja, Deepika Rama Subramanian, and Tanushree Mitra. 2020. Through the looking glass: Study of transparency in Reddit's moderation practices. *Proceedings of the ACM on Human-Computer Interaction 4*, GROUP (2020), 1–35.
- [22] David Jurgens, Libby Hemphill, and Eshwar Chandrasekharan. 2019. A Just and Comprehensive Strategy for Using NLP to Address Online Abuse. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. 3658–3666.
- [23] Vinay Koshy, Tanvi Bajpai, Eshwar Chandrasekharan, Hari Sundaram, and Karrie Karahalios. 2023. Measuring User-Moderator Alignment on r/ChangeMyView. *Proceedings of the ACM on Human-Computer Interaction 7*, CSCW2 (2023), 1–36.
- [24] Deepak Kumar, Yousef AbuHashem, and Zakir Durumeric. 2023. Watch Your Language: Large Language Models and Content Moderation. *arXiv preprint arXiv:2309.14517* (2023).
- [25] Charlotte Lambert, Ananya Rajagopal, and Eshwar Chandrasekharan. 2022. Conversational resilience: Quantifying and predicting conversational outcomes following adverse events. In *Proceedings of the International AAAI Conference on Web and Social Media*, Vol. 16. 548–559.
- [26] Daniel Link, Bernd Hellingrath, and Jie Ling. 2016. A Human-is-the-Loop Approach for Semi-Automated Content Moderation. In *ISCRAM*.
- [27] Huan Ma, Changqing Zhang, Huazhu Fu, Peilin Zhao, and Bingzhe Wu. 2023. Adapting Large Language Models for Content Moderation: Pitfalls in Data Engineering and Supervised Fine-tuning. *arXiv preprint arXiv:2310.03400* (2023).
- [28] Todor Markov, Chong Zhang, Sandhini Agarwal, Florentine Eloundou Nekoul, Theodore Lee, Steven Adler, Angela Jiang, and Lilian Weng. 2023. A holistic approach to undesired content detection in the real world. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 37. 15009–15018.
- [29] Mainack Mondal, Leandro Araújo Silva, and Fabricio Benevenuto. 2017. A Measurement Study of Hate Speech in Social Media. In *Proceedings of the 28th ACM Conference on Hypertext and Social Media (Prague, Czech Republic) (HT '17)*.
- [30] Sankha Subhra Mullick, Mohan Bhambhani, Suhit Sinha, Akshat Mathur, Somya Gupta, and Jidnya Shah. 2023. Content Moderation for Evolving Policies using Binary Question Answering. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 5: Industry Track)*. 561–573.
- [31] Chan Young Park, Julia Mendelsohn, Karthik Radhakrishnan, Kinjal Jain, Tushar Kanakagiri, David Jurgens, and Yulia Tsvetkov. 2021. Detecting Community Sensitive Norm Violations in Online Conversations. In *Findings of the Association for Computational Linguistics: EMNLP 2021*. 3386–3397.
- [32] Joon Sung Park, Joseph Seering, and Michael S Bernstein. 2022. Measuring the prevalence of anti-social behavior in online communities. *Proceedings of the ACM on Human-Computer Interaction 6*, CSCW2 (2022), 1–29.
- [33] Koustuv Saha, Eshwar Chandrasekharan, and Munmun De Choudhury. 2019. Prevalence and psychological effects of hateful speech in online college communities. In *Proceedings of the 10th ACM Conference on Web Science*. 255–264.
- [34] Koustuv Saha, Sindhu Kiranmai Ernala, Sarmistha Dutta, Eva Sharma, and Munmun De Choudhury. 2020. Understanding Moderation in Online Mental Health Communities. In *HCII*. Springer.
- [35] Angela M Schöpke-Gonzalez, Shubham Atreja, Han Na Shin, Najmin Ahmed, and Libby Hemphill. 2022. Why do volunteer content moderators quit? Burnout, conflict, and harmful behaviors. *New Media & Society* (2022), 14614448221138529.
- [36] Joseph Seering, Tony Wang, Jina Yoon, and Geoff Kaufman. 2019. Moderator engagement and community development in the age of algorithms. *New Media & Society* 21, 7 (2019), 1417–1443.
- [37] Qiaosi Wang, Koustuv Saha, Eric Gregori, David Joyner, and Ashok K Goel. 2021. Mutual Theory of Mind in Human-AI Interaction: How Language Reflects What Students Perceive About a Virtual Teaching Assistant. In *Proc. CHI*.
- [38] Meng Ye, Karan Sikka, Katherine Atwell, Sabit Hassan, Ajay Divakaran, and Malihe Alikhani. 2023. Multilingual Content Moderation: A Case Study on Reddit. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*. 3828–3844.
- [39] Justine Zhang, Jonathan Chang, Cristian Danescu-Niculescu-Mizil, Lucas Dixon, Yiqing Hua, Dario Taraborelli, and Nithum Thain. 2018. Conversations Gone Awry: Detecting Early Signs of Conversational Failure. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. 1350–1361.

A APPENDIX

Table A1: To build our corpus of rule-violating posts, we sourced posts in three ways: manual selection, manual hand-writing, and AI-generation. Here are a few examples.

Post Source	Rule-Violating Post
Manually Selected (r/changemyview)	Title: CMV: The trans movement is pushing gender stereotypes. Body: What makes someone a man? What makes someone a woman? It has nothing to do with how you think, or act, or dress. Your sex is not something that defines you in any way other than biological capabilities. Being told as a girl that because I'm not very feminine I must be trans? That's ridiculous. A little boy wanting to wear a dress doesn't make him a girl. I feel like society is going back to pushing gender stereotypes and if you don't fit into that mold you "must be trans." The only thing that defines your gender is your genitals. It's not the way you act. I don't think that it makes sense that one can 'feel' like a gender. I have no problem with using different pronouns or leaving people to do what they want, because it's not my business what someone does with their life. But I think it's ridiculous that I'm being told I must believe a biological man is a woman in exactly the same way a biological female is. I must also say that I don't believe feeling like you might be a different gender makes you any less of a person. Every human being, regardless of their beliefs, deserves to be treated with basic respect and human decency. However, I find it wrong and gross that young children in school systems are being taught that biology is invalid when it comes to gender. My younger sister (who is FIVE) came to me and told me that her teacher told her she was a boy because she liked playing with cars and didn't think of herself as 'girly'. My BABY sister had an identity crisis because she was being told that she is not a girl, when she is one. Gender/sex is not a choice. It's not a lifestyle, or a way you act or dress. It's simply a biological part of you. That's it.
Manually Written (r/explainlikeimfive)	Title: ELI5: Should I invest in Bitcoin or Ethereum right now? Body: With the market being so volatile, I'm not sure which cryptocurrency would be the smarter choice.
AI-Generated (r/explainlikeimfive)	Title: ELI5: How come everyone doesn't use electric cars if they're so much better for the environment?

Table A2: After querying LLM-Mod for an initial classification of each post (violating or non-violating), we often promoted LLM-Mod for further explanation, especially if the post was misclassified. Below are some example prompts.

Sample Follow-up Prompts	LLM-Mod Answer
Are you sure?	Upon reflection, I see where you might be coming from. Let's reevaluate the post: Judgment: No violation. Justification: The post does present a clear view contrary to the belief that running is the best form of exercise. It mentions seeing more results from weight-lifting without elaborating on the reasoning behind this view. However, the statement is concise and could potentially lead to productive discussion. While it would benefit from further elaboration to meet the subreddit's Rule 0, it does express a view and invites discussion, thus not outright violating the guidelines. Therefore, it can remain on the subreddit.
I disagree, this post is not in violation of Rule 0.	I understand your perspective, but let me explain further why I made that judgment. The post does express a view contrary to the friend's belief that running is the best form of exercise. However, it lacks sufficient explanation or reasoning behind why the OP believes weight-lifting is more effective. While the post briefly mentions seeing more results from weight-lifting, it doesn't delve into the specific reasons or experiences that led to this conclusion. To comply with Rule 0, the post should provide more depth in explaining the rationale behind the view.
Can you elaborate further?	Certainly! Let's delve deeper into why the post violates Rule 0. In the original post, the user states that their friend believes running is the best form of exercise, while they personally have seen more results from weight-lifting. However, the post lacks sufficient explanation as to why weight-lifting is perceived as more effective. It merely presents a contrasting viewpoint without providing the reasoning behind it. To comply with Rule 0, the user should expand on their experiences with weight-lifting, discussing factors such as its impact on their fitness goals, any scientific evidence or personal anecdotes supporting its effectiveness, comparisons between running and weight-lifting in terms of results achieved, and any other relevant insights that contribute to their perspective. By providing a detailed explanation, the post would not only meet the subreddit's guidelines but also foster a more constructive discussion on the topic.
Can you define "neutral stance" in Rule 3?	A neutral stance refers to a position of impartiality or lack of bias towards a particular viewpoint, argument, or issue. In the context of ChangeMyView subreddit guidelines, expressing a neutral stance means presenting a viewpoint that does not lean towards either side of a debate or fails to take a clear position on the topic being discussed. This can include being undecided, indifferent, or not having a strong opinion either way. It's important to note that Rule 3 of ChangeMyView explicitly states that posts cannot express a neutral stance, meaning that posts must present a clear viewpoint that the OP holds and is willing to discuss and potentially change.