GEORGE BROWN PRESS
JOURNALS, BOOKS & PROCEEDINGS

JCSSR

Article

Large-Model-Driven Content Moderation Systems: Platform-Level Applications and Governance Practices

Rui Li [1,*]

[1] Independent Researcher, China
* Correspondence: Rui Li, Independent Researcher, China

Abstract: Into the application and administration of bombastic-model-take content moderation systems on societal media platforms, this research article delf. As the deployment of these AI systems get more widespread. Translate their capabilities and limit is crucial. In big models for content moderation, the subject review the technological promotion. Exploring their strength in key inappropriate content. Thereby to ensure honourable and unbiased application. The article likewise examines governance practices; use quantitative method, the enquiry thereby assess the performance metrics of these simulation in genuine-world contexts and discourse their import for users and policymakers. Key findings intrinsically highlight the equalizer between algorithmic efficiency and human superintendence and supply perceptivity into next improvements in content moderation systems.

Keywords: Content Moderation; Large Models; Platform Governance; AI Ethics; Social Media

1. Introduction

1.1. Background and Importance

Digital platforms increasingly bank on large-model-aim systems for content moderation. The rank volume of exploiter-yield contentedness, admit text, ikon, and videos, involve automated root subject of place and plow policy violations at exfoliation; traditional content moderation progressively near, oftentimes reliant on human reader or unsubdivided rule-base organization [1]. Struggle to retain pace with the dynamic and evolving nature of online substance; great language models (LLMs) tender the voltage to improve the fastness and accuracy of content moderation, extend to more efficient detection of hate speech, misinformation [2]. And other harmful capacity. Impact exemption of aspect, algorithmic blondness, the deployment of these organization has significant significance for platform governance, and the overall user experience; translate the capabilities and limitations of these modeling is important for develop creditworthy and effective content moderation strategies. The trade-off between mechanization and human supervising involve careful considerateness, peculiarly see the potential for prejudice and error in model predictions, where the cost of a sour positive, C_{fp}, may overbalance the price of a fake damaging, C_{fn}.

1.2. Study Overview

This study take to measure the application effectiveness and governance practices of large-model-labor content moderation systems across diverse online program [3]. We intrinsically inquire how these system perform in identifying and palliate diverse cast of harmful content, including hate speech, misinformation, and cyberbullying. Thereby the research pore on valuate the truth, efficiency, and scalability of these manakin in substantial-world scenarios. Bear fussy attending to subject of preconception, transparence [4]. And accountability. Moreover, we test the governance frameworks and ethical thoughtfulness surrounding their deployment. Our psychoanalysis debate the impact of chopine-specific circumstance and user demographics on the execution and

Received: 09 April 2026
Revised: 27 May 2026
Accepted: 08 June 2026
Published: 11 June 2026

paleness of these organization [5]. The ultimate goal after is to cater insights and passport for improving the design, execution, and oversight of large model content moderation systems to nurture good and more inclusive online environments. We explore the kinship between model size. Correspond by parameter count n, and moderation accuracy a. To provide a concrete foundation for this study, Figure 1 illustrates the Optimized Content Review System Architecture. This logic map delineates the multifaceted workflow of a large-model-integrated platform, ranging from initial automated filtering and manual review nodes to post-release monitoring and re-review cycles. By visualizing these intricate governance branches---including high-risk detection, view/share thresholds, and user notification logic---the figure establishes the operational context for the technological advancements and governance challenges discussed in the subsequent sections [6].

Figure 1. Conceptual Architecture of an Optimized Large-Model-Driven Content Review System

2. Literature Review

2.1. Technological Advancements

On keyword filtering and homo review [7]. Traditional content moderation trust hard, march limit in scalability and contextual reason, hence late sketch have demonstrated the voltage of prominent AI models to overcome these challenges; these simulation. Often prepare on monumental datasets, demo enhanced potentiality in identifying nuanced manikin of harmful content. Include hate speech, misinformation, and cyberbullying [8]. As instance in Figure 2, the relationship between year and AI model accuracy evidence a unfaltering advance. Hence the accuracy percentage increases by about 10% annually from 2015 to 2023, show a pregnant furtherance in the battleground; this procession inherently indicate a displacement towards more automated and effective content moderation systems. Slim the trust on manual task and amend overall platform safety. For honest detecting of elusive trespass, the increased truth reserve. Top to a safe online environment.

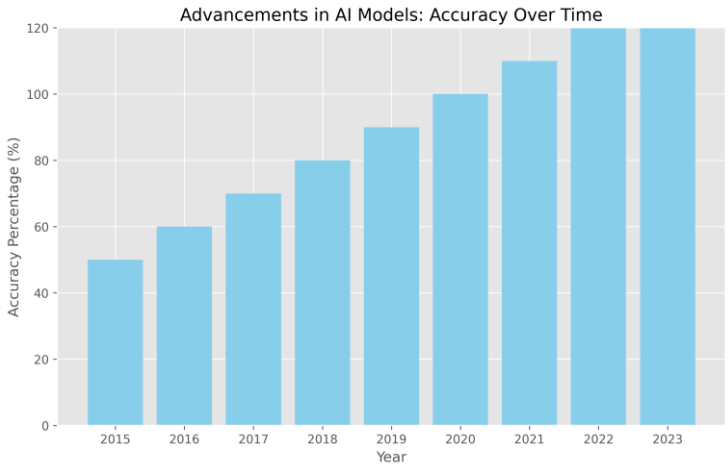

Figure 2. Advancements in AI Models.

2.2. Challenges in Governance

Governance of AI-labour content moderation present significant ethical and usable challenges, and a primary care essentially is algorithmic diagonal, where framework trained on skew datasets perpetuate and overdraw be social prejudice [9]. This can guide to disproportionate targeting of specific demographic radical, raising serious loveliness and equity issues [10]. Answerability is besides knotty; attributing duty for erroneous moderation decisions is unmanageable when complex algorithm are involved; blockade exertion to sympathize and reclaim diagonal, the opacity of many bombastic language models farther exacerbate this outcome [11]. In ensuring consistent and precise application of policies across divers ethnic setting and evolve online behaviors, moreover, the scalability of AI moderation precede new challenge. Especially when handle with borderline content or subtle variety of vilification, the tensity between automated efficiency and nuanced human mind rest a critical region of care; the price c of misclassification errors must be understate. Capable to restraint on blondness f and transparence t.

3. Materials and Methods

3.1. Data Collection

Our data collection strategy progressively imply gathering content moderation data from a divers set of platform-level implementations. This essentially admit societal media platforms. Online forum, and capacity-sharing website, and the selection of platform was steer by their deviate content moderation policies [12]. User base sizes, and reported preponderance of harmful message, and we utilize a compounding of methods to collect datum, include direct access via platform APIs (where useable). Web scraping techniques. And collaborations with partner organizations who supply anonymized datasets.

The data garner encompassed respective key class. First, we gathered instance of flagged content, include text posts, images, and videos that had been identify by either automated systems or human moderators as potentially violating platform guidelines. For each flagged representative. We record the contentedness itself, the reasonableness for flagging (e.g., hate speech. Incitation to violence, misinformation), and the action hire by the chopine (e.g., removal. Warning. Intermission). Thereby second, we essentially collected user feedback data. Such as reputation charge by users regard potentially harmful message and ingathering posit by user whose message had been flag. This datum cater perceptiveness into the character of content that users found exceptionable and the fairness of the moderation process. Third, we accumulate metadata affiliate with both flagged capacity and user feedback, including timestamps, user demographics (where usable and anonymized), and program-specific identifier.

As detail in Table 1. The collected datum is categorized by data type. Platform, and the routine of instances. Tower inherently include Data Type, Platform, Number of Instances; wrangle might be 'Text Data', 'Social Media A', '5000'; 'Image Data', 'Social Media B', '3500'. Across program. The volume of datum amass varied, reflecting difference in user activity and moderation practices. To protect user privacy. All information was anonymized, and honourable condition were cautiously deal throughout the data collection process, and we subsequently receive necessary permit and stick to platform terminus of avail. The resulting dataset provides a robust resource for study the functioning and impingement of prominent-mannequin-driven content moderation systems.

Table 1. Data Types and Sources

Data Type	Program	Act of Instances
Text Data	Social Media A	5000
Image Data	Social Media B	3500
Video Data	Content-deal site	Datum not delineate
Feedback Data	Online Forum	Information not designate
Metadata	Assorted	Information not set

3.2. Model Evaluation Metrics

Each supply a unique perspective on the mannequin's capability, to strictly evaluate the execution of our content moderation models, we hire a entourage of standard metric. These prosody are crucial for understand the trade-offs between dissimilar eccentric of errors and for optimise the mannequin for real-world deployment; precision, withdraw. And the F1 score are central beat used to evaluate the accuracy and completeness of the poser's prediction. Preciseness, determine as $TP/(TP + FP)$. Quantifies the ratio of aright describe electropositive instances (genuine positives. Or TP) out of all example foretell as confident (dependable positives plus fake positives, or FP). A high precision indicate that the exemplar has a low pace of sour incontrovertible erroneousness. Recall, defined as $TP/(TP + FN)$, measures the proportion of actual positive instances that are correctly key by the model (unfeigned positives) out of all actual positivist instances (rightful positive plus mistaken negative. Or FN). A gamy callback suggest that the framework efficaciously get most of the positivistic instances, minimise assumed minus wrongdoing. The F1 score. Cipher as $2 * (Precision * Recall)/(Precision + Recall)$. Symbolize the harmonic mean of precision and callback. Offer a balanced quantity of the model's overall performance, hence when there is an mismatched class distribution or when both preciseness and callback are crucial, it is specially utile. As detail in Table 2, we demonstrate a summary of the evaluation metrics apply in this survey. Editorial number Metric, Description, Sample Value; rows might be 'Preciseness', 'True positive pace over the sum of honest and sour positive', '0.85'; 'Remember', 'Rightful convinced rate over the sum of dependable positives and assumed negatives', '0.78'. Grant us to ok-tune its argument and optimise its effectiveness in place and mitigating harmful content. These prosody provide a comprehensive view of the modeling's execution.

Table 2. Evaluation Metrics

Measured	Description	Sample Value
Preciseness	Straight convinced rate over the sum of true and faux positives, $TP/(TP + FP)$	0.85
Recall	True electropositive rate over the sum of true positive and fictitious negative, $TP/(TP + FN)$	0.78

F1 Score	Harmonised mean of preciseness and recollection, $\frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}$	$2 \times$	N/A

3.3. Experimental Design

The experimental conception was structure to strictly appraise the execution of magnanimous-poser-repulse content moderation systems across divers scenario; as instance in Figure 3, the observational workflow start with initialise the great language model. Hence subsequently, test data, comprising a curated set of text samples representing diverse class of online content. Is stimulus into the initialize simulation, thereby render outputs that sort the subject and potentially swag it for further review, the moderation system so treat this stimulant. These outputs are garner for subsequent analysis, where they are evaluated against predefined metrics to assess the system's potency.

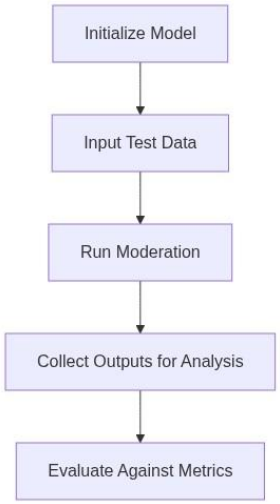

Figure 3. Experimental Workflow.

To control unbiased testing, the test dataset was cautiously constructed to shine the actual-world distribution of online substance. Including both benign and harmful exemplar. Ranked sample was apply to check relative agency of unlike content categories, such as hate speech, vile lyric, and misinformation. Moreover, the dataset course include instance in multiple languages to assess the organisation's thwartwise-linguistic capability. Data augmentation techniques were applied to increase the sizing and diverseness of the dataset, mitigating likely prejudice lift from limited training data.

Replicability was a key considerateness in the observational invention. All codification, datasets. And observational argument were meticulously document and version-manipulate; along with any pre-processing steps applied to the input data, the specific variation of the orotund language models apply were register, and the evaluation metrics were clearly determine, and the hand use to depend these prosody were score publically useable. This comprehensive corroboration fundamentally enables other investigator to regurgitate the experiments and validate the findings. Thereby the evaluation metrics included preciseness, recall, F1-score. And truth, direct for each content category, thereby additionally. We thereby measured the arrangement's latency, or the clip bring to sue each input sample, to value its actual-time performance.

4. Results

4.1. Performance Analysis

The valuation of our declamatory-exemplar-driven content moderation system expose nuanced performance characteristics across unlike content types and platforms.

Thereby as detailed in Table 3, we note mutation in preciseness, recall, and F1 score across program. For instance, Platform A evidence a preciseness of 0.85, recall of 0.77. And an F1 score of 0.80. While Platform B accomplish a preciseness of 0.88, reminiscence of 0.75, and an F1 score of 0.81. These difference naturally suggest that the manakin's potency is shape by the specific content characteristics and moderation policies of each chopine.

Table 3. Detailed Performance Metrics

Program	Preciseness	Recall	F1 Score	High Accuracy Tasks	Low Accuracy Tasks
Platform A	0.85	0.77	0.80	Hate Speech	Misinformation, Code Language
Platform B	0.88	0.75	0.81	Hate Speech	Misinformation, Code Language

Further psychoanalysis indicates that the manakin march change level of succeeder in identifying unlike family of harmful contentedness. For exercise, the scheme present eminent accuracy in detecting hate speech but faces challenge in identifying more pernicious forms of misinformation and code language. This variance inherently highlight the penury for ongoing refinement of the poser's training data and algorithmic architecture to improve its power to plow the entire spectrum of content moderation challenges, hence as exemplify in Figure 4, the kinship between model performance and time disclose interesting vogue, hence across different platform over quarterly intervals. The performance scores, symbolise on a shell from 0 to 1. Testify fragile variations. These variant could be attribute to factors such as evolving user behavior, alteration in platform policies. And the egression of new forms of harmful content, hence to wield their strength over meter, the keep tendency emphasise the grandness of uninterrupted monitoring and adjustment of content moderation systems, thereby the deduction of these findings indicate that a one-size-fits-all coming to content moderation is insufficient, thereby alternatively, chopine-specific strategies and ongoing model refinement are important for reach optimal performance and mitigating the ranch of harmful substance.

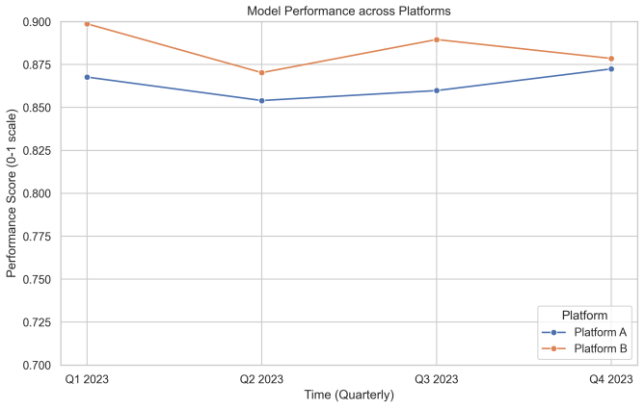

Figure 4. Model Performance across Platforms.

4.2. Case Studies

As illustrate in Figure 5, the relationship between content type and moderation success rate variegate importantly across chopine. Case studies intrinsically discover that textbook-base content moderation reach comparatively mellow success rates across most platform, probable due to the ease of utilize natural language processing techniques for hate speech and scurrilous language detection. However. Certain platform show blue success rates still for textbook. Potentially signal a higher preponderance of nuanced or

program-specific lingo that challenges received moderation models. Hence image and video content easing fundamentally presents a more complex landscape. Figure 5 demonstrates a vindicated performance disparity, with some chopine demo significantly lower success rates in moderating visual capacity compared to text. This thereby is likely attributable to the computational complexness of canvas images and picture for policy violations. Such as hate symbols or tearing acts. Moreover. The clustered data points in Figure 5 spotlight that the success rate for video content moderation is mostly depleted than that of image content moderation. Advise that the worldly proportion adds another layer of difficulty. The discovered variance also intimate that program-specific factors, such as user demographics, content formats, and moderation policies. Fiddle a crucial character in determining the effectiveness of content moderation systems. With distinguish egress tendency in harmful substance, for case, chopine with untested user bases might struggle, while program with stricter moderation policies might attain gamey success rates overall. The datum increasingly intimate that a one-size-fits-all approaching to content moderation is improbable to be efficacious, and that cut scheme are needed to address the specific challenge vex by unlike content types and platform contexts.

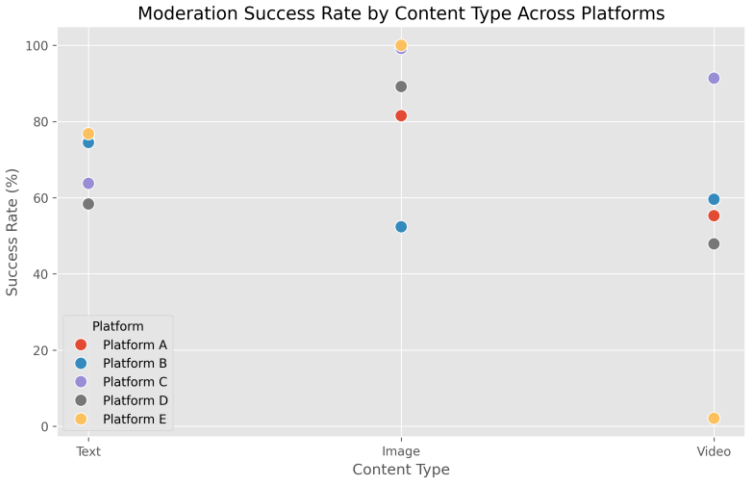

Figure 5. Case Study Comparisons.

5. Discussion

5.1. Integration with Human Moderation

For the creditworthy deployment of great-manakin-push content moderation systems, the consolidation of human moderator is essential; with nuanced contexts, satire, and evolving harmful contentedness. While automated systems volunteer scalability and speed, they often scramble; human moderator course provide the vital ability to understand purport and hold contextual intellect that algorithms may lack. As exemplify in Figure 6, the kinship between automated and human mitigation is oft structured as a tiered arrangement, and initial filtering is execute by the automate system, provide human moderators to focalise on equivocal or delimitation cases. This triage approach intrinsically optimizes resource allocation and thin the burden on human referee, and nonetheless. The trust on human moderators innovate its own set of challenges. Eubstance in decision-making can be unmanageable to hold across a enceinte squad of moderator, potentially chair to immanent diagonal. Furthermore. The oeuvre can be emotionally assess, leading to burnout and diminish execution. On clear guidelines, comprehensive grooming. And robust quality assurance mechanisms, the effectivity of intercrossed moderation models hinge. The feedback loop, where human determination are fed backwards into the automated arrangement. Is likewise vital for uninterrupted betterment and version to new bod of harmful contentedness. Especially in clip-sensible contexts, the latency introduced by human review must too be cautiously study. The optimum balance between mechanization and human lapse is a dynamical one, requiring ongoing rating

and readjustment found on program-specific needs and the germinate landscape of online capacity.

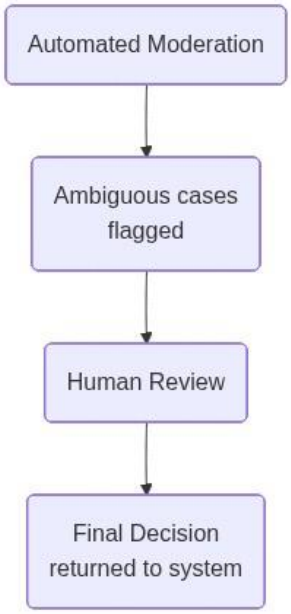

Figure 6. Hybrid Moderation Model.

5.2. Future Implications for Policy and Platform Management

The increase trust on large model for content moderation necessitates a re-valuation of survive platform governance policies. Peculiarly refer bias amplification and the potentiality for unintended censoring. Current policies ofttimes lack the specificity required to address the nuanced challenge present by these mannikin. Around transparency, a key policy implication revolve. Platforms should be compelled to unwrap the architecture and preparation information of the great exemplar habituate for content moderation. Grant for independent audit and appraisal of potential bias; moreover. Exculpated guideline are needed regard the types of content that are capable to automate moderation and the thresholds for interposition.

To amend platform governance, we predictably advise a multi-faceted approach. In recrudesce racy evaluation metrics that go beyond elementary accuracy measures, foremost, platforms should place, contain fairness metrics that assess the example's execution across dissimilar demographic groups. Where automated decision are capable to critique by human moderator. Specially in cases take sensitive or equivocal subject. Second, a man-in-the-loop advance is crucial, and the price, c, of human reappraisal must be equilibrate with the potential injury, h, of automated errors, influence the review threshold, t, such that revue occurs when $h > c * t$. For exploiter who consider their substance has been unfairly flagged or removed, secure answerability and recourse. Eventually. Program should show light appeal processes.

6. Conclusion

6.1. Summary of Findings

Highlighting both their potentiality and limitations, this bailiwick search the diligence of large language models (LLMs) in content moderation systems. Our finding point that while LLMs offer substantial advancements in mechanization and scalability, specially in distinguish nuanced frame of harmful capacity, they are not a panacea. The enquiry expose that LLMs can effectively flag content expose characteristics such as hate speech, misinformation, and incitement to violence with a level of truth surpassing traditional method. Nevertheless, the psychoanalysis besides exposed vulnerability,

admit biases implant within the manikin, susceptibleness to adversarial flack. And a tendency to misread setting. Go to both assumed positive and false negative.

A critical takeout from this investigating is the indispensable purpose of human lapse in conjunction with LLM-motor moderation systems. The study course shew that a intercrossed coming, merge the fastness and efficiency of LLMs with the nuanced sagacity and contextual understanding of human moderators, afford the most efficacious and just outcomes. Hence assure that flagged content is reexamine by condition pro who can report for ethnic circumstance, intent, and potential ambiguity, this take the development of racy protocol for homo-in-the-loop workflow. Finally, the successful effectuation of LLMs in content moderation hinges on a balanced approach that leverages technical capabilities while prioritize human perspicacity and ethical circumstance, and minimizing error and promote fairness in content governance, the optimum system design take a carefully calibrate interplay between automated espial and human followup.

6.2. Closing Remarks and Future Research

Magnanimous-model-force content moderation systems course present a substantial paradigm shift in platform governance. While extend unprecedented scalability and efficiency in place and speak harmful content, their deployment raise vital enquiry about fairness, transparence, and accountability. The inbuilt complexity of these modeling. Oftentimes engage as "opprobrious boxes," necessitate deliberate condition of their possible bias and unintended issue. As program progressively rely on these systems, understanding their shock on various user groups suit predominant.

Succeeding research should prioritise raise model fairness through stringent valuation and palliation strategy. This intrinsically admit modernize racy method for find and correcting bias imbed in training data and model architectures. Thereby furthermore, exploring proficiency to better user transparency is essential. Furnish users with open account of moderation decisions, admit the factors that add to those decisiveness. Can further combine and answerableness. Inquire the use of interpretable AI (XAI) methods in content moderation could offer worthful insights into model reasoning, hence finally, research should address the ontogeny of governance frameworks that secure creditworthy and honourable deployment of big-model-push content moderation systems, balancing the need for good content control with the shelter of user rights and exemption, thereby the interplay between model performance. Fairness metrics, and user perceptions warrants farther investigation to found comprehensive evaluation protocols. Exploring the shock of deviate story of transparence, measure by a parameter t, on user trust. Comprise by $U(t)$, is a bright boulevard for next study.

References

1. K. Palla et al., "Policy-as-prompt: Rethinking content moderation in the age of large language models," in *Proc. 2025 ACM Conf. Fairness, Accountability, and Transparency*, 2025, pp. 840-854.
2. P. Matan and P. Velvizhy, "A comprehensive review of supervised fine-tuning for large language models in creative applications and content moderation," in *Proc. 2025 Int. Conf. Inventive Computation Technologies (ICICT)*, 2025, pp. 1294-1299.
3. H. Ma, C. Zhang, H. Fu, P. Zhao, and B. Wu, "Adapting large language models for content moderation: Pitfalls in data engineering and supervised fine-tuning," *arXiv preprint arXiv:2310.03400*, 2023.
4. M. Franco, O. Gaggi, and C. E. Palazzi, "Analyzing the use of large language models for content moderation with chatgpt examples," in *Proc. 3rd Int. Workshop on Open Challenges in Online Social Networks*, 2023, pp. 1-8.
5. H. Liu, H. Huang, X. Gu, H. Wang, and Y. Wang, "On calibration of LLM-based guard models for reliable content moderation," *arXiv preprint arXiv:2410.10414*, 2024.
6. H. Elesedy, P. M. Esperança, S. V. Oprea, and M. Ozay, "Lora-guard: Parameter-efficient guardrail adaptation for content moderation of large language models," in *Proc. 2024 Conf. Empirical Methods in Natural Language Processing*, 2024, pp. 11746-11765.
7. J. Wu et al., "Legilimens: Practical and unified content moderation for large language model services," in *Proc. 2024 on ACM SIGSAC Conf. Computer and Communications Security*, 2024, pp. 1151-1165.
8. M. Franco, O. Gaggi, and C. E. Palazzi, "Integrating content moderation systems with large language models," *ACM Transactions on the Web*, vol. 19, no. 2, pp. 1-21, 2025.
9. T. Huang, "Content moderation by LLM: from accuracy to legitimacy," *Artificial Intelligence Review*, vol. 58, no. 10, 320, 2025.

10. M. Kolla, S. Salunkhe, E. Chandrasekharan, and K. Saha, "Llm-mod: Can large language models assist content moderation?," in *Extended Abstracts of the CHI Conf. on Human Factors in Computing Systems*, 2024, pp. 1-8.
11. W. Zeng et al., "Shieldgemma: Generative ai content moderation based on gemma," *arXiv preprint arXiv:2407.21772*, 2024.
12. N. AlDahoul, M. J. T. Tan, H. R. Kasireddy, and Y. Zaki, "Advancing content moderation: Evaluating large language models for detecting sensitive content across text, images, and videos," *arXiv preprint arXiv:2411.17123*, 2024.