

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

Dipto Das
Department of Computer Science
University of Toronto
Toronto, Ontario, Canada
dipto.das@utoronto.ca

Achhiya Sultana
Independent University Bangladesh
Dhaka, Bangladesh
achhiyaset@iub.edu.bd

Ankit Singh Chauhan
Indiana University Indianapolis
Indianapolis, Indiana, United States
ankichau@iu.edu

Saadia Binte Alam
Independent University Bangladesh
Dhaka, Bangladesh
saadiabinte@iub.edu.bd

Mohammad Shidujaman
Independent University Bangladesh
Dhaka, Bangladesh
shidujaman@iub.edu.bd

Shion Guha
Faculty of Information
University of Toronto
Toronto, Ontario, Canada
shion.guha@utoronto.ca

Sunandan Chakraborty
Indiana University Indianapolis
Indianapolis, Indiana, United States
sunchak@iu.edu

Syed Ishtiaque Ahmed
Department of Computer Science
University of Toronto
Toronto, Ontario, Canada
ishtiaque@cs.toronto.edu

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—language that disregards or marginalizes the cultural and religious perspectives of historically underrepresented communities, often through implicit erasure, misrepresentation, or normative framing, rather than overt hostility. 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.



This work is licensed under a Creative Commons Attribution 4.0 International License. *COMPASS '26, Virtual Event*
© 2025 Copyright held by the owner/author(s).
ACM ISBN 978-x-xxxx-xxxx-x/YYYY/MM
<https://doi.org/10.1145/nnnnnnn.nnnnnnn>

CCS Concepts

• **Social and professional topics** → **Race and ethnicity; Religious orientation; Cultural characteristics**; • **Human-centered computing** → *Interactive systems and tools; Empirical studies in collaborative and social computing*; • **Applied computing** → Document management and text processing.

Keywords

Minority, LLM, RAG, content moderation, ethics

ACM Reference Format:

Dipto Das, Achhiya Sultana, Ankit Singh Chauhan, Saadia Binte Alam, Mohammad Shidujaman, Shion Guha, Sunandan Chakraborty, and Syed Ishtiaque Ahmed. 2025. *Mod-Guide: An LLM-based Content Moderation Feedback System to Address Insensitive Speech toward Indigenous Ethnic and Religious Minority Communities*. In *ACM SIGCAS/SIGCHI Conference on Computing and Sustainable Societies (COMPASS) (COMPASS '26)*, July 27–31, 2026, Virtual Event. ACM, New York, NY, USA, 14 pages. <https://doi.org/10.1145/nnnnnnn.nnnnnnn>

1 Introduction

Language is more than a means of communication and is a form of power [87]. 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. [25]. Linguistic marginalization and injuries manifest in online communities through hate speech, bullying, political incitement, and other forms of insensitive speech. In

the context of this study, we define insensitive speech as the linguistic acts that—while not overtly hateful or profane—disregards, misrepresents, trivializes, or marginalizes the cultural, religious, or epistemic values of historically underrepresented communities. Unlike hate speech, which is often explicit in its hostility or incitement, insensitive speech manifests through dismissive framings, culturally uninformed generalizations, or normative assumptions rooted in majority worldviews. Its harm arises not only from intent or content, but from its failure to recognize and respect the situated meanings, lived experiences, and interpretive frameworks of minority groups.

Most platforms respond to harmful content by enforcing moderation policies through a combination of human moderators and algorithmic systems [49, 71]. Recent advances in large language models (LLMs) have enabled more scalable moderation [52, 101], but these models are predominantly shaped by and reinforce majority perspectives [60]. 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 would likely reinforce the societal barrier between the majority and minority groups in the case of understanding each other’s perspectives.

We focus on Bangladesh, where the Hindu and Chakma communities are the largest religious and Indigenous ethnic minorities, respectively [14]. Motivated by concepts of hermeneutical injustice [28] and the divide between majority and minority consciousness [22], 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 [59], the significance of our work lies in grounding RAG in epistemically marginalized perspectives and evaluating its implications for 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 that are widely recognized in HCI scholarship [99]:

- **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 discourse in human-computer interaction, AI ethics, and social computing 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, the construction of our dataset, the design of the LLM-RAG pipeline, and the empirical evaluation. We conclude by reflecting on the challenges of scale, the normativity of dataset curation, and the implications for design toward community-centered justice and fair content moderation systems.

2 Literature Review

This section situates our work at the intersection of linguistic marginalization, epistemic injustice, and automated content moderation. First, we discuss linguistic marginalization and conceptualize *insensitive speech* as a form of harm shaped by cultural and historical context. Next, we examine epistemic barriers between majority and minority communities through the lenses of Du Bois’ notion of the *veil* and Fricker’s *hermeneutical injustice*. Finally, we review research on automated content moderation and large language models, highlighting limitations in addressing culturally contextual harms and motivating our community-grounded RAG-based approach.

2.1 Linguistic Marginalization as Insensitive Speech

Language plays a crucial role in shaping social hierarchies and power dynamics. It establishes normative and non-normative identities [15]. 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 [15]. 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 conventions—can 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 [77]. The large-scale communal riots and the disproportionate targeting of Hindus during the Liberation War illustrate this pattern [6, 80]. In recent decades, assaults on Hindu communities have increased, often fueled by social media rumors of religious insults against the majority [30, 43, 79], such as the violence during the 2021 Durga Puja [38]. Furthermore, political instability worsens the persecution, leading to targeted attacks on Hindus [5, 45], Christians [44], and atheists [27, 85]. 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 [17, 40]. Despite a peace accord in 1997, they continue to struggle for autonomy and basic recognition of indigeneity to this day [16, 68].

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, [77] 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 [95]. 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 [90]. 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.

2.2 Epistemic Barriers among Majority and Minority

Marginalization of minorities often stems from entrenched tribal stigma surrounding attributes like ethnicity, religion, language, and cultural practices [35]. For example, in many contexts, misunderstandings of minority religions' practices and beliefs lead to unsubstantiated fear (e.g., Islamophobia [4]), misrepresentation (e.g., depicting non-Abrahamic faiths as satanic or pagan [89]), 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 [57]. 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 [35], 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 [77].

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" [22] and Miranda Fricker's notion of hermeneutical injustice [28]. 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

self-perception and hinders mutual comprehension across racial divides [22]. Recent work [77] 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 [28], 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 [77, 90, 91]. 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 [22], while dominant social norms and unconscious biases perpetuate injustice against marginalized communities [28]. 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.

2.3 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 [50, 81]. While platforms' "institutional ethics" [81] 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 [50], 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) [21, 49].

Particularly focused on moderation philosophy, Seering et al. [83] 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 [49]. Language technologies have become central to automated content moderation systems [92, 96]. In terms of complexity and sophistication, these systems range from simple keyword filters [47, 48], to task-specific models for sentiment analysis and hate speech detection [20, 72], to foundational large language models (LLMs) deployed at scale [42, 52, 101]. While multilingual LLMs have shown promising results in detecting explicit hate speech, fake news, and discriminatory language [51, 74, 75], 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 [13, 34]. Prior research has shown that these models exhibit demographic (e.g., race, gender, nationality, religion, caste) [20, 31–33, 36], socioeconomic [8], and political biases [1], 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 [9, 42], wherein they have examined how data selection and fine-tuning impacted LLMs' economic and political biases [1], how model responses vary with culturally sensitive prompts [73], and found that persona-based prompting can improve alignment with specific moderation goals [54]. Studies highlighted how crowd-sourced data annotation is subject to limited annotator expertise [53], dismissal of religious faiths [78], minorities' underrepresentation [88, 93], and disproportionate association of toxicity with minorities [98]. Retrieval augmented generation (RAG)—a method to enhance language model outputs by retrieving relevant external documents while generating responses [59], can be an effective technique to address the concerns of LLM biases affecting content moderation [58, 94]. 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 [54, 83] and RAG enhances factual accuracy and contextual grounding [46, 86]. 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.

3 Methods

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 [21, 78, 90]. Here, we build on our findings and community relationships fostered during the earlier phases of our research. Our study proceeded in three stages (see Figure 1): (1) corpus preparation, (2) development of the Mod-Guide moderation system, and (3) evaluation of moderation feedback.

3.1 Overview

First, we collaborated with 22 members from two minority communities in Bangladesh—Hindu and Chakma—using the asynchronous research community method to construct a corpus of culturally insensitive speech containing 132 instances and accompanying explanations grounded in community perspectives. Second, we integrated this corpus into a moderation pipeline that combines persona-based prompting GPT-4 with retrieval-augmented generation (RAG) to ground moderation feedback in community-authored explanations. Third, we evaluated the system through a mixed-method approach consisting of (a) quantitative analysis of generated responses using text embeddings, (b) assessment of factual accuracy of generated texts by 2 experts, and (c) a user study examining the perceived usefulness of moderation feedback with 15 participants from majority and minority communities. The following sections offer further details about each of these stages.

3.2 Author Positionality

Prior research has highlighted how the researchers' identities may reflexively address inevitable tensions and bring affinities into perspective in studying marginalized communities [61, 82]. Therefore, we consider it essential to situate this work on marginalized minority communities in the Global South in relation to the researchers' positionality. Among all authors (2 women and 6 men), five were born and raised in Bangladesh, while the other three were 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.

4 Corpus Preparation to Understand Minority Hermeneutics

We collected the corpus through the Asynchronous Remote Community (ARC) method [63]. Prior research has used this method to engage with participants when in-person communication is difficult to arrange due to population distribution [62], stigma [64], or fear of isolation [97]. Over a month, we nudged discussions from those

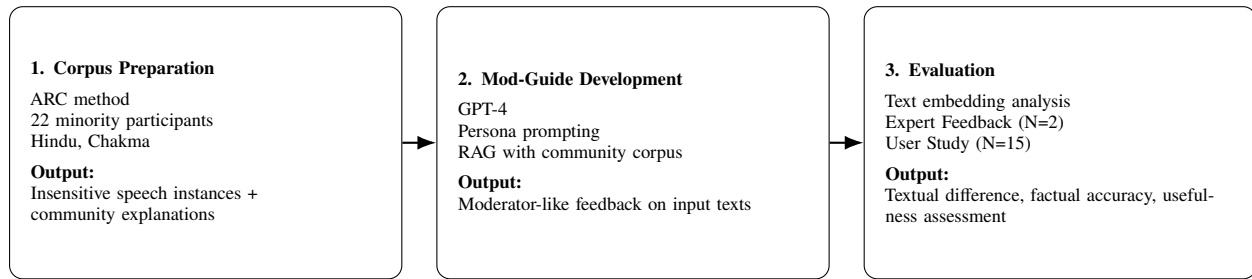


Figure 1: Methods overview.

groups weekly to sustain engagement while allowing flexibility for participants to share instances of social media posts they found culturally insensitive, about which they have found that the religious 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.

4.1 ARC 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 [77, 90]. 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. While our religious minority participants came from various parts of the country, most of our ethnic minority participants were from the Chittagong Hill Tracts (CHT) region, where most Indigenous ethnic minority communities live. Most of our Hindu participants were from underprivileged scheduled castes (*tafsili jati*) [84]. In addition to reflecting the general demographic pattern of Hindu communities in Bangladesh, the higher representation of participants from underprivileged castes also resists the Brahminical and casteist interpretations of Hindu beliefs and practices in our corpus.

4.2 Procedure

Similar to previous ARC studies [63, 64, 97], 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 [39, 62]. All participants had existing Facebook and WhatsApp accounts that they used to participate in the study, thus maintaining platform-related 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 [76, 97]. 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 non-problematic 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 insensitive 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 non-textual content (e.g., HTML tags, URLs). While allowing the participants to share screenshots, URLs, and videos made it easier for them to share the examples of insensitive speech they encountered, using OCR, transcription, and cleaning non-natural language components enabled textual standardization, allowing us to convert image-based content into analyzable text for inclusion in both the RAG corpus and the moderation evaluation pipeline and allowed us to preserve contextually rich, vernacular examples that participants considered important.

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 (*later referred as Insensitive Speech Example-1*) that Hindu participants in our ARCs found to be culturally insensitive.

কিছু মানবতার ফেলিওয়ালাদের দেখতেছি, মূর্তি পাহারা 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 well-circulating 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, an ARC participant shared the following explanations for why the above text was insensitive based on different schools of thought within Hinduism:

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.

5 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, which is often based on stereotypes and deepens the cultural divide between the majority and minority communities in the country. Drawing on Du Bois [22], we refer to that as the veil. We examined OpenAI’s GPT-4 in particular. Additionally, we explore retrieval-augmented generation (RAG) [59] based on community insights, with content moderation in mind. We chose RAG over other approaches, such as few-shot prompting or fine-tuning, to ensure interpretability, adaptability, and alignment with community perspectives. RAG allows generated texts to be directly grounded in retrievable, community-authored explanations, preserving traceability and cultural nuance [18]. Unlike fine-tuning, which embeds knowledge irreversibly into model weights, RAG supports modular updates as community insights evolve. This approach preserves traceability, allows the corpus to evolve as communities contribute additional insights, and supports modular updates as new examples are collected. Although RAG introduces computational overhead compared to simple prompting, it offers an interpretable mechanism for integrating minority hermeneutics into moderation feedback, which aligns with the exploratory and design-oriented objectives of this work. 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.

5.1 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. [83], 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. Compared to other metaphoric conceptualizations of moderation [65, 102], this offers a computationally tractable and interactionally diverse set of metaphors that can be readily translated into promptable roles for language models. Hence, based on Seering et al.’s recommendations [83], 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.

5.2 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 2). 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 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 2). 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

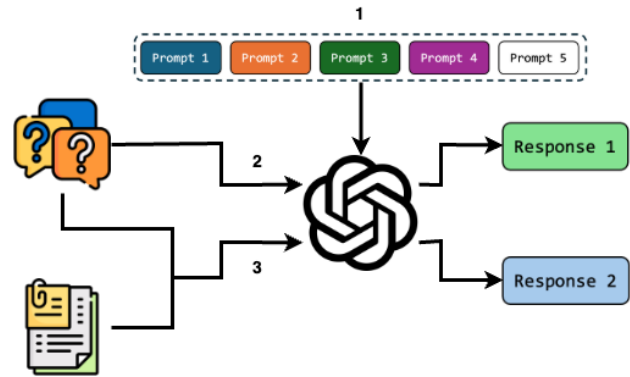


Figure 2: Prompt, LLM, and RAG pipeline.

the LLM (see path 2 in Figure 2). 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 2), 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.

5.3 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 *Insensitive Speech Example-1* 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 3 that translates as:

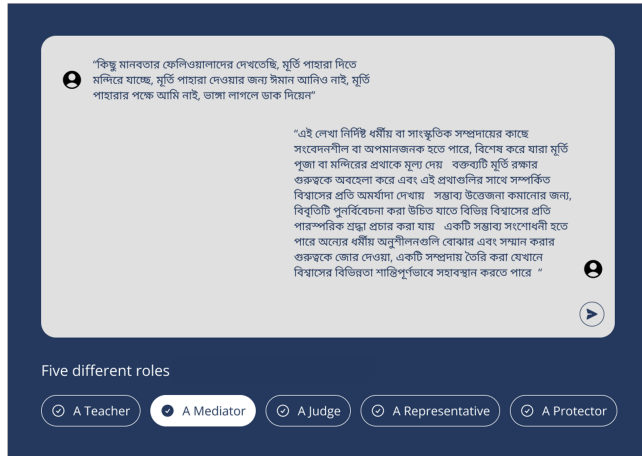


Figure 3: Feedback from Mod-Guide's in Mediator role.

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.

6 Evaluation of Moderation Feedback

We adopted a mixed-method evaluation approach in our study, where we considered content moderation persona, whether the 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 LLM-based 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?

6.1 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 [103]. 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 ten 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 off-the-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 community-sourced corpus affecting the generated texts.

6.2 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 [67]. 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.

6.2.1 Expert Participants. We recruited two expert participants, one from each minority community, through convenience sampling [26]. The expert (E1) from the religious minority Hindu community was 35 years old man. He was from an underprivileged Hindu caste. He obtained a (*kabyotirtho*) certification from the Bangladesh Sanskrit and Pali Education Board, demonstrating his extensive knowledge of Hindu beliefs and scriptures. In addition, he was knowledgeable about local Hindu practices and experiences through his role as an administrator and moderator, and his involvement in various social welfare initiatives aimed at religious minorities. His background positions him as an expert who could evaluate Mod-guide's outputs without reinforcing casteist perspectives. The expert (E2) from the ethnic minority community was a 32-year-old man. He has worked on issues affecting Indigenous ethnic minority communities. Besides collaborating with community members in the CHT region and the activist groups in the national capital, he has also served as a young representative on Indigenous rights at international venues. These participants did not take part in the earlier corpus generation phase but were well familiar with their respective 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 [12, 66]. 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.

6.2.2 Expert Feedback. The expert participant from the religious minority community (E1) 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 community-sourced data. For example, for the statement "*Hindus should not worship idols*", 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: ("*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: "*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) 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: *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 4) 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 [91]. Participant E2 also focused on the text's use of vulgar language ("cdi" is a Romanized Bengali internet slang that means "fuck") 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.

6.3 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.



Figure 4: A screenshot shared by an ARC participant.

6.3.1 User Study. 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 [19], 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 with $\alpha = 0.05$.

6.3.2 Usefulness of Persona and RAG-based Feedback. 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.

7 Limitations and Future Work

While this paper makes conceptual, technical, and methodological contributions to the design of culturally sensitive moderation systems, it has several limitations that warrant acknowledgment. Since this paper is one of the initial outcomes of a larger project focused on

minority communities’ experiences with computing technologies in the Global South, we also outline later in this section how we plan to address those shortcomings in our future work.

First, the dataset used in Mod-Guide, while rich in contextual and narrative depth, is relatively small. Such a limited size may constrain the diversity of insensitive speech patterns captured and reduce the recall capacity of semantic retrieval in the RAG pipeline. It may also limit generalizability to other minority communities in Bangladesh or to other communities across the Global South. Second, the effectiveness of RAG depends on the semantic quality of retrieved documents. While we used multilingual embeddings to enable retrieval in Bengali, concerns about uneven embedding quality remain, especially given the low-resource status of Bengali in NLP. Moreover, our RAG-based pipeline’s computational overhead for tasks like vector stores, chunking, and retrieval tools may not be readily available in resource-constrained settings. Third, the factual accuracy assessment in our paper involved only two expert participants, which, while insightful, may introduce subjectivity and reduce the evaluation’s robustness. Similarly, the usefulness study involved a small participant pool, and demographic coverage was uneven across ethnic and religious groups. Our future work to improve the tool will expand these evaluations by including more diverse participants, and employing both qualitative and quantitative measures (e.g., inter-rater reliability, Likert ratings) to triangulate user perceptions. Fourth, we also acknowledge our concerns about using OCR to extract Bengali text from screenshots submitted by participants. While it was necessary to incorporate real-world content that often circulates as images, this process may introduce errors or mistranscriptions (e.g., OCR limitations on low-resolution images).

Finally, while our focus on two specific minority communities in Bangladesh—Hindus and Chakmas enables rich, context-aware analysis, it limits the applicability of findings to other religious, ethnic, or linguistic groups. Additionally, even within the focal communities, there exists internal diversity (e.g., caste, gender, regional dialects) that our sample may not fully capture. Thus, while our system demonstrates promise, its outputs should be interpreted as community-situated rather than universally representative. We also recognize that moderation decisions, even when community-informed, may reproduce power asymmetries or unintentionally essentialize minority identities. Interpretations of what constitutes “insensitive” speech are context-dependent and contested. Thus, while Mod-Guide foregrounds community narratives, it must remain adaptable to revision, contestation, and critique through ongoing participatory design. Our future work will expand the community-sourced corpus to include additional minority groups in Bangladesh, including Buddhist and Christian communities, as well as other Indigenous ethnic minority groups, such as the Marma and Santal peoples. Further engagement within the Hindu and Chakma populations could also examine intra-community variations, as identified above, to avoid essentializing minority perspectives. Additionally, future studies should investigate how corpus size and composition affect the quality and contextual accuracy of RAG-generated feedback in faith-based and culturally sensitive cases.

8 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.

8.1 Rethinking Dataset on Minorities as Prototypical Resources

Compared to the vast amount of data traditionally used to train LLMs [11], 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 [7] 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 [2] are being guided by and whose intelligence the AI systems, particularly those used for content moderation, reinforce [3]. 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” [55]. 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 community-sourced corpus compared to ours, which focuses solely on the Hindu and Chakma communities. Therefore, while we

recognize the need for future work to expand and diversify these corpora through sustained community partnerships, we emphasize that 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.

8.2 Content Moderation for Restorative Justice

Scholars in social computing have studied content moderation on online platforms as an exercise of discipline and punishment [21, 83]. 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 [100]. 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.

8.3 Bias to Explainability in Algorithmic Audits

Scholarships across different fields, including human-computer interaction, social computing, algorithmic fairness, and natural language processing, have increasingly focused on biases in language technologies [20, 70] and how they manifest in downstream applications [37, 56]. Many of these studies use algorithmic audits as a methodological approach—empirical investigations that examine public algorithmic systems for potentially problematic behaviors [10]. 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 [29], 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 [21]. 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 [23, 69] to improve interpretability. Keeping this concern in mind, in our study, we chose RAG compared to few-shot prompting since the former offers greater transparency and scalability, especially in low-resource settings where examples must remain auditable and epistemically accountable [9, 24]. Furthermore, audits should broaden their focus beyond identifying and addressing bias to also include explainability metrics [41], particularly in the downstream applications of LLMs, like in content moderation.

9 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.

References

- Ahmed Agiza, Mohamed Mostagir, and Sherief Reda. 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*, Vol. 7. 2–12.
- Syed Ishtiaque Ahmed. 2022. Situating ethics: A postsecular perspective for HCI. *Interactions* 29, 4 (2022), 84–86.
- Syed Ishtiaque Ahmed. 2022. Whose intelligence? Whose ethics?: Ethical pluralism and decolonizing AI. <https://www.youtube.com/watch?v=ReSbgRSJ4WY>. last accessed: Feb 22, 2025.
- Chris Allen. 2016. *Islamophobia*. Routledge.
- The Prothom Alo. 2024. 5–20 August: 1068 minority homes and businesses attacked (translated). <https://www.prothomalo.com/bangladesh/6bm2lfn7bz>. last accessed: Feb 21, 2025.
- Tahmima Anam. 2013. Pakistan’s State of Denial. <https://www.nytimes.com/2013/12/27/opinion/anam-pakistans-overdue-apology.html>. Last accessed: July 7, 2023.
- Arjun Appadurai. 2015. Fear of Small Numbers. *Writing Religion: The Case for the Critical Study of Religion* (2015), 73–95.
- Mina Arzaghi, Florian Carichon, and Golnoosh Farnadi. 2024. Understanding Intrinsic Socioeconomic Biases in Large Language Models. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, Vol. 7. 49–60.
- Anne Arzberger, Stefan Buijsman, Maria Luce Lupetti, Alessandro Bozzon, and Jie Yang. 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*, Vol. 7. 61–73.
- Jack Bandy. 2021. Problematic machine behavior: A systematic literature review of algorithm audits. *Proceedings of the ACM on human-computer interaction* 5, CSCW1 (2021), 1–34.
- Emily M Bender, Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. 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.
- Robert Bowman, Camille Nadal, Kellie Morrissey, Anja Thieme, and Gavin Doherty. 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.
- Venetia Brown, Retno Larasati, Aisling Third, and Tracie Farrell. 2024. A Qualitative Study on Cultural Hegemony and the Impacts of AI. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, Vol. 7. 226–238.
- Bangladesh Statistics Bureau BSB. 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](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].
- Judith Butler. 2021. *Excitable speech: A politics of the performative*. routledge.
- Bhumitra Chakma. 2008. Assessing the 1997 Chittagong hill tracts peace accord. *Asian Profile* 36, 1 (2008), 93.
- Bhumitra Chakma. 2010. The post-colonial state and minorities: ethnocide in the Chittagong Hill Tracts, Bangladesh. *Commonwealth & comparative politics* 48, 3 (2010), 281–300.
- Jiawei Chen, Hongyu Lin, Xianpei Han, and Le Sun. 2024. Benchmarking large language models in retrieval-augmented generation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 38. 17754–17762.
- Robert A Cummins and Eleonora Gullone. 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*, Vol. 74. 74–93.
- Dipto Das, Shion Guha, Jed R Brubaker, and Bryan Semaan. 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.
- Dipto Das, Carsten Østerlund, and Bryan Semaan. 2021. “Jol” or “Pani”? How Does Governance Shape a Platform’s Identity? *Proceedings of the ACM on Human-Computer Interaction* 5, CSCW2 (2021), 1–25.
- William Edward Burghardt Du Bois. 2015. *Souls of black folk*. Routledge.
- Upol Ehsan, Q Vera Liao, Michael Muller, Mark O Riedl, and Justin D Weisz. 2021. Expanding explainability: Towards social transparency in ai systems. In *Proceedings of the 2021 CHI conference on human factors in computing systems*. 1–19.
- Upol Ehsan and Mark O Riedl. 2020. Human-centered explainable ai: Towards a reflective sociotechnical approach. In *International Conference on Human-Computer Interaction*. Springer, 449–466.
- Sheena Erete, Aarti Israni, and Tawanna Dillahunt. 2018. An intersectional approach to designing in the margins. *Interactions* 25, 3 (2018), 66–69.
- Ilker Etikan, Sulaiman Abubakar Musa, Rukayya Sunusi Alkassim, et al. 2016. Comparison of convenience sampling and purposive sampling. *American journal of theoretical and applied statistics* 5, 1 (2016), 1–4.
- Agence France-Presse. 2015. American atheist blogger hacked to death in Bangladesh — theguardian.com. <https://www.theguardian.com/world/2015/feb/27/american-atheist-blogger-hacked-to-death-in-bangladesh>. Last accessed July 7, 2023.
- Miranda Fricker. 2007. *Epistemic injustice: Power and the ethics of knowing*. Oxford University Press.
- Batya Friedman and Helen Nissenbaum. 1996. Bias in computer systems. *ACM Transactions on information systems (TOIS)* 14, 3 (1996), 330–347.
- Sumit Ganguly. 2021. Bangladesh’s Deadly Identity Crisis. <https://foreignpolicy.com/2021/10/29/bangladesh-communal-violence-hindu-muslim-identity-crisis/>. Last accessed: July 7, 2023.
- Sourojit Ghosh. 2024. Interpretations, Representations, and Stereotypes of Caste within Text-to-Image Generators. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, Vol. 7. 490–502.
- Sourojit Ghosh and Aylin Caliskan. 2023. Chatgpt perpetuates gender bias in machine translation and ignores non-gendered pronouns: Findings across bengali and five other low-resource languages. In *Proceedings of the 2023 AAAI/ACM Conference on AI, Ethics, and Society*. 901–912.
- Sourojit Ghosh and Aylin Caliskan. 2023. ‘Person’== Light-skinned, Western Man, and Sexualization of Women of Color: Stereotypes in Stable Diffusion. *arXiv preprint arXiv:2310.19981* (2023).
- Sourojit Ghosh, Pranav Narayanan Venkit, Sanjana Gautam, Shomir Wilson, and Aylin Caliskan. 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*, Vol. 7. 476–489.

- [35] E. Goffman. 2009. *Stigma: Notes on the Management of Spoiled Identity*. Touchstone.
- [36] Kimia Hamidieh, Haoran Zhang, Walter Gerych, Thomas Hartvigsen, and Marzyeh Ghassemi. 2024. Identifying implicit social biases in vision-language models. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, Vol. 7. 547–561.
- [37] David Hartmann, Amin Oueslati, and Dimitri Stauffer. 2024. Watching the Watchers: A Comparative Fairness Audit of Cloud-based Content Moderation Services. *arXiv preprint arXiv:2406.14154* (2024).
- [38] Mubashar Hasan. 2021. Minorities under attack in Bangladesh. <https://www.lowyinstitute.org/the-interpreter/minorities-under-attack-bangladesh>. Last accessed: July 7, 2023.
- [39] Emma Heywood, Beatrice Ivey, and Sacha Meuter. 2024. Reaching hard-to-reach communities: using WhatsApp to give conflict-affected audiences a voice. *International Journal of Social Research Methodology* 27, 1 (2024), 107–121.
- [40] Glen Hill and Kabita Chakma. 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 (2022), 519–543.
- [41] Robert R Hoffman, Shane T Mueller, Gary Klein, and Jordan Litman. 2018. Metrics for explainable AI: Challenges and prospects. *arXiv preprint arXiv:1812.04608* (2018).
- [42] Sedat Inan, Hasan Çetin, and Nurettin Yakupoğlu. 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 (2024), 397–409.
- [43] Amnesty International. 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.
- [44] Minority Rights Group International. 2018. Christians. <https://minorityrights.org/minorities/christians-6/>. Last accessed: July 7, 2023.
- [45] The Daily Ittefaq. 2014. Attacks on minorities continue. <https://web.archive.org/web/20140110191737/http://www.clickititefaq.com/more-stories/attacks-minorities-continue/>. Last accessed: July 7, 2023.
- [46] Gautier Izcard and Edouard Grave. 2020. Leveraging passage retrieval with generative models for open domain question answering. *arXiv preprint arXiv:2007.01282* (2020).
- [47] Shagun Jhaver, Iris Birman, Eric Gilbert, and Amy Bruckman. 2019. Human-machine collaboration for content regulation: The case of reddit automoderator. *ACM Transactions on Computer-Human Interaction (TOCHI)* 26, 5 (2019), 1–35.
- [48] Shagun Jhaver, Quan Ze Chen, Detlef Knauss, and Amy X Zhang. 2022. Designing word filter tools for creator-led comment moderation. In *Proceedings of the 2022 CHI conference on human factors in computing systems*. 1–21.
- [49] Jialun Aaron Jiang, Peipei Nie, Jed R Brubaker, and Casey Fiesler. 2023. A trade-off-centered framework of content moderation. *ACM Transactions on Computer-Human Interaction* 30, 1 (2023), 1–34.
- [50] Jialun Aaron Jiang, Morgan Klaus Scheuerman, Casey Fiesler, and Jed R Brubaker. 2021. Understanding international perceptions of the severity of harmful content online. *PLoS one* 16, 8 (2021), e0256762.
- [51] Hellen Koka, Solomon Langat, Francis Mulwa, James Mutisya, Samuel Owaka, Millicent Sifuna, Juliette R Ongus, Joel Lutemiah, and Rosemary Sang. 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* (2024).
- [52] 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*. 1–8.
- [53] Shanu Kumar, Gauri Kholkar, Saish Mendke, Anubhav Sadana, Parag Agrawal, and Sandipan Dandapat. 2024. Socio-Culturally Aware Evaluation Framework for LLM-Based Content Moderation. *arXiv preprint arXiv:2412.13578* (2024).
- [54] Louis Kwok, Michal Bravansky, and Lewis D Griffin. 2024. Evaluating cultural adaptability of a large language model via simulation of synthetic personas. *arXiv preprint arXiv:2408.06929* (2024).
- [55] George Lakoff. 2007. Cognitive models and prototype theory. *The cognitive linguistics reader* (2007), 130–167.
- [56] Michelle S Lam, Mitchell L Gordon, Danaë Metaxa, Jeffrey T Hancock, James A Landay, and Michael S Bernstein. 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 (2022), 1–34.
- [57] Erika Lee. 2019. *America for Americans: A history of xenophobia in the United States*. Basic Books.
- [58] Maxyn Leitner, Rebecca Dorn, Fred Morstatter, and Kristina Lerman. 2025. Characterizing Network Structure of Anti-Trans Actors on TikTok. *arXiv preprint arXiv:2501.16507* (2025).
- [59] Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, et al. 2020. Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in neural information processing systems* 33 (2020), 9459–9474.
- [60] Tianlin Li, Xiaoyu Zhang, Chao Du, Tianyu Pang, Qian Liu, Qing Guo, Chao Shen, and Yang Liu. 2024. Your large language model is secretly a fairness proponent and you should prompt it like one. *arXiv preprint arXiv:2402.12150* (2024).
- [61] Calvin A Liang, Sean A Munson, and Julie A Kientz. 2021. Embracing four tensions in human-computer interaction research with marginalized people. *ACM Transactions on Computer-Human Interaction (TOCHI)* 28, 2 (2021), 1–47.
- [62] Haley MacLeod, Grace Bastin, Leslie S Liu, Katie Siek, and Kay Connelly. 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.
- [63] Haley MacLeod, Ben Jelen, Annu Prabhakar, Lora Oehlberg, Katie A Siek, and Kay Connelly. 2016. Asynchronous remote communities (ARC) for researching distributed populations. In *PervasiveHealth*. 1–8.
- [64] Juan F Maestre, Haley MacLeod, Ciabhan L Connelly, Julia C Dunbar, Jordan Beck, Katie A Siek, and Patrick C Shih. 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.
- [65] J Nathan Matias. 2019. The civic labor of volunteer moderators online. *Social Media+ Society* 5, 2 (2019), 2056305119836778.
- [66] Nora McDonald, Sarita Schoenebeck, and Andrea Forte. 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 (2019), 1–23.
- [67] Sewon Min, Kalpesh Krishna, Xinxin Lyu, Mike Lewis, Wen-tau Yih, Pang Wei Koh, Mohit Iyer, Luke Zettlemoyer, and Hannaneh Hajishirzi. 2023. Factscore: Fine-grained atomic evaluation of factual precision in long form text generation. *arXiv preprint arXiv:2305.14251* (2023).
- [68] Mashfiq Mizan and Arifat Rahaman. 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].
- [69] Sina Mohseni, Niloofar Zarei, and Eric D Ragan. 2021. A multidisciplinary survey and framework for design and evaluation of explainable AI systems. *ACM Transactions on Interactive Intelligent Systems (TiiS)* 11, 3-4 (2021), 1–45.
- [70] Jakob Mökander, Jonas Schuett, Hannah Rose Kirk, and Luciano Floridi. 2024. Auditing large language models: a three-layered approach. *AI and Ethics* 4, 4 (2024), 1085–1115.
- [71] Maria D Molina and S Shyam Sundar. 2022. When AI moderates online content: effects of human collaboration and interactive transparency on user trust. *Journal of Computer-Mediated Communication* 27, 4 (2022), zmac010.
- [72] Marzieh Mozafari, Reza Farahbakhsh, and Noël Crespi. 2020. Hate speech detection and racial bias mitigation in social media based on BERT model. *PLoS one* 15, 8 (2020), e0237861.
- [73] Abhijit Mukherjee, Poulomee Coomar, Soumyajit Sarkar, Karen H Johanneson, Alan E Fryar, Madeline E Schreiber, Kazi Matin Ahmed, Mohammad Ayaz Alam, Prosun Bhattacharya, Jochen Bundschuh, et al. 2024. Arsenic and other geogenic contaminants in global groundwater. *Nature Reviews Earth & Environment* 5, 4 (2024), 312–328.
- [74] Richard R Orlandi, Todd T Kingdom, Timothy L Smith, Benjamin Bleier, Adam DeConde, Amber U Luong, David M Poetker, Zachary Soler, Kevin C Welch, Sarah K Wise, et al. 2021. International consensus statement on allergy and rhinology: rhinosinusitis 2021. In *International forum of allergy & rhinology*, Vol. 11. Wiley Online Library, 213–739.
- [75] Flor Miriam Plaza-del Arco, Debora Nozza, Dirk Hovy, 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.
- [76] Annu Sible Prabhakar, Lucia Guerra-Reyes, Vanessa M Kleinschmidt, Ben Jelen, Haley MacLeod, Kay Connelly, and Katie A Siek. 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.
- [77] Mohammad Rashidujjaman Rifat, Dipto Das, Arpon Poddar, Mahiratul Jannat, Robert Soden, Bryan Semaan, and Syed Ishtiaque Ahmed. 2024. 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 (2024), 1–32.
- [78] Mohammad Rashidujjaman Rifat, Abdullah Hasan Safir, Sourav Saha, Jahedul Alam Junaed, Maryam Saleki, Mohammad Ruhul Amin, and Syed Ishtiaque Ahmed. 2024. 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.
- [79] Sajal Roy, Ashish Kumar Singh, et al. 2023. Sociological perspectives of social media, rumors, and attacks on minorities: Evidence from Bangladesh. *Frontiers in Sociology* 8 (2023), 1067726.
- [80] Tanika Sarkar and Sekhar Bandyopadhyay. 2017. *Calcutta: The stormy decades*. Taylor & Francis.
- [81] Morgan Klaus Scheuerman, Jialun Aaron Jiang, Casey Fiesler, and Jed R Brubaker. 2021. A framework of severity for harmful content online. *Proceedings of the ACM on Human-Computer Interaction* 5, CSCW2 (2021), 1–33.
- [82] Ari Schlesinger, W Keith Edwards, and Rebecca E Grinter. 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.
- [83] Joseph Seering, Geoff Kaufman, and Stevie Chancellor. 2022. Metaphors in moderation. *New Media & Society* 24, 3 (2022), 621–640.
- [84] Dwaipayan Sen. 2018. *The decline of the caste question: Jogendranath Mandal and the defeat of Dalit politics in Bengal*. Cambridge University Press.
- [85] Samira Shackle. 2018. Atheist bloggers in Bangladesh are still under threat — New Humanist. <https://newhumanist.org.uk/articles/5386/atheist-bloggers-in-bangladesh-are-still-under-threat>. Last accessed July 7, 2023.
- [86] Weijia Shi, Sewon Min, Michihiro Yasunaga, Minjoon Seo, Rich James, Mike Lewis, Luke Zettlemoyer, and Wen-tau Yih. 2023. Replug: Retrieval-augmented black-box language models. *arXiv preprint arXiv:2301.12652* (2023).
- [87] Divyanshu Kumar Singh, Dipto Das, and Bryan Semaan. 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)*. Association for Computing Machinery, New York, NY, USA. doi:10.1145/3706598.3714073
- [88] Shanshan Song, Micaela Ashton, Rebecca Hahn Yoo, Zoljargal Lkhagvajav, Robert Wright, Debra JH Mathews, and Casey Overby Taylor. 2025. Participant Contributions to Person-Generated Health Data Research Using Mobile Devices: Scoping Review. *Journal of medical Internet research* 27 (2025), e51955.
- [89] Sharada Sugirtharajah. 2004. *Imagining Hinduism: A postcolonial perspective*. Routledge.
- [90] Achhiya Sultana, Dipto Das, Saadia Binte Alam, Mohammad Shidujaman, and Syed Ishtiaque Ahmed. 2024. A Civics-oriented Approach to Understanding Intersectionally Marginalized Users' Experience with Hate Speech Online. *arXiv preprint arXiv:2410.14950* (2024).
- [91] Sharifa Sultana, Rokeya Akter, Zinnat Sultana, and Syed Ishtiaque Ahmed. 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.
- [92] Heng Sun and Wan Ni. 2022. Design and Application of an AI-Based Text Content Moderation System. *Scientific Programming* 2022, 1 (2022), 2576535.
- [93] James Thorne. 2022. Data-efficient autoregressive document retrieval for fact verification. *arXiv preprint arXiv:2211.09388* (2022).
- [94] Dimitris Tsirmpas, Ion Androutsopoulos, and John Pavlopoulos. 2025. Scalable Evaluation of Online Moderation Strategies via Synthetic Simulations. *arXiv preprint arXiv:2503.16505* (2025).
- [95] The World In Us. 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].
- [96] Sahaj Vaidya, Jie Cai, Soumyadeep Basu, Azadeh Naderi, Donghee Yvette Wohn, and Aritra Dasgupta. 2021. Conceptualizing visual analytic interventions for content moderation. In *2021 IEEE Visualization Conference (VIS)*. IEEE, 191–195.
- [97] Ashley Marie Walker and Michael A DeVito. 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.
- [98] Michael Wiegand, Josef Ruppenhofer, and Elisabeth Eder. 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.
- [99] Jacob O Wobbrock. 2012. Seven research contributions in HCI. *Intelligence* 174, 12-13 (2012), 910–950.
- [100] Sijia Xiao, Shagun Jhaver, and Niloufar Salehi. 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 (2023), 1–36.
- [101] Wenjun Zeng, Yuchi Liu, Ryan Mullins, Ludovic Peran, Joe Fernandez, Hamza Harkous, Karthik Narasimhan, Drew Proud, Piyush Kumar, Bhaktipriya Radharapu, et al. 2024. Shieldgemma: Generative ai content moderation based on gemma. *arXiv preprint arXiv:2407.21772* (2024).
- [102] Alice Qian Zhang, Kaitlin Montague, and Shagun Jhaver. 2023. Cleaning up the streets: Understanding motivations, mental models, and concerns of users flagging social media posts. *arXiv preprint arXiv:2309.06688* (2023).
- [103] Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q Weinberger, and Yoav Artzi. 2019. BERTscore: Evaluating text generation with bert. *arXiv preprint arXiv:1904.09675* (2019).