Analyzing the Use of Large Language Models for Content Moderation with ChatGPT Examples

Mirko Franco
Department of Mathematics
University of Padua
Padua, Italy
mirko.franco@math.unipd.it

Ombretta Gaggi
Department of Mathematics
University of Padua
Padua, Italy
gaggi@math.unipd.it

Claudio E. Palazzi
Department of Mathematics
University of Padua
Padua, Italy
cpalazzi@math.unipd.it

ABSTRACT

Content moderation systems are crucial in Online Social Networks (OSNs). Indeed, their role is to keep platforms and their users safe from malicious activities. However, there is an emerging consensus that such systems are unfair to fragile users and minorities. Furthermore, content moderation systems are difficult to personalize and lack effective communication between users and platforms. In this context, we propose an enhancement of the current framework of content moderation, integrating Large Language Models (LLMs) in the enforcing pipeline.

CCS CONCEPTS

• **Social and professional topics** → **User characteristics**; **Censoring filters**; • **Human-centered computing** → **Social networks**.

KEYWORDS

content moderation, harmful content, large language models

ACM Reference Format:
Mirko Franco, Ombretta Gaggi, and Claudio E. Palazzi. 2023. Analyzing the Use of Large Language Models for Content Moderation with ChatGPT Examples. In *3rd International Workshop on Open Challenges in Online Social Networks (OASIS '23), September 04–08, 2023, Rome, Italy*. ACM, New York, NY, USA, 8 pages. https://doi.org/10.1145/3599696.3612895

1 INTRODUCTION

The spread of mobile devices, the ubiquitous availability of wireless connectivity, and the advent of social media platforms have empowered people all over the world, allowing access to information anywhere and anytime, as well as the creation of new businesses and relationships. Furthermore, social media platforms have become the place where life happens for many (young) people [5, 14]. On the other hand, this has raised unprecedented concerns about user privacy and safety [9, 13]. Some examples are the unauthorized forwarding of sex-related content without the owner's consent (i.e., revenge porn) [10–12], spread of information shared on online social networks (OSNs) [18], presence of disturbing and/or harmful content [9, 25], misinformation, cyberbullying, online grooming, etc.

In this context, we have also to consider that people's willingness to share content depends on the social circle with which they are sharing them [19]. Indeed, the perception of privacy is influenced by several factors, e.g., personal preferences, sociocultural background, etc. [1, 17]. Similarly, people may be unwilling to receive content from one or more particular senders, while they may feel comfortable receiving the same content from another person.

Social media platforms heavily rely on content moderation systems to enforce integrity, i.e., to keep platforms and their users safe from malicious activities and remove violating content [16]. Although many enforcing pipelines involve human moderators in the loop (e.g., Facebook), they are far from working in a safe and secure way. In particular, there is an emerging consensus that content moderation systems are unfair to fragile users (in terms of age, digital literacy, level of education, etc.) and minorities (e.g., LGBTQ+, etc.) [15, 27]. This is a serious issue if we consider that, for example, content and account removals experienced by these people limit their online participation and freedom of speech. Furthermore, as mentioned above, the content people are willing to receive and visualize depends on the sender, and the visualization of disturbing material (e.g., texts, images, videos, etc.) can cause severe health consequences, such as post-traumatic stress disease (PSTD), especially for those with past traumatic experiences [25]. This represents a problem since, if properly designed, social media platforms embody a tool to achieve (some of) the Sustainable Development Goals (SDGs), the global agenda for the next decade of the United Nations (UN) [10]. Some of them are indeed strongly linked with online social networks: Good Health and Well-being (3), Gender Equality (5), Sustainable Cities and Communities (11), and Peace, Justice and Strong Institutions (16). Therefore, including fair content moderation systems in online social media platforms to minimize harm to fragile users and make online communities safer and more inclusive is urgent.

In recent years, Large Language Models (LLMs) (e.g., GPT [7], LLaMa [26], etc.) - which model the generative likelihood of word sequences and predict future words - have received a lot of attention from academia, industry and our society, also thanks to the launch of *ChatGPT*[1] [29]. These models have the potential to profoundly change our interactions with applications, websites, robots, and computing systems, thus having a significant impact on our lives. Indeed, for example, they could be applied in healthcare as an interface for patients and healthcare providers, in law to help people access justice systems, in programming, and in many other

[1] https://openai.com/blog/chatgpt

OASIS '23, September 04–08, 2023, Rome, Italy

ACM ISBN 979-8-4007-0225-9/23/09.
https://doi.org/10.1145/3599696.3612895

scenarios [6]. A first attempt has been made by Qin *et al.* [21], who proposed an LLM-based system to detect depression on social media. But little is known about the applications of this novel technology to online social networks and content moderation systems.

This paper aims to discuss issues about the current framework of content moderation on Online Social Networks (OSNs) and propose an enhanced enforcing pipeline that can adapt to different personal preferences. Thanks to the integration with an LLM, besides categorizing content as sensitive (or not), the proposed pipeline provides in-depth explanations to users on the reasons behind a detected violation. Moreover, through the explanations, it can better support the work of human moderators. This is crucial in creating safer and more inclusive platforms for fragile users and minorities. In addition, our contribution is in line with the SDGs. To demonstrate the effectiveness of using an LLM in this context, we show some examples considering three case studies: sex-related texts, texts containing gender stereotypes, and those that are offensive to people with disabilities. To this aim, we use *ChatGPT* as a representative example of LLM, but others can be considered. To the best of our knowledge, this is the first attempt in this direction.

The rest of this paper is organized as follows. Section 2 reviews the relevant literature. Section 3 presents the proposed architecture for content moderation and its differences from the usual pipeline. We present and discuss some examples of dialogues in Section 4. Finally, we draw our conclusion and present some future directions in Section 5.

2 RELATED WORK

Content moderation is one of the pillars of Online Social Networks (OSNs). In fact, content moderation systems fight malicious activities (e.g., online child exploitation, misinformation, hate speech, etc.) and remove violating content, thus protecting platforms and their users, as well as fostering a good user experience [16]. Thanks to the increasing popularity of online social media platforms, this topic has gained a lot of interest from researchers. For instance, Franco *et al.* presented some guidelines for developers to build messaging systems safer by design for sexting in [11] and described the proposed platform, named *SafeSext*, in detail in [12]. Furthermore, in [10], they discussed two possible decentralized approaches for preventing the unauthorized forwarding of private content and their extension with blockchain and NFT technologies. Instead, Stratta *et al.* [25] proposed a Chrome extension to hide sensitive content automatically on the Web, thus making a step forward towards the well-being of people on the Internet. Ali *et al.* [2] explored which indicators are most helpful in detecting and mitigating online harm on social media platforms for young people, acknowledging that the implementation of end-to-end encryption will limit the amount of information available to platforms for content moderation. Halevy *et al.* [16] reviewed the recent technical progress and ongoing challenges to preserve the integrity of online social networks. Arora et al. [3] discussed methods for harmful content detection and argued about the differences between the research efforts and the actual needs of social media platforms.

In recent years, awareness about the harm caused by content moderation systems and their unfairness to gender, race, religion, etc. have raised both in the scientific community [15, 23, 27] and among citizens [4]. Haimson *et al.* [15] analyzed and presented the results of two surveys regarding social media content and takedown experiences, focusing on conservative, transgender, and black social media users. They figured out that there are significant differences in the type of removed content between groups and in the consequences of the removals. Indeed, while conservative users' removed content often actually violated the guidelines, those of the other two considered categories of users often followed the policies or fell into a gray area of content moderation. This phenomenon, however, limits the online participation of marginalized people and minorities and their ability to communicate and take advantage of the unprecedented opportunities provided by social media platforms. In the same direction, Vaccaro *et al.* [27] investigated how users can influence the automated decisions of content moderation systems (i.e., contestability) through some participatory design workshops. Instead, Shahid *et al.* [23] interviewed 19 Bangladeshi users who had received restrictions for violating Facebook community guidelines to understand better the interaction of humans with moderation systems and, in particular, inform researchers about the experiences of users from the Global South. They showed that these systems are unfair and centered on Western norms, even amplifying historical power relations, thus perpetuating harm to marginalized people.

These studies provided insightful suggestions for the design of social media platforms and content moderation systems. For instance, blurring potentially problematic content allows the platform to avoid a binary decision (e.g., to remove or not) and better manage content moderation gray areas. Another design guideline consists in applying different content moderation policies for different online spaces (e.g., timeline vs. private conversation) [15]. Vaccaro *et al.* [27] suggested that users and platforms would benefit from improved communication. In particular, participants of the surveys advocated for more transparency and explanations of the decisions of content moderation systems, along with the possibility to co-design such systems. According to [15] and [23], this is particularly important since involving minorities and marginalized people in defining policies and designing these systems would allow us to obtain more equitable and inclusive platforms. Besides them, local stakeholders, policymakers, human rights organizations, etc., should also be included to integrate local norms, values, and sensitiveness in content moderation systems. Finally, content moderation should have an educative aim (e.g., no consequences for the first violation) rather than being punitive (e.g., deleting the account).

Researchers have recently focused on detecting private and sensitive content and adapting content moderation systems to different sensitiveness. For instance, Li *et al.* [17] proposed a new method to collect sensitive content and employed it to generate a taxonomy of sharing preferences of 116 users involved in their study. Acknowledging that a one-size-fits-all approach would not be suitable for detecting private photos, Vishwamitra *et al.* [28] employed the method proposed in [17] to collect a dataset of images and develop an ML-based model able to detect private images in a user-specific manner automatically. However, this methodology is difficult to apply with disturbing content, especially with marginalized people (in terms of race, religion, age, level of education, etc.), because of

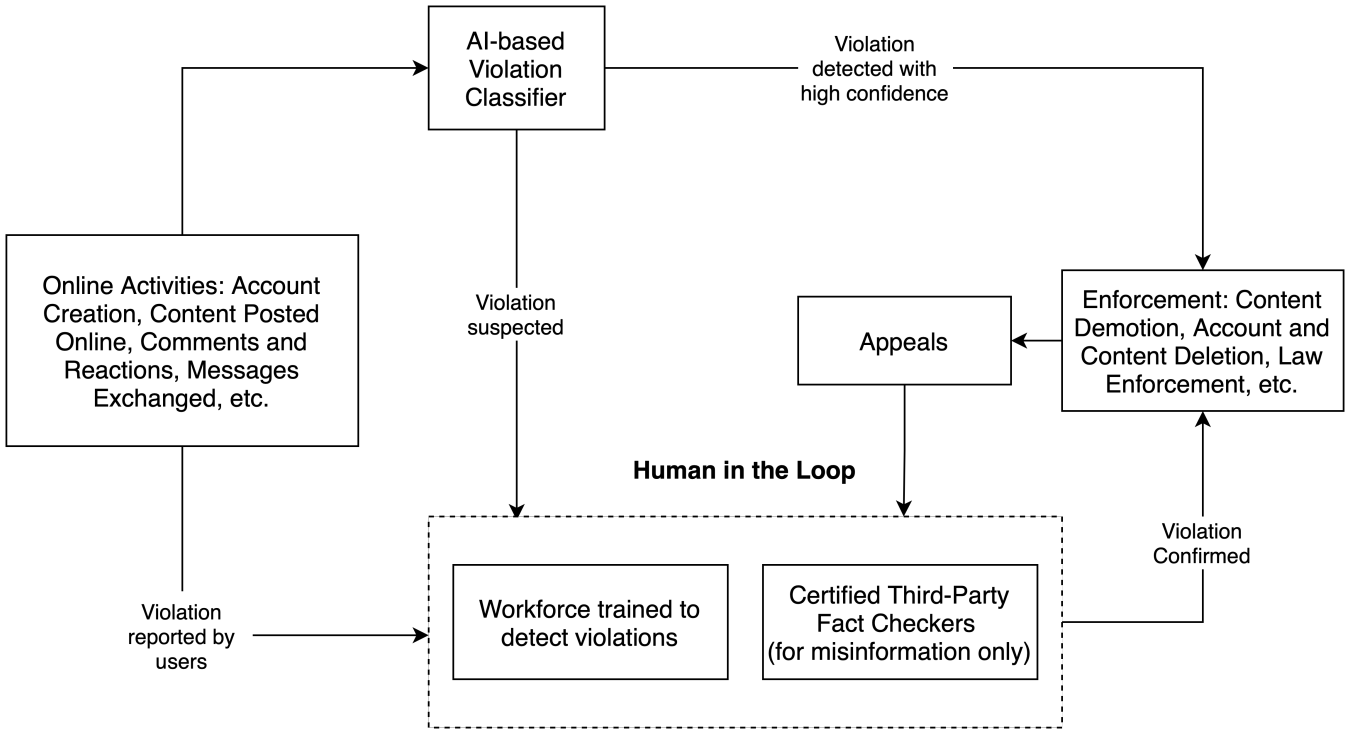

Figure 1: Facebook's Content Moderation Pipeline

the numerous ethical concerns and the possibility of causing harm to users.

Large Language Models (LLMs) are expected to impact our interactions with technology, including social media platforms. Some studies have already explored the use of language interfaces on online social networks. For instance, Falduti *et al.* [8] developed a chatbot to facilitate access to justice systems for victims of Non-Consensual Intimate Images (NCII) abuse.

In this direction, we propose an enhanced content moderation pipeline able to provide better explanations to users about decisions and, if necessary, to chat with them. The proposed pipeline can consider personal preferences by applying different rules (i.e., by specifying different policies in the prompt provided to the LLM), thus better supporting the needs of fragile users and minorities. This work is the first step in showing how social media platforms and content moderation systems can benefit from the integration with LLMs.

3 INTEGRATING CONTENT MODERATION SYSTEMS WITH LARGE LANGUAGE MODELS

Figure 1 represents Facebook's content moderation pipeline, which is similar to the one implemented by other social media platforms. Online Social Networks (OSNs) can detect potential violations and harmful content in two possible ways: through AI-based systems that continuously analyze the content (e.g., as it is uploaded) or reports of users who visualize the content on the platforms (e.g., in the feed, private conversations, etc.) or experience potentially dangerous situations (e.g., online grooming, spam, blackmailing, etc.). Violating content detected by the AI with very high confidence may be removed immediately and/or the platform may take other actions without the intervention of human reviewers. Conversely, if the violation is just suspected, human moderators are involved. In this case, the platform handles potential violations differently depending on whether they are community guidelines violations or misinformation (e.g., fake news). In the former scenario, the content is reviewed by a pool of paid moderators trained in recognizing prohibited material and behaviors. The platform removes the content (or performs another action) if multiple reviewers agree that the content is violating. Instead, in the latter case, misinformation content are sent to and reviewed by third-party fact-checkers. Once the violation is confirmed, the platform will take action (e.g., content deletion, limitation of the account, etc.) and communicate it to the user. Then, he/she can appeal the decision and ask for a re-evaluation by human moderators. More details on the flow of integrity enforcement at Facebook can be found in [16].

Recently, this framework of content moderation has shown its weaknesses. For instance, besides the inertia of social media platforms in shaping their community guidelines to consider minorities, marginalized people, and fragile users, the machine learning algorithms underlying content moderation pipelines struggle to interpret different languages and consider the needs of minorities and fragile users [4]. Moreover, human moderators are often English speakers and, even when local moderators are involved, they work in unethical conditions. Just to mention some examples, their

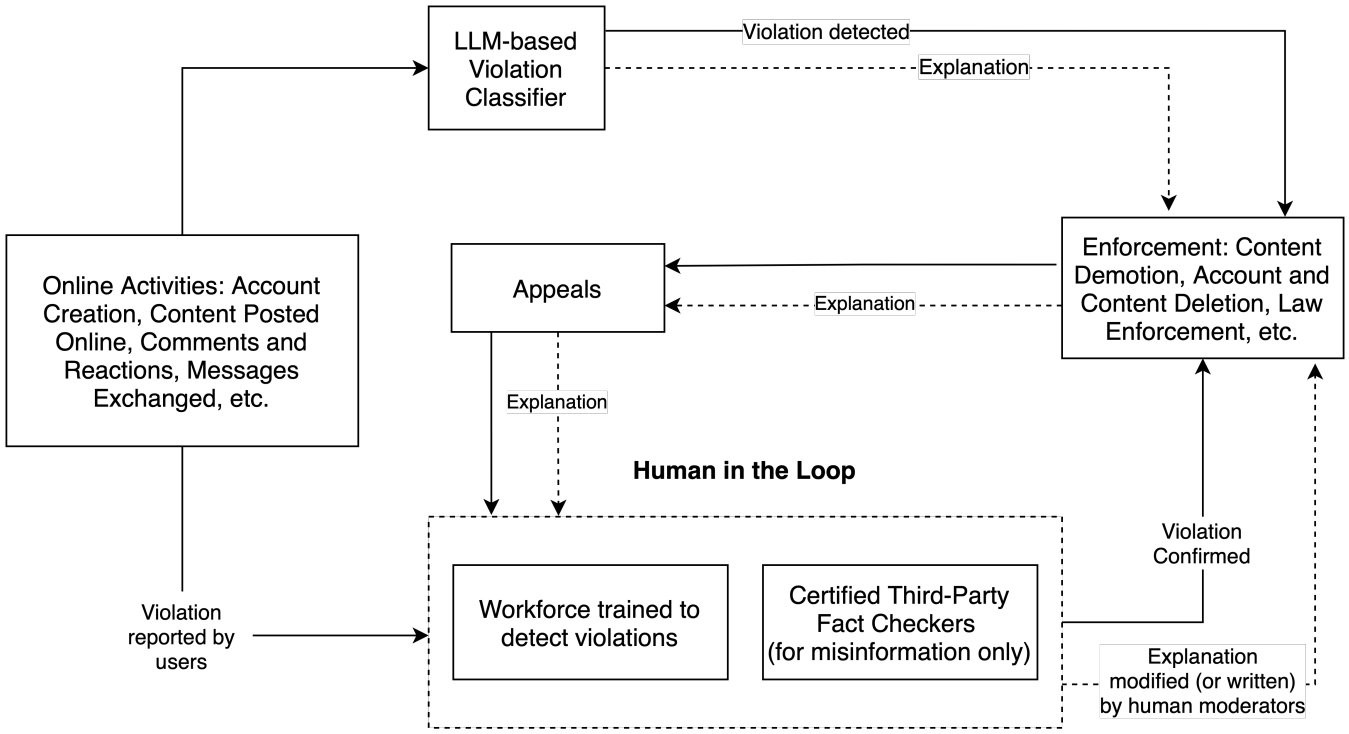

Figure 2: The Proposed Content Moderation Pipeline

salary is low, their psychological well-being is continuously at risk because of exposure to harmful content, and they are constantly pressured by the power hegemony of platforms (e.g., online companies can easily find cheaper labor). In addition, the platforms do not always provide proper explanations about their decisions, leaving users alone to understand the reason behind a specific action. This prevents the users from actually challenging the decisions of content moderation systems, making these platforms far from being really inclusive.

In the rest of the paper, we will focus on the text and, in particular, on detecting harmful and disturbing texts, even though most of our findings can be generalized to other kinds of content. In particular, we proposed an enhanced content moderation pipeline integrated with an LLM that enables a more user-centric experience and an improved interaction with such systems. The use of an LLM not only allows us to obtain a classification of the content (to detect a possible violation) but also, by designing appropriate prompts, to easily influence our systems to consider different sets of rules, each of them identifying the personal preferences of a person. In addition, we can ask the language model to explain its decision, considering that even if it is not real reasoning in a human-like (or logic) sense, it will appear as such to the user. This explanation can be helpful to human moderators as a support for their work, but more importantly to the user, who can understand the rationale behind a specific action taken by the platform. Moreover, we believe that even the appeal process would benefit from the integration of content moderation pipelines with LLMs, letting the user better explain the reasons for his/her request for re-evaluation. Finally, as mentioned in [8], language interfaces may help report malicious behavior (e.g., NCII abuse, etc.) and access justice procedures. Therefore, augmenting content moderation with LLMs allows social media platforms to improve the user experience, providing a new way of interaction, and be more inclusive, minimizing the harm caused to minorities and fragile users through the application of user-specific rules (e.g., to hide disturbing content). We show the proposed pipeline in Figure 2.

Unfortunately, at the time of writing, LLMs have limited mathematical capabilities and obtaining a numeric confidence value of the decision of the model is not possible (we can consider it as an open research problem). Therefore, the proposed content moderation pipeline cannot differentiate between a suspected violation and one detected with high confidence; it only generates a binary outcome with a violation alert raised or not.

We report here a possible structure of dialogue with an LLM and explain how to exploit its answers to perform content moderation, i.e., identify harmful text considering rules provided by the user (and hence his/her sensitiveness). First of all, we must provide the language model with the rules it has to consider while evaluating the content.

Prompt 1: Assuming the following rules:
- First rule
- Second rule
- Third rule
- ...

> Evaluate whether the following text complies with the rules. Answer only with "Yes." or "No.". The text is: `[TEXT]`

By explicitly asking the model to answer only with *Yes* or *No*, we can easily process it automatically and figure out the binary decision of the LLM. Moreover, our method can consider different sensitiveness by employing different rules in evaluating the text, thus better satisfying the needs of fragile users and minorities. Once the language model has answered our request, we can ask to explain its decision by providing the following prompt.

> **Prompt 2**: Provide an explanation.

As we mentioned earlier, although the limited reasoning capabilities of LLMs, their answers sound reasonable to humans. So, even if the explanation is not the result of real reasoning in a human-like sense, it can appear convincing to users, thus enhancing the interaction with content moderation systems, especially in case of error or disagreement.

4 CASE STUDIES

In this section, we report and discuss some examples generated by using *ChatGPT* to provide a preliminary assessment of the effectiveness of our proposal, focusing on three case studies: sex-related content, text containing gender stereotypes, and text that is offensive to people with disabilities. These topics have received a lot of attention in the last years from academia and citizens, and therefore represent interesting case studies [10–12, 22, 24].

4.1 Sex-related Text

As mentioned earlier, our approach can consider sociocultural differences (and/or personal preferences) by using different rules to evaluate the text. With the support of an appropriate user interface, these rules can be selected by the user and inserted in the prompt. In this way, the content moderation system can flag the same text as a violation or not, depending on the specified rules, as shown in the following two examples.

> **Prompt 1:** Assuming the following rules:
> - We do not tolerate content related to politics.
> - We do not tolerate violence, hatred, and dehumanization of individuals or groups based on race, ethnic or national origin, religion, gender, gender identity, age, or sexual orientation.
> - We do not tolerate content that promotes self-harm.
>
> Evaluate whether the following text complies with the rules. Answer only with "Yes." or "No.". The text is: ugh, just asking, but I have some new requests on lovesense... is it someone from here?
> **Answer:** Yes.
> **Prompt 2:** Provide an explanation.
> **Answer:** The provided text does not violate any of the mentioned rules. It does not contain any political content, violence, hatred, dehumanization, or promotion of self-harm. It appears to be a casual inquiry about "lovesense," which does not violate the stated rules.

As expected, if we provide the same text to the system with different rules, the output is different from the previous case. We can leverage this behavior to make content moderation systems more inclusive, allowing the users to adapt to some extent the rules employed by these systems.

> **Prompt 1:** Assuming the following rules:
> - We do not tolerate sex-related content except for medical purposes.
> - We do not tolerate content related to politics.
> - We do not tolerate violence, hatred, and dehumanization of individuals or groups based on race, ethnic or national origin, religion, gender, gender identity, age, or sexual orientation.
> - We do not tolerate content that promotes self-harm.
>
> Evaluate whether the following text complies with the rules. Answer only with "Yes." or "No.". The text is: ugh, just asking, but I have some new requests on lovesense... is it someone from here?
> **Answer:** No.
> **Prompt 2:** Provide an explanation.
> **Answer:** The given text does not comply with the rules stated. Although it doesn't contain explicit sexual content or violence, it does refer to a specific product related to sexual activity ("lovesense") and asks if someone from the current conversation is associated with it. This implies a potential intention to engage in or discuss sexual content, which violates the first rule.

Social media platforms are also a place where people search for information and/or share informative content, even on sensitive topics, such as sexuality. This represents an additional difficulty for content moderation systems, especially for the automated classifiers, because of the possible ambiguity and overlay between allowed and not allowed content. Our approach can handle even this scenario, as shown by the example below, where we ask the LLM to evaluate a sex-related text about sexual health and sexually transmitted diseases. Clearly, such content is reasonable in every online social network.

> **Prompt 1:** Assuming the following rules:
> - We do not tolerate sex-related content except for medical purposes.
> - We do not tolerate content related to politics.
> - We do not tolerate violence, hatred, and dehumanization of individuals or groups based on race, ethnic or national origin, religion, gender, gender identity, age, or sexual orientation.
> - We do not tolerate content that promotes self-harm.
>
> Evaluate whether the following text complies with the rules. Answer only with "Yes." or "No.". The text is: Condoms are a must-have accessory when it comes to sex that is pleasurable and safe. They're your best defence against sexually transmitted diseases (some of which can lead to cancer or AIDS) and unwanted pregnancies.[2]
> **Answer:** Yes.

[2] Source: https://www.durex.com.my/blogs/explore-sex/how-to-put-on-a-condom/

Prompt 2: Provide an explanation.
Answer: The provided text complies with the given rules. It discusses the importance of using condoms for safe and pleasurable sex, emphasizing their role in preventing sexually transmitted diseases and unwanted pregnancies. The content is focused on the medical aspect of sexual health and does not contain any explicit or inappropriate sexual content. It does not violate the rules related to sex-related content, violence, hatred, or promotion of self-harm.

4.2 Text containing Gender Stereotypes

Unfortunately, online social networks are also a place for sharing misogynist content, enforcing traditional gender stereotypes and biases, and perpetuating harm to minorities, such as the LGBTQ+ community. As shown in the example below, our approach is able to detect this violation of the second rule correctly.

Prompt 1: Assuming the following rules:
- We do not tolerate sex-related content except for medical purposes.
- We do not tolerate content containing sexism and gender stereotypes (e.g., those related to professions or activities), even indirectly.

Evaluate whether the following text complies with the rules. Answer only with "Yes." or "No.". The text is: The mum - solely responsible for domestic chores - bought some house appliances in order to do her housework: a dishwasher, a washing machine, and an iron. How many house appliances did she buy?
Answer: No.
Prompt 2: Provide an explanation.
Answer: The provided text does not comply with the given rules. It contains gender stereotypes by portraying the mother as solely responsible for domestic chores, which reinforces traditional gender roles. Additionally, the mention of domestic chores and specific house appliances can be seen as indirectly related to gender stereotypes. Therefore, the text does not adhere to the rules that prohibit content containing sexism and gender stereotypes.

Unfortunately, we receive the wrong answer if we remove the part of the sentence "*solely responsible for domestic chores*".

Prompt 1: Assuming the following rules:
- We do not tolerate sex-related content except for medical purposes.
- We do not tolerate content containing sexism and gender stereotypes (e.g., those related to professions or activities), even indirectly.

Evaluate whether the following text complies with the rules. Answer only with "Yes." or "No.". The text is: The mum bought some house appliances in order to do her housework: a dishwasher, a washing machine, and an iron. How many house appliances did she buy?
Answer: Yes.
Prompt 2: Provide an explanation.
Answer: Yes, the text complies with the given rules. It does not contain any sex-related content or promote sexism or gender stereotypes. The sentence simply states that the mum bought some house appliances (dishwasher, washing machine, and iron) to do her housework. It focuses on the actions and tasks related to household chores without making any explicit or implicit reference to gender roles or stereotypes.

These stereotypes are also sometimes present in the exercises for children (e.g., math, grammar, etc.), causing a collective awareness about these topics.

Prompt 1: Assuming the following rules:
- We do not tolerate sex-related content except for medical purposes.
- We do not tolerate content containing sexism and gender stereotypes (e.g., those related to professions or activities), even indirectly.

Evaluate whether the following text complies with the rules. Answer only with "Yes." or "No.". The text is: If the mum has to iron three shirts and has already ironed two of them, how many more does she have to iron? The dad is a mechanic, and he usually repairs three cars in a day. How many cars may he repair in 2 days?
Answer: Yes.
Prompt 2: Provide an explanation.
Answer: Yes, the provided text complies with the rules. The text is a simple mathematical problem involving ironing shirts and repairing cars. It does not contain any sex-related content, and there are no elements of sexism or gender stereotypes present in the text. The activities mentioned (ironing shirts and repairing cars) are not tied to any specific gender, and the text does not imply any gender bias. It solely focuses on mathematical problem-solving without any inappropriate or offensive content.

Unfortunately, the violation is not recognized. However, although we initially wrote this example considering it as violating, the output of the model is not surprising. Indeed, this topic is highly controversial, and an agreement is not present even among humans. Moreover, the provided explanation is reasonable. However, if we provide the following prompt to the model, the answer confirms that the problem arises when someone (e.g., mum, dad, etc.) is always associated with a specific activity and/or profession.

Prompt 3: Is there some potential violation of the rules in case of repeated similar text?
Answer: Yes, there is a potential violation of the rules in case of repeated similar text. While the specific text you provided in your previous question does not violate the rules, if similar texts were repeated excessively or in a pattern that could be interpreted as promoting gender stereotypes, it could potentially be considered a violation. Context and intent play a role in determining compliance, so it is important to be mindful of the overall content and avoid any patterns

that could be seen as perpetuating sexism or gender biases.

This is also the case of the first example where, if we substitute *mum* with *dad*, the violation is detected as well since the dad becomes solely responsible for domestic chores.

4.3 Text that is Offensive to People with Disabilities

The use of inclusive language has gained importance and plays a key role in shaping perceptions, attitudes, and social inclusion for people with disabilities. By adopting an inclusive and people-first language, we contribute to building a more inclusive and equitable society. Unfortunately, there is a lack of awareness among users of social media about disability, and many contents are not inclusive at all. Therefore, detecting non-inclusive and/or offensive content is important. The following example shows that people-first language should be used.

> **Prompt 1:** Assuming the following rules:
> • The use of inclusive language is mandatory.
> Evaluate whether the following text complies with the rules. Answer only with "Yes." or "No.". The text is: An interesting case study is represented by disabled people.
> **Answer:** No.
> **Prompt 2:** Provide an explanation.
> **Answer:** The text does not comply with the rule of using inclusive language. The term "disabled people" is not considered inclusive because it focuses on the disability rather than the person. A more inclusive and person-first language would be "people with disabilities."

However, using people-first language is not sufficient, as shown in the following example.

> **Prompt 1:** Assuming the following rules:
> • The use of inclusive language is mandatory.
> Evaluate whether the following text complies with the rules. Answer only with "Yes." or "No.". The text is: People with disabilities cannot have a satisfying sexual life.
> **Answer:** No.
> **Prompt 2:** Provide an explanation.
> **Answer:** No, the provided text does not comply with the rule of using inclusive language. The phrase "People with disabilities cannot have a satisfying sexual life" uses a negative generalization, implying that all individuals with disabilities are unable to have satisfying sexual lives. This statement is both incorrect and perpetuates stereotypes about people with disabilities. Inclusive language aims to avoid making sweeping generalizations or assumptions about any group of people based on their characteristics, including disabilities. A more inclusive and respectful way to express this idea would be to use language that acknowledges the diversity of experiences among individuals with disabilities, such as "Some people with disabilities may face unique challenges in their sexual lives, but many can and do have satisfying sexual experiences."

5 CONCLUSION

Content moderation systems are one of the fundamental parts of social media platforms. Indeed, keeping platforms and their users safe from malicious activities and harmful content is crucial, especially for maintaining a pleasant online space. However, these systems are unfair to fragile users and minorities, preventing social media platforms from being safe and inclusive and limiting the online participation of these users.

In this scenario, we have proposed an enhancement of the current content moderation systems, integrating Language Models (LLMs) in the enforcing pipeline. Our approach improves the personalization of content moderation pipelines, allowing users to specify some rules in the prompt of an LLM, thus influencing its decisions and, through explanations, the communication between users and platforms. The enhanced interaction with content moderation systems improves the user experience and represents the first step toward more inclusive and safer online social networks.

Unfortunately, content moderation systems, even if integrated with large language models, may still need help in considering all the possible personal preferences and scenarios. Large language models themselves, even if powerful, have some issues, such as hallucinations and knowledge recency. However, we believe that our proposal is still an improvement of the current framework of content moderation. Indeed, online communities, in particular minorities and fragile users, can benefit from improved communication with the platform and increased personalization.

We plan to extend our research in several directions. First of all, considering that most of the content shared on social media is visual, with the advent of multimodal models (e.g., GPT-4 [20]), we would like to generalize our approach to images and videos. Moreover, designing an appropriate interface for the users to select (some of) the content moderation rules, i.e., to personalize the system, is necessary. We plan to perform an extensive test campaign to prove the effectiveness of our approach. In particular, we would like to implement and test our system involving real users from different cultures. In addition, considering real (and long) community guidelines would be interesting and allow us to design a more comprehensive system. Finally, we would like to investigate the privacy implications of our approach and explore the possibilities of decentralization (e.g., deployment on mobile devices) of the content moderation pipeline.

ACKNOWLEDGMENTS

The authors of this paper acknowledge the support of the European Space Agency's (ESA) project "QUIC over Satellite" (QUICoS) https://artes.esa.int/projects/quicos, contract n. 4000138640/22/NL/AF and of the Italian Ministry of University and Research (MUR) under the PON ex DM 1061 and the PNRR CN-MOST initiatives. 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* (Brisbane, QLD, Australia) *(DIS '16)*. Association for Computing Machinery, New York, NY, USA, 672–683. https://doi.org/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. https://doi.org/10.1145/3579608

[3] Arnav Arora, Preslav Nakov, Momchil Hardalov, Sheikh Muhammad Sarwar, Vibha Nayak, Yoan Dinkov, Dimitrina Zlatkova, Kyle Dent, Ameya Bhatawdekar, Guillaume Bouchard, and Isabelle Augenstein. 2023. Detecting Harmful Content on Online Platforms: What Platforms Need vs. Where Research Efforts Go. *ACM Comput. Surv.* (jun 2023). https://doi.org/10.1145/3603399 Just Accepted.

[4] Katie Canales. 2021. Facebook's AI moderation reportedly can't interpret many languages, leaving users in some countries more susceptible to harmful posts. https://www.businessinsider.com/facebook-content-moderation-ai-cant-speak-all-languages-2021-9. Retrieved 19 June 2023.

[5] Scott Counts and Karen E. Fisher. 2010. Mobile social networking as information ground: A case study. *Library & Information Science Research* 32, 2 (2010), 98–115. https://doi.org/10.1016/j.lisr.2009.10.003

[6] Rishi Bommasani et al. 2022. On the Opportunities and Risks of Foundation Models. arXiv:2108.07258 [cs.LG]

[7] Tom B. Brown et al. 2020. Language Models Are Few-Shot Learners. In *Proceedings of the 34th International Conference on Neural Information Processing Systems* (Vancouver, BC, Canada) *(NIPS'20)*. Curran Associates Inc., Red Hook, NY, USA, Article 159, 25 pages.

[8] 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* (Limassol, Cyprus) *(GoodIT '22)*. Association for Computing Machinery, New York, NY, USA, 96–102. https://doi.org/10.1145/3524458.3547247

[9] Mirko Franco, Salah A. Falyoun, 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* (Limassol, Cyprus) *(GoodIT '22)*. Association for Computing Machinery, New York, NY, USA, 138–145. https://doi.org/10.1145/3524458.3547259

[10] 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 (2023), 211–222. https://doi.org/10.1016/j.future.2023.03.025

[11] 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. https://doi.org/10.1109/CCNC49033.2022.9700555

[12] Mirko Franco, Ombretta Gaggi, and Claudio E. Palazzi. 2023. Can Messaging Applications Prevent Sexting Abuse? A Technology Analysis. *IEEE Transactions on Mobile Computing* (2023), 1–14. https://doi.org/10.1109/TMC.2023.3238189

[13] Barbara Guidi. 2020. When Blockchain meets Online Social Networks. *Pervasive and Mobile Computing* 62 (2020), 101131. https://doi.org/10.1016/j.pmcj.2020.101131

[14] Barbara Guidi, Andrea Michienzi, and Andrea De Salve. 2020. Community evaluation in Facebook groups. *Multimedia Tools and Applications* 79 (2020), 19. https://doi.org/10.1007/s11042-019-08494-0

[15] 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. https://doi.org/10.1145/3479610

[16] 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. https://doi.org/10.1145/3462671

[17] 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* (Honolulu, HI, USA) *(CHI '20)*. Association for Computing Machinery, New York, NY, USA, 1–14. https://doi.org/10.1145/3313831.3376498

[18] Chi Liu, Tianqing Zhu, Jun Zhang, and Wanlei Zhou. 2022. Privacy Intelligence: A Survey on Image Privacy in Online Social Networks. *ACM Comput. Surv.* 55, 8, Article 161 (dec 2022), 35 pages. https://doi.org/10.1145/3547299

[19] 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* (Portland, OR, USA) *(CHI EA '05)*. Association for Computing Machinery, New York, NY, USA, 1985–1988. https://doi.org/10.1145/1056808.1057073

[20] OpenAI. 2023. GPT-4 Technical Report. arXiv:2303.08774 [cs.CL]

[21] Wei Qin, Zetong Chen, Lei Wang, Yunshi Lan, Weijieying Ren, and Richang Hong. 2023. Read, Diagnose and Chat: Towards Explainable and Interactive LLMs-Augmented Depression Detection in Social Media. arXiv:2305.05138 [cs.CL]

[22] Elisa Rubegni, Monica Landoni, Antonella De Angeli, and Letizia Jaccheri. 2019. Detecting Gender Stereotypes in Children Digital StoryTelling. In *Proceedings of the 18th ACM International Conference on Interaction Design and Children* (Boise, ID, USA) *(IDC '19)*. Association for Computing Machinery, New York, NY, USA, 386–393. https://doi.org/10.1145/3311927.3323156

[23] 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* (Hamburg, Germany) *(CHI '23)*. Association for Computing Machinery, New York, NY, USA, Article 391, 18 pages. https://doi.org/10.1145/3544548.3581538

[24] Ather Sharif, Aedan Liam McCall, and Kianna Roces Bolante. 2022. Should I Say "Disabled People" or "People with Disabilities"? Language Preferences of Disabled People Between Identity- and Person-First Language. In *Proceedings of the 24th International ACM SIGACCESS Conference on Computers and Accessibility* (Athens, Greece) *(ASSETS '22)*. Association for Computing Machinery, New York, NY, USA, Article 10, 18 pages. https://doi.org/10.1145/3517428.3544813

[25] 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* (Honolulu, HI, USA) *(CHI EA '20)*. Association for Computing Machinery, New York, NY, USA, 1–8. https://doi.org/10.1145/3334480.3383029

[26] Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurelien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. 2023. LLaMA: Open and Efficient Foundation Language Models. arXiv:2302.13971 [cs.CL]

[27] 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. https://doi.org/10.1145/3476059

[28] 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* (Baltimore, MD, USA) *(CODASPY '22)*. Association for Computing Machinery, New York, NY, USA, 65–76. https://doi.org/10.1145/3508398.3511517

[29] Wayne Xin Zhao, Kun Zhou, Junyi Li, Tianyi Tang, Xiaolei Wang, Yupeng Hou, Yingqian Min, Beichen Zhang, Junjie Zhang, Zican Dong, Yifan Du, Chen Yang, Yushuo Chen, Zhipeng Chen, Jinhao Jiang, Ruiyang Ren, Yifan Li, Xinyu Tang, Zikang Liu, Peiyu Liu, Jian-Yun Nie, and Ji-Rong Wen. 2023. A Survey of Large Language Models. arXiv:2303.18223 [cs.CL]