# Decoding the Rule Book: Extracting Hidden Moderation Criteria from Reddit Communities

**WARNING: The content contains model outputs that are offensive and toxic.**

**Anonymous ACL submission**

## Abstract

Effective content moderation systems require explicit classification criteria, yet online communities like subreddits often operate with diverse, implicit standards. This work introduces a novel approach to identify and extract these implicit criteria from historical moderation data using an interpretable architecture. We represent moderation criteria as score tables of lexical expressions associated with content removal, enabling systematic comparison across different communities. Our experiments demonstrate that these extracted lexical patterns effectively replicate the performance of neural moderation models while providing transparent insights into decision-making processes. The resulting criteria matrix reveals significant variations in how seemingly shared norms are actually enforced, uncovering previously undocumented moderation patterns including community-specific tolerances for language, features for topical restrictions, and underlying subcategories of the toxic speech classification.

## 1 Introduction

Content moderation is essential for fostering healthy online discourse, but remains challenging due to the diverse and often implicit norms that govern different communities. While platforms like Reddit[1] host numerous micro-communities (subreddits) with distinct norms and moderation practices, the specific criteria used to enforce these norms often remain opaque.

Previous research has examined emergent norms across Reddit communities, identifying both generic norms (shared across many communities) and community-specific norms (Table 1) (Fiesler et al., 2018; Chandrasekharan et al., 2018; Park et al., 2024). However, even when communities share similar norm categories (e.g., "Be civil"), the precise criteria for violations differ between communities based on moderator decisions, topic domains, and user expectations. These criteria are rarely fully captured in stated rules alone, instead requiring substantial domain knowledge and familiarity with community practices.

While existing approaches have leveraged historical moderation data to train classifiers that predict norm violations (Chandrasekharan et al., 2019; Park et al., 2021), these models often function as black boxes, making it difficult to understand the specific patterns being used for classification decisions. This opacity presents significant challenges for practical implementation, as moderation systems must reflect moderators' intent (Kolla et al., 2024) rather than merely detecting superficial features.

In this work, we aim to explicitly discover the implicit criteria used in moderation decisions across Reddit communities. We introduce an approach that extracts and represents these criteria as score tables of lexical expressions associated with content removal. We propose to build this CriteriaMatrix by employing Partial Attention Transformer (PAT) (Kim et al., 2023), an interpretable architecture that can predict moderation probabilities on any lexical expressions. . For each of individual communities, a PAT model is trained with corresponding data, and used to predict moderation scores for a given lexical expression in the corresponding community. This enables comparison of moderation patterns across communities.

Our experiments on the Reddit moderation dataset (Chandrasekharan et al., 2018) show that interpretable model PAT can effectively replicate the performance of neural moderation models, achieving comparable results to ChatGPT. PAT is further used to build analysis using PAT provides various insights of implicit criteria that drive moderation decisions, providing hints of potential risks and future directions for better moderation system.

[1] https://www.reddit.com/

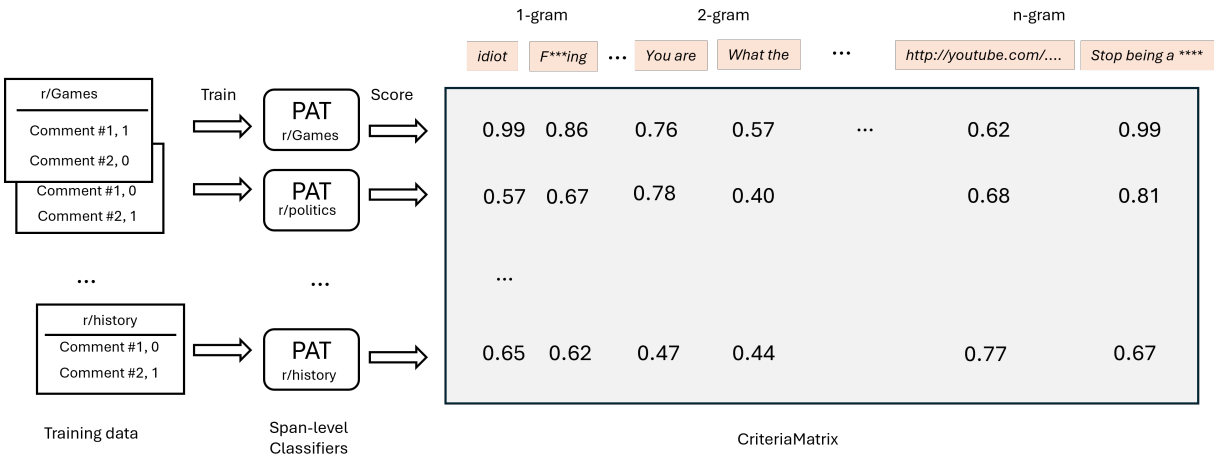


Figure 1: Overview of criteria discovery. For each subreddit, a text classifier (PAT) is trained from past moderation data of comment deletion. PAT models for each subreddit score each term in the vocabulary, allowing us to compare across different subreddits.

Our contributions include: (1) a novel representation of moderation criteria as scores assigned to lexical expressions, providing unambiguous and verifiable insights; (2) demonstrating how PAT can effectively identify global moderation patterns by assigning well-calibrated scores to text spans; and (3) revealing previously unrecognized characteristics of moderation practices across Reddit communities, including varying tolerances for similar content types and community-specific enforcement patterns.

| Violation Category | Example Rule |
|---|---|
| Incivility | *"Be civil"* |
| Harassment, Doxxing | *"Don't harass others"* |
| Spam, Reposting, Copyright | *"No excessive posting"* |
| Format, Images, Links | *"Use the correct tags"* |
| Low-quality, Spoilers | *"No low-quality posts"* |
| Off-topic, Politics | *"Only relevant posts"* |
| Hate Speech | *"No racism, sexism"* |
| Trolling, Personal Army | *"No trolls or bots"* |
| Meta-Voting | *"No downvoting"* |

Table 1: Common subreddit norm violations and paraphrased rules (Park et al., 2021)

## 2 Related works

### 2.1 Criteria extraction for content moderation.

Research on online community standards has identified underlying norms that guide content moderation across platforms (Chandrasekharan et al., 2018; Park et al., 2024; Neuman and Cohen, 2023). These norms typically represent high-level concepts, while community rules provide more explicit, fine-grained enforcement guidance (Park et al., 2021; Rao et al., 2023).

Prior works on norm discovery exhibit several notable limitations. First, many approaches are constrained by limited categories of norms, such as (e.g., supportiveness, politeness, humor) (Park et al., 2024; Goyal et al., 2024), which restricts the discovery of novel norm types.

Second, both norms and rules remain abstract, with the specific criteria used in violation decisions often remaining vague or implicit. Previous approaches to determine whether a given text violates a particular norm have relied on assessments from crowd-workers or generic LLMs (Neuman and Cohen, 2023; Park et al., 2024). However, these assessment methods likely deviate from the actual criteria applied by moderators.

Lastly, the extracted norms are not evaluated if they are actually predictive of moderation outcomes (Chandrasekharan et al., 2018; Neuman and Cohen, 2023), while only parts of extraction pipeline is evaluated.

Interviewing or surveying on actual moderators and users are informative (Lambert et al., 2024; Weld et al., 2024), but they focused on commonality of aspects and less on identifying fine-grained criteria differences.

### 2.2 Model interpretability

Our work approaches criteria extraction as a feature extraction problem. We train a model and interpret it to gain insights about the data, which is relevant to the broader field of model interpretability. Specifically, we choose PAT (Kim et al., 2023), which is

an interpretable architecture and demonstrated to be capable of discovering dataset or model biases in information retrieval domain (Kim et al., 2024). There are a few reasons that we chose PAT over alternatives. While most interpretability methods like LIME (Ribeiro et al., 2016) and SHAP (Lundberg and Lee, 2017) focus on local explanations for specific instances (Burkart and Huber, 2021), we require global explanations that characterize the entire model (Phillips et al., 2018; Kim et al., 2024). Local explanation methods have significant drawbacks for our purposes: they are computationally expensive (Ahmed et al., 2024), require multiple forward passes per instance, and produce scores that are not meaningful as a probability.

### 2.3 Content moderation systems

Large portion of works in content moderation are focused on incivility (toxicity) and hate speech (Park et al., 2021), which are not sufficient to capture diverse norms in micro-communities like Reddit. A notable approach for capturing diverse norms involves using moderators' rule violation comments as training signals, allowing models to learn from explicit moderation decisions (Park et al., 2021). However, it is limited to few communities where explanation comments are provided. Moreover, it is not clear if a violation criteria for a norm in one community is applicable to others.

Recent work has explored using instruction-following LLMs like ChatGPT for moderation rule enforcement (Kolla et al., 2024), but our experiments confirm previously reported limitations: these models achieve only moderate accuracy and struggle with community-specific rules beyond universal violations like incivility and hate speech.

There were concern that content moderation models with seemingly high performance are indeed relying on the mentions of the ethnic group and spurious features (Röttger et al., 2021; Hartvigsen et al., 2022). Our work address this issue by discovering actual criteria used for classifications, which allow us to identify spurious features or potential limitations of models.

## 3 Criteria discovery

### 3.1 Overview

We conceptualize the challenge of understanding community-specific moderation as a vocabulary scoring problem. While moderation decisions are complex and contextual, we hypothesize that they can be meaningfully represented through predictive lexical patterns that signal rule violations. Our approach aims to build explicit, unambiguous representations of subreddit-specific moderation criteria by extracting and scoring phrasal expressions for each communities. This vocabulary-based representation offers several advantages: interpretable, verifiable, and actionable.

**Task definition:** Given datasets $\{(X_i, Y_i)\}_{i=1}^{N}$ from $N$ communities (subreddits), where $X_i = \{x_i^1, x_i^2, ..., x_i^{n_i}\}$ represents the set of texts from community $i$ and $Y_i = \{y_i^1, y_i^2, ..., y_i^{n_i}\}$ denotes the corresponding binary labels, our goal is to build a vocabulary $V = \{v_1, v_2, ..., v_{|V|}\}$ and community-specific score matrices $S_i \in \mathbb{R}^{|V|}$ for each community $i$, where each element $s_i^j$ indicates the contribution of term $v_j$ to moderation decisions in community $i$.

### 3.2 Dataset

We use the Reddit moderation data that was used in prior works in norms discovery Chandrasekharan et al. (2018), so that one can compare our finding with theirs. The dataset was collected from May 2016 to March 2017. First, the researchers streamed comments via Reddit's API, then after a 24-hour delay checked which ones had been removed by moderators, and retrieved the original content from their logs. Through this process, about 2.8 million moderated comments from 100 top subreddits. As this dataset only contains moderated comments, we augmented it with comments collected from dumps of 2016 Oct to Nov. We built a balanced dataset for each subreddits, where median size of the training data is 18,000 an min size was 8,000 instances.

### 3.3 Method overview

The traditional way of building a machine learning model based on n-gram features suffers from the curse of dimensionality and data sparsity. Thus, we propose to use the novel approach PAT to train a neural classifier to extract n-gram features.

We employ Partial Attention Transformer (PAT) (Kim et al., 2023), which is a model designed for model explanations for text-pair classification tasks, such as natural language inference, and query-document relevance (Kim et al., 2024). The key strength of PAT lies in that it is trained with labels for full texts, while it is forced to predict the label based on scores from two parts of the texts, which gives it ability to assign well-calibrated (0-1 range) probability values.

### 3.4 PAT training

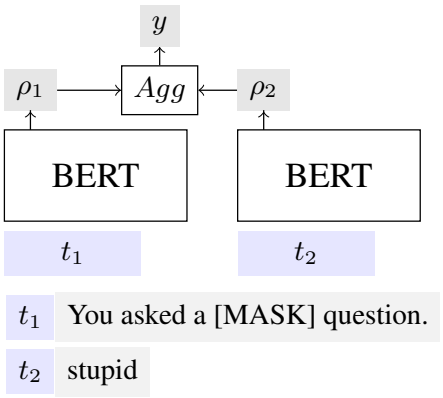


Figure 2: PAT training architecture. A comment text "You asked a stupid question" is partitioned into two sequences and encoded by BERT. The model is supervised with final $y$ label, while encouraging the model to generate corresponding scores $\rho_1$ and $\rho_2$ for each sequences.

We are given a input text $x$ and a label $y$. we creates two partial sequences $(t_1, t_2)$ from $x$, where one sequence $t_1$ is formed by extracting a continuous span from $x$, while $t_2$ contains the remaining tokens with a mask token in place of the extracted span.

PAT encode each partial sequence through a BERT encoder (Devlin et al., 2019). The CLS token representation is projected to obtain scores for both classes

$$\text{PAT}(t_i) = W \cdot \text{BERT}_{CLS}(t_i) + b. \quad (1)$$

Two outputs from PAT for each of $t_1$ and $t_2$ are aggregated by element-wise sum followed by softmax:

$$\hat{y} = \text{softmax}(\text{PAT}(t_1) + \text{PAT}(t_2)) \quad (2)$$

We use cross-entropy loss on predicted probabilities $\hat{y}$ and gold label $y$.

Given an arbitrary span $t$, we can use PAT predict a probability that a text is moderated if it contain $t$, which is given as

$$P(y|t) = \text{softmax}(\text{PAT}(t)). \quad (3)$$

### 3.5 CriteriaMatrix construction

For each subreddit, we train one PAT model with the training data from the subreddit. We then build a vocabulary that is shared across different subreddits, which allow us to compare moderation criteria between communities.

| | mean | politics | Games | history |
|---|---|---|---|---|
| Russian | 0.56 | 0.96 | 0.73 | 0.69 |
| do you speak english | 0.59 | 0.94 | 0.88 | 0.55 |
| Tesla | 0.49 | 0.36 | 0.84 | 0.82 |
| Asian man | 0.68 | 0.49 | 0.96 | 0.76 |
| fucked up | 0.65 | 0.51 | 0.81 | 0.93 |
| holy shit | 0.50 | 0.51 | 0.78 | 0.84 |
| Trump | 0.79 | 0.58 | 1.00 | 0.93 |

Table 2: Sample entries from CriteriaMatrix. First two terms have high scores for being moderated in the politics subreddit, as users used them to accuse others as a Russian spy. Terms having scores around 0.8 are moderated for being off-topic. The topic about the last term was moderated in most subreddits at 2016 (Chandrasekharan et al., 2018).

Since the potential space of all possible token sequences is prohibitively large, we selectively identify the spans that likely influence moderation decisions. For each subreddit $i$, we sampled 1,000 comments and applied the corresponding model, $\text{PAT}_i$ to these comments. We then extracted spans of texts that received high scores. For each span, we tokenized it and collected n-grams (sequences of $n$ consecutive tokens, where $n$ ranges from 1 to 9) that are substrings of it.

We applied this process across 60 subreddits and built a candidate vocabulary. For each $n$ values, we selected the top 10,000 most frequent n-gram terms, based on probability scoring from an off-the-shelf large language model, Llama-3 (Grattafiori et al., 2024).

Finally, we apply each $PAT_i$ to score all terms in vocabulary, to get a score matrix $M$, where $M_{i,j}$ indicates the score that PAT for subreddit $i$ has assigned to a term $j$. We refer to $M$ as CriteriaMatrix throughout the paper. Table 2 shows a selected example entries that demonstrate significant differences between subreddits.

### 3.6 Span-based classification

Note that PAT is used differently in building CriteriaMatrix $M$ than how it is trained.

First, when PAT scores vocabulary terms, it only has access to at most $n$ tokens at a time, whereas during training, at least one of the two towers processes a longer context.

Second, during training, each dimension of the logit values is separately summed (Equation 2) before applying softmax, while in the span extraction, probability is calculated per individual span.

To evaluate the validity of these span-level

scores, we designed a classifier that operates on short text segments. Given a text, we tokenize it by whitespace and enumerate all three-token windows. Each three-token span is scored by PAT, and these scores are averaged to produce a final prediction. This approach allows us to quantify how much predictive performance is maintained when the model is constrained to shorter spans than those used during training.

## 4 Evaluation

In this section, we compare the classification performance of various models, including PAT, to evaluate their effectiveness and gain insights about the data. To accept the output of PAT as moderation criteria, we want to make sure that PAT is predictive of moderation outcomes.

Since we intend to use span-level scores to understand moderation criteria, we must first verify if these scores effectively predict moderation outcomes by evaluating the PAT classifier (subsection 3.6). We include a variant called PAT (Bipartite) which divides the input text into only two parts rather than multiple spans. This better align with training set up as in Figure 2, and demonstrate that PAT could get higher performance with longer contexts.

We evaluated performance across models with varying degrees of subreddit awareness:

**Subreddit aware models:** Models that have knowledge of the specific subreddit being classified, including BERT fine-tuned (FT) on each target subreddit's training data, PAT (Bipartite) and standard PAT trained on subreddit-specific data, ChatGPT prompted with the subreddit name, ChatGPT prompted with both subreddit name and official rules collected via API, LlamaGuard2 (Inan et al., 2023) with subreddit name in the input, and LlamaGuard2 (Toxic) with toxicity definitions and subreddit name in the prompt.

**Subreddit agnostic models:** Models trained or prompted without information about which specific subreddit the content belongs to, including BERT (FT) trained on aggregated data from 60 subreddits, PAT trained on aggregated data without subreddit identifiers, ChatGPT without subreddit-specific context, LlamaGuard2 with default configuration, and LlamaGuard2 (Toxic) with only toxicity definitions in the prompt.

All models were evaluated across 97 subreddits[2], with 100 instances per subreddit. F1 and AUC scores were computed for each subreddit and then averaged across all communities.

| | F1 | AUC |
|---|---|---|
| **Subreddit aware** | | |
| BERT (FT) | 0.81 | 0.90 |
| PAT (Bipratite) | 0.80 | 0.89 |
| PAT | 0.69 | 0.83 |
| ChatGPT | 0.70 | 0.72 |
| ChatGPT + Rule | 0.68 | 0.63 |
| LlamaGuard2 | 0.18 | 0.41 |
| LlamaGuard2 (Toxic) | 0.27 | 0.35 |
| **Subreddit agnostic** | | |
| BERT (FT) | 0.73 | 0.80 |
| PAT | 0.70 | 0.74 |
| ChatGPT | 0.56 | 0.59 |
| LlamaGuard2 | 0.17 | 0.42 |
| LlamaGuard2 (Toxic) | 0.26 | 0.35 |

Table 3: Comparison of classification performance between BERT and PAT using F1 and AUC metrics.

To compute AUC for generative models, we used the token probability for the answer token-the first token that differentiate the classification outcomes-as the prediction score, following the approach used in LlamaGuard (Inan et al., 2023).

Table 3 demonstrates that while PAT does not achieve the same performance level as BERT (which has a full text view), it still attains reasonable scores comparable to other methods. Its F1 and AUC metrics are on par with ChatGPT despite the latter having an order of magnitude more parameters.

Notably, PAT (Bipartite) maintains nearly identical performance to BERT (FT), demonstrating the robustness of the PAT training approach. These results confirm the effectiveness of PAT on Reddit moderation data, which is consistent with its strong performance across diverse tasks in previous work (Kim et al., 2023, 2024). The architecture's success across multiple datasets and tasks indicates that our findings are not limited to this specific Reddit dataset.

Across all models, subreddit-aware variants consistently outperformed their subreddit-agnostic counterparts.

Interestingly, ChatGPT with access to explicit subreddit rules did not outperform the version that only knew the subreddit names. This can be at-

[2]We excluded three subreddits that no longer exist: Incels, soccerstreams, and The_Donald.

| Subreddit | BERT (FT) | ChatGPT (+Subreddit) | ChatGPT (+Rule) |
|---|---|---|---|
| churning | 0.97 | 0.83 | 0.85 |
| conspiracy | 0.90 | 0.72 | 0.80 |
| DIY | 0.83 | 0.71 | 0.73 |
| fantasyfootball | 0.82 | 0.32 | 0.48 |
| Games | 0.85 | 0.57 | 0.7 |

Table 4: Classification F1 scores on selected subreddits for BERT (FT), ChatGPT run which is provided with subreddit name, and provided with rules of subreddit

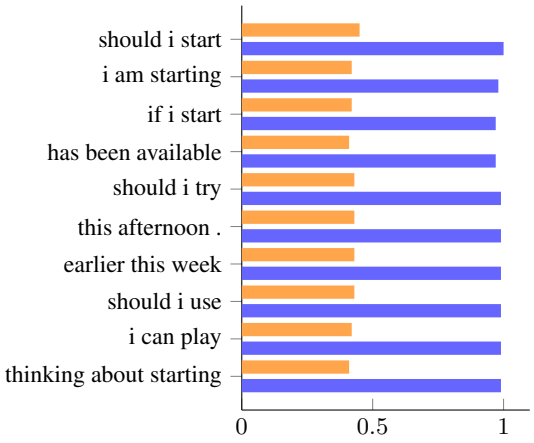


Figure 3: Three-token terms that has highest score difference between the fantasyfootball subreddit (blue) and average over subreddits (orange).

tributed to two factors: First, the majority (70%) of violations (Park et al., 2021) are widely recognized problematic behaviors such as incivility, hate speech, or spam, which ChatGPT already identifies as inappropriate. Second, the actual moderation criteria for more specific rules often cannot be inferred from the officially stated rules alone.

LlamaGuard 2 models performed poorly across all configurations, indicating a significant mismatch between Reddit and their original purpose of safeguarding LLM inputs and outputs.

## 5 CriteriaMatrix Analysis

CriteriaMatrix constructed using PAT scores provides insights into content moderation patterns across different subreddits. In this section, we address three critical questions about these patterns: 1) What patterns are used for prediction when an official rule is not sufficient for deciding moderation. 2) Which norms exhibit unexpected tolerance levels across different communities? 3) What meaningful subcategories exist within broadly defined norms?

### 5.1 Subreddit specific norms

Table 4 presents F1 scores for selected subreddits, highlighting cases where in-domain supervised models like BERT (FT) significantly outperform both the standard ChatGPT and the rule-enhanced ChatGPT variant. Although providing official rules improves ChatGPT's performance in some communities, substantial performance gaps remain in many subreddits.

These performance disparities suggest that BERT (FT) may rely on implicit criteria not captured in official rules. We now examine what specific criteria BERT (FT) models learn to predict moderation outcomes in these communities, focusing on patterns that may not be evident from the official rules alone.

We analyzed CriteriaMatrix for r/fantasyfootball to understand the performance gap between the in-domain supervised model and other models, including GPT variants.

Figure 3 shows the three-token terms with the largest score differences between r/fantasyfootball and the mean across other subreddits. Phrases such as "thinking about starting" and "should i use" appeared as particularly predictive of moderation decisions in this community.

Examining the official rules of r/fantasyfootball, we found a rule that says "No individual threads of any kind specific to your team or league." We can infer that these high-scoring phrases typically appear when users ask questions specific to their own fantasy teams. Since our dataset lacks features to determine whether these comments appeared in individual threads or within valid threads, we cannot directly verify rule violations.

GPT models likely predicted these comments as rule-compliant due to insufficient contextual evidence. In contrast, the supervised BERT model learned that these linguistic patterns alone strongly predict moderation, possibly due to undersampling of non-moderated content in the training data. This reveals how supervised models can learn community-specific moderation patterns that may not be strictly aligned with the literal interpretation of official rules.

### 5.2 Blaming Moderators

Criticizing moderators on Reddit is known to frequently result in content removal across most subreddits (Chandrasekharan et al., 2018; Fiesler et al., 2018). We investigated whether there are differences in how communities determine which

moderator-related comments warrant moderation. CriteriaMatrix showed that unigram terms like "moderator" or "mod" have high scores at many subreddits. Specifically the term "mod" had scores ranging from 0.18 to 0.99, with mean of 0.68, and in 12, "mod" received scores over 0.9.

Based on these observations, we hypothesized that moderation classifiers for many subreddits might be sensitive to any mention of moderators, potentially regardless of context or intent. To test this hypothesis, we constructed a dataset of 50 synthetic comments generated by ChatGPT and Claude that contained the keyword "mod" without expressing blame or criticism toward moderators.

Applying our trained classifiers to this synthetic dataset, we found that 16 out of 60 classifiers (26.7%) predicted **all** 50 instances as requiring moderation, despite their non-critical nature. This suggests several possibilities: (1) people expressed sarcastic descriptions about moderation, so that any mention about moderation is considered sarcastically blaming moderator. (2) synthetic sentences are not natural, which actually triggered moderation.

To confirm this finding with real-world data, we subsequently analyzed the actual moderation rates of comments containing the term "mod" across our dataset. The results strongly supported our hypothesis: in 16% of subreddits, more than 90% of comments mentioning "mod" were removed, and in 34% of subreddits, the removal rate exceeded 80%. These high moderation rates for comments containing a simple reference to moderators support the hypothesis that many communities may have low tolerance for moderator discussions of any kind, potentially explaining why our classifiers flagged even neutral mentions.

### 5.3 Tolerance to personal attack

We hypothesized that there would be meaningful difference in tolerance to a personal attack. Personal attack is very broad categories and there are numerous lexicons. To avoid inspecting all vocabulary, we implemented a clustering approach based on score similarity across subreddits.

We consider that two terms similar in their contribution to moderation decision if their scores over subreddits are highly correlated. Using Pearson correlation coefficients as our similarity metric we performed k-means clustering (k=100) on the term space. We additionally filtered terms that are far from the corresponding centroid to ensure purity of clusters. The resulting Silhouette score of 0.15 indicates a weak but present clustering structure.

| | # of clusters |
|---|---|
| **Personal attack** | 10 |
| **Hate speech** | 6 |
| **Topical** | 16 |
| **URL** | 18 |
| **Markdown** | 5 |
| **Others** | 45 |

Table 5: Distribution of term clusters categorized by content moderation types, derived from k-means clustering (k=100).

We manually categorized these clusters according to the following criteria: Personal attacks included terms potentially offensive toward the second person ("you") or containing general slurs; hate speech includes terms offensive to demographic groups; topical clusters contained topically coherent terms not inherently offensive; URL and markdown clusters focused on structural elements; and the remainder were classified as "other."

We focused on clusters classified as personal attacks. To differentiate between these clusters, we assigned a name to each based on what is common among the terms. We generated descriptive name using Claude [3]. Note that these names are intended to make distinguishing clusters easy and do not precisely describe the clusters.

Table 6 shows the these personal attack clusters with their average scores over subreddits and standard deviation. The clustering has successfully differentiated between distinct personal attack types ranging from the most offensive ones like "direct intelligence insults" (scoring highest at 0.84) to more subtle forms like epistemic competence undermining and boundary-crossing advice (scoring 0.56-0.57).

Figure 4 shows how moderation scores for each personal attack cluster vary across six different subreddits. While the overall pattern shows that some subreddits consistently maintain higher or lower moderation thresholds across all clusters, we observe exceptions in some subreddits. Two subreddits show much lower scores for clusters 37 and 52, respectively. This pattern suggests community-specific tolerance for certain types of personal address. For instance, the lower scores for cluster 52 (boundary-crossing advice) likely indicate that in advice-focused communities, direct guidance that might be considered intrusive elsewhere is instead

[3] https://claude.ai/

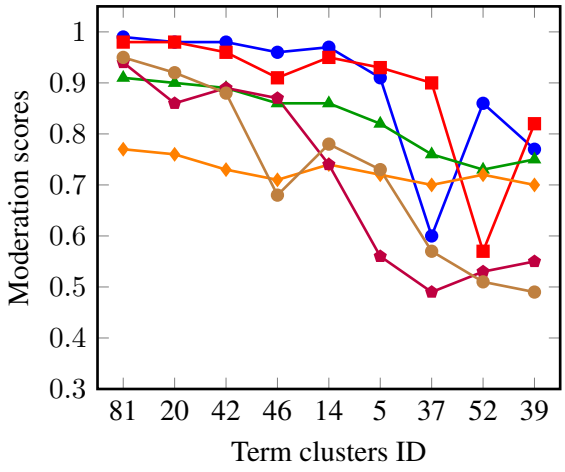


Figure 4: Average moderation scores for personal attack term clusters across six different subreddits. Each line represents a distinct subreddit (anonymized), showing how communities vary in their tolerance for different types of personal attacks.

viewed as appropriate. Similarly, the reduced sensitivity to cluster 37 (second-person framing) in another subreddit suggests that direct addressing of other users is more acceptable within its conversational norms.

| ID | Score (SD) | Name |
|---|---|---|
| 81 | 0.84 (0.15) | Direct intelligence insults |
| 20 | 0.81 (0.16) | Profane command attacks |
| 42 | 0.78 (0.16) | Behavioral mockery |
| 46 | 0.72 (0.16) | Indirect intelligence attacks |
| 14 | 0.71 (0.16) | Dismissive commands |
| 5 | 0.64 (0.15) | Identity questioning |
| 37 | 0.57 (0.13) | Second-person framing |
| 52 | 0.57 (0.13) | Boundary-crossing advice |
| 39 | 0.56 (0.14) | Competence undermining |

Table 6: Clusters of personal-attack (toxic) language with average scores and standard deviations. Terms for each clusters are listed in Table 7.

Our analysis of gives a revealed several important patterns in content moderation.

### 5.4 Summary

Our analysis of CriteriaMatrix reveals three key insights about content moderation across Reddit communities.

First, for community-specific rules, we found patterns that strongly predict moderation decisions in our dataset but may not represent objectively sufficient conditions for removal. This suggests supervised models may learn spurious correlations rather than moderators' true intent.

Second, our "blaming moderators" analysis demonstrated how historical moderation data can lead classifiers to flag even neutral mentions of moderators as violations. This pattern might not align with moderators' future expectations, highlighting needs for potential disconnects between learned patterns and intended policy.

Finally, our clustering of personal attacks revealed that toxic content exists on a spectrum, with communities showing varying tolerance levels for different types of attacks. While prior work often treats toxicity as having universal criteria and thresholds, our findings suggest that community-specific calibration of tolerance for different attack types could improve moderation effectiveness.

## 6 Conclusion

In this paper, we introduced a novel approach to understanding content moderation criteria across different online communities by leveraging an interpretable architecture PAT, to extract lexical expressions predictive of moderation decisions. These expressions provide insights into classification criteria while functioning effectively as classifiers themselves. Our methodology could benefit other classification tasks, which has multiple sub-domains with possible different classifications criteria by assisting developers to understand the underlying criteria and discover possible biases of the dataset.

## Limitations

Our analysis and dataset have several limitations that could affect the generalizability of our findings.

First, our approach only captures patterns observable in short textual expressions, excluding toxic behaviors that require broader context to identify. Our data also lacks important contextual elements such as previous comments or parent posts that often influence moderation decisions.

Second, our use of balanced datasets with equal proportions of moderated and non-moderated content differs significantly from the natural distribution on Reddit, where only about 5% of comments are typically moderated. This sampling approach, while standard for classification tasks, may amplify certain patterns that would be less prominent in real-world applications.

Third, the dataset's limited time window and size may be critical to the resulting models having suboptimal performance and biases. Some patterns identified in our analysis might not persist or might

appear different with more comprehensive data collection spanning longer periods.

These limitations suggest caution in interpreting our findings and highlight opportunities for future work with richer contextual data, more representative sampling, and longitudinal analysis of moderation patterns.

## References


Shamim Ahmed, M Shamim Kaiser, Mohammad Shahadat Hossain, and Karl Andersson. 2024. A comparative analysis of lime and shap interpreters with explainable ml-based diabetes predictions. *IEEE Access*.

Nadia Burkart and Marco F Huber. 2021. A survey on the explainability of supervised machine learning. *Journal of Artificial Intelligence Research*, 70:245–317.

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

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.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. 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 4171–4186.

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*, volume 12.

Agam Goyal, Charlotte Lambert, Yoshee Jain, and Eshwar Chandrasekharan. 2024. Uncovering the internet's hidden values: An empirical study of desirable behavior using highly-upvoted content on reddit. *arXiv preprint arXiv:2410.13036*.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, et al. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

Thomas Hartvigsen, Saadia Gabriel, Hamid Palangi, Maarten Sap, Dipankar Ray, and Ece Kamar. 2022. Toxigen: A large-scale machine-generated dataset for adversarial and implicit hate speech detection. *arXiv preprint arXiv:2203.09509*.

Hakan Inan, Kartikeya Upasani, Jianfeng Chi, Rashi Rungta, Krithika Iyer, Yuning Mao, Michael Tontchev, Qing Hu, Brian Fuller, Davide Testuggine, et al. 2023. Llama guard: Llm-based input-output safeguard for human-ai conversations. *arXiv preprint arXiv:2312.06674*.

Youngwoo Kim, Razieh Rahimi, and James Allan. 2023. Conditional natural language inference. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 6833–6851, Singapore. Association for Computational Linguistics.

Youngwoo Kim, Razieh Rahimi, and James Allan. 2024. Discovering biases in information retrieval models using relevance thesaurus as global explanation. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 19530–19547.

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*, pages 1–8.

Charlotte Lambert, Frederick Choi, and Eshwar Chandrasekharan. 2024. " positive reinforcement helps breed positive behavior": Moderator perspectives on encouraging desirable behavior. *Proceedings of the ACM on Human-Computer Interaction*, 8(CSCW2):1–33.

Scott M Lundberg and Su-In Lee. 2017. A unified approach to interpreting model predictions. *Advances in neural information processing systems*, 30.

Yair Neuman and Yochai Cohen. 2023. Ai for identifying social norm violation. *Scientific Reports*, 13(1):8103.

Chan Young Park, Shuyue Stella Li, Hayoung Jung, Svitlana Volkova, Tanu Mitra, David Jurgens, and Yulia Tsvetkov. 2024. Valuescope: Unveiling implicit norms and values via return potential model of social interactions. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 16659–16695.

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*, pages 3386–3397.

Richard Phillips, Kyu Hyun Chang, and Sorelle A Friedler. 2018. Interpretable active learning. In *Conference on fairness, accountability and transparency*, pages 49–61. PMLR.

Kavel Rao, Liwei Jiang, Valentina Pyatkin, Yuling Gu, Niket Tandon, Nouha Dziri, Faeze Brahman, and Yejin Choi. 2023. What makes it ok to set a fire? iterative self-distillation of contexts and rationales for disambiguating defeasible social and moral situations. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 12140–12159, Singapore. Association for Computational Linguistics.

Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. 2016. "why should i trust you?": Explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '16, page 1135–1144, New York, NY, USA. Association for Computing Machinery.

Paul Röttger, Bertie Vidgen, Dong Nguyen, Zeerak Waseem, Helen Margetts, and Janet Pierrehumbert. 2021. Hatecheck: Functional tests for hate speech detection models. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 41–58.

Galen Weld, Amy X Zhang, and Tim Althoff. 2024. Making online communities 'better': a taxonomy of community values on reddit. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 18, pages 1611–1633.

# A Baseline implementations

### ChatGPT - Prompt

For ChatGPT runs in section 4, we concatenated a task specific instruction to a target text to be classified.

**ChatGPT with rule**

For each subreddit, we construct a prompt by concatenating its moderation rules. We used the both summary and detail parts of rules which is collected by the official API.

The prompt is formatted as follows:

```
[Subreddit moderation rules]
The above rule describes
prohibited contents. Classify if
the following text is prohibited.
If prohibited, output 'unsafe'
as a first token. If not, output
'safe'.
```

The model is instructed to output the token `unsafe` if the input violates any subreddit rule and `safe` otherwise.

**ChatGPT - subreddit aware**

```
If the following text is posted
in {subreddit} subreddit, will it
be moderated (deleted)?
Answer Yes or No, as a single
token.
```

**ChatGPT - subreddit agnostic**

```
If the following text is posted
on Reddit, will it be moderated
(deleted)?
Answer Yes or No, as a single
token.
```

# B PAT model implementations

### Training Hyperparameters

The following hyperparameters were used during model fine-tuning:

- **Number of training epochs:** 3
- **Learning rate:** 5e-5
- **Training batch size:** 16
- **Warmup ratio:** 0.1
- **Weight decay:** 0.01

- **Learning rate scheduler:** Linear with warmup

No extensive hyperparameter tuning were used throughout the deveopment process.

# C Disclaimers

## C.1 Reddit Content Moderation Dataset

The dataset of Reddit moderation (Chandrasekharan et al., 2018), used as the artifact in this paper, has been carefully curated and anonymized by its creators to protect user privacy and prevent the inclusion of personally identifying information. The dataset consists of comments text with personal identifiable meta data removed.

## C.2 AI Assistance

We acknowledge the use of AI assistants, Claude by Anthropic [4] and GPT by OpenAI [5], in the writing process of this paper. These AI assistants provided support in drafting and refining the contents of the paper. However, all final decisions regarding the content, structure, and claims were made by the human authors, who carefully reviewed and edited the generated content.

## C.3 Computational cost

We used one of the following GPUs for training: NVIDIA RTX A6000, A40, or V100. With any of these devices, training took less than two hours. Note that all implementations are designed to run within 16 GB of VRAM, and the computational cost is typical compared to the standard practice of fine-tuning the BERT-base-uncased model from Hugging Face's Transformers library.

[4] https://www.anthropic.com/claude
[5] https://chat.openai.com/

| ID | Score | Name | Sample Terms |
|---|---|---|---|
| 81 | 0.84 | Direct intelligence insults | pussy; whore; are you gay; you are an idiot; you are a moron; trump is a racist; you are a loser; are you a moron; are you gay?; you are a stupid; you are so stupid; what a fucking idiot; are you an idiot; stop being a bitch; stop being an idiot; |
| 20 | 0.81 | Profane command attacks | motherfucker; motherfuckers; shut the fuck; shut up you; kiss my ass; fuck you,; shut the fuck up; get the fuck out; eat shit and die; calm the fuck down; fucked in the ass; blow your brains out; you are a fucking; get off your ass; |
| 42 | 0.78 | Behavioral mockery | douchebag; dickhead; slut; idiot; cunt; douchebags; slander; bitches; dick,; dumb and dumber; stupid stupid stupid; what a stupid; a piece of shit; you are a fool; dumb, stupid,; are you a nerd; you are a hypocrite; |
| 46 | 0.72 | Indirect intelligence attacks | shut up; yourself a; shut up and; shut up,; change your life; live your life; get yourself a; teach your child; shut up.; eat your heart; "shut up; go away,; stop being a; in your mouth; shut the hell; stop being so; sick of your; |
| 14 | 0.71 | Dismissive commands | slutty; stupidest; stupid,; stupid people; stupidity is; this is stupid; this is stupid and; what is this stupid; too stupid to be; stupid...; there is no stupid; complete and utter bullshit; kind of an idiot; |
| 5 | 0.64 | Identity questioning | you a; are you a; you are a; are you an; you are an; you were a; you are one; like you are; you are such; like you were; you is a; you must be a; you are one of; that you are a; looks like you are; because you are a; |
| 37 | 0.57 | Second-person framing | you; youll; youd; you are; are you; you don; you must; - you; you were; all you; because you; you just; maybe you; you,; (you; now you; you.; you all; , you; you never |
| 52 | 0.57 | Boundary-crossing advice | get your; find your; keep your; things you; let your; put your; your personal; save your; what your; where your; please dont; leaving your; please do not; for all your; speak to a; with all your; let it go; for your personal; part of your; speak to an |
| 39 | 0.56 | Competence undermining | you may not; you should have; you have no; you dont know; the reason you; you should just; you had no; you dont get; you want to be; you do not need; you have no idea; you are not going; |

Table 7: Clusters of personal-attack (toxic) language with sample terms