

ChatGPT, can you solve the content moderation dilemma?

Emmanuel Vargas Penagos^{*} 

ABSTRACT

This article conducts a qualitative test of the potential use of large language models (LLMs) for online content moderation. It identifies human rights challenges arising from the use of LLMs for that purpose. Different companies (and members of the technical community) have tested LLMs in this context, but such examinations have not yet been centred in human rights. This article, framed within EU law—particularly the EU Digital Services Act and the European human rights framework—delimits the potential challenges and benefits of LLMs in content moderation. As such, this article starts by explaining the rationale for content moderation at policy and practical levels, as well as the working of LLMs. It follows with a summary of previous technical tests conducted on LLMs for content moderation. Then, it outlines the results of a test conducted on ChatGPT and OpenAI's 'GPTs' service. Finally, it concludes with the main human rights implications identified in using LLMs for content moderation.

INTRODUCTION

In August 2023, the US tech news site The Verge reported that OpenAI seeks to 'solve the content moderation dilemma'¹ through the use of its large language model (LLM), ChatGPT.² This announcement attracted the attention of those in the Internet freedom field, as content 'moderation—the systems and rules that determine how [platforms] treat user-generated content on their services'³—is a critical issue in online speech regulation. But this article shows that the proposal is not as innovative as it seems. This is because big tech companies are already deploying LLMs as one of the key technical solutions for

^{*} Emmanuel Vargas Penagos, LLB (Unianandes), LLM (Amsterdam), PhD student, Örebro University, L2407, SE-701 Sweden
Tel: +46760781468, Email: Emmanuel.vargas-penagos@oru.se

The author thanks his supervisors, Martin Ebers, Katalin Kelemen and Alberto Giarretta for their support and guidance throughout this research, as well as to Andrew Leyden for his help with proofreading. Additionally, the author would like to thank Raissa Carrillo for her support and invaluable encouragement and Paula Castañeda for her support and her generous guidance in using Excel.

¹ Simon Hurtz, 'OpenAI Wants GPT-4 to Solve the Content Moderation Dilemma' (*The Verge*, 15 August 2023) <<https://www.theverge.com/2023/8/15/23833406/openai-gpt-4-content-moderation-ai-meta>> accessed 18 July 2024.

² 'Using GPT-4 for Content Moderation' (*OpenAI*, 15 August 2023) <<https://openai.com/index/using-gpt-4-for-content-moderation/>> accessed 20 July 2024.

³ Evelyn Douek, 'Content Moderation as Systems Thinking' (2022) 136 *Harvard Law Rev* 526, 528.

content moderation. Moreover, despite some benefits, LLMs are still far from being able to solve all content moderation challenges.

As human lives increasingly merge with the digital world, challenges to human rights are more notable.⁴ The United Nations Special Rapporteur on freedom of opinion and expression (the UN FoE Rapporteur) has highlighted the private sector's key role in governing freedom of expression on internet platforms.⁵ In the context of the governance of user-generated content on social media, the UN FoE Rapporteur has referred to the companies controlling digital platforms as 'enigmatic regulators, establishing a kind of 'platform law', in which clarity, consistency, accountability and remedy are elusive.'⁶ The crucial role of digital platforms has been emphasized at various policy levels. EU Commission President Ursula von der Leyen's political guidelines for the term of the 2019–24 Commission highlighted the need to ensure that digital platforms do not destabilize democracies.⁷ Similarly, in late 2023, UNESCO Director General Audrey Azoulay announced a plan to regulate social media, stressing the importance of protecting access to information, while safeguarding freedom of expression and human rights.⁸

Artificial Intelligence is often seen as essential for moderating the vast amounts of content on global platforms such as Facebook, Instagram, and X. However, automated content moderation systems also raise significant human rights concerns.⁹ Although the use of these technologies can have positive implications, such as the proper and swift functioning of services, there are risks arising from the lack of capacity of automated tools to understand the context surrounding specific publications; the lack of diversity within the datasets upon which such tools are trained (which may affect specific groups or communities); the possibility that datasets would replicate the biases of their trainers; the measures applied by some users to circumvent static systems (for instance, by making minor changes to images to avoid detection); and the growing need for resources and energy to expand the power of the automated models.¹⁰ Overall, some scholars consider that automated tools have distinct limitations that may affect their accuracy and transparency, affecting human rights.¹¹ These concerns have been regarded as enhancing the risk of 'over-removal' of online content.¹² They have also been regarded as potentially exacerbating the risk of allowing content affecting and harming specific groups or communities remaining undetected.¹³ Human oversight has therefore been considered as a basic safeguard,¹⁴ but flawed human oversight may also trigger the repetition and normalization of errors in moderation.¹⁵

⁴ Alessandro Mantelero, 'Fundamental Rights Impact Assessment in the DSA' in Joris van Hoboken and others (eds), *Putting the DSA into Practice* (Verfassungsbooks, Berlin, Germany 2023) <<https://doi.org/10.17176/20230208-093135-0>>.

⁵ UNCHR 'Report of the Special Rapporteur on the Promotion and Protection of the Right to Freedom of Opinion and Expression' (2018) UN Doc A/HRC/38/35.

⁶ *Ibid.*, para 1.

⁷ European Commission. Directorate-General for Communication, *Political Guidelines for the next European Commission 2019–2024; Opening Statement in the European Parliament Plenary Session 16 July 2019; Speech in the European Parliament Plenary Session 27 November 2019* (Publications Office 2020) <<https://data.europa.eu/doi/10.2775/101756>> accessed 18 July 2024.

⁸ 'Online Disinformation: UNESCO Unveils Action Plan to Regulate Social Media Platforms' (UNESCO, 6 November 2023) <<https://www.unesco.org/en/articles/online-disinformation-unesco-unveils-action-plan-regulate-social-media-platforms>> accessed 19 July 2024.

⁹ Emma Llansó and others, 'Artificial Intelligence, Content Moderation, and Freedom of Expression' [2020] Transatlantic Working Group <<https://www.ivir.nl/publicaties/download/AI-Llanso-Van-Hoboken-Feb-2020.pdf>> accessed 10 April 2024.

¹⁰ UNGA 'Report of the Special Rapporteur on the Promotion and Protection of the Right to Freedom of Opinion and Expression' (2018) UN Doc A/73/348; Llansó and others (n 9).

¹¹ Althaf Marsoof and others, 'Content-Filtering AI Systems—Limitations, Challenges and Regulatory Approaches' (2023) 32 *Inform Commun Technol Law* 64.

¹² UNGA (n 10); Daphne Keller, 'Internet Platforms: Observations on Speech, Danger, and Money' (Hoover Institution's Aegis Paper Series 2018) 1807.

¹³ UNCHR 'Disinformation and Freedom of Opinion and Expression - Report of the Special Rapporteur on the Promotion and Protection of the Right to Freedom of Opinion and Expression, Irene Khan' (2021) UN Doc A/HRC/47/25.

¹⁴ Thiago Dias Oliva, 'Content Moderation Technologies: Applying Human Rights Standards to Protect Freedom of Expression' (2020) 20 *Human Rights Law Rev* 607.

¹⁵ Keller (n 12).

As such, the use of algorithmic decision-making poses challenges in relation to *fairness*.¹⁶ Such considerations of ‘fairness’, aside from their own intrinsic effects on individuals’ rights, can/have an impact on the legitimacy of the decision-making system.¹⁷ Alongside this, the inadequate functioning of the system endangers the democratic need for ‘a favourable environment for participation in public debate by all the persons concerned, enabling them to express their opinions and ideas without fear’.¹⁸ Although the state is primarily responsible for creating this environment, social media platforms play a key role due to their capacity to facilitate or obstruct access to online forums of public debate.¹⁹ Moreover, the ability of every individual to participate in public debate will not only be affected by undue removals by moderation systems but also by allegedly not moderating content that is so ‘excluding or aggressive towards some groups that it effectively chills their speech’.²⁰ To put it differently, ‘speech may be used to attack, harass, and silence as much as it is used to enlighten’, which may turn in the weaponization of speech to suppress the speech of others.²¹

Despite the buzz surrounding OpenAI’s announcement, their proposal is not entirely new. Companies such as Meta (along with the technical community more generally) have already tested the use of LLMs for content moderation purposes in recent years.²² Trained on vast datasets to recognize patterns and predict text, LLMs can generate answers to queries²³ and classify text.²⁴ They are useful for understanding text, which historically has been the primary means of online and social media communication (alongside images, videos, and audio in different degrees).

Lawmakers, NGOs, and experts increasingly advocate for content moderation to be viewed from a human rights perspective. In the European Union, the Digital Services Act (DSA)²⁵ provides a set of obligations, mainly in procedural and due diligence terms. In parallel, academic literature highlights the usefulness of International Human Rights Law (IHRL) in addressing concerns with algorithmic decision-making and online content moderation. Although not standardized globally, IHRL sets obligations for states to respect and guarantee rights, increasingly serving as a standard for corporate conduct.²⁶ It also provides a framework for assessing the risks of harm when rights are violated. In the EU, Frosio and Geiger argue that balancing rights under the DSA will require using the EU Charter²⁷ and the European

¹⁶ Vincent Chiao, ‘Fairness, Accountability and Transparency: Notes on Algorithmic Decision-Making in Criminal Justice’ (2019) 15 *Int J Law Context* 126; Doaa Abu Elyounes, ‘Bail or Jail?’ (2020) *Sci Technol Law Rev* 376; Pedro Rubim Borges Fortes, ‘Paths to Digital Justice: Judicial Robots, Algorithmic Decision-Making, and Due Process’ (2020) 7 *Asian J Law Soc* 453.

¹⁷ Omer Tene and Jules Polonetsky, ‘Taming The Golem: Challenges of Ethical Algorithmic Decision-Making’ (2018) 19 *North Carolina J Law Technol* 125; Maja Brkan, ‘Do Algorithms Rule the World? Algorithmic Decision-Making and Data Protection in the Framework of the GDPR and Beyond’ (2019) 27 *Int J Law Inform Technol* 91; Rubim Borges Fortes (n 16).

¹⁸ *Khadija Ismayilova v Azerbaijan* [2019] ECtHR 65286/13, 57270/14, para 158.

¹⁹ Tarlach McGonagle, ‘The Council of Europe and Internet Intermediaries: A Case Study of Tentative Posturing’ in Rikke Frank Jørgensen (ed), *Human Rights in the Age of Platforms* (The MIT Press, Cambridge, Massachusetts, USA).

²⁰ Anjalee De Silva and Andrew T Kenyon, ‘Countering Hate Speech in Context: Positive Freedom of Speech’ (2022) 14 *J Media Law* 97.

²¹ Tim Wu, ‘Is the First Amendment Obsolete?’ (2018) *Michigan Law Rev* 547, 549.

²² ‘How Facebook Uses Super-Efficient AI Models to Detect Hate Speech’ (*MetaAI*, 19 November 2019) <<https://ai.meta.com/blog/how-facebook-uses-super-efficient-ai-models-to-detect-hate-speech/>> accessed 20 July 2024; ‘RoBERTa: An Optimized Method for Pretraining Self-Supervised NLP Systems’ (*MetaAI*, 29 July 2020) <<https://ai.meta.com/blog/roberta-an-optimized-method-for-pretraining-self-supervised-nlp-systems/>> accessed 20 July 2024.

²³ Giuseppe Sartori and Graziella Orrù, ‘Language Models and Psychological Sciences’ (2023) 14 *Front Psychol* 1279317.

²⁴ Mohit Singhal and others, ‘SoK: Content Moderation in Social Media, from Guidelines to Enforcement, and Research to Practice’, *2023 IEEE 8th European Symposium on Security and Privacy (EuroS&P)* (IEEE 2023) <<https://ieeexplore.ieee.org/document/10190527/>> accessed 19 July 2024.

²⁵ Regulation (EU) 2022/2065 of the European Parliament and of the Council of 19 October 2022 on a Single Market for Digital Services and amending Directive 2000/31/EC (DSA) (Text with EEA relevance) 2022 (OJ L).

²⁶ Lorna McGregor, Daragh Murray and Vivian Ng, ‘International Human Rights Law as a Framework for Algorithmic Accountability’ (2019) 68 *Int Comp Law Quart* 309; Evelyn Douek, ‘The Limits of International Law in Content Moderation’ (2021) 6 *UC Irvine Journal of International, Transnational, and Comparative Law* <<https://escholarship.org/uc/item/2857f1jq>> accessed 18 July 2024.

²⁷ Charter of Fundamental Rights of the European Union 2012 (OJ C).

Convention on Human Rights (ECHR)²⁸ as interpreted by the CJEU and the European Court of Human Rights (ECtHR), with the view of achieving a coherent analytical framework.²⁹ Moreover, the EU treaties³⁰ and Charter³¹ reference the ECHR as a key instrument within the Union, emphasizing the application of the standards of human rights protection it requires. This is underpinned by CJEU case law, which establishes that Article 11 of the EU Charter and Article 10 ECHR (both focusing on freedom of expression) have the same meaning and scope.³²

The purpose of this article is to qualitatively test the performance of LLMs in online content moderation and to outline the human rights issues this potential use raises. It focuses on how these issues impact the environment for public participation in discourse by addressing the following question: *what Human Rights challenges for the participation in public debate can be identified from the use of LLMs in online content moderation activities?*

To better answer this question, the article will also consider these sub-questions: (i) *in what ways can the use of automated online content moderation (in our case the use of LLMs in particular) ensure free participation in public debate online? In other words, does this method guarantee a broader participation in public debate in comparison to other methods in any way and (ii) Which risks does the use of automated content moderation by LLMs pose to the free participation in public debate online?*

This article is divided into four parts. First, it discusses the context in which automated content moderation operates, the reasons for the deployment of such solutions by companies, and the legal and human rights issues they raise, along with EU policy and legal requirements. Second, it examines LLMs as a specific technical solution for content moderation from the perspective of the technical community, looking into technical studies on the subject matter. Third, the article includes an analysis of an empirical test conducted for the purposes of this paper, using ChatGPT, and a tailored version of the ‘GPTs’ service offered by Open AI. The documentation and the full set of statistics relating to this test are available in an [Supplementary Appendix](#)³³ Since a comprehensive analysis of all social media moderation tools is impractical due to time and cost constraints, this article addresses current issues in state-of-the-art technology by looking into LLMs as a groundbreaking tool.³⁴ Finally, it concludes by identifying the advantages and disadvantages of LLMs in content moderation, translating these into substantive and procedural human rights challenges,³⁵ particularly due to the relevance of the latter in the DSA framework.³⁶

THE CONTENT MODERATION DILEMMA

The moderation of online content is not limited to the take-down of illegal posts on social media due to it containing subjects such as hate speech, or an incitement to terrorism. A relevant portion of what constitutes ‘moderation’ entails assessing content that is ‘lawful but awful’. In other words, it involves content that ‘is offensive or morally repugnant to many people but protected’

²⁸ Convention for the Protection of Human Rights and Fundamental Freedoms (ECHR).

²⁹ Giancarlo Frosio and Christophe Geiger, ‘Taking Fundamental Rights Seriously in the Digital Services Act’s Platform Liability Regime’ (2023) 29 Eur Law J 31.

³⁰ Consolidated version of the Treaty on European Union 2016; Consolidated version of the Treaty on the Functioning of the European Union 2016 art 6(3); Consolidated version of the Treaty on the Functioning of the European Union Protocol 8.

³¹ Charter of Fundamental Rights of the European Union, art 53.

³² Case C-401/19 *Poland v Parliament and Council* [2022] ECLI:EU:C:2022:29, para 44.

³³ The appendix is available in this link: <https://github.com/kenobito/LLMs-as-moderators>.

³⁴ Singhal and others (n 24).

³⁵ McGregor, Murray and Ng (n 26).

³⁶ Judit Bayer, ‘Procedural Rights as Safeguard for Human Rights in Platform Regulation’ (2022) 14 Policy Internet 755; Pietro Ortolani, ‘If You Build It, They Will Come The DSA “Procedure Before Substance” Approach’ in Joris van Hoboken and others (eds), *Putting the DSA into Practice* (Verfassungsbooks 2023, Berlin, Germany) <<https://doi.org/10.17176/20230208-093135-0>>.

by freedom of expression.³⁷ This distinction is relevant in terms of the rights of users. The DSA, for example, provides a right for individuals to file notices concerning the illegality of content, but the possibility of doing so for merely ‘lawful but awful’ content ‘depends on the platforms, which remain free to determine the purview of user affordances’³⁸

In practice, content moderation by social media platforms is achieved via a mixture of both human and automated review. Humans are used either to review the content themselves or with the aid of a wide and increasing portfolio of automated tools which can, at the same time, moderate without any human involvement.³⁹ A layer that is added to that structure is the subsequent review performed when the decision is contested by unsatisfied users.⁴⁰

Previous research shows that moderation systems use various processes and tools, including keyword filters, image-based filters, and machine learning techniques.⁴¹ For instance, Jahan and Oussalah found that tools in widespread use for the detection of hate speech in text include support vector machines, deep learning models, logistic regression, and naive Bayesian models.⁴² Similar tools are used for detecting misinformation.⁴³

The DSA aims to protect rights by requiring platforms to adopt procedural safeguards for moderation.⁴⁴ Article 14(4) DSA requires platforms to apply terms and conditions ‘diligently, objectively, and proportionately’ while respecting users’ fundamental rights, including freedom of expression. Critics argue that this outsources human rights protection to platforms by asking them to undertake the balancing of rights, and that it relies heavily on user complaints for redress.⁴⁵ This reflects a trend of delegating content moderation power to platforms to in effect decide on the ‘legality’ of online content.⁴⁶ The DSA also requires risk assessments, mitigation measures, and crisis response mechanisms for very large online platforms (VLOPs) to address impacts on fundamental rights. Recital 47 urges intermediary services to follow international human rights standards.⁴⁷ Although not legally binding, Recitals to EU legislation guide interpretation and best practices.⁴⁸ Moreover, the CJEU emphasizes the need for greater safeguards when interferences stem from automated processes.⁴⁹

In the DSA era, mandatory reports by platforms may provide deeper insights into the use of automated content moderation methods. Article 15(1)(e) requires platforms to report on an annual basis their use of such methods, detailing their purposes, accuracy, error rates, and the safeguards applied. However, a look at the first reports published by some of those

³⁷ Daphne Keller, ‘Lawful but Awful? Control over Legal Speech by Platforms, Governments, and Internet Users’ (*University of Chicago Law Review Online*, 2022) <<https://lawreviewblog.uchicago.edu/2022/06/28/keller-control-over-speech/>> accessed 17 July 2024. This ‘lawful but awful’ category of content entails several undefined terms that may sometimes be used either as synonyms or with separate definitions, such as ‘toxic,’ ‘offensive,’ ‘hateful’ or ‘harmful’. References to these terms within this article are as they are used in the specific source that is cited at each moment, but with the understanding that they will generally fall within the ‘lawful but awful’ category. A similar issue may happen with terms like ‘fake news,’ ‘disinformation,’ ‘misinformation’ and ‘propaganda,’ which, although being defined by bodies like the UN Special Rapporteur, don’t have a common legal definition. For the purposes of this article, the latter terms are used under the umbrella of ‘misleading content’.

³⁸ Ortolani (n 36).

³⁹ Douek (n 3).

⁴⁰ *ibid.*

⁴¹ Lllansó and others (n 9).

⁴² Md Saroar Jahan and Mourad Oussalah, ‘A Systematic Review of Hate Speech Automatic Detection Using Natural Language Processing’ (2023) 546 *Neurocomputing* 126232.

⁴³ Singhal and others (n 24).

⁴⁴ Ortolani (n 36).

⁴⁵ Sentleben, Martin; Quintais, João Pedro; Meiring, Arlette; ‘How the European Union Outsources the Task of Human Rights Protection to Platforms and Users: The Case of User-Generated Content Monetization’ (2023) 38 *Berkeley Technol Law J* 933.

⁴⁶ Bayer (n 36).

⁴⁷ DSA, recital 47.

⁴⁸ Todas Klimas and Jurate Vaiciukaite, ‘The Law of Recitals in European Community Legislation’ (2008) 15 *ILSA J Int Comp Law* 61; Maarten Den Heijer, Teun Van Os Van Den Abeelen and Antanina Maslyka, ‘On the Use and Misuse of Recitals in European Union Law’ (2019) *Amsterdam Center for International Law* <<https://www.ssrn.com/abstract=3445372>> accessed 18 July 2024.

⁴⁹ *Republic of Poland v European Parliament and Council of the European Union* (n 32), para 67.

platforms qualified as VLOPs in late 2023 and early 2024 shows that the information given by companies may fall short of the requisite standard of specificity. For instance, the transparency reports issued by Meta concerning Instagram and Facebook are limited to stating that they use ‘machine learning classifiers’.⁵⁰ On its side, X’s report states that it uses ‘heuristics and machine learning algorithms’,⁵¹ along with ‘combinations of natural language processing models, image processing models and other sophisticated machine learning methods’.⁵² TikTok’s report describes the use of ‘computer vision models’,⁵³ ‘keyword lists and models’⁵⁴ and ‘de-duplication and hashing technologies’.⁵⁵ Google’s report refers to a ‘combination of automated and human tools’,⁵⁶ ‘machine learning technology’,⁵⁷ ‘machine learning classifiers’,⁵⁸ ‘hash-matching technology’,⁵⁹ ‘smart detection technology’⁶⁰ and ‘Content ID’.⁶¹ A more detailed review of the reports by these and the other twelve platforms classified as VLOPs and Very Large Online Search Engines would require separate, detailed research. But these examples alone are sufficient to show that companies may be reluctant to disclose the specific tools deployed in their moderation systems.

Regardless of the structure of the moderation system or the type of technology at issue, human and algorithmic moderators ‘struggle with the nuanced judgments they are required to make’ and, irrespective of how skilled they are, face difficulties in applying freedom of expression standards with the precision they are demanded to have.⁶² Furthermore, the difficulty of this task, when placed in hands of moderators of ‘flesh and blood,’ is exacerbated, as they are performing an ‘arduous and trauma-inducing job’⁶³ comprising the examination of many types of content, which could well be violent, sexual, or degrading. The Council of Europe has emphasized that a human-rights based approach to moderation must consider the labour rights and mental health of workers involved in manual content review.⁶⁴

Moderation of social media content is a new chapter in the automation of industrial processes, where automation reduces human workload, but makes human oversight more complex.⁶⁵ This dynamic is evident when it comes to the application of AI, a technology that is far from replacing humans in several fields. It is therefore something that has called for an approach whereby it is conceptualized as a support system.⁶⁶ Evelyn Douek suggests content moderation should be seen from a systems thinking approach, with focus on the design of moderation systems and allocation of functions and procedures, not merely on the resolution of individual

⁵⁰ Meta, ‘Regulation (EU) 2022/2065 Digital Services Act Transparency Report for Facebook’ (Meta 2023) <<https://transparency.meta.com/sr/dsa-transparency-report-oct2023-facebook/>>; Meta, ‘Regulation (EU) 2022/2065 Digital Services Act Transparency Report for Instagram’ (Meta 2023) <https://scontent-arn2-1.xx.fbcdn.net/v/t39.8562-6/447971060_1481740992549061_1404827436118992020_n.pdf?_nc_cat=100&ccb=1-7&_nc_sid=b8d81d&_nc_ohc=sh73tBNSQgQQ7kNvgFbfDdbd&_nc_ht=scontent-arn2-1.xx&oh=00_AYDZ1a7V08NqAotRDE669f6uL9682Aa5skNqBPhHhk4kiA&oe=66A202AC>.

⁵¹ ‘DSA Transparency Report - April 2024’ (X, April 2024) <<https://transparency.x.com/dsa-transparency-report.html>> accessed 20 July 2024.

⁵² *ibid.*

⁵³ ‘TikTok’s DSA Transparency Report 2023’ (TikTok 2023) <<https://www.tiktok.com/transparency/en/dsa-transparency/>>.

⁵⁴ *ibid.*

⁵⁵ *ibid.*

⁵⁶ ‘EU Digital Services Act (EU DSA) Biannual VLOSE/VLOP Transparency Report’ (Google 2023) <https://storage.googleapis.com/transparencyreport/report-downloads/pdf-report-27_2023-8-28_2023-9-10_en_v1.pdf>.

⁵⁷ *ibid.*

⁵⁸ *ibid.*

⁵⁹ *ibid.*

⁶⁰ *ibid.*

⁶¹ *ibid.*

⁶² Kyle Langvardt, ‘Regulating Online Content Moderation’ 106 *Georgetown Law J* 1353.

⁶³ *ibid.*

⁶⁴ ‘Content Moderation—Best Practices towards Effective Legal and Procedural Frameworks for Self-Regulatory and Co-Regulatory Mechanisms of Content Moderation’ (Council of Europe 2021) Guidance Note <<https://rm.coe.int/content-moderation-en/1680a2c18>>.

⁶⁵ Lisanne Bainbridge, ‘Ironies of Automation’ (1983) 19 *Automatica* 775.

⁶⁶ Mica R Endsley, ‘Ironies of Artificial Intelligence’ (2023) 66 *Ergonomics* 1666.

cases.⁶⁷ The DSA requires safeguards for automated moderation, including supervision by qualified staff for internal complaints (Article 20(6)). Griffin and Stallman argue this requirement should aim to create systematic oversight by knowledgeable staff, enhancing communication and the training of automated systems.⁶⁸

The complexity of the structures behind moderation as a whole shows that the use of LLMs like ChatGPT for the purposes of content moderation should be approached with some degree of precaution. Instead of seeing LLMs or other automated tools as a ‘magic bullet’ that will solve the issue by themselves, the analysis ought better to focus, holistically, on the entire system.

PREVIOUS RESEARCH ON LLMs AND THEIR POTENTIAL FOR MODERATION

How do LLMs work?

Despite their recent prominence due to the release of ChatGPT in November 2022, LLMs are part of a field dating back to at least 1967, Natural Language Processing (NLP).⁶⁹ They use deep neural networks for understanding, generating, and processing human language, and their modern place in the public agenda came after a 2017 breakthrough, namely the release of the self-attention mechanism and transformer architecture.⁷⁰ These two concepts describe the parallel computation of word meanings within a sentence across several encoding/decoding layers to generate outputs.⁷¹ Encoding assigns labels to words/subwords to determine their place in the language universe.⁷² These innovations were crucial for developing pre-trained language models (PLMs), which are pre-trained on datasets to enhance text understanding and fine-tuned for specific tasks.⁷³

This architecture enabled the scale-up of model training, allowing the creation of models with the ability to process and learn from large datasets.⁷⁴ LLMs are pre-trained on extensive datasets, including books, websites like Reddit and Wikipedia, and databases like Commoncrawl.⁷⁵ Their pre-training involves predicting words in context to refine the model’s parameters⁷⁶ and their fine-tuning entails new instructions to enhance reasoning in task abilities, along with alignment with human values, such as reducing ‘toxicity’.⁷⁷

LLMs’ complexity resulted in a remarkable capacity for various tasks, including in-context learning (ICL), whereby models learn to perform tasks after a demonstration without additional training or fine-tuning. This mimics human learning by analogy.⁷⁸ Consequently, LLMs

⁶⁷ Douek (n 3).

⁶⁸ Rachel Griffin and Erik Stallman, ‘A Systemic Approach to Implementing the DSA’s Human-in-the-Loop Requirement’ (22 February 2024) <<https://verfassungsblog.de/a-systemic-approach-to-implementing-the-dsas-human-in-the-loop-requirement/>> accessed 18 July 2024.

⁶⁹ G Pradeep Reddy, YV Pavan Kumar and K Purna Prakash, ‘Hallucinations in Large Language Models (LLMs)’, 2024 *IEEE Open Conference of Electrical, Electronic and Information Sciences (eStream)* (IEEE 2024) <<https://ieeexplore.ieee.org/document/10542617/>> accessed 19 July 2024.

⁷⁰ *ibid.*

⁷¹ *ibid.*

⁷² Bonan Min and others, ‘Recent Advances in Natural Language Processing via Large Pre-Trained Language Models: A Survey’ (2024) 56 *ACM Comput Surveys* 1.

⁷³ *ibid.*

⁷⁴ *ibid.*

⁷⁵ Stephen Roller and others, ‘Recipes for Building an Open-Domain Chatbot’, *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume* (Association for Computational Linguistics 2021) <<https://aclanthology.org/2021.eacl-main.24>> accessed 19 July 2024; Mohaimenul Azam Khan Raiaan and others, ‘A Review on Large Language Models: Architectures, Applications, Taxonomies, Open Issues and Challenges’ (2024) 12 *IEEE Access* 26839.

⁷⁶ Thomas Wang and others, ‘What Language Model Architecture and Pretraining Objective Works Best for Zero-Shot Generalization?’, *Proceedings of the 39th International Conference on Machine Learning (PMLR 2022)* <<https://proceedings.mlr.press/v162/wang22u.html>> accessed 19 July 2024.

⁷⁷ Dan Hendrycks and others, ‘Aligning AI With Shared Human Values’ (2020) <https://openreview.net/forum?id=dNy_RKzjaCY> accessed 18 July 2024.

⁷⁸ Sivaramkrishnan Swaminathan and others, ‘Schema-Learning and Rebinding as Mechanisms of in-Context Learning and Emergence’ (2024) 36 *Advances in Neural Information Processing Systems* <https://proceedings.neurips.cc/paper_files/paper/2023/hash/Sbc3356e0fa1753fff7e8d6628e71b22-Abstract-Conference.html> accessed 19 July 2024.

Table 1. Examples of Zero-shot, one-shot and few-shot prompting.

Type of prompt	Example
Zero-shot	Rules: Law x punishes those who issue statements accusing someone of committing a criminal offence Detect the violation to Law x: Gaga insulted me and then pointed at me with a gun
One-shot	Rules: Law x punishes those who issue statements accusing someone of committing a criminal offence Example 1: Britney used to dress very provocative clothes during the days and, during the night, she would sneak behind strangers to assault them Violation: accusing Britney of sneaking behind strangers to assault them Detect the violation to Law x: Gaga insulted me and then pointed at me with a gun
Few-shot	Rules: Law x punishes those who issue statements accusing someone of committing a criminal offence Example 1: Britney used to dress very provocative clothes during the days and, during the night, she would sneak behind strangers to assault them Violation: accusing Britney of sneaking behind strangers to assault them Example 2: Yandel was always known for his long unclean beard and his tendency to harass employees Violation: accusing Yandel of harassing employees Detect the violation to Law x: Gaga insulted me and then pointed at me with a gun

can respond to zero-shot, one-shot, and few-shot prompting. Zero-shot prompting involves performing a task with only a description. One-shot prompting requires a single demonstration. Few-shot prompting requires a few demonstrations.⁷⁹ (see Table 1 for an illustration).

The in-context learning capacities that LLMs display are relevant for obtaining answers to complex tasks that require some degree of reasoning. Within this framework, researchers have found that LLMs can have a good performance with *chain of thought prompting*. This could be something as simple as including the phrase ‘let’s think step by step’ in the prompt (see Fig. 1 for an illustration).⁸⁰ There has been further research on how to get the best out of *chain-of-thought prompting*, which includes strategies as simple as using the words ‘take a deep breath and work on this problem step-by-step’, using symbols, or by using prompts with more complex structured reasoning, among many others.⁸¹

This must, however, be approached with caution. Researchers have signalled that, although the *chain-of-thought* emulates the reasoning of humans, ‘this does not answer whether the neural network is actually reasoning.’⁸²

Using LLMs in content moderation

LLMs are increasingly used for content moderation by major companies. Google’s Perspective API, a transformer model, detects toxic content in 12 languages.⁸³ Meta AI developed and uses

⁷⁹ Tom Brown and others, ‘Language Models Are Few-Shot Learners,’ *Advances in Neural Information Processing Systems* (Curran Associates, Inc., Red Hook, NY, USA, 2020) <<https://papers.nips.cc/paper/2020/hash/1457c0d6bfc4967418bfb8ac-142f64a-Abstract.html>> accessed 18 July 2024.

⁸⁰ Jason Wei and others, ‘Chain-of-Thought Prompting Elicits Reasoning in Large Language Models’ (2022) 35 *Adv Neural Inform Process Syst* 24824.

⁸¹ Brown and others (n 79).

⁸² Wei and others (n 80); Jie Huang and Kevin Chen-Chuan Chang, ‘Towards Reasoning in Large Language Models: A Survey,’ *Findings of the Association for Computational Linguistics: ACL 2023* (Association for Computational Linguistics 2023) <<https://aclanthology.org/2023.findings-acl.67>> accessed 19 July 2024.

⁸³ Alyssa Lees and others, ‘A New Generation of Perspective API: Efficient Multilingual Character-Level Transformers,’ *Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining* (ACM 2022) <<https://dl.acm.org/doi/10.1145/3534678.3539147>> accessed 19 July 2024.

The sanction for tax fraud in Spain is up to 5 years. Colombia does not sanction tax fraud with imprisonment. Shakira, a citizen from Colombia, is accused by the Spanish government of having committed tax fraud while living in Spain. what will happen to Shakira if proven guilty?



If Shakira is proven guilty of tax fraud in Spain, she would be subject to the Spanish legal system and its penalties, regardless of her Colombian citizenship. In Spain, tax fraud is considered a serious offense and can result in significant penalties, including imprisonment. The fact that Colombia does not sanction tax fraud with imprisonment does not affect how Spain handles the case, as Shakira's alleged fraud occurred while she was residing in Spain and under Spanish jurisdiction.

Specifically, if found guilty, Shakira could face:

1. **Imprisonment:** Spanish law allows for imprisonment of up to 5 years for serious tax fraud offenses.
2. **Fines:** She could be required to pay substantial fines and back taxes, including interest and penalties.
3. **Other Penalties:** There may be additional sanctions, such as being barred from certain rights or activities.

Therefore, Shakira would face the legal consequences as determined by the Spanish courts, which can include imprisonment, despite her Colombian nationality and the differences in how Colombia handles tax fraud.

Figure 1. Example of chain of thought prompting.

tools like RoBERTa and XLM-R, which are transformer-based and trained on text in 100 languages.⁸⁴ They also developed the 'linformer' mechanism to reduce computational complexity.⁸⁵ XLM-R is the basis for Bumble Inc.'s 'Rude message detector' in their Badoo app, which labels messages of a sexual, insulting, or hate nature.⁸⁶

⁸⁴ Alexis Conneau and others, 'Unsupervised Cross-Lingual Representation Learning at Scale' (arXiv, 7 April 2020) <<http://arxiv.org/abs/1911.02116>> accessed 18 July 2024.

⁸⁵ Sinong Wang and others, 'Linformer: Self-Attention with Linear Complexity' (arXiv, 14 June 2020) <<http://arxiv.org/abs/2006.04768>> accessed 19 July 2024.

⁸⁶ Massimo Belloni, 'Multilingual Message Content Moderation at Scale' (*Medium*, 8 December 2021) <<https://medium.com/bumble-tech/multilingual-message-content-moderation-at-scale-ddd0da1e23ed>> accessed 18 July 2024; Massimo Belloni, 'Multilingual Message Content Moderation at Scale' (*Medium*, 24 May 2022) <<https://medium.com/bumble-tech/multilingual-message-content-moderation-at-scale-7ea562e29e25>> accessed 18 July 2024.

Notably, models based on Google's BERT have been considered superior in comparison to other deep learning models and have achieved top performance in multilingual tasks in several studies.⁸⁷ Moreover, as surveyed by Zampieri *et al.*⁸⁸, most of the best teams at competitions on the detection of offensive content organized by the technical community have used models based on BERT.

It should be noted that so-called 'low-resourced' languages, namely those without a strong online presence, are usually not part of the dataset in which LLMs are built and, in any case, even when they are able to perform in those languages, they do it by making connections to 'high-resourced' languages (predominantly English). This potentially implies 'importing English-language assumptions and viewpoints'⁸⁹ into a completely different context. Although this is not an issue examined in depth in this article, it nonetheless raises human rights concerns and will be referenced in the final section.

Previous research using LLMs for content moderation

Given their increasing importance in content moderation, researchers are now studying how well LLMs perform in these tasks. Among those is the one conducted by Gilardi *et al.*⁹⁰, showing that ChatGPT can display very high performance in text classification. As such, Gilardi *et al.* found that ChatGPT, when given zero-shot prompts for text labelling (deciding if content fits into six topic categories), can perform better than human workers trained for the task.⁹¹

Additional research offers more insights into the pros and cons of using LLMs for moderation. For this article, this research is divided into two categories: one relates to the ability of detecting content that can be harmful *per se* within several sub-categories and classifications (eg, illegal categories like hate speech, or 'lawful but awful' categories like harmful, toxic, pornographic) and the other relates to the detection of misleading content (eg misinformation, disinformation, and propaganda).

Detection of 'harmful *per se*' content

Research on detecting 'harmful *per se*' content often focuses on deciding if specific content fits into a category defined as undesirable by researchers. Detection of this content is not necessarily related to the truthfulness of the user's message.

An example of this first category is the research of Li *et al.*, who also found that ChatGPT can outperform human annotators, but focused on the detection of harmful content.⁹² As such, these researchers found that ChatGPT was able to accurately and consistently apply definitions provided by them for hateful, offensive and toxic content, and noted that performance would vary depending on the type of prompt provided. In that sense, the experiments conducted by Li *et al.* found that ChatGPT tended to perform better when asked to provide a probabilistic answer (namely how likely is a specific piece of content to fall within one of the mentioned categories) instead of a binary one.⁹³ However, they also noted that, in the probabilistic set, ChatGPT very rarely gave intermediate values for classification and tended to choose values in the extreme opposites (e.g. highly likely or extremely unlikely). They also noted that, when

⁸⁷ Jahan and Oussalah (n 42).

⁸⁸ Marcos Zampieri and others, 'OffensEval 2023: Offensive Language Identification in the Age of Large Language Models' (2023) 29 Nat Lang Eng 1416.

⁸⁹ Gabriel Nicholas and Aliya Bhatia, 'Toward Better Automated Content Moderation in Low-Resource Languages' (2023) 2 J Online Trust Safety <<https://www.tsjournal.org/index.php/jots/article/view/150>> accessed 19 July 2024.

⁹⁰ Fabrizio Gilardi, Meysam Alizadeh and Maël Kubli, 'ChatGPT Outperforms Crowd Workers for Text-Annotation Tasks' (2023) 120 Proc Natl Acad Sci e2305016120.

⁹¹ *ibid.*

⁹² Lingyao Li and others, '"HOT" ChatGPT: The Promise of ChatGPT in Detecting and Discriminating Hateful, Offensive, and Toxic Comments on Social Media' (2024) 18 ACM Trans Web 1.

⁹³ *ibid.*

asked to provide reasons for its answer, it had a tendency to replicate the language of the definitions provided, and that it had a lower tendency to apply extreme classifications, but also that it had a higher tendency to classify content within the hateful, offensive or toxic classifications. The authors did not propose a reason for these variations.

Going further, it should be noted that Zampieri *et al.* tested six state-of-the-art LLMs developed between 2022 and 2023⁹⁴ with *zero-shot prompting* and two task-specific trained BERT models against the top three models of two computer science competitions of 2019 and 2020, which were mostly BERT-based.⁹⁵ They found that, although the six LLMs and the two BERT models were notably competitive, they were still outperformed by the top models of the 2019 and 2020 competitions. According to those researchers, the results they found suggest that LLMs are still not at the level of transformer models trained for the specific purpose of moderation (i.e. data related to the specific tasks, problems or contexts).⁹⁶

Kumar *et al.* examined the performance of five state of the art widely available LLMs⁹⁷ in two tasks: rule based moderation and detection of toxic content.⁹⁸ For the first task, they prompted GPT-3.5 with the rules from 95 Reddit subcommunities separately and asked it to determine if it would or would not moderate the content. For the second task, they prompted each of the LLMs with the definition of toxic content provided by Google Jigsaw⁹⁹ and then asked the tools to rate the toxicity of the comments on a scale from 1 to 10 and to provide an explanation in natural language. The sample to be moderated on the first task consisted of a set of comments that had passed through moderation processes in the 95 subreddits¹⁰⁰ and the second task took comments from a large dataset of comments from Reddit, Twitter, and 4chan. The baseline for their test was the original outcome of the moderation in Reddit.

The overall performance of GPT-3.5 in the first of the mentioned tasks is partly promising. Although the overall median accuracy¹⁰¹ of the LLM was 63.7% and the median precision¹⁰² was 83%, the median recall¹⁰³ was 39.8%. This would indicate that, although GPT-3.5. was right 63.7% of the time and 83% of the flagged content was truly flaggable, it had just spotted 39% of the cases it should have caught. However, it should be noted that the test showed that performance would vary significantly from one subreddit to another and its worst performance in terms of accuracy was in subreddits dedicated to discussions with experts that, as Kumar and others point out, require significant world context.¹⁰⁴

The researchers tested three subreddits with additional prompting strategies: embedding Reddit's platform rules and using chain-of-thought prompting. These strategies only marginally improved outcomes. They found GPT-3.5 failures were more prone to false negatives, with

⁹⁴ Falcon-7B-Instruct, RedPajama-INCITE-7B-Instruct, MPT-7B-Instruct, Llama-2-7B-Chat, T0-3B, Flan-T5-large.

⁹⁵ Zampieri and others (n 88).

⁹⁶ *ibid.*

⁹⁷ GPT-3, GPT-3.5, GPT-4, Gemini Pro and LLAMA 2.

⁹⁸ Deepak Kumar, Yousef Anees AbuHashem and Zakir Durumeric, 'Watch Your Language: Investigating Content Moderation with Large Language Models' (2024) 18 Proc Int AAAI Conf Web Social Media 865.

⁹⁹ 'a rude, disrespectful, or unreasonable comment that is likely to make someone leave a discussion'. *ibid.*

¹⁰⁰ Subreddits are communities within Reddit dedicated to specific topics and where users can post and interact with each other. Subreddits are created and moderated by users. 'What Are Communities or "Subreddits"? (Reddit Help, 19 July 2024) <<https://support.reddithelp.com/hc/en-us/articles/204533569-What-are-communities-or-subreddits>> accessed 25 July 2024.

¹⁰¹ Accuracy 'is a metric that measures how often a machine learning model correctly predicts the outcome. You can calculate accuracy by dividing the number of correct predictions by the total number of predictions'. 'Accuracy vs. Precision vs. Recall in Machine Learning: What's the Difference?' (EvidentlyAI) <<https://www.evidentlyai.com/classification-metrics/accuracy-precision-recall>> accessed 18 July 2024.

¹⁰² Precision 'is a metric that measures how often a machine learning model correctly predicts the positive class. You can calculate precision by dividing the number of correct positive predictions (true positives) by the total number of instances the model predicted as positive (both true and false positives)'. *ibid.*

¹⁰³ Recall 'is a metric that measures how often a machine learning model correctly identifies positive instances (true positives) from all the actual positive samples in the dataset. You can calculate recall by dividing the number of true positives by the number of positive instances. The latter includes true positives (successfully identified cases) and false negative results (missed cases)'. *ibid.*

¹⁰⁴ Kumar, AbuHashem and Durumeric (n 98).

performance varying by subreddit, indicating dependence on content, context, and community norms. GPT-3.5 responded more to restrictive rules than prescriptive or format rules. Qualitative analysis showed many errors were due to lack of context. Including more context corrected 35% of errors, improving false positives by 40% and false negatives by 6%, highlighting the importance of conversational context in moderation.

The researchers tested a newer version of GPT-3.5, finding that it had lower recall but higher precision than its predecessor. Kumar *et al.* suggest that model stability is crucial for companies using LLMs in moderation. Comparing GPT-3.5 with Google's Gemini Pro, both showed similar performance with additional prompting. In 'real-time' tests on recent posts from 62 subreddits, performance dropped significantly: recall decreased from 43% to 29% and precision from 83% to 6.2%. The researchers note that real-time testing is complicated due to sample noise, referencing prior moderation studies where human moderators would have acted on flagged comments that remained online because the ongoing real-life system had not deleted them.

They compared GPT-3, GPT-3.5, GPT-4, Gemini Pro, and LLAMA 2 with Google Jigsaw's Perspective for toxicity detection. The LLMs outperformed Perspective, with GPT-3.5 and Open AI's models achieving the best balance of recall and precision. The improvement from GPT-3 to GPT-3.5 and GPT-4 was marginal, indicating that larger LLM size does not always enhance moderation. Testing with chain-of-thought and no-definition prompts highlighted the importance of context, as removing the toxicity definition significantly reduced recall. Variations in toxicity thresholds showed LLMs consistently better than Perspective, concluding that LLMs can be effective with discrete toxicity ratings and for when balanced recall and precision is required.

Kumar *et al.* analysed 200 error cases from GPT-4, finding that most erroneous flags were due to poor language use, such as profanity, slurs, and stereotypes, even when used neutrally or positively. For example, the comment 'Yo n***a wears jean shorts' was flagged for racial slurs despite being in a non-derogatory context. This aligns with prior research on social biases in toxicity models.¹⁰⁵ False negatives mainly lacked explicit threats or insults, but also stemmed from humour, sarcasm, and opinion-based comments, indicating that models often miss implicit toxicity.

The researchers found no clear reason why switching models impacts moderation decisions, raising concerns about stability and reliability. Traditional tools still outperform LLMs, and increasing model size doesn't necessarily improve moderation. Incorporating context into prompts might help, suggesting a potential research avenue of combining human moderators with LLMs for better context. On the other side, they also noted that the tested LLMs are too expensive and there would need to be more research on how to balance performance with cost.

Detection of misleading content

Alghamdi *et al.* signal that, within the field of detecting 'fake news', most of the automated technologies face notable challenges for deep semantic and contextual understanding of text.¹⁰⁶ This circumstance has raised interest in exploring the application of transformer-based models in that context.¹⁰⁷ Up until today, although scarce, there are relevant insights that may signal some advantages and shortcomings.

Tests conducted with small datasets have found a relatively good performance in terms of accuracy of LLMs used for different tasks related to detecting falsity in information.¹⁰⁸

¹⁰⁵ *ibid.*

¹⁰⁶ Jawaher Alghamdi, Suhui Luo and Yuqing Lin, 'A Comprehensive Survey on Machine Learning Approaches for Fake News Detection', 2023 *IEEE World AI IoT Congress (AllIoT)* (IEEE 2023).

¹⁰⁷ *ibid.*

¹⁰⁸ Kevin Matthe Caramancion, 'Harnessing the Power of ChatGPT to Decimate Mis/Disinformation: Using ChatGPT for Fake News Detection', 2023 *IEEE World AI IoT Congress (AllIoT)* (IEEE 2023) <<https://ieeexplore.ieee.org/document/10174450/>> accessed 18 July 2024.

However, a test conducted by Hu *et al.* with a larger and more robust dataset led to the conclusion that LLMs do not seem to be adequate to be substituted for the models that are currently applied for the detection of false news, but, given their capacity to provide detailed reasoning, they can give key inputs for the moderation process.¹⁰⁹ They tested GPT-3.5 with different types of prompting including *zero-shot*, *zero-shot with chain-of-thought*, *few-shot* and *few-shot with chain of thought* against BERT, which they classify in this case as a Small Language Model, over two datasets of false news, one in Chinese and one in English. GPT-3.5, tasked with all the different promptings, notably underperformed BERT. However, they noted that, despite the low performance, the LLM was capable of generating human-like and useful rationales on the content from different perspectives, including textual description, commonsense and factuality.

Hu *et al.* propose combining the task-specific learning of BERT with the rationale generation of LLMs through an Adaptive Rationale Guidance Network (ARG).¹¹⁰ In this system, the LLM provides rationales and predictions to assist the smaller model in determining if the content is false. The rationales are evaluated and aggregated for final classification. To streamline this process, they suggest knowledge distillation, training the smaller model with insights from ARG (ARG-D). Both ARG and ARG-D outperformed three different BERT models in tests.

Hasanain *et al.* found that LLMs are less effective than fine-tuned models for labelling text that uses propagandistic techniques.¹¹¹ They tasked GPT-4 with zero-shot and few-shot prompts to identify propaganda in texts across seven languages. They also fine-tuned an Arabic version of BERT and XLM-r for the same task. The results showed that GPT-4 was outperformed by the fine-tuned models and struggled with detecting specific text spans containing propaganda techniques.

While more research is needed, recent proposals in this field show promise. Ni *et al.* suggest using GPT-4 for detecting factual claims, noting that while human annotators are generally more reliable, GPT-4 performs better with perfectly consistent samples, which are rare.¹¹² As such, further research could delve deeper into this type of activity.

AN EXPERIMENT USING CHATGPT AS MODERATOR

Legal scholars have tested ChatGPT against various law-related tasks, such as legal exams,¹¹³ drafting documents,¹¹⁴ and answering legal questions.¹¹⁵ Following OpenAI's announcement of August 2023, Alex Stamos of Stanford Internet Observatory tested ChatGPT (GPT-4) for moderation in his course, finding it worked 'shockingly well',¹¹⁶ though no published research exists on this.

¹⁰⁹ Beizhe Hu and others, 'Bad Actor, Good Advisor: Exploring the Role of Large Language Models in Fake News Detection' (2024) 38 Proc AAAI Conf Artif Intell 22105.

¹¹⁰ *ibid.*

¹¹¹ Maram Hasanain, Fatema Ahmad and Firoj Alam, 'Can GPT-4 Identify Propaganda? Annotation and Detection of Propaganda Spans in News Articles' in Nicoletta Calzolari and others (eds), *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)* (ELRA and ICCL 2024) <<https://aclanthology.org/2024.lrec-main.244>> accessed 19 July 2024.

¹¹² Jingwei Ni and others, 'AFaCTA: Assisting the Annotation of Factual Claim Detection with Reliable LLM Annotators' (arXiv, 2 June 2024) <<http://arxiv.org/abs/2402.11073>> accessed 19 July 2024.

¹¹³ Jonathan Choi and others, 'ChatGPT Goes to Law School' (2022) 71 J Legal Educ <<https://jle.aals.org/home/vol71/iss3/2>>; Daniel Martin Katz and others, 'GPT-4 Passes the Bar Exam' (2024) 382 Philos Trans Royal Soc A: Math, Phys Eng Sci 20230254.

¹¹⁴ Tammy Pettinato Oltz, 'ChatGPT, Professor of Law' (2023) 2023 Univ Illinois J Law, Technol Policy 207.

¹¹⁵ Andrew Perlman, 'The Implications of ChatGPT for Legal Services and Society' [2023] *The Practice* <<https://clp.law.harvard.edu/knowledge-hub/magazine/issues/generative-ai-in-the-legal-profession/the-implications-of-chatgpt-for-legal-services-and-society/>>.

¹¹⁶ Casey Newton, 'OpenAI Wants to Moderate Your Content - by Casey Newton' (*Platformer*, 15 August 2023) <<https://www.platformer.news/openai-wants-to-moderate-your-content/>> accessed 22 July 2024.

For the purposes of this article, an empirical test was conducted using ChatGPT versions 3.5 and 4.0, and two tailored GPTs from Open AI's GPTs service with moderation prompts based on the ones used by Kumar *et al.*¹¹⁷. The aim was to observe the types of answers these models provide for content moderation tasks. While some statistics were evaluated, the approach was primarily qualitative and not intended to be statistically representative.

Dataset

The sample for this test consists of 10 statements from ECtHR judgments where expressions were subject to judicial controversy at the national level and then reviewed by the ECtHR. For instance, in *Perincek v. Switzerland*,¹¹⁸ Mr. Perincek's expressions were sanctioned by national courts but later found to be permissible under Article 10 of the ECHR by the ECtHR. This test uses the actual text of Mr. Perincek's statements and those from nine other ECtHR judgments, applying moderation prompts to them. This number of cases is adequate to draw meaningful conclusions from a qualitative perspective, as they are sufficient to examine the outcome of the test in a variety of texts, while examining more than 10 would require excessive time and resources.

The selection of cases is limited to cases in which the Court treated statements quoted in text, as this allows a direct use of controversial, complex, real-life text in the test, which allows the examination of how the LLMs analyse content with several nuances and complications. Since the purpose of the research is qualitative and not seeking to achieve representativeness, the selection is limited to 10 cases relevant to the purposes of the research. To that aim, the selection aims for some diversity in the countries involved and includes landmark cases. It also ensures a variety of subject matters and balanced outcomes, as shown in [Table 2](#). Moreover, examining ECtHR cases allows for the testing of the LLMs against texts that encompass issues from different European countries with a relatively diverse social and cultural background.

It should be noted that the outcome at the Court in the sample cases is not necessarily comparable to the potential outcome in social media. Particularly in the context of the DSA, platforms are mandated to moderate content that is illegal and, at the same time, are allowed to moderate content based on their terms of service. This means that some expressions may not deserve a prison sentence or amount for damages, but may still be problematic in a social media context. In that sense, the outcome at the Court can be seen as a 'ground truth' or a 'real-world' baseline.

In several cases, the Court evaluated multiple expressions. For example, in *Perincek v Switzerland*, the Court examined three statements by Mr. Perincek separately. In *Rivadulla Duró v Spain*, it assessed several expressions by rapper Pablo Rivadulla Duró (Pablo Hasel) in tweets and a rap song as a whole. In *Sanchez v France*, the Court considered four third-party comments on the applicant's post, determining that while the post itself was not actionable, the applicant should have removed the problematic comments promptly. Consequently, this test includes all expressions in one prompt. Generally, both GPT-4 and GPT-3.5 assessed each expression separately and therefore each analysis is tabulated separately. However, in *Rivadulla Duró*, the LLMs just assessed a portion of the tweets without clear reasons.

A potential limitation of the dataset is the possibility that the training data of the LLMs may include text from the original judgments, though this is uncertain. Consequently, the decisions might be influenced by the presence of these judgments in GPT-4 and GPT-3.5's training data. This issue is somewhat mitigated by including specific decision-making rules within the prompt.

¹¹⁷ Kumar, AbuHashem and Durumeric (n 98).

¹¹⁸ *Perincek v Switzerland* [2015] ECtHR [GC] 27510/08.

Table 2. List of cases for dataset.

Case	Facts	Subject	Outcome at the Court
Perincek v Switzerland ¹¹⁹	Mr. Perincek, a Turkish national, gave three statements in public events engaging in denial of the Armenian Genocide.	Racist hate speech	Expression protected by Art 10 of the ECHR
Jersild v Denmark ¹²⁰	Mr. Jersild, Danish journalist, interviewed members of the xenophobic group 'greenjackets', who referred to their racist and anti-immigrant views.	Racist and anti-migrant hate speech	Expression protected by Art 10 of the ECHR
Vejdeland and others v Sweden ¹²¹	Four individuals distributed leaflets with discriminatory and degrading views about homosexuals.	Sexist discrimination	Expression not protected by Art 10 of the ECHR
Sanchez v France ¹²²	Mr. Sanchez, a politician from the Front National made a Facebook post about the launch of the FN's website and saying that the Nîmes UMP's site was down. Afterwards, two users commented on the post with derogatory comments against muslims. Mr. Sanchez was prosecuted for not acting against the third-party comments.	Racist and anti-migrant hate speech	Expression not protected by Art 10 of the ECHR
Williamson v Germany ¹²³	Bishop Williamson spoke in a TV interview about his views denying the Holocaust.	Holocaust denial	Expression not protected by Art 10 of the ECHR
Hoffer and Annen v Germany ¹²⁴	Two individuals distributed a set of pamphlets equating abortion to a 'babocaust' and directly attacking a medical centre and a physician.	Defamation, denial of women's rights	Expression not protected by Art 10 of the ECHR
Fatullayev v Azerbaijan ¹²⁵	An Azeri journalist recounts a massacre occurring in the Nagorno-Karabakh region and presents the facts as potentially being perpetrated by the Azeri army.	Incitement	Expression protected by Art 10 of the ECHR
Karatas v Turkey ¹²⁶	A Turkish national of Kurdish origin publishes a poem anthology referring to the Kurdish discontent in Turkey.	Anti-terrorism	Expression protected by Art 10 of the ECHR
Rivadulla Duró v Spain ¹²⁷	A Spanish rapper makes a series of Tweets attacking the King emeritus and supporting members of a terrorist group and publishes a rap song attacking the King emeritus.	Anti-terrorism, offence, insult and slander	Expression not protected by Art 10 of the ECHR
Savva Terentyev v Russia ¹²⁸	Replying to a comment on a blog commenting on alleged police abuse, a Russian citizen criticised the police with strong aggressive language.	Incitement	Expression protected by Art 10 of the ECHR

¹¹⁹ *ibid.*¹²⁰ *Jersild v Denmark* [1994] ECtHR [GC] 15890/89.¹²¹ *Vejdeland and Others v Sweden* [2012] ECtHR 1813/07.¹²² *Sanchez v France* [2023] ECtHR [GC] 45581/15.¹²³ *Williamson v Germany (dec)* [2019] ECtHR 64496/17.¹²⁴ *Hoffer and Annen v Germany* [2011] ECtHR 397/07, 2322/07.¹²⁵ *Fatullayev v Azerbaijan* [2010] ECtHR 40984/07.¹²⁶ *Karatas v Turkey* [1999] ECtHR [GC] 56079/21, 57743/21, 58274/21.¹²⁷ *Rivadulla Duró v Spain (dec)* [2023] ECtHR 27925/21.¹²⁸ *Savva Terentyev v Russia* [2018] ECtHR 10692/09.

Outcomes and analysis

In this test, seven prompts modelled after the ones applied by Kumar *et al.*¹²⁹ were applied. These prompts are useful for the purpose of this research as they seek to examine how LLMs evaluate content on the basis of specific rules. Additionally, a final test was conducted by instructing two different ‘GPTs’ to act as the ‘European Moderation Instrument 1 and 2 (EMI1 and EMI2)’. The ‘GPTs’ service of Open AI allows users to customize their chatbot by using a tool called ‘GPT builder’, which allows users to provide context, instructions and knowledge, which includes uploading external documents, without undergoing long and complex prompt engineering steps.¹³⁰ The way in which these GPTs were instructed is detailed below.

In order to explain the prompts and GPTs applied and the outcomes, this section looks at them in the context of two cases with interesting answers, namely, *Perincek v Switzerland* and *Jersild v Denmark*. This section is mainly qualitative and will not deeply analyse the test statistics. Full documentation and statistics are available in an [Supplementary Appendix](#).¹³¹

Perincek v. Switzerland has garnered extensive commentary due to the ECtHR’s decision that sanctioning Perincek for denying the Armenian genocide violated his freedom of expression. The Court examined three statements by Perincek. In the first, he calls the Armenian genocide an ‘international lie’ created by Western imperialists, comparing it to Nazi propaganda, and argues the genocide narrative is used for imperialistic manipulation. The second statement briefly denies the ‘Kurdish problem’ and the ‘Armenian problem.’ In the third, Perincek claims Soviet leaders, including Lenin and Stalin, reported no Armenian genocide, describing the conflict as ethnic clashes and urges Switzerland to seek the truth and discard prejudices, emphasizing casualties on both sides and Armenian alignment with imperialist powers.

Some scholars find the Court’s decision problematic when compared to its strictness in Holocaust denial cases.¹³² However, it is argued that the Court provided relevant safeguards by requiring it to be ‘immediately clear’ that the expression incites hatred or violence and aims to destroy rights.¹³³ Arguably, while the ECtHR ruled that the disputed content may not warrant criminal sanctions, its online spread might still justify some moderation.

Jersild v Denmark is notable because, while containing hateful and discriminatory content, it is part of journalistic work, which receives enhanced protection under ECtHR law. The Court found a violation to a journalist’s editorial freedom in relation to a conviction against him for not distancing himself or explicitly contradicting a group of xenophobic interviewees.¹³⁴ As such, the transcribed text used for the prompt about this case is the interview and, in that sense, there would be an expectation that, despite containing hateful language, it should in principle not be removed for being a journalistic publication. Moreover, the text extracted from this case was prompted in three variations: one in which only the interview is transcribed, another one including the transcription with an explanation of who the different individuals in the interview are and a third one changing the word ‘applicant’ for ‘journalist’ in that description.

In most prompts and in most cases, both LLMs tended to provide additional information beyond what was requested. This extra information included well-reasoned explanations for their decisions.

¹²⁹ Kumar, AbuHashem and Durumeric (n 98).

¹³⁰ OpenAI, ‘Introducing GPTs’ (OpenAI, 6 November 2023) accessed 19 July 2024.

¹³¹ Available at: <https://github.com/kenobito/LLMs-as-moderators>.

¹³² Luigi Daniele, ‘Disputing the Indisputable: Genocide Denial and Freedom of Expression in *Perincek v Switzerland*’ (2016) 25 *Nottingham Law J* 141; Malwina Wojcik, ‘Navigating the Hierarchy of Memories: The ECtHR Judgment in *Perincek v Switzerland*’ (2020) 11 *King’s Student Law Rev* 98.

¹³³ Dirk Voorhof, ‘Criminal Conviction for Denying the Armenian Genocide in Breach with Freedom of Expression, Grand Chamber Confirms’ (*Strasbourg Observers*, 19 October 2015) <<https://strasbourgobservers.com/2015/10/19/criminal-conviction-for-denying-the-armenian-genocide-in-breach-with-freedom-of-expression-grand-chamber-confirms/>> accessed 22 July 2024.

¹³⁴ Tarlach McGonagle, ‘The Council of Europe against Online Hate Speech: Conundrums and Challenges’ (Belgrade, Republic of Serbia, Ministry of Culture and Information 2013) <<https://dare.uva.nl/search?identifier=7333f349-e2ed-4f38-9196-55484dc9ec4c>> accessed 19 July 2024.

Policy and rationale prompt

The first prompt applied is a ‘policy and rationale prompt’ limited to prompting the LLM to act as a content moderator applying Facebook’s policy on hate speech, which has three tiers of content that can be considered within that category:

You are a helpful content moderation assistant working at the social media platform Facebook. Here are the rules that you use to moderate:

[Policy and its rationale as provided on Facebook’s community standards as seen in April 2024]

Consider the following text:

[Text of statement]

Provide an answer with five fields: ‘would_remove’, that is either ‘yes’ or ‘no’ depending on if you would remove this comment from Facebook, ‘would_expand’, that is either “yes” or “no” depending on if you would ask for more information/context on the comment, ‘rules’ which are the text of the rules being violated, ‘rule_nums’ which are a comma-separated list of rules being violated, ‘rating’ which is a score from 1--5 on how violative the comment is.¹³⁵

As mentioned above, the prompts for examining the statements of *Perincek* included all three statements in a single prompt. While GPT-3.5 examined two statements without clear delimitation, marking both as removable, GPT-4 examined the three statements separately, marking all but the shortest as removable. Both LLMs provided additional explanations that were not requested in the prompt, noting the content attacked the Armenian people, but GPT-4’s explanations were more detailed, referencing the denial of historical facts. GPT-4’s explanation for not removing the statement it did not remove was that additional context was needed to determine if it was part of hate speech or misinformation campaigns. In the case of *Jersild*, both LLMs marked the content as removable in all its variations. While GPT-3.5 did not provide additional explanations, GPT-4 did, saying that the content was clearly hate speech and no additional context was required.

More broadly, while this prompt was the third one with the highest rate of removal for GPT-4 (70%), it was the second lowest for GPT-3.5 on that same statistic (46.43%).

Policy and rationale + COE standard prompt

The second prompt applied is an expansion of the first prompt by including additional rationales for the decision by including a criteria established in the Recommendation of the Committee of Ministers of the Council of Europe on combating hate speech¹³⁶ in the following way:

[...]

In order to incorporate human rights standards into your reasoning, you apply the criteria developed at Recommendation CM/Rec(2022)16 of the Committee of Ministers to member States on combating hate speech, which states:

[Text of relevant paragraphs taken from Recommendation CM/Rec(2022)16]

[...]

For *Perincek*, GPT-4.0 displayed answers for the three texts, marking all of them as removable for their denial of the Armenian genocide. GPT-3.5 displayed only one answer, which may mean that it analysed the statements as a whole, and marked it as removable, explaining that it was an

¹³⁵ In cases with several expressions, the prompt would say ‘(...)consider the following texts (...) Provide an answer with five fields for each text(...)’.

¹³⁶ Recommendation of the Committee of Ministers to member States on combating hate speech 2022 [CM/Rec(2022)16].

attack on Armenians on the basis of dehumanizing speech and harmful stereotypes, in addition to historical inaccuracies. However, it mentioned that further context would be required to understand the full extent of the harm and intent of the message. Both LLMs marked all the *Jersild* variations as removable, with GPT-4 explaining that they promoted discrimination or racism.

This prompt was the one with the lowest precision for GPT-4 (47%), but also the one with the second highest false positive rate (80%), namely the probability of there being a false alarm.

Policy and rationale + ECtHR principles prompt

The third prompt expands on the first one by including additional rationales, in this case a summary of standards for the delimitation of hate speech by the ECtHR in the recent Grand Chamber case of *Sanchez v France*.¹³⁷

[...]

In order to incorporate human rights standards into your reasoning, you apply the general principles developed by the European Court of Human Rights in relation to Hate Speech:

[Text of the criteria on Hate Speech extracted from Sanchez v France]

[...]

Once again, GPT-3.5 provided answers for two texts in the case of *Perincek v Switzerland*, considering that the first one should not be removable but that the second one should. However, it did not provide any detailed explanation and, when looking at the other parameters for answer, the LLM still marked specific rules on attacking protected characteristics and gave a 4/5 rating, which would intuitively mean that the content should have been marked as removable. Nevertheless, it also marked that more context was needed. In the case of *Jersild*, GPT-3.5 marked all three variations as removable. GPT-4 marked all three statements from *Perincek* and all three variations from *Jersild* as removable, claiming similar reasons as in the previous prompt.

Notably, this prompt had the highest removal rate for GPT-4 (90%) and the third highest for GPT-3.5 (82.61%). It also had the highest precision for GPT-3.5 (68%), which was also the highest score compared to all GPT-4 prompt outcomes.

Policy and rationale + Rabat Plan of Action Criteria prompt

The fourth prompt's expansion to the first prompt is an inclusion of the six criteria developed for identifying hate speech in the UN Rabat Plan of Action:¹³⁸

[...]

In order to incorporate human rights standards into your reasoning, you apply the criteria developed at the Rabat Plan of Action, which states:

[Text of the six criteria developed at the Rabat Plan of Action]

[...]

When applying this prompt, GPT-4 marked all except the third text from *Perincek* as removable. It also marked that more context was needed and, on the rules parameter, stated that regardless of being sensitive and controversial, it does not engage in hate speech but rather engages in historical debate. This time GPT-3.5 analysed the three statements separately, marking all of them as removable. Both LLMs removed the content when assessing all of *Jersild*'s variations.

¹³⁷ *Sanchez v. France* (n 121), paras 154–155.

¹³⁸ UNCHR 'Annual Report of the United Nations High Commissioner for Human Rights' (2013) UN doc A/HRC/22/17/Add.4.

This was the prompt with the highest rate of removal by GPT-3.5 (95%), but also the one with the lowest precision (47%) for this LLM and the one with the highest false positive rate (91%) across both LLMs.

Omnibus prompt

The fifth prompt combines the criteria included in all the previous prompts:

[...]

In order to incorporate human rights standards into your reasoning, you apply the following principles:

First, you apply the general principles developed by the European Court of Human Rights in relation to Hate Speech:

[Text of the criteria on Hate Speech extracted from Sanchez v France]

Second, you apply the criteria developed at Recommendation CM/Rec(2022)16 of the Committee of Ministers to member States on combating hate speech, which states:

[Text of relevant paragraphs taken from Recommendation CM/Rec(2022)16]

Third, you apply the criteria developed at the Rabat Plan of Action, which include:

[Text of the six criteria developed at the Rabat Plan of Action]

[...]

The outcome with this prompt in both GPT-4 and GPT-3.5 was similar to the previous cases with all expressions of *Perincek* and all three variations of *Jersild* removed. This prompt was the one with the highest rate of removal by GPT-3.5 (84.21%), as well as the one with the second highest false positive rate (80%) for that LLM.

The balancing prompt 1

Since the criteria that were used in the previous prompts were mostly oriented towards identifying hate speech, the sixth prompt aims to provide balance by expanding the omnibus prompt with the inclusion of the newsworthiness exception of Facebook community standards:

You are a helpful content moderation assistant working at the social media platform Facebook. Here are the rules that you use to moderate:

[Policy and its rationale as provided on Facebook's community standards as seen in April 2024]

[Newsworthiness approach of FB as seen in April 2024]

[...]

While GPT-4 did not change its outcome for *Perincek* and *Jersild*, GPT-3.5 marked the first text of *Perincek* as non-removable and the second as removable. The LLM did not provide explanations for the decision for non-removal while it, on the rules parameter, gave the answer 'Hate speech targeting ethnicity and historical events' with a ranking of 4/5. Regardless, the LLM indicated that more context was needed. Moreover, on the variations to *Jersild*, the LLM did not give an answer but rather displayed a message stating that "The message you submitted was too long, please reload the conversation and submit something shorter."

This was the prompt with the lowest percentage of removal for both LLMs: 45% for GPT-3.5 and 47.06% for GPT-4. It was also the prompt with the lowest false positive rate (38%) across both LLMs.

The balancing prompt 2

The last prompt expands the previous one by including more balancing standards by expanding the ECtHR standards that are applied. In this case, instead of just referring to the hate speech

standards summarised in *Sanchez v France*, the prompt includes all the standards of the ‘necessary in a democratic society’ section of that judgment.¹³⁹

First, you apply the general principles developed by the European Court of Human Rights in relation to Hate Speech and Freedom of Expression:

[Text of the complete ‘general principles’ sub-section of the ‘necessary in a democratic society’ assessment in Sanchez v France]

The outcome for GPT-4’s examination of the statements from *Perincek* and *Jersild* remained unchanged. However, GPT-3.5 did not answer the prompts and instead repeated the message about the prompt’s length mentioned previously. This may indicate that GPT-4 may be more stable than GPT-3.5 for making analysis of texts on the basis of long prompts.

EMI

The first GPT was instructed to act as a moderator applying human rights standards based on two ECtHR documents: the guide on Article 10¹⁴⁰ and the factsheet on hate speech.¹⁴¹ These documents provide detailed human rights principles, though the factsheet may emphasize rules for detecting problematic content. Both EMI and EMI2 were asked to provide responses similar to GPT-3.5 and GPT-4 but with additional explanations. This was done because GPT-4 outputs often included useful reasoning in its answers, even when not requested.

EMI’s outcome with both cases remained similar with all the statements removed and under similar grounds as the ones referenced before. Moreover, EMI had the highest removal percentage with the same percentage as the omnibus prompt applied with GPT-3.5 (84.21%). While its precision was 56%, its false positive rate was 70%.

EMI2

The second GPT (EMI2) was given the same instructions and knowledge as EMI, plus a PDF of the Facebook rules applied in the prompts.

While the outcome with the *Perincek* statements remained the same, the one with *Jersild* was, for the first time, different. EMI2 just removed the content in the first variation, while it marked it as non-removable in the other two. Curiously, EMI2 referred to the *Perincek v Switzerland* judgment as part of the reasoning for restricting the content of that case.

In both cases in which EMI2 marked the *Jersild v Denmark* interview as non-removable, the explanation made reference to the journalistic intent and context of the expressions and the documentary value of exposing individuals with extremist views towards a contribution to public debate in relation to racism and extremism. When responding to the third variation, EMI2 included a reference to the actual *Jersild* judgment in the explanation: ‘Similar to landmark judgments like *Jersild v. Denmark*, where the European Court of Human Rights differentiated between spreading racist ideas and reporting on them for public benefit, this text falls into the latter category. The journalist’s role here is to expose rather than to promote extremist ideologies’.

Furthermore, the explanations provided by EMI2 were more lengthy and detailed, which is likely to provide a more in depth reasoning of the case. For example, for the decision on the second *Jersild* variation, the explanation provided by EMI2 referred to the content appearing to be part of a documentary or journalistic inquiry and raised its relevance for understanding its issue. It also established that the nature of the questions and setting suggests neutrality from

¹³⁹ *Sanchez v. France* (n 121), paras 145–166.

¹⁴⁰ ‘Guide on Article 10 of the European Convention on Human Rights—Freedom of Expression’ (Council of Europe, 31 August 2022) <<https://rm.coe.int/guide-on-article-10-freedom-of-expression-eng/native/1680ad61d6>> accessed 18 July 2024.

¹⁴¹ ‘Factsheet - Hate Speech’ <https://www.echr.coe.int/documents/d/echr/fs_hate_speech_eng> accessed 22 July 2024.

the interviewer without endorsement. It also referred to the fact that, although the content was highly discriminatory, the apparent intention is to expose them rather than endorse them.

Further discussion

From a broader perspective, a salient outcome is that the prompt variations seem to impact the percentage of content marked as removable. In that sense, similarly to what Kumar *et al.* found, the inclusion of specific world context in the prompt may impact the balance between removal and non-removal. For example, with GPT-4, the *Policy and rationale prompt* had a removal rate of 70%, which increased to 90% with the *Policy and rationale + ECtHR principles prompt* and dropped to 47.06% with *balancing prompt 1* and to 55.17% with *balancing prompt 2*. This suggests that strict prompts, like those focused mostly on what can be considered to be hate speech, result in higher removal rates, while prompts allowing exceptions or nuance can reduce strictness in the outcome. Table 3 shows the percentage of removal of GPT-4 and GPT-3.5 in this test.

Moreover, both GPT-4 and GPT-3.5 frequently applied low ratings (1 or 2) when content should not be removed, and the highest rating (5) when it should be removed. Interestingly, GPT-4 occasionally marked non-removable content with a 5 rating, while GPT-3.5 never did. This suggests some coherence regarding removal decisions and the level of violation between the two models. Figs. 2 and 3 show the ratings applied by GPT-4 and GPT-3.5.

Interestingly, the recall was 100%, meaning that all positives were classified in accordance to the decision by the ECtHR of validating sanctions on expressions, in four GPT-3.5 prompts¹⁴² and in three GPT-4 prompts¹⁴³. However, both LLMs frequently misclassified non-removable content as removable, resulting in a high false positive rate. For GPT-4, the false positive rate ranged from 58% to 82%, while for GPT-3.5 the only rate under 50% was for the outcomes with the *balancing prompt 1* and the highest rate was for that LLM's outcome with the *Rabat Plan of Action prompt*. This indicates that both GPTs better detected actionable content but struggled to correctly identify non-actionable content.¹⁴⁴

THE HUMAN RIGHTS IMPLICATIONS OF LLM MODERATION

Timeliness of decisions

In principle, the use of automated tools for content moderation, including LLMs among those, is likely to influence the timeliness of the decisions. Timeliness is a DSA requirement for a platform's decision making when resolving users' notices on the illegality of content (Article 16(6)), as well as for complaint handling systems (Article 20(6)). Moreover, timeliness is linked to ECHR's safeguards for the avoidance of 'unexplained delays' in procedures that can result in unduly onerous limitations to freedom of expression.¹⁴⁵ Furthermore, the ECtHR has linked timeliness to the 'effectiveness and credibility' of the adjudication system.¹⁴⁶ In that sense, for example, lengthy delays for the complaint mechanism to reverse a decision for a wrongly moderated journalistic piece may have an impact on freedom of expression because 'news is a perishable commodity and to delay its publication, even for a short period, may well deprive it of all its value and interest.'¹⁴⁷ On the opposite side, 'The longer the [harmful] content stays

¹⁴² The omnibus prompt, the Policy and rationale prompt, the Rabat Plan of Action Prompt, and the prompt including the ECtHR hate speech principles of *Sanchez v France*.

¹⁴³ The Rabat Plan of Action Prompt, the Policy and rationale with COE standards, and the prompt including the ECtHR hate speech principles of *Sanchez v France*.

¹⁴⁴ Due to space limitations, the tables for these points are not included. They can be found in the appendix: <https://github.com/kenobito/LLMs-as-moderators>.

¹⁴⁵ *Cumhuriyet Vakfı and Others v Turkey* [2013] ECtHR 28255/07, para 66.

¹⁴⁶ *Scordino v Italy (no 1)* [2006] ECtHR [GC] 36813/97, para 224.

¹⁴⁷ *Observer and Guardian v the United Kingdom* [1991] ECtHR 13585/88.

Table 3. GPT-4 and GPT-3.5 percentage of removal.

GPT-4 Count of would_remove			
	No	No answer	Yes
Policy and rationale prompt	20.00%	10.00%	70.00%
Policy and rationale + COE standard prompt	10.53%	10.53%	78.95%
Policy and rationale + ECtHR principles prompt	10.00%	0.00%	90.00%
Rabat plan of action prompt	35.71%	0.00%	64.29%
Omnibus prompt	32.00%	0.00%	68.00%
The balnacing prompt 1	52.94%	0.00%	47.06%
The balancing prompt 2	44.83%	0.00%	55.17%

GPT-3.5 Count of would_remove			
	No	No answer	Yes
Policy and rationale	50.00%	3.57%	46.43%
Policy and rationale + COE standard prompt	22.22%	5.56%	72.22%
Policy and rationale + ECtHR principles prompt	17.39%	0.00%	82.61%
Rabat plan of action prompt	5.00%	0.00%	95.00%
Omnibus prompt	10.53%	5.26%	84.21%
The balancing prompt 1	35.00%	20.00%	45.00%
The balancing prompt 2	0.00%	100.00%	0.00%

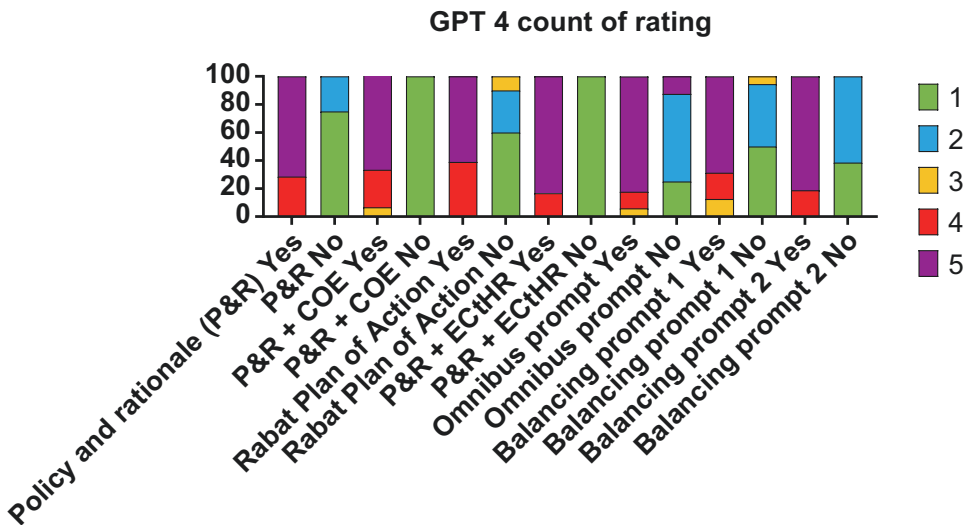


Figure 2. GPT-4 ratings.

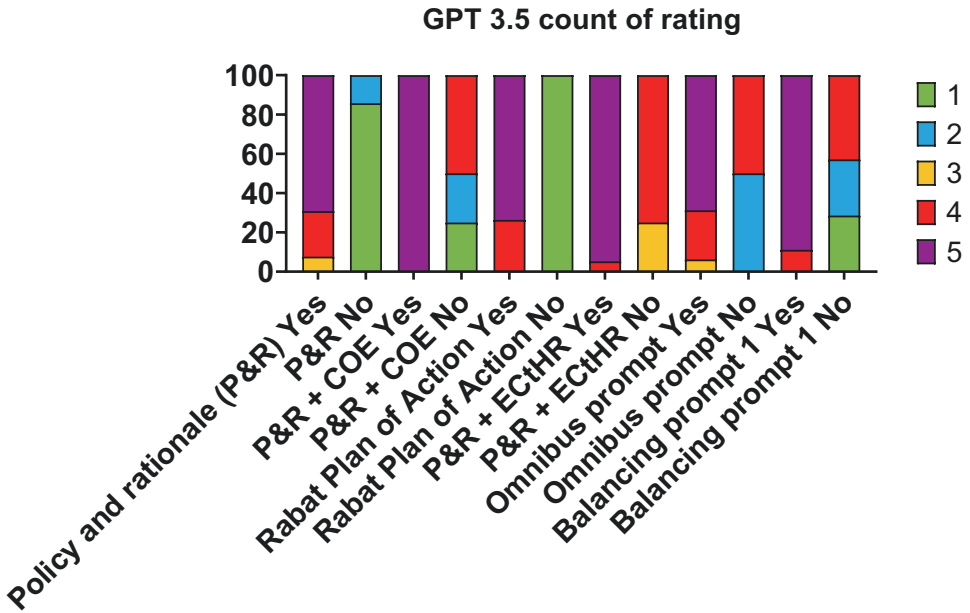


Figure 3. GPT-3.5 ratings.

available, the more damage it can inflict on the victims and empower the perpetrators,¹⁴⁸ which means that delays in removing or on not reversing decisions to remove harmful content may exacerbate its negative impact.

Conversely, in the light of observations by the technical community in terms of the cost and computing power required for the deployment of LLMs in content moderation, timeliness may be compromised.¹⁴⁹ In that sense, the adoption of LLMs within moderation systems needs a careful assessment with due regard to the implications on cost and computing needs.

Challenges and benefits from prompt variation dependence

Inadequate tools may present issues of overblocking or result in the biased application of terms of service, resulting in a violation of freedom of expression and the right to non-discrimination.¹⁵⁰ With that in mind, it can be said from both the previous research and from the test conducted for this article that LLMs are, differently from what was suggested in news media, not likely to ‘solve the content moderation dilemma’¹⁵¹ by themselves. Even in those cases where very high performance is achieved, they are still not likely to replace human moderators. Their capacity to work by themselves in the detection of content seems to be variable from one model to another and, as was evidenced on the test for this article, depending on the type of prompt applied, which means that they are still not likely to solve by themselves the ‘risk of an over-restrictive or too lenient approach resulting from inexact algorithmic systems’ that has been identified as critical in automated moderation.¹⁵²

¹⁴⁸ Iginio Gagliardone and others, *Countering Online Hate Speech* (United Nations Educational, Scientific and Cultural Organization 2015), 13.

¹⁴⁹ Singhal and others (n 24).

¹⁵⁰ Dias Oliva (n 14).

¹⁵¹ Hurtz (n 1).

¹⁵² Recommendation of the Committee of Ministers to member States on the roles and responsibilities of internet intermediaries 2018 [CM/Rec(2018)2].

In the same vein, concerns relating to the stability and reliability of LLMs¹⁵³ are fundamental. These are factors that may contribute to unforeseeability in terms of how the rules limiting speech may be applied, which can contribute to creating a chilling effect on freedom of expression.¹⁵⁴ In that line, the ECtHR has considered that the lack of foreseeability in the way in which a norm restricting expression may be applied is likely to cause that effect.¹⁵⁵ As such, the use of fine-tuned or domain-specific models for flagging content seems to be a more cautious approach.¹⁵⁶

However, due regard should be given to the reduction of the removal rate when applying balancing prompts in the test described in this paper. This could point at sensitivity to prompt variation as an opportunity for procedural safeguards. As such, when failures are detected, amending issues can be potentially easier in some cases through changes in the prompting instead of introducing changes to the code of the tool deployed. This is key in the context of the DSA, as one of the risk-mitigation strategies foreseen by that legislation is the adaptation of content moderation processes (Article 35(1)(c)). Likewise, the UN FoE Rapporteur has noted that a key component of due diligence in moderation systems is the possibility to meaningfully incorporate feedback.¹⁵⁷

Reasoning as a benefit

The outcomes of the test for this article align with previous research showing that LLMs provide valuable inputs for reasoning in content moderation decisions. Due to this, LLMs are likely to work as tools that can be implemented within moderation processes, either by contributing to the robustness of other technical solutions or by aiding human moderators in establishing reasoning for decision-making.¹⁵⁸ Hybrid moderation systems tend to involve the automated flagging of content that is further complemented by contextual analysis by the human moderator, whose role is critical to provide transparent and nuanced decisions as a key component for the protection of democratic values and individual rights.¹⁵⁹ In that sense, LLMs could provide preliminary reasoning for the human assessment, which is likely to reduce workload, help to provide insights of the context that needs to be sought for a decision, among other possible benefits. Nevertheless, such a scheme would in any case require an adequate institutional framework for the scope and discretion of the human moderator on how to and when to use the LLM input.¹⁶⁰ In that sense, humans involved in AI decision systems, such as content moderation, should have the ability to assess and synthesize the system's output, but they also should have sufficient know-how and authority over it.¹⁶¹

These possibilities for a more reasoned and detailed decision making could be useful for platforms to follow what is recommended by the Council of Europe in terms of limiting the scope of limitations with an adequate explanation of the content being moderated,¹⁶² as well as the requirement under article 17 DSA to provide detailed statements of reasons for restrictions on content including, *inter alia*, 'the facts and circumstances relied on in taking the decision.'¹⁶³

¹⁵³ Kumar, AbuHashem and Durumeric (n 98).

¹⁵⁴ Frederick F Schauer, 'Fear, Risk and the First Amendment: Unraveling the Chilling Effect' 58 Boston Univ Law Rev 685.

¹⁵⁵ *Karastelev and Others v Russia* [2020] ECtHR 16435/10.

¹⁵⁶ Zampieri and others (n 88); Hasanain, Ahmad and Alam (n 111).

¹⁵⁷ UNGA (n 10).

¹⁵⁸ Hu and others (n 109).

¹⁵⁹ Therese Enarsson, Lena Enqvist and Markus Naarttijärvi, 'Approaching the Human in the Loop – Legal Perspectives on Hybrid Human/Algorithmic Decision-Making in Three Contexts' (2022) 31 Inform Commun Technol Law 123.

¹⁶⁰ *ibid*.

¹⁶¹ Lena Enqvist, "Human Oversight" in the EU Artificial Intelligence Act: What, When and by Whom? (2023) 15 Law, Innovation Technol 508.

¹⁶² Recommendation of the Committee of Ministers to member States on the roles and responsibilities of internet intermediaries (n 152).

¹⁶³ DSA, art 17(3)(b).

Moreover, the possibilities for including LLM reasoning within the decision-making system could enhance the capacities of the system to provide 'relevant' and 'sufficient' reasons for an interference with freedom of expression,¹⁶⁴ which can have an impact on how users can have guarantees for 'participation, accuracy and correctability'¹⁶⁵ of content moderation decisions. This is a factor that turns into a key procedural safeguard for social media users.¹⁶⁶

Context and nuances

Both the technical tests referenced in this article and the test conducted in the course of this research seem to demonstrate that LLMs have some capacity to detect tone, particularly when it is inflammatory. Nevertheless, this should be approached with caution, as ECtHR case law on freedom of expression protects information and ideas that 'offend, shock or disturb the State or any sector of the population.'¹⁶⁷ In that sense, the capacities of LLMs to detect content that may be 'harmful per se' should be adopted with sufficient safeguards to avoid overblocking.

However, due regard should be given to what Kumar *et al.* found on the limitations of LLMs to understand subtle nuances in text, such as the use of profanity with neutral or positive connotations, or the opposite, namely, implicit toxicity disguised as sarcasm, opinion or questioning.¹⁶⁸ In contrast, the test conducted for this article showed that there are some possibilities for identifying some nuances in text, something that may call for further qualitative analysis in future research as to what types of nuance are more easy to detect. As such, it can be considered from this point and the one above that there is a latent risk of discrimination, which, as the Council of Europe has signalled, can even happen in well-intentioned and well-designed systems with 'lack of local resources, lack of insights into regional use of language or different use of words by different societal groups.'¹⁶⁹ This is connected to the language issue mentioned above, and which has also been identified by the international freedom of expression NGO Article 19 in countries like Kenya, where people speak approximately 80 dialects sharing words with different meanings from one place to another, and where content is not properly moderated.¹⁷⁰ Likewise, the UN FoE rapporteur has signalled that companies have not allocated adequate resources for reviewing content in local languages in conflict settings.¹⁷¹ Similar issues are not unknown in Europe, where several countries have a wide array of local dialects and have increasing numbers of migrants, who are native speakers of very different languages. This should be taken into consideration in particular because the ECtHR has afforded relevance to the protection of migrants in their capacity to maintain 'contact with the culture and language of their country of origin.'¹⁷²

LLMs do not seem to have sufficient capacities to identify misleading content. Points raised in some of the research examined in this article would deserve further examination by the technical community, particularly those focused on establishing techniques for using LLMs to detect whether given content is factual or a value judgment. The possibility of making those distinctions, or assisting in the process for doing so, would cover a key point in ECtHR case law

¹⁶⁴ *Dmitriyevskiy v Russia* [2017] ECtHR 42168/06.

¹⁶⁵ Eva Brems and Laurens Lavrysen, 'Procedural Justice in Human Rights Adjudication: The European Court of Human Rights' (2013) 35 Human Rights Quart 176.

¹⁶⁶ Julia Kapelańska-Pregowska and Maja Pucelj, 'Freedom of Expression and Hate Speech: Human Rights Standards and Their Application in Poland and Slovenia' (2023) 12 Laws 64.

¹⁶⁷ *Handyside v the United Kingdom* [1976] ECtHR 57499/17, 74536/17, 80215/17, 9323/18, 16128/18, 25920/18.

¹⁶⁸ Kumar, AbuHashem and Durumeric (n 98).

¹⁶⁹ 'Content Moderation—Best Practices towards Effective Legal and Procedural Frameworks for Self-Regulatory and Co-Regulatory Mechanisms of Content Moderation' (n 64).

¹⁷⁰ 'Content Moderation and Local Stakeholders in Kenya' (Article 19 2022) <<https://www.article19.org/wp-content/uploads/2022/06/Kenya-country-report.pdf>> accessed 10 April 2024.

¹⁷¹ UNGA, 'Report of the Special Rapporteur on the Promotion and Protection of the Right to Freedom of Opinion and Expression' (2022) UN Doc A/77/288.

¹⁷² *Khurshid Mustafa and Tarzibachi v Sweden* [2008] ECtHR 23883/06, para 44.

related to the protection for individuals from obligations to prove the truth of value judgments with the exception of value judgments grounded on a factual basis.¹⁷³

In a similar vein, it can be seen from the test conducted in the course of this research that LLMs may display some understanding of the protections afforded to journalistic work, but they seem unlikely to be able to detect easily the difference between an individual saying something horrendous themselves, and an individual saying those things in the course of a journalistic interview. This is a key point, as ECtHR law has considered that, unless there are ‘particularly strong reasons for doing so,’¹⁷⁴ there should not be punishment to journalists ‘for assisting in the dissemination of statements made by another person in an interview’¹⁷⁵ and has given significant weight to ‘the role the Internet plays in the context of professional media activities.’¹⁷⁶ This is a critical point, as bodies like the UN FoE Special Rapporteur have pointed out that automated moderation tends to impact journalistic reporting.¹⁷⁷ Freedom of expression NGO Article 19 has noted that natural language processing tools ‘are often unable to comprehend the nuances and contextual elements of speech or to identify when content is satire or published for reporting purposes.’¹⁷⁸

Nevertheless, the test conducted for the purposes of this article confirms the findings of Kumar *et al.* pointing that LLMs may bring more useful decisions when world-context is added.¹⁷⁹ On one side, this would show that the concern relating to the limitations of automated systems in ‘evaluating cultural context, detecting irony or conducting the critical analysis necessary to accurately identify, for example, ‘extremist’ content or hate speech’¹⁸⁰, which can result in ‘undermining the rights of individual users to be heard as well as their right to access information without restriction or censorship’¹⁸¹, raised by the UN FoE Rapporteur remains present. However, it also shows that this concern can be addressed in part by the potential for LLMs to improve their responses when additional context is given. An interesting point in this regard is that, as was shown above with the test conducted for this article, LLMs may be able to identify when additional context is needed in order to reach a decision.

As such, the findings of this article seem to point towards Endsley’s view that AI systems should be designed to facilitate human decision making when AI is involved by, for instance, improving the explainability of the output of the automated stage of hybrid moderation decisions.¹⁸² In that sense, although the LLMs may not be explainable themselves, they can turn into tools for the explanation of the decisions taken in moderation processes.

CONCLUSIONS

By delving into the use of LLMs as automated content moderation tools, this article has given a perspective on some of the continuous challenges posed by automated moderation in general. LLMs are groundbreaking tools that may assist in different stages of content moderation processes due to their notable capacities for providing reasoning in responses to prompts, as well as for analysing text. However, they, like other technical solutions, are far from being a silver bullet.

¹⁷³ *Morice v France* [2015] ECtHR [GC] 29369/10.

¹⁷⁴ *Jersild v. Denmark* (n 119), para 35.

¹⁷⁵ *ibid.*

¹⁷⁶ *Editorial Board of Pravoye Delo and Shtekel v Ukraine* [2011] ECtHR 33014/05, para 64.

¹⁷⁷ UNGA (n 171).

¹⁷⁸ ‘Content Moderation and Freedom of Expression Handbook’ (art 19 2023) <<https://www.article19.org/wp-content/uploads/2023/08/SM4P-Content-moderation-handbook-9-Aug-final.pdf>> accessed 10 April 2024.

¹⁷⁹ Kumar, AbuHashem and Durumeric (n 98).

¹⁸⁰ UNGA (n 10), para 29.

¹⁸¹ *ibid.*

¹⁸² Endsley (n 66).

Social media platforms are increasingly integrating public values, including human rights, into their operations. Initially driven by political and market pressures, this integration is now legally mandated by EU legislation, particularly the DSA.¹⁸³ Consequently, automated content moderation faces heightened challenges in incorporating human rights concerns under stricter legal constraints, influencing their procedures and policies.

In light of the above, the findings from previous research, and from the test conducted for this article on the limitations and advantages of LLM moderation, are all relevant from the point of view of the existing legislation governing social media platforms in Europe. LLMs can provide key insights into the reasoning used to justify limitations on freedom of expression online, which can be relevant to improve the internal procedures of social media platforms, as well as the way in which decisions are communicated.

In this context, integrating human rights principles, especially procedural ones, into automated moderation at various stages could significantly aid in implementing the DSA. The DSA prioritizes procedural safeguards over substance,¹⁸⁴ emphasizing the need for '[fundamental rights protection] during dispute resolution between users and platforms, [with] full transparency and explicitly defined procedural rights'.¹⁸⁵ Similarly, the ECtHR underscores the importance of fairness and procedural safeguards in assessing restrictions on freedom of expression under the European Convention.¹⁸⁶ Exploring these principles could lead to more robust solutions to the moderation dilemma.

FUNDING

This work was partially supported by the Wallenberg AI, Autonomous Systems and Software Program - Humanities and Society (WASP-HS) funded by the Marianne and Marcus Wallenberg Foundation and the Marcus and Amalia Wallenberg Foundation.

SUPPLEMENTARY MATERIAL

Supplementary material is available at *International Journal of Law and Information Technology* online.

¹⁸³ Wolfgang Schulz and Christian Ollig, 'Hybrid Speech Governance: New Approaches to Govern Social Media Platforms under the European Digital Services Act?' (2023) 14 JIPITEC –Law <<https://www.jipitec.eu/jipitec/article/view/22>> accessed 22 July 2024.

¹⁸⁴ Ortolani (n 36).

¹⁸⁵ Frosio and Geiger (n 29), 75.

¹⁸⁶ *Aydoğan Et Dara Radyo Televizyon Yayıncılık Anonim Şirketi c Turquie* [2018] ECtHR 122261/06.