Algorithmic Gatekeepers:
The Human Rights Impacts of LLM Content Moderation

II. Technical primer

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 their thoughtful 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

II. Technical primer

Content Moderation Landscape

Content moderation

Content moderation is the overall process of monitoring, screening, reviewing, and removing user-generated content in accordance with the digital platform's policies. Social media platforms moderate content by deciding what content remains online and what is removed. This binary "take down/leave up" model can be extended with the option to downrank or demote problematic content or, conversely, to amplify content with recommender systems. Under the EU Digital Services Act (DSA),[1] content moderation "means the activities, whether automated or not, undertaken by providers of intermediary services, that are aimed, in particular, at detecting, identifying and addressing illegal content or information incompatible with their terms and conditions, provided by recipients of the service, including measures taken that affect the availability, visibility, and accessibility of that illegal content or that information, such as demotion, demonetisation, disabling of access to, or removal thereof, or that affect the ability of the recipients of the service to provide that information, such as the termination or suspension of a recipient's account."[10]

Recommender systems

Recommender systems are algorithms that drive what users see online (e.g. on their social media feed). They rank and prioritise information typically by demoting or amplifying content, sometimes personalising content to the user. The DSA defines recommender systems as a "fully or partially automated system used by an online platform to suggest in its online interface specific information to recipients of the service or prioritise that information, including as a result of a search initiated by the recipient of the service or otherwise determining the relative order or prominence of the information displayed."[11]

Trust and Safety (T&S)

Trust and safety is "the umbrella term to describe the teams at internet companies and service providers that work to ensure users are protected from harmful and unwanted experiences."[12] Trust and safety workers typically develop and enforce content policies and tend to operate cross-functionally across policy, operations, product, and engineering teams. Some examples of trust and safety include moderating user-generated content (e.g., hate speech, misinformation, harassment, terrorist content, child sexual abuse material, etc.) or behaviour (e.g., impersonation, coordinated disinformation campaigns).

1 European Commission. (n.d.). The Digital Services Act (DSA). Retrieved from https://commission.europa.eu/strategy-and-policy/priorities-2019-2024/europe-fit-digital-age/digital-services-act_en

Distinctions between algorithmic models

Generative AI

Generative AI systems create new content that resembles its training data, which can include text, image, or video formats.[9] It's an umbrella term that includes LLMs as a type of Generative AI model. As a popular application, generative AI is the dominant feature of chat bots, such as OpenAI's ChatGPT, Anthropic's Claude, or DeepSeek's R1. Generative AI can be distinguished from discriminative AI, which focuses on categorising data and is typically used for classification (e.g., spam filtering or facial recognition). Advanced models, such as OpenAI' GPT-4o and o1 or Google's Gemini, include aspects of both generative AI and discriminative AI as they allow a wide range of tasks to be performed.

Foundation Models / General Purpose AI Systems (GPAI)

Foundation models,[1] also referred to as general-purpose AI models, are trained on vast amounts of largely unlabelled data, which can include text, image, or video formats. Foundation models provide broad knowledge and can perform a wide range of tasks. These models can be adapted for downstream tasks,[2] often through minimal fine-tuning.[3]

Large language models (LLMs)

Large language models (LLMs) are a subset of foundation models trained on vast textual data and capable of processing and generating language.[4] They can generally perform a wide range of tasks and can be adapted for specific purposes.[5]

Multi-modal LLMs (MLLMs)

Multi-modal LLMs are trained on any combination of modalities—text, image, video, and audio—as inputs and outputs. The underlying LLM architecture makes the model capable of understanding and generating language content across modalities.[6][7]

Multilingual language models

Multilingual language models are trained on text data from dozens to hundreds of languages simultaneously, which make them capable of processing and generating inputs and outputs in multiple languages. "Multilingual language models infer connections between languages, allowing them to apply word associations and underlying grammatical rules learned from languages with more text data available to train on (in particular English) to those with less."[8]

LLM developers and deployers

LLM-powered systems for content moderation are becoming more popular, but still seem to be developed primarily by existing dominant companies developing foundation models themselves (i.e. OpenAI, Meta, Anthropic, and Google). With the exceptions of a few newcomers such as the Chinese AI company DeepSeek or the French startup Mistral, companies that don't develop their own LLMs typically use third party foundation model and then fine-tune and/or prompt the model for their own in-house content moderation purposes. Additionally, several smaller AI startups act as intermediaries by fine-tuning. LLMs for content moderation and marketing them to third parties.[2]

There's hardly any available information on who develops and uses LLMs for content moderation and recommender systems, but it appears that the companies who develop foundation models have begun to deploy these systems to moderate content on their platforms. It's important to note the use of LLMs for content moderation is still in its early stages, and the field is largely experimental and unproven. Now, some Trust & Safety (T&S) vendors have started offering these models, which may lead to wider adoption among smaller companies that previously lacked the resources to develop their own machine learning capabilities.[3] However, disclosure about the use and reliability is still lacking.

In February 2024, Meta stated they "started testing Large Language Models (LLMs) by training them on [their] Community Standards to help determine whether a piece of content violates [their] policies. These initial tests suggest the LLMs can perform better than existing machine learning models."[4] They also claim to use "LLMs to remove content from review queues in certain circumstances when [they're] highly confident it doesn't violate [their]policies."[5] Potential use cases for independent researchers include using LLMs for content governance to evaluate the accuracy, bias, and limitations of existing models. Additional potential users could be governments, especially those who moderate internet access. The following sections of this report assesses the human rights impacts of such use cases.

For more information, read BSR's Human Rights Across the Generative AI Value Chain report.[6]

2 Bernard, T. (2023, July 24). *The evolving trust and safety vendor ecosystem.* Tech Policy Press. Retrieved from https://www.techpolicy.press/the-evolving-trust-and-safety-vendor-ecosystem/

3 TELUS International. (2024, April 22). *TELUS International launches Fine-Tune Studio to deliver high-quality datasets that improve the performance, adaptability, and safety of generative AI models.* Business Wire. Retrieved from https://www.businesswire.com/news/home/20240422616380/en/;
Labelbox. (n.d.). *How to build a content moderation model to detect disinformation.* Retrieved from https://labelbox.com/guides/how-to-build-a-content-moderation-model-to-detect-disinformation/

4 Clegg, N. (2024, February 6). *Labeling AI-generated images on Facebook, Instagram, and Threads*. Meta. Retrieved from https://about.fb.com/news/2024/02/labeling-ai-generated-images-on-facebook-instagram-and-threads/

5 Ibid.,

6 Hoh, J. Y., & Nigam, S. (2025, February). *Human rights across the generative AI value chain.* BSR. Retrieved from https://www.bsr.org/en/reports/human-rights-across-the-generative-ai-value-chain

Potential LLM applications for content moderation

Automated content removal

LLMs could be used to automatically identify and/or remove violative content. Proponents claim that GenAI has the potential to perform better than traditional machine learning models and outsourced human reviewers across abuse types.

Content flagging

LLMs could be used to flag potentially violative content to aid human reviewers with detailed explanations. Proponents claim that LLMs can be trained to recognise content violations better than existing machine learning models and human reviewers.

Detecting unusual patterns and behaviours

Multi-modal models could detect patterns of abuse through text, images, or videos. LLMs' language and parsing capabilities could highlight and analyse pattern trends between individual moderation cases. Proponents claim that this could aid human reviewers and investigators by linking cases together to determine patterns of abuse such as fraud.

User notifications and "nudges"

LLMs could inform users on platforms' reasoning behind content governance decisions.[29] For example, LLMs might be able to provide instant, personalised messages to users following the removal of their content and informing them of any account or content action taken against them, and why. Moreover, an LLM-based intervention could prompt a user to reframe a post before publishing it, so that the content doesn't violate the platform's policies or the law.

Real-time moderation at scale

LLMs could improve the scalability of automated moderation beyond the capacities of existing human moderation. Proponents claim that LLMs have the potential to update and moderate content policies in real time. For example, OpenAI claims that "a content moderation system using GPT-4 results in much faster iteration on policy changes, reducing the cycle from months to hours. GPT-4 is also able to interpret rules and nuances in long content policy documentation and adapt instantly to policy updates, resulting in more consistent labelling."[30] There could also be a potential to expand automatic detection of non-English contexts through multilingual models.[31] Another use case could be to improve the quality of user reports at scale, as it is difficult to parse out legitimate reports and appeals. LLMs could help human moderators in this process.

Customised and personalised moderation

By tailoring the pre-trained models to specific content policies through fine-tuning and continued learning, LLMs could potentially enable or improve customised moderation. For example, LLMs could support trust and safety policy development by testing existing policies for consistency[32] or by providing instant feedback to policy changes. LLMs could furthermore make localised content moderation at scale potentially feasible. This would entail an LLM to have customised content moderation rules for respective regions and countries (albeit with implications on access to information and issues related to the splintering of the internet).

Recommender systems

LLMs could potentially analyse each piece of content more accurately and decide how to rank it according to the platform's policies. Incorporating LLMs into recommendation systems could broaden enforcement responses from the binary 'leave up/take down' decisions to alternatives such as downranking or upranking. However, currently, LLMs are too expensive to fully implement in recommendation systems, though they might be embedded within certain subcomponents without the 'chat' interface.

Potential opportunities for LLM Moderation

Accuracy

As the system can perform real-time web searches and stay up-to-date through regular retraining,[4][5] accuracy levels could improve as a result. LLMs are significantly more complex models capable of parsing language, allowing them, in theory, to deliver more accurate assessments of nuanced content compared to traditional machine learning approaches. LLMs might help process content better for classification purposes, for example by better interpreting text or teaching the LLM to improve labelling[7] and improving the ability to classify appropriately (e.g. to conduct screening of hard cases from easy cases),[8] thus supporting human content moderators.

AI companies claim that better accuracy is one of the strongest benefits of LLM moderation. For example, Open AI claims that compared to traditional approaches, LLMs might provide more consistent labels, faster feedback loops, and reduced mental burden.[9] Google and DeepMind researchers released a paper where they claim that "LLMs can achieve 90% accuracy when compared to human verdicts" from a dataset of 50,000 comments.[10] [11] That said, to our knowledge, there haven't been any papers written by independent researchers sharing findings that would validate these claims.

As with traditional machine learning-based content moderation, the accuracy level falls dramatically when these systems moderate content where context is critical, such as hate speech, reclaimed speech, terrorist content, or satire, among others. A researcher argued that accuracy alone is inadequate and deceptive, as it overlooks the critical difference between straightforward and complex cases, while also failing to account for the unavoidable compromises involved in pursuing higher accuracy.[12] Importantly, their accuracy levels vary tremendously between languages, as LLMsand multilingual models perform poorly in non-English, non-dominant languages, known as low-resource languages. Finally, as most machine learning moderation is not currently trained on the policy-text itself, but the results of people labelling policy against that policy text, many inconsistencies emerge from this additional level of abstraction.[13]

7 Dave Willner, personal communication, August 12, 2024.
8 Huang, T. (2024, September). *Content moderation by LLM: From accuracy to legitimacy* [PDF]. arXiv. https://www.arxiv.org/pdf/2409.03219
9 Using GPT-4 for Content Moderation. 2023. OpenAI.
10 Thomas, K., Kelley, P. G., Tao, D., Meiklejohn, S., Vallis, O., Tan, S., Bratanič, B., Ferreira, F. T., Eranti, V. K., & Bursztein, E. (2024). *Supporting human raters with the detection of harmful content using large language models* (p. 1). arXiv. https://arxiv.org/pdf/2406.12800
11 Ibid., Additionally, the authors piloted their "proposed techniques in a real-world review queue yielded a 41.5% improvement in optimizing available human rater capacity, and a 9–11% increase (absolute) in precision and recall for detecting violative content." https://arxiv.org/pdf/2406.12800 P.1
12 Huang, T. (2024, September). *Content moderation by LLM: From accuracy to legitimacy* [PDF]. arXiv. https://www.arxiv.org/pdf/2409.03219
13 Kurzgesagt – In a Nutshell. (2024, February 18). Why alien life would be our doom – The great filter [Video]. YouTube. https://www.youtube.com/watch?v=JMq49FZ5qmY

Accessibility

Developing LLMs by using a pre-trained API lowers the barrier to entry, not only by eliminating the need for machine learning expertise but also by not requiring any specific technical skills, training examples, or even model training. This increased accessibility allows content moderation systems to be built through simple prompt design (also known as 'prompt engineering').

Efficiency

The effectiveness of content moderation systems may improve because LLMs enable the automation of more complex tasks, allowing for greater productivity. Basic forms of automation will likely continue to play a role in areas where they already perform effectively—such as spam detection—due to their significantly lower cost. The vision for LLM integration is to position them as a foundational layer within a stack of content moderation tools. In this model, simpler systems would handle straightforward tasks, while more complex and nuanced issues would be escalated to more sophisticated LLM-based systems.

User notification

LLMs could potentially enable explanations of content moderation decisions in real-time, for example through a conversational agent. This could lead to better quality explanations for moderation decisions.[14] Such use could also include the option to appeal instantly, instead of immediately taking down content. Finally, it could support compliance with emerging legislation on requirements to notify users about content moderation decisions (e.g. art. 17 EU Digital Services Act[15]).

Personalisation

LLMs could be used to personalise content moderation itself. They could also potentially enhance content moderation by predicting user behaviour, ranking content, and addressing complex, unfamiliar data through advanced text analysis. For example, parts of LLMs might be embedded within the recommendation stack, potentially yielding improvements in performance. This does not necessarily imply that the LLM itself would analyse sequences of previously consumed content—recommender systems are already highly optimised for this task. Nonetheless, partial integration could help better predict future user recommendations, such as identifying which posts or media users might engage with next. These models could also help rank content more effectively and predict how users might rate or respond to it, aiding in curating and filtering content based on user preferences. Finally, LLMs could use text embeddings in posts to process and interpret new or less familiar data. This capability could allow them to handle nuances and context in posts more effectively, even when the data is ambiguous or previously unseen. Indeed, LLMs could improve recommendation quality by generating textual representations and drawing on external knowledge to identify and build connections between items and users.[16]

---

14 Huang, T. (2024, September). *Content moderation by LLM: From accuracy to legitimacy* [PDF]. arXiv. https://www.arxiv.org/pdf/2409.03219

15 European Union. (n.d.). *Digital Services Act – Article 17.* EU Digital Services Act. Retrieved from https://www.eu-digital-services-act.com/Digital_Services_Act_Article_17.html

16 Wu, L., Zheng, Z., Qiu, Z., Wang, H., Gu, H., Shen, T., Qin, C., Zhu, C., Zhu, H., Liu, Q., Xiong, H., & Chen, E. (2024, June 18). A survey on large language models for recommendation (Version 5). arXiv. https://arxiv.org/abs/2305.19860

Generalisation

Recent studies have shown that LLMs can have helpful generalisation and reasoning skills, enabling them to handle new tasks and domains with minimal adjustments. Instead of requiring extensive fine-tuning for each task, they could adapt by using instructions or a few examples only. Advanced methods like in-context learning could improve their performance further without specific training.

Additionally, these capabilities could make LLMs promising for transforming recommender systems, prompting strategies like chain-of-thought that could allow LLMs to transparently illustrate their reasoning when making moderation decisions.[17] Instead of delivering opaque or unexplained binary judgments ("leave up" or "take down"), LLMs could articulate clear, step-by-step explanations for each moderation action, outlining the rationale according to specific policies. There's also an argument that the generalisation capability of LLMs could improve accuracy, as they may be better equipped to identify emergent patterns of abuse and edge cases compared to traditional machine learning models, which are typically constrained by labelled training data. However, as we'll show later, this advantage is not absolute, and challenges with accuracy remain.

LLM Methods

Pre-training on large datasets

The pre-training process is the initial phase of training an LLM. The model is trained on a large dataset primarily with the goal of next-token prediction—that is, statistically predicting upcoming words and phrases.[13] While the training may result in a broad understanding of language as a byproduct, developing such an understanding is not the explicit goal. In the context of content moderation, pre-trained models learn to identify patterns, including those associated with violative content such as hate speech or incitement to violence. Pre-trained models are trained on millions to trillions of parameters, which makes the process computationally costly and lengthy. Today's dominant pre-training models (e.g. Open AI's GPT-4.5, Google's Gemini 2.5 Pro Experimental, Anthropic's Claude 3.7 Sonnet and Meta's LlaMa 3.3) have become a sort of infrastructure, known as "foundation models."[14]

Human-in-the-Loop systems (HITL)

Human-in-the-loop systems (HITL) leverage the automated functions of LLMs while ensuring human oversight and involvement within the process.[23] HITL is especially important for tasks that require expertise (e.g. cultural nuance) and judgment within higher risk systems. Related to content governance, an HITL system could for example automatically flag harmful content, but have human moderators make the final decision on content removal.

17 Zhao, Z., Fan, W., Li, J., Liu, Y., Mei, X., Wang, Y., Wen, Z., Wang, F., Zhao, X., Tang, J., & Li, Q. (2024, April 29). *Recommender systems in the era of large language models (LLMs)* (Version 6). arXiv. https://arxiv.org/html/2307.02046v6

Reinforcement learning from human feedback (RLHF)

Reinforcement learning from human feedback (RLHF) is the process of AI development in which humans provide feedback on LLM-produced content.[26] RLHF is used to further train the model to better align with the desired outcome and future decisions. In content moderation, human moderators may provide feedback on LLM classifications of content based on their own expertise and review.

Fine-tuning

Fine-tuning is the secondary process of training an LLM. Using a smaller, specialised dataset, the model is trained for a customised purpose.[15] Tuning or fine-tuning is "the process of adapting a model to a new domain or set of custom use cases by training the model on new data"[16] and it can update the parameters or weights in the core LLM. Technically, this is where RLHF occurs, during which LLMs are trained specifically to produce helpful, context-aware responses rather than simply auto completing text, as they do immediately following pre-training. For content moderation, this means training on labelled data demonstrating violative and non-violative content according to platforms' policies. During this stage, fine-tuned models can be tailored to identify specific violative content such as hate speech, harassment, or incitement to violence. The fine-tuning stage is typically less computationally expensive and quicker. As a result, it is more accessible for smaller teams to fine-tune an already pre-trained model for a specialised function.

Prompt engineering

"Prompt engineering is the process of creating a prompt that is designed to improve performance."[17] It aims to figure out the best way to prompt the model to behave in a certain way. As a simpler method of determining if a post is violative, a user could prompt the LLM with the relevant policy and ask if it's violative.[18]

Other natural language processing methods (NLP)

Natural language processing (NLP) is the broader umbrella term for all aspects of computer-based language processing, including LLMs, or how machine learning can "enable computers to understand and communicate with human language."[19] While LLMs can be helpful in advanced content moderation, other NLP techniques provide the foundation for LLM development and can be used for simpler moderation tasks.

Examples of such tasks include toxicity detection,[20] which analyses data through the assignment of confidence scores to determine "toxicity" (i.e. profanity, insult, graphic, threat).[21] Sentiment analysis produces an emotional score for text data (e.g. tone detection, emotion recognition).[22] LLMs can perform more nuanced sentiment analysis than traditional NLP techniques. Keyword filtering identifies and flags text based on keywords, terms, or phrases. The flagging of specific terms is strengthened by contextual filtering. Text classification assigns text data into predefined categories (e.g. hate speech, spam, and child sexual abuse material). Entity recognition classifies entities within text data to delineate internal context and relationships (e.g. people, locations, dates, organisations). Sequence modelling analyses the order and relationship of words in text. This function is most relevant to LLMs (i.e. predicting the next element in a data sequence).

Algorithmic Gatekeepers: II. Technical primer

European Center for
Not-for-Profit Law