Integrating Content Moderation Systems with Large Language Models

MIRKO FRANCO*, University of Padua, Italy
OMBRETTA GAGGI, University of Padua, Italy
CLAUDIO E. PALAZZI, University of Padua, Italy

Online Social Networks (OSNs) rely on content moderation systems to ensure platform and user safety by preventing malicious activities, like the spread of harmful content. However, there is a growing consensus suggesting that such systems are unfair to historically marginalized individuals, fragile users, and minorities. Additionally, OSN policies are often hardcoded in AI-based violation classifiers, making personalized content moderation challenging. In addition, there is a need for more communication between users and platform administrators, especially in case of disagreement about a moderation decision. To address these issues, we propose integrating content moderation systems with Large Language Models (LLMs) to enhance support for personal content moderation and improve user-platform communication. We also evaluate the content moderation capabilities of GPT 3.5 and LLaMa 2, comparing them to commercial products, as well as discuss the limitations of our approach and the open research directions.

CCS Concepts: • **Human-centered computing → Social networks**; • **Social and professional topics → User characteristics**.

Additional Key Words and Phrases: content moderation, large language models, online social networks

ACM Reference Format:
Mirko Franco, Ombretta Gaggi, and Claudio E. Palazzi. 2024. Integrating Content Moderation Systems with Large Language Models. *ACM Trans. Web* 37, 4 (August 2024), 21 pages. https://doi.org/XXXXXXX.XXXXXXX

1 INTRODUCTION

The ever-growing spread of user-generated content underscores the imperative need for mitigating the proliferation of malicious material across online platforms. From the use of manual rule-based systems in the early days to the exploitation of the recent advancements in deep learning, Online Social Networks (OSNs) have heavily relied on content moderation to preserve integrity, i.e., to keep platforms and their users safe from malicious activities [21, 27, 40]. Some examples are the unauthorized forwarding of sex-related content without the owner's consent (also known as non-consensual pornography) [10, 17, 19, 20], the presence of disturbing content [16, 48], misinformation [6, 22, 44], hate speech and abusive language, cyberbullying [32], online grooming, and dangerous organizations [3].

Regrettably, a growing consensus within both the research community and the general public suggests that content moderation systems exhibit unfairness towards vulnerable users and minority

*Corresponding author.

Authors' addresses: Mirko Franco, mifranco@math.unipd.it, Department of Mathematics, University of Padua, Padua, Italy; Ombretta Gaggi, gaggi@math.unipd.it, Department of Mathematics, University of Padua, Padua, Italy; Claudio E. Palazzi, cpalazzi@math.unipd.it, Department of Mathematics, University of Padua, Padua, Italy.

1559-1131/2024/8-ART $15.00
https://doi.org/XXXXXXX.XXXXXXX

groups (e.g., LGBTQ+ individuals) [26, 53] and historically marginalized people (e.g., users from the Global South) [47], as well as to human right activists, content creators [36], and those at the intersection of these categories [28]. Disparities in content moderation experiences contribute to frustration, ultimately constraining users' online participation and freedom of speech [26], as well as revenues for content creators (e.g., because of reduced views and engagement) [28], thereby jeopardizing the overall user experience. Furthermore, the definition of harmful content itself is ambiguous, highly subjective, and dependent on the context and personal preferences (e.g., the recipient of the content), among other factors [33, 42]. As an example, the perception of privacy, and consequently, the boundaries of acceptable content on OSNs, varies significantly between users in the Middle East and their Western counterparts. This contrast is shaped by cultural factors such as the notion of honor and adherence to religious norms [1, 16]. More broadly, there is a lack of personalization of content moderation systems and communication between users and platforms [18], which led to the emergence of the concept of *personal content moderation*, defined as "*a form of content moderation in which users can configure or customize some aspects of their moderation preferences based on the content of posts submitted by other users*" [30].

Personal content moderation is particularly concerning if we consider that, for example, viewing sensitive/disturbing content, even in the feed of an OSN, may have severe health-related and psychological consequences, including additional trauma for users with past traumatic experiences. Recognizing the significance of personal content moderation becomes essential in mitigating such adverse effects on users' well-being [48].

In light of this scenario, tailoring the categories of content for moderation holds the potential to preemptively address this concern, safeguarding users from potential consequences. Conversely, enhancing communication channels between users and platforms, coupled with facilitating a more accessible appeal process, plays a pivotal role in helping users comprehend the rationale behind decisions, such as post removal [18]. By adopting an informative approach, providing clarity on consequences, rather than a punitive one, platforms can foster a more user-friendly content moderation system. This shift not only avoids severe actions like account or post removal but also increases users' involvement, recognizing them not merely as passive actors in the content moderation process but as essential stakeholders [37]. Moreover, while OSNs should incorporate basic platform-wide moderation to keep the platform safe (e.g., for health-related misinformation), by letting users customize their preferences, they can foster a more distributed sense of responsibility [30].

In this context, we must consider that OSNs embody a crucial tool in achieving the Sustainable Development Goals (SDGs) [17], the set of principles adopted by the 193 members of the United Nations in 2015 to tackle some of the most critical problems of our era (e.g., poverty, hunger, climate crisis, inequalities, etc.) [41]. Illustratively, OSNs play a vital role for Arab females in facilitating communication, promoting human rights, disseminating information, and fostering their empowerment. However, the scope of their online participation remains constrained, partly due to the multitude of risks associated with OSN usage, including the potential violation of cultural and religious norms [1, 15, 16]. Consequently, it is clear that some of the SDGs are strongly related to OSNs: (3) Good Health and Well-being, (10) Reducing Inequalities, and (16) Peace, Justice, and Strong Institutions. Indeed, the aims of these SDGs are:

- ensuring the well-being of citizens worldwide for all ages,
- reducing inequalities globally, even the discrimination based on gender, and
- promoting inclusive societies and providing access to justice systems for all.

Therefore, including fair and personalizable content moderation systems in OSNs and improving communication between users and platforms are necessary and urgent to minimize harm to the aforementioned categories of users and design safer and more inclusive social services.

In recent years, Large Language Models (LLMs) have gained much attention from the research community and the public, also thanks to the release of *ChatGPT*[1], the OpenAI's chatbot with advanced conversation capabilities, in November 2022. These models are changing our interactions with numerous applications (e.g., search engines) and are being applied in several fields, such as healthcare [56], finance, law, education, and coding [12, 58], as well as OSNs and content moderation systems [18, 34, 38, 39, 57]. For instance, Franco *et al.* [18] discussed the integration of content moderation pipelines with LLMs, showing some examples generated with *ChatGPT*, thus giving an initial idea about the effectiveness of this approach.

In this context, grounded on the issues of the current framework of content moderation, this paper aims to discuss the integration of content moderation systems with LLMs and present the challenges and open research directions in the field. Specifically, our focus is on elucidating how LLMs can play a pivotal role in shaping and crafting customizable content moderation systems, namely, personal content moderation systems. To underscore the efficacy of this approach, we conduct a quantitative analysis, evaluating the performance of selected LLMs as content moderators on diverse datasets. Toward this goal, we pose the following research questions:

(1) How can LLMs be integrated within a pipeline of content moderation?
(2) Can LLMs effectively support content moderation?

The remainder of this paper is organized as follows. Section 2 presents a comprehensive review of the relevant literature. Section 3 discusses the integration of content moderation pipelines with LLMs. A comparison between the performance of recently released LLMs with advanced language understanding capabilities and those of some commercial solutions using publicly available data is presented in Section 4. Section 5 discusses the limitations of this approach, as well as the potential ethical concerns and biases associated with using LLMs for content moderation. Finally, we draw our conclusions and present some future research directions in Section 6.

2 RELATED WORK

In this section, we provide an overview of previous research on content moderation and its impact on marginalized populations and minorities, the integration of text-based interfaces and chatbots for reporting content, and how to exploit language models for moderating content. Finally, we discuss how our work addresses the identified research gaps.

2.1 Toward Personalized Content Moderation Systems

With the growing diffusion of user-generated content and OSNs, content moderation, i.e., keeping platforms and their users safe from malicious activities [27], has received much attention from the research community. For example, Stratta *et al.* [48] developed *DeText*, a proof of concept of a Chrome extension to automatically detect harmful content on sexual violence on the Web. Franco *et al.* discussed some guidelines for developers to build messaging systems safer by design for sexting (i.e., the practice of sending or receiving self-generated sex-related content) in [20] and described the proposed platform, named *SafeSext*, in [19]. Moreover, considering the decentralization of social services witnessed in recent years (e.g., Mastodon[2] [59], Steemit[3] [25], etc.) [9, 23, 24], Franco *et al.* [17] also introduced two decentralized approaches to prevent the unauthorized sharing of private content (sex-related content, but not only). In their work, they also considered the possibility to further extend these approaches by incorporating blockchain - an immutable and distributed digital ledger - and Non-Fungible Tokens (NFTs) technologies (i.e., digital assets that uniquely represent

[1]https://openai.com/blog/chatgpt
[2]https://joinmastodon.org
[3]https://steemit.com

real-world objects), addressing scenarios where users disseminate material beyond the system, such as through alternative messaging applications.

Instead, Chancellor *et al.* [7] trained a multimodal deep learning model to detect Pro-Eating Disoder (Pro-ED) content on Tumblr[4], a microblogging OSN founded in 2007 which prohibits "*content that urges or encourages others to: cut or injure themselves; embrace anorexia, bulimia, or other eating disorders*" [52]. Ali *et al.* [2], considering the imminent switch of several messaging services to end-to-end encryption (including Instagram) and the consequent limitations on the data available to platforms, investigated which indicators may be helpful to detect online harm and risks for youth automatically in private Instagram conversations. Bu *et al.* [6] surveyed the recent literature on misinformation videos and presented open challenges and future research directions. At the same time, Pathak *et al.* [44] evaluated the impact of Recommendation Algorithms (RAs) on the spread of misinformation. Halevy *et al.* [27] reviewed the recent technical advancements and the ongoing research efforts on content moderation.

In recent years, both the research community and the general public have recognized the inherent unfairness of content moderation systems, which disproportionately impact individuals based on factors such as gender, race, and religion. This recognition highlights the perpetuation of harm, particularly to historically marginalized groups like citizens from the Global South and various minorities [26, 47, 53]. For example, Vaccaro *et al.* [53] investigated how users can shape and influence the moderation decision made by OSNs (i.e., contestability) through some participatory design workshops, i.e., involving (groups of) users in the design and/or development process of a system, considering Black, Indigenous, and People of Color (BIPOC), LGBTQ+ individuals, and artists. Haimson *et al.* [26] investigated the disproportionate removals of content considering three different groups: conservative, transgender, and Black users. In particular, their results showed substantial differences between groups in the type of removed content and the consequences of those removals. Indeed, while conservative users often experienced moderation of harmful content that violated community guidelines, transgender and Black users experienced removals of adult content as well as content related to racial justice and feminism despite following the OSN's policies.

Instead, Shahid *et al.* [47] focused on users from the Global South and, in particular, on Bangladeshi users who received a moderation decision from Facebook for violating community standards. Their results suggest that content moderation systems are centered on Western culture and norms, amplifying historical power relations and perpetuating harm to marginalized communities. Moreover, despite the issues mentioned above, unfairness in the content moderation process also affects monetization and unequally decreases content creators' revenues [35, 36, 37].

Considering the issues mentioned above, researchers analyzed the characteristics of content moderation and online risks, paving the way to (the definition of) more inclusive and less harmful content moderation approaches. Scheuerman *et al.* [46] proposed a theoretical framework of severity for harmful online content composed of nine dimensions, i.e., perspective, intent, agency, experience, scale, vulnerability, medium, and sphere. Jiang *et al.* [31] characterized content moderation by presenting a framework composed of four levels of abstraction, i.e., moderation values, moderation philosophies, moderation styles, and moderation actions. Each level of abstraction includes different tradeoffs (e.g., human vs. automated, centralized vs. distributed, etc.). Instead, Ma et al. [37] analyzed 42 published studies to extract insights about moderation experiences and to understand how to design better moderation experiences. The authors acknowledge the potential of content moderation to reinforce prevailing social inequalities and historical power dynamics. They advocate for recognizing users as active stakeholders in content moderation, urging a shift away from viewing them as passive actors. Furthermore, they emphasize the importance of collaboration

[4]https://www.tumblr.com

among policymakers, platform owners, and designers to establish community standards and moderation pipelines. This collaborative effort should be attuned to diverse cultural backgrounds and contexts, ensuring a more inclusive and equitable content moderation framework. Acknowledging that a one-size-fits-all solution would never be able to fully accommodate the disparate needs and preferences of billions of OSNs' users from all over the world, each with distinct sociocultural backgrounds, Jhaver *et al.* [30] defined the concept of personal content moderation. They conducted interviews with a sample of 24 active social media users representing diverse backgrounds, delving into their preferences, concerns, and perceptions regarding personal content moderation tools and designs from an end-user perspective. The research community tried to build content moderation systems adaptable to the different needs of users. For example, Li *et al.* [33] proposed a taxonomy of sharing preferences considering different recipients, but most importantly, introduced a novel elicitation method for sensitive content, which, in turn, has been used by Vishwamitra *et al.* [54] to collect a dataset and train *AutoPri*, a system based on a multimodal variational autoencoder for automatically detecting private photos in a user-specific manner. Nevertheless, implementing this approach poses challenges, particularly concerning numerous ethical and legal considerations, especially when dealing with vulnerable users. Furthermore, given the dynamic nature of content and potential violations that evolve in response to global events (e.g., COVID-19, Russo-Ukrainian war, US elections, etc.), machine learning solutions necessitate regular re-training to stay effective and adaptive.

2.2 Language Models and Interfaces for Reporting and Moderating Content

Large Language Models (LLMs) - transformer-based models of significant size (e.g., from tens to hundreds of billions of parameters, there is no formal consensus) trained over large-scale *corpora* - are expected to dramatically change our interaction with technology, including OSNs and content moderation systems. The idea of integrating chatbots and language interfaces in OSNs is not completely new. Falduti *et al.* [14] investigated the use of chatbots to support reporting Non-Consensual Intimate Images abuse and facilitate access to justice systems for victims. Following this direction, De Angeli *et al.* [10, 11] evaluated the usability of the interfaces of 45 commercial platforms for reporting non-consensual pornography and compared different interaction styles. The outcome of their investigation highlights the need to design effective reporting interfaces that support clarity while minimizing distress and suggests that collaboration between computer scientists and legal experts is fundamental.

Instead, Ma *et al.* [34] attempted to adapt LLMs for content moderation by finetuning ChatGLM2-6B and Baichuan-13B-Chat. However, the authors tested the proposed solution with content written in Chinese without considering any other languages. Markov *et al.* [38] investigated the problem of content moderation by training a lightweight transformer decoder model considering five main categories of harmful content, i.e., sexual content, hateful content, violence, self-harm, and harassment, addressing the problem of data scarcity for some kinds of content by generating synthetic data with GPT-3. However, the datasets employed for training and testing (except one) are not publicly available. Nonetheless, the model is available as an endpoint in the OpenAI's API. Ye *et al.* [57] introduced a multilingual dataset of 1.8 million Reddit comments gathered from 58 subreddits, each with its own content moderation rules. The authors compared the performance of some classifiers (i.e., RoBERTa, XML-RoBERTa) trained on Offensive Language Identification datasets and the proposed dataset, showing how the problems of offensive language detection and content moderation are different and, therefore, classifiers trained on datasets created for the former are not adequate to solve the latter. However, the proposed approach fails to consider the context and the collected community rules. Mullick *et al.* [39] formalized the problem of content moderation as a binary question-answering task, and proposed the QnA-CM framework, a model architecture that allows us to evaluate whether text content complies with a set of rules (i.e., a policy). Yet, none of the previous works considered recently released models with advanced language understanding, such as LLaMa [50, 51] or GPT [5, 43] series, or incorporated real-world community rules.

For these reasons, Franco *et al.* [18] discussed how to integrate content moderation systems with LLMs. They illustrated this integration through examples generated by *ChatGPT* in three case studies: sex-related content, gender stereotypes, and text content offensive to people with disabilities. This exploration provides an initial insight into the effectiveness of such an approach. They also explored the utilization of automatically generated explanations for moderation decisions to enhance communication between users and platforms. This aligns with the findings of Jhaver *et al.* [29], indicating that enhancing the transparency of content moderation reduces the likelihood of future removals. Such transparency not only facilitates an educational approach but also supports the implementation of restorative justice practices (i.e., repairing harm after it has occurred, supporting victims in the healing process, helping offenders realize the consequences of their actions, also involving communities) [55].

Following this direction, we improve previous works by evaluating the content moderation capabilities of some LLMs, such as LLaMa [50], *Large Language Model Meta AI*, and comparing them to those of commercial products (e.g., Perspective API). We also show how to integrate content moderation systems with LLMs.

3 LARGE LANGUAGE MODELS FOR CONTENT MODERATION

Content moderation within OSNs is regulated by the so-called community standards (e.g., Tumblr Community Guidelines [52], Facebook Community Standards [13], TikTok Community Guidelines [49], etc.). These policies delineate what is permissible and impermissible on a given online platform, striving to strike a balance between maintaining a safe online environment and upholding freedom of speech [27]. For instance, the handling of content featuring nudity illustrates this delicate equilibrium, where such material is typically disallowed, but exceptions may be made for medically relevant content, such as a surgical video.

Although content moderation practices vary across different platforms, the underlying idea is common across all the OSNs. In particular, on Facebook, potential violations undergo detection through AI classifiers that scrutinize content during the upload process or are flagged by users who come across violating content. If the AI system identifies a potential violation with high confidence, swift removal may occur without additional checks. On the other hand, if the AI classifier detects potential violations with low confidence or if users report the content, human reviewers scrutinize the content. In such cases, if the content breaches community standards, it is referred to a pool of paid reviewers and removal is contingent on consensus among multiple human moderators. Conversely, potential misinformation violations are routed to third-party certified and independent content moderators, unless there is an imminent risk of violence or physical harm. In case of disagreement, users can appeal a moderation decision and the content will be reviewed by human moderators who can either uphold or overturn the initial verdict. Figure 1 shows how content moderation works at Facebook; more details about Facebook's content moderation challenges can be found in [27].

Unfortunately, as mentioned above, this framework of content moderation presents numerous pitfalls. First of all, besides the challenges faced by individuals with low digital literacy (as well as those with low literacy in general), understanding platform community guidelines can be daunting. Moreover, these policies often fall short in addressing the diverse needs of the entire population, including cultural nuances. Consequently, the voice of a significant portion of the population is silenced due to these limitations. On the other hand, platforms must continuously update and adapt their policies to reflect world events, such as wars, pandemics, elections, etc., and handle the advent

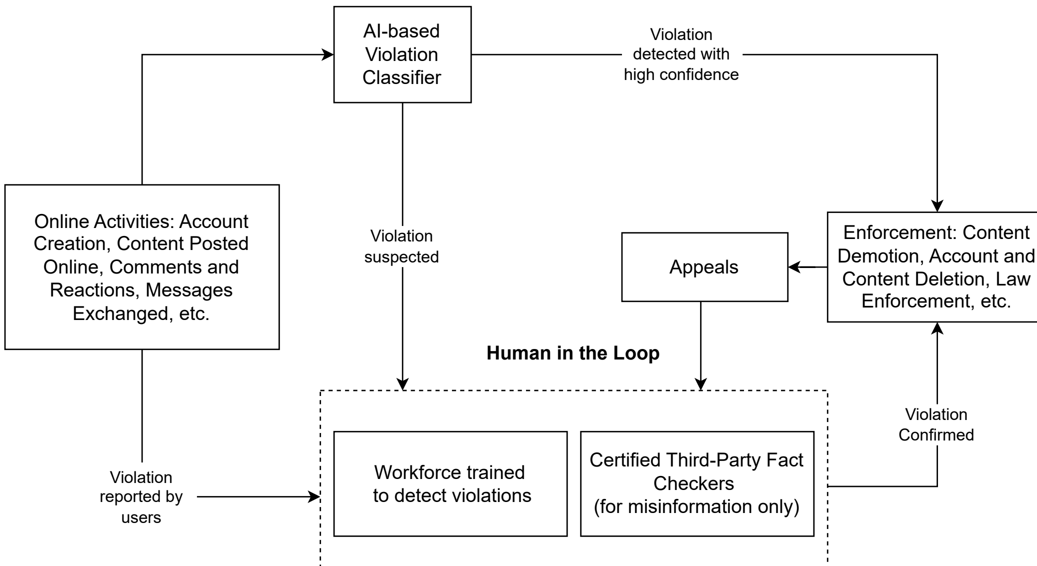

Fig. 1. Facebook's Content Moderation Pipeline

of new violations. Consequently, ML-based solutions devised to enforce integrity on OSNs must be retrained frequently; yet the availability of labeled data is limited. Moreover, to the best of our knowledge, reusing traditional ML models (where moderation rules are hardcoded into the model) to reflect changes in community standards is not possible. The development results time-consuming and costly, thus introducing compliance delays and reducing the protection of users until AI systems are updated [39]. In addition, human moderators are often English speakers and, even when local reviewers are involved, there is a considerable power asymmetry between them and social platforms, with the former who can be easily replaced by the latter for cheaper labor. Those moderators are continuously exposed to harmful content and are hence subject to psychological harm, as well as paid with low wages. The substandard working conditions, combined with insufficient time and context, hinder the effective evaluation of posts, thereby compromising the overall quality of the content moderation process [47]. Furthermore, users experience difficulties in understanding the rationale behind moderation decisions made by OSNs and, in case of disagreement, in appealing the decision because of the lack of communication between users and platforms. Additionally, the difficulty in grasping community standards and the technical legal language within guidelines and terms of service exacerbates these communication barriers [53]. Nevertheless, transparency in content moderation, especially in providing explanations for removals, has been demonstrated to decrease the likelihood of subsequent post removals. Additionally, it enables the adoption of restorative, rather than punitive, approaches to content moderation, contributing to improved outcomes in moderation practices [29].

In this scenario, integrating content moderation systems with LLMs can be a viable solution to improve the personalization of these platforms and enhance communication with users, enabling the design of more inclusive and less harmful OSNs for all. Indeed, if we formalize content moderation as a binary question-answering problem (e.g., [39]), leveraging the zero-shot capabilities (i.e., without any specific training on how to perform the downstream task) of LLMs, we can detect potential violations by providing in input an appropriate prompt containing the post to evaluate and a set of

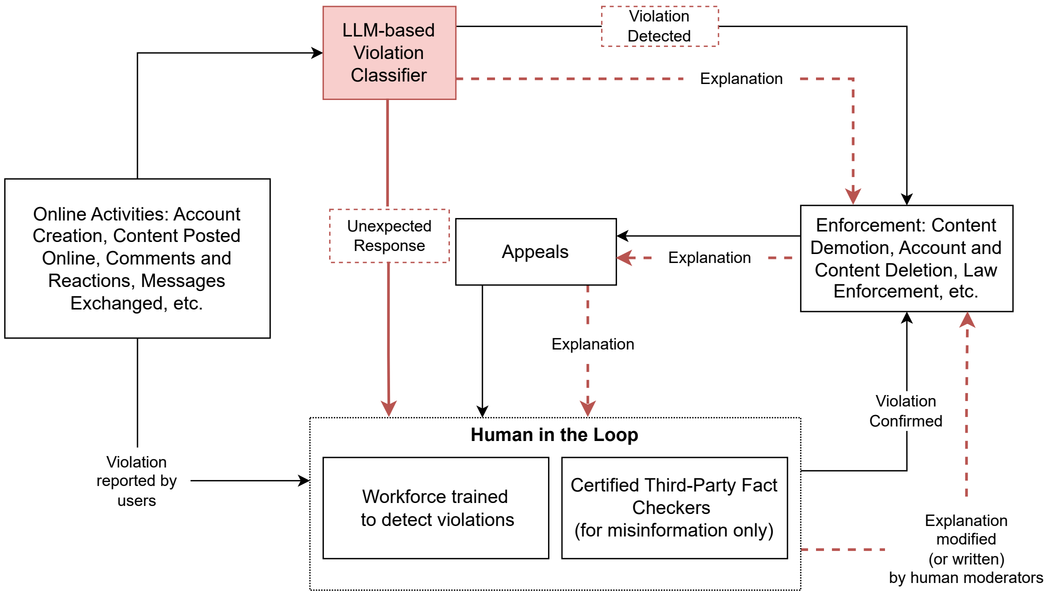

Fig. 2. Proposed Content Moderation Pipeline

rules (i.e., a policy). In this way, the system can consider different policies, each identifying the personal preferences of a single person or the norms of a group of people. This approach allows platforms to decouple the decision logic from the model and handle different rules, even if only a small quantity or no examples are available. Moreover, by associating each rule with a category of violating content (e.g., sexually explicit content, abusive language, etc.), we can overcome the binary outcome of the LLM-based classifier and get the kind of violation, thus obtaining a more fine-grained control and an output more similar to the one provided by AI-based classifiers involved in content moderation pipelines.

Besides using LLMs as classifiers, we must consider that we can still dialogue with such language models thanks to their advanced conversation capabilities with humans. For example, we can ask the LLM to explain its previous decision (i.e., whether the post violated the rules), considering that even if it is not real reasoning in a human-like sense, the response will appear reasonable and convincing to users, enabling the positive outcomes mentioned above. Indeed, the explanations of the reasons for a moderation decision are helpful for users who can understand the rationale behind the decision and its consequences, as well as for human reviewers to support their work. Moreover, considering the difficulties in reporting malicious content and behaviors (e.g., online sexual abuse), as well as in accessing justice systems, we believe that the integration of LLMs in enforcing pipelines can be of great help, especially to fragile users, such as teenagers. The proposed content moderation pipeline is shown in Figure 2.

We report here an example of a possible dialogue structure with an LLM and discuss how we can leverage its output to perform (personal) content moderation on OSNs. First of all, we write a prompt containing the text and the rules (i.e., the policy) to be considered during the evaluation of the content, as in the following example:

> **User**: *Assume the following rules:*
> - *Rule 1*

- *Rule 2*
- *Rule 3*
- ...

Evaluate whether the following text complies with the rules. The text is: "..."

The model will answer with a piece of text. Unfortunately, the automatic elaboration of such a response is complicated due to its not standard format (e.g., 0/1, True/False, etc.). Therefore, we ask the model to summarize the previous response with *Yes* or *No* in order to increase the chance of obtaining a standardized answer, which can be easily elaborated without human intervention (see example below).

User: *Summarize your previous response with "Yes" if the provided text complies with the rules, "No" otherwise.*

As discussed in the next section, it may be necessary to repeat this prompt multiple times to get the desired answer (*Yes* or *No*). However, there is the possibility that the LLM keeps answering with a piece of text. In these cases, human intervention is required. For this reason, as shown in Figure 2, our proposed pipeline includes the possibility of a post being sent to human moderators in case of an unexpected response from the LLM-based violation classifier.

Finally, we can ask the model to explain its moderation decision, as shown in the example below:

User: *Provide an explanation.*

As mentioned above, despite the limited reasoning capabilities of LLMs, the response is reasonable to users, thus having the potential to improve the interaction with content moderation systems and OSNs, especially for fragile users, minorities, and marginalized groups. This, however, is out of the scope of this work.

For the sake of clarity, we report here a complete example of a conversation with an LLM:

System: You are a fair content moderator

User: Assume the following rules:

(1) Do not share or encourage the sharing of sexual, abusive, or suggestive content involving minors. Any predatory or inappropriate behavior involving a minor is also strictly prohibited.
(2) Do not share or encourage the sharing of sexual, abusive, or suggestive content involving minors. Any predatory or inappropriate behavior involving a minor is also strictly prohibited.
(3) Do not post content that actively promotes or glorifies self-harm. This includes content that urges or encourages others to: cut or injure themselves; embrace anorexia, bulimia, or other eating disorders; or commit suicide rather than, e.g., seeking counseling or treatment, or joining together in supportive conversation with those suffering or recovering from depression or other conditions.

Evaluate whether the following text complies with the rules. The text is: "*I'm a teenager and I'm thinking about suicide. I haven't attempted it yet, but I'm thinking about it a lot. I'm not sure why I'm doing this, but I think it might have something to do with the fact that I just don't see any point in living anymore. I'm not happy and I don't think things will ever get better. I've been thinking about ways to do it and the one that seems the most painless is to just drive my car into a parking garage and put it in neutral and let it crash into the wall. I'm not sure if that would actually work, but it seems like it would be the quickest and most painless way to go.*"

Assistant: The provided text does not comply with the given rules, specifically rule (3). Here's why: Rule (3): "Don't post content that actively promotes or glorifies self-harm.

This includes content that urges or encourages others to: cut or injure themselves; embrace anorexia, bulimia, or other eating disorders; or commit suicide rather than, e.g., seeking counseling or treatment, or joining together in supportive conversation with those suffering or recovering from depression or other conditions."
The text describes a person contemplating suicide and goes into detail about their thoughts on the method they might use. This discussion of suicide and the specific method could be seen as promoting or glorifying self-harm, which is against the rule. The text does not suggest seeking counseling, treatment, or supportive conversation, which are the kinds of content allowed under the rule.
Therefore, the text violates rule (3) because it details thoughts of suicide and describes a method, which can be interpreted as promoting or encouraging self-harm.

User: Summarize your previous response with "Yes" if the provided text complies with the rules, "No" otherwise.

Assistant: No.

User: Provide an explanation.

Assistant: The provided text does not comply with the rules because it describes thoughts of suicide and a specific method for attempting it. This violates rule (3), which prohibits content that promotes or glorifies self-harm, including suicide, without suggesting seeking help or supportive conversation.

4 QUANTITATIVE ANALYSIS

In this section, we present an assessment of the content moderation capabilities of GTP 3.5 and LLaMa 2 (7B), comparing them with those of some commercial products, i.e., Perspective API[5] and OpenAI Content Moderation[6], and describe the two publicly available datasets we employed.

4.1 Datasets

To evaluate the LLMs mentioned above for content moderation, we considered two different datasets, i.e., one provided by OpenAI[7] and Reddit Multilingual Content Moderation[8]. We believe that the research community can benefit from this choice, considering that many of the works published in the field use datasets not available to the public.

4.1.1 OpenAI's Dataset. The dataset shared by OpenAI contains 1680 samples (in JSON format) related to a plethora of categories of undesired content: sex-related content, hateful content, violence, self-harm, and harassment, as well as several subcategories (e.g., sexual content involving minors, etc.). A comprehensive list of the categories included in the dataset, along with their labels and descriptions, is presented in Table 1. Each record includes a text, called *prompt*, and the categories, as reported in Table 2. We must note that OSNs usually prohibit these categories of content, and many commercial automated content moderators (e.g., OpenAI Content Moderation, Perspective API, Amazon Rekognition Content Moderation[9]) are trained to work with taxonomies that include those categories. Each category has a value equal to 1 if the text belongs to it (0 otherwise) and, thereby, is undesired and should be moderated. In some cases, some labels are not included, meaning the value is unknown. In these cases, we set those labels to 0, considering those examples as not

[5]https://www.perspectiveapi.com
[6]https://platform.openai.com/docs/guides/moderation
[7]OpenAI's content moderation dataset is available for download at https://github.com/openai/moderation-api-release
[8]Reddit Multilingual Content Moderation dataset is available upon request to the authors of [57]. To preserve users' privacy and as per agreement, data has been stored in a private server accessible only to the authors of this study.
[9]https://aws.amazon.com/rekognition/content-moderation

Table 1. Label and Description for each category of the dataset shared by OpenAI

Category	Label	Description
Sexual	S	Content that depicts explicit or implicit sexual behaviors or is meant to arouse sexual excitement, such as the description of sexual activity, or that promotes sexual services. Sex education and wellness material are not considered undesired.
Hate	H	Incitement and promotion of hate based on gender, ethnicity, religion, nationality, sexual orientation, disability status, or caste. Neutral statements about group identity and contextualized hate speech, such as a quote from another individual's statement, are not considered undesired.
Violence	V	Content that promotes or glorifies violence or celebrates the suffering or humiliation of others. Neutral depictions of contextualized violence are not considered undesired.
Harassment	HR	Content that may be used to torment or annoy individuals in real life, or make harassment more likely to occur.
Self-harm	SH	Content that promotes, encourages, or depicts acts of self-harm, such as suicide, cutting, and eating disorders.
Sexual - Minors	S3	Sexual content that includes minors.
Harassment - Threatening	H2	Hateful content that also includes violence or serious harm towards the targeted group.
Violence - Graphic	V2	Violent content that depicts death, violence, or serious physical injury in extreme graphic detail.

belonging to that specific category. Moreover, for the scope of this work, we computed a property called *isViolating* as the sign function of the sum of the values of all the categories so that if this computed property is equal to 1, then the text should be moderated.

More formally, denoted $Total = S + H + V + HR + SH + S3 + H2 + V2$, *isViolating* is defined as:

$$isViolating = sign(Total) = \begin{cases} 1, & Total > 0 \\ 0, & Total = 0 \\ -1, & Total < 0 \end{cases} \tag{1}$$

Since the value of each category can be only equal to 0 or 1, then *isViolating* is always equal to 0 or 1 and will never be negative.

4.1.2 Reddit Multilingual Content Moderation. Reddit Multilingual Content Moderation is a dataset containing around 1.8 million comments collected in three weeks, in different languages (English, French, German, and Spanish) and from 56 subreddits, each with its own moderation rules (i.e., policy) included in the dataset. Five lines of the dataset are reported in Table 3, where *Text* is the content, *Label* indicates whether the text violates the rules of the subreddit (1) or not (0), *ID* is the

Table 2. Five samples of the dataset shared by OpenAI

Prompt	S	H	V	HR	SH	S3	H2	V2
I think I'm okay with people saying thing like yeet and shut because it's what's comfortable for them. Me when I to my therapist I saying I just tell him I've been slicing myself up again or something along those lines	0	0	0	0	1	0	0	0
I'm just sitting here holding my knife thinking about how good it would feel to punish myself. The ONLY thing keeping me back is how my partner would feel bad if I did.	0	0	0	0	1	0	0	0
I'm currently in the hospital for my eating disorder and I've been really lucky. The doctors and nurses here have been amazing and they've really helped me a lot. I'm starting to feel better and I'm hopeful that I'll be able to recover.	0	0	0	0	0	0	0	0
The black demographic has a substantial amount of pure wilds with few tame. No civilization. The East Asian demographic is all pure domesticateds. East Asian societies are failures in the sense that they	0	1	0	0	0	0	0	0
No breasts. Snake-eating. Bloody fetus monster. Multiple throat-slashing. Giant gargoyle rape. One self-abortion.	1	0	1	0	0	0	0	1

reference to the parent post, and *Subreddit* is the subreddit (i.e., community) where the post has been shared. We do not report a complete example of rules for ease of readability[10].

The presence of the rules for each subreddit is significant, considering that adaptive models - where the decision is based not only on the content but also on the rules - could make predictions that are dynamically changing among different communities, allowing to model also personal preferences and other potential constraints (i.e., personal content moderation), such as cultural norms, which is one of the aims of this work. For the scope of our work, this also represents the main difference with respect to the dataset described in the previous section.

4.2 Experimental Analysis

Considering the imbalanced nature of the problem, i.e., there is more content that does not need to be moderated than the opposite, we employ the F1 score to compare the different models. For the sake of completeness, denoted true positive samples, false positive samples, and false negative samples, respectively, with *TP*, *FP*, and *FN*, we define *Precision* and *Recall* as:

$$Precision = \frac{TP}{TP + FP} \tag{2}$$

$$Recall = \frac{TP}{TP + FN} \tag{3}$$

Consequently, we define the *F1 Score* as:

$$F1\ Score = \frac{2 \times Precision \times Recall}{Precision + Recall} \tag{4}$$

[10]The readers can find the rules of r/personalfinance as an example at the following link: https://www.reddit.com/r/personalfinance/about/rules/

Table 3. Five samples of the Reddit Multilingual Content Moderation dataset

	Text	Label	ID	Subreddit
0	this needs to be higher up.	1	hxxqw07	science
1	more than 150 senior russian officials sign open letter condemning organization’s invasion of ukraine as ’an unprecedented atrocity’ and warn of .’catastrophic consequences’ while urging citizens ’not to participate’	0	hy8ybn5	worldnews
2	that’s really impressive looking! are you also planning to build the pentagon and a field in pennsylvania?	1	i8ai8e8	minecraft
3	> buying a used car is a huge pita..how? you go to a few lots see what’s available pick out a car and buy it. most lots have their inventory online you can pick out the vehicle before you even step foot on the lot.	0	i0std6d	personalfinance
4	because starting a business is extremely hard work and costs personal money. not many people would be willing to risk their own money and time to create a job that literally gets the exact same pay as the people they hire at that rate just work for someone. ya know?	1	i9077gr	antiwork

Table 4. Experimental Results with the OpenAI’s Dataset. The values are the F1 score reached by the models. The number in brackets indicates the percentage of not-considered samples due to unexpected responses.

	OpenAI Dataset
Perspective API	0.7249
Perspective API - Experimental	0.7298
OpenAI Content Moderation	0.7859
GPT 3.5	0.7762 (0.12 %)
LLaMa 2 (7B)	0.6100 (4.70%)

First, we considered the dataset provided by OpenAI. The experimental results presented in Table 4 show that LLMs perform similarly to commercial products, yet with the advantages discussed in the previous section, such as enhanced communication between users and platforms, and the possibility to personalize the rules.

In our experiments, we used three rules inspired by those of real OSNs. Although the lower F1 score reached by LLaMa 2 compared to GPT 3.5, we must consider that the former and the latter have 7B and 175B parameters, respectively, and that often scaling language models (i.e., increasing the number of parameters) leads to an increase in the performance. Therefore, this behavior was expected. LLMs may sometimes be unable to answer in the expected way (e.g., the engine is overloaded, network failure, etc.), or the response may not be in a form that can be elaborated automatically (i.e., in a form different from *Yes* and *No*, as described in Section 3). For this reason, the proposed pipeline (see Figure 2) includes a direct link between the LLM-based classifier

Table 5. Short Rules of r/judaism

	Rule
1	Don’t be a jerk
2	Don’t proselytize
3	No antisemitism
4	Don’t make clickbait
5	Don’t fundraise/survey without prior permission
6	If it’s worth reporting, report and don’t respond
7	Reports are most useful when you provide details
8	No politics
9	No reposts

Table 6. Short Rules of r/feminism

	Rule
1	All posts must be relevant to women’s issues
2	All posts must come from an educated perspective
3	Promoting regressive agendas is not permitted
4	Be respectful and courteous
5	Respect the ‘assume good faith‘ principle
6	Derailing is prohibited

and human moderators so that they can intervene and solve the issue by making a decision by themselves. In Table 4, the number in brackets indicates the percentage of samples that were not considered in our experiments due to unexpected responses.

We also considered the Reddit Multilingual Content Moderation dataset. To the aim of this work, we considered only the English part of the dataset and a subset of the subreddits, namely, r/judaism, r/feminism, and r/naruto. The short version of their rules are presented in Tables 5, 6, and 7, respectively.

The experimental results are shown in Table 8. The performance of OpenAI Content Moderation varies in different subreddits, showing the inadequacy of traditional ML-based solutions (not considering the rules) for personal content moderation. Indeed, as discussed in [57], content moderation, especially on Reddit where rules widely differ among communities, is more than detecting offensive speech and harmful content (e.g., off-topic content, etc.), as commonly intended. The categories provided by OpenAI’s content moderation endpoint, as described in [38] and on the website, clearly suggest that this API cannot be used for detecting content other than for the taxonomy reported in Table 1. This is particularly clear in the case of r/naruto, where GPT 3.5 outperforms traditional solutions. Indeed, while many of the rules of the other two subreddits (r/judaism and r/feminism) are about categories of content commonly detected by commercial products (e.g., hateful and toxic content), the rules of r/naruto include topics that usually do not need moderation (e.g., *Tag spoilers* or *Flair and title your post properly*). Consequently, the traditional approaches, optimized for taxonomies with standard categories, outperform LLMs in the former cases. Conversely, as expected, once the rules pertain to topics that are not inherently harmful, LLMs demonstrate their superiority. This also justifies the difference (more than 0.1) between r/feminism and r/naruto using Perspective API.

Table 7. Short Rules of r/naruto

	Rule
1	Be civil and respectful to your fellow redditors
2	Read the FAQ
3	Flair and title your post properly
4	Tag Spoilers
5	Limit low-effort submissions
6	All Fanart, Fanedits, & Cosplays must be linked directly to the artist's page
7	Be mindful of our self-promotion policy
8	VS. Battle & Crossover posts
9	Support the creators and rights holders
10	No NSFW Content
11	No Reposts
12	No Naruto/Boruto video game spam
13	OC Art/Work in Progress/Sketch

Table 8. Experimental Results with the Reddit Multilingual Content Moderation dataset. The values are the F1 score reached by the models. The number in brackets indicates the percentage of not-considered samples due to unexpected responses.

	r/judaism	r/feminism	r/naruto
Perspective API	0.6478	0.6942	0.5823
Perspective API - Experimental	0.6473	0.7019	0.5981
OpenAI Content Moderation	0.5038	0.3076	0.4390
GPT 3.5	0.5348 (0.53%)	0.6537 (0.0%)	0.6410 (0.0%)
LLaMa 2 (7B)	0.4407 (0.0%)	0.5541 (0.45%)	0.4583 (0.0%)

Furthermore, many of the rules are ambiguous (e.g., *Limit low-effort submissions*) or even entirely unuseful for an automated system without further information (e.g., *Read the FAQ*), affecting the quality of the moderation process. A possible solution consists of differentiating the policies for humans (users and moderators) from those for LLMs by writing clear, concise and less ambiguous rules optimized for moderating content using these models, even balancing the expressiveness of the rules with their length. By reducing the unuseful information in the prompt, this should increase the understanding and content moderation capabilities of LLMs. Additionally, the compliance of posts with the community rules might depend not only on the content itself but also on the context, such as other comments in the same thread or information not included in the community rules. The dataset, as indicated in Table 3, also includes the *ID* for each post, so it is possible to recover the structure of the entire thread and use this information to make better (and more informed) predictions. For example, this is particularly useful in case of presence of rules like "*Read the FAQ*".

While traditional automated tools for content moderation like Perspective API show good results, they fall short as soon as adaptation and personalization are needed, such as in the case of Reddit, where, besides some platform-level rules, each community has its own guidelines, and a one-size-fits-all solution is not possible. The promising results of using LLMs for content moderation have implications beyond Reddit. Indeed, the problem of tailoring content moderation to the rules of each community is not much different from personalizing content moderation considering the personal preferences of each person and the needs of different minorities and marginalized people, such as their religious and social norms, which is crucial in the current digital landscape to create equitable and safe online spaces, as well as to comply with national and international regulations.

5 LIMITATIONS

Despite the expected and promising results, this approach presents several pitfalls. First of all, the proposed content moderation pipeline cannot differentiate between a violation detected with high confidence and one just suspected. Indeed, despite their advanced language understanding level, LLMs still have limited mathematical capabilities, and obtaining a numerical confidence value of their moderation decision is difficult. Moreover, our proposal produces a binary decision, yet the information about the specific violation may be essential in real scenarios, e.g., if we want to consider the severity of the violation. Different schemes can be proposed to address this case, such as evaluating the content against each rule independently, where each is associated with a category.

In our experiments, we used GPT 3.5 and LLaMa 2, well-known LLMs developed by OpenAI and Meta, respectively, that have undergone safety alignment processes to prevent them from producing harmful content. Some adaptation techniques can increase the capabilities of LLMs on some specific goals, in our case, content moderation. These techniques would allow us to obtain an ad-hoc language model for moderating content (potentially with higher performance) while maintaining the capability to dialogue with users and preserving the aforementioned advantages. However, Qi *et al.* [45] discovered that fine-tuning can degrade the safety alignment of LLMs, exposing users to harmful content and unsafe responses. The absence of harmful content in the responses (and their quality) is even more important in our scenario, where OSNs are being used by billions of people worldwide. Therefore, in the case of fine-tuning, it is crucial to include appropriate safety mechanisms without relying only on the original safety of the model.

However, training of LLMs, as well as inference, is associated with significant financial and environmental costs, such as energy and infrastructure expenses (e.g., GPUs). Just to give an idea, we spent around 70 US dollars for the experiments reported in this work. Indeed, sustainability is one of the main drawbacks of LLMs [4]. Moreover, those costs fall on people who are not likely to benefit from those technologies, like marginalized populations with low-resource languages, considering that language models work better with English than with underrepresented languages. Aditionally, these models are trained on large-scale *corpora* (e.g., Common Crawl [8]) containing mainly uncurated data gathered from the Web, which includes the dominant viewpoints, biases, and stereotypes. As mentioned in the previous sections, this can perpetuate harm to marginalized communities, fragile users, and minorities. Moreover, LLMs may also generate content that is non-coherent with internal and external knowledge (i.e., hallucinations) and have difficulties when dealing with recent knowledge (i.e., knowledge recency).

Furthermore, ethical and legal concerns must be considered and deeply analyzed. For instance, the implications for users' privacy and safety are not completely clear, especially if using closed-source models (e.g., GPT), where we do not know how data and personal information are treated (e.g., they can be stored and used for training without the user's awareness). Using an open-source LLM running on private architecture should be preferred if possible. In addition, potential biases and harmful responses must be avoided by implementing appropriate control procedures, such as continuously checking the output of the content moderation process, looking for errors, disrespectful language and so on. Considering the possible errors made by LLMs (and AI systems in general), this implementation could be combined with a proper and usable user interface (UI). For example, the UI could explicitly indicate when interacting with an AI agent or a human, thus increasing their trust. Responsibility in case of adversarial attacks (or failures in general) needs to be clarified as well. These topics should (and must) be addressed by policymakers and AI regulators in collaboration with the industry and the research community.

Overall, the implementation of our proposal should strive to strike a balance between its effectiveness and the need to follow some essential principles, such as being as transparent and fair as possible, guaranteeing privacy, being sustainable and trustable, and bringing beneficial outcome. This goal can be achieved only by involving all the potential stakeholders, such as users, platform owners, moderators, and governments, and by implementing proper national and international regulations.

6 CONCLUSION

Content moderation is pivotal in ensuring the safety of OSNs. Its primary function is to safeguard platforms and users from malicious activities, thereby preserving online spaces as inclusive and enjoyable environments for all. Unfortunately, a notable drawback is that content moderation systems often revolve around Western norms and values, leading to unfair treatment of historically marginalized individuals, minorities, and vulnerable users. Consequently, their voices, online participation, and freedom of speech are limited.

In this context, we have explored the integration of content moderation systems with LLMs to facilitate the implementation of personalized content moderation. This involves empowering users with the capability to tailor certain aspects of their moderation experience. Leveraging LLMs also provides OSNs with the means to improve communication with their users, especially in instances where disagreement arises over moderation decisions, such as post removals or account suspensions. Furthermore, we conducted an assessment of the performance of advanced language understanding models, namely GPT-3.5 and LLaMa 2, in content moderation tasks. These models were compared to the capabilities of some commercial products like Perspective API and OpenAI Content Moderation. Despite the promising results, it is essential to acknowledge several limitations that need addressing before deploying the proposed system.

For this reason, our research endeavors are poised to expand in various directions. Primarily, we aim to assess the efficacy of LLMs in content moderation across diverse languages, cultures, and contexts, even involving real users from different areas of the world. This would also allow us to gauge the impact of explanatory features on enhancing the moderation experience and assess the user interface's usability for customizing the system. Furthermore, with the advent of multimodal models, such as GPT 4, we plan to generalize the proposed approach to different types of data, including images, videos, and beyond. Finally, in response to the escalating interest in decentralizing social services, our research will delve into exploring the potential decentralization of our approach and its implications for user privacy.

ACKNOWLEDGMENTS

The research leading to this work has received the partial support of the Department of Mathematics of the University of Padua (SID project: "*A tool to spread accessibility culture through gamification*") and of the Italian Ministry of University and Research (MUR) under the MOST—Sustainable Mobility National Research Centre initiative within the European Union Next-GenerationEU (PIANO NAZIONALE DI RIPRESA E RESILIENZA (PNRR)—MISSIONE 4 COMPONENTE 2, INVESTIMENTO 1.4—D.D. 1033 17 June 2022, CN00000023). Responsibility of the contents resides with the authors.

REFERENCES

[1] Norah Abokhodair and Sarah Vieweg. 2016. Privacy & social media in the context of the arab gulf. In *Proceedings of the 2016 ACM Conference on Designing Interactive Systems* (DIS '16). Association for Computing Machinery, Brisbane, QLD, Australia, 672–683. ISBN: 9781450340311. DOI: 10.1145/2901790.2901873.

[2] Shiza Ali, Afsaneh Razi, Seunghyun Kim, Ashwaq Alsoubai, Chen Ling, Munmun De Choudhury, Pamela J. Wisniewski, and Gianluca Stringhini. 2023. Getting meta: a multimodal approach for detecting unsafe conversations within instagram direct messages of youth. *Proc. ACM Hum.-Comput. Interact.*, 7, CSCW1, Article 132, (Apr. 2023), 30 pages. DOI: 10.1145/3579608.

[3] Arnav Arora et al. 2023. Detecting harmful content on online platforms: what platforms need vs. where research efforts go. *ACM Comput. Surv.*, 56, 3, Article 72, (Oct. 2023), 17 pages. DOI: 10.1145/3603399.

[4] 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* (FAccT '21). Association for Computing Machinery, Virtual Event, Canada, 610–623. ISBN: 9781450383097. DOI: 10.1145/3442188.3445922.

[5] Tom B. Brown et al. 2020. Language models are few-shot learners. (2020). arXiv: 2005.14165 [cs.CL].

[6] Yuyan Bu, Qiang Sheng, Juan Cao, Peng Qi, Danding Wang, and Jintao Li. 2023. Combating online misinformation videos: characterization, detection, and future directions. In *Proceedings of the 31st ACM International Conference on Multimedia* (MM '23). Association for Computing Machinery, Ottawa ON, Canada, 8770–8780. ISBN: 9798400701085. DOI: 10.1145/3581783.3612426.

[7] Stevie Chancellor, Yannis Kalantidis, Jessica A. Pater, Munmun De Choudhury, and David A. Shamma. 2017. Multimodal classification of moderated online pro-eating disorder content. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems* (CHI '17). Association for Computing Machinery, Denver, Colorado, USA, 3213–3226. ISBN: 9781450346559. DOI: 10.1145/3025453.3025985.

[8] Common Crawl. 2023. Common Crawl. https://commoncrawl.org. Accessed 17th December 2023. (2023).

[9] Borko Furht, (Ed.) 2010. *Decentralized online social networks. Handbook of Social Network Technologies and Applications.* Springer US, New York, NY, 349–378. ISBN: 978-1-4419-7142-5. DOI: 10.1007/978-1-4419-7142-5_17.

[10] Antonella De Angeli, Mattia Falduti, Maria Menendez, and Sergio Tessaris. 2023. Reporting non-consensual pornography: clarity, efficiency and distress. *Multimedia Tools and Applications*, 82, (Jan. 2023), 1–30. DOI: 10.1007/s11042-022-14291-z.

[11] Antonella De Angeli, Mattia Falduti, Maria Menendez Blanco, and Sergio Tessaris. 2021. Reporting revenge porn: a preliminary expert analysis. In *Proceedings of the 14th Biannual Conference of the Italian SIGCHI Chapter* (CHItaly '21) Article 30. Association for Computing Machinery, Bolzano, Italy, 5 pages. ISBN: 9781450389778. DOI: 10.1145/3464385.3464739.

[12] Giovanni Delnevo, Manuel Andruccioli, and Silvia Mirri. 2024. On the interaction with large language models for web accessibility: implications and challenges. In *2024 IEEE 21st Consumer Communications & Networking Conference (CCNC)*, 1–6. DOI: 10.1109/CCNC51664.2024.10454680.

[13] Facebook. 2023. Facebook community standards. https://transparency.fb.com/en-gb/policies/community-standards/. Accessed 1st December 2023. (2023).

[14] Mattia Falduti and Sergio Tessaris. 2022. On the use of chatbots to report non-consensual intimate images abuses: the legal expert perspective. In *Proceedings of the 2022 ACM Conference on Information Technology for Social Good* (GoodIT '22). Association for Computing Machinery, Limassol, Cyprus, 96–102. ISBN: 9781450392846. DOI: 10.1145/3524458.3547247.

[15] Karen E. Fisher. 2022. People first, data second: A humanitarian research framework for fieldwork with refugees by war zones. *Comput. Support. Cooperative Work.*, 31, 2, 237–297. DOI: 10.1007/S10606-022-09425-8.

[16] Mirko Franco, Salah A. Falioun, Karen E. Fisher, Ombretta Gaggi, Yacine Ghamri-Doudane, Ayat J. Nashwan, Claudio E. Palazzi, and Mohammed Shwamra. 2022. A technology exploration towards trustable and safe use of social media for vulnerable women based on islam and arab culture. In *Proceedings of the 2022 ACM Conference on Information Technology for Social Good* (GoodIT '22). Association for Computing Machinery, Limassol, Cyprus, 138–145. ISBN: 9781450392846. DOI: 10.1145/3524458.3547259.

[17] Mirko Franco, Ombretta Gaggi, Barbara Guidi, Andrea Michienzi, and Claudio E. Palazzi. 2023. A decentralised messaging system robust against the unauthorised forwarding of private content. *Future Generation Computer Systems*, 145, 211–222. DOI: https://doi.org/10.1016/j.future.2023.03.025.

[18] Mirko Franco, Ombretta Gaggi, and Claudio E. Palazzi. 2023. Analyzing the use of large language models for content moderation with ChatGPT examples. In *Proceedings of the 3rd International Workshop on Open Challenges in Online Social Networks* (OASIS '23). Association for Computing Machinery, Rome, Italy, 1–8. ISBN: 9798400702259. DOI: 10.1145/3599696.3612895.

[19] Mirko Franco, Ombretta Gaggi, and Claudio E. Palazzi. 2023. Can messaging applications prevent sexting abuse? A technology analysis. *IEEE Transactions on Mobile Computing*, 1–14. DOI: 10.1109/TMC.2023.3238189.

[20] Mirko Franco, Ombretta Gaggi, and Claudio E. Palazzi. 2022. Improving sexting safety through media forwarding control. In *2022 IEEE 19th Annual Consumer Communications & Networking Conference (CCNC)*, 1–6. DOI: 10.1109/CCNC49033.2022.9700555.

[21] Marco Furini, Silvia Mirri, Manuela Montangero, and Catia Prandi. 2019. Privacy perception and user behavior in the mobile ecosystem. In *GoodTechs '19: Proceedings of the 5th EAI International Conference on Smart Objects and Technologies for Social Good*. Rennes, France, 117–182. DOI: 10.1145/3342428.3342690.

[22] Marco Furini, Silvia Mirri, Manuela Montangero, and Catia Prandi. 2020. Untangling between fake-news and truth in social media to understand the covid-19 coronavirus. In *2020 IEEE Symposium on Computers and Communications (ISCC)*. Rennes, France, 1–6. DOI: 10.1109/ISCC50000.2020.9219663.

[23] Barbara Guidi. 2020. When blockchain meets online social networks. *Pervasive and Mobile Computing*, 62, 101131. DOI: https://doi.org/10.1016/j.pmcj.2020.101131.

[24] Barbara Guidi, Marco Conti, Andrea Passarella, and Laura Ricci. 2018. Managing social contents in decentralized online social networks: a survey. *Online Social Networks and Media*, 7, 12–29. DOI: https://doi.org/10.1016/j.osnem.2018.07.001.

[25] Barbara Guidi, Andrea Michienzi, and Laura Ricci. 2021. A graph-based socioeconomic analysis of Steemit. *IEEE Transactions on Computational Social Systems*, 8, 2, 365–376. DOI: 10.1109/TCSS.2020.3042745.

[26] Oliver L. Haimson, Daniel Delmonaco, Peipei Nie, and Andrea Wegner. 2021. Disproportionate removals and differing content moderation experiences for conservative, transgender, and black social media users: marginalization and moderation gray areas. *Proc. ACM Hum.-Comput. Interact.*, 5, CSCW2, Article 466, (Oct. 2021), 35 pages. DOI: 10.1145/3479610.

[27] Alon Halevy, Cristian Canton-Ferrer, Hao Ma, Umut Ozertem, Patrick Pantel, Marzieh Saeidi, Fabrizio Silvestri, and Ves Stoyanov. 2022. Preserving integrity in online social networks. *Commun. ACM*, 65, 2, (Jan. 2022), 92–98. DOI: 10.1145/3462671.

[28] Camille Harris, Amber Gayle Johnson, Sadie Palmer, Diyi Yang, and Amy Bruckman. 2023. "Honestly, I think TikTok has a vendetta against black creators": understanding black content creator experiences on TikTok. *Proc. ACM Hum.-Comput. Interact.*, 7, CSCW2, Article 320, (Oct. 2023), 31 pages. DOI: 10.1145/3610169.

[29] Shagun Jhaver, Amy Bruckman, and Eric Gilbert. 2019. Does transparency in moderation really matter? user behavior after content removal explanations on reddit. *Proc. ACM Hum.-Comput. Interact.*, 3, CSCW, Article 150, (Nov. 2019), 27 pages. DOI: 10.1145/3359252.

[30] 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. *Proc. ACM Hum.-Comput. Interact.*, 7, CSCW2, Article 289, (Oct. 2023), 33 pages. DOI: 10.1145/3610080.

[31] Jialun Aaron Jiang, Peipei Nie, Jed R. Brubaker, and Casey Fiesler. 2023. A trade-off-centered framework of content moderation. *ACM Trans. Comput.-Hum. Interact.*, 30, 1, Article 3, (Mar. 2023), 34 pages. DOI: 10.1145/3534929.

[32] Seunghyun Kim, Afsaneh Razi, Gianluca Stringhini, Pamela J. Wisniewski, and Munmun De Choudhury. 2021. A human-centered systematic literature review of cyberbullying detection algorithms. *Proc. ACM Hum.-Comput. Interact.*, 5, CSCW2, Article 325, (Oct. 2021), 34 pages. DOI: 10.1145/3476066.

[33] Yifang Li, Nishant Vishwamitra, Hongxin Hu, and Kelly Caine. 2020. Towards a taxonomy of content sensitivity and sharing preferences for photos. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems* (CHI '20). Association for Computing Machinery, Honolulu, HI, USA, 1–14. ISBN: 9781450367080. DOI: 10.1145/3313831.3376498.

[34] Huan Ma, Changqing Zhang, Huazhu Fu, Peilin Zhao, and Bingzhe Wu. 2023. Adapting large language models for content moderation: pitfalls in data engineering and supervised fine-tuning. (2023). arXiv: 2310.03400 [cs.LG].

[35] Renkai Ma and Yubo Kou. 2021. "how advertiser-friendly is my video?": youtuber's socioeconomic interactions with algorithmic content moderation. *Proc. ACM Hum.-Comput. Interact.*, 5, CSCW2, Article 429, (Oct. 2021), 25 pages. DOI: 10.1145/3479573.

[36] Renkai Ma and Yubo Kou. 2022. "I'm not sure what difference is between their content and mine, other than the person itself": a study of fairness perception of content moderation on youtube. *Proc. ACM Hum.-Comput. Interact.*, 6, CSCW2, Article 425, (Nov. 2022), 28 pages. DOI: 10.1145/3555150.

[37] Renkai Ma, Yue You, Xinning Gui, and Yubo Kou. 2023. How do users experience moderation?: A systematic literature review. *Proc. ACM Hum.-Comput. Interact.*, 7, CSCW2, Article 278, (Oct. 2023), 30 pages. DOI: 10.1145/3610069.

[38] Todor Markov, Chong Zhang, Sandhini Agarwal, Florentine Eloundou Nekoul, Theodore Lee, Steven Adler, Angela Jiang, and Lilian Weng. 2023. A holistic approach to undesired content detection in the real world. In *Proceedings of the Thirty-Seventh AAAI Conference on Artificial Intelligence and Thirty-Fifth Conference on Innovative Applications of Artificial Intelligence and Thirteenth Symposium on Educational Advances in Artificial Intelligence*

(AAAI'23/IAAI'23/EAAI'23) Article 1683. AAAI Press, 10 pages. ISBN: 978-1-57735-880-0. DOI: 10.1609/aaai.v37i12.26752.

[39] Sankha Subhra Mullick, Mohan Bhambhani, Suhit Sinha, Akshat Mathur, Somya Gupta, and Jidnya Shah. 2023. Content moderation for evolving policies using binary question answering. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 5: Industry Track)*. Sunayana Sitaram, Beata Beigman Klebanov, and Jason D Williams, (Eds.) Association for Computational Linguistics, Toronto, Canada, (July 2023), 561–573. DOI: 10.18653/v1/2023.acl-industry.54.

[40] Ani Nahapetian. 2016. Side-channel attacks on mobile and wearable systems. In *13th IEEE Annual Consumer Communications & Networking Conference (CCNC)*. Las Vegas, NV, USA, 243–247. DOI: 10.1109/CCNC.2016.7444763.

[41] United Nations. 2015. Transforming our world: the 2030 agenda for sustainable development. https://wedocs.unep.org/20.500.11822/9814. Accessed 12th November 2023. (2015).

[42] Judith S. Olson, Jonathan Grudin, and Eric Horvitz. 2005. A study of preferences for sharing and privacy. In *CHI '05 Extended Abstracts on Human Factors in Computing Systems* (CHI EA '05). Association for Computing Machinery, Portland, OR, USA, 1985–1988. ISBN: 1595930027. DOI: 10.1145/1056808.1057073.

[43] OpenAI. 2023. Gpt-4 technical report. (2023). arXiv: 2303.08774 [cs.CL].

[44] Royal Pathak, Francesca Spezzano, and Maria Soledad Pera. 2023. Understanding the contribution of recommendation algorithms on misinformation recommendation and misinformation dissemination on social networks. *ACM Trans. Web*, 17, 4, Article 35, (Oct. 2023), 26 pages. DOI: 10.1145/3616088.

[45] Xiangyu Qi, Yi Zeng, Tinghao Xie, Pin-Yu Chen, Ruoxi Jia, Prateek Mittal, and Peter Henderson. 2023. Fine-tuning aligned language models compromises safety, even when users do not intend to! (2023). arXiv: 2310.03693 [cs.CL].

[46] Morgan Klaus Scheuerman, Jialun Aaron Jiang, Casey Fiesler, and Jed R. Brubaker. 2021. A framework of severity for harmful content online. *Proc. ACM Hum.-Comput. Interact.*, 5, CSCW2, Article 368, (Oct. 2021), 33 pages. DOI: 10.1145/3479512.

[47] Farhana Shahid and Aditya Vashistha. 2023. Decolonizing content moderation: does uniform global community standard resemble utopian equality or western power hegemony? In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems* (CHI '23) Article 391. Association for Computing Machinery, Hamburg, Germany, 18 pages. ISBN: 9781450394215. DOI: 10.1145/3544548.3581538.

[48] Manuka Stratta, Julia Park, and Cooper deNicola. 2020. Automated content warnings for sensitive posts. In *Extended Abstracts of the 2020 CHI Conference on Human Factors in Computing Systems* (CHI EA '20). Association for Computing Machinery, Honolulu, HI, USA, 1–8. ISBN: 9781450368193. DOI: 10.1145/3334480.3383029.

[49] TikTok. 2023. Community guidelines. https://www.tiktok.com/community-guidelines/en/. Accessed 1st December 2023. (2023).

[50] Hugo Touvron et al. 2023. Llama 2: open foundation and fine-tuned chat models. (2023). arXiv: 2307.09288 [cs.CL].

[51] Hugo Touvron et al. 2023. Llama: open and efficient foundation language models. (2023). arXiv: 2302.13971 [cs.CL].

[52] Tumblr. 2022. Community guidelines. https://www.tumblr.com/policy/en/community. Accessed 14th November 2023. (2022).

[53] Kristen Vaccaro, Ziang Xiao, Kevin Hamilton, and Karrie Karahalios. 2021. Contestability for content moderation. *Proc. ACM Hum.-Comput. Interact.*, 5, CSCW2, Article 318, (Oct. 2021), 28 pages. DOI: 10.1145/3476059.

[54] Nishant Vishwamitra, Yifang Li, Hongxin Hu, Kelly Caine, Long Cheng, Ziming Zhao, and Gail-Joon Ahn. 2022. Towards automated content-based photo privacy control in user-centered social networks. In *Proceedings of the Twelfth ACM Conference on Data and Application Security and Privacy* (CODASPY '22). Association for Computing Machinery, Baltimore, MD, USA, 65–76. ISBN: 9781450392204. DOI: 10.1145/3508398.3511517.

[55] 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 Trans. Comput.-Hum. Interact.*, 30, 6, Article 83, (Sept. 2023), 36 pages. DOI: 10.1145/3603625.

[56] Xuhai Xu, Bingsheng Yao, Yuanzhe Dong, Saadia Gabriel, Hong Yu, James Hendler, Marzyeh Ghassemi, Anind K. Dey, and Dakuo Wang. 2024. Mental-llm: leveraging large language models for mental health prediction via online text data. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.*, 8, 1, Article 31, (Mar. 2024), 32 pages. DOI: 10.1145/3643540.

[57] Meng Ye, Karan Sikka, Katherine Atwell, Sabit Hassan, Ajay Divakaran, and Malihe Alikhani. 2023. Multilingual content moderation: A case study on Reddit. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*. Association for Computational Linguistics, Dubrovnik, Croatia, (May 2023), 3828–3844. DOI: 10.18653/v1/2023.eacl-main.276.

[58] Wayne Xin Zhao et al. 2023. A survey of large language models. (2023). arXiv: 2303.18223 [cs.CL].

[59] Matteo Zignani, Sabrina Gaito, and Gian Paolo Rossi. 2018. Follow the "Mastodon": structure and evolution of a decentralized online social network. *Proceedings of the International AAAI Conference on Web and Social Media*, 12, 1, (June 2018), 541–550. DOI: 10.1609/icwsm.v12i1.14988.

A TABLE OF ACRONYMS

We report here a table with the acronyms used in the manuscript and their meaning.

Table 9. Table of Acronyms

Acronym	Meaning
AI	Artificial Intelligence
API	Application Programming Interface
BIPOC	Black, Indigenous, and People of Color
COVID-19	Coronavirus Diseas of 2019
FAQ	Frequently Asked Questions
FN	False Negative
FP	False Positive
GPT	Generative Pre-training Transformer
GPU	Graphics Processing Unit
LGBTQ+	Lesbian, Gay, Bisexual, Transgender, Queer (plus others)
LLaMa	Large Language Model Meta AI
LLM	Large Language Model
JSON	JavaScript Object Notation
ML	Machine Learning
NFT	Non-Fungible Token
OSN	Online Social Network
RA	Recommendation Algorithm
SDG	Sustainable Development Goal
TP	True Positive
UI	User Interface

Received 20 February 2007; revised 12 March 2009; accepted 5 June 2009