Mod-Guide: An LLM-based Content Moderation Feedback System to Address Insensitive Speech toward Indigenous Ethnic and Religious Minority Communities

Anonymous submission

Abstract

Language operates as a mechanism of both marginalization and resistance, especially for minority communities navigating insensitive and harmful speech online. As content moderation increasingly depends on large language models (LLMs), concerns arise about whether these systems can recognize culturally insensitive speech from the perspectives of historically underrepresented groups. Focusing on Bangladesh's Hindu and Chakma communities-the country's largest religious and Indigenous ethnic minorities, respectively-this paper investigates the epistemic limits of LLM-based moderation systems and explores methods for incorporating minority perspectives. We co-created a culturally grounded corpus of insensitive speech with community members and integrated their narratives into moderation pipelines using retrieval-augmented generation (RAG). Our tool, Mod-Guide, improves LLM sensitivity to minority viewpoints by leveraging contextual cues derived from lived experience. Through mixed-method evaluations involving both minority and majority participants, we demonstrate that RAG-enhanced moderation responses are more contextually accurate and perceived differently across ethnic lines. This work advances research in human-computer interaction, AI ethics, and social computing by foregrounding restorative justice and hermeneutical inclusion in the design of content moderation systems.

Introduction

Language is more than a means of communication and is a form of power (Singh, Das, and Semaan 2025). It shapes social hierarchies, legitimizes authority, and enables the marginalization-a process through which individuals and groups are pushed to the periphery of society based on attributes like race, gender, ethnicity, religion, caste, nationality, language, sexual orientation, etc. (Erete, Israni, and Dillahunt 2018). Linguistic marginalization and injuries manifest in online communities through hate speech, bullying, political incitement, and other forms of insensitive speech. Most platforms respond by enforcing content moderation policies through a combination of human moderators and algorithmic systems (Jiang et al. 2023; Molina and Sundar 2022). Recent advances in large language models (LLMs) have enabled more scalable moderation (Kolla et al. 2024; Zeng et al. 2024), but these models are predominantly shaped by and reinforce majority perspectives (Li et al. 2024). Given the epistemic underrepresentation of the religious and Indigenous ethnic minorities, whose perspectives and experiences with insensitive or harmful speech might significantly differ from those of the majority in those LLM-based content moderation systems, it likely would reinforce the societal barrier between the majority and minority groups in the case of understanding each other's perspectives.

We focus on Bangladesh, where Hindu and Chakma communities represent the largest religious and Indigenous ethnic minorities (BSB 2022). Motivated by concepts of hermeneutical injustice (Fricker 2007) and the divide between majority and minority consciousness (Du Bois 2015), we collaborated with members from those communities to curate a corpus of culturally insensitive statements. Participants described why specific speech acts were hurtful and problematic, grounding their explanations in religious texts, oral histories, cultural practices and rituals, lived experiences, and documents from rights organizations. These insights reflect interpretive resources that are typically excluded from LLM training data.

To operationalize these perspectives, we introduce Mod-Guide, an LLM-based moderation feedback tool that uses retrieval-augmented generation (RAG) to ground moderation responses in this community-sourced corpus. While RAG has shown strong performance across a range of NLP tasks (Lewis et al. 2020), the significance of our work lies in grounding RAG with epistemically marginalized perspectives and evaluating its implications in the context of culturally sensitive moderation. We evaluate Mod-Guide using a mixed-method study with participants from the majority and minority communities, comparing its outputs to responses from the off-the-shelf GPT-4 model. Our analysis shows that grounding LLM responses in minority perspectives through RAG significantly affects how harmful speech is interpreted and moderated. We also find that the perceived usefulness of these moderation outputs varies by ethnicity but not by religion. This work makes two key contributions in line with recognized HCI contribution types (Wobbrock 2012):

- **Dataset contribution**: a curated and annotated corpus of culturally insensitive speech from minority perspectives.
- Artifact contribution: the design and evaluation of Mod-Guide, a feedback system that integrates these perspectives into the workflow of LLM-based moderation.

This research contributes to ongoing human-computer interaction, AI ethics, and social computing discourse by centering epistemically marginalized communities in data curation and system design. It demonstrates how LLMs can be made more sensitive to pluralistic norms through community participation and socio-technical design. The following sections detail the sociolinguistic framing of marginalization, our dataset's construction, the LLM-RAG pipeline's design, and the empirical evaluation. We conclude by reflecting on challenges of scale, normativity in dataset curation, and the implications for design toward community-centered justice and fair content moderation systems.

Literature Review

Linguistic Marginalization as Insensitive Speech

Language plays a crucial role in shaping social hierarchies and power dynamics. It establishes normative and nonnormative identities (Butler 2021). As such, people are marginalized through language, often in the form of bullying, hate speech, and threats. Similarly, religious and ethnic minorities are also vulnerable to linguistic injuries. Such injury arises not only from offensive speech targeting certain religions and ethnicities but also from the mode or ways those identities are positioned as dismissed and devalued (Butler 2021). In this paper, we focus on linguistic injuries and vulnerabilities, where exact words may not be explicitly offensive (e.g., name-calling), yet their conventional bearing-how words derive power from historical and social conventionscan come across as disregarding or diminishing the experiences, identities, practices, and contexts of religious and ethnic minorities, which we dub as insensitive speech.

To study the linguistic marginalization of religious and ethnic minorities in Bangladesh, we need to understand their sociopolitical contexts. Religious minorities in Bangladesh, particularly Hindus, have long faced marginalization characterized by both historical and ongoing violence (Rifat et al. 2024a). The large-scale communal riots and the disproportionate targeting of Hindus during the Liberation War illustrate this pattern (Sarkar and Bandyopadhyay 2017; Anam 2013). In recent decades, assaults on Hindu communities have increased, often fueled by social media rumors of religious insults against the majority (Ganguly 2021; Roy, Singh et al. 2023; International 2021), such as the violence during the 2021 Durga Puja (Hasan 2021). Furthermore, political instability worsens the persecution, leading to targeted attacks on Hindus (Ittefaq 2014; Alo 2024), Christians (International 2018), and atheists (France-Presse 2015; Shackle 2018). Similarly, the Indigenous ethnic minorities in Bangladesh (known as Adivasi) face marginalization due to their ethnic and cultural differences from the majority Bengali population. These communities, particularly in the Chittagong Hill Tracts, have experienced displacement, settlement, encroachment on their ancestral lands, ethnocide, and violence due to the region's militarization since before the country's independence (Chakma 2010; Hill and Chakma 2022). Despite a peace accord in 1997, they continue to struggle for autonomy and basic recognition of indigeneity to this day (Chakma 2008; Mizan and Rahaman 2025).

Recent scholarships in social computing and ICT for development have looked into how these sociopolitical experiences of religious and ethnic minorities in Bangladesh manifest as everyday linguistic marginalization in online communities in their interaction with other users and content moderation. For example, (Rifat et al. 2024a) explained how social psychology shapes the participation of religious minorities online, who, due to a fear of isolation, fall into a spiral of silence, negotiate through the future uncertainties and present impression of fear, and accommodate their communication with religious majority communities. Among the Indigenous communities in Bangladesh, many share religious minority identities, such as Chakma, Santhals, and Garo, who follow Buddhism, Hinduism, and Christianity, respectively (Us n.d.). Users from these communities have markedly different experiences with hate speech on online platforms compared to their peers from the majority community. The lack of urgency in addressing their experience with explicitly profane speech creates a clear disparity concerning membership, rights, and participation as users of online platforms (Sultana et al. 2024). Taking that into account, efforts to address insensitive speech with conventional bearing are more likely to be influenced by majoritarianism and, hence, require additional contextual content moderation and depend on increased awareness among majority religious and ethnic groups, such as the Bengali Muslims in Bangladesh.

Epistemic Barriers among Majority and Minority

Marginalization of minorities often stems from entrenched tribal stigma surrounding attributes like ethnicity, religion, language, and cultural practices (Goffman 2009). For example, in many contexts, misunderstandings of minority religions' practices and beliefs lead to unsubstantiated fear (e.g., Islamophobia (Allen 2016)), misrepresentation (e.g., depicting non-Abrahamic faiths as satanic or pagan (Sugirtharajah 2004)), or exclusion. Similarly, immigrants who speak different languages often face suspicion or hostility, as their speech is perceived as secretive or exclusionary, reinforcing their marginalization in the form of xenophobia (Lee 2019). Scholars argue that such stigma and marginalization are not the victims' attributes but a feature of the society that imposes it. Through various social processes, minorities' symbols, beliefs, practices, and physical conditions are made non-normative in society and are devalued or discredited to such an extent that they adopt different coping mechanisms (Goffman 2009), such as hiding their identities, avoiding sharing their experiences or withdrawing from social interactions out of fear of isolation and the desire to conform to norms in both online and offline settings (Rifat et al. 2024a).

In this paper, we seek to understand the experiences of religious and ethnic minorities being marginalized, ridiculed, and misunderstood in the Bangladeshi social media sphere by combining W.E.B. Du Bois' concept of "the veil" (Du Bois 2015) and Miranda Fricker's notion of hermeneutical injustice (Fricker 2007). These theoretical angles provide complementary lenses for understanding and addressing the underlying processes that lead to the minorities' marginalization. Du Bois' conceptualization of the "veil" highlights how racial minorities in the United States experience an imposed separation that distorts their selfperception and hinders mutual comprehension across racial divides (Du Bois 2015). Recent work (Rifat et al. 2024a) in the context of Bangladesh has highlighted how the religious minority communities feel a comparable divide between themselves and the religious majority, particularly in how their identities and practices are misinterpreted, leading to alienation and marginalization. That metaphorical veil between the majority and minority groups in terms of ethnicity or religion functions as an epistemic barrier, preventing adequate and effective intergroup understanding.

Drawing from Fricker's work (Fricker 2007), this epistemic difference could be dubbed hermeneutical injustice, where minority groups struggle to make sense of their experiences due to the lack of necessary conceptual resources within normative epistemic frameworks shaped by religious and ethnic majorities' beliefs and practices. For example, theological interpretations (e.g., the role of idols in worship for Hindu minorities) and dietary practices of the ethnic minority communities (e.g., consumption of pork, frog, and alcohol) are considered wrong from the perspective of the majority Bengali Muslims' standpoint (Rifat et al. 2024a; Sultana et al. 2024, 2022). When members of the majority community talk about those beliefs and practices, the minority groups might deem such comments as stereotypical, condescending, insulting, and overall insensitive, which reinforces division and further marginalizes minorities online.

Divisions between majority and minority groups are sustained by institutionalized ignorance and a lack of empathy (Du Bois 2015), while dominant social norms and unconscious biases perpetuate injustice against marginalized communities (Fricker 2007). In online communities where religious and ethnic minorities encounter insensitive speech, different moderation and feedback mechanisms could be implemented with careful attention to the epistemologies of these groups. More broadly, dismantling these barriers demands inclusive epistemic practices-encompassing knowledge production, recognition, and validation-to value minorities' perspectives and foster interfaith communication and mutual understanding. These practices would ultimately address the power asymmetries experienced by religious and ethnic minorities online by shaping the design and governance of sociotechnical systems like online platforms.

Language Models in Moderating Insensitive Speech

With the global adoption of online platforms and the diverse communities they host, moderating harmful and insensitive speech has become a complex sociotechnical challenge. Existing scholarship has shown that perceptions of what constitutes harmful content and its severity vary significantly across cultural and social contexts (Jiang et al. 2021; Scheuerman et al. 2021). While platforms' "institutional ethics" (Scheuerman et al. 2021) do not want to implement the perspectives of users who think anything that does not pertain to a particular religious belief should be removed, they rarely make an active effort in addressing the hermeneutical injustice (Jiang et al. 2021), i.e., the structural exclusion of minority perspectives in defining what counts as harmful. As online communities grow, platforms must negotiate competing moderation values (e.g., community identity), philosophies (e.g., nurturing vs. punishing), and implementation styles (e.g., human vs. algorithmic moderation) (Jiang et al. 2023; Das, Østerlund, and Semaan 2021).

Particularly focused on moderation philosophy, Seering et al. (Seering, Kaufman, and Chancellor 2022) examined how moderation can be conceptualized through different metaphors, such as mentoring, law enforcement, and custodianship. These metaphors shape how platforms and moderators perceive their roles, influencing decisions and ethics about intervention, the balance between users' autonomy and governance, and the prioritization of different cultures and values. As the platforms adopt algorithmic moderation for the sake of efficiency, these societal complexities are often pawned off to algorithmic systems (Jiang et al. 2023). Language technologies have become central to automated content moderation systems(Sun and Ni 2022; Vaidya et al. 2021). In terms of complexity and sophistication, these systems range from simple keyword filters(Jhaver et al. 2019, 2022), to task-specific models for sentiment analysis and hate speech detection (Das et al. 2024; Mozafari, Farahbakhsh, and Crespi 2020), to foundational large language models (LLMs) deployed at scale (Kolla et al. 2024; Zeng et al. 2024; İnan, Çetin, and Yakupoğlu 2024). While multilingual LLMs have shown promising results in detecting explicit hate speech, fake news, and discriminatory language (Plazadel Arco et al. 2023; Koka et al. 2024; Orlandi et al. 2021), they often struggle with more subtle forms of disinformation and culturally coded insensitivity.

However, LLMs reflect and reinforce dominant cultural norms, which can lead to representational harms, particularly for non-Western communities (Ghosh et al. 2024; Brown et al. 2024). Prior research has shown that these models exhibit demographic (e.g., race, gender, nationality, religion, caste) (Ghosh and Caliskan 2023a,b; Hamidieh et al. 2024; Ghosh 2024; Das et al. 2024), socioeconomic (Arzaghi, Carichon, and Farnadi 2024), and political biases (Agiza, Mostagir, and Reda 2024), raising concerns about how automated moderation disproportionately impacts marginalized communities. Hence, recent works have attempted to reconceptualize moderation by embedding safety paradigms directly into LLM pipelines (İnan, Çetin, and Yakupoğlu 2024; Arzberger et al. 2024), wherein they have examined how data selection and fine-tuning impacted LLMs' economic and political biases (Agiza, Mostagir, and Reda 2024), how model responses vary with culturally sensitive prompts (Mukherjee et al. 2024), and found that persona-based prompting can improve alignment with specific moderation goals (Kwok, Bravansky, and Griffin 2024). Studies highlighted how crowdsourced data annotation is subject to limited annotator expertise (Kumar et al. 2024), dismissal of religious faiths (Rifat et al. 2024b), minorities' underrepresentation (Thorne 2022; Song et al. 2025), and disproportionate association of toxicity with minorities (Wiegand, Ruppenhofer, and Eder 2021). Retrieval augmented generation (RAG)-a method to enhance language model outputs by retrieving relevant external documents while generating responses (Lewis et al. 2020), can be an effective technique to address the concerns of LLM biases affecting content moderation (Leitner et al. 2025; Tsirmpas, Androutsopoulos, and Pavlopoulos 2025). However, there is a dearth of literature that has examined its effectiveness in moderating content around minority identity, especially in non-English languages and the Global South contexts.

Our work advances research at the intersection of content moderation, LLMs, and low-resource language communities in two key ways. First, we address the dataset challenge by constructing a culturally grounded corpus of insensitive speech in Bengali, annotated and contextualized by members of underrepresented religious and ethnic minority communities in Bangladesh. Rather than relying on crowd-sourced or majority-labels that often obscure minority perspectives, our approach centers the lived experiences, interpretive frameworks, and rationales of those most affected by marginalization. Second, we build on insights from prior literature that persona-based prompting may help align LLM outputs with specific moderation philosophies (Seering, Kaufman, and Chancellor 2022; Kwok, Bravansky, and Griffin 2024) and RAG enhances factual accuracy and contextual grounding (Izacard and Grave 2020; Shi et al. 2023). We implemented this insight in our content moderation feedback system, Mod-Guide, in which we prompt an LLM to adopt various moderation roles and ground its responses in the minority community-sourced corpus using RAG. We evaluated which configurations-combinations of prompts and the presence/absence of RAG-produce more contextually sensitive, factually accurate, and epistemically inclusive feedback.

Corpus Preparation to Understand Minority Hermeneutics

This paper is part of a broader study to understand minority communities' experiences with content moderation in online communities and develop tools to make those spaces more inclusive and accessible for these communities [citations redacted for review]. Here, we build on our findings and community relationships fostered during the earlier phases of our research. We collected the corpus through the Asynchronous Remote Community (ARC) method (MacLeod et al. 2016). Prior research has used this method to engage with participants where in-person communication can be difficult to arrange due to population distribution (MacLeod et al. 2017), stigma (Maestre et al. 2018), or fear of isolation (Walker and DeVito 2020). Over a month, we weekly elicited those groups to sustain engagement while allowing flexibility for participants to share instances of social media posts that they found culturally insensitive, about which they have found that the religious majority and ethnic majority communities have different perceptions. In doing so, our corpus prioritizes minority hermeneutics-interpretation of their practices, experiences, values, and beliefs from their own perspectives, over being shaped or constrained by majoritarian normative societal views.

Participants

In this paper, we focus on the religious minority Hindu community and the Indigenous ethnic minority Chakma community in Bangladesh. We recruited participants aged 18 years and older by sharing the recruitment materials and additional information with our personal networks, through Facebook advertisements, and by reaching out to participants from our previous studies involving these communities. We also contacted the administrators and moderators of local Facebook groups dedicated to these minority communities, asking their permission to post the call for participation in those groups. We asked the respondents to the study's advertisements to self-identify key characteristics such as gender, caste, age, and their places of upbringing and current residence, which prior studies found to have differing experiences within the Hindu and Chakma communities (Sultana et al. 2024; Rifat et al. 2024a). Our ARCs with these participants included 11 from the religious minority Hindu community (7 male and 4 female) and 11 from the Indigenous ethnic minority Chakma community (2 male, 4 female, and 5 did not respond to the question asking their gender). Both ARCs had more members, but those who did not post at least once in the groups were excluded from the reported counts. Most of our Hindu participants were from underprivileged scheduled castes (tafsili jati) (Sen 2018), reflecting the general demographic pattern of Hindu communities in Bangladesh.

Procedure

Similar to previous ARC studies (MacLeod et al. 2016; Maestre et al. 2018; Walker and DeVito 2020), based on our participants' preferences, we used a secret Facebook group and a secret WhatsApp group, respectively, to interact with the former and the latter minority groups. Hosting the ARCs on these online platforms minimized the need to familiarize participants with a new system (MacLeod et al. 2017; Heywood, Ivey, and Meuter 2024). All participants had existing Facebook and WhatsApp accounts that they used to participate in the study, thus maintaining platformrelated risks similar to those participants regularly assume while using these communication channels. After completing our informed consent procedure and orienting them with a code of conduct, we invited them to join the groups. From 25/10/2024 to 23/11/2024, we maintained engagement through weekly elicitation while allowing for flexibility.

However, a few participants either did not actively engage or ceased participating after the first couple of weeks in those Facebook and WhatsApp groups, which is a pattern of attrition and participation consistent with previous ARC studies (Prabhakar et al. 2017; Walker and DeVito 2020). The other participants responded to our prompts by sharing examples of textual posts, comments, images, and videos they perceived as insensitive to their religious and Indigenous ethnic identities, cultures, rituals, and practices. We specifically sought instances that were often dismissed as nonproblematic by the religious and ethnic majority communities, as the participants experienced through interacting with friends and acquaintances in those communities or having their reports of such content overlooked by content moderation systems on online platforms. We also asked the participants to explain why they found the contents insensitive, referencing sources such as the scriptures of the religious minority communities, national and international resolutions regarding the rights and concerns of the Indigenous ethnic minorities, and their lived experiences and understanding of their respective communities. The participants also engaged with and built upon each others' responses. The first two authors monitored the groups to ensure compliance with the code of conduct and asked follow-up questions to nudge the participants to share additional details.

Our participants participated in the ARCs using both Bengali and English. To streamline the corpus, we translated all written communication into Bengali, the national language of Bangladesh. Participants occasionally shared screenshots of social media posts they considered insensitive. We utilized optical character recognition (OCR) to convert those images into Bengali text. Similarly, any videos shared by participants as examples were transcribed into Bengali text. For web URLs provided by participants as instances of hate speech, we transcribed the content into Bengali. We anonymized all these contents before OCR and transcription. For common examples of insensitive speech, some participants shared links to online repositories containing writings and references explaining why such remarks are inappropriate. We scraped the web pages in those cases, excluding nontextual content (e.g., HTML tags, URLs).

We gathered 53 instances of insensitive speech directed at the religious minority Hindu community and 79 instances targeting the indigenous ethnic minority Chakma community, organizing them into two separate spreadsheets. Each spreadsheet contains two columns: one listing examples of insensitive speech and the other explaining their inappropriateness. Let's consider the following example text that Hindu participants in our ARCs found to be culturally insensitive.

কিছু মানবতার ফেলিওয়ালাদের দেখতেছি, মূর্তি পাহা-রা দিতে মন্দিরে যাচ্ছে। মূর্তি পাহারা দেওয়ার জন্য ঈমান আনিও নাই, মূর্তি পাহারার পক্ষে আমি নাই। ভাঙ্গা লাগলে ডাক দিয়েন (I have been seeing some vendors of humanism who are going to temples to guard the idols. I did not bring *imaan* (faith in Islam) for guarding the idols, [and] I am not in favor of guarding the idols. Call [me] if those [idols] need to be broken.)

Since this example text was collected from the post of a user belonging to the religious majority, it reflects their cultural value and belief: the prohibition on idol worship in Islam. In contrast, in the Hindu faith, idols are viewed as a medium for worship. Consequently, a few of our participants pointed out the aforementioned text that was recently wellcirculating in the Bangladeshi social media sphere as insensitive speech. They also explained why they consider it culturally insensitive from different angles. For example, while some participants explained the relevance of idols in Hindu rituals based on references from Hindu scriptures, some others presented arguments informed by their observations of social practices in different religions. For example, a snippet of the explanation for the above text being insensitive was:

জগতে নিরাকারবাদী বহু সম্প্রদায় আছে, যাহারা অবতা-রবাদ মানেন না এবং উপসনা কিংবা সাধণার জন্য কো-নরূপ সাকার বিগ্রহ বা প্রতীকের প্রয়োজন হয় না। অনে-কে আবার নিরাকারবাদ গ্রহণ করেও অবস্থাবিশেষে প্র-তীকের (অউম, ধর্মচক্র, স্টার অফ ডেবিড) আবশ্যকতা স্বীকার করে। তাঁরা ঈশ্বরের বাহ্যিক মূর্তি স্বীকার করেন না কিন্তু তাঁরাও মনে মনে কোন না কোন মূর্তি কল্পনা করে থাকেন অথবা হৃদাকাশে কোন প্রতীকের কল্পনা করেন।

... এই বিষয়ে স্বামী বিবেকানন্দ একটা কথা বলেছিলেন, "দুই প্রকার মানুষের রুপ-কল্পনার বা মূর্তির প্রয়োজন হয় না-যে ধর্মের কোন ধার ধারে না; আর সিদ্ধপুরুষের, যিনি এই-সকল অবস্থার মধ্য দিয়া গিয়াছেন। আমরা এই দুই অবস্থার মধ্যে রহিয়াছি। ভিতরে ও বাহিরে আমাদের কোন-না-কোনরুপে মূর্তির প্রয়োজন। ..." (There are many formless-theist communities in the world who do not believe in incarnations and do not require any tangible deity or symbol for worship or spiritual practice. Again, some who accept formless-theism still acknowledge the necessity of symbols (such as Om, the Dharma Wheel, or the Star of David) in certain contexts. While they do not accept an external image/idol of God, they still mentally envision some form or symbol within their hearts. On this matter, Swami Vivekananda once said: "Two types of people do not require forms or idols-those who have no concern for religion at all, and the enlightened beings who have transcended all such states. We exist somewhere in between these two conditions. Internally and externally, we need some form of an idol or image.")

We emphasize that our work does not seek to evaluate different theological beliefs and practices. Rather, we aim to highlight how various cultural, religious, and social values influence people's perceptions of content sensitivity and the roles they expect moderators to fulfill. Hence, we will use this corpus of speech the minority communities viewed as culturally insensitive and the rationales behind such perceptions to inform LLM-based automated content moderation.

Mod-Guide: Persona-based LLM Prompting and RAG Pipeline for Moderation Feedback

This paper investigates the effectiveness of large language models (LLMs) in moderating insensitive speech directed at religious and ethnic minority communities in Bangladesh that is often based on stereotypes and deepens the cultural divide between the majority and minority communities in the country. Drawing on Du Bois (Du Bois 2015), we refer to that as the veil. We particularly examined OpenAI's GPT-4. Additionally, we explore retrieval-augmented generation (RAG) based on community insights with the application for content moderation in mind. We prompted these systems in different ways to shape their responses by defining their role as content moderators (e.g., nurturing, governing) and persona (e.g., teacher, judge) to build a moderation feedback system called "Mod-Guide." In this section, we will first discuss the different prompts, then explain the LLM and RAG pipeline, and finally discuss the interface with examples.

Prompting

We used five prompts in the pipeline. In all these cases, the objective was the same: to understand whether a piece of text would potentially be perceived as hate speech, hurtful, or culturally sensitive. Those prompts featured different roles of moderators of online interaction. Drawing on Seering et al. (Seering, Kaufman, and Chancellor 2022), we considered five different roles content moderators play in moderating

online communities, such as (a) nurturing and supporting, (b) overseeing and facilitating, (c) governing and regulating, (d) managing, and (e) fighting for communities. Based on the recommendations of the same study (Seering, Kaufman, and Chancellor 2022), for these different roles, we chose the metaphors teacher, mediator, judge, representative, and protector, respectively.

- *Prompt 1:* As a supportive and nurturing content moderator like a teacher, evaluate whether this text could be perceived as hate speech, hurtful, or culturally insensitive. Consider if it marginalizes, reinforces stereotypes, or excludes any group. If so, provide constructive feedback by identifying concerns, explaining why they may be problematic, and suggesting more inclusive alternatives.
- *Prompt 2:* As a moderator who oversees and facilitates like a mediator, consider whether this text could be perceived as hate speech, hurtful, or culturally insensitive during interactions between majority and minority communities based on religion and ethnicity. If it might escalate tension, explain different perspectives and suggest revisions that promote respectful and inclusive dialogue.
- *Prompt 3:* As a content moderator who governs and regulates like a judge, determine whether this text constitutes hate speech, is hurtful, or culturally insensitive. Decide if it violates principles of fairness, dignity, or inclusivity, and provide a clear ruling to retain or remove the content.
- *Prompt 4:* As a moderator familiar with religious and ethnic relationships in Bangladesh, assess whether this text represents the country's broader societal values. Consider if it could be perceived as hate speech, hurtful, or culturally insensitive to members of any community. Provide feedback by highlighting potential issues and suggesting ways to foster respectful and inclusive dialogue.
- *Prompt 5:* As a content moderator who protects, advocates, and looks out for religious and ethnic minorities like Hindus and Chakmas, examine if this text could be perceived as hate speech, hurtful, or culturally insensitive to them. Instead of reinforcing stereotypes, erasing voices, or contributing to harm against these marginalized groups, explain how it can center respect and inclusion.

We added an extra instruction to all five prompts— "Answer briefly and translate that in the Bengali language before responding"—after observing that the LLMs, with or without RAG, tended to respond primarily in English even when prompted in Bengali. This addition was intended to ensure that the feedback would be generated in Bengali.

LLM and RAG Pipeline

The RAG and LLM pipeline consisted of a data preprocessing and ingestion phase, a prompting step to define the tasks of the content moderator, and the LLM or RAG component (see Figure 1). We developed and operated the pipeline between December 2024 and January 2025.

To evaluate the LLM and RAG, we designed five prompts, as described above, that embodied distinct moderator metaphors, each reflecting a different moderation approach. The off-the-shelf LLM we are using is GPT-4 from



Figure 1: Prompt, LLM, and RAG pipeline.

OpenAI, which supports controlled retrieval, where it is up to the language model to decide if retrieval is necessary. We designed the script to do forced retrieval using a separate system prompt, where we used the five prompts outlined earlier to define the persona of the LLM-based content moderation (see path 1 in Figure 1). Under the hood, OpenAI generates a small query based on the prompt that triggers a retrieval tool call. Next, we generated evaluation questions, where we asked if an example from our corpus could be considered insensitive speech by religious or ethnic minorities. Then, we asked these evaluation questions to the LLM (see path 2 in Figure 1). The retrieval tool performs a similarity search against this query in the vector store, which contains embeddings of knowledge collected from the minority communities. The corpus collected from the minority communities provides additional cultural and situational context, along with explanations of why these communities perceive certain example texts as insensitive. The retrieved information is then processed based on the system prompt from earlier to generate an output. The data is then processed through a pipeline to build a retrieval-augmented generation (RAG) component using LangChain, allowing the LLM to reference it during inference. Based on the general length of our pairs of example text and the corresponding explanation of that being culturally insensitive, we used recursive character text splitting with chunk size=512 and k=2 so that the embeddings do not lose context, and both the text and the explanation are retrieved if the pair is split between two different chunks. We asked the same evaluation questions to the LLM (see path 3 in Figure 1), but this time, it could utilize RAG. Thus, we obtained two sets of responses-one from the standalone LLM and another from the RAG-enhanced system, enabling a comparative evaluation of their effectiveness.

Interactive Interface

We developed an interactive user interface (UI) around our LLM pipeline, enabling users to receive feedback on their texts while leveraging RAG based on the community-sourced corpus and exploring different moderation personas without requiring prior knowledge of these mechanisms and prompt engineering. We chose a web-based interactive interface due to its platform independence and ease of access across different devices. First, we created a high-fidelity prototype in Figma, which served as a blueprint and guided the UI's development process and maintained design consistency throughout the project. Then, we developed the final interface using React.js. Its use in the front end enhances performance due to the framework's virtual document object model and facilitates seamless updates, resulting in a dynamic and responsive user experience. In the back end, we handled server-side logic and API calls using Python.

We refer to the integrated platform—comprising the corpus, the LLM pipeline with RAG and prompt variations, and the user interface—as **Mod-Guide**. This tool assists users in online communities to identify and avoid culturally insensitive speech, simulating the role of a content moderator. For instance, when we input the example text discussed in the previous section and asked Mod-Guide to respond in the role of a mediator, it generated the Bengali feedback shown in Figure 2 that translates as:



Figure 2: Feedback from Mod-Guide's in Mediator role.

This text could be regarded as insensitive or offensive to certain religious or cultural communities, specifically those who value idol worship or temple practices. The statement dismisses the significance of idol protection and implies disrespect towards the faith associated with these practices. To defuse potential tensions, it is advisable to rephrase the statement to focus on promoting mutual respect for diverse beliefs. A possible revision could be emphasizing the importance of understanding and respecting each other's religious practices, fostering a community where diversity in beliefs can coexist peacefully.

This feedback adopts some high-level insights and similar wording from the explanations provided by the minority community members in the corpus. However, the questions remain whether the responses become significantly different if LLM uses RAG based on the community-sourced corpus, whether the responses are factually correct, and how users from minority and majority religions and ethnicities find those responses useful.

Evaluation of Moderation Feedback

We adopted a mixed-method evaluation approach in our study, where we considered content moderation persona, whether community knowledge corpus was provided for RAG, and which LLM model was used as independent factors. We compared the effectiveness of their combinations in moderating insensitive speech toward religious and ethnic minorities, in other words, addressing hermeneutical differences of these communities with the majority religious and ethnic group in the country. We evaluated the moderation feedback based on three criteria by asking the following questions in the evaluation phase:

1. Difference in textual response:

- (a) How do various prompts impact text generation in LLM-based content moderation?
- (b) How does the use of RAG impact text generation in LLM-based content moderation?
- 2. **Factual accuracy:** Is the feedback generated in LLMbased content moderation, both without and with RAG, factually accurate?
- 3. Users' perceived usefulness: How do people's demographic backgrounds and the persona of LLM-based content moderation influence the perceived usefulness of the feedback?

Quantitative Analysis of Textual Differences

To analyze textual differences and similarities between responses generated by off-the-shelf LLM GPT-4 and those generated through RAG with community-generated knowledge as context, we employed BERTScore, which leverages contextual embeddings to measure token similarity to offer strong alignment with human judgments and greater robustness to adversarial paraphrases compared to traditional text generation metrics (Zhang et al. 2019). However, there is a dearth of research on whether a metric like BERTScore works well for low-resource languages like Bengali. While future NLP research should look into the cross-language applicability of this metric, our evaluation tried to address this concern by using a multilingual BERT model.

To compare whether and how five different content moderation personas (reflected through prompts) influence the generated responses from the LLM, we analyzed the responses' variance across different prompts. First, we used the distiluse-base-multilingual sentence encoder to find the embeddings of the responses generated for prompts reflecting different moderation personas. Then, we calculated the Euclidean distances of the embeddings for different pairs of prompts. Based on whether or not the distance scores maintained normality in the Shapiro-Wilk test, we used a series of parametric paired t-tests or non-parametric Wilcoxon signed-rank tests, respectively, to compare responses for $\binom{5}{2} = 10$ pairs of persona prompts based on the Euclidean distances of their embeddings.

In answering evaluation question 1(a), our null hypothesis was: "There is no significant difference in the text generated by LLMs, measured by the Euclidean distance of their embeddings, for prompts reflecting different content moderation personas." With Bonferroni correction, our results for all pairs of prompts ($p < 10^{-22}$) provided strong evidence that there is a significant difference in the text generated by LLMs for prompts reflecting different moderation personas.

To answer question 1(b), we tested the influence of the use of RAG on text generation using a similar approach. Since the distances of the embeddings of texts generated by offthe-shelf GPT-4 from OpenAI and with RAG did not follow a normal distribution, we used the Wilcoxon signed-rank test. Assuming a null hypothesis: "There is no significant effect of using RAG on the responses of the LLMs". We obtained p =3.3e - 54, based on which we rejected the null hypothesis, i.e., we found strong evidence of RAG based on communitysourced corpus affecting the generated texts.

Qualitative Analysis Responses' Factual Accuracy

There exist few studies focused on evaluating the factual accuracy of long-form text generated by LLMs without any human effort (Min et al. 2023). Due to considerable disparities in resources and online presence, these approaches remain unusable in non-English languages, like Bengali. Moreover, especially in contexts of minority religious faiths and Indigenous ethnic practices, where interpretations are crucial, evaluation of models by human participants is more appropriate.

In this part of the evaluation, we recruited two expert participants, one from each minority community, through convenience sampling (Etikan et al. 2016). These participants did not participate in the earlier corpus generation phase but were well familiar with their corresponding communities' cultures. We presented them with ten randomly selected posts' responses and explanations generated in LLM-based content moderation, from GPT-4 without and with RAG, and inquired whether the explanations were factually accurate and where the LLMs' responses were lacking. Following sharing the random sample of responses as a spreadsheet, the first author regularly communicated with the participants asynchronously over a week. To analyze their feedback, we used iterative thematic coding, which is widely used in human-computer interaction research (McDonald, Schoenebeck, and Forte 2019; Bowman et al. 2023). In this approach, we identified codes-identities, groups, topics, or issues that appeared repetitively across multiple iterations. We later aggregated the related codes into broader themes.

The expert participant (E1, male, 35) believed that the information provided in most responses from the LLMs was somewhat correct. However, the responses obtained directly from GPT-4 were shallow compared to the ones generated by augmenting its responses through retrieval from communitysourced data. For example, for the statement "হিন্দু ধর্মাবল-ম্বীদের মূর্তিপূজা করা উচিত নয় কারণ "ন তস্য প্রতিমা অস্তি" (যজুর্বেদ ৩২/৩) (*"Hindus should not worship idols [reference to Sanskrit verse from Hindu holy text the Vedas]"*), participant E1 said,

I find the first response [from GPT-4] to somewhat lack in depth. It correctly emphasizes the need to respect and understand religious beliefs but does not address the central topic [role of idols]. [But,] the second response, "কিছু হিন্দু ধর্মের অনুসারীরা মূর্তিপূজা-কে তাদের ভক্তি প্রকাশের একটি উপায় হিসেবে বিবে-চনা করেন, যা তাদের আত্মার সাথে ঈশ্বরের সংযুক্তির একটি মাধ্যম।": ("Some Hindus consider idol worship as a way of expressing their devotion, a means of connecting their souls to God.") [from RAG] provides a more nuanced perspective. While the verse mentioned is correct, it is translated literally. It could be interpreted to recognize different theological traditions within Hinduism regarding the role of idols in worshiping.

The expert acknowledged that there are a few schools of Hindu faith that do not permit idols in worship, but a general critique of that comes across as culturally insensitive for Hindu minorities, which is correctly mentioned in responses from LLM without and with RAG. His explanation highlights that while the cited Vedic verse mentioned in the text is accurate in a literal sense, its interpretation varies across different Hindu sects. Therefore, prescribing a singular viewpoint, especially by someone who is not a religious minority, can be insensitive to diverse Hindu cultures.

However, in several other instances, the RAG based on the community-sourced corpus improved the information quality of the responses, which directly influenced the accuracy of the feedback. For example, evaluating the following statement made by a user from Bengali majority community: "যারা পাহাড থেকে সেনা হটানোর দাবি জানাচ্ছেন, আপনা-রা এইটাও জানেন না যে আপনারা অনেককিছুই জানেন না!" ("Those of you who are demanding that the army be removed from the hills, you don't even know that you don't know much!"), GPT-4 decided that "This text does not constitute hate speech, is not hurtful, nor culturally insensitive." Our expert from the ethnic minority community (E2, male, 32) argued that this response does not take the experiences and perspectives of the Indigenous ethnic communities in the Chittagong hill tracks. In contrast, the response from LLM with RAG was "This statement could be perceived as dismissive or condescending towards those requesting the withdrawal of forces from hill areas. ... For a more respectful dialogue, consider framing it as a call for understanding and dialogue rather than an outright dismissal: "যারা পাহাড় থে-কে সেনা হটানোর দাবি জানাচ্ছেন, আসুন সবাই মিলে আলো-চনা করি এবং আরও জানার চেষ্টা করি।" (Those who are demanding to withdraw the army from the mountains, let us all discuss together and try to know more.)", which E2 found more insightful and accurate.

In some cases, LLM, without and with RAG, fails to capture different problematic aspects of insensitive speech. Let's consider the following screenshot (see Figure 3) shared by one of our ARC participants. After applying OCR on this, we retained only the text but not the image. The text uses the term, 'upojati' ('উপজাতি', literal translation: sub-nation), which is often used as a slang for the ethnic minorities, which they find offensive (Sultana et al. 2022). Participant E2 also focused on the text's use of vulgar language targeted at Indigenous women. LLM's response discussed and reflected on the former issue and recommended the "use [of] precise and accepted terminology that members of these communities identify with. In Bangladesh, 'Adibashi' or 'Indigenous Peoples' might be more appropriate than 'upojati'." However, neither the use of GPT-4 nor the use of RAG on top of that focused on the latter issue. This shortcoming might be a result of not having enough context possibly obtainable from the image or LLM's systematic overlooking of Indigenous women's concerns, which should be explored further and more rigorously in future work.



Figure 3: A screenshot shared by an ARC participant.

Quantitative Analysis of Perceived Usefulness

We conducted a quantitative evaluation to understand whether individuals from various religious and ethnic backgrounds find the feedback from LLM-based content moderation useful and which persona they prefer. For this phase, we recruited a combination of 15 participants from the ethnic and religious majority and minority communities, such as Bengalis, non-Bengali Indigenous groups, Muslims, and Hindus. Among those from the minority communities, three participants also took part in the corpus collection or factual accuracy evaluation phases. For a randomly selected sample of texts, we presented the participants with feedback from LLMs with five different content moderation prompts. To avoid possible inconsistencies among participants in interpreting Likert scale levels (Cummins and Gullone 2000), we asked them to identify the feedback they perceived to be the most useful and explain why they found those more useful compared to others. We analyzed how the demographic background and the content moderation persona adopted (reflected through the prompts) influence the perceived usefulness of the LLMs' feedback using the χ^2 test (let, $\alpha = 0.05$).

In two separate tests focusing on demographic attributes, religion and ethnicity, we considered Bengali Hindus as the religious minority and ethnic majority, respectively. Based on our data, we did not find evidence (p = 0.596) to claim that there is a significant relationship between the participants' religious identity and responses from which persona they found useful. However, our data suggested that there is a relationship (p = 0.0104) between whether the participants were from the ethnic majority or the ethnic minority Indigenous community and the response resulting from which moderation persona prompt they found the most useful. We

allowed the participants to include small notes about the criteria they considered to decide the "usefulness" of the responses. Our participants shared that they prioritized factors such as empathic and inclusive language, promoting education and contextual awareness, etc. However, deeper qualitative studies in the future should look into whether and how different linguistic and informative aspects are prioritized across demographic variations.

Discussion

We have described how we collaborated with two religious and ethnic minority communities in Bangladesh to collect a corpus of insensitive speech, how we used different moderation personas to generate decisions and feedback on those examples of insensitive speech from GPT-4 model and how we informed the LLM through a RAG pipeline regarding the community-sourced explanations about why those examples might come across as culturally insensitive for Bangladeshi Hindu and Chakma communities, and evaluated the impact of different persona and community-sourced explanation on LLMs' text generation and their truthfulness and usefulness for users from different demographic backgrounds. Mirroring that flow, in this section, we are going to reflect on how we should regard the sizes and labeling of datasets collected through collaboration with minority communities, why moderating, be it human-run or LLM-based, content related to minority identities and experiences should adopt a restorative justice perspective, and how algorithmic audits should adopt explainability measures besides their focus on biases.

Rethinking Dataset on Minorities as Prototypical Resources

Compared to the vast amount of data traditionally used to train LLMs (Bender et al. 2021), our corpus sourced from religious and ethnic minority communities could be characterized as quite small and could be viewed as a limitation of our study. However, dismissing these community contributions solely because of their size risks reinforcing epistemic erasure, where marginalized voices are systematically excluded from the development and evaluation of AI systems. This exclusion aligns with what Appadurai (Appadurai 2015) describes as ideocide-the systematic annihilation of the ethical and epistemological frameworks of marginalized groups. For example, how the interpretation and labeling of a text about idol worship as "culturally insensitive" vary between Hindu communities and Muslim communities based on their distinct religious values and beliefs. Let's think of moderation in online communities as determining the permissibility of content based on morality and ethics. We need to consider whose ethics (Ahmed 2022a) are being guided by and whose intelligence the AI systems, particularly those used for content moderation, reinforce (Ahmed 2022b). In the context of LLM training, the scarcity of data from minorities is not just a technical issue but also a reflection of broader socio-political inequalities in knowledge production. Recognizing the limited number of example social media posts in our corpus that Bangladeshi religious and Indigenous ethnic minorities find culturally insensitive, along with the corresponding explanations of these views in our corpus, we argue that the size of such a community-sourced corpus should be viewed as a "prototype-based category" (Lakoff 2007). This definition should not depend on straightforward rules about whether a corpus is categorized as big or small based on the number of data instances; instead, it should focus on their prototypical members-similar to how a robin is a better example of a bird than an emu or penguin. Similarly, a corpus that includes examples of culturally insensitive speech according to a wider range of religious minorities, such as Hindus, Buddhists, and Christians, as well as Indigenous ethnic groups like the Chakma, Marma, Garo, and Santhal, would be a more comprehensive communitysourced corpus compared to ours, which focuses solely on the Hindu and Chakma communities. Therefore, datasets and corpora obtained through collaboration with minority communities should be viewed as prototypical examples that can be enhanced rather than dismissed due to their small size.

Content Moderation for Restorative Justice

Scholars in social computing have studied content moderation on online platforms as an exercise of discipline and punishment (Seering, Kaufman, and Chancellor 2022; Das, Østerlund, and Semaan 2021). However, recent works with Bangladeshi minority communities recommend that the design and interaction in online communities should promote restorative justice-an approach to addressing harm that emphasizes healing, accountability, and repairing relationships rather than focusing solely on punishment (Xiao, Jhaver, and Salehi 2023). This approach involves dialogue among those affected-victims, offenders, and the community to foster understanding and find mutually agreed-upon resolutions. It can provide an effective framework for addressing the lack of intercultural knowledge between majority and minority groups and for building trust among them. Rather than relying on stereotypes and overlooking hermeneutical differences, our approach to educating the majority religious and ethnic groups about the perspectives and experiences of minorities can help build trust and lead toward restorative justice. Recognizing diverse epistemologies instead of privileging majority worldviews through LLM-based content moderation, community-sourced corpora, such as those used to refine LLM-based moderation for reflecting the perspectives of Bangladeshi religious minority Hindus and Indigenous ethnic minority Chakmas, can act as a form of restorative intervention, fostering intercultural knowledge-sharing and shared meaning-making. Additionally, different moderation personas (e.g., teacher, mediator) would facilitate conversations within the community and enhance cultural awareness instead of viewing users from different religions and ethnicities through a dichotomy of victims and offenders. By integrating restorative justice with AI ethics, social computing research can conceptualize LLM-based content moderation systems that protect minority groups, repair epistemic harms, and foster online communities that promote trust and reconciliation across cultural and religious veils.

Bias to Explainability in Algorithmic Audits

Scholarships across different fields, including humancomputer interaction, social computing, algorithmic fairness, and natural language processing, have increasingly focused on biases in language technologies (Das et al. 2024; Mökander et al. 2024) and how they manifest in downstream applications (Hartmann, Oueslati, and Staufer 2024; Lam et al. 2022). Many of these studies use algorithmic audits as a methodological approach-empirical investigations that examine public algorithmic systems for potentially problematic behaviors (Bandy 2021). A central criterion these audits focus on is bias, defined as the systematic and unfair discrimination by computing systems against certain individuals or groups in favor of others (Friedman and Nissenbaum 1996), with mitigation often framed as the relevant objective. When algorithmic systems, like LLM, are used in content moderation, it is essential to identify and address biases related to religious and ethnic identities. However, ensuring transparency in decision-making is equally important. Without clear explanations for moderation choices, perceptions of favoritism may arise. For example, Das and colleagues found that given the postcolonial relationship among different religions in the region, when there is not enough clarification, users from Bengali Hindu communities accused Quora's moderation of favoring Bengali Muslims, while users from the latter group believed the platform's decisions were influenced by and privileged the former (Das, Østerlund, and Semaan 2021). This challenge of addressing biases with adequate explanation becomes even more complex when moderating discussions about religious beliefs and cultural rituals. Given this complexity, automated content moderation systems that rely on AI should incorporate principles of explainable AI (Ehsan et al. 2021; Mohseni, Zarei, and Ragan 2021) to improve interpretability. Furthermore, audits should broaden their focus beyond identifying and addressing bias to also include explainability metrics (Hoffman et al. 2018), particularly in the downstream applications of LLMs, like in content moderation.

Conclusion

Our paper develops a corpus of insensitive speech that may not be directly hostile like hate speech but reinforces stereotypes, disregards cultural values or marginalizes the perspectives of religious and ethnic minorities in Bangladesh. Through a tool we developed called "Mod-Guide" that poses different moderation roles and personas, we evaluated whether augmenting GPT-4's text generation by retrieving information from community-sourced explanations can provide significantly different, accurate, and more useful insights for users from diverse backgrounds compared to directly using OpenAI's GPT-4. While our approach offers a promising pathway for fostering pluralistic understanding among religious and ethnic majorities and minorities, challenges remain, including the scalability of incorporating diverse perspectives. Future work should examine reasoning in RAG, explore interdisciplinary collaborations, and expand participatory approaches to improve alignment between LLMs and other marginalized minority communities.

Optional: Author Positionality Statement

Prior research has highlighted how the researchers' identities may reflexively address inevitable tensions and bring affinities into perspective in studying marginalized communities (Schlesinger, Edwards, and Grinter 2017; Liang, Munson, and Kientz 2021). Among all authors (two women and six men), five were born and raised in Bangladesh, while the other three are from India. Except for one author (who is from a North Indian ethnic background), all authors belong to the Bengali ethnolinguistic group. Three authors identify as Bengali Hindus (the lead author from an underprivileged caste in Bangladesh, the rest from a dominant caste in West Bengal, India), and four authors were born in Muslim communities. In addition to their varied sociocultural perspectives, all authors' backgrounds in computer science, with different authors' prior research with marginalized communities, text mining, and data science, have informed and guided the motivation and execution of this study.

References

Agiza, A.; Mostagir, M.; and Reda, S. 2024. Politune: Analyzing the impact of data selection and fine-tuning on economic and political biases in large language models. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, volume 7, 2–12.

Ahmed, S. I. 2022a. Situating ethics: A postsecular perspective for HCI. *Interactions*, 29(4): 84–86.

Ahmed, S. I. 2022b. Whose intelligence? Whose ethics?: Ethical pluralism and decolonizing AI. https://www.youtube.com/watch?v=ReSbgRSJ4WY. Last accessed: Feb 22, 2025.

Allen, C. 2016. Islamophobia. Routledge.

Alo, T. P. 2024. 5–20 August: 1068 minority homes and businesses attacked (translated). https://www.prothomalo. com/bangladesh/6bm2lfn7bz. Last accessed: Feb 21, 2025.

Anam, T. 2013. Pakistan's State of Denial. https://www.nytimes.com/2013/12/27/opinion/anam-

pakistans-overdue-apology.html. Last accessed: July 7, 2023.

Appadurai, A. 2015. Fear of Small Numbers. Writing Religion: The Case for the Critical Study of Religion, 73–95.

Arzaghi, M.; Carichon, F.; and Farnadi, G. 2024. Understanding Intrinsic Socioeconomic Biases in Large Language Models. In *Proceedings of the AAAI/ACM Conference on AI*, *Ethics, and Society*, volume 7, 49–60.

Arzberger, A.; Buijsman, S.; Lupetti, M. L.; Bozzon, A.; and Yang, J. 2024. Nothing Comes Without Its World–Practical Challenges of Aligning LLMs to Situated Human Values through RLHF. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, volume 7, 61–73.

Bandy, J. 2021. Problematic machine behavior: A systematic literature review of algorithm audits. *Proceedings of the acm on human-computer interaction*, 5(CSCW1): 1–34.

Bender, E. M.; Gebru, T.; McMillan-Major, A.; and Shmitchell, S. 2021. On the dangers of stochastic parrots: Can language models be too big? In *Proceedings of the*

2021 ACM conference on fairness, accountability, and transparency, 610–623.

Bowman, R.; Nadal, C.; Morrissey, K.; Thieme, A.; and Doherty, G. 2023. Using thematic analysis in healthcare HCI at CHI: A scoping review. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, 1–18.

Brown, V.; Larasati, R.; Third, A.; and Farrell, T. 2024. A Qualitative Study on Cultural Hegemony and the Impacts of AI. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, volume 7, 226–238.

BSB, B. S. B. 2022. Preliminary Report on Population and Housing Census 2022 : English Version. https: //sid.portal.gov.bd/sites/default/files/files/sid.portal.gov. bd/publications/01ad1ffe_cfef_4811_af97_594b6c64d7c3/ PHC_Preliminary_Report_(English)_August_2022.pdf.

[Accessed: Jan 25, 2025].

Butler, J. 2021. *Excitable speech: A politics of the performative*. routledge.

Chakma, B. 2008. Assessing the 1997 Chittagong hill tracts peace accord. *Asian Profile*, 36(1): 93.

Chakma, B. 2010. The post-colonial state and minorities: ethnocide in the Chittagong Hill Tracts, Bangladesh. *Commonwealth & comparative politics*, 48(3): 281–300.

Cummins, R. A.; and Gullone, E. 2000. Why we should not use 5-point Likert scales: The case for subjective quality of life measurement. In *Proceedings, second international conference on quality of life in cities*, volume 74, 74–93.

Das, D.; Guha, S.; Brubaker, J. R.; and Semaan, B. 2024. The"Colonial Impulse" of Natural Language Processing: An Audit of Bengali Sentiment Analysis Tools and Their Identity-based Biases. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems*, 1–18.

Das, D.; Østerlund, C.; and Semaan, B. 2021. "Jol" or" Pani"?: How Does Governance Shape a Platform's Identity? *Proceedings of the ACM on Human-Computer Interaction*, 5(CSCW2): 1–25.

Du Bois, W. E. B. 2015. Souls of black folk. Routledge.

Ehsan, U.; Liao, Q. V.; Muller, M.; Riedl, M. O.; and Weisz, J. D. 2021. Expanding explainability: Towards social transparency in ai systems. In *Proceedings of the 2021 CHI conference on human factors in computing systems*, 1–19.

Erete, S.; Israni, A.; and Dillahunt, T. 2018. An intersectional approach to designing in the margins. *Interactions*, 25(3): 66–69.

Etikan, I.; Musa, S. A.; Alkassim, R. S.; et al. 2016. Comparison of convenience sampling and purposive sampling. *American journal of theoretical and applied statistics*, 5(1): 1–4.

France-Presse, A. 2015. American atheist blogger hacked to death in Bangladesh — theguardian.com. https://www.theguardian.com/world/2015/feb/27/americanatheist-blogger-hacked-to-death-in-bangladesh. Last accessed July 7, 2023.

Fricker, M. 2007. *Epistemic injustice: Power and the ethics of knowing*. Oxford University Press.

Friedman, B.; and Nissenbaum, H. 1996. Bias in computer systems. *ACM Transactions on information systems (TOIS)*, 14(3): 330–347.

Ganguly, S. 2021. Bangladesh's Deadly Identity Crisis. https://foreignpolicy.com/2021/10/29/bangladesh-communal-violence-hindu-muslim-identity-crisis/. Last accessed: July 7, 2023.

Ghosh, S. 2024. Interpretations, Representations, and Stereotypes of Caste within Text-to-Image Generators. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, volume 7, 490–502.

Ghosh, S.; and Caliskan, A. 2023a. Chatgpt perpetuates gender bias in machine translation and ignores non-gendered pronouns: Findings across bengali and five other lowresource languages. In *Proceedings of the 2023 AAAI/ACM Conference on AI, Ethics, and Society*, 901–912.

Ghosh, S.; and Caliskan, A. 2023b. 'Person'== Lightskinned, Western Man, and Sexualization of Women of Color: Stereotypes in Stable Diffusion. *arXiv preprint arXiv:2310.19981*.

Ghosh, S.; Venkit, P. N.; Gautam, S.; Wilson, S.; and Caliskan, A. 2024. Do Generative AI Models Output Harm while Representing Non-Western Cultures: Evidence from A Community-Centered Approach. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, volume 7, 476–489.

Goffman, E. 2009. *Stigma: Notes on the Management of Spoiled Identity*. Touchstone.

Hamidieh, K.; Zhang, H.; Gerych, W.; Hartvigsen, T.; and Ghassemi, M. 2024. Identifying implicit social biases in vision-language models. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, volume 7, 547–561.

Hartmann, D.; Oueslati, A.; and Staufer, D. 2024. Watching the Watchers: A Comparative Fairness Audit of Cloud-based Content Moderation Services. *arXiv preprint arXiv:2406.14154*.

Hasan, M. 2021. Minorities under attack in Bangladesh. https://www.lowyinstitute.org/the-interpreter/minoritiesunder-attack-bangladesh. Last accessed: July 7, 2023.

Heywood, E.; Ivey, B.; and Meuter, S. 2024. Reaching hardto-reach communities: using WhatsApp to give conflictaffected audiences a voice. *International Journal of Social Research Methodology*, 27(1): 107–121.

Hill, G.; and Chakma, K. 2022. Muscular nationalism, masculinist militarism: the creation of situational motivators and opportunities for violence against the Indigenous peoples of the Chittagong Hill Tracts, Bangladesh. *International Feminist Journal of Politics*, 24(4): 519–543.

Hoffman, R. R.; Mueller, S. T.; Klein, G.; and Litman, J. 2018. Metrics for explainable AI: Challenges and prospects. *arXiv preprint arXiv:1812.04608*.

Inan, S.; Çetin, H.; and Yakupoğlu, N. 2024. Spring water anomalies before two consecutive earthquakes (M w 7.7 and M w 7.6) in Kahramanmaraş (Türkiye) on 6 February 2023. *Natural Hazards and Earth System Sciences*, 24(2): 397– 409. International, A. 2021. Bangladesh: Protection of Hindus and others must be ensured amid ongoing violence. https: //www.amnesty.org/en/latest/news/2021/10/bangladesh-protection-of-hindus-and-others-must-be-ensured-amid-ongoing-violence/. Last accessed: July 7, 2023.

International, M. R. G. 2018. Christians. https:// minorityrights.org/minorities/christians-6/. Last accessed: July 7, 2023.

Ittefaq, T. D. 2014. Attacks on minorities continue. https://web.archive.org/web/20140110191737/http://www. clickittefaq.com/more-stories/attacks-minorities-continue/. Last accessed: July 7, 2023.

Izacard, G.; and Grave, E. 2020. Leveraging passage retrieval with generative models for open domain question answering. *arXiv preprint arXiv:2007.01282*.

Jhaver, S.; Birman, I.; Gilbert, E.; and Bruckman, A. 2019. Human-machine collaboration for content regulation: The case of reddit automoderator. *ACM Transactions on Computer-Human Interaction (TOCHI)*, 26(5): 1–35.

Jhaver, S.; Chen, Q. Z.; Knauss, D.; and Zhang, A. X. 2022. Designing word filter tools for creator-led comment moderation. In *Proceedings of the 2022 CHI conference on human factors in computing systems*, 1–21.

Jiang, J. A.; Nie, P.; Brubaker, J. R.; and Fiesler, C. 2023. A trade-off-centered framework of content moderation. *ACM Transactions on Computer-Human Interaction*, 30(1): 1–34.

Jiang, J. A.; Scheuerman, M. K.; Fiesler, C.; and Brubaker, J. R. 2021. Understanding international perceptions of the severity of harmful content online. *PloS one*, 16(8): e0256762.

Koka, H.; Langat, S.; Mulwa, F.; Mutisya, J.; Owaka, S.; Sifuna, M.; Ongus, J. R.; Lutomiah, J.; and Sang, R. 2024. Combining Morphological and Molecular Tools Can Enhance Tick Species Identification for Improved Tick-Borne Disease Surveillance Among Pastoral Communities in Kenya. *Vector-Borne and Zoonotic Diseases*.

Kolla, M.; Salunkhe, S.; Chandrasekharan, E.; and Saha, K. 2024. Llm-mod: Can large language models assist content moderation? In *Extended Abstracts of the CHI Conference on Human Factors in Computing Systems*, 1–8.

Kumar, S.; Kholkar, G.; Mendke, S.; Sadana, A.; Agrawal, P.; and Dandapat, S. 2024. Socio-Culturally Aware Evaluation Framework for LLM-Based Content Moderation. *arXiv* preprint arXiv:2412.13578.

Kwok, L.; Bravansky, M.; and Griffin, L. D. 2024. Evaluating cultural adaptability of a large language model via simulation of synthetic personas. *arXiv preprint arXiv:2408.06929*.

Lakoff, G. 2007. Cognitive models and prototype theory. *The cognitive linguistics reader*, 130–167.

Lam, M. S.; Gordon, M. L.; Metaxa, D.; Hancock, J. T.; Landay, J. A.; and Bernstein, M. S. 2022. End-user audits: A system empowering communities to lead large-scale investigations of harmful algorithmic behavior. *Proceedings of the ACM on Human-Computer Interaction*, 6(CSCW2): 1–34. Lee, E. 2019. America for Americans: A history of xenophobia in the United States. Basic Books.

Leitner, M.; Dorn, R.; Morstatter, F.; and Lerman, K. 2025. Characterizing Network Structure of Anti-Trans Actors on TikTok. *arXiv preprint arXiv:2501.16507*.

Lewis, P.; Perez, E.; Piktus, A.; Petroni, F.; Karpukhin, V.; Goyal, N.; Küttler, H.; Lewis, M.; Yih, W.-t.; Rocktäschel, T.; et al. 2020. Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in neural information processing systems*, 33: 9459–9474.

Li, T.; Zhang, X.; Du, C.; Pang, T.; Liu, Q.; Guo, Q.; Shen, C.; and Liu, Y. 2024. Your large language model is secretly a fairness proponent and you should prompt it like one. *arXiv* preprint arXiv:2402.12150.

Liang, C. A.; Munson, S. A.; and Kientz, J. A. 2021. Embracing four tensions in human-computer interaction research with marginalized people. *ACM Transactions on Computer-Human Interaction (TOCHI)*, 28(2): 1–47.

MacLeod, H.; Bastin, G.; Liu, L. S.; Siek, K.; and Connelly, K. 2017. "Be Grateful You Don't Have a Real Disease" Understanding Rare Disease Relationships. In *Proceedings* of the 2017 CHI Conference on Human Factors in Computing Systems, 1660–1673.

MacLeod, H.; Jelen, B.; Prabhakar, A.; Oehlberg, L.; Siek, K. A.; and Connelly, K. 2016. Asynchronous remote communities (ARC) for researching distributed populations. In *PervasiveHealth*, 1–8.

Maestre, J. F.; MacLeod, H.; Connelly, C. L.; Dunbar, J. C.; Beck, J.; Siek, K. A.; and Shih, P. C. 2018. Defining through expansion: conducting asynchronous remote communities (arc) research with stigmatized groups. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*, 1–13.

McDonald, N.; Schoenebeck, S.; and Forte, A. 2019. Reliability and inter-rater reliability in qualitative research: Norms and guidelines for CSCW and HCI practice. *Proceedings of the ACM on human-computer interaction*, 3(CSCW): 1–23.

Min, S.; Krishna, K.; Lyu, X.; Lewis, M.; Yih, W.-t.; Koh, P. W.; Iyyer, M.; Zettlemoyer, L.; and Hajishirzi, H. 2023. Factscore: Fine-grained atomic evaluation of factual precision in long form text generation. *arXiv preprint arXiv:2305.14251*.

Mizan, M.; and Rahaman, A. 2025. Removal of word 'adivasi': Indigenous group attacked at NCTB; 20 hurt — thedailystar.net. https://www.thedailystar.net/news/bangladesh/news/removal-word-adivasi-indigenous-group-attacked-nctb-20-hurt-3799851. Last accessed 21-02-2025].

Mohseni, S.; Zarei, N.; and Ragan, E. D. 2021. A multidisciplinary survey and framework for design and evaluation of explainable AI systems. *ACM Transactions on Interactive Intelligent Systems (TiiS)*, 11(3-4): 1–45.

Mökander, J.; Schuett, J.; Kirk, H. R.; and Floridi, L. 2024. Auditing large language models: a three-layered approach. *AI and Ethics*, 4(4): 1085–1115. Molina, M. D.; and Sundar, S. S. 2022. When AI moderates online content: effects of human collaboration and interactive transparency on user trust. *Journal of Computer-Mediated Communication*, 27(4): zmac010.

Mozafari, M.; Farahbakhsh, R.; and Crespi, N. 2020. Hate speech detection and racial bias mitigation in social media based on BERT model. *PloS one*, 15(8): e0237861.

Mukherjee, A.; Coomar, P.; Sarkar, S.; Johannesson, K. H.; Fryar, A. E.; Schreiber, M. E.; Ahmed, K. M.; Alam, M. A.; Bhattacharya, P.; Bundschuh, J.; et al. 2024. Arsenic and other geogenic contaminants in global groundwater. *Nature Reviews Earth & Environment*, 5(4): 312–328.

Orlandi, R. R.; Kingdom, T. T.; Smith, T. L.; Bleier, B.; De-Conde, A.; Luong, A. U.; Poetker, D. M.; Soler, Z.; Welch, K. C.; Wise, S. K.; et al. 2021. International consensus statement on allergy and rhinology: rhinosinusitis 2021. In *International forum of allergy & rhinology*, volume 11, 213–739. Wiley Online Library.

Plaza-del Arco, F. M.; Nozza, D.; Hovy, D.; et al. 2023. Respectful or toxic? using zero-shot learning with language models to detect hate speech. In *The 7th workshop on online abuse and harms (woah)*. Association for Computational Linguistics.

Prabhakar, A. S.; Guerra-Reyes, L.; Kleinschmidt, V. M.; Jelen, B.; MacLeod, H.; Connelly, K.; and Siek, K. A. 2017. Investigating the suitability of the asynchronous, remote, community-based method for pregnant and new mothers. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*, 4924–4934.

Rifat, M. R.; Das, D.; Poddar, A.; Jannat, M.; Soden, R.; Semaan, B.; and Ahmed, S. I. 2024a. The Politics of Fear and the Experience of Bangladeshi Religious Minority Communities Using Social Media Platforms. *Proceedings of the ACM on Human-Computer Interaction*, 8(CSCW2): 1–32.

Rifat, M. R.; Safir, A. H.; Saha, S.; Junaed, J. A.; Saleki, M.; Amin, M. R.; and Ahmed, S. I. 2024b. Data, Annotation, and Meaning-Making: The Politics of Categorization in Annotating a Dataset of Faith-based Communal Violence. In *Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency*, 2148–2156.

Roy, S.; Singh, A. K.; et al. 2023. Sociological perspectives of social media, rumors, and attacks on minorities: Evidence from Bangladesh. *Frontiers in Sociology*, 8: 1067726.

Sarkar, T.; and Bandyopadhyay, S. 2017. *Calcutta: The stormy decades*. Taylor & Francis.

Scheuerman, M. K.; Jiang, J. A.; Fiesler, C.; and Brubaker, J. R. 2021. A framework of severity for harmful content online. *Proceedings of the ACM on Human-Computer Interaction*, 5(CSCW2): 1–33.

Schlesinger, A.; Edwards, W. K.; and Grinter, R. E. 2017. Intersectional HCI: Engaging identity through gender, race, and class. In *Proceedings of the 2017 CHI conference on human factors in computing systems*, 5412–5427.

Seering, J.; Kaufman, G.; and Chancellor, S. 2022. Metaphors in moderation. *New Media & Society*, 24(3): 621–640.

Sen, D. 2018. *The decline of the caste question: Jogendranath Mandal and the defeat of Dalit politics in Bengal.* Cambridge University Press.

Shackle, S. 2018. Atheist bloggers in Bangladesh are still under threat — New Humanist. https: //newhumanist.org.uk/articles/5386/atheist-bloggers-inbangladesh-are-still-under-threat. Last accessed July 7,

2023. Shi, W.; Min, S.; Yasunaga, M.; Seo, M.; James, R.; Lewis,

M.; Zettlemoyer, L.; and Yih, W.-t. 2023. Replug: Retrievalaugmented black-box language models. *arXiv preprint arXiv:2301.12652.*

Singh, D. K.; Das, D.; and Semaan, B. 2025. The Power of Language: Resisting Western Heteropatriarchal Normative Writing Standards. In *Proceedings of the CHI Conference on Human Factors in Computing Systems (CHI '25)*. New York, NY, USA: Association for Computing Machinery.

Song, S.; Ashton, M.; Yoo, R. H.; Lkhagvajav, Z.; Wright, R.; Mathews, D. J.; and Taylor, C. O. 2025. Participant Contributions to Person-Generated Health Data Research Using Mobile Devices: Scoping Review. *Journal of medical Internet research*, 27: e51955.

Sugirtharajah, S. 2004. *Imagining Hinduism: A postcolonial perspective*. Routledge.

Sultana, A.; Das, D.; Alam, S. B.; Shidujaman, M.; and Ahmed, S. I. 2024. A Civics-oriented Approach to Understanding Intersectionally Marginalized Users' Experience with Hate Speech Online. *arXiv preprint arXiv:2410.14950*.

Sultana, S.; Akter, R.; Sultana, Z.; and Ahmed, S. I. 2022. Toleration Factors: The Expectations of Decorum, Civility, and Certainty on Rural Social Media. In *Proceedings of the* 2022 International Conference on Information and Communication Technologies and Development, 1–14.

Sun, H.; and Ni, W. 2022. Design and Application of an AI-Based Text Content Moderation System. *Scientific Programming*, 2022(1): 2576535.

Thorne, J. 2022. Data-efficient autoregressive document retrieval for fact verification. *arXiv preprint arXiv:2211.09388.*

Tsirmpas, D.; Androutsopoulos, I.; and Pavlopoulos, J. 2025. Scalable Evaluation of Online Moderation Strategies via Synthetic Simulations. *arXiv preprint arXiv:2503.16505*.

Us, T. W. I. n.d. Indigenous Peoples of Bangladesh — The World in Us — theworldinus.org. https://www.theworldinus.org/blog/indigenous-peoples-of-bangladesh. [Accessed 21-02-2025].

Vaidya, S.; Cai, J.; Basu, S.; Naderi, A.; Wohn, D. Y.; and Dasgupta, A. 2021. Conceptualizing visual analytic interventions for content moderation. In *2021 IEEE Visualization Conference (VIS)*, 191–195. IEEE.

Walker, A. M.; and DeVito, M. A. 2020. "More gay'fits in better": Intracommunity Power Dynamics and Harms in Online LGBTQ+ Spaces. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, 1–15.

Wiegand, M.; Ruppenhofer, J.; and Eder, E. 2021. Implicitly abusive language–what does it actually look like and why are we not getting there? In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 576–587.

Wobbrock, J. O. 2012. Seven research contributions in HCI. *Intelligence*, 174(12-13): 910–950.

Xiao, S.; Jhaver, S.; and Salehi, N. 2023. Addressing interpersonal harm in online gaming communities: The opportunities and challenges for a restorative justice approach. *ACM Transactions on Computer-Human Interaction*, 30(6): 1–36.

Zeng, W.; Liu, Y.; Mullins, R.; Peran, L.; Fernandez, J.; Harkous, H.; Narasimhan, K.; Proud, D.; Kumar, P.; Radharapu, B.; et al. 2024. Shieldgemma: Generative ai content moderation based on gemma. *arXiv preprint arXiv:2407.21772*.

Zhang, T.; Kishore, V.; Wu, F.; Weinberger, K. Q.; and Artzi, Y. 2019. Bertscore: Evaluating text generation with bert. *arXiv preprint arXiv:1904.09675*.