Evaluating the Potentials of LLMs in User-Controlled Content Filtering on Social Media

Anonymous Author(s)

Abstract

Content moderation on social media comes with several challenges and open questions. Specifically, the moderation of hate speech proves to be intricate as it is highly dependent on context. User-controlled filtering could constitute one building block to allow for better context-sensitive hate speech filtering and foster safer online environments. Here, we present the importance of giving a subset of moderation decisions to the user and evaluate the potentials of LLMs within user-controlled content filtering on social media.

CCS Concepts

• **Human-centered computing → Empirical studies in collaborative and social computing**; **Empirical studies in HCI**.

Keywords

Content Moderation, Hate Speech, Social Media, LLMs, User-Controlled Content Filtering

ACM Reference Format:
Anonymous Author(s). 2025. Evaluating the Potentials of LLMs in User-Controlled Content Filtering on Social Media. In *Proceedings of Make sure to enter the correct conference title from your rights confirmation email (CHI '25).* ACM, New York, NY, USA, 4 pages. https://doi.org/10.1145/nnnnnnn.nnnnnnn

1 Motivation

Hate speech on social media remains a persistent challenge [2, 3]. While content moderation aims to foster safer online environments, existing approaches struggle with the complexity and context-dependence of hate speech. Automated moderation tools offer scalability but often fail to capture nuance, leading to over- or under-moderation [1]. One-size-fits-all solutions risk marginalizing vulnerable communities or failing to protect those most affected by online hostility.

A promising alternative lies in user-controlled content filters, which allow users to tailor moderation according to their specific needs. This approach holds potential to improve context-sensitivity and mitigate the pitfalls of rigid, top-down enforcement. My work investigates how large language models (LLMs) can support this paradigm. Unlike previous moderation tools, LLMs enable dynamic filtering capabilities such as expanding keyword filters with adjacent terms, generating counter-speech, or rephrasing harmful content in more neutral language. Yet, their role in user-centered moderation remains underexplored.

This paper proposes integrating LLMs into user-controlled filtering as a novel contribution to the content moderation literature. I argue that LLMs can meaningfully augment personalization, provided they are evaluated for fairness, usability, and effectiveness. I outline design and implementation pathways that consider participatory methods, including user studies, A/B testing, and cognitive burden assessments.

The theme of this workshop, *Mind the Context*, aligns closely with my work: first, by emphasizing the need for moderation decisions that respect the sociolinguistic and cultural contexts in which hate speech occurs; second, by investigating how emerging technologies like LLMs shift the power dynamics and ethical considerations of moderation.

I hope to contribute to ongoing conversations about how LLMs can support diverse user needs in a transparent and participatory manner. My aim is to design moderation tools that are both adaptive and accountable, and I welcome the opportunity to engage with others exploring the intersection of AI, ethics, and human-centered design.

2 The Case for User-Controlled Content Filters

The prevalence of hate speech on social media has become a persistent concern for both users and platform administrators [12–14]. While social media platforms facilitate engagement and information exchange, they also provide a space for harmful content to spread. Content moderation—the process of managing user-generated content to foster collaboration and prevent misuse—is a central mechanism for mitigating these harms [7, 9]. However, moderating hate speech remains highly challenging due to biases in automated detection, limited contextual awareness, and inconsistent definitions of harmful content [8]. These limitations often lead to disproportionate impacts on marginalized communities, either through over-enforcement that silences counter-speech [1] or under-enforcement that fails to curb harassment [4].

Content moderation can be categorized into three primary types: platform moderation, community moderation, and personal moderation [10]. Platform moderation, typically enforced by automated systems and content reviewers, plays a crucial role in enforcing policy guidelines but struggles with bias and lack of contextual nuance. Automated tools often fail to distinguish between different uses of slurs (e.g., hate speech vs. reclaimed language) [1, 5] and exhibit inconsistent performance across languages and dialects [15]. These shortcomings highlight the inadequacy of one-size-fits-all moderation approaches, particularly in diverse and dynamic online communities.

User-controlled content filtering has emerged as a promising alternative, allowing individuals to tailor content filtering to their specific needs [10]. This approach offers flexibility in adapting to

CHI '25, Yokohama, Japan

ACM ISBN 978-x-xxxx-xxxx-x/YYYY/MM
https://doi.org/10.1145/nnnnnnn.nnnnnnn

evolving language and social norms while potentially addressing implicit hate speech that automated systems struggle to detect [10]. Studies suggest that many users prefer user-controlled filtering over rigid, platform-driven moderation [11]. However, this model introduces new challenges, including concerns about increased cognitive burden, the risk of creating ideological echo chambers, and the difficulty of designing intuitive filtering mechanisms that align with user intent [10].

Despite growing interest in user-controlled filtering, research on its effectiveness remains limited. Existing third-party tools offer some degree of customization, but their impact on mitigating hate speech has not been comprehensively studied. This gap underscores the need for further research into user preferences, usability challenges, and the development of robust filtering tools that balance customization with accessibility.

My work explores user-controlled filtering, supported by LLMs, as a means to reduce exposure to hate speech. By integrating LLMs into user-controlled filtering, I aim to investigate how personalized moderation can enhance user agency while addressing the limitations of platform-driven approaches.

LLMs offer new opportunities for improving content moderation by enhancing context-sensitivity and allowing more nuanced filtering decisions. Traditional automated moderation often struggles with implicit hate speech, reclaimed slurs, and cultural context, leading to both over- and under-enforcement. LLMs, when carefully fine-tuned, could provide more adaptable filtering options, helping users customize moderation settings to better reflect their needs. However, the use of LLMs in this domain also raises critical challenges, such as bias in model outputs, computational costs, and the risk of amplifying existing moderation disparities.

To advance this research, I investigate the following key questions:

- How can LLMs improve the context-sensitivity of user-controlled filtering for hate speech moderation?
- What are the usability challenges and ethical concerns in integrating LLMs into user-controlled content filtering?
- How can such filtering be designed to balance user agency with concerns about the (ir)reliability and potential biases of LLM outputs and cognitive burden?

By addressing these questions, my work seeks to develop fairer, more transparent, and user-driven moderation systems that leverage LLMs to enhance safety while preserving free expression.

3 Human-Centered Evaluation of LLMs in User-Controlled Content Filtering

A human-centered evaluation and auditing approach is crucial for understanding how user-controlled content filtering systems operate in real-world contexts. Unlike platform-driven moderation, which applies top-down enforcement, user-controlled filtering shifts decision-making to individuals, allowing them to personalize their exposure to potentially harmful content. While this approach offers greater autonomy, it also introduces new risks and trade-offs, necessitating rigorous evaluation that goes beyond traditional accuracy-based metrics.

Human-centered evaluation should examine how filtering systems affect users cognitively, emotionally, and socially, ensuring that these tools are not only effective in reducing harm but also usable, fair, and transparent. This requires moving beyond binary classification metrics (e.g., precision and recall in hate speech detection) and incorporating qualitative and participatory insights into system evaluation.

I advocate that a comprehensive human-centered evaluation framework of LLMs in User-Controlled Content Filtering should integrate the following dimensions:

- **Effectiveness in Harm Reduction**: To what extent does the system reduce users' exposure content that is assessed as harmful or distressing by the respective user? Does it achieve this without over-filtering, which could limit access to critical discourse or counter-speech?
- **Balance between User Agency and Cognitive Burden**: How intuitive and manageable is the filtering system for diverse users? Does it offer meaningful control without overwhelming users with complexity? Are users able to fine-tune the filtering criteria in a way that aligns with their needs, or does it require excessive labor?
- **Fairness and Inclusion**: Does the system disproportionately silence certain groups, particularly marginalized communities? How does it handle linguistic and cultural nuances in hate speech detection, particularly for reclaimed language or counter-speech? Are its filtering mechanisms biased against specific political viewpoints, identities, or communities?
- **Transparency and Explainability**: Can users understand why certain content was filtered? Does the system provide clear feedback and control mechanisms to adjust filtering settings? Are the filtering decisions interpretable, particularly in cases where LLMs are used for moderation?
- **Context Sensitivity**: How well does the system account for the context of online interactions, such as satire, irony, or community norms? Can it distinguish between hate speech, counter-speech, and legitimate critique? Does it consider evolving social and political dynamics in hate speech discourse?
- **Longitudinal Impact and Adaptability**: How do user perceptions of filtering change over time? Does the system allow for adaptation based on feedback and evolving user needs? Are there mechanisms to prevent unintended consequences, such as the reinforcement of filter bubbles or desensitization to harmful content?

Unlike platform-wide moderation policies, which can be evaluated through aggregate content trends, user-controlled filtering requires individualized assessment that considers personal preferences, situational contexts, and shifting online dynamics. This introduces several challenges.

While giving users control over their filtering settings enhances autonomy, self-imposed filtering choices can have broader societal consequences. If filtering leads to excessive content avoidance, it may limit exposure to critical discussions, counter-speech, or alternative perspectives, raising concerns about self-reinforcing echo chambers. Evaluation must examine whether user-controlled content filters enable informed decision-making rather than overly insulating users from diverse viewpoints.

LLM-assisted filtering mechanisms must be evaluated for biases that may disproportionately affect marginalized communities. Certain phrases or linguistic styles—particularly those used in activist discourse, reclaimed language, or cultural vernaculars—may be misclassified as harmful [1]. Evaluation must include intersectional evaluations, ensuring that filtering models do not replicate discriminatory biases found in training data.

Another challenge of AI-assisted moderation is opacity in decision-making. If users do not understand how their filters operate, they may lose trust in the system or fail to make adjustments that align with their needs. An evaluation must assess whether filtering explanations are accessible and actionable and how well users can modify their settings if unintended filtering occurs.

While personalization enhances user control, excessive configurability may result in decision fatigue [6]. Evaluation should assess whether users can navigate filtering options without extensive technical knowledge and whether default settings provide a useful starting point for non-expert users.

Hate speech is context-dependent and constantly evolving, making static filtering rules insufficient. User-controlled content filters must be evaluated in terms of adaptability to new forms of online abuse and effectiveness refining moderation over time.

To ensure meaningful evaluation, I propose a mixed-methods approach that incorporates:

- **User Studies and Usability Testing**: Conducting lab-based and field studies with diverse participants to assess usability, perceived control, and satisfaction with filtering outcomes.
- **Qualitative Interviews and Participatory Feedback**: Gathering narratives from marginalized users to understand how moderation affects them personally and culturally, especially across language and regional contexts.

4 The Role of Participatory Design

Participatory Design (PD) is essential in ensuring that user-controlled filtering tools are context-sensitive and reflective of the lived experiences of those most affected by online hate speech. PD fosters active user involvement in shaping filtering mechanisms, ensuring that solutions accommodate diverse risks and sensitivities rather than imposing a one-size-fits-all approach. It also helps uncover contextual nuances—for example, how certain communities reclaim language or balance visibility with safety.

My research contributes to this agenda by developing participatory design methods that integrate user perspectives into the development and evaluation of LLM-assisted filtering. Standard evaluation approaches often fail to capture the lived experiences of those most affected by online harm. To address this gap, I explore participatory frameworks that incorporate the needs of communities disproportionately targeted by hate speech. By centering their insights, I aim to refine evaluation methodologies that account for linguistic nuance, shifting social contexts, and the unintended consequences of automated filtering.

My research examines how users interact with filtering tools, the extent to which they understand and trust moderation decisions, and whether current design approaches support meaningful engagement without overwhelming users with complexity.

Finally, I am interested in the fairness and transparency of filtering mechanisms, particularly in their impact on marginalized communities. Many automated moderation tools struggle with bias, misclassifying reclaimed language or activist discourse as harmful while failing to recognize more covert forms of abuse. Evaluating filtering systems requires not only auditing their immediate effects but also considering their long-term implications.

5 Conclusion

I want to develop human-centered evaluation methods for user-controlled filtering that enhance safety while preserving user autonomy and discourse diversity. By integrating participatory auditing approaches, I aim to refine how LLM-assisted filtering is assessed, ensuring it aligns with the needs of those most affected by online harm. My goal is to establish best practices for evaluating AI-driven moderation, balancing usability, fairness, and transparency. I would be happy to discuss participatory auditing methods, ethical challenges, and regulatory considerations for LLM moderation. Collaborating with researchers focused on AI-driven moderation would help strengthen participatory auditing methods, explore best practices for evaluating filtering mechanisms, and contribute to broader conversations on ethical and regulatory frameworks for LLM moderation.

References

[1] Adeela Arshad-Ayaz, M. Naseem, and Hedia Hizaoui. 2022. Perspectives of Muslim and Minority Canadian Youth on Hate Speech and Social Media. *Journal of Contemporary Issues in Education* (2022). doi:10.20355/jcie29489

[2] S. Castaño-Pulgarín, Natalia Suárez-Betancur, L. M. T. Vega, and Harvey Mauricio Herrera López. 2021. Internet, social media and online hate speech. Systematic review. *Aggression and Violent Behavior* 58 (2021), 101608. doi:10.1016/J.AVB.2021.101608

[3] Melisa Castellanos, Alexander Wettstein, Sebastian Wachs, Julia Kansok-Dusche, Cindy Ballaschk, Norman Krause, and Ludwig Bilz. 2023. Hate speech in adolescents: A binational study on prevalence and demographic differences. In *Frontiers in Education*, Vol. 8. Frontiers Media SA, Lausanne, Switzerland, 1076249.

[4] Lucas Dixon, John Li, Jeffrey Sorensen, Nithum Thain, and Lucy Vasserman. 2018. Measuring and mitigating unintended bias in text classification. In *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society*. 67–73.

[5] Rebecca Dorn, Lee Kezar, Fred Morstatter, and Kristina Lerman. 2024. Harmful Speech Detection by Language Models Exhibits Gender-Queer Dialect Bias. In *Proceedings of the 4th ACM Conference on Equity and Access in Algorithms, Mechanisms, and Optimization* (San Luis Potosi, Mexico) *(EAAMO '24)*. Association for Computing Machinery, New York, NY, USA, Article 6, 12 pages. doi:10.1145/3689904.3694704

[6] Andrew Flangas, Alexis R Tudor, Frederick C Harris Jr, and Sergiu Dascalu. 2021. Preventing decision fatigue with aesthetically engaging information buttons. In *International Conference on Human-Computer Interaction*. Springer, 28–39.

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

[8] David Hartmann, Amin Oueslati, and Dimitri Staufer. 2024. Watching the Watchers: A Comparative Fairness Audit of Cloud-based Content Moderation Services. *arXiv preprint arXiv:2406.14154* (2024).

[9] 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.

[10] Shagun Jhaver, Alice Qian Zhang, Quan Ze Chen, Nikhila Natarajan, Ruotong Wang, and Amy X. Zhang. 2023. Personalizing Content Moderation on Social Media: User Perspectives on Moderation Choices, Interface Design, and Labor. *Proceedings of the ACM on Human-Computer Interaction* 7, CSCW2 (Sept. 2023), 1–33. doi:10.1145/3610080

[11] Shagun Jhaver and Amy X Zhang. 2023. Do users want platform moderation or individual control? Examining the role of third-person effects and free speech support in shaping moderation preferences. *New Media & Society* (2023), 14614448231217993.

[12] Sandra Miranda, Fabio Malini, Branco Di Fátima, and Jorge Cruz. 2022. I love to hate!: the racist hate speech in social media. *European Conference on Social Media* (2022). doi:10.34190/ecsm.9.1.311

[13] Sandra Lopes Miranda. 2023. Analyzing Hate Speech Against Women on Instagram. *Open Information Science* 7 (2023). doi:10.1515/opis-2022-0161

[14] Zewdie Mossie and Jenq-Haur Wang. 2020. Vulnerable community identification using hate speech detection on social media. *Inf. Process. Manag.* 57 (2020), 102087. doi:10.1016/J.IPM.2019.102087

[15] Tian Xiang Moy, Mafas Raheem, and R. Logeswaran. 2021. Hate Speech Detection in English and Non-English Languages: A Review of Techniques and Challenges. *Webology* (2021). doi:10.14704/web/v18si05/web18272