How University Students Perceive Hate-Speech Moderation by Large Language Models

Samantha Flores, *University of Chicago,* Joshua Enebo, *University of Chicago*
Stella Lee, *University of Chicago,* Zayna Cheema, *University of Chicago*

Abstract

Large Language Models (LLMs) have been used to automate and assist human raters with identifying harmful content, such as hate speech. While these tools have been deemed largely successful, it is unclear whether user trust and perceptions of these tools align or diverge from the performance metrics that these models bring to content moderation. In this study, we sampled 1,000 tweets from the Davidson et al. (2017) hate speech corpus and used GPT 3.5-turbo to classify them into "hate-speech", "offensive language," or "neutral", and selected 10 low-confidence examples (confidence < 0.80) to use for a counterbalanced online survey with 21 University of Chicago students, collecting general opinions about LLM use to moderate content, agreement judgements on the LLM's labels, and independent classification of tweets. We find that University students are generally skeptical about LLM content moderation. Despite this distrust, however, not only were there no significant differences in demographics in the LLM classification task, but across all demographics, they also agreed with the LLM's classifications when presented with its labeling.

1. Introduction

Large language models have been increasingly incorporated into a variety of tasks, one such task that they have gained prevalence in is social media content moderation [1][2][3]. The incorporation of these models offers a promising alternative to sole manual moderation through its reduction in labor demands. While speed and accuracy demonstrate the merits of utilizing such models for content moderation, LLMs are not immune to bias themselves. There is increasing evidence that LLMs exhibit bias, and this is a consequence of bias present in their training data [8]. Instead of maintaining a position of neutrality, these systems reflect or amplify societal biases, raising pertinent questions regarding fairness, accountability, and trust in their application.

1.2 Motivation

Although LLMs have demonstrated high performance in hate speech detection within controlled settings [4] [5], their real-world effectiveness should not be dependent on the general accuracy of these models themselves, but rather, platforms should also consider how users perceive the fairness, transparency, and bias of these systems. A model may accurately classify content based on a labeled dataset, but if users disagree with its judgment or believe that moderation decisions are arbitrary or biased, regardless of the model's underlying performance, the platform may lose credibility and fail to maintain legitimacy and trust.

This issue is especially true when moderating a sensitive topic such as hate speech. Social media platforms have varying policies for content moderation. For instance, Twitter has adopted a policy of "freedom of speech, not freedom of reach," which reduces the visibility of tweets that violate platform guidelines without outright removing them. Their approach is an attempt to strike a balance between aggressive censorship and the promotion/ spread of harmful content. This method, while reducing the risk of over-censorship (false positives), however, increases the risk of harmful content persisting and reaching other users (false negatives). Other platforms, like Facebook or Instagram, instead favor content removal, even for borderline content. Here, they prioritize safety over inclusivity. In this context, allowing even a small number of hateful messages to pass through can cause significant harm. The differing strategies that different platforms use demonstrate that content moderation cannot be treated as something purely technical, especially for LLM-based classifiers; rather, the design of thresholds and interpretability mechanisms must also be guided by the sociotechnical goals of the platform itself. Understanding how users perceive such systems, especially amongst those individuals who are highly active on social media, is imperative to developing responsible deployment.

1.3 Research Objective

This study investigates how university students perceive the use of LLMs to moderate hate speech on social media (specifically X in our case). Specifically, we aim to answer three primary questions:

1. Do university students perceive LLM-based content moderation as fair and unbiased?

2. Is there any correlation between the perception of LLM content moderation of hate speech and the participant demographics

3. How do university students believe hate speech should be moderated when presented with concrete examples?

By exploring these three questions, we aim to inform the design and deployment of LLM moderation tools that are not only sound but also socially legitimate.

2. Related Work

2.1 LLM Content Classification

Prior work in LLM content classification has largely focused on the performance capabilities of these models, either on their own or alongside human moderators [1] [2] [3]. For instance, Kolla et al. (2024) [1] explored the feasibility of using LLMs to identify rule violations on Reddit and found that while LLM moderators generally had a high true-negative rate of 92.3%, their true positive performance was much lower at 43.1%. Their work highlights some of the limitations that LLMs have when it comes to moderating posts with higher complexities. Another similar approach from Thomas et al. (2024) [2] explored how feasible it was to utilize these models to automate and assist human raters in detecting harmful content, and, similarly, they found that while LLMs had a 90% overall accuracy rate, mild noise (i.e., incorrect labeling) generally reduced the models overall accuracy. Their results were similar to Kolla et al. (2024) in that the model underperformed when given longer text input.

Seering et al. (2024) [3] contrast the two prior approaches taken regarding LLM content moderation. Their Chillbot system introduced a backchannel moderation tool for moderators on Discord to send anonymous nudges to users who could potentially be exhibiting rule-breaking behavior. This specific study tackled the edge cases found in both [1] and [2] where the models had low-confidence scores. Their work shows that moderation accuracy is only part of the picture for a successful implementation.

2.2. Hate Speech Detection

In addition to general content classification, several studies have recently begun investigating how well LLMs perform in moderating hate speech content [4] [5] [6] [9] [10]. Both Chiu et al. (2021) [4] and Davidson et al. (2017) [5] demonstrate the limitations that come with LLM classifiers for hate speech and how they must be paired with sociolinguistic sensitivity. Their studies demonstrate that LLMs perform best when they are given curated examples and explicit instructions, but they also tend to fail in cases involving ambiguous language, slang, and subtle bigotry. Huang (2024) [10] builds on these limitations by arguing that accuracy alone is not a sufficient evaluation metric and that it is misleading in terms of its failure to distinguish between easy cases and hard cases. They propose a new legitimacy framework that differentiates between these easy and hard cases, where this framework emphasizes justification, user participation, and contextual awareness. They argue that LLMs are better suited to supporting reviewers and enhancing transparency rather than acting as independent moderators.

To better contextualize the current state of this category of LLM content moderation, Fortuna and Nunes (2018) [6] show a comprehensive literature review of automated hate speech detection techniques and data sources. Their work highlights the difficulties that come with not only defining hate speech but also the inconsistencies that come with annotations and contextual misunderstandings within these datasets. More recently, Huang et al. (2023) [9] evaluated ChatGPT's ability to detect and explain implicit hate speech to users. They showed that despite having 80% correct classification rates, these explanations swayed user judgment even when the explanation was incorrect. Their work suggests that LLMs inherit challenges from ambiguous hateful speech, and that it introduces new risk when natural language explanations are perceived as authoritative.

2.3 Bias

Despite these promising capabilities for content classification, and more specifically, hate speech classification, literature has also exposed potential systemic biases that come with using LLMs for content moderation. For instance, Davidson et al. (2019) [8] explore how five widely used hate speech and abuse datasets contain racial bias. Their work showed that across all the datasets, Black-aligned tweets were 2.7 times more likely to be classified as hate speech, harassment, or abuse. This demonstrates an underlying issue regarding how data is sampled and labeled, and these concerns transfer over directly to LLMs, regardless of the high accuracy that they exhibit when moderating content.

These studies collectively provide insight into the technical and ethical boundaries of using LLMs for content moderation, but they give very little attention to how end users interpret, experience, and evaluate these systems. While recall and accuracy are important, they do not capture the effects that this form of moderation has in cases where users perceive

decisions as inconsistent or biased. Seering et al. (2024) [3] began addressing this gap, yet their focus was still mostly on moderator behavior rather than on user trust. As more and more platforms begin to incorporate LLMs for content moderation, user perceptions must be understood, regardless of how well these LLMs are performing.

2.4 User Perceptions and Personal Moderation Tools

While prior work has largely focused on performance and fairness, less focus has been given to how users interact with automated moderation systems. Jhaver et al. (2023) [7] aim to address this methodological gap through a qualitative study that explored end-user perceptions of personal content moderation tools. Through a simulated social media feed with interactive controls, they found that users showed varying moderation preferences that were shaped by both cultural context and concerns about misclassified content. Not only does their work show that users desire transparency, but it also shows their need for control in terms of moderating hate speech and offensive content. This work differs from prior research in that it uncovers underlying tensions that exist between usability and explainability. These user insights are important to evaluating LLM moderation systems, not just because they may influence overall model accuracy, but because they influence user decisions. Through their framing of moderation as not only a technical challenge but also a human-factors challenge, the authors push for user-centered evaluation of content moderation tools, especially when considering how LLMs are increasingly being used for this space.

Our work builds directly on these insights by continuing to focus solely on the user perspective instead of focusing on the performance of LLMs or the nuances that underlie utilizing LLMs to moderate hateful content. Unlike Jhaver et al., however, we seek to gain insight into current user consensus regarding LLM content moderation, and if users perceive LLM moderation decisions as biased, either politically slanted or culturally insensitive.

3. Methods

3.1. Dataset

To gauge differences between LLM raters and university students' ratings of hate speech content, we utilized a real-world dataset of hate-speech content compiled by Davidson et al. (2017) under the MIT license [5]. This dataset is available in both CSV and Python pickle format and contains 25,000 posts drawn from public Twitter streams, where CrowdFlower (CF) workers manually coded each Tweet as "Hate speech", "Offensive", or "Neither." For this study's purposes, we treated this dataset's labels as *ground truth*.

3.2 Classifying Tweets

We randomly sampled 1,000 tweets from the full dataset to limit API costs. All subsequent analyses (CSV outputs, label comparisons, and survey stimuli) derive from this subset.

To simulate automated content moderation, we implemented a Python Script (Appendix A) that employed the GPT-3.5-turbo model via the OpenAI Python SDK to classify this subset of tweets. Each Tweet was then classified one at a time via the **Classify_Tweet ()** function. This function performed the following: prompt construction, API invocation, retry logic, and JSON parsing.

3.3 Prompt Design:

```
You are a content moderation assistant.
Classify the following tweet into one of
three categories: "hate_speech", "offen-
sive_language", or "neutral". Then, esti-
mate your confidence in your classifica-
tion as a float between 0.0 and 1.0.

Tweet: "<TWEET_TEXT>"

Respond only in JSON in this format:
{
  "label": "...",
  "confidence": ...
}
```

After classification, we collected a DataFrame with the following columns: tweet, gpt_label, and confidence, and saved it under **classified_tweets_gpt35_confidence.csv.** We then filtered out any tweets classified as "offensive speech", leaving 300 tweets classified as either "hate speech" or "neither". Following that, we compared the GPT-assigned labels with the original labels in Davidson et al. annotations.

From those 300 classified tweets, the research team collectively decided on 10 tweets to be used in the perception study:

- Six tweets were classified as hate speech, of which all tweets had a confidence interval < 0.80
- Four tweets were classified as neutral, of which all of the tweets had a confidence interval < 0.80

We decided to select tweets with confidence scores below 0.80 to emphasize examples where the model exhibited moderate uncertainty. This decision was rooted in the idea of increasing the likelihood of divergent human judgments.

The full dataset, and subsequent subsets of it, along with the full script used to simulate automated content moderation, can be found in the following GitHub repository: https://github.com/slflores0911/LLMHateSpeechContent

3.4 Study Design The online survey was divided into five sections to assess university student perceptions of LLM moderation decisions on hate speech content.

Section 1: Informed Consent We first presented participants with an informed consent form describing the purpose of the study, risks of exposure to offensive content, data usage, and mental health resources. Only individuals who explicitly agreed to the consent form were allowed to continue.

Section 2: Demographic Questionnaire Participants then answered a brief set of demographic questions, including political ideology, voting history, religious identity, ethnicity, gender identity, income, and age.

Section 3: General Attitudes Toward LLM Moderation We then asked participants to indicate their beliefs about the effectiveness and perceived bias of LLMs in moderating hate speech. This section included "yes-no" questions (e.g., "LLMs are effective at moderating hate speech") and Likert-scale statements (e.g., "I believe LLMs are successful at accurately moderating hate speech on the internet").

Section 4: Participant Evaluation of LLM Classifications We then showed participants five tweets, each with the LLM's classification, and they were then asked whether they agreed or disagreed with the model's label. At each tweet, participants had the option to provide any additional comments/justification for their responses.

Section 5: Participants' Classification of Tweets In this section, we showed participants five tweens selected from the dataset that had been classified by the LLM but were presented without labels. Participants were then asked to classify each tweet themselves as either:

- Hate Speech
- Not Hate Speech
- Other (optional write-in)

Once more, participants had the option to provide any additional comments/justification for their responses.

To account for any potential priming and ordering effects, we implemented the following design: For sections 4 and 5, we alternated the order in which they were presented for about half of the participants. That is, some participants completed the classification task themselves before they evaluated the LLM-labeled tweets. This manipulation allows us to assess whether prior exposure to model classifications had any influence on participants' judgments in the subsequent task. This counterbalancing was done during survey distribution using separate survey links.

4. Results

4.1 Participant Demographics and LLM Evaluations

We gathered results from 21 University of Chicago graduate and undergraduate students. We collected demographic data, which included the following: participants' gender, political affiliation, political ideology, religion, ethnicity, annual household income, and age. All response data and analysis can be found under the analysis folder of this GitHub repository: https://github.com/slflores0911/LLMHateSpeechContent
The results show demographic disparities in some of these demographic categories. As shown in **Figure 1a**. 95.5% of our participants identified as left-leaning, with the remaining 4.5% identifying as independent.

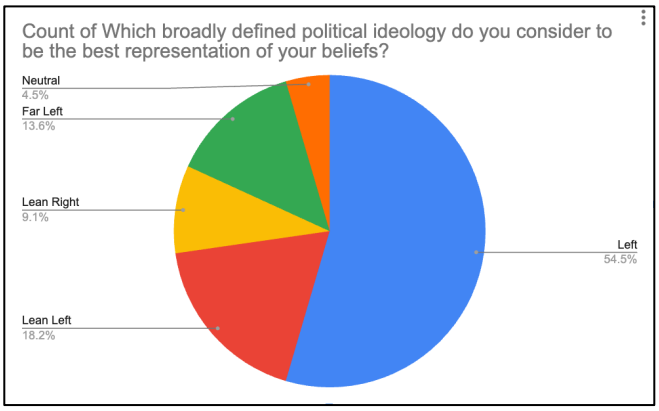

Figure 1a

With regards to party affiliation, we observe similar patterns within our participants, where 77.3 percent of our subject pool identified as Democrats (**Figure 1b)**.

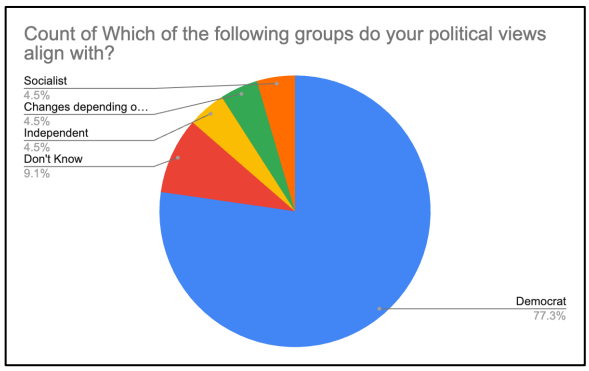

Figure 1b

Despite the data being skewed in favor of left-leaning individuals, we do observe user discrepancies with general agreement with LLM content classification. For instance, when participants were asked to independently classify tweets, we observed a near-even split between agreement and disagreement with how they classified one of the Tweets, where 54.5% of users agreed with the LLM's classification (**Figure 1c).** User justifications indicated uncertainty about the context of the given tweet, i.e, "not sure what a sp*ear chucker is".

Despite participants' divergent opinions, we did not observe any statistically significant correlations between political affiliation or political ideology and how users rated their classifications, and their perceptions of LLM content moderation of hate speech.

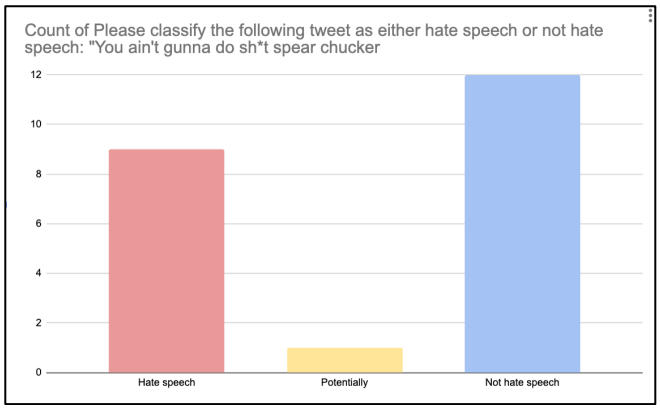

Figure 1c

4.2 General Attitudes Towards LLM Moderation

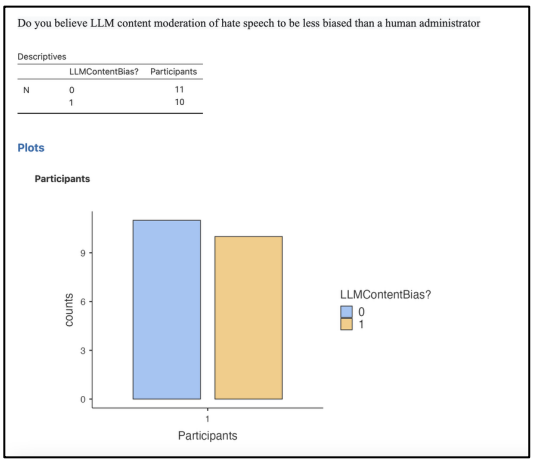

Figure 2a (0 indicates No, 1 indicates Yes)

We observe that participants generally expressed skeptical attitudes towards the use of LLMs for moderating hate speech, the results in Figure 2a. *Do you believe LLM content moderation of hate speech to be less biased than a human administrator?* Showed a near-even split between beliefs of LLMs exhibiting bias.

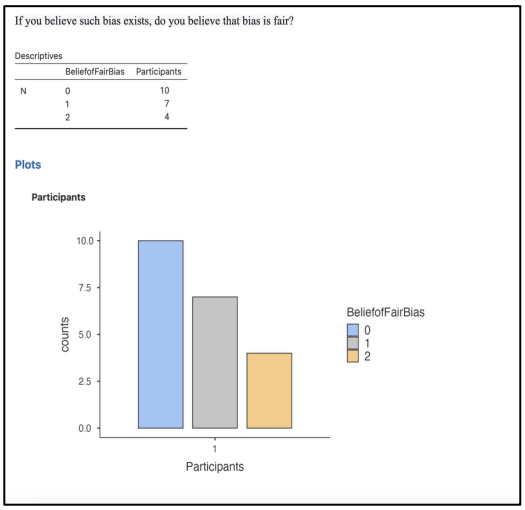

Figure 2b (0=bias is unfair, 1=bias is fair, 2=no bias existed)

Further, while some participants acknowledged that bias may exist in LLM outputs, the majority of our participants believed that LLMs exhibited bias and that the bias they exhibited was unfair. (see Figure 2b), as "0" indicates belief of unfair bias, "1" indicates they believe the bias is fair, and "2" indicates there is no bias.

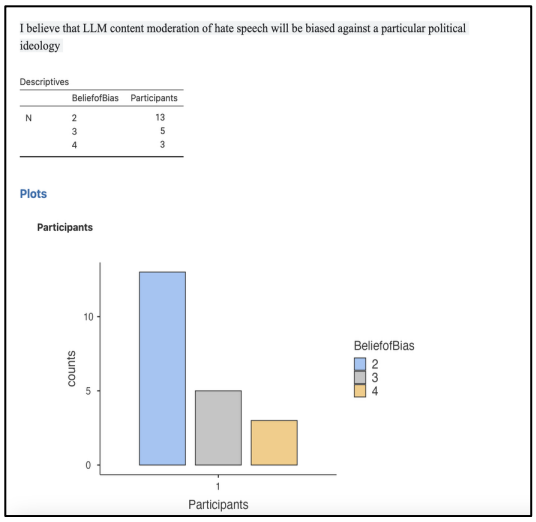

Figure 2c (2=Agree, 3=Neutral, 4=Disagree)

We also observed that many of our participants (N = 11) expressed concern that LLMS would be biased against particular political ideologies.

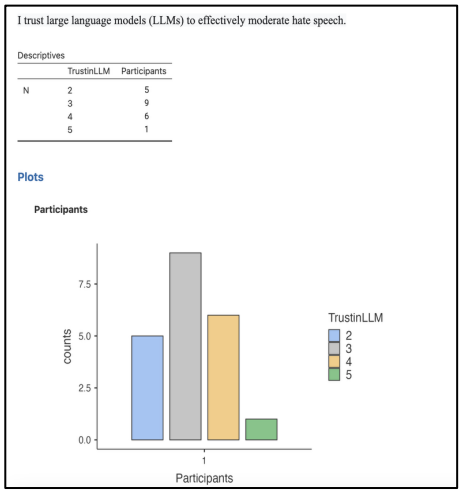

Figure 2d (2=Agree, 3=Neutral, 4=Disagree, 5=Strongly Disagree)

We also observed that participants' responses to *I trust LLMS to effectively moderate hate speech* showed that participants had a high rating of neutrality (N = 8) and a leaning towards disagreement (N = 8 for disagreement v N = 5 for agreement). Few participants had strong confidence in LLMs (N = 5).

4.3 Paired Samples T-test results

A series of paired samples t-tests were conducted to observe differences in perceptions of LLM.

Paired Samples T-Test

			statistic	df	p
BeliefofFairBias	BeliefofLLMSuccess	Student's t	−9.98	20.0	< .001

Note. H_a $\mu_{Measure\ 1 - Measure\ 2} \neq 0$

Table 3a

We observed that participants' beliefs in Fair Bias and their Beliefs in LLM Success were highly correlated ($p < .001$).

Paired Samples T-Test

			statistic	df	p
BeliefofBias	BeliefofFairBias	Student's t	12.20	20.0	< .001
	BeliefofLLMSuccess	Student's t	−3.18	20.0	0.005

Note. H_a $\mu_{Measure\ 1 - Measure\ 2} \neq 0$

Table 3b

We also observe statistically significant correlations between participants' beliefs of bias and their perceptions of bias being fair ($p < 0.001$) and successful ($p = 0.005$) (Table 3b). This aligns with general participants' responses, considering that many participants believed that bias was exhibited and that bias was unfair.

Paired Samples T-Test

			statistic	df	p
TrustinLLM	BeliefofBias	Student's t	2.15	20.0	0.044
	BeliefofFairBias	Student's t	8.17	20.0	< .001
	BeliefofLLMSuccess	Student's t	−1.16	20.0	0.258

Note. H_a $\mu_{Measure\ 1 - Measure\ 2} \neq 0$

Table 3c

Trust in LLMs was significantly related to perceptions of bias and fairness (Table 3c). We observe that participants reporting trusting LLMs were correlated to their ultimate perceptions of them being biased ($p = .044$) and more likely to believe that any bias present is fair ($p < .001$). However, trust in LLMs was not a significant predictor of users' beliefs of LLM success, suggesting that perceived fairness and bias are more influential in shaping trust than perceived effectiveness.

Paired Samples T-Test

			statistic	df	p
LLMContentBias?	TrustinLLM	Student's t	−10.58	20.0	< .001
	BeliefofBias	Student's t	−11.66	20.0	< .001
	BeliefofFairBias	Student's t	−1.31	20.0	0.204
	BeliefofLLMSuccess	Student's t	−10.95	20.0	< .001

Note. H_a $\mu_{Measure\ 1 - Measure\ 2} \neq 0$

Table 3d

We observe similar results in users' general perceptions of LLMs exhibiting potential bias. Table 3d demonstrates that participants who viewed LLM content moderation as biased were highly correlated with their likelihood of trust in LLMs (t (20) = -10.58, $p < .001$), and with their perceptions of these LLMs being biased (t (20) = -11.66, $p < .001$). It was also highly correlated with their beliefs about whether the models would be successful (t (20) = -10.95, $p < .001$). These findings suggest that participants who perceive LLMs as politically or ideologically biased tend to distrust and devalue them, regardless of performance metrics.

Paired Samples T-Test

			statistic	df	p
LLM Effectiveness V Human	LLMContentBias?	Student's t	−1.75	20.0	0.096
	TrustinLLM	Student's t	−12.20	20.0	< .001
	BeliefofBias	Student's t	−12.39	20.0	< .001
	BeliefofFairBias	Student's t	−2.68	20.0	0.014
	BeliefofLLMSuccess	Student's t	−12.49	20.0	< .001

Note. H_a $\mu_{Measure\ 1 - Measure\ 2} \neq 0$

Table 3e

Table 3e compares LLM effectiveness versus human moderators across key user beliefs. We observe significant correlations in users' general perceptions of LLM content moderation effectiveness and users' trust, belief in bias, fair bias, and belief in LLM success. Perceptions of content bias were not significantly correlated with users' general perceptions of LLMs exhibiting bias ($p = .096$).

4.4 Correlation Matrix

A correlation analysis was also conducted to explore relationships between key beliefs about LLMs (Table 1, Appendix). We observe a strong positive correlation between "Trust in LLMs" and "Belief in LLM Success" ($r = 0.667$, $p < .001$. "Belief in Bias" and "Fairness of Bias" were also positively correlated: $r = 0.608$, $p = 0.003$. Several other relationships, including those between LLM effectiveness and other variables such as trust, success, and bias, showed weak or non-significant correlations.

These results suggest that perceptions of trust, fairness, and success are closely connected. Even when participants acknowledge that LLMs can be accurate, concerns about bias and fairness still influence how much they trust the system in general.

5. Discussion

Our study set out to investigate how university students perceive the use of large language models to moderate hate speech content, and if their perception of LLM performance, as well as their judgment about what is classified as hate speech, correlated with demographic data about the students. Specifically, we wondered if students perceived LLM content moderation to be fair, if there existed any correlation between certain demographic groups and perception of bias within LLM content moderation, and how students would believe hate speech should be moderated, given the chance to classify themselves. We presented our participants with questions about LLM content moderation and then had them do two classification exercises. The first asked them to determine if an LLM's classification was correct or not for a particular set of tweets. The second asked them to classify tweets on their own. Given our research questions, we curated the following three hypotheses:

H1: Political affiliation will correlate with perceptions of LLM content moderation being biased.

H2: Students will believe LLMs can do a better job than manual administrators at moderating hate speech.

H3: Students will not perceive there to be substantial bias in LLM moderation of hate speech.

The results suggest that students were skeptical of LLM hate speech moderation (Figure 1b). It also suggests that they were skeptical about LLM hate speech moderation being more effective than human content moderation (Figure 1a). Despite this, our results shed light on the widespread skepticism across the political and demographic spectrum towards LLMs in their current form. Despite the limited sample size of the study, registering this skepticism among college students helps determine how social media apps should proceed with hate speech moderation if trust is a primary goal of theirs.

Regarding **H1**, which stated that political affiliation would correlate with perceptions of LLM content moderation as biased, we found that this hypothesis could not be adequately tested due to the high political homogeneity of our sample. This could mean that we did not have a large enough sample, considering that over 95% of our participants identified as left-leaning or Democratic, preventing a statistically meaningful comparison between ideological groups. This limitation underscores the need for more ideologically diverse samples in future research if political effects on trust in LLMs are to be validly assessed. If it were the case that we had a more diverse sample, this could be reflective of strong priors on LLM content moderation, and that the strength of those priors was stronger than political orientation. This would be an unsurprising interpretation given that the current population of

survey participants was drawn from friends of four computer science majors.

As it relates to the survey questions about their beliefs regarding LLMs and hate speech moderation, skepticism was widespread and the primary result of the survey. When asked if LLMs accurately moderated hate speech on the internet, responses skewed towards tentatively disagreeing that they could accurately moderate. When asked to compare LLM content moderation accuracy to human administrators, the survey respondents indicated they did not think LLMs would do as good a job as humans. In a related question about trusting LLMs to moderate content, there was a very uncertain result, with many remaining neutral on the question (N = 8). The only correlation we could identify with certainty was that respondents who reported they distrusted LLMs indeed were less confident in their ability to *successfully* moderate content, thus nullifying **H2**. This indicates that there is some connection between trust and efficacy for survey participants, and how much they trust them will impact how successful they are perceived. This is important as companies move forward with LLM content moderation, as our results indicate they should invest in building trust so that their community more positively perceives the introduction of LLM content moderation of hate speech.

Moving to questions about bias, the survey indicated that students felt like bias against political ideologies did exist in LLM content moderation of hate speech, thus nullifying **H3**, and that bias is not fair. If they perceived it to be fair, that could have been because they thought a particular political ideology was tied to more hateful speech, and thus the LLM could moderate them more harshly, but that could also be justified given the political circumstances. Students seemed to reject this conclusion, instead mostly concluding there was political bias in moderation, but thinking that the bias was unjustified. Thus, there is some lamentable behavior when LLMs moderate one political ideology more than another, according to the survey participants. This points more to the belief that fairness for the survey participants means moderating speech in equal amounts distributed across the political spectrum, even if hate speech is more prominent in some corners of the political universe. It also implies that the survey participants normatively assume bias to be bad in any form. When asked if LLMs could be trusted to be less biased than humans, the results were very uncertain. This could be interpreted in two ways. Either there is equally distributed skepticism towards both systems of moderating hate speech, or they believe the data itself is tainted by human bias, and the bias of human moderating systems taints the ability for LLMs to do so fairly. This seems to nullify H2, meaning that people do not trust it more than humans, definitively. Interestingly, the respondents seemed more certain that LLMs were more biased than they were certain LLMs would do a worse job of moderating content. This means there is another factor that we did not study that differentiates efficacy from bias, and there is some delta in how people perceive LLMs on both metrics.

As previously noted, there are limitations to the extent of our conclusions. The first limitation is the subject pool. We were only able to obtain 21 responses, a relatively small sample size for this study. We also had a large amount of concentration within the demographic questions, with many people sharing the same political ideologies and practices. This posed a challenge when it came to delineating much about whether certain demographic groups are strongly correlated with certain responses.

Relatedly, the researchers of this study note that due to recruitment largely being conducted through word of mouth, the participant pool was likely to be a more homogenous group of people than is ideal. This assumes people gravitate towards others who share similar beliefs. Thus, the implications of our results could have reflected the demographic homogeneity, and it could be that this particular pool of participants is skeptical of LLMs' ability to moderate hate speech content because many of them were in computer science, which may have made them more aware of algorithmic limitations, model opacity, or sociotechnical risks, leading to greater skepticism. It may be that, of a more average population, these results would be substantially different. At the very least, we can make these claims about UChicago students who have friends who are in computer science or computer science-adjacent majors, which says something about how that segment of society may view the LLM moderation of hate speech, but does not necessarily reflect more than that.

Future researchers should aim to recruit a larger and more demographically diverse sample to better capture variations in perception across political, cultural, and academic lines. Incorporating open-ended questions or interviews could provide deeper insight into why participants hold certain beliefs about LLM bias and effectiveness. Additionally, future studies could compare LLM moderation results with official platform policies and human moderator choices to evaluate alignment and consistency. Exploring interactive moderation tasks may also help assess how users respond to AI decisions in more natural settings.

With regards to **H2** and **H3**, we initially hypothesized that students would view LLMs as more effective than human moderators, and that they would not perceive substantial bias in LLM moderation. However, neither hypothesis was supported by our results. Upon reflection, this may be less a reflection of participants' misconceptions and more a misalignment in how we initially framed these expectations.

Prior literature and our introduction highlighted existing skepticism toward algorithmic fairness from a research perspective, particularly in sensitive domains like hate speech

detection. In this light, expecting participants to prefer LLMs over human moderators may have been unrealistic, especially when most users already lack transparency into how these models work or are trained. Similarly, **H3** underestimated participants' sensitivity to bias. In hindsight, a more appropriate hypothesis might have focused on the relationship between perceived bias and trust, or the conditions under which users are willing to defer to LLM decisions, rather than having an optimistic view about these systems. In terms of lessons learned, we as researchers would need to align hypotheses with both the theoretical literature and observed public attitudes towards LLMs

Our findings suggest that university students remain uncertain of LLMs in moderating hate speech and remain deeply skeptical of their fairness, current effectiveness, and political neutrality. As LLM moderation becomes more common, these perceptions highlight the need for greater transparency, accountability, and public engagement in how such systems are designed and deployed.

References

[1] Mahi Kolla, Siddharth Salunkhe, Eshwar Chandrasekharan, and Koustuv Saha. 2024. LLM-Mod: Can Large Language Models Assist Content Moderation? In Extended Abstracts of the CHI Conference on Human Factors in Computing Systems (CHI EA'24). Association for Computing Machinery, New York, NY, USA, Article 217, 1–8. https://doi.org/10.1145/3613905.3650828

[2] Thomas, K., Kelley, P. G., Tao, D., Meiklejohn, S., Vallis, O., Tan, S., Bratanič, B., Ferreira, F. T., Eranti, V. K., & Bursztein, E. (2024). Supporting human raters with the detection of harmful content using large language models (arXiv:2406.12800). arXiv. https://arxiv.org/abs/2406.12800

[3] Joseph Seering, Manas Khadka, Nava Haghighi, Tanya Yang, Zachary Xi, and Michael Bernstein. 2024. Chillbot: Content Moderation in the Backchannel. Proc. ACM Human Computer Interaction 8, CSCW2, Article 402 (November 2024), 26 pages. https://doi.org/10.1145/368694

[4] Chiu, K.-L., Collins, A., & Alexander, R. (2022). Detecting hate speech with GPT-3 (arXiv:2103.12407). arXiv. https://arxiv.org/abs/2103.12407

[5] Davidson, T., Warmsley, D., Macy, M., & Weber, I. (2017). Automated Hate Speech Detection and the Problem of Offensive Language. In *Proceedings of the 11th International AAAI Conference on Web and Social Media (ICWSM'17) * (pp. 512–515). Montreal, Canada.

[6] Paula Fortuna and Sérgio Nunes. 2018. A Survey on Automatic Detection of Hate Speech in Text. ACM Comput. Surv. 51, 4, Article 85 (July 2019), 30 pages. https://doi.org/10.1145/3232676

[7] Shagun Jhaver, Alice Qian Zhang, Quan Ze Chen, Nikhila Natarajan, Ruotong Wang, and Amy X. Zhang. 2023. Personalizing Content Moderation on Social Media: User Perspectives on Moderation Choices, Interface Design, and Labor. Proc. ACM Hum.-Comput. Interact. 7, CSCW2, Article 289 https://doi.org/10.1145/3610080

[8] Davidson, T., Bhattacharya, D., & Weber, I. (2019). Racial bias in hate speech and abusive language detection datasets (arXiv:1905.12516). arXiv. https://arxiv.org/abs/1905.12516

[9] Huang, F., Kwak, H., & An, J. (2023, April). Is ChatGPT better than human annotators? Potential and limitations of ChatGPT in explaining implicit hate speech. In Companion Proceedings of the ACM Web Conference 2023 (WWW'23, pp. 294–297). ACM. https://doi.org/10.1145/3543873.3587368

[10] Huang, Tao. 2024. *Content Moderation by LLM: From Accuracy to Legitimacy*. arXiv. https://arxiv.org/abs/2409.03219

Appendix

Correlation Matrix

		LLM Effectiveness V Human	TrustinLLM	LLMContentBias?	BeliefofFairBias	BeliefofBias	BeliefofLLMSuccess
LLM Effectiveness V Human	Pearson's r	—					
	df	—					
	p-value	—					
TrustinLLM	Pearson's r	−0.364	—				
	df	19	—				
	p-value	0.104	—				
LLMContentBias?	Pearson's r	0.139	−0.392	—			
	df	19	19	—			
	p-value	0.549	0.078	—			
BeliefofFairBias	Pearson's r	0.209	−0.384	0.232	—		
	df	19	19	19	—		
	p-value	0.364	0.085	0.313	—		
BeliefofBias	Pearson's r	0.058	−0.357	0.230	0.608	—	
	df	19	19	19	19	—	
	p-value	0.802	0.112	0.317	0.003	—	
BeliefofLLMSuccess	Pearson's r	−0.198	0.667	−0.236	0.066	0.092	—
	df	19	19	19	19	19	—
	p-value	0.390	< .001	0.303	0.776	0.691	—

Table 1a - Correlation Matrix

All analysis and data subsets used for this study can be found here: https://github.com/slflores0911/LLMHateSpeechContent

The two surveys used to conduct the study: https://docs.google.com/forms/d/1BeFrfpn-119N1oeNezemRSnsRC3LNdLWZOlC44NOm7U/edit

https://docs.google.com/forms/d/1-6y2wtJV-ViP4klMj10T9yi94T65i0xiuHMwk4L8IKwU/edit