SLM-Mod: Small Language Models Surpass LLMs at Content Moderation

Anonymous ACL submission

Abstract

Large language models (LLMs) have shown promise in many natural language understanding tasks, including content moderation. However, these models can be expensive to query in real-time and do not allow for a community-specific approach to content moderation. To address these challenges, we explore the use of open-source small language models (SLMs) for community-specific content moderation tasks. We fine-tune and evaluate SLMs (less than 15B parameters) by comparing their performance against much larger open- and closed-sourced models. Using 150K comments from 15 popular Reddit communities, we find that SLMs outperform LLMs at content moderation—11.5% higher accuracy and 25.7% higher recall on average across all communities. We further show the promise of cross-community content moderation, which has implications for new communities and the development of cross-platform moderation techniques. Finally, we outline directions for future work on language model based content moderation.[1]

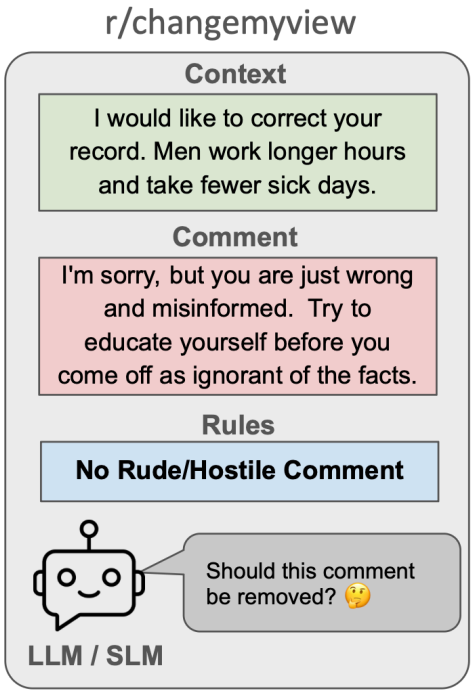

Figure 1: **Online Moderation with Language Models.** Given a comment from a subreddit *r/changemyview*, the preceding context, and community rules, we compare LLMs and SLMs on moderation performance.

1 Introduction

Content moderation has become a growing area of interest for the NLP community (Jurgens et al., 2019) due to the rapid growth and use of social media. The primary challenge in content moderation lies in detecting undesirable, norm-violating behavior amidst vast amounts of content posted by users. In order to deal with large volumes of content, most platforms rely on automated tools to either directly remove norm-violating content or triage undesirable content for manual review by human moderators (Chandrasekharan et al., 2019).

Although moderation involves dealing with a wide range of undesirable behaviors, current computational approaches tend to adopt narrow definitions of abuse (e.g., toxicity, hate speech) (Jurgens et al., 2019). However, content moderation is a nuanced and contextual task that varies across platforms, communities, and over time—e.g., a post deemed desirable in one community may be undesirable in another (Chandrasekharan et al., 2018). Moderation requires an understanding of community norms, and depend on the manual labor and judgment of human moderators, who are often overworked and uncompensated (Li et al., 2022). Despite these variations between communities, current approaches do not effectively incorporate community norms when detecting undesirable content.

Due to the natural language understanding capabilities of large language models (LLMs) (Radford et al., 2019; Brown, 2020), recent work has explored the use of LLMs for content moderation (Ku-

[1] Code & models will be publicly released after acceptance.

mar et al., 2024; Kolla et al., 2024). Although LLMs can be used off-the-shelf and generate explanations for moderation decisions, moderators would still require the ability to dynamically adapt these models to suit the preferences and norms of their community. In other words, there is a need for specialized models that provide moderators with more control and configurability via fine-tuning.

Current methods for fine-tuning LLMs are constrained by the substantial computational resources and costs required, making it nearly impossible for moderators to tailor LLMs for community-specific content moderation. Given these limitations, a viable alternative could be the use of small language models (SLMs), which provide a lightweight and cost-effective option compared to LLMs. Recent research has shown that despite their smaller size, SLMs can be fine-tuned to achieve performance comparable to LLMs on various natural language understanding tasks (Schick and Schütze, 2021).

This work explores the potential for using SLMs in content moderation, evaluating their performance relative to LLMs in terms of accuracy, recall, and precision across multiple online communities. In particular, we focus on whether fine-tuned SLMs can offer a more resource-efficient approach without compromising on moderation quality. We aim to determine if SLMs can strike a balance between cost-effectiveness and moderation accuracy, ultimately providing a viable alternative to the computationally expensive LLMs for large-scale, community-specific moderation.

Findings. Our findings reveal that SLMs outperform LLMs at in-domain content moderation tasks having both higher accuracy and recall, with slightly lower precision. We present a case study of *r/changemyview* and conduct an error analysis on false positives and false negatives to identify trade-offs when using SLMs over LLMs. Next, we highlight the potential of SLMs for cross-domain content moderation tasks and investigate possible reasons for their high performance. Finally we discuss the implications of our findings for content moderation and highlight future directions for improving cross-community approaches to moderation as well as dynamically adapting community-specific models over time.

2 Related Work

Automated Approaches to Content Moderation Human moderation is limited in scalability, especially as platforms like Facebook, YouTube, and Reddit experience exponential growth in content volume. To address these limitations, automated approaches have been increasingly adopted. Automated approaches typically use natural language processing and machine learning techniques to flag potentially harmful content for human review (Chandrasekharan et al., 2019). For example, n-gram models and sentiment analysis techniques have been used as classic automated approaches to classify content as toxic or non-toxic (Davidson et al., 2017; Vigna et al., 2017; Warner and Hirschberg, 2012). These methods rely on identifying patterns of word usage and sentiment to make judgments about the harmfulness of content, providing a foundational approach for content moderation. More advanced methods, including deep learning models networks to automatically learn multi-layers of abstract features from raw data have been explored for detecting offensive content (Nobata et al., 2016; Badjatiya et al., 2017; Zhang and Luo, 2018). Recently, Jha et al. (2024) proposed an LLM and VLM-based framework for online content moderation via meme interventions, and Maity et al. (2023) developed a generative framework for explainable cyber-bullying detection.

LLM-Assisted Content Moderation The rise of large language models (LLMs) and their successors has transformed content moderation by enhancing the ability to detect harmful content with greater contextual awareness. These models excel at processing longer texts and capturing nuanced meanings, making them highly effective for identifying hate speech, misinformation, and abusive language on social platforms (Kolla et al., 2024). However, despite their advancements, LLMs are computationally expensive to run, making them less feasible for real-time moderation at scale (Kumar et al., 2024; Lai et al., 2022) In light of these limitations, our study investigates whether SLMs, which are more lightweight and cost-effective (Irugalbandara et al., 2024), can be fine-tuned to handle moderation tasks with comparable accuracy, potentially offering a balance between cost and effectiveness in online content moderation.

3 Experimental Setup

3.1 Data Curation

We curate our data from the publicly available dataset of Reddit comment removals between

May 10th, 2016 and February 4th, 2017 by Chandrasekharan et al. (2018) by sampling 10K comments (5K moderated and 5K unmoderated) for 15 popular subreddits from Reddit's landing page. For each subreddit, we split its data into 80/20 train/test sets. Appendix A includes the complete list, description, and subscriber statistics of the subreddits.

3.2 Models and Configuration

For our study, we evaluate the $4-$bit quantized versions of three small language models (SLMs): Llama-3.1-8b (Dubey et al., 2024), Gemma-2-9b (Team et al., 2024), and Mistral-nemo-instruct (Mistral AI, 2024), and three large language models (LLMs): Cohere's Command R+ (Cohere For AI, 2024), OpenAI's GPT-4o; and GPT-4o-mini (OpenAI, 2024). For the LLMs, a temperature of 0 was used to ensure consistency across runs, and a `top_p` of 0.75 was used.

We fine-tune 15 subreddit-specific SLMs using rank 16 Low-Rank Adaptation (LoRA) (Hu et al., 2021) for 1 epoch on a balanced sample of total 8,000 moderated and unmoderated comments.

3.3 Task

The task for the language models is to determine a moderation outcome for a given comment when provided with the comment, its preceding context, and subreddit rules. For a language model $\mathcal{M}$ and a comment $\mathcal{T}$ along with its context $\mathcal{C}$, and subreddit rules $\mathcal{R}$, we prompt the model with a prompt p. Consequently, the moderation task is:

$$\mathcal{D}_{\mathcal{M}} \leftarrow \mathcal{M}_p(\mathcal{T}, \mathcal{C}, \mathcal{R}) \quad (1)$$

where $\mathcal{D}_{\mathcal{M}}$ represents the moderation decision generated by the language model $\mathcal{M}$.

We define two sets of tasks: (a) **In-domain Moderation**, which involves moderating comments from the same subreddit on which the model was fine-tuned; (b) **Cross-domain Moderation**, which involves moderating comments from a subreddit different from the one on which the model was fine-tuned. Note that this distinction is primarily useful for SLMs. For off-the-shelf LLMs, the prompt is subreddit-specific but the models themselves are not, so we always prompt with the correct rules.

3.4 Prompts

We use rule-based prompting, drawing on prior work (Kumar et al., 2024), by querying the models with text, context (preceding comment), and rules from each subreddit using the following prompt p:

```
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.
### Text: {TEXT}
### Context: {CONTEXT}
### Rules: {RULES}
Determine whether the provided text is undesirable or not. Answer with 'True' or 'False'.
```

We obtained rules for each subreddit by querying Reddit using the PRAW API.[2] For in-domain tasks, we use the rules of the original subreddit, whereas, for cross-domain tasks, we use the rules of the source subreddit since we assume that we do not have rules for the target subreddit.

3.5 Evaluation Metrics

For our evaluation tasks, we focus on the metrics of *accuracy*, *precision*, and *recall*. These give us a holistic picture of the efficacy of different models on content moderation tasks, along with an insight into how different models handle violating comments in terms of the precision/recall trade-off.

4 In-domain Moderation

4.1 Performance Comparison

In this section, we compare the performance of SLMs versus LLMs on in-domain content moderation tasks. Figure 2 provides a visual depiction of the accuracy, recall, and precision across models.

Accuracy We evaluate the content moderation performance of each model by comparing it to human moderators' decisions. We find that the fine-tuned SLMs outperform the LLMs by an average 11.5% in accuracy across 15 subreddits. Among the fine-tuned SLMs, Mistral-Nemo-instruct, Gemma-2-9b, and Llama-3.1-8b show accuracies of 77.87%, 77.2%, and 72.5% respectively. Among LLMs, GPT-4o and GPT-4o-mini are the highest-performing models, both with an average accuracy of 65.8%. SLMs' performance is consistently superior, exhibiting higher accuracy than the LLMs across all subreddits. The highest and lowest average accuracy is achieved by the SLMs on *r/changemyview* (90.97%) and *r/AskReddit* (60.4%), and by the LLMs on *r/movies* (77.4%) and *r/askscience* (53.4%).

[2] https://praw.readthedocs.io/en/stable/

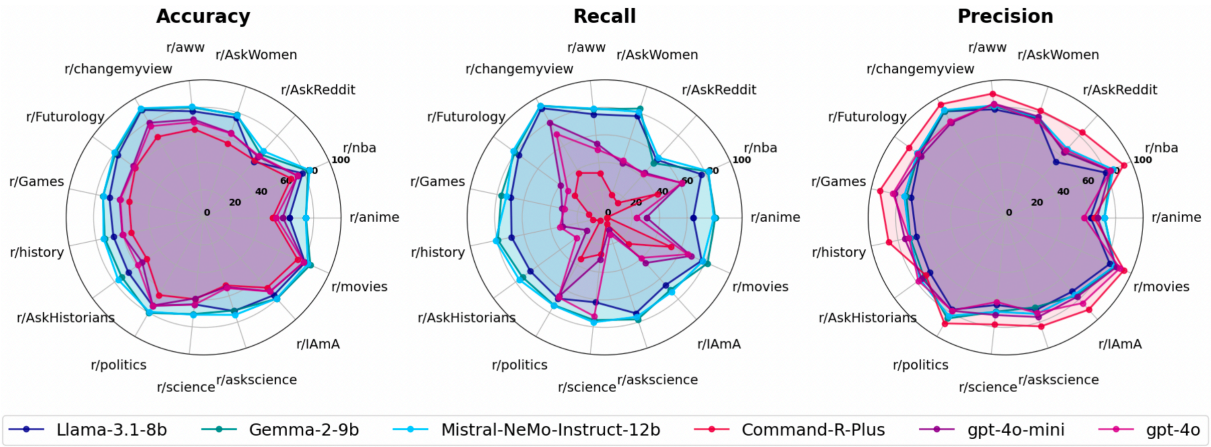

Figure 2: **In-domain Moderation Performance.** Comparing the performance of SLMs versus LLMs on accuracy, recall, and precision for in-domain content moderation performance. Best performing SLMs outperform LLMs on accuracy and recall across all subreddits, while LLMs outperform SLMs on precision.

Precision/Recall Trade-off We find that SLMs demonstrate the highest recall in all 15 subreddits, having an average recall of 77.7% for Mistral, 77.5% for Gemma, and 71.5% for Llama. On the other hand, LLMs achieve the highest precision in 14 out of 15 subreddits, with the highest average precision being for Command R+ at 85.5%. When comparing the top performers in both metrics, LLMs show an average 8% advantage in precision, meaning they are more accurate in identifying non-harmful content and avoiding false positives. On the other hand, SLMs have a 22% lead in recall, indicating their superior ability to identify harmful content and reduce false negatives. This contrast reveals a notable trade-off: *LLMs excel in minimizing over-flagging of content, whereas SLMs prioritize flagging harmful content even at the expense of more false positives*.

4.2 Error Analysis: A case study on `r/changemyview`

In order to get a deeper understanding of how SLMs tackle content moderation and where they falter, we complement our quantitative results with a qualitative analysis. However, due to the scale and complexity of the data, qualitative analysis of the entire dataset is infeasible. Therefore, we focus on one specific community, *r/changemyview* that has been a community of interest in various prior content moderation works (Srinivasan et al., 2019; Koshy et al., 2023; Jhaver et al., 2017). *r/changemyview* is a subreddit with 3.7M subscribers for debating opinions where users invite others to challenge their perspectives with thoughtful counterarguments, and if their view is changed, then they can award a *delta* (Δ) to a commenter. Moreover, our fine-tuned SLMs showed strong overall performance on this subreddit, achieving an average of 91% accuracy, 90% precision, and 93% recall, so it would be valuable to further examine the few niche error cases to conduct additional exploration. These factors make it an interesting subreddit to qualitatively compare the differences between content moderation with SLMs versus LLMs, where looking at the false positives and negatives offers valuable insights into the types of errors made by the models.

We retrieved all comments from the test set where the SLMs made an error and inspected them both manually and computationally. Overall, there were 15.2% (152/1000) false positives and 11.5% false negatives (115/1000) in the moderation outcomes where at least one SLMs made a mistake.

Impact of Content Length Figure 3 depicts the probability of occurrences of false positive and false negative instances as the number of words in the comment increases. For false positives, we observe that at median comment length, the probability of an SLM moderating the comment incorrectly is around 0.6, while for GPT-4o and GPT-4o-mini, this probability is only 0.4. Command R+ is an exception to the case of LLMs with a probability around 0.65. This means that SLMs are likely to make mistakes and moderate short comments more aggressively compared to LLMs. Qualitatively, we also observe that shorter comments seem to confuse the models more often in terms of false positives. For example, a comment *"Very succinctly stated"*, which was in reply to a comment that has

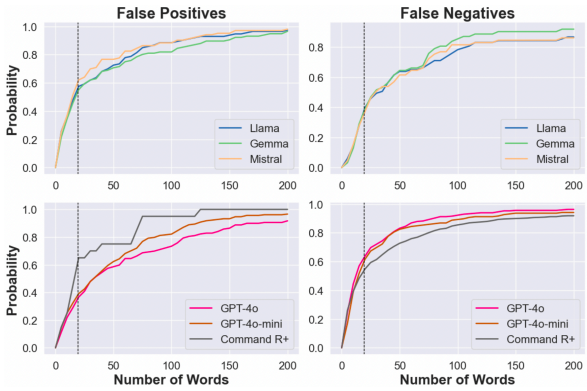

Figure 3: **Impact of content length.** Probabilities of the mistakes (FP and FN) made by SLMs and LLMs on varying comment length (in words) in `r/changemyview` reveals that SLMs tend to over-moderate shorter comments whereas LLMs are more forgiving for the same. The vertical bar indicates median length of comments in *r/changemyview* at 19 words.

more than 300 words was incorrectly moderated by the SLM, and while a human moderator would understand that this comment is probably said in jest, the SLM might confuse these comments with violating a rule like *"Don't be rude or hostile to other users"*. LLMs, on the other hand, are better suited at handling these kinds of comments and are able to perceive the intended meaning, as all three LLMs correctly left this comment unmoderated.

On the other hand, looking at the false negatives provides us with an opposite observation. The probability of an SLM getting a false negative at median comment length is around 0.4 while for LLMs this probability is around 0.6, with Command R+ having a slightly lower probability. This indicates that short and undesirable comments are well moderated by the SLMs, but are missed by LLMs at a much higher rate.

In both cases, we see that Command R+ performs a bit more like SLMs versus LLMs. While differences in the training data could play a role, it is possible that the model size plays a role in content moderation performance, and with 104B parameters Command R+ can be perceived as being closer to SLMs than to LLMs.

Impact of Content Topic Next, we investigate the impact of content topics that might cause a difference in the performance of the SLMs and LLMs. In order to do so, we perform Latent Dirichlet Allocation (LDA) (Blei et al., 2003) to extract topics from content in the FPs and FNs for the SLMs. Due to the small number of comments under consideration, we only extract four topics, and follow it up with manual assessment of the comments.

For FPs, three interesting topics we found were (1) web-links in the comments, (2) short comments tagging the comment in the context to respond to a specific point, and (3) comments mentioning *delta* (e.g., *"sounds like you owe him (or me) a delta"*, "Δ"). Upon inspecting the rules, we found rules that might confuse the model in each of these three instances: (1) *"Doesn't Contribute Meaningfully"* and *"No Neutral/Transgender/Harm a specific person/**Promo**/Meta"*, (2) *"No Rude/Hostile Comment"*, and (3) *"No Delta Abuse/Misuse or Should Award Delta"*.

We find that 13.7% of the FP comments (10/73) contained links, and while all three SLMs got these incorrectly, at least one LLM got each of these 10 moderation outcomes right. Similarly for comments tagging and replying to the previous comment, we find that LLMs correctly identify 51% (36/73) comments as non-violating.

On the other hand, for comments mentioning *delta* we actually found that there were some comments in the training set nearly identical to the examples that the SLMs moderated (e.g., "Δ", "!delta rip", "Then delta this mannn") which were removed by human moderators. Further, the rule *"No Delta Abuse/Misuse or Should Award Delta"* is a clear dismissal of these kind of comments. Therefore, our judgment in this case was that the SLMs were correct at moderating them, and perhaps these comments were missed by human moderators.

For FNs, we found that there were many comments related to topics such as wars, political controversies, and gender fluidity which the LLMs always moderate correctly, whereas the SLMs do not. This might be an expected outcome, as all the LLMs in our study have undergone extensive RLHF (Ouyang et al., 2022) and are therefore more cautious when it comes to controversial and sensitive topics, whereas the SLMs are much smaller and do not have the same safety standards.

5 Cross-domain Moderation

5.1 Performance Comparison

In this section, we investigate the cross-domain performance of SLMs on moderation tasks. Figure 4 shows a visual depiction of the findings. We choose three communities examined in prior work (Kumar et al., 2024), namely *r/IAmA*, *r/askscience*, and *r/movies* as our three subreddits for the discussion

of cross-domain performance. Appendix D provides complete cross-domain results.

We observe that *r/askscience* is the subreddit with the lowest average cross-domain performance at 58% with the highest performance of 65% given by the Mistral fine-tuned for *r/science* and *r/history*. On the other hand, *r/movies* was the easiest of the three subreddits, with the highest average cross-domain performance at 68.6% and a highest performance of 80% given by the Mistral fine-tuned for *r/aww* and *r/nba*, and Gemma model fine-tuned for *r/aww*. Finally, for *r/IAmA*, the average accuracy was 63.7% and the highest performance came from the Mistral model fine-tuned for *r/politics* at 75%.

It is noteworthy that while the moderation accuracy of cross-domain models is not at par with in-domain SLM models, the best-performing cross-domain SLMs outperform best-performing LLMs for these subreddits, with relative advantages of 6.5% for *r/IAmA*, 11.6% for *r/askscience*, and 0.1% for *r/movies*, indicating the promise of cross-community based approach for content moderation using SLMs over using LLMs.

Furthermore, we see that there are some models that show a promising overall cross-domain performance. Specifically, models trained on *r/changemyview*, *r/nba*, and *r/movies* show an average cross-domain accuracy between 64.6 and 67.8%, and can therefore be used as *'meta-experts'* to moderate communities that would benefit more from a cross-domain moderation as compared to in-domain moderation. One example is *r/AskReddit*, which would benefit from the cross-domain models from 10 other communities compared to its in-domain model based on Llama-3.1-8b and Gemma-2-9b, and 7 others for Mistral-NeMo-Instruct.

5.2 Exploring factors that affect cross-domain performance of SLMs

Due to the promising performance of cross-domain models, we further investigate why some subreddits may benefit from cross-domain moderation, and whether there is a pattern to the cross-domain accuracy of an SLM on a target subreddit by testing the impact of subreddit size, description and rules. Our hypothesis is that relative sizes of subreddits can play a role in determining content moderation performance across communities as similarly-sized communities may have similar underlying norms. Accordingly, the similarity of topics as well as community rules between subreddits may have a direct impact on cross-community content moderation outcomes, due to the similarity in content or in the way in which the content is moderated.

For size, we construct a matrix of relative number of subscribers of the source subreddit w.r.t to the target subreddit, and for description and rules, we get the subreddit description and rules using PRAW and use Cohere's `embed-english-v3.0` (Cohere For AI, 2023) to generate embeddings and compute two cosine-similarity matrices between pairs of source and target subreddit descriptions and rules.

We then conduct a column-wise (in this case, subreddit-wise) t–test for non-correlation between the relative-size matrix, and the cosine similarity matrices for description and rules, with the cross-domain accuracy matrices for each model, using the Pearson correlation coefficient (Schober et al., 2018) as our test-statistic (r). The alternative hypothesis was that there exists a positive correlation.

We find that there was no statistically significant ($\alpha = 0.05$) positive correlation between relative size and the performance of cross-domain models on a subreddit. For subreddit description, we find two statistically significant positive correlations with Llama-3.1-8b on *r/AskHistorians* ($t(12) = 0.459, p = .049$) and Gemma-2-9b on *r/askscience* ($t(12) = 0.535, p = .024$). Finally, for correlation between subreddit rules and cross-domain performance we again find only two significant positive correlations with Llama-3.1-8b on *r/nba* ($t(12) = 0.462, p = .048$) and Gemma-2-9b on *r/anime* ($t(12) = 0.459, p = .049$). Full results of the t-test can be found in Appendix E.

This result signifies that while the topic of the subreddit and the community rules may play some role in determining the cross-domain performance, the overall notions of what determines the cross-community performance go beyond just subreddit sizes, description, and rules.

6 Discussion and Implications

In this section, we discuss the implications of our work for online content moderation.

6.1 Shift from generalist LLMs to specialist SLMs for content moderation

Our findings indicate a significant advantage in employing SLMs for content moderation tasks over LLMs, consistently showing superior performance in terms of both accuracy and recall. Our findings suggest that SLMs adopt a more aggressive

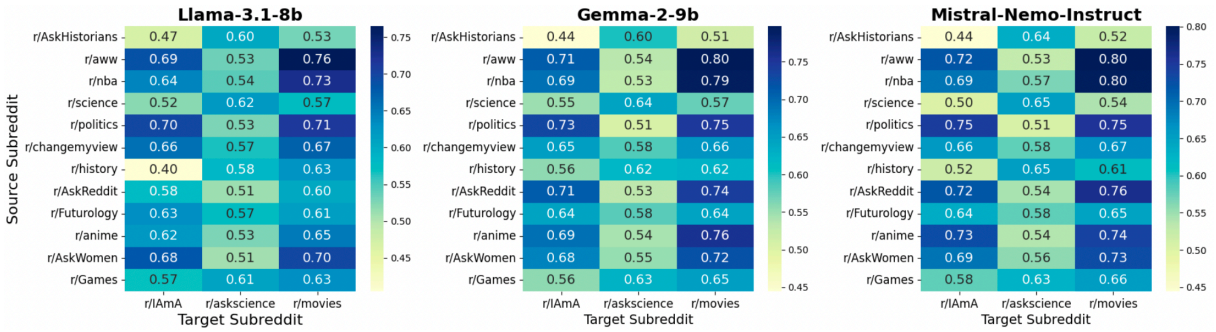

Figure 4: **Cross-domain Moderation Performance.** Comparison of performance of SLMs in terms of accuracy for cross-domain content moderation performance on three target subreddits: `r/IAmA`, `r/askscience`, `r/movies`. `mistral-nemo-instruct` gives the best cross-domain performance, with 75% accuracy for `r/IAmA` by the model fine-tuned for `r/politics`, 65% accuracy for `r/askscience` by the model fine-tuned for `r/science` and `r/history`, and 80% for `r/movies` by the model fine-tuned for `r/aww` and `r/nba`.

approach compared to LLMs, resulting in significantly higher recall for SLMs, though at the cost of slightly lower precision relative to LLMs. However, prior work has shown that moderators, who are overburdened already, may not be able to attend to all instances of norm-violating and undesirable behavior (Park et al., 2022; Chandrasekharan et al., 2019). Hence, the higher recall of SLMs compared to LLMs could actually provide a benefit for more reliable detection of harmful content that could otherwise be overlooked and stay on the platform for longer—potentially leading to further undesirable outcomes (Lambert et al., 2022).

Apart from performance, a key strength of SLMs lies in their ability to be fine-tuned for specific communities, allowing for moderation that aligns closely with the unique needs of individual subreddits. Additionally, specialized communities and those serving sensitive populations (Saha et al., 2020) may require extra safeguards and considerations when using automated content moderation tools, which can be more easily fine-tuned with SLMs. This approach improves moderation accuracy by better reflecting the decisions of community-specific moderators. In addition, community norms can evolve over time, and performing continual pretraining (Ke et al., 2023) on SLMs with incrementally collected data can help the models stay updated in accordance.

Finally, SLMs for content moderation are a cheaper and more scalable option for platforms like Reddit, which manage large volumes of user-generated content and provide them agency over their models and data without the reliance on third-party APIs. LLMs, on the other hand, are expensive to query, rate-limited, and mostly closed-source.

6.2 Automated tools for removals vs. triaging

The precision/recall trade-off between LLMs and SLMs discussed in the previous section also has implications for their potential usage from a moderation design perspective. Specifically, content moderation can either be done in an automated manner with no moderator involvement or in a moderator-in-the-loop manner where triaged and reported comments are manually moderated.

Since LLMs are better at accurately identifying violating comments while minimizing false positives, they are more suited to be used as automated moderation tools as they are less likely to wrongfully penalize community members by moderating potentially benign comments due to their cautious approach. SLMs, on the other hand, are more aggressive at flagging comments with higher recall, which makes them more suited to scenarios where the priority is to ensure that potentially harmful content is quickly triaged and sent for further review by human moderators. This will ensure the flagging of seemingly undesirable behavior in a time-sensitive manner and not leave such content visible on the platform for long periods, and upon manual inspection, if the comment is benign, it can be allowed to exist by the moderator.

6.3 Cross-domain approaches to moderation

We observe that along with great in-domain performance, SLMs also perform well in cross-domain settings which suggests that norm violation representations learned by fine-tuned SLMs can generalize effectively across different online communities, making them a viable option for moderating content in new or growing communities.

Moreover, we saw that certain *'meta-expert'*

models have high average cross-domain performance, often providing certain communities with more performance than their in-domain models. This indicates that SLMs have the capability to learn shared norm violation representations, and by leveraging cross-community similarities, cross-domain SLMs can be used for moderation without needing specific training data for every new community, thus enabling faster and more cost-efficient deployment of automated moderation tools across platforms. Specifically, a new community developing its rules can make use of cross-domain experts to identify which expert provides the community with the highest content moderation performance, and use that to inform their own community rules.

However, we show that determining which cross-domain expert would provide a community with the highest benefit is a nuanced task that goes beyond measures like similarity in subreddit sizes, descriptions, and rules. This provides an avenue for future research to explore the development of strategies to identify the right set of cross-community experts as well as meta-moderation models. Using a mixture of experts (MoE) framework, for example, could allow a community to draw from specialized models trained on different clusters of subreddits with shared themes, rules, or behaviors. This would enable automated moderation systems to dynamically route content to the most appropriate expert and ensure that moderation decisions are well-aligned with specific community norms.

6.4 Instability of Closed-source LLMs

Table 1: **Instability of GPT-4o.** Average performance metrics for two different releases of GPT-4o.

Metric	GPT-4o (2024-05-13)	GPT-4o (2024-08-06)
Precision	0.736	**0.763**
Recall	**0.557**	0.447
Accuracy	**0.677**	0.658

Since closed-source models undergo frequent updates, it is crucial to assess the stability of their performance in content moderation tasks. To analyze how different releases of the same model can affect performance, we compare the initial release of OpenAI's GPT-4o model (May 13, 2024) with the latest version (August 06, 2024).

From Table 1, we observe that the accuracy of GPT-4o declined from 67.7% to 65.8%, accompanied by a huge reduction in recall from 55.7% to 44.7%. Conversely, the average precision improved from 73.6% to 76.3%. Crucially, 10.7% or 2,166 comments across all subreddits that were correctly moderated by the May release of the model, were no longer moderated by the August version, highlighting a lack of trust that could be placed in these closed-source LLMs for content moderation.

Given the black-box nature of closed-source models and the lack of knowledge of upcoming updates, ensuring consistent moderation performance would pose a challenge when using closed-source models. Moderators may instead consider using open-source language models, which tend to be more stable since moderators can choose not to update the models and can be fine-tuned more easily to meet specific moderation requirements.

7 Conclusion

This paper examined the effectiveness of small language models (SLMs) and large language models (LLMs) in content moderation tasks. We found that SLMs with less than 15b parameters, such as Gemma-2-9b and Mistral-Nemo-Instruct, consistently outperformed significantly larger LLMs like GPT-4o at identifying undesirable content. SLMs showed a higher recall, indicating their superior ability to flag a wider range of undesirable content, which can be crucial for effective moderation on large-scale platforms. In addition to in-domain moderation tasks, we uncovered the potential of SLMs to be used in cross-community moderation tasks. Cross-community moderation can benefit smaller communities with fewer resources to train their own in-domain models or newer communities where community norms are still emerging. We provided qualitative insights into the tradeoffs arising from content length and topic that could impact SLMs and LLMs differently. Overall, SLMs offer an effective, scalable, and cost-effective solution for content moderation, achieving a strong balance between accuracy, recall, and stability. Future work can further improve cross-domain performance by training models that learn shared notions of norms and values, and explore hybrid models that leverage complementary strengths of SLMs and LLMs to enhance moderation capabilities across diverse online communities and platforms.

8 Limitations

8.1 Scale of Subreddit Selection

We chose 15 highly popular subreddits for our study with a good mix of science, AMA, enter-

tainment, history, sports, and political subreddits in order to capture a wide spectrum of underlying norms and subreddit characteristics. We therefore believe that our findings are representative and scalable to a wider range of communities. However, future studies could expand the scope to incorporate a larger number of subreddits for increased robustness of our findings and more open-sourced models for content moderation.

8.2 Text-Based Content Moderation

Due to availability of publicly available text-based datasets for Reddit comment removals, we resorted to purely text-based content moderation. However, undesirable behavior may occur in multiple other modes like images and future studies can expand on our insights to explore the performance of vision-language models in content moderation settings.

8.3 Continual Updating of Models

Community norms and notions of undesirable behavior may change over time and models fine-tuned on a specific date range of posting activity may not necessarily generalize to newer comment removals or an evolved representation of content that moderators would remove. Therefore, future studies can evaluate the adaptability of models on temporal distribution shifts in online content and determine the efficacy of techniques like domain-adaptive pretraining (DAPT) (Gururangan et al., 2020) and continual pretraining (Ke et al., 2023) in keeping SLMs update with the latest norms.

Ethical Considerations

Our work explores the use of language models for online content moderation and shows the promise of small-scale fine-tuned models to achieve superior performance. While small language models provide freedom to fine-tune the models as required, this can potentially have consequences if used by communities and moderators in an adversarial manner to moderate comments from particular users or factions of society therefore leading to unintended ethical consequences in terms of freedom of rightful expression on social media platforms. Finally, since we work with the OpenAI[3] and Cohere[4] APIs, we ensure to comply with their terms of use policies.

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

[4]https://cohere.com/terms-of-use

References

Pinkesh Badjatiya, Shashank Gupta, Manish Gupta, and Vasudeva Varma. 2017. Deep learning for hate speech detection in tweets. In *Proceedings of the 26th International Conference on World Wide Web Companion*, WWW '17 Companion, page 759–760, Republic and Canton of Geneva, CHE. International World Wide Web Conferences Steering Committee.

David M Blei, Andrew Y Ng, and Michael I Jordan. 2003. Latent dirichlet allocation. *Journal of machine Learning research*, 3(Jan):993–1022.

Tom B Brown. 2020. Language models are few-shot learners. *arXiv preprint arXiv:2005.14165*.

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, 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.

Cohere For AI. 2023. Introducing Embed v3 — cohere.com. https://cohere.com/blog/introducing-embed-v3. [Accessed 13-10-2024].

Cohere For AI. 2024. c4ai-command-r-plus-08-2024.

Thomas Davidson, Dana Warmsley, Michael Macy, and Ingmar Weber. 2017. Automated hate speech detection and the problem of offensive language. *Preprint*, arXiv:1703.04009.

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*.

Suchin Gururangan, Ana Marasović, Swabha Swayamdipta, Kyle Lo, Iz Beltagy, Doug Downey, and Noah A Smith. 2020. Don't stop pretraining: Adapt language models to domains and tasks. *arXiv preprint arXiv:2004.10964*.

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*.

Chandra Irugalbandara, Ashish Mahendra, Roland Daynauth, Tharuka Kasthuri Arachchige, Jayanaka Dantanarayana, Krisztian Flautner, Lingjia Tang, Yiping Kang, and Jason Mars. 2024. Scaling down to scale up: A cost-benefit analysis of replacing openai's llm

with open source slms in production. In *2024 IEEE International Symposium on Performance Analysis of Systems and Software (ISPASS)*, pages 280–291.

Prince Jha, Raghav Jain, Konika Mandal, Aman Chadha, Sriparna Saha, and Pushpak Bhattacharyya. 2024. Memeguard: An llm and vlm-based framework for advancing content moderation via meme intervention. *arXiv preprint arXiv:2406.05344*.

Shagun Jhaver, Pranil Vora, and Amy Bruckman. 2017. Designing for civil conversations: Lessons learned from changemyview. *Georgia Institute of Technology*.

David Jurgens, Eshwar Chandrasekharan, and Libby Hemphill. 2019. A Just and Comprehensive Strategy for Using NLP to Address Online Abuse. arXiv. ArXiv:1906.01738 [cs].

Zixuan Ke, Yijia Shao, Haowei Lin, Tatsuya Konishi, Gyuhak Kim, and Bing Liu. 2023. Continual pre-training of language models. *arXiv preprint arXiv:2302.03241*.

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.

Vinay Koshy, Tanvi Bajpai, Eshwar Chandrasekharan, Hari Sundaram, and Karrie Karahalios. 2023. Measuring user-moderator alignment on r/changemyview. *Proceedings of the ACM on Human-Computer Interaction*, 7(CSCW2):1–36.

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.

Vivian Lai, Samuel Carton, Rajat Bhatnagar, Q. Vera Liao, Yunfeng Zhang, and Chenhao Tan. 2022. Human-ai collaboration via conditional delegation: A case study of content moderation. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*, CHI '22, New York, NY, USA. Association for Computing Machinery.

Charlotte Lambert, Ananya Rajagopal, and Eshwar Chandrasekharan. 2022. Conversational resilience: Quantifying and predicting conversational outcomes following adverse events. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 16, pages 548–559.

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.

I Loshchilov. 2017. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*.

Krishanu Maity, Raghav Jain, Prince Jha, Sriparna Saha, and Pushpak Bhattacharyya. 2023. Genex: A commonsense-aware unified generative framework for explainable cyberbullying detection. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 16632–16645.

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

Chikashi Nobata, Joel R. Tetreault, Achint Oommen Thomas, Yashar Mehdad, and Yi Chang. 2016. Abusive language detection in online user content. *Proceedings of the 25th International Conference on World Wide Web*.

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

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. 2022. Training language models to follow instructions with human feedback. *Advances in neural information processing systems*, 35:27730–27744.

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.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. 2019. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9.

Koustuv Saha, Sindhu Kiranmai Ernala, Sarmistha Dutta, Eva Sharma, and Munmun De Choudhury. 2020. Understanding moderation in online mental health communities. In *HCII*. Springer.

Timo Schick and Hinrich Schütze. 2021. It's not just size that matters: Small language models are also few-shot learners. *Preprint*, arXiv:2009.07118.

Patrick Schober, Christa Boer, and Lothar A Schwarte. 2018. Correlation coefficients: appropriate use and interpretation. *Anesthesia & analgesia*, 126(5):1763–1768.

Kumar Bhargav Srinivasan, Cristian Danescu-Niculescu-Mizil, Lillian Lee, and Chenhao Tan. 2019. Content removal as a moderation strategy: Compliance and other outcomes in the changemyview community. *Proceedings of the ACM on Human-Computer Interaction*, 3(CSCW):1–21.

Gemma Team, Morgane Riviere, Shreya Pathak, Pier Giuseppe Sessa, Cassidy Hardin, Surya Bhupatiraju, Léonard Hussenot, Thomas Mesnard, Bobak Shahriari, Alexandre Ramé, et al. 2024. Gemma 2: Improving open language models at a practical size. *arXiv preprint arXiv:2408.00118*.

Fabio Del Vigna, Andrea Cimino, Felice Dell'Orletta, Marinella Petrocchi, and Maurizio Tesconi. 2017. Hate me, hate me not: Hate speech detection on facebook. In *Italian Conference on Cybersecurity*.

William Warner and Julia Hirschberg. 2012. Detecting hate speech on the world wide web. In *Proceedings of the Second Workshop on Language in Social Media*, pages 19–26, Montréal, Canada. Association for Computational Linguistics.

Ziqi Zhang and Lei Luo. 2018. Hate speech detection: A solved problem? the challenging case of long tail on twitter. *Preprint*, arXiv:1803.03662.

A Subreddits Statistics

In Table 2 we list the number of subscribers in each of the subreddits we stufy in our work along with thei rofficial public description from Reddit.

B LoRA Hyperparameters

In order to fine-tune our community-specific LLMs, we perform Low-Rank Adaptation (LoRA) (Hu et al., 2021) for 1 epoch on 8000 balanced samples from publicly available content moderation datasets. We use rank $r = 16$ LoRA with an $\alpha = 32$ and no dropout. We use 5 warmup steps, a linear learning rate schedule with $lr = 2e-4$, AdamW (Loshchilov, 2017) optimizer with a weight decay of 0.01.

C Compute Resources

All experiments on open-source models were run on a GPU server equipped with 1xNVIDIA A100. The experiments with the OpenAI models cost about 250 USD and experiments with Cohere Command R+ and Cohere Embedv3 English cost about 30 USD.

D Detailed Cross-Domain Results

In this section we provide the model performances on cross-domain content moderation tasks across different subreddits. Figure 5, Figure 6, and Figure 7 represent cross-domain accuracy of all subreddits for Llama-3.1-8b, Gemma-2-9b, and Mistral-NeMo-Instruct.

E Detailed Cross-domain Correlation Test Results

In this section, we provide a detailed set of results from the Pearson Correlation Coefficient t-test for non-correlation that we conducted to identify any patterns in cross-domain model performance. Table 6 shows test results for relative subreddit size, Table 7 shows the test results for embeddings of subreddit descriptions, and Table 8 shows test results for embeddings of subreddit rules.

Table 2: Subreddit subscriber sizes and descriptions

Subreddit	Size	Subreddit Description
r/askscience	26M	A subreddit for people to ask a science question, and get a science answer
r/IAmA	23M	A Q&A subreddit featuring interactive interviews with individuals of various backgrounds
r/movies	34M	A space for inclusive discussions on films, including reviews and news about major releases
r/anime	11M	A subreddit for Reddit's premier anime community
r/AskHistorians	2.1M	A subreddit for well-researched, expert-level answers on historical questions
r/AskReddit	49M	A subreddit to ask and answer thought-provoking questions
r/AskWomen	5.5M	A subreddit for women to share their perspectives and experiences on various topics
r/aww	37M	A subreddit for cute and cuddly pictures
r/changemyview	3.7M	A subreddit for users to present opinions they are open to having challenged through reasoned debate
r/Futurology	21M	A subreddit devoted to the field of Future(s) Studies and evidence-based speculation about the development of humanity, technology, and civilization
r/Games	3.3M	A subreddit to provide a place for informative and interesting gaming content and discussions.
r/history	18M	A subreddit for discussions about history
r/nba	13M	A subreddit for NBA discussion
r/politics	8.7M	A subreddit for news and discussion about U.S. politics.
r/science	33M	A subreddit to share and discuss new scientific research.

Table 3: Model Accuracy by Subreddit

Subreddit	llama-3.1-8b	gemma-2-9b	mistral-nemo-instruct-12b	Command-R-Plus	gpt-4o-mini	gpt-4o	gpt-4o-2024-05-13	gpt3.5-turbo
r/askscience	0.716	0.712	**0.745**	0.520	0.533	0.541	0.550	0.541
r/IAmA	0.762	**0.796**	0.794	0.689	0.724	0.725	0.731	0.604
r/movies	0.807	**0.850**	0.832	0.748	0.799	0.798	0.809	0.752
r/anime	0.623	**0.745**	0.743	0.504	0.576	0.529	0.574	0.576
r/AskHistorians	0.678	0.737	**0.769**	0.511	0.553	0.592	0.593	0.563
r/AskReddit	0.545	0.618	**0.649**	0.556	0.593	0.602	0.598	0.577
r/AskWomen	0.762	**0.785**	0.781	0.567	0.646	0.642	0.669	0.620
r/aww	0.776	0.804	**0.810**	0.645	0.716	0.697	0.753	0.691
r/changemyview	0.902	0.911	**0.916**	0.676	0.794	0.766	0.768	0.623
r/Futurology	0.771	0.793	**0.805**	0.613	0.635	0.623	0.646	0.649
r/Games	0.701	**0.748**	0.747	0.554	0.625	0.621	0.675	0.610
r/history	0.671	0.735	**0.744**	0.537	0.604	0.601	0.630	0.597
r/nba	0.788	0.832	**0.844**	0.696	0.746	0.748	0.778	0.699
r/politics	0.742	**0.803**	0.793	0.652	0.740	0.742	0.736	0.692
r/science	0.636	0.707	**0.712**	0.598	0.592	0.638	0.641	0.561
Avg.	0.725	0.772	**0.779**	0.605	0.658	0.658	0.677	0.624

Table 4: Model Precision by Subreddit

Subreddit	llama-3.1-8b	gemma-2-9b	mistral-nemo-instruct-12b	Command-R-Plus	gpt-4o-mini	gpt-4o	gpt-4o-2024-05-13	gpt3.5-turbo
r/askscience	0.709	0.687	0.738	**0.833**	0.762	0.730	0.771	0.617
r/IAmA	0.725	0.767	0.754	**0.903**	0.777	0.838	0.736	0.508
r/movies	0.830	0.872	0.869	**0.943**	0.884	0.910	0.855	0.759
r/anime	0.618	0.719	**0.720**	0.643	0.667	0.571	0.640	0.579
r/AskHistorians	0.682	0.735	0.769	0.717	0.754	**0.785**	0.765	0.584
r/AskReddit	0.543	0.642	0.671	**0.834**	0.636	0.659	0.599	0.583
r/AskWomen	0.756	0.761	0.766	**0.816**	0.771	0.739	0.708	0.633
r/aww	0.790	0.812	0.818	**0.905**	0.832	0.828	0.815	0.716
r/changemyview	0.890	0.890	0.901	**0.949**	0.793	0.805	0.721	0.616
r/Futurology	0.770	0.784	0.794	**0.867**	0.759	0.792	0.734	0.662
r/Games	0.702	0.737	0.754	**0.936**	0.827	0.838	0.821	0.754
r/history	0.665	0.710	0.715	**0.870**	0.750	0.718	0.718	0.604
r/nba	0.799	0.838	0.856	**0.939**	0.833	0.837	0.829	0.690
r/politics	0.775	0.849	0.828	**0.890**	0.784	0.782	0.721	0.685
r/science	0.640	0.690	0.690	**0.783**	0.714	0.618	0.609	0.600
Avg.	0.726	0.766	0.776	**0.855**	0.770	0.763	0.736	0.639

Table 5: Model Recall by Subreddit

Subreddit	llama-3.1-8b	gemma-2-9b	mistral-nemo-instruct-12b	Command-R-Plus	gpt-4o-mini	gpt-4o	gpt-4o-2024-05-13	gpt3.5-turbo
r/askscience	0.733	**0.780**	0.760	0.050	0.096	0.130	0.140	0.225
r/IAmA	0.663	0.712	**0.729**	0.259	0.446	0.396	0.529	0.723
r/movies	0.773	**0.820**	0.781	0.528	0.689	0.661	0.748	0.744
r/anime	0.645	**0.805**	0.795	0.018	0.304	0.232	0.334	0.564
r/AskHistorians	0.667	0.739	**0.769**	0.038	0.159	0.252	0.274	0.448
r/AskReddit	0.564	0.531	**0.585**	0.141	0.436	0.423	0.570	0.521
r/AskWomen	0.774	**0.832**	0.808	0.173	0.417	0.438	0.571	0.571
r/aww	0.754	0.791	**0.796**	0.325	0.540	0.496	0.654	0.632
r/changemyview	0.917	**0.938**	0.935	0.373	0.795	0.701	0.895	0.661
r/Futurology	0.772	0.808	**0.823**	0.267	0.396	0.332	0.497	0.612
r/Games	0.697	**0.769**	0.731	0.117	0.316	0.300	0.450	0.330
r/history	0.692	0.794	**0.811**	0.087	0.312	0.334	0.439	0.551
r/nba	0.771	0.822	**0.827**	0.419	0.614	0.616	0.702	0.717
r/politics	0.681	0.737	**0.740**	0.348	0.662	0.671	0.767	0.710
r/science	0.621	0.751	**0.767**	0.270	0.307	0.723	0.789	0.367
Avg.	0.715	0.775	**0.777**	0.228	0.433	0.447	0.557	0.558

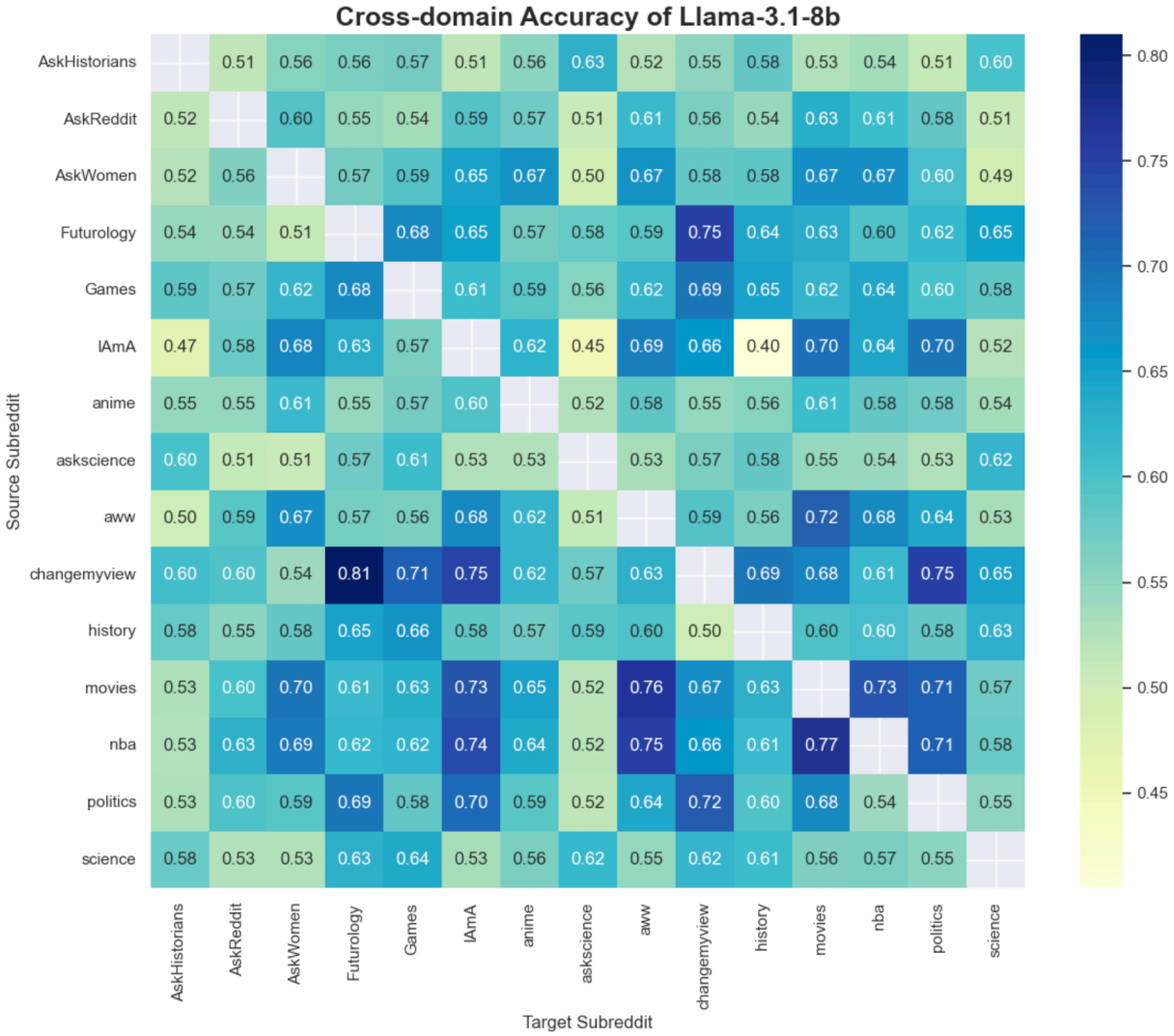

Figure 5: **Cross-domain Moderation Performance for Llama-3.1-8b**

Table 6: **Relative-Size Results of t-test for cross-domain setting.** Testing the null hypothesis of non-correlation between the relative sizes of the source and target subreddits with the cross-domain accuracy of SLMs. Values in the table represent the Pearson's Correlation Coefficient r and statistically significant values with p-value < 0.05 are marked with $(*)$. We see no statistically significant positive correlation.

	llama-3.1-8b	gemma-2-9b	mistral-nemo-instruct
AskHistorians	-0.068	-0.331	-0.180
AskReddit	0.010	-0.033	0.099
AskWomen	0.256	0.153	0.275
Futurology	-0.340	-0.421	-0.388
Games	-0.218	-0.280	-0.225
IAmA	0.033	-0.106	-0.036
anime	0.102	-0.138	0.045
askscience	-0.026	-0.231	-0.190
aww	0.073	0.063	0.033
changemyview	-0.151	-0.108	-0.314
history	-0.319	-0.309	-0.143
movies	-0.029	-0.003	-0.080
nba	0.296	0.259	0.295
politics	0.062	-0.033	-0.101
science	-0.309	-0.287	-0.291

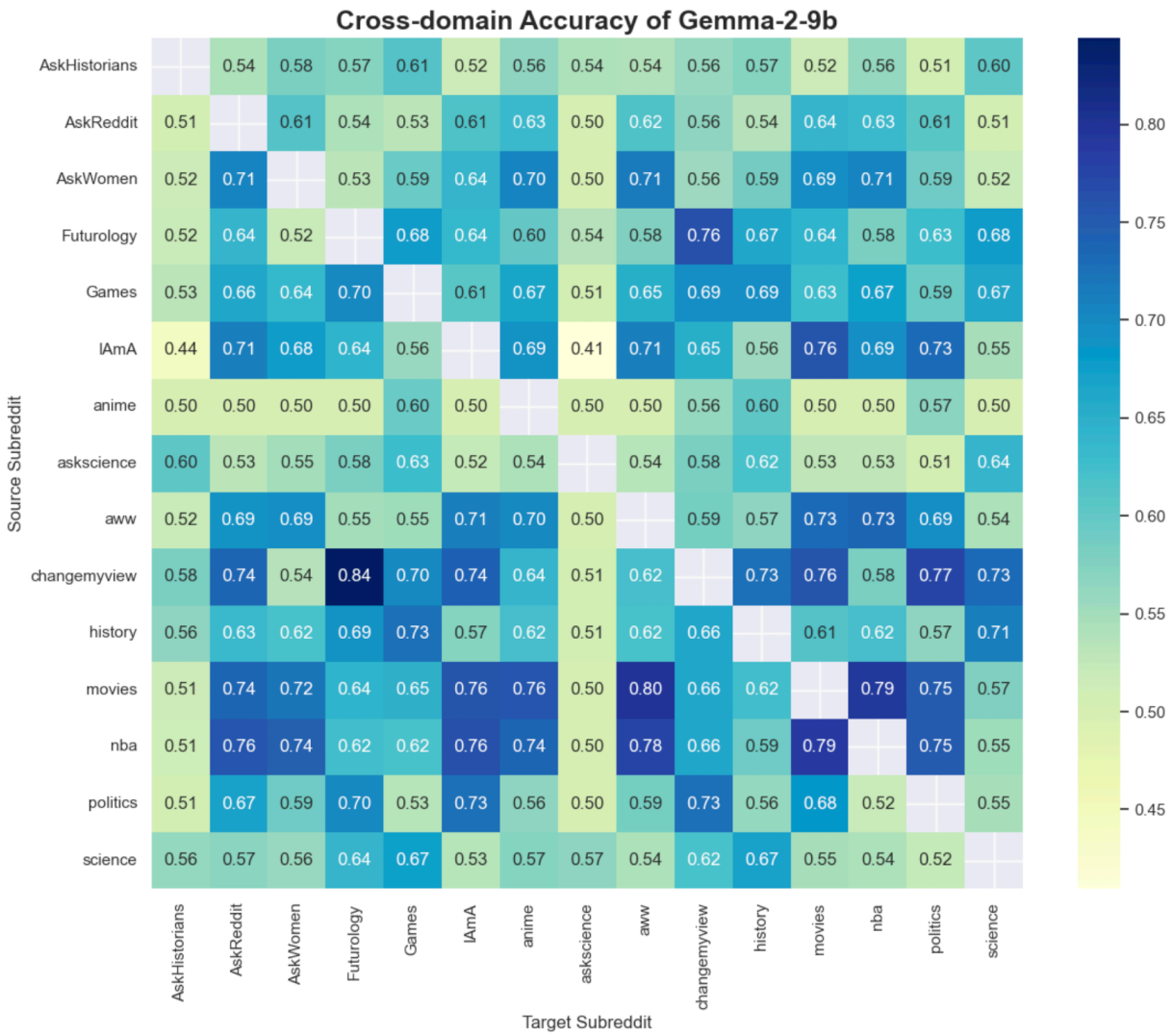

Figure 6: **Cross-domain Moderation Performance for Gemma-2-9b**

Table 7: **Subreddit Description Embeddings Results of t-test for cross-domain setting.** Testing the null hypothesis of non-correlation between the between the cosine similarity matrix of pairwise subreddit rule embeddings between the source and target subreddits with the cross-domain accuracy of SLMs. Values in the table represent the Pearson's Correlation Coefficient r and statistically significant values with p-value < 0.05 are marked with $(*)$. We see only two instances of statistically significant positive correlation for Llama-3.1-8b on *r/AskHistorians* and Gemma-2-9b on *r/askscience*.

	llama-3.1-8b	**gemma-2-9b**	**mistral-nemo-instruct**
AskHistorians	0.459 $(*)$	0.370	0.264
AskReddit	0.009	-0.063	-0.037
AskWomen	-0.074	0.030	-0.078
Futurology	0.099	0.152	0.116
Games	0.196	0.224	0.195
IAmA	-0.004	0.045	-0.020
anime	0.160	0.172	0.135
askscience	0.314	0.536 $(*)$	0.329
aww	-0.070	0.138	0.042
changemyview	0.166	0.114	0.131
history	0.284	0.018	-0.046
movies	-0.168	-0.086	-0.139
nba	-0.073	-0.264	-0.179
politics	-0.155	-0.202	-0.371
science	0.377	0.273	0.367

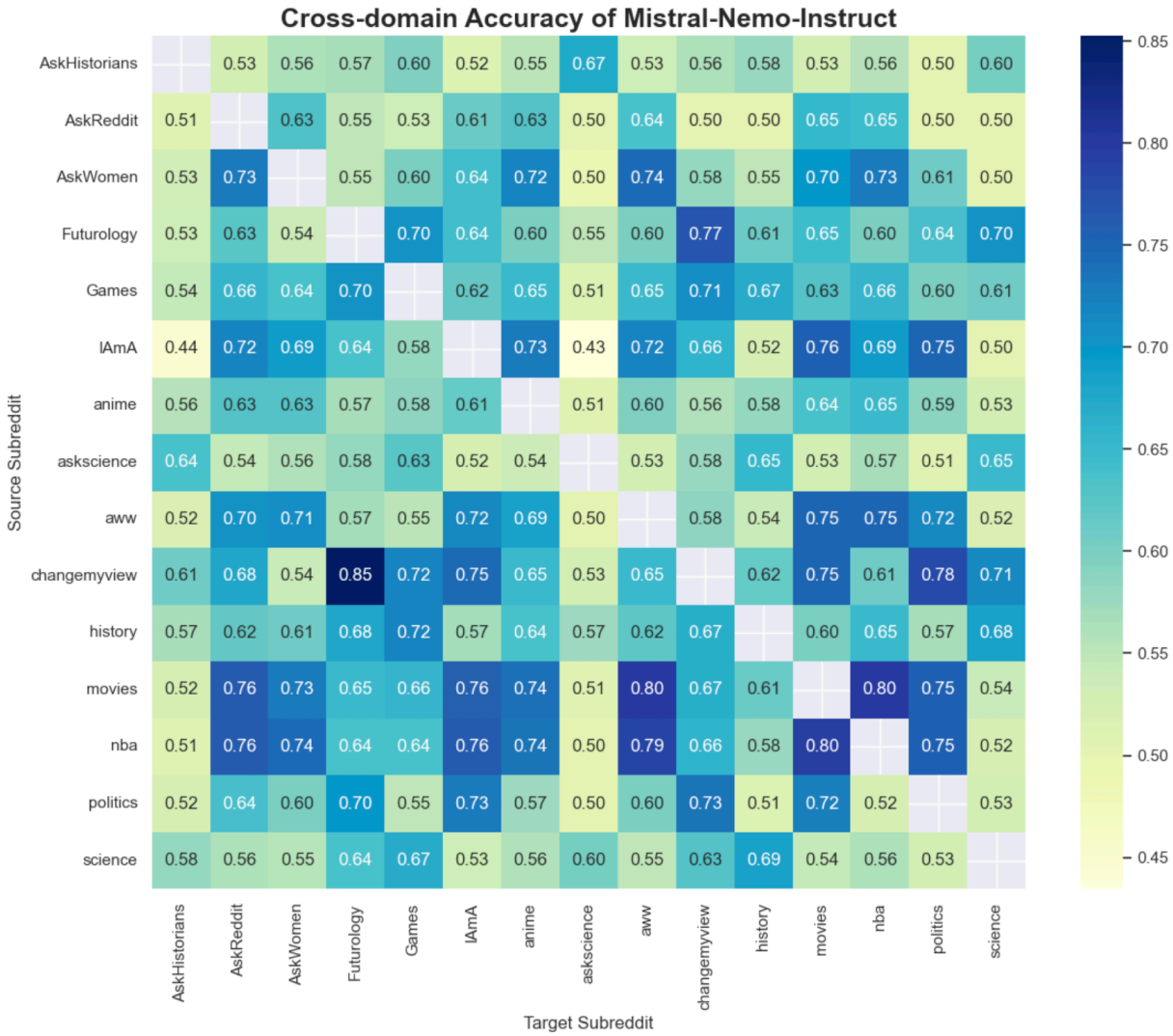

Figure 7: **Cross-domain Moderation Performance for Mistral-NeMo-Instruct**

Table 8: **Subreddit Rule Embeddings Results of t-test for cross-domain setting.** Testing the null hypothesis of non-correlation between the cosine similarity matrix of pairwise subreddit rule embeddings of the source and target subreddits with the cross-domain accuracy of SLMs. Values in the table represent the Pearson's Correlation Coefficient r and statistically significant values with p-value < 0.05 are marked with $(*)$. We see only two instances of statistically significant positive correlation for Llama-3.1-8b on *r/nba* and Gemma-2-9b on *r/anime*.

	llama-3.1-8b	**gemma-2-9b**	**mistral-nemo-instruct**
AskHistorians	0.009	0.008	-0.103
AskReddit	0.173	-0.007	0.308
AskWomen	0.307	0.244	0.298
Futurology	-0.116	-0.058	-0.057
Games	0.004	-0.084	-0.102
IAmA	0.025	-0.021	0.065
anime	0.266	0.459 $(*)$	0.357
askscience	-0.223	0.020	-0.361
aww	0.257	0.201	0.229
changemyview	0.134	-0.029	-0.018
history	0.077	-0.327	-0.337
movies	0.344	0.262	0.322
nba	0.462 $(*)$	0.439	0.388
politics	0.416	0.454	0.330
science	-0.054	-0.007	-0.010