Algorithmic Gatekeepers:
The Human Rights Impacts of LLM Content Moderation

VIII. Preliminary Recommendations

European Center for
Not-for-Profit Law

Acknowledgements:

Author: Marlena Wisniak, ECNL.

We extend special thanks to Isabelle Anzabi, who contributed significantly during the early stages of the research process. We express gratitude to the ECNL team—Karolina Iwanska, Vanja Skoric, and Francesca Fanucci—for theirthoughtful review and feedback on the report.

Valuable input and review were provided by Evani Radiya-Dixit from the American Civil Liberties Union (ACLU); Lindsey Andersen from Business for Social Responsibility (BSR); Mona Elswah and Aliya Bhatia from the Center for Democracy and Technology (CDT); independent researcher and policy expert Luca Belli; and Roya Pakzad from Taraaz.

Insightful contributions through interviews and consultations came from representatives of Meta's Human Rights Team, the Policy and Safety Machine Learning Teams at Discord, and the Research Team at Jigsaw.

We extend our sincere gratitude to everyone who generously contributed their invaluable time, insights, and expertise to the preparation of this report. Your thoughtfulness and creativity have greatly enriched the quality and depth of our findings. We thank Betsy Popken of the UC Berkeley Human Rights Center; Corynne McSherry from the Electronic Frontier Foundation (EFF); Daniel Leufer and Eliska Pirkova from Access Now; Dave Willner of Stanford University; Dunstan Alison Hope; Jonathan Stray from UC Berkeley; Justin Hendrix of Tech Policy Press; Mike Masnick of Techdirt; Paul Barrett from New York University; Sabina Nong of Stanford University; Tarunima Prabhakar from Tattle; and Vladimir Cortes.

Design for the publication was created by Sushruta Kokkula and Andrea Judit Tóth. The illustrations featured in the report are based on the work of Balázs Milánik, Rozalina Burkova (The Greats) and Daniela Yankova (The Greats).

We thank the Omidyar Network for their generous support.

April 2025

Table of contents

Preliminary recommendations to policy makers 4

Preliminary recommendations to AI developers and deployers 6

Recommendations specific to LLM developers 7

Recommendations specific to LLM deployers 7

Recommendations to both LLM developers and deployers 8

Preliminary recommendations

These preliminary recommendations are based on our initial assessment of the potential human rights risks associated with using LLMs for content moderation. Throughout the year, we will further refine and develop these recommendations in collaboration with civil society, researchers, and industry stakeholders to make them more actionable and effective.

Additionally, we will conduct a deeper analysis of AI value chains, with a particular focus on the relationship between LLM developers and deployers, and identify key areas for policymakers to address.

Preliminary recommendations to policy makers

Refrain from mandating the use of AI systems, including LLMs, for content moderation

Policymakers should ensure that no law requires or incentivises platforms to use LLMs or automated content analysis tools for detecting or removing illegal content.[1] Automated systems, while useful in some contexts, should not replace human oversight in decisions affecting human rights and their use should never be mandated. Indeed, the former UN Special Rapporteur on Freedom of Expression has warned against laws requiring proactive content filtering, emphasising that such measures undermine privacy and risk leading to pre-publication censorship.[2]

Mandate human overview

Policymakers should establish regulatory frameworks requiring meaningful human oversight of LLM systems to ensure accountability for high-stakes AI decisions. This governance should mandate multi-tiered human supervision processes, including comprehensive review during development, deployment-level controls for system operators, and accessible feedback mechanisms for end-users, including an obligation to document and report on oversight protocols. For LLM moderation, platforms must prohibit fully automated, irreversible high-consequence decisions and establish robust appeal mechanisms where all automated actions remain human-reviewable and reversible. Rather than relying on industry self-regulation alone, clear regulatory standards with appropriate enforcement mechanisms are essential to guarantee LLM systems operate with proper human guidance in applications impacting human rights.

---

1 Freedom House has reported that at least 22 countries currently mandate or encourage platforms to use machine learning to remove political, social, or religious content deemed unfavorable
Freedom House. (2023, October 4). New report: Advances in artificial intelligence are amplifying a crisis for human rights online. https://freedomhouse.org/article/new-report-advances-artificial-intelligence-are-amplifying-crisis-human-rights-online#:~:text=applications%20comply%20with%20or%20strengthen,supporters%20must%20adapt%20the%20lessons

2 Office of the United Nations High Commissioner for Human Rights. (2018). A human rights approach to online content regulation: Thematic report. https://www.ohchr.org/sites/default/files/Documents/Issues/Expression/Factsheet_2.pdf

Focus regulatory efforts on platform transparency and accountability

Instead of granting platforms greater authority to assess and remove user-generated content, regulations should target the structural issues within social media business models, particularly recommender systems.[3] Policymakers should require greater transparency from platforms about how their recommendation algorithms function (including their accuracy, precision, and recall rate, among others). They should mandate human rights due diligence in the design and deployment of AI-based moderation tools. Importantly, they should ensure transparency in content moderation decisions and improve mechanisms for dispute resolution. Where data protection or transparency laws are in effect, they should meaningfully enforce those.

The EU DSA aims to address some of these issues, albeit insufficiently, according to civil society groups in Europe.[4] Still, similar, and more ambitious approaches should be encouraged globally.

Require developers and deployers to conduct human rights impact assessments[5]

Policymakers should mandate HRIAs for the development and use of LLM moderation. HRIAs should be conducted at all stages of the AI lifecycle, starting with the ideation stage and running through post-deployment, including a process for reviewing the impacts in an iterative and ongoing way. Appropriate resources and capacity must be allocated for this purpose to ensure adequate classification and assessment. Policymakers should clearly identify which events or circumstances would require that an HRIA be undertaken. HRIAs should always prioritise harm reduction and the adverse human rights impacts on marginalised groups, taking a holistic approach and assessing the impacts of AI systems on a wide range of human rights.

To promote transparency, HRIA results should be publicly accessible via registers, press releases, social media, and public archives. Legal frameworks should balance public access with protecting trade secrets and intellectual property. An external oversight body should be established to review HRIAs regularly, with public access to all oversight information. This will ensure accountability and prevent developers from self-assessing impacts without external validation.

Policymakers should promote methods that involve marginalised and affected communities in the HRIA process. Civil society organisations and community representatives should help shape priorities and decisions regarding AI deployment.

Finally, policymakers should integrate HRIAs with other accountability mechanisms, by harmonising HRIAs with data protection, environmental due diligence, algorithmic audits, and citizen review boards to enhance comprehensive accountability and harm mitigation.

---

3 ARTICLE 19. (2023, August). (p.35). Content moderation and freedom of expression handbook. https://www.article19.org/wp-content/uploads/2023/08/SM4P-Content-moderation-handbook-9-Aug-final.pdf

4 EDRi. (2022, November 16). The DSA fails to reign in the most harmful digital platform businesses – but it is still useful. European Digital Rights (EDRi). https://edri.org/our-work/the-dsa-fails-to-reign-in-the-most-harmful-digital-platform-businesses-but-it-is-still-useful/

5 European Center for Not-for-Profit Law (ECNL). (2021, November 23). Recommendations for incorporating human rights into AI impact assessments. https://ecnl.org/publications/recommendations-incorporating-human-rights-ai-impact-assessments

Meaningfully include external stakeholders, especially civil society and affected communities, in AI policy and development

Policymakers should m**eaningfully include external stakeholders, especially civil society** and affected communities, in AI policy. They should allocate adequate resources. For more information om meaningful engagement in policymaking, see ECNL's research on participatory tools and models.[6]

Policymakers should also require LLM developers and deployers to meaningfully engage a diverse range of stakeholders, including civil society, researchers, and marginalised communities, at all stages of the LLM lifecycle.

Preliminary recommendations to AI developers and deployers

Relationship between LLM developers and deployers

Most recommendations in this section apply to both LLM developers and deployers who adapt and fine-tune these models specifically for content moderation on their platforms. LLM developers create and train foundational models designed with broad capabilities, establishing the initial performance standards and classification limitations. Deployers then take these general models and fine-tune or adjust them to align closely with their internal content moderation policies, community standards, and regulatory obligations.

This relationship generates shared responsibilities for rights-based LLM moderation. Developers provide the underlying technology and frameworks that allow customisation and calibration, thus enabling deployers to adapt these models effectively. Deployers, in turn, are responsible for carefully selecting suitable models, fine-tuning them according to their unique needs and contexts, and implementing appropriate human oversight systems to mitigate and remedy harms specific to their platforms.

When reading recommendations addressed to both LLM developers and deployers, the dynamic between both stakeholders should be considered to contextualise their responsibilities. For more information, read BSR's Human Rights Across the Generative AI Value Chain report.[7]

6 European Center for Not-for-Profit Law (ECNL). (2023, November 13). New dimensions for public participation. https://ecnl.org/publications/new-dimensions-public-participation

7 Hoh, J. Y., Nigam, S., Andersen, L., & Darnton, H. (2025). Human rights across the generative AI value chain: Human rights assessment of the generative AI value chain and responsible AI practitioner guides. BSR. https://www.bsr.org/en/reports/human-rights-across-the-generative-ai-value-chain

Recommendations specific to LLM developers

LLM developers should establish clear and comprehensive Acceptable Use Policies (AUPs) for deployers seeking to use LLMs for content moderation. An AUP outlines the permitted ways in which a product or service can be used, as well as prohibited activities and consequences for violations. As advised by BSR, the AUP should include an explicit human rights commitment, emphasising the rights most severely impacted by LLM moderation (i.e. salient human rights risks) and meaningful human rights impact assessments. In the context of using LLMs for content moderation, the AUPs could for instance define clear boundaries between acceptable automation of content review and situations requiring human oversight, a requirement to establish robust appeal mechanisms for content decisions made entirely or partially by LLMs, mandatory transparency about when content decisions involve LLM systems, and rigorous measures to address potential risks of bias, inconsistency, and over- and under-moderation, especially of marginalised groups. Furthermore, they can include concrete metrics to assess policy effectiveness and human rights impacts, noting that metrics should be both quantitative and qualitative.

LLM developers should meaningfully engage with external stakeholders, especially civil society and affected communities, in the development and ongoing review of the AUP. LLM developers must effectively enforce the AUP, ensuring that it's applied consistently across all jurisdictions, even when local laws conflict with international human rights standards. They should continuously review and update the AUP in response to evolving moderation technologies, emerging threats, user feedback, and shifting human rights contexts.

Recommendations specific to LLM deployers

LLM deployers should thoroughly review vendor AUPs and documentation to ensure alignment with their platform's content policies. They should assess model performance across diverse content categories, regions, and languages relevant to their user base, and request transparency regarding training data demographics and known biases to anticipate potential moderation blind spots.

LLM deployers should conduct comprehensive scenario testing with platform-representative content before implementing the LLM for content moderation purposes. They should pay special attention to edge cases involving marginalised communities and culturally-specific content that may be prone to misinterpretation or bias, and evaluate false positive and negative rates across different content types and user demographics to identify potential disparate impacts.

LLM deployers should negotiate robust contractual protections with developers, including minimum performance standards across diverse content types, commitments for regular model updates, explicit audit rights for model performance on their specific content, and clear liability frameworks for systematic errors that might negatively impact users, especially marginalised groups, or create legal exposure.

LLM deployers should implement redundancy systems to mitigate single-source dependence risks. They should consider deploying multiple LLMs from different providers for high-stakes moderation decisions to cross-validate results, develop fallback procedures for system unavailability or underperformance, and maintain sufficient human moderation capacity that can scale during system transitions or disruptions.

Recommendations to both LLM developers and deployers

Minimise data collection and embed privacy-by-design into product development and use

LLM developers and deployers should minimise data collection, processing, and sharing by adopting privacy-by-design principles and implementing rigorous data protection measures into every stage of LLM lifecycle. It's important to avoid indiscriminate scraping of user data; instead, LLM developers should curate training datasets to exclude sensitive personal information or at least anonymise it.

This means applying the EU GDPR's data minimisation principle – personal data should be "limited to what is necessary."[8] This includes collecting only essential data, anonymising or pseudonymising it wherever possible, and limiting data retention to the minimum duration necessary. For example, an LLM moderation system might only need to store content features (like embeddings for policy-relevant attributes) rather than full user identifiers. Moreover, employing privacy-preserving techniques (such as hashing identifiers or using differential privacy on training data) can allow the model to detect patterns (e.g. repeat abusive behaviour) without exposing individuals' identities.

Developers and deployers should use secure, encrypted storage and ensure robust access control mechanisms (including security audits) to protect sensitive data. Additionally, sharing data with third parties should be strictly limited and governed by comprehensive data-sharing agreements, emphasising user consent and transparency. Regular audits and adherence to data protection regulations, such as the EU GDPR and California Consumer Privacy Act (CCPA), are essential to maintain accountability and protect users' rights.

Ensure human oversight

LLM developers should implement granular oversight mechanisms that enable multiple tiers of human intervention: first, internal red team review during development, second, API-level controls for downstream deployers, and third, end-user feedback channels. These systems should provide appropriate visibility into model decision processes, facilitate timely intervention at critical decision points, and incorporate feedback loops that improve model behaviour over time.

LLM deployers should for their part incorporate human oversight mechanisms into their content moderation work flows. Systems should be designed such that decisions—especially borderline, complex, or high-stakes cases where the content calls for more nuanced assessments—automatically get flagged for review by a trained human moderator before final action.[9] For example, if an LLM generates an output with low confidence or if the content involves context that the model might not grasp (such as satire, political criticism, or culturally specific language), a human should make the call. In any case, all automated decisions should be reversible by humans, not

8 GDPR.eu. (n.d.). Article 5 – Principles relating to processing of personal data. General Data Protection Regulation (GDPR). Retrieved from https://gdpr-info.eu/art-5-gdpr/#:~:text=purposes%20%28%E2%80%98purpose%20limitation%E2%80%99%29%3B%203,which%20the%20personal%20data%20are

9 Digital Trust & Safety Partnership. (2024, September). Best practices for AI and automation in trust and safety. (p. 2) Digital Trust & Safety Partnership. https://dtspartnership.org/best-practices-for-ai-and-automation-in-trust-and-safety/

only for deciding what content to take down or leave up, but also from broader content moderation decisions. The moderation staff should be properly trained on when and how to intervene in automated decisions, including by LLMs.

Conduct meaningful human rights due diligence, including impact assessments

LLM developers and deployers should identify and assess the potential human rights implications of using LLMs for content moderation as consistent with Guiding Principle 17 (UNGPs). The process should focus on individuals and groups who may be impacted, considering factors such as demographic and identity characteristics and geographic location, centering impacts to marginalised groups.[10] It is important to also assess indirect and cumulative impacts that may arise as a consequence of the AI system's implementation.

After identifying potential impacts, LLM developers and deployers should assess the specific impacts on the human rights of likely affected individuals and groups. This assessment should consider the severity and likelihood of potential harm, combining qualitative and quantitative data to substantiate the findings. Stakeholder engagement is essential at this stage to ensure diverse perspectives and expertise are considered. Following the impact assessment, LLM developers should implement appropriate mitigation measures to prevent and minimise adverse impacts on human rights. These measures may involve adjusting the design or functionality of the LLM for the purposes of content moderation, implementing technical safeguards, enhancing user controls, and providing transparent documentation to users and affected individuals. Addressing hallucinations and factual inaccuracies in particular will require targeted mitigation strategies and scalable solutions.[11]

LLM Developers and deployers should document the entire HRIA process to ensure both internal and external transparency. This documentation should include the methods used for identifying and assessing impacts, the mitigation measures taken, consultations and stakeholder engagements conducted, and the decisions made throughout the process. Documentation should be accessible to supervisory authorities and the public when appropriate to demonstrate compliance and inform decision-making.

Finally, the results of the HRIA and the AI system itself should be continuously monitored and reviewed, particularly if significant changes occur or new concerns arise. Developers and deployers should establish mechanisms for regular updates to the HRIA and ongoing stakeholder engagement. For example, they should do iterative testing to assess accuracy, emerging harms, and evolving user behaviors, with the goal to identify unanticipated behaviors and unintended consequences. As recommended by the DTSP, LLM developers and deployers should "tak[e] a deliberate, iterative process to GenAI experimentation and roll-out."[12] They recommend increasing the frequency and depth of risk assessments for generative AI due to its complexity and potential for unpredictable outputs, especially for LLM deployers.

---

10 We acknowledge that most proposed algorithmic fairness techniques require access to and process of demographic attribute data. To learn more about this tension, and data collection practices and governance frameworks in the public interest, see Partnership on AI's "Demographic Data" project. Partnership on AI. (n.d.). Demographic data. Partnership on AI. Retrieved from https://partnershiponai.org/work stream/demographic-data/

11 Digital Trust & Safety Partnership. (2024, September). Best practices for AI and automation in trust and safety. (p. 40-41) Digital Trust & Safety Partnership. https://dtspartnership.org/best-practices-for-ai-and-automation-in-trust-and-safety/

12 Ibid.

Carry out red-teaming and adversarial testing

First, LLM developers and deployers should conduct continuous red-teaming exercises to simulate real-world misuse and uncover vulnerabilities in LLM moderation tools. To test LLM robustness, they should specialised testing techniques that deliberately try to make platform's LLMs fail or produce problematic outputs. This includes identifying ways the model can be manipulated to produce incorrect, biased statements, or harmful content, and test how the model responds to edge cases or unusual requests. These can then be used to refine moderation systems and improve coverage.[13] LLM developers and deployers should then leverage findings from adversarial testing to iteratively strengthen the LLM's safeguards against manipulation and misuse.

To complement the human rights impact assessments, LLM developers and deployers should conduct continuous audits, focusing on areas where LLM responses may be particularly problematic or harmful to users and affected communities. The audits should meaningfully include external stakeholders, and test LLM performance across diverse demographics and linguistic contexts to identify and mitigate risks of discrimination.

Audits must have clear definitions for key concepts such as hate speech, incitement to violence, disinformation, and terrorist content online. Auditing methodologies should allow for flexibility and responsiveness to emerging issues, i.e. be adaptable to assess LLM performance on a range of human rights concerns and contexts. At the same time, auditing metrics should enable consistent evaluation across LLM systems and tools, setting a high bar for standardisation.[14]

To the extent possible, audits should integrate both quantitative and qualitative assessments to provide robust and actionable insights. If auditors rely on automation to conduct some aspects of the audit, they should ensure that any automated approaches are designed to capture nuanced and uncommon issues.[15] Integrating monitoring and quality control mechanisms to ensure high quality is necessary as well.

Audit findings should be applicable beyond controlled testing environments, offering insights into real-world implications. LLM developers should assess how the identified risks impact users directly, and what downstream applications they could have for those who deploy LLMs in their products. Clear frameworks should be established for measuring and addressing recurrent issues across different deployment contexts, while at the same time being applicable to uncommon and exceptional circumstances, too.

Importantly, LLM developers and deployers should ensure that all auditing processes include independent oversight and external validation.

13 Microsoft Research. (2022, May). (De)ToxiGen: Leveraging large language models to build more robust hate speech detection tools. Microsoft. https://www.microsoft.com/en-us/research/blog/detoxigen-leveraging-large-language-models-to-build-more-robust-hate-speech-detection-tools/

14 Allen, D., Denkovski, O., & Giannaccini, F. (2024, October 14). Ensuring AI accountability: Auditing methods to mitigate the risks of large language models. Democracy Reporting International. https://democracy-reporting.org/en/office/EU/publications/ensuring-ai-accountability-auditing-methods-to-mitigate-the-risks-of-large-language-models#UsingLLMstoEvaluateLLMs%E2%80%99Outputs

15 Ibid.

Meaningfully engage external stakeholders in LLM design, development, and use

LLM Developers and deployers should build strong relationships with diverse stakeholders, including civil society organisations, researchers, and affected communities, and meaningfully include them throughout the LLM lifecycle—including in the impact assessment and audit processes. Developers should regularly update these groups on new product and policy development related to LLM moderation, and integrate their perspectives into their products, policies, and governance. For more information about how to achieve this, read ECNL's framework for meaningful engagement.[16]

As recommended by the DTSP, LLM developers and deployers should collaborate across teams and functions, ensuring that subject matter experts, engineers, trust and safety advisors, lawyers, and human rights specialists are involved in the design, development, and deployment of AI systems. For example, they could establish cross-functional committees to assess risks before developing or using LLMs for content moderation. Through a collaborative process, policy and operations teams should drive model development, including the creation of classifiers and dataset labelling for training and evaluation, ensuring that any identified risks are effectively mitigated or prevented. When setting model performance thresholds for sensitive applications (such as severe human rights impacts or deployment in high-risk geographic areas), teams with special thematic and regional expertise should be consulted.[17]

Up skilling teams in AI governance and human rights is critical. LLM developers and deployers should invest in training staff on best practices in AI policy enforcement, prompt engineering, and fine-tuning generative AI models for trust and safety applications, ensuring that users and affected communities' rights are protected. [18]

Ensure adequate language and cultural context

LLM developers and deployers should urgently improve LLM performance in low-resource languages, as LLM systems must be designed to perform effectively across high- and low-resource languages, rather than defaulting to English-centric approaches. Developers and deployers must ensure that LLM tools account for regional dialects, slang, and nuanced cultural expressions, which are often misclassified by Global North-centric models.[19]

To achieve this, LLM developers and deployers should consult language experts, annotators, and community leaders from the regions and communities where their models are deployed. This approach can significantly improve the contextual understanding needed to moderate content including hate speech, satire, reclaimed language, and "code mixing" (where speakers switch between languages), reducing systemic biases in AI moderation and NLP applications. [20] LLM developers and deployers should partner with local researchers, universities, and linguistic institutions

16 European Center for Not-for-Profit Law. (2023). Framework for meaningful engagement in human rights impact assessments of AI. European Center for Not-for-Profit Law. https://ecnl.org/publications/framework-meaningful-engagement-human-rights-impact-assessments-ai

17 Digital Trust & Safety Partnership. (2024, September). Best practices for AI and automation in trust and safety. (p. 30) Digital Trust & Safety Partnership. https://dtspartnership.org/best-practices-for-ai-and-automation-in-trust-and-safety/

18 Ibid., p. 40-41.

19 Radiya-Dixit, E., & Bogen, M. (2024, October). Beyond English-centric AI: Lessons on community participation from non-English NLP groups (p. 4). Center for Democracy & Technology. https://cdt.org/wp-content/uploads/2024/10/2024-10-18-AI-Gov-Lab-Beyond-English-Centric-AI-brief-final.pdf.

20 Ibid.

to source high-quality, contextually relevant training data around the world. They should invest in local (underfunded) NLP communities, e.g., African NLP and Arabic NLP researchers, as these groups can foster innovation and improve AI models while addressing concerns about Global North dominance in AI development.[21]

Generating new, culturally relevant, and community-driven multilingual data – rather than relying solely on translation from high-resource languages – is key.[22] LLM developers and deployers should then iterate on dataset development by engaging diverse community members in the content creation, annotation, and quality assurance processes. In doing so, they should center rights-based data collection practices, ensuring community-informed consent, transparency, and equitable compensation for contributors.[23] When direct expert engagement is limited, LLM developers could look to crowdsourcing methods or semi-automatic dataset creation methods, where LLMs generate initial datasets that native speakers refine and annotate manually. However, these methods come with their own challenges and limitations (out of scope for this paper), which must be addressed.[24]

Finally, LLM deployers should fine-tune models for regional linguistic diversity rather than relying on machine-translated text, which often lacks cultural nuance. They should enable an iterative feedback process where community members participate in defining the use and purpose of LLMs tools and refining model prompts.[25]

Refine LLMs through continuous feedback and retraining

Developers should implement feedback loops that allow human reviewers to label and refine training datasets for ongoing model improvement, for instance by incorporating updated examples of flagged content into training datasets. They should then regularly retrain models with newly identified violative content to ensure that LLMs adapt to evolving abuse patterns and emerging risks, with LLM deployers having to do this more frequently at the fine-tuning level (instead of the training one).[26]

LLM developers and deployers can explore using AI to sample enforcement decisions made by both human moderators and automated models to identify inconsistencies, enforcement gaps, or model drift. These could potentially be addressed by LLM developers updating the training data.[27] However, using AI for such purpose should be complimented by human review loops to validate model outputs, particularly for sensitive content.

Finally, LLM deployers should carefully balance precision and recall depending on the type of abuse being moderated.[28] For example, models detecting child sexual abuse material may prioritise recall over precision, as the consequences of missing harmful content (false negatives) outweigh undue content take down (false positives). In

---

21 Mona Elswah, personal communication, July 31, 2024 ; See also Cohere's Aya MLLM https://cohere.com/de/research/aya.

22 Mona Elswah, personal communication, July 31, 2024.

23 For instance, they should recognise dataset creation as labor, as data workers perform essential but often invisible tasks in maintaining and improving NLP systems. Datasets contributors are entitled to fair compensation for dataset contributors, aligning payments with local wage standards and ensuring ongoing informed consent and decent working conditions.

24 Mona Elswah, personal communication, July 31, 2024.

25 Promising examples include examples include developing AI models to detect propaganda in Arabic media or combat misinformation in underrepresented languages (e.g., AraIEval, https://araieval.gitlab.io/). Mona Elswah, personal communication, July 31, 2024.

26 Digital Trust & Safety Partnership. (2024, September). Best practices for AI and automation in trust and safety. (p. 31) Digital Trust & Safety Partnership. https://dtspartnership.org/best-practices-for-ai-and-automation-in-trust-and-safety/

27 Ibid.

28 Ibid.

contrast, misinformation or hate speech detection may prioritise higher precision to reduce over-moderation of legitimate speech. In any case, LLM deployers should analyze flagged false positives and false negatives to adjust moderation thresholds and refine decision-making processes.

Evaluate LLMs rigorously

Effective evaluation, including both quantitative and qualitative evaluation, is critical to ensuring that LLMs perform accurately, fairly, and reliably across diverse use cases and groups. As recommended by the DTSP, LLM evaluation should at minimum include key performance indicators such as precision, recall, false positive and false negative rates, impression rate, time-to-detection, and violative view rate, and fairness across groups.[29]

Furthermore, LLMs should be assessed for fairness by analysing performance discrepancies across datasets that represent marginalised or at-risk groups. If bias is identified, LLM developers should adjust training datasets and model thresholds to improve equitable outcomes, while LLM deployers should adjust these at the fine-tuning level. Continuous bias audits and fairness evaluations should be embedded into LLM development and use work flows.

In any case, developers and deployers should evaluate LLMs models against rigorous, well-defined benchmarks to validate performance and ensure robustness before deployment. Industry standards, regulatory compliance, and context-specific benchmarks should be considered to align LLMs with human rights and legal frameworks. Quality control measures should be regularly updated based on real-world feedback and external stakeholder input, external audits, and error analysis.

Implement internal grievances mechanisms

To ensure accountability in content moderation, LLMs deployers should implement robust internal grievance mechanisms that provide users with clear avenues for human review in appeal and redress specific to harms caused by LLM content moderation. As noted in DTSP's best practices, they should develop user-friendly processes for individuals to appeal moderation decisions, such as content removal, account suspension, or termination.[30] LLM developers should also establish grievance mechanisms for errors and harms at the foundation level.

A corollary to effective grievance mechanisms, at minimum for LLM deployers, is the need to notify users and provide them with detailed explanations when their content is flagged or removed because of an AI system. Upon appeal, a human moderator (or panel) should promptly review the content and the AI's rationale. LLM deployers are encouraged to build an easy in-app appeals interface for users. This aligns with the Santa Clara Principles on Transparency and Accountability in Content Moderation, which list notice and appeal as fundamental to due process in moderation.[31] The EU DSA also requires internal complaint systems and even independent review bodies for disputed cases (Article 21 DSA).

Overall, LLM developers and deployers must ensure that appeal processes are easily accessible, with clear explanations of the reasoning behind enforcement actions. They should provide users with timely responses to their appeals and offer escalation options if they contest a decision.

---

29 Ibid.
30 Ibid.
31 The Santa Clara Principles on Transparency and Accountability in Content Moderation. (n.d.). The Santa Clara Principles. Retrieved from https://santaclaraprinciples.org/

Ensure meaningful transparency

LLM developers and deployers should publish transparency reports accessible to researchers, civil society, and users about how their LLM moderation systems work, how they're used, and how well they perform. [32] In these reports, they should publicly document methodologies used in risk assessments, results, and mitigation strategies, among others. Notably, LLM developers and deployers should disclose how they included stakeholders in their impact assessment process, how they integrated external feedback into product development and/or use, what feedback they discarded, and why. Annex IV of the EU AI Act outlines the minimum technical documentation applicable to an AI system that should be disclosed.[33]

LLM deployers' transparency reports should effectively inform external stakeholders about the content moderation practices and outcomes on their platform. At minimum, reports should include metrics such as the volume of content evaluated by LLMs, the number of removals the LLM triggered, the false positive and negative rates discovered (including via appeals), breakdowns of enforcement by content category and language, and steps taken to improve the LLM's performance.[34] Similar to how social media platforms publish quarterly "community standards enforcement reports," platforms should publish LLM moderation statistics, such as the number of posts that were auto-flagged and of those, how many were mistakes (e.g. overturned on appeal). [35]

LLM developers and deployers should publicly share audit methodologies, findings, and any resulting actions taken with relevant stakeholders, including civil society, affected communities, researchers, and policymakers. LLM developers and deployers must openly address areas where LLMs fall short, and where they've implemented corrective measures to prevent or minimise these shortcomings.

As recommended by BSR,[36] disclosures should occur at various levels, with content tailored to meet the specific needs of different audiences. Corporate-level disclosures are usually conducted annually, following formal reporting standards, and are intended for stakeholders such as investors, civil society organisations, and company analysts. [37]

32 Narayanan, A., & Kapoor, S. (2023, June 26). Generative AI companies must publish transparency reports: The debate about AI harms is happening in a data vacuum. Knight First Amendment Institute. https://knightcolumbia.org/blog/generative-ai-companies-must-publish-transparency-reports

33 European Union. (2024, July 12). Official Journal of the European Union: Annex IV. EUR-Lex. https://eur-lex.europa.eu/legal-content/EN/TXT/HTML/?uri=OJ:L_202401689#anx_IV

34 Narayanan, A., & Kapoor, S. (2023, June 26). Generative AI companies must publish transparency reports: The debate about AI harms is happening in a data vacuum. Knight First Amendment Institute. https://knightcolumbia.org/blog/generative-ai-companies-must-publish-transparency-reports

35 Narayanan, A., & Kapoor, S. (2023, June 26). Generative AI companies must publish transparency reports: The debate about AI harms is happening in a data vacuum. Knight First Amendment Institute. https://knightcolumbia.org/blog/generative-ai-companies-must-publish-transparency-reports; Digital Trust & Safety Partnership. (2024, September). Best practices for AI and automation in trust and safety. (p. 31) Digital Trust & Safety Partnership. https://dtspartnership.org/best-practices-for-ai-and-automation-in-trust-and-safety/

36 Business for Social Responsibility (BSR). (2025, February). Aligning transparency and disclosure practices with human rights responsibilities: Guide 7 of the Responsible AI Practitioner Guides for taking a human rights-based approach to generative AI. https://www.bsr.org/files/BSR-Aligning-Transparency-and-Disclosure-Practices-with-Human-Rights-Responsibilities.pdf

37 Ibid.

Model, product, or system-level disclosures, such as model cards[38] or dataset datasheets,[39] target audiences with specialised AI knowledge and focus on principles of interpretability and explainability. These reports should at minimum include a description of the model's intended use (and non-intended use), its training data (in general terms, noting any major gaps or skews), its performance on key evaluation sets (accuracy, precision/recall, bias metrics, etc.), and its known limitations.[40] For example, if the LLM moderation system encounters difficulties with memes or certain dialects, deployers should communicate this openly. Transparency in such areas helps users and auditors better understand the model's limitations and scope. Relatedly, LLM deployers should ensure transparency regarding their content policies and the mechanisms through which the LLM enforces them. Ideally, they should publicly disclose the guidelines that the LLM has been trained to follow, enabling external observers to assess consistency and accountability.

User-focused disclosures are designed for the general public and aim to present information in a clear and accessible way. For more information about emerging best practices related to documenting and reporting across the LLM and Generative AI value chain, see BSR's Responsible AI Practitioner Guides for Taking a Human Rights-Based Approach to Generative AI, Guide 7.[41]

Finally, as consistent with Article 40 DSA, AI developers and deployers should enable third-party scrutiny, by sharing data with academic and civil society researchers under proper privacy protections. Where possible, they should open-source certain components or datasets, or at minimum use model cards and dataset nutrition labels.[42]

38 Google. (n.d.). Gemma model card. Google AI. Retrieved April 3, 2025, from https://ai.google.dev/gemma/docs/core/model_card

39 Gebru, T., Morgenstern, J., Vecchione, B., Vaughan, J. W., Wallach, H., Daumé III, H., & Crawford, K. (2021). Datasheets for datasets (v8). arXiv. https://arxiv.org/abs/1803.09010

40 Business for Social Responsibility (BSR). (2025, February). Aligning transparency and disclosure practices with human rights responsibilities: Guide 7 of the Responsible AI Practitioner Guides for taking a human rights-based approach to generative AI. https://www.bsr.org/files/BSR-Aligning-Transparency-and-Disclosure-Practices-with-Human-Rights-Responsibilities.pdf

41 Ibid.

42 Belli, L., & Wisniak, M. (2023, August 22). What's in an algorithm? Empowering users through nutrition labels for social media recommender systems. Knight First Amendment Institute. https://knightcolumbia.org/content/whats-in-an-algorithm-empowering-users-through-nutrition-labels-for-social-media-recommender-systems

Algorithmic Gatekeepers: VIII. Preliminary Recommendations

European Center for
Not-for-Profit Law