MoMoE: Mixture of Moderation Experts Framework for AI-Assisted Online Governance

Agam Goyal, Xianyang Zhan*, Yilun Chen*, Koustuv Saha‡, Eshwar Chandrasekharan‡
Siebel School of Computing and Data Science
University of Illinois Urbana-Champaign
{agamg2, zhan39, yilunc3, ksaha2, eshwar}@illinois.edu

Abstract

Large language models (LLMs) have shown great potential in flagging harmful content in online communities. Yet, existing approaches for moderation require a separate model for every community and are opaque in their decision-making, limiting real-world adoption. We introduce **M**ixture **o**f **Mo**deration **E**xperts (MoMoE), a modular, cross-community framework that adds post-hoc explanations to scalable content moderation. MoMoE orchestrates four operators— **Allocate**, **Predict**, **Aggregate**, **Explain** —and is instantiated as seven community-specialized experts ($\text{MoMoE}_{\text{Community}}$) and five norm-violation experts ($\text{MoMoE}_{\text{NormVio}}$). On 30 unseen subreddits, the best variants obtain Micro-F1 scores of 0.72 and 0.67, respectively, matching or surpassing strong fine-tuned baselines while consistently producing concise and reliable explanations. Although community-specialized experts deliver the highest peak accuracy, norm-violation experts provide steadier performance across domains. These findings show that MoMoE yields scalable, transparent moderation without needing per-community fine-tuning. More broadly, they suggest that lightweight, explainable expert ensembles can guide future NLP and HCI research on trustworthy human-AI governance of online communities.[1]

1 Introduction

A persistent challenge that online communities face is identifying content that violates community norms. This challenge is particularly crucial on platforms like Reddit, which hosts over 125,000 active communities called subreddits with diverse norms and moderation needs, placing significant burden on unpaid moderators (Li et al., 2022).

To alleviate this burden, various sociotechnical tools for content moderation have been proposed in prior work. These include keyword-based moderation using simple regular expression filters (Long et al., 2017; Jhaver et al., 2019, 2022), traditional ML-based moderation, which range from embedding-based classifiers (Chandrasekharan et al., 2017, 2019) to language model (LM)-based moderation approaches, which have recently gained popularity as they show promising performance and can enhance transparency. (Kumar et al., 2024; Kolla et al., 2024; Zhan et al., 2025).

However, existing LM-based approaches for content moderation face some key challenges that hinder their deployment in real-world scenarios.

First, while Zhan et al. (2025) demonstrated that fine-tuned SLMs can outperform off-the-shelf LLMs on content moderation, they require substantial community-specific training data for fine-tuning models. This creates significant barriers for new communities, as they may lack historical moderation data required to fine-tune these models.

Second, research has identified that different online communities operate under shared yet distinct norms and values (Chandrasekharan et al., 2018; Goyal et al., 2024). Yet, existing LM-based approaches rely on instantiating a single model per community, which hinders the ability of these models to cater to a large number of communities that may share a similar kind of norms violations, to enable a cross-community moderation approach.

Third, while existing approaches focus solely on accuracy, recent work has called for improved transparency in order to improve moderator trust (Huang, 2024; Palla et al., 2025; Moran et al., 2025) and ensure that moderator workload doesn't increase due to difficulty in identifying inconsistencies within these systems (Ashkinaze et al., 2024).

Finally, current AI-based approaches treat content moderation as a fully automated task, overlooking the crucial role of human moderators who

*Both authors contributed equally.
‡Both authors are advisors of this work.
[1]Code: https://github.com/scuba-illinois/MoMoE

Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing, pages 12657–12672
November 4-9, 2025

possess contextual understanding and community expertise. Effective moderation systems should not aim to replace human moderators but rather augment and complement their capabilities by providing transparent justifications that allow for human oversight and intervention (Selbst et al., 2019a; Kolla et al., 2024; Huang et al., 2024). Therefore, there is a critical need for frameworks that can show how to efficiently operate with these AI-based approaches by leveraging the complementary strengths of humans and LLMs.

In this paper, we introduce MoMoE (**M**ixture **o**f **Mo**deration **E**xperts), a novel ensemble framework for cross-community content moderation that addresses these limitations. MoMoE is a modular framework that is composed of four operators: **(1)** **`Allocate`:** Operator that decides how to pick relevant experts and weigh their decisions for a specific instance of the moderation task (e.g., classification-based, similarity-based, etc.); **(2)** **`Predict`:** Operator that leverages a mixture of fine-tuned small language models ("experts") representing either community-based experts ($\text{MoMoE}_{\text{Community}}$) or norm violation-based experts ($\text{MoMoE}_{\text{NormVio}}$) to provide a moderation outcome; **(3)** **`Aggregate`:** Operator that decides how to combine predictions of the individual experts (e.g., dot-product composition, majority voting, etc.); and **(4)** **`Explain`:** Operator that provides simplified post-hoc LLM-based explanations for MoMoE decisions.

We evaluate the effectiveness of MoMoE using a comment removal dataset (Chandrasekharan and Gilbert, 2019) by simulating a real-time content moderation scenario, and perform an extensive quantitative and qualitative analysis. We find that MoMoE performs competitively against strong baselines on 30 unseen communities. Specifically, the best configurations of $\text{MoMoE}_{\text{Community}}$ and $\text{MoMoE}_{\text{NormVio}}$ achieve Micro-F1 scores of 0.72 and 0.67, respectively. While $\text{MoMoE}_{\text{Community}}$ achieves a wider range of performance depending on the target community, $\text{MoMoE}_{\text{NormVio}}$ provides consistently strong performance across communities. Through case studies, we provide a detailed analysis of the complementary strengths of different operator configurations. Further, through manual inspection we find that the explanations provided by `Explain` reliably reflect the decision-making trace of MoMoE.

By integrating multiple expert perspectives and providing transparent explanations, MoMoE aims to create a more generalizable approach to AI-assisted governance of online forums that upholds community-specific norms while leveraging cross-community knowledge. The modular nature of MoMoE provides human moderators the agency to intervene and perform recalibration at the level of each operator, and moreover it also provides opportunity for individual components to be enhanced with advancements in NLP. Our goal is to enhance the potential for human-AI collaborative moderation by contributing a *framework for AI-based tools that complement rather than replace human expertise*, while still performing competitively in comparison to strong baselines.

2 Related Work

AI-assisted content moderation: Content moderation on most online platforms is primarily done manually by either commercial moderators or unpaid volunteers (Gillespie, 2018; Roberts, 2019; Li et al., 2022). Prior work has proposed many AI-based approaches to content moderation. This includes both embedding-based classifiers (Chandrasekharan et al., 2019; Park et al., 2021) and LLM-based approaches (Mullick et al., 2023; Kolla et al., 2024; Kumar et al., 2024; Vishwamitra et al., 2024; Zhan et al., 2025). However, it has been found that in highly contextual tasks such as moderation, human judgment is often superior to automated judgment (Jurgens et al., 2019; Gorwa et al., 2020). Due to the dichotomous nature of this problem, there has been a lack of studies on this front, except that of Park et al. (2025) which proposes a human-LLM pipeline for cross-cultural hate speech moderation. *Our work is a step in this direction of enhancing AI-assisted moderation. We focus on rule-based content moderation, which encompasses hate-speech moderation but is broader and more reflective of real-world moderation processes.*

Human-AI collaborative decision making: Human decision-making is highly nuanced and contextual. With the rise of LLMs, there has been an increasing body of research that proposes to use LLMs for high-stakes decision making in domains of healthcare (Benkirane et al., 2025), moderation (Koshy et al., 2024), etc. Key considerations in these collaborative tools is to assist human decision-making without replacing them, and the final judgment is that of humans (Steyvers and Kumar, 2024). As a result, there has recently been a growing body of research (Li et al., 2023; Vereschak et al., 2024; Li et al., 2025; Castañeda et al., 2025) building tools and approaches that facilitate these decision-

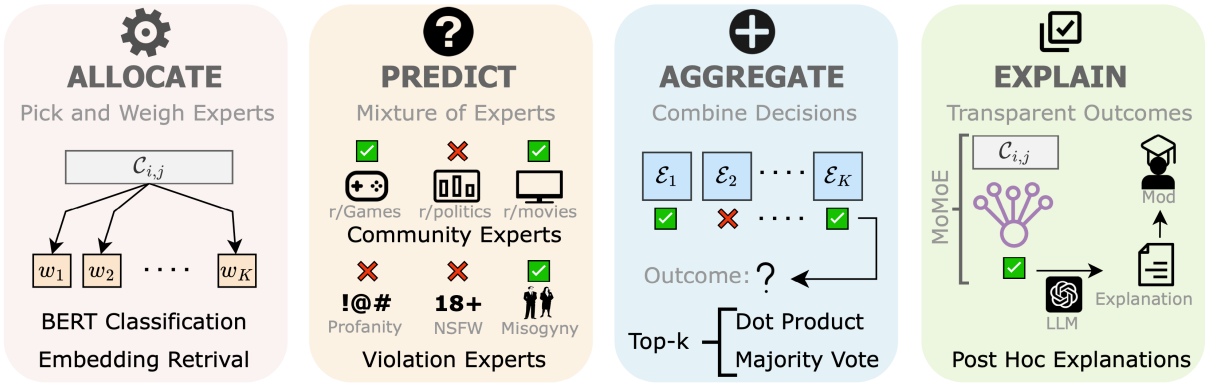

Figure 1: **MoMoE** is composed of four modular operators— **(1)** **Allocate** **:** Determines how to pick the relevant experts and weigh the predictions they provide using softmax probabilities from classification models or similarity-based scoring; **(2)** **Predict** **:** Determines individual expert predictions from two kind of ensembles, with community-specific experts or norm-violation experts; **(3)** **Aggregate** **:** Determines how to aggregate the predictions of individual experts into a single outcome using strategies like dot product between allocated weights and expert predictions or majority voting; and **(4)** **Explain** **:** Uses a post hoc LLM-based approach to summarize and explain MoMoE's decision output to help moderators understand outcomes and rectify potential inconsistencies.

making processes. However, despite the growing interest in using LLMs for content moderation, *there is a lack of research in developing approaches to enhance online governance which our work aims to address.*

3 Preliminaries

We now detail the communities we examine and the datasets we curate.

Communities: We categorize communities (subreddits) of interest into two groups for our study: **(1) Source subreddits:** These subreddits serve as the foundation for our expert models. We select 7 popular subreddits with a wide spectrum of topics, moderation styles, and community norms[2]: *r/AskHistorians*, *r/AskReddit*, *r/Games*, *r/anime*, *r/changemyview*, *r/politics*, and *r/science*. **(2) Target subreddits:** These subreddits are used for testing the performance and generalization capabilities of MoMoE compared to other baseline approaches. We select 30 diverse subreddits (listed in Appendix A) chosen for their variety in topics, community sizes, and community norms.

Datasets: We curate our data from the publicly available dataset of Reddit comment removals collected between May 10, 2016 and February 4, 2017 by Chandrasekharan et al. (2018). We create two kinds of datasets for our tasks:

(1) Community Dataset ($\mathcal{D}_{\text{Community}}$): This dataset consists of *subreddit*, *comment*, *context*, and *label*, where the *context* is the parent-comment of the original comment, and the *label* is a binary value of 'True' or 'False' to indicate whether the comment was removed by moderators. $\mathcal{D}_{\text{Community}}$ contains a total of 70,000 entries (7 source subreddits $\times$ 10K comments/subreddit)

(2) Norm Violation Dataset ($\mathcal{D}_{\text{NormVio}}$): This dataset consists of *norm-violation*, *subreddit*, *comment*, *context*, and *label*, where the *norm-violation* column represents the specific kind of norm the original comment violates or does not violate, as noted by the *label*. We create the labels for this column through an LLM-based approach. Specifically, we prompt GPT-4o (OpenAI, 2024) with the *context*, *comment*, and the rules of the subreddit and ask it to label each removed *comment* in our source datasets with the *subreddit* rule that it violates. Next, we manually categorize the set of rules from different subreddits into 5 broader norm-violation themes.[3] We then use these mappings as the final label for the *norm-violation* column. Overall, $\mathcal{D}_{\text{NormVio}}$ contains 81,262 entries, balanced across norm-violation categories. The prompt to obtain rule-to-norm violation mappings, and their validity in terms of accuracy (87% accuracy) and coders' inter-rater reliability (Krippen-

[2]Although we select these subreddits, our framework can be extended to any other set of subreddits and number of experts at the time of deployment. (See Section 10)

[3]The 5 themes are (1) 'Bad Faith or Unsubstantiated Arguments', (2) 'Civility and Respect', (3) 'Low Effort, Off-Topic, or Non-Substantive Contributions', (4) 'Rule Enforcement and Structural Integrity of Discussions', and (5) 'Spam, Solicitation, Misinformation, and Machine-Generated Content'.

dorff's $\alpha = 0.82$) can be found in Appendix B.

4 MoMoE Framework

We now explain each component of MoMoE and the rationale behind our choices.

`Allocate`: Given an incoming context-comment pair, this operator identifies the appropriate experts within MoMoE and determines their relative importance. We implement two distinct approaches for this allocation:

(1) Classification-based allocation: We create an 80:20 train-test split for the source datasets and fine-tune two separate RoBERTa-base model (Liu et al., 2019) given the concatenated "context" and "comment" as input to (i) predict the source *subreddit* label from $\mathcal{D}_{\text{Community}}$ and (ii) predict the *norm-violation* label from $\mathcal{D}_{\text{NormVio}}$. These classifiers for $\mathcal{D}_{\text{Community}}$ and $\mathcal{D}_{\text{NormVio}}$ achieve a test accuracy of 78% and 62% respectively.[4] Next, for allocation, we compute the softmax probabilities from logits of the penultimate layer of the model, and use them directly as expert weights. This approach leverages the model's ability to *identify subreddit-specific linguistic patterns and discussion topics*.

(2) Similarity-based allocation: We utilize the SentenceBERT model (Reimers and Gurevych, 2019) `all-mpnet-base-v2` to compute embeddings for all comments in both source and target subreddits. For each comment in the target datasets, we compute two types of averaged cosine similarities: (i) between the embedding of the target comment and the embeddings of all comments in each source subreddit; and (ii) between the embedding of the target comment and the embeddings of all comments in each norm-violation category in the source subreddits. This process yields either: (i) 7 similarity scores (one per source subreddit) or (ii) 5 similarity scores (one per norm-violation category) that each lies in $[-1, 1]$. We apply a softmax function ($\tau = 0.1$) to convert these scores into probability distributions, which serve as the weights for our experts. This approach captures *semantic similarity between comments across communities*.[5]

`Predict`: This operator is the core component of MoMoE that uses the mixture-of-experts inspired framework to determine moderation outcomes. This component takes the context-comment pair as input and produces binary moderation decisions from multiple specialized experts. For our expert models, following Zhan et al. (2025), we leverage two state-of-the-art open-source small language models (SLMs): *Llama-3.1-8B* (Dubey et al., 2024), and *Mistral-Nemo-Instruct* (Mistral AI, 2024). Each model is fine-tuned using Low-Rank Adaptation (LoRA) (Hu et al., 2021) to create specialized experts for content moderation. We fine-tune these models using rule-based prompting. The detailed prompts used for LoRA fine-tuning as well as hyperparameter details can be found in Appendix C. We explore two distinct approaches to expert specialization:

(1) Community-based Experts: This approach ($\text{MoMoE}_{\text{Community}}$) creates subreddit-specific experts by fine-tuning SLMs on data from each source subreddit. Using an 80:20 train-test split of $\mathcal{D}_{\text{Community}}$ stratified by *subreddit*, we fine-tune separate expert models for each source subreddit. Each expert specializes in the specialized rules and moderation patterns of its respective community.

(2) Norm-violation Experts: This approach ($\text{MoMoE}_{\text{NormVio}}$) creates subreddit-specific experts by fine-tuning SLMs on data from each norm-violation category. Using an 80:20 train-test split of $\mathcal{D}_{\text{NormVio}}$ stratified by *norm-violation*, we fine-tune separate expert models for each of the 5 categories, where each expert is specialized in detecting particular kinds of norm violations.

These complementary approaches offer distinct advantages for content moderation. We hypothesize that *community-based experts would excel at capturing the nuanced, community-specific norms* that may vary significantly across different subreddits. In contrast, *norm-violation experts should generalize better across communities* by focusing on fundamental categories of problematic content that tend to be universally unacceptable across most online spaces, albeit to possibly varying extents.

`Aggregate`: This operator is responsible for combining the decisions of multiple experts using their allocated weights to produce a final outcome. We implement this component with a "Top-K" approach that selects the K experts with the highest allocation weights. Within this framework, we explore two distinct aggregation strategies:

(1) Dot Product: We compute a weighted composition score by taking the dot product between two vectors of dimension K: (i) a binary decision vector from the experts; and (ii) the allocation weight

[4] See Appendix C for fine-tuning hyperparameters.

[5] Note that learning allocation weights through backpropagation is another alternative which we do not explore in this work as one of our key goals is transparent allocations.

	LLM		Base SLM		Fine-tuned SLM		MoMoE$_{\text{Community}}$		MoMoE$_{\text{NormVio}}$	
Subreddit	**GPT-4o-mini**	**GPT-4o**	**Llama**	**Mistral**	**Llama**	**Mistral**	**Llama**	**Mistral**	**Llama**	**Mistral**
r/anime	41.8	33.0	56.6	12.6	63.1	75.5	61.4	72.9	58.7	70.6
r/AskHistorians	26.3	38.2	54.9	8.3	67.4	76.9	66.9	74.5	63.6	72.6
r/AskReddit	51.7	51.5	56.3*	40.6	55.3	62.5	55.7*	60.8	49.7	60.1
r/changemyview	79.4	74.9	57.7	55.4	90.3	91.8	84.8	86.7	83.5	85.4
r/Games	45.7	44.2	55.6	22.9	69.9	74.3	70.3*	72.4	68.4	71.5
r/politics	71.8	72.2	54.3	53.8	72.5	78.1	72.8	78.3	70.2	73.6
r/science	42.9	66.6	52.2	30.2	63.0	72.7	62.2	69.4	61.3	67.4

Table 1: **MoMoE Performance on Source Subreddits.** Micro-F1 scores (higher is better) are colored by relative drop vs. the corresponding fine-tuned SLM (for MoMoE$_{\text{Community}}$, MoMoE$_{\text{NormVio}}$, and Base SLM) or vs. the best fine-tuned SLM (for LLMs): ≤ 2.5 % drop , 2.5–10 % drop , > 10 % drop , and improvement ($^*p < 0.05$). The MoMoE models use the classification-based strategy and dot-product based aggregation.

vector determined by the **Allocate** operator. We apply a threshold at 0.5—if the composition score exceeds this threshold, it returns 'True' (comment should be removed); otherwise, it returns 'False'. **(2) Majority Vote:** We determine the final outcome by taking a simple majority vote among the Top-K chosen experts. The decision supported by more than half of the experts is final.

These aggregation strategies allow us to evaluate the trade-offs between *relying on a few highly relevant experts versus incorporating a broader consensus*, and different aggregation methods.

Explain: This operator is the final component of MoMoE, aimed at using a strong LLM like GPT-4o for providing transparent justifications for moderation decisions to human moderators who would use such a framework in practice (See Appendix D for prompt design). We generate explanations that detail the reasoning behind MoMoE's decisions.

These explanation strategies have many benefits. First, it enables moderators to easily understand which experts were most relevant to a particular moderation decision and why. Second, it helps moderators identify the specific types of norm violations or community standards that were considered while making these decisions, which helps in facilitating more consistent and fair moderation decisions across similar cases that may have otherwise been treated differently or overlooked. Most importantly, the transparent nature of these explanations allows moderators to identify potential biases or miscalibrations within the MoMoE framework.

Evaluation Baselines: **(A) Performance:** We evaluate MoMoE against the following baselines: **(1) Detoxify** (Hanu and Unitary team, 2020): a model that computes toxicity and is thresholded at 0.5 to determine the moderation outcome. **(2) Global SLMs:** Small language models fine-tuned separately on entire train-split of $\mathcal{D}_{\text{Community}}$ (G-Llama$_{\text{Community}}$ and G-Mistral$_{\text{Community}}$) and $\mathcal{D}_{\text{NormVio}}$ (G-Llama$_{\text{NormVio}}$ and G-Mistral$_{\text{NormVio}}$). **(3) Global LLMs:** Zero-shot prompted LLMs (GPT-4o and GPT-4o-mini) to predict whether a comment will be removed. Note that we do not perform few-shot ICL as prior work (Zhan et al., 2025) has shown that it does not reliably improve performance due to lack of ways to incorporate examples relevant to specific communities. **(B) Explainability:** We perform manual validation to check that the generated explanations reliably reflect the course of MoMoE's decision-making trace.

5 Results

We now evaluate **Predict** in terms of Micro-F1 (F1 hereafter), highlighting the key tradeoffs compared to community-specific SLMs and LLMs, and demonstrate the benefits of MoMoE operators.

MoMoE Performance on Source Subreddits: We first evaluate how MoMoE$_{\text{Community}}$ and MoMoE$_{\text{NormVio}}$ perform on the test splits of the source subreddits, the train-data of which their experts were fine-tuned on. This evaluation acts as a sanity check to ensure that shifting from community-specific SLMs to a mixture-of-moderation-experts does not lead to a large drop in performance and at the same time can still outperform off-the-shelf SLMs and LLMs. From Table 1, we see that both MoMoE approaches show only moderate performance drops compared to community-specific fine-tuned SLMs, with MoMoE$_{\text{Community}}$ experiencing smaller drops (typically $\leq$2.5%) than MoMoE$_{\text{NormVio}}$ across most subreddits. All drops are significant by Welch's t-test (Welch, 1947) for MoMoE (p<0.05) and LLMs (p<0.001). Notably, MoMoE$_{\text{Community}}$ even outperforms the source SLMs in two cases (p<0.05), highlighting that the ensemble approach can sometimes benefit from shared moderation patterns across communities. Furthermore, both MoMoE variants

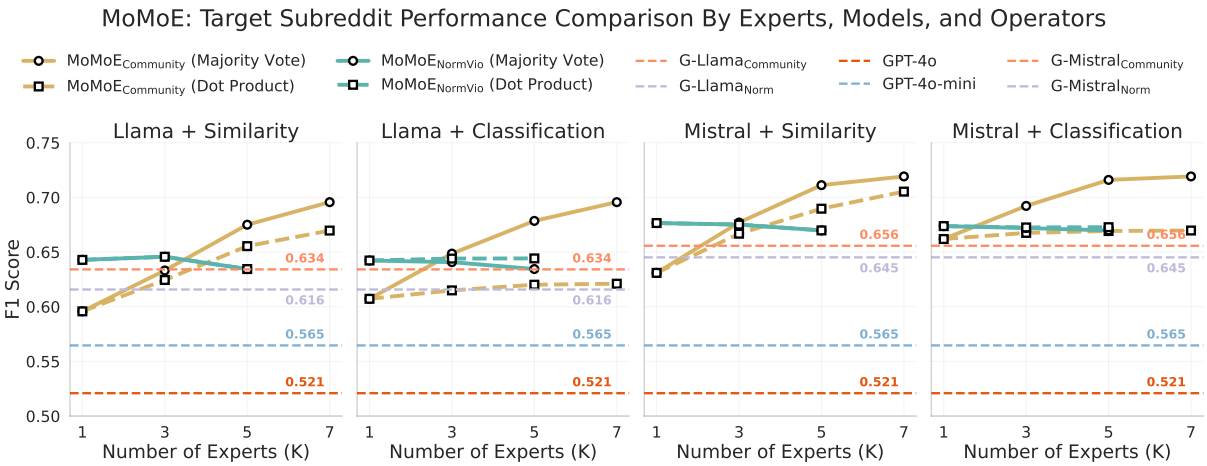

Figure 2: Performance of MoMoE on Target Subreddits reveals that both MoMoE$_{Community}$ and MoMoE$_{NormVio}$ perform competitively against baselines either matching or outperforming them in terms of Micro-F1 score.

substantially outperform off-the-shelf LLMs GPT-4o-mini and GPT-4o, which show dramatic performance drops (often >10%) compared to specialized SLMs. This shows that MoMoE performs well on source data, encouraging us to apply MoMoE to a more challenging setting where we adopt a community-agnostic approach without prior knowledge of comment origins.

MoMoE Performance Against Baselines: In our subreddit-agnostic setting where we use comments from 30 target subreddits, Figure 2 demonstrates that MoMoE consistently matches or outperform baselines. The strongest baselines are the Global SLMs G-Mistral$_{Community}$ and G-Llama$_{Community}$ with F1 scores of 0.656 and 0.634 respectively, while the LLMs GPT-4o and GPT-4o-mini show much weaker performance with F1 scores of 0.521 and 0.565 respectively. The toxicity classifier based on Detoxify (not plotted) showed the worst performance with an F1 score of 0.38. The Mistral-based MoMoE$_{Community}$ with both classification and similarity allocation strategies with majority-voting based aggregation is the best performing configuration with F1 scores of 0.72. All improvements for MoMoE$_{Community}$ against the strongest baseline—community-based Global SLMs—are significant at K=5 and 7 ($p < 0.001$), except dot-product aggregation for Llama with classification allocation. For MoMoE$_{NormVio}$, all improvements at K=1 and K=3, and at K=5 only improvements by Mistral, are significant ($p < 0.05$).

We find that for MoMoE$_{Community}$, increase in the number of experts K generally leads to a notable increase in performance, while for MoMoE$_{NormVio}$, increasing the number of experts maintains or slightly drops performance. Based on our rule-to-norm mappings (Appendix B), we hypothesize that in the case of norm-violation experts even if one of the experts deems the comment as violating, it should be removed as the violation of any of these norms is harmful across most communities. Incorporating more experts could therefore lead to misclassification in some cases. We expand on these findings with a precision-recall trade-off analysis comparing MoMoE$_{Community}$ and MoMoE$_{NormVio}$ in Appendix F. Also see Appendix G for performance of `Predict` on imbalanced test-split in terms of AUC and Macro-F1.

Impact of Operator Choice: In terms of the `Allocate` operators, for low number of experts (K = 1 or 3) classification-based allocation slightly outperforms similarity-based allocation (≈ 0.04 F1 on average, $p < 0.05$), while for higher number of experts (K = 5 or 7), both allocation strategies are equivalent with negligible differences in F1 scores. In terms of `Aggregate` operators, we find that for MoMoE$_{Community}$, majority voting consistently outperforms dot-product based aggregation. On the other hand, for MoMoE$_{NormVio}$ the two aggregation strategies are roughly equivalent with dot product slightly outperforming majority vote in the case of classification-based allocation (≈ 0.01 F1 on average, $n.s.$). We discuss potential reasons for some of these observations in Section 5.

Dissecting Performance of `Predict` on Target Subreddits: We now dive deeper into the performance of Llama-based MoMoE on the target subreddits, providing insights into the key differences

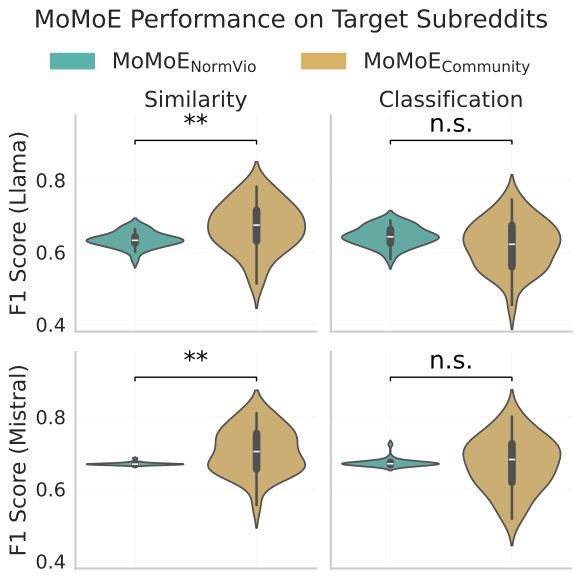

Figure 3: Comparing F1 score performance with dot-product based aggregation we observe that while $\texttt{MoMoE}_{\texttt{Community}}$ provides a wider range of performance across subreddits ($\approx 0.45 - 0.8$), $\texttt{MoMoE}_{\texttt{NormVio}}$ gives consistent moderate performance across subreddits (≈ 0.65). ($^{*}p < 0.05$, $^{**}p < 0.01$, $^{***}p < 0.001$)

between $\texttt{MoMoE}_{\texttt{Community}}$ versus $\texttt{MoMoE}_{\texttt{NormVio}}$ with the dot-product based aggregation strategy for this case study, although these trends are consistent with majority-voting as well. All significance levels are from Welch's $t-$test (Welch, 1947).

Figure 3 shows key differences in performance variability between our two MoMoE approaches. The Llama-based $\texttt{MoMoE}_{\texttt{Community}}$ demonstrates a much wider performance range across target subreddits, with F1 scores spanning from as low as 0.45 for *r/hillaryclinton* with the classification allocation strategy to as high as 0.78 for *r/Overwatch* with the similarity allocation strategy. Overall, $\texttt{MoMoE}_{\texttt{Community}}$ with similarity-based allocation achieves a mean F1 score of 0.67 ($\pm$0.07), while with classification allocation it achieves 0.62 ($\pm$0.07) ($p < 0.01$). In contrast, $\texttt{MoMoE}_{\texttt{NormVio}}$ shows consistent performance across all subreddits, with a much narrower range of F1 scores from 0.58 for *r/pokemontrades* with similarity-based allocation to 0.69 for *r/DestinyTheGame* with classification-based allocation. For $\texttt{MoMoE}_{\texttt{NormVio}}$, both allocation strategies yield similar overall performance with mean F1 scores of 0.64 ($\pm$0.02) for similarity- and 0.64 ($\pm$0.03) for classification-based allocation ($n.s.$). Mistral MoMoE shows very similar trends (See Appendix E).

How do `Allocate` and `Aggregate` operators affect outcomes of `Predict`? In principle, the `Allocate` operator is similar in functionality to a *jury allocator* in the jury learning setting (Gordon et al., 2021) by helping identify which experts and in what proportion should determine MoMoE's prediction. As a result, a natural followup question after observing the strong performance of MoMoE is analyzing what kind of expert compositions are facilitated by the two strategies of classification- and similarity-based allocation. One notable thing about our framework is that if any single expert is allocated a weight of more than 0.5, the decision taken by that expert would be the final one, and therefore in such cases only that expert is required to arrive at an outcome.

We find that since classification-based allocation is essentially a prediction task, on average for $\texttt{MoMoE}_{\texttt{Community}}$ and $\texttt{MoMoE}_{\texttt{NormVio}}$, 72.7% and 69.3% of all predictions on the target subreddits are guided by just one expert respectively. For $\texttt{MoMoE}_{\texttt{Community}}$ the most and least solely-utilized experts were *r/AskReddit* (23.1% of cases) and *r/AskHistorians* (3.2% of cases), while for $\texttt{MoMoE}_{\texttt{NormVio}}$ they were the *'Civility and Respect'* expert (38.5% of cases) and *'Bad Faith or Unsubstantiated Arguments'* expert (2.8% of cases). Similarity-based allocation on the other hand shows the opposite pattern, where at least 3 experts are needed in 87% of all predictions, and the allocation weights are much more uniform in comparison to classification-based allocation.

These insights also provide an explanation for the performance differences between majority voting and dot product based `Aggregate` operators. Specifically for $\texttt{MoMoE}_{\texttt{Community}}$, we see in Figure 2 that majority vote aggregation performs very similarly across both similarity- and classification-based allocation, as despite the difference in allocation distributions, the top experts remain similar across strategies. On the other hand, dot product aggregation is more sensitive to allocation distributions and as a result we see a clear drop in F1 performance from similarity-based to classification-based allocation by ≈ 0.05 and ≈ 0.04 for all experts with Llama and Mistral respectively. This further indicates that with dot-product based aggregation, our ensemble framework can leverage knowledge from multiple experts to get large benefits over a single, dominant expert.

False Positives	
Error Category	**# Instances**
Sarcasm and Jokes	16
Civil Disagreement / Critique	7
Pop Culture / Meme Reference	9
Innocuous Factual Statement	7
Self-Deprecation / Light-weight Banter	3
Ambiguous / Need More Context	8
False Negatives	
Error Category	**# Instances**
Not-so-civil Disagreement	10
Missed Hate Speech and Slurs	8
Social Stereotypes including Sexism	9
Endorsement of Illegal Piracy	11
Off-topic Comments	12

Table 2: **Error Taxonomy Summary.** Prevalence of representative patterns in false positives (FP) and false negatives (FN) from the manual audit of *r/movies*.

6 Error Analysis of MoMoE Decisions

In order to identify and analyze potential failure cases of the moderation decisions made by MoMoE, we perform an error analysis. Understanding of failure modes is crucial to improve the system and deploy it safely in real-world settings.

Specifically, we analyzed results on the target community *r/movies* for the best-performing setting: Mistral-based $\text{MoMoE}_{\text{Community}}$ with classification-based allocation and majority vote based aggregation. We sample 100 instances where MoMoE got the moderation outcome incorrect (50 False Positives and 50 False Negatives) to perform a qualitative error analysis, open-coding each error instance into fine-grained categories.

From Table 2, we find that most false positives arise from sarcasm, jokes, and innocuous or civil disagreement-cases where tone and intent are playful or critical but non-toxic. False negatives on the other hand cluster around not-so-civil disagreement, nuanced hate speech/stereotypes, and policy-violating yet context-light content (e.g., piracy or off-topic posts). This pattern indicates that the model often over-penalizes borderline or humorous language without sufficient pragmatic/contextual cues, yet still under-detects subtle toxicity and norm violations that require fine-grained semantic and socio-linguistic reasoning.

These errors suggest a need for stronger contextualization in the form of conversation/thread history, community context, and better modeling of pragmatic phenomena such as sarcasm, stance, politeness. While the NLP community has long targeted these capabilities, our results underscore that distinguishing civil versus toxic disagreement and detecting nuanced hate speech remain open challenges, warranting future work in this area.

7 MoMoE Explanations

The final component of MoMoE is the **Explain** operator that turns raw model signals into concise, moderator-facing rationales. The operator is designed around three key principles inspired by HCI research in human-AI collaborative systems: **(i) Progressive Disclosure:** provide a one-line verdict first and allow cascading when needed (Choi et al., 2023); **(ii) Reliability:** the explanation is based on the same evidence driving the decision (Selbst et al., 2019b); and **(iii) Actionability:** by surfacing disagreements or low confidence so that moderators know when to intervene (Koshy et al., 2024).

During inference, we log a trace of every expert containing its vote and confidence. Next, we prompt GPT-4o to convert the trace into a three-level JSON explanation with keys **Summary**, **Key Points**, and **Trace**. The *Summary* provides a simple actionable insight with a 'Remove', 'Keep', 'Review' decision, a brief reason inferred from the top expert, and level of 'High' or 'Low' consensus among experts. The *Key Points* provides information about the top expert along with its allocation, and details about consensus on the decision. The *Trace* represents the original trace for audit including the decision and MoMoE confidence-level, along with LLM-inferred salient spans that could possibly indicate problematic areas in the comment. See Appendix D for the prompts, in which we include three-shot exemplars and query GPT-4o with *temperature*=0 to generate explanations.

Summary: Review: Hate Speech; Low Consensus
Key Points:
1. Top expert: 'Civility and Respect' (0.35)
2. Low consensus: 2/5 experts – Review
Trace:
1. Decision: "REMOVE"
2. Confidence: 0.58
3. Salient Spans: ["go back to your country", "lmao"]

We manually sample four explanation samples from each target subreddit totaling 120 examples, and we obtain a 100% reliability, which indicates that given MoMoE's trace, GPT-4o can generate nearly perfect explanations in terms of reliably reflecting the decision processes. The **Explain** operator is therefore model-agnostic, and enhances transparency without overwhelming moderators.

8 Discussion and Implications

Cross-community Content Moderation: MoMoE improves cross-community moderation by pooling knowledge from a small set of specialized experts instead of fine-tuning one model per subreddit. Its architecture leverages the `Allocate` and `Aggregate` operators adapt to unfamiliar communities while retaining high F1 on known ones. This design lowers the data barrier for new or low-resource communities and shows that performance need not be sacrificed for good generalization capabilities. For NLP researchers, our results show the power of lightweight expert ensembles without resorting to generalist models, highlighting the need for research on efficient transfer and dynamic selection of experts.

Complementary Strengths of 'Community' and 'Norm-violation' Experts: With the `Predict` operator, we show that community experts can provide a wider range of performance across unseen communities while norm-violation experts provide consistently strong performance. This shows that while a community ensemble can prove beneficial when the target community has similar content or norms to the source communities, they may struggle to adapt to completely different kinds of communities. An ensemble based on broader norm-violations, on the other hand, may offer consistent performance, as research has shown that such violations are often shared across communities (Chandrasekharan et al., 2018). This suggests that content deemed undesirable in one community is also likely to be considered norm-violating in others.

AI-assisted Content Moderation: Beyond raw accuracy, MoMoE illustrates a practical blueprint for AI-assisted moderation. The `Explain` operator converts model traces into layered, human-readable rationales that surface confidence and expert disagreement, giving moderators clear cues about when to intervene. This workflow shifts the narrative from *automation* to *augmentation*, letting moderators handle edge cases while the system filters routine decisions. For HCI researchers, the framework highlights the value of progressive disclosure and actionable transparency in sociotechnical tools, opening avenues for studying trust calibration and interface design in mixed-initiative governance systems, and also real-time deployment and user-studies using MoMoE.

Directions for future work: Although our offline metrics are promising, they provide an incomplete view of moderator support. Systems such as MoMoE should be assessed with user-centric evaluations that foreground human outcomes (Selbst et al., 2019b). Useful study designs could include within-subjects experiments to assess outcomes such as decision quality, time-to-decision, workload, reliance, and perceived transparency and trust.

Further, moderation systems should not focus solely on removals, but also on supporting *proactive* approaches that elevate identify and reinforce desirable content (Grimmelmann, 2015; Lambert et al., 2024). One direction is to add "prosocial experts" trained to detect prosocial (Bao et al., 2021) and high-quality discourse (Goyal et al., 2025), enabling ranking/boosting policies alongside removal recommendations.

9 Conclusion

We introduce MoMoE, a modular ensemble framework that scales content moderation of online communities beyond resource-intensive community-specific approaches in a transparent manner. MoMoE attains strong F1 scores on 30 unseen communities, matching or surpassing fine-tuned SLM baselines and comprehensively outperforming zero-shot LLMs. MoMoE explanations turn raw decision traces into concise, moderator-ready rationales that were judged reliable in manual inspection. These findings highlight expert ensembles as a viable path toward data-efficient, human-AI collaborative governance. We outline directions for future work on adaptive expert selection, real-time user studies, and deployment of moderation systems at scale.

10 Limitations

Our work has limitations, which also suggest interesting future directions.

(1) Specific choice of in-domain subreddits: We evaluate MoMoE on seven subreddits that were deliberately varied in size, topic, and rule complexity, but we do not test how alternative in-domain sets influence performance. Other community selections or a different mix or count of experts might improve or lower the performance we observe. Because our primary goal was to establish MoMoE as a viable framework, we chose breadth over exhaustive tuning; the reported results therefore serve as a proof-of-concept. Future work should replicate our study with additional subreddit collections and

systematically vary the number and granularity of both community and norm-violation experts.

(2) No exploration of multi-agent LLM frameworks: We deliberately restrict MoMoE to lightweight, single-pass SLM experts rather than a multi-agent setup in which several large LLMs interact or debate. This choice was informed by recent work that found that fine-tuned SLMs surpass zero- and few-shot LLMs in moderation while remaining far cheaper to deploy (Zhan et al., 2025), and it keeps our design focused on data-efficient generalization to unseen communities rather than a multi-agent orchestration. Future work could build human-AI collaborative, multi-agent systems where each agent embodies a community, a moderator persona, or a norm-violation category.

(3) Label noise and annotation bias: Our training and test labels are derived from ground truth moderator actions collected by prior work, which can be inconsistent and influenced by local norms, human biases, or fatigue. This noise may both inflate and depress measured performance. Similarly, while our LLM-based approach for constructing $\mathcal{D}_{\text{NormVio}}$ shows high accuracy and inter-annotator agreement, it is imperfect which could lead to some amount of performance drop.

(4) English-only evaluation: All subreddits in our study are English-speaking. MoMoE's experts and especially the `Explain` operator therefore rely on English language cues. Generalizing to multilingual or code-switched communities will likely demand new experts and prompt adaptation, and is something that future work could explore.

(5) Latency and resource overhead: Although each of our experts is a lightweight $4-$bit quantized SLM, invoking multiple experts plus GPT-4o for explanations adds latency and compute relative to a single classifier. While this is not an issue for deployment as such systems would likely be hosted on a Reddit backend, in high-traffic settings this could raise deployment costs.

(6) Lack of user-centric evaluation: We measure explanation quality with manual validation but do not study how MoMoE affects actual moderation workflows on Reddit, and the trust or decision time of Reddit moderators. Controlled user studies and longitudinal field deployments are needed to validate practical utility and uncover such findings.

Ethical Considerations

MoMoE is targeted at the reduction of harmful content, yet its deployment could raise several ethical questions. First, moderation labels inherited from Reddit may encode community-specific biases. We mitigate this by releasing our code and allowing researchers to audit or retrain experts on alternative annotations. Second, false positives in moderation can censor legitimate speech while false negatives can expose community users to harm, and therefore we design the Explain operator to surface confidence and disagreement so that humans remain in the loop for contentious cases. Finally, since we use GPT-4o-mini and GPT-4o we ensure to comply with the OpenAI API's terms of use policies.[6] We believe that our transparent reporting of limitations, along with the open release of artifacts upon publication will ensure that we minimize introducing any new harms.

Acknowledgments

A.G. was supported by compute credits from the OpenAI Researcher Access Program. This work used the Delta system at the National Center for Supercomputing Applications through allocation #240481 from the Advanced Cyberinfrastructure Coordination Ecosystem: Services & Support (ACCESS) program, which is supported by National Science Foundation grants #2138259, #2138286, #2138307, #2137603, and #2138296.

The authors thank the members of the Social Computing Laboratory (SCUBA) at the University of Illinois and anonymous reviewers of the ACL Rolling Review for their insightful suggestions that helped improve the work.

References

Joshua Ashkinaze, Ruijia Guan, Laura Kurek, Eytan Adar, Ceren Budak, and Eric Gilbert. 2024. Seeing like an ai: How llms apply (and misapply) wikipedia neutrality norms. *arXiv preprint arXiv:2407.04183*.

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*, pages 1134–1145, Ljubljana Slovenia. ACM.

Kenza Benkirane, Jackie Kay, and Maria Perez-Ortiz. 2025. How can we diagnose and treat bias in large

[6] https://openai.com/policies/terms-of-use/

language models for clinical decision-making? In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 2263–2288, Albuquerque, New Mexico. Association for Computational Linguistics.

Chance Castañeda, Jessica Mindel, Will Page, Hayden Stec, Manqing Yu, and Kenneth Holstein. 2025. Supporting ai-augmented meta-decision making with indecision. *arXiv preprint arXiv:2504.12433*.

Eshwar Chandrasekharan, Chaitrali Gandhi, Matthew Wortley Mustelier, and Eric Gilbert. 2019. Crossmod: A Cross-Community Learning-based System to Assist Reddit Moderators. In *Proceedings of the ACM on Human-Computer Interaction*, volume 3, pages 1–30.

Eshwar Chandrasekharan and Eric Gilbert. 2019. Hybrid Approaches to Detect Comments Violating Macro Norms on Reddit. *arXiv preprint*. ArXiv:1904.03596 [cs].

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. In *Proceedings of the ACM on Human-Computer Interaction*, volume 2, pages 1–25.

Eshwar Chandrasekharan, Mattia Samory, Anirudh Srinivasan, and Eric Gilbert. 2017. The Bag of Communities: Identifying Abusive Behavior Online with Preexisting Internet Data. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*, pages 3175–3187, Denver Colorado USA. ACM.

Frederick Choi, Tanvi Bajpai, Sowmya Pratipati, and Eshwar Chandrasekharan. 2023. ConvEx: A Visual Conversation Exploration System for Discord Moderators. In *Proceedings of the ACM on Human-Computer Interaction*, volume 7, pages 262:1–262:30.

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

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

Mitchell L. Gordon, Kaitlyn Zhou, Kayur Patel, Tatsunori Hashimoto, and Michael S. Bernstein. 2021. The Disagreement Deconvolution: Bringing Machine Learning Performance Metrics In Line With Reality. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, pages 1–14, Yokohama Japan. ACM.

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

Agam Goyal, Charlotte Lambert, 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*.

Agam Goyal, Charlotte Lambert, and Eshwar Chandrasekharan. 2025. The language of approval: Identifying the drivers of positive feedback online. *arXiv preprint arXiv:2509.10370*.

James Grimmelmann. 2015. The virtues of moderation. *Yale JL & Tech.*, 17:42. Publisher: HeinOnline.

Laura Hanu and Unitary team. 2020. Detoxify. Github. https://github.com/unitaryai/detoxify.

Andrew F Hayes and Klaus Krippendorff. 2007. Answering the call for a standard reliability measure for coding data. *Communication methods and measures*, 1(1):77–89.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2021. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*.

Evey Jiaxin Huang, Abhraneel Sarma, Sohyeon Hwang, Eshwar Chandrasekharan, and Stevie Chancellor. 2024. Opportunities, tensions, and challenges in computational approaches to addressing online harassment. In *Proceedings of the 2024 ACM Designing Interactive Systems Conference*, DIS '24, page 1483–1498, New York, NY, USA. Association for Computing Machinery.

Tao Huang. 2024. Content moderation by llm: From accuracy to legitimacy. *arXiv preprint arXiv:2409.03219*.

Shagun Jhaver, Iris Birman, Eric Gilbert, and Amy Bruckman. 2019. Human-machine Collaboration for Content Regulation: The Case of Reddit Automoderator. *ACM Transactions on Computer-Human Interaction (TOCHI)*, 26(5):1–35. Publisher: ACM New York, NY, USA.

Shagun Jhaver, Quan Ze Chen, Detlef Knauss, and Amy X Zhang. 2022. Designing Word Filter Tools for Creator-led Comment Moderation. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*, pages 1–21.

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*, pages 3658–3666.

Mahi Kolla, Siddharth Salunkhe, Eshwar Chandrasekharan, and Koustuv Saha. 2024. Llm-mod: Can large language models assist content moderation? In *Extended Abstracts of the 2024 CHI Conference on Human Factors in Computing Systems*, CHI EA '24, New York, NY, USA. Association for Computing Machinery.

Vinay Koshy, Frederick Choi, Yi-Shyuan Chiang, Hari Sundaram, Eshwar Chandrasekharan, and Karrie Karahalios. 2024. Venire: A machine learning-guided panel review system for community content moderation. *arXiv preprint arXiv:2410.23448*.

Deepak Kumar, Yousef Anees AbuHashem, and Zakir Durumeric. 2024. Watch your language: Investigating content moderation with large language models. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 18, pages 865–878.

Charlotte Lambert, Fred 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*, (CSCW).

Hanlin Li, Brent Hecht, and Stevie Chancellor. 2022. Measuring the Monetary Value of Online Volunteer Work. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 16, pages 596–606.

Zhuoyan Li, Zhuoran Lu, and Ming Yin. 2023. Modeling human trust and reliance in ai-assisted decision making: A markovian approach. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pages 6056–6064.

Zhuoyan Li, Hangxiao Zhu, Zhuoran Lu, Ziang Xiao, and Ming Yin. 2025. From text to trust: Empowering ai-assisted decision making with adaptive llm-powered analysis. *arXiv preprint arXiv:2502.11919*.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Kiel Long, John Vines, Selina Sutton, Phillip Brooker, Tom Feltwell, Ben Kirman, Julie Barnett, and Shaun Lawson. 2017. " could you define that in bot terms"? requesting, creating and using bots on reddit. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*, pages 3488–3500.

Ilya Loshchilov and Frank Hutter. 2017. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*.

Mistral AI. 2024. Mistral NeMo — mistral.ai. https://mistral.ai/news/mistral-nemo/. [Accessed 12-10-2024].

Rachel Elizabeth Moran, Joseph Schafer, Mert Bayar, and Kate Starbird. 2025. The end of trust and safety?: Examining the future of content moderation and upheavals in professional online safety efforts. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*, pages 1–14.

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)*, pages 561–573.

OpenAI. 2024. Hello GPT-4o. https://openai.com/index/hello-gpt-4o/.

Konstantina Palla, José Luis Redondo García, Claudia Hauff, Francesco Fabbri, Henrik Lindström, Daniel R Taber, Andreas Damianou, and Mounia Lalmas. 2025. Policy-as-prompt: Rethinking content moderation in the age of large language models. *arXiv preprint arXiv:2502.18695*.

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. *arXiv preprint arXiv:2110.04419*.

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):1–29.

Junyeong Park, Seogyeong Jeong, Seyoung Song, Yohan Lee, and Alice Oh. 2025. Llm-c3mod: A human-llm collaborative system for cross-cultural hate speech moderation. *arXiv preprint arXiv:2503.07237*.

Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence embeddings using Siamese BERT-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3982–3992, Hong Kong, China. Association for Computational Linguistics.

Sarah T Roberts. 2019. *Behind the screen*. Yale University Press.

Andrew D. Selbst, Danah Boyd, Sorelle A. Friedler, Suresh Venkatasubramanian, and Janet Vertesi. 2019a. Fairness and abstraction in sociotechnical systems. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*, FAT* '19, page 59–68, New York, NY, USA. Association for Computing Machinery.

Andrew D. Selbst, Danah Boyd, Sorelle A. Friedler, Suresh Venkatasubramanian, and Janet Vertesi. 2019b. Fairness and Abstraction in Sociotechnical

Systems. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*, pages 59–68, Atlanta GA USA. ACM.

Mark Steyvers and Aakriti Kumar. 2024. Three challenges for ai-assisted decision-making. *Perspectives on Psychological Science*, 19(5):722–734.

Oleksandra Vereschak, Fatemeh Alizadeh, Gilles Bailly, and Baptiste Caramiaux. 2024. Trust in ai-assisted decision making: Perspectives from those behind the system and those for whom the decision is made. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems*, pages 1–14.

Nishant Vishwamitra, Keyan Guo, Farhan Tajwar Romit, Isabelle Ondracek, Long Cheng, Ziming Zhao, and Hongxin Hu. 2024. Moderating new waves of online hate with chain-of-thought reasoning in large language models. In *2024 IEEE Symposium on Security and Privacy (SP)*, pages 788–806. IEEE.

Bernard L Welch. 1947. The generalization of 'student's' problem when several different population varlances are involved. *Biometrika*, 34(1-2):28–35.

Xianyang Zhan, Agam Goyal, Yilun Chen, Eshwar Chandrasekharan, and Koustuv Saha. 2025. SLM-mod: Small language models surpass LLMs at content moderation. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 8774–8790, Albuquerque, New Mexico. Association for Computational Linguistics.

A Target Subreddits

In this section, we list the 30 subreddits we used as our target communities for downstream evaluation of MoMoE. These communities were randomly sampled from the original comment removal dataset released by Chandrasekharan et al. (2019) after removing the source communities our in-domain experts were fine-tuned on. Sampling these subreddits from the original dataset was crucial as we needed ground truth removal labels for evaluation.

r/food, r/PoliticalDiscussion, r/hearthstone, r/OldSchoolCool, r/gonewild, r/spacex, r/WTF, r/pokemongo, r/DestinyTheGame, r/BlackPeopleTwitter, r/nottheonion, r/Overwatch, r/pokemontrades, r/explainlikeimfive, r/IAmA, r/personalfinance, r/hillaryclinton, r/news, r/leagueoflegends, r/funny, r/toronto, r/depression, r/pcmasterrace, r/OutOfTheLoop, r/HistoryPorn, r/ShitRedditSays, r/asoiaf, r/relationships, r/nba, r/movies.

B Norm Violation Dataset Creation

As outlined in the main text, we use an LLM-based approach to create our norm-violations dataset $\mathcal{D}_{\text{NormVio}}$ using GPT-4o. We use the prompt below asking the LLM to classify each instance of rule-violating comment into a specific set of rules it violated. We then augment this "positive" class with a sample of non-violating comments to form a balanced dataset for fine-tuning experts with an 80% split, while the rest is used for testing.

PROMPT

```
You are an expert moderator for the
r/{SUBREDDIT} subreddit on Reddit.
Here is a description of the subreddit:
{SUBREDDIT_DESCRIPTION}
You are given a comment, the preceding
context which it is replying to, and a list
of rules for the subreddit.
This comment was removed by the moderators
of the subreddit, and your task is
to determine which rule(s) the comment
violates.

Context:
{CONTEXT}

Comment:
{COMMENT}

Rules:
{RULES}

Please return the list of rule number(s)
that the comment violates in a list format:
e.g., 5, 7, 9.
If the comment violates only one rule,
return a list with one element: e.g., 9.
Even if you think the comment violates no
rules or you are not sure, return the rule
it is most likely to violate and nothing
else.
```

Once the LLM classified comments into specific rules, we then manually grouped the rules from all in-domain subreddits in broader norm-violation categories, ending up with 5 themes listed below. To ensure annotation quality, we have the first and second authors manually validate a sample of 140 comments (20 from each source subreddit) obtaining an accuracy of 87%, and an inter-rater reliability (IRR) Krippendorff's $\alpha = 0.82$ which denotes high agreement (Hayes and Krippendorff, 2007).

Norm-Violation Mapping

Civility and Respect

- *r/science*: 2. No abusive or offensive comments
- *r/politics*: 3. No incivility or personal attacks towards users
- *r/politics*: 4. No flaming, baiting, or trolling
- *r/AskReddit*: 2. Be respectful to other users at all times
- *r/AskHistorians*: 1. Users shall behave with courtesy

- *r/changemyview*: 2. No rude/hostile comment
- *r/AskHistorians*: 9. No racist or bigoted comments
- *r/politics*: 1. No hateful speech
- *r/Games*: 2. No attacks / witch-hunts / bigotry / inflammatory language

Low-Effort, Off-Topic, or Non-Substantive Contributions

- *r/science*: 1. No off-topic comments, memes, low-effort jokes
- *r/Games*: 3. No off-topic or low-effort content
- *r/anime*: 2. No memes, reaction images, shitposts
- *r/anime*: 1. Everything posted must be anime-specific
- *r/Games*: 1. No content primarily for humor or entertainment
- *r/changemyview*: 5. No comment that doesn't contribute meaningfully
- *r/AskHistorians*: 2. Comments must be in-depth and comprehensive

Bad-Faith or Unsubstantiated Arguments

- *r/science*: 3. Non-professional personal anecdotes will be removed
- *r/science*: 4. Criticism must assume basic competence of researchers
- *r/science*: 5. Dismissing established findings must provide evidence
- *r/AskHistorians*: 3. Comments should reflect topic familiarity
- *r/AskHistorians*: 4. No speculative or anecdotal comments
- *r/changemyview*: 3. No bad-faith accusation
- *r/anime*: 4. Do not post heavily NSFW content

Spam, Solicitation, Misinformation, Machine-Generated Content

- *r/politics*: 2. No spam or solicitation
- *r/AskReddit*: 4. No spam, machine-generated content, or karma farming
- *r/anime*: 5. No spam, low-effort comments, or karma farming
- *r/AskHistorians*: 6. Comments should not be only links or quotations
- *r/AskHistorians*: 7. Must cite all quotes; no plagiarism
- *r/AskHistorians*: 8. Comments should not consist solely of jokes
- *r/AskReddit*: 3. No harmful misinformation

Rule Enforcement and Structural Integrity of Discussions

- *r/AskReddit*: 1. No begging for goods, services, or awards
- *r/anime*: 3. Do not link to illegal content
- *r/AskHistorians*: 5. No political agendas or moralising
- *r/changemyview*: 4. No delta abuse or misuse
- *r/changemyview*: 1. Top-level comments must challenge OP
- *r/science*: 6. No medical advice

C Fine-tuning Details

For community-based and norm-violation-based SLMs, we perform Low-Rank Adaptation (LoRA) fine-tuning (Hu et al., 2021) for 1 epoch on balanced samples from the positive and negative class labels. We use rank $r = 16$, $\alpha = 32$. We do not use any dropout. Further, we use a learning rate $lr = 2e - 4$ with a linear schedule and 5 warmup steps, and the AdamW (Loshchilov and Hutter, 2017) optimizer with a weight decay of 0.01.

For the Global SLM baselines, due to the significantly larger amount of data these models are fine-tuned on, we perform 2 epochs of fine-tuning, keeping all other hyperparameters intact. We now outline our fine-tuning prompts for each scenario, inspired by Zhan et al. (2025).

Community Expert SLM Prompt

```
You are acting as a moderator for the
r/{SUBREDDIT} subreddit. You will be given
a comment from Reddit and the rules deemed
suitable to arrive at a moderation outcome,
and your task is to determine if the given
text is undesirable or not based on the
information provided to you.

Here is a comment from a Reddit conversation
thread, the context (preceding comment),
and the associated subreddit rules.

Context:
{CONTEXT}

Comment:
{COMMENT}

Rules:
{RULES}

Determine whether the provided text is
undesirable or not. Answer with 'True' or
'False'.

### Your Response:
```

Norm-Violation Expert SLM Prompt

```
You are acting as a moderator for the
r/{SUBREDDIT} subreddit. You will be
given a comment from Reddit and the
community norm deemed suitable to arrive
at a moderation outcome, and your task is
to determine if the given text violates
the provided norm or not based on the
information provided to you.

Here is a comment from a Reddit conversation
thread, the context (preceding comment),
and the associated community norm.

Context:
{CONTEXT}

Comment:
```

```
{COMMENT}

Norm:
{NORM}

Determine whether the provided text is
undesirable or not. Answer with 'True' or
'False'.


### Your Response:
```

Finally, for the RoBERTa-base models fine-tuned for classification as part of the `Allocate` operators, we use fine-tune the model for 3 epochs with a learning rate $lr = 1e-5$ and maximum sequence length of 512 tokens.

D Explanation Prompt Design

In this section, we report our prompt used for generating MoMoE explanations:

MoMoE$_{\text{NormVio}}$ Explain Prompt

System:
You are "MoMoE-Explain", an assistant that writes short, moderator-facing rationales for AI-based content moderation decisions.
- Audience: Experienced Reddit moderators.
- Style: concise, neutral, no technical jargon, no private model thoughts.
- Output JSON keys in this exact order: Summary, Key Points, Trace.

User:
Here is the decision trace for a comment:
{TRACE}

Generate:
1. **Summary**: ≤ 25 words stating outcome recommendation ('Remove', 'Review', 'Keep'), Key norm violated, Consensus-level among experts ('Low', 'High').

2. **Key Points:** 2 bullet points (≤ 10 words each) covering:
- Top expert: <Name_of_Expert> (<Weight>)
- <Level_of_Consensus> consensus: X/5 experts - <Recommendation>

3. **Trace:**
- Decision: "<Decision>"
- Confidence: <MoMoE confidence_score>
- Salient Spans: ["<span_1>", "<span_2"]

For 'Salient Spans', identify upto three specific sequence within the comment that likely led to the moderation outcome, keeping in mind the top experts that voted for its removal. If the outcome is to 'Keep' the comment, leave the Salient Span list empty.

Respond only with valid JSON.

We provide the model with three-shot exemplars of a TRACE and generated explanations, covering all decision cases in order to provide the model with the precise format expected.

E Further Discussion on MoMoE Performance Across Target Subreddits

In this section, we provide a deeper discussion of the performance and differences between Mistral-based MoMoE$_{\text{Community}}$ and MoMoE$_{\text{NormVio}}$ on target communities, shown in Figure 3. All significance levels are from Welch's $t-$test (Welch, 1947).

The Mistral-based MoMoE$_{\text{Community}}$, similar to the case of Llama, shows a much wider range of performance, with F1 scores ranging from 0.52 for *r/hillaryclinton* with the classification allocation strategy to as high 0.81 for *r/Overwatch* with the similarity allocation strategy. Overall, MoMoE$_{\text{Community}}$ with similarity allocation achieves a mean F1 score of 0.71 (± 0.06), while with classification allocation it achieves 0.67 (± 0.07) ($p < 0.05$). MoMoE$_{\text{NormVio}}$ again shows contrasting, consistent performance across all subreddits, with F1 scores from 0.66 for *r/gonewild* with classification-based allocation to 0.72 for *r/PoliticalDiscussion* with classification-based allocation. Both allocation strategies for MoMoE$_{\text{NormVio}}$ yield similar overall performance with mean F1 scores of 0.67 (± 0.00) for similarity and 0.67 (± 0.01) for classification-based allocation ($n.s.$).

F Precision-Recall Trade-offs

We saw that both MoMoE$_{\text{Community}}$ and MoMoE$_{\text{NormVio}}$ perform competitively in comparison to baselines, while MoMoE$_{\text{Community}}$ generally outperforms MoMoE$_{\text{NormVio}}$. We discuss here two kinds of precision-recall trade-offs with MoMoE in Figure 4.

First, we observe that the recall of both MoMoE$_{\text{Community}}$ and MoMoE$_{\text{NormVio}}$ is higher than their precision, which in combination with our existing results highlights that MoMoE may be overaggressive, flagging potentially violating comments rather than erring on the side of caution. In contrast, the precision and recall of LLMs on the target communities was 0.78 and 0.44 for GPT-4o-mini, and 0.82 and 0.38 for GPT-4o, respectively. This observation is in line with that of Zhan et al. (2025), who found that SLMs prioritize potentially harmful content even at the cost of over-flagging.

Second, within the two types of ensembles MoMoE$_{\text{Community}}$ and MoMoE$_{\text{NormVio}}$ we observe that on recall, MoMoE$_{\text{NormVio}}$ outperforms

	Subreddit	Imb. (%)	Llama AUC	Llama F1	Mistral AUC	Mistral F1	G-Llama AUC	G-Llama F1	G-Mistral AUC	G-Mistral F1
Maj. Vote	r/Overwatch	5	0.87 ± 0.02	0.46 ± 0.01	**0.90 ± 0.02**	**0.49 ± 0.01**	0.82 ± 0.02	0.44 ± 0.01	0.85 ± 0.02	0.47 ± 0.01
	r/hillaryclinton	5	0.60 ± 0.06	0.44 ± 0.02	**0.71 ± 0.04**	**0.47 ± 0.02**	0.55 ± 0.05	0.41 ± 0.02	0.66 ± 0.04	0.45 ± 0.02
	r/Overwatch	10	0.86 ± 0.01	0.52 ± 0.01	**0.89 ± 0.02**	**0.56 ± 0.01**	0.81 ± 0.02	0.50 ± 0.01	0.84 ± 0.02	0.54 ± 0.01
	r/hillaryclinton	10	0.60 ± 0.03	0.47 ± 0.01	**0.70 ± 0.02**	**0.51 ± 0.01**	0.55 ± 0.03	0.44 ± 0.01	0.65 ± 0.02	0.48 ± 0.01
Dot Prod.	r/Overwatch	5	0.74 ± 0.01	0.43 ± 0.01	**0.76 ± 0.01**	**0.44 ± 0.01**	0.69 ± 0.01	0.40 ± 0.01	0.71 ± 0.01	0.42 ± 0.01
	r/hillaryclinton	5	0.57 ± 0.05	0.34 ± 0.01	**0.63 ± 0.02**	**0.31 ± 0.01**	0.52 ± 0.04	0.31 ± 0.02	0.58 ± 0.03	0.28 ± 0.02
	r/Overwatch	10	0.73 ± 0.01	0.49 ± 0.01	**0.75 ± 0.01**	**0.51 ± 0.01**	0.68 ± 0.02	0.46 ± 0.02	0.70 ± 0.01	0.48 ± 0.02
	r/hillaryclinton	10	0.58 ± 0.02	0.38 ± 0.01	**0.63 ± 0.02**	**0.36 ± 0.01**	0.53 ± 0.02	0.35 ± 0.01	0.58 ± 0.01	0.33 ± 0.01

Table 3: Performance in terms of AUC and Macro-F1 of Llama- and Mistral-based $\texttt{MoMoE}_{\texttt{Community}}$ models under majority-vote and dot-product aggregation on class-imbalanced test splits (5 % and 10 % "removed" labels). "Global SLMs" denote the strongest single-model baselines, fine-tuned on $\mathcal{D}_{\text{Community}}$. Values are mean ± sd over 10 runs.

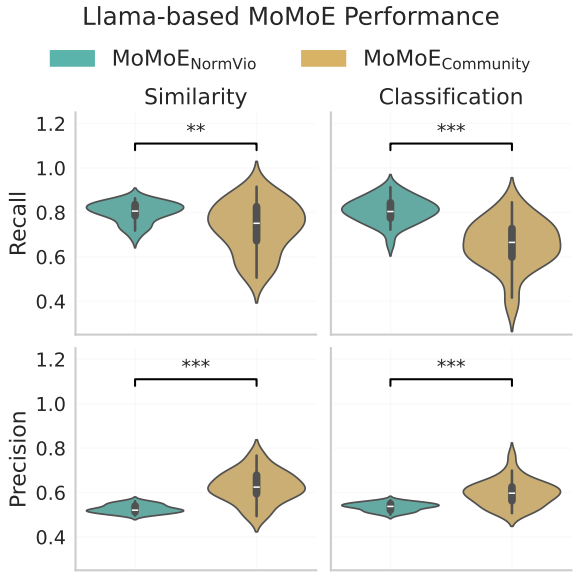

Figure 4: Comparison of precision-recall trade-offs with Llama-based $\texttt{MoMoE}_{\texttt{Community}}$ with $\texttt{MoMoE}_{\texttt{NormVio}}$ using a dot-product aggregation. We observe that $\texttt{MoMoE}_{\texttt{NormVio}}$ has higher recall compared to $\texttt{MoMoE}_{\texttt{Community}}$ (mean difference $\approx$ 0.06), whereas in terms of precision, $\texttt{MoMoE}_{\texttt{Community}}$ outperforms $\texttt{MoMoE}_{\texttt{NormVio}}$ (mean difference ≈ 0.08). ($^{*}p < 0.05$, $^{**}p < 0.01$, $^{***}p < 0.001$)

$\texttt{MoMoE}_{\texttt{Community}}$, with a recall of $0.80(\pm 0.05)$ compared to that of $\texttt{MoMoE}_{\texttt{Community}}$ at $0.73(\pm 0.11)$ for similarity-based allocation, and $0.81(\pm 0.06)$ in comparison to $0.66(\pm 0.11)$ for $\texttt{MoMoE}_{\texttt{Community}}$. With precision on the other hand, we observe that $\texttt{MoMoE}_{\texttt{Community}}$ outperforms $\texttt{MoMoE}_{\texttt{NormVio}}$, with a precision of $0.63(\pm 0.07)$ compared to that of $\texttt{MoMoE}_{\texttt{NormVio}}$ at $0.53(\pm 0.02)$ for similarity-based allocation, and $0.60(\pm 0.06)$ in comparison to $0.54(\pm 0.02)$ for $\texttt{MoMoE}_{\texttt{NormVio}}$. All differences are significant by Welch's $t-$test. This highlights that although both ensembles are over-aggressive at removing comments, this tendency is particularly enhanced in $\texttt{MoMoE}_{\texttt{NormVio}}$.

For practitioners, this means that the `Predict` component $\texttt{MoMoE}_{\texttt{NormVio}}$ in a standalone manner is more suited for a comment triaging scenario where a human moderator will oversee decisions, rather than automated moderation settings. This would ensure that community members are not wrongfully punished with their benign comments being removed. We see the same trend with Mistral-based MoMoE as well.

G Performance on Imbalanced Test Split

In this section we report the performance of the best performing `Predict` configuration, $\texttt{MoMoE}_{\texttt{Community}}$ on imbalanced test split on the worst (*r/hillaryclinton*) and best performing (*r/Overwatch*) subreddits. Table 3 shows performance of Llama- and Mistral-based $\texttt{MoMoE}_{\texttt{Community}}$ compared to the best performing baseline of G-$\text{Llama}_{\text{Community}}$ and G-$\text{Mistral}_{\text{Community}}$ in terms of AUC and Macro-F1 scores. Prior work has shown that the proportion of comments that actually violate community norms in the real world are around 6-7% of all comments (Park et al., 2022). We therefore test on two imbalance thresholds of 5% and 10% violating comments, and the remaining non-violating comments over 10 random seeds. Since both classification and similarity-based allocation performed very similarly in the case of $\texttt{MoMoE}_{\texttt{Community}}$, we report here the results for classification-based allocation.

Similar to the case of balanced test split, we observe that Mistral-based $\texttt{MoMoE}_{\texttt{Community}}$ performs the best in terms of both AUC and F1 scores, followed by Llama-based $\texttt{MoMoE}_{\texttt{Community}}$, both of which outperform the Global SLM baselines. We also again observe that majority vote based aggregation works better than dot product aggregation. These results indicate that MoMoE continues to show superior performance on target communities even under a more realistic imbalanced data scenario.

H Compute Resources

All experiments on open-source models were run on internal GPU servers equipped with 4xNVIDIA A100 and 3xNVIDIA A40. The experiments with the OpenAI models cost about 100 USD.