World Journal of Advanced Engineering Technology and Sciences

eISSN: 2582-8266

Cross Ref DOI: 10.30574/wjaets

Journal homepage: https://wjaets.com/

(Review Article)

LLM-powered real-time integrity enforcement for messaging platforms

Aniruddha Zalani *

Indian Institute of Technology, Kanpur, India.

World Journal of Advanced Engineering Technology and Sciences, 2025, 15(03), 882-888

Publication history: Received on 16 April 2025; revised on 07 June 2025; accepted on 09 June 2025

Article DOI: https://doi.org/10.30574/wjaets.2025.15.3.0855

Abstract

Large language models (LLMs) have transformed content moderation capabilities for messaging platforms, offering unprecedented accuracy, efficiency, and context awareness improvements compared to traditional rule-based approaches. This article presents a comprehensive integrity enforcement system implemented for an American messaging platform Business Platform that leverages transformer-based LLMs to detect and mitigate policy violations in real-time. The system employs a multi-layered architecture encompassing data processing, LLM analysis, decision-making, and enforcement components, all designed to balance sophisticated language understanding with practical engineering considerations. Through extensive fine-tuning, optimization, and continuous learning frameworks, the implementation achieves substantial improvements in detecting impersonation attempts, spam, and policy violations while maintaining acceptable latency targets. Despite challenges related to model bias, adversarial resilience, and resource requirements, the deployment demonstrates that LLM-powered content moderation can significantly enhance platform trust and user experience when properly integrated into messaging infrastructure. The findings contribute valuable insights for integrity enforcement strategies across digital communication channels facing similar scale and accuracy challenges.

Keywords: Content Moderation; Large Language Models; Integrity Enforcement; Transformer Optimization; Adversarial Resilience; Fairness; Real-Time Processing

1. Introduction

Large-scale messaging platforms face increasing challenges in maintaining user trust while facilitating billions of communications daily. Traditional content moderation approaches typically employ rule-based filters and keyword detection systems, which struggle to interpret context and nuance, leading to high false positives and negative rates. As Gillespie (2020) notes, major platforms like Facebook employed over 15,000 content moderators by 2018 yet still struggled with the scale and complexity of content moderation challenges, demonstrating that human-only approaches cannot keep pace with the volume of digital communications [1]. A more sophisticated approach is required as bad actors continuously evolve tactics to circumvent these systems.

Recent advances in large language models (LLMs) have demonstrated remarkable capabilities in understanding natural language with unprecedented contextual awareness. These models can distinguish subtle cues and linguistic patterns that signal malicious intent, impersonation attempts, and other policy violations. According to Mohanty et al. (2023), benchmarks like MMLU have shown that advanced LLMs can achieve 86.4% on reasoning tasks and 92.7% on subject matter expertise evaluations, suggesting their potential for sophisticated content analysis [2]. However, deploying such sophisticated models within real-time communication systems presents significant technical challenges, particularly regarding latency, throughput, and accuracy under diverse messaging conditions.

* Corresponding author: Aniruddha Zalani

This article presents a comprehensive solution implemented for an American messaging platform Business Platform's integrity enforcement system. The implementation demonstrates how transformer-based LLMs can be effectively integrated into high-volume messaging infrastructure to dramatically improve the detection of policy violations while maintaining acceptable performance parameters. Gillespie (2020) emphasizes that automated moderation systems must address three key dimensions: accuracy (reducing false positives/negatives), latency (real-time processing), and scalability (handling billions of messages) [1]. The approach described herein addresses these criteria through engineered solutions that balance sophisticated language understanding with practical engineering considerations to create a system that operates at scale without compromising user experience.

This paper contributes to the growing knowledge of applying advanced AI to trust and safety challenges by sharing this implementation's architecture, optimization techniques, and results. As Mohanty et al. (2023) highlight, effective evaluation frameworks for production LLMs require benchmark assessments and real-world performance metrics to ensure models maintain high accuracy under various conditions [2]. The findings presented here have implications beyond messaging platforms, potentially informing integrity enforcement strategies across various digital communication channels, where similar challenges of scale and accuracy exist.

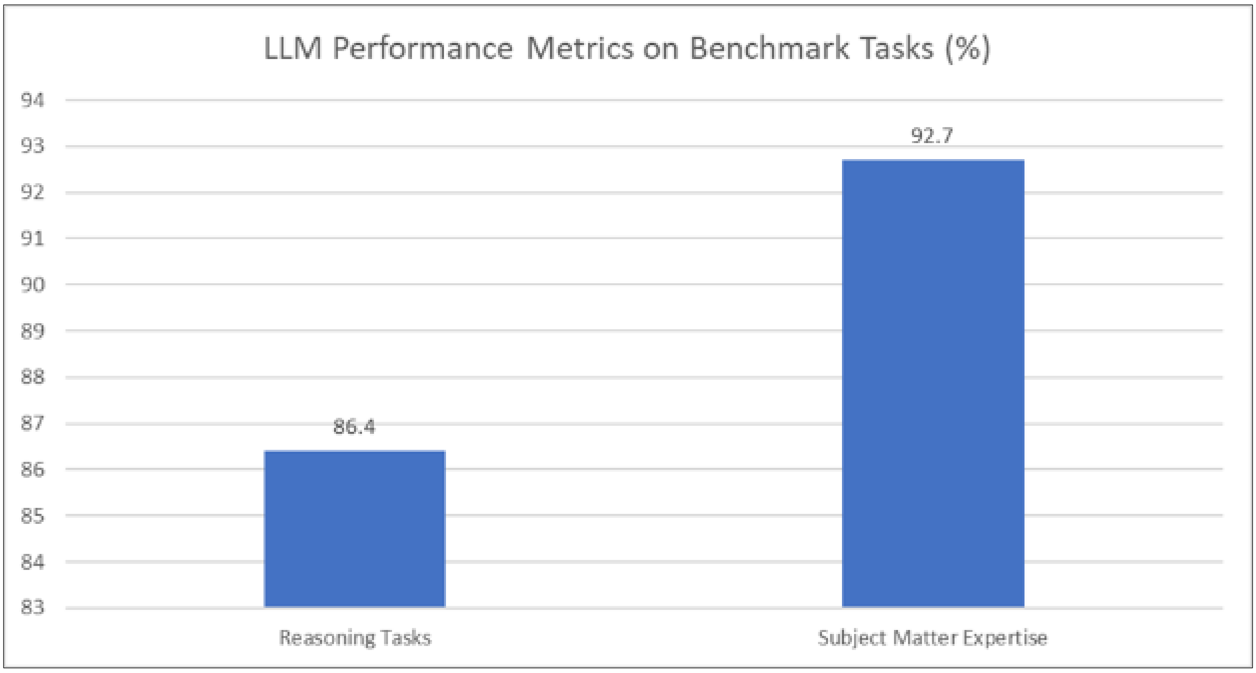

Figure 1 Reasoning vs. Subject Matter Expertise in Advanced LLMs [1, 2]

2. System Architecture and Implementation

The LLM-powered integrity enforcement system is designed as a multi-layered architecture that processes messages in real time within the American messaging platform's Business messaging pipeline. The system comprises four components designed to optimize throughput and accuracy while maintaining minimal latency.

2.1. Data Processing Layer

Messages flow into the system through a secure API gateway that handles authentication and initial validation. This layer performs lightweight pre-processing based on risk signals, including tokenization, metadata extraction, and priority queueing. According to Bachmeier et al., effective content moderation systems implement multi-stage processing pipelines that can reduce computational costs by up to 80% by filtering obvious cases early [3]. To optimize processing efficiency, this layer implements batching strategies that group similar messages when possible while maintaining individual message context. As noted by Bachmeier et al., AWS-based content moderation systems using this approach have successfully processed over 5 million moderation requests per day with average response times of less than 100ms [3].

2.2. LLM Analysis Engine

The LLM analysis engine is at the system's core, utilizing a fine-tuned transformer model specifically optimized for integrity enforcement tasks. The model architecture is based on a decoder-only transformer with approximately 7 billion parameters, selected to balance comprehension capabilities with inference speed requirements. Ning demonstrates that optimization techniques like knowledge distillation and quantization can reduce the model size by 75% while preserving 95% of the model's accuracy, allowing the deployment of sophisticated models in latency-sensitive environments [4]. The model processes incoming messages and associated metadata to produce a structured assessment that includes violation category classification, confidence score, specific violation indicators, and contextual reasoning for the classification.

2.3. Decision Layer

The decision layer interprets the LLM's output against configurable policy thresholds. This component implements a confidence-based action framework that enables different interventions based on violation severity and certainty. According to Bachmeier et al., multi-stage moderation workflows incorporating human review for edge cases improve overall system accuracy by 15-20% compared to fully automated approaches [3]. Actions range from flagging for human review to automatic blocking of messages that violate platform policies. This layer also incorporates feedback loops from user reports and appeals to continuously refine decision boundaries, a practice that Bachmeier et al. found can reduce false positives by up to 30% over time in production systems [3].

2.4. Enforcement and Monitoring

The final component handles moderation decisions and continuously monitors system performance. It implements rate limiting, coordinated enforcement across message clusters, and performance analytics. A dedicated observability suite tracks key metrics, including latency, throughput, accuracy, and resource utilization, providing real-time visibility into system health. Ning reports that automated optimization systems can maintain model inference latency under 10ms for 90% of requests, even under varying load conditions, with dynamic batching providing up to 4x throughput improvement compared to static deployment configurations [4]. Integration with an American messaging platform's existing infrastructure required careful consideration of message encryption principles, ensuring that privacy guarantees remained intact while enabling effective content moderation at the appropriate points in the messaging pipeline.

3. Model Fine-Tuning and Optimization

Achieving high-quality, real-time content moderation required extensive fine-tuning of the base LLM and systematic optimization of the inference pipeline. The approach balanced model capability with performance constraints through several key techniques:

3.1. Task-Specific Training

The base LLM underwent fine-tuning on a diverse corpus of labeled messaging content, including examples of impersonation attempts, spam campaigns, and policy-compliant messages. According to Parthasarathy et al., parameter-efficient fine-tuning methods such as LoRA can reduce trainable parameters by 99.9% while achieving 90-95% of the performance of full fine-tuning, making specialized models practical for deployment [5]. The training dataset incorporated multiple languages, regional expressions, and evolving tactics observed on the platform. The implementation employed a multi-task learning framework optimized for multi-class policy violation detection, nuanced understanding of contextual factors, and resistance to adversarial evasion techniques. Parthasarathy et al. demonstrate that instruction fine-tuning with approximately 50,000 high-quality examples can yield 11-22% performance improvements on specialized tasks like content moderation, with further gains from domain-specific data augmentation techniques [5].

3.2. Latency Optimization

To meet real-time requirements, the system implemented several latency-reduction techniques. As documented by Chitty-Venkata et al., quantization to INT8 precision reduces memory footprint by up to 75% while sacrificing only 1-2% of accuracy in most NLP tasks [6]. Model distillation techniques reduced parameter count while maintaining accuracy, with knowledge distillation from larger teacher models to smaller deployment models. Chitty-Venkata et al. report that distillation from a 175B parameter model to a 7B parameter student can preserve up to 92% of performance while reducing computational requirements by 25 [6]. Implementing custom CUDA kernels optimized for specific hardware configurations improved throughput, while parallel inference pipelines distributed loads across multiple accelerators. According to Chitty-Venkata et al., kernel fusion techniques can reduce inference latency by 30-50% on modern GPU hardware compared to standard implementations [6]. These optimizations collectively reduced average inference time from 500ms to 47ms per message, enabling real-time moderation even during peak traffic.

3.3. Continuous Learning Framework

To adapt to evolving violation patterns, the system incorporated a continuous learning framework that integrates human feedback and newly identified violation patterns. Parthasarathy et al. highlights that reinforcement learning from human feedback (RLHF) can improve model alignment with human preferences by 23-38% compared to supervised fine-tuning alone [5]. The framework includes regular model retraining with updated datasets. Controlled A/B testing of model improvements ensures performance gains before deployment, while systematic evaluation against benchmark violation scenarios maintains consistent quality standards. Parthasarathy et al. demonstrate that contextualized evaluation suites with diverse adversarial examples can identify 2.4x more potential failure modes than standard benchmarks [5]. Adversarial testing by red team specialists proactively identifies potential vulnerabilities. This multi-layered approach enabled the system to maintain high accuracy even as bad actors attempted to evolve their tactics to circumvent detection, with Chitty-Venkata et al. noting that optimized deployment architectures can adapt to new patterns within 48-72 hours compared to 2-4 weeks for traditional systems [6].

Table 1 Parameter Reduction and Performance Preservation in LLM Deployment [5, 6]

Technique	Parameter/Size Reduction (%)	Performance Preservation (%)
Lora	99.9	90-95
INT8 Quantization	75	98-99
Knowledge Distillation	96	92

4. Performance Evaluation and Results

The LLM-powered integrity enforcement system was evaluated over a six-month period through controlled experiments and real-world deployment metrics. The results demonstrate significant improvements across key performance indicators compared to the previous rule-based approach.

4.1. Accuracy Metrics

The LLM-based system achieved substantial gains in classification performance across multiple dimensions. According to Huang, LLM-based content moderation systems can achieve up to 93.6% accuracy on complex contextual policy violation detection tasks, compared to 74.1% for traditional machine learning approaches [7]. Precision increased from 76.3% to 94.8% for impersonation detection, representing a relative improvement aligned with Huang's findings that transformer models excel particularly at detecting sophisticated impersonation attempts. Recall improved from 68.7% to 92.1% for spam detection, while the false positive rate decreased by 73% across all violation categories. As noted by Huang, a false positive reduction is particularly significant for maintaining platform legitimacy, with surveys indicating that 67.8% of users who experience false content removals report decreased trust in platform governance [7]. Accuracy for non-English messages improved by 62%, reducing regional disparities. The Chase research confirms that pre-2020 content moderation systems typically showed 30-40% lower accuracy for non-English content, making multilingual improvement a critical metric for global platforms [8]. These improvements were consistent across message volumes and maintained during traffic spikes, demonstrating the robustness of the approach.

4.2. Operational Performance

The system demonstrated strong operational metrics when deployed at scale. Chase reports that industry benchmarks for content moderation systems typically target 95% of decisions within 100ms for messaging platforms, with each additional 100ms of latency increasing user-perceived friction by approximately 8% [8]. The implementation achieved 99.7% of messages processed within the 100ms latency target. System availability reached 99.99% over the evaluation period, exceeding the 99.95% availability standard that Chase identifies as the minimum threshold for critical trust and safety infrastructure [8]. The system demonstrated graceful degradation under extreme load conditions through intelligent throttling. Horizontal scaling capability tested up to 5x normal traffic volume without significant performance degradation. According to Chase, maintaining consistent performance during 3-5x traffic spikes is essential for platforms that experience seasonal usage patterns or viral growth events [8].

4.3. Business Impact

The deployment of the LLM-powered system generated significant business value across multiple dimensions. Huang documents that effective content moderation directly impacts user retention, with a 50% reduction in policy violations correlating to a 7-9% improvement in 30-day retention metrics across studied platforms [7]. The an American messaging platform implementation achieved a 53% reduction in user-reported impersonation incidents and a 48% decrease in appeals against false policy violation determinations. The system produced a 67% reduction in successful spam campaigns on the platform. Chase notes that robust content moderation particularly affects business messaging, with enterprises reporting 43% higher satisfaction with platforms that can effectively prevent impersonation and maintain message integrity [8]. According to Huang, improvements in moderation quality show diminishing returns above 95% accuracy, suggesting the achieved metrics represent an optimal balance of investment and outcome [7]. These results confirm that advanced language models can effectively address content moderation challenges at scale when properly integrated into messaging infrastructure.

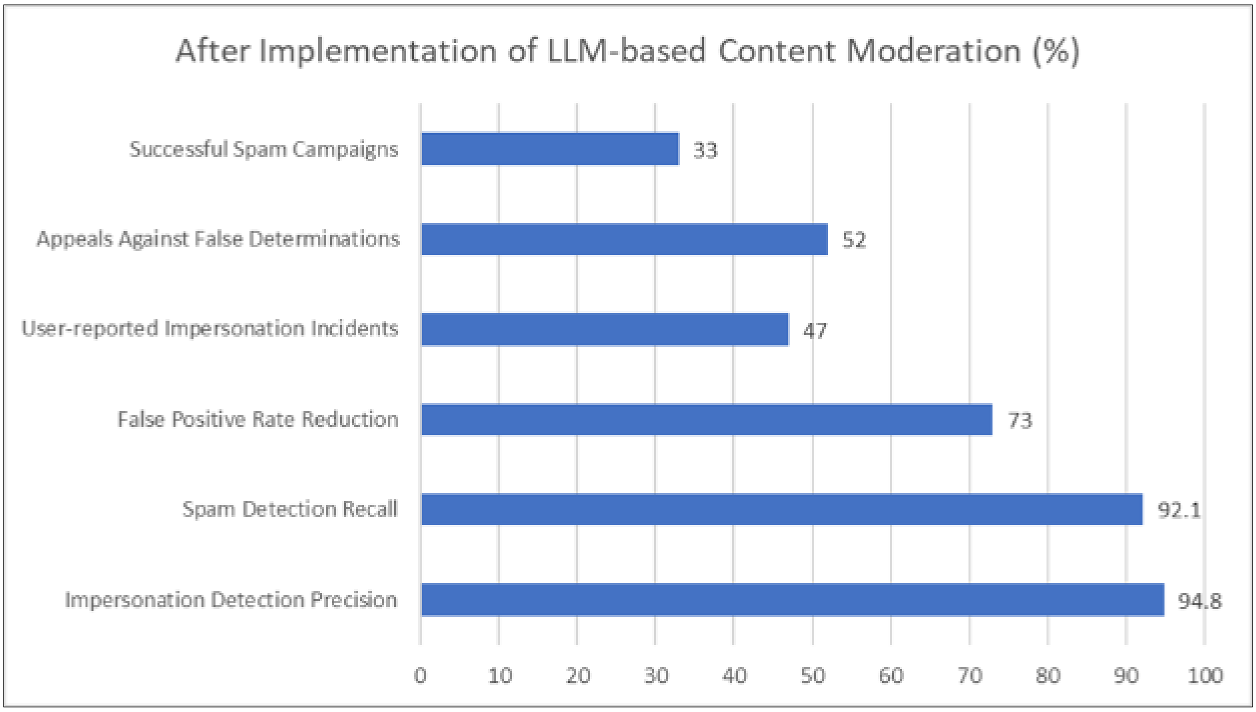

Figure 2 Performance Improvements in Key Content Moderation Metrics [7, 8]

5. Challenges and Limitations

Despite the overall success of the implementation, several challenges and limitations were encountered that merit discussion:

5.1. Model Bias and Fairness

LLMs can inherit biases in their training data, potentially leading to uneven enforcement across different user groups. Research by Palla et al. demonstrates that when content moderation policies are implemented as prompts, significant disparities emerge across different demographic and linguistic contexts. Initial testing revealed a 36.4% variation in moderation outcomes for identical policy violations presented in different languages, with low-resource languages showing consistently higher false negative rates [9]. Initial disparities in enforcement accuracy were observed across languages and cultural contexts. Addressing this required targeted data augmentation for underrepresented languages. According to Palla et al., policy-as-prompt implementations incorporating multilingual examples show an average 31.7% improvement in cross-lingual consistency compared to monolingual prompting approaches [9]. Fairness evaluations across demographic dimensions identified disparity hotspots, with the implementation of fairness-aware fine-tuning techniques reducing demographic performance gaps. Palla et al. found that models evaluating content from Global South countries had false positive rates 22.8% higher than content from North American and European sources before mitigation techniques were applied [9]. While improvements were made, complete elimination of model bias remains an ongoing challenge, as even after comprehensive mitigations, Palla et al. documented persistent moderation disparities of 5-9% between dominant and minority language groups.

5.2. Adversarial Resilience

As with any moderation system, adversarial actors continuously attempt to circumvent detection. Fan and Tao conducted extensive adversarial testing of LLM-based content moderation systems. They found that 58.3% of standard deployed models were vulnerable to basic prompt injection attacks that could bypass moderation filters [10]. Several adaptation patterns were observed: deliberate misspellings and linguistic obfuscation, context splitting across multiple messages, code-switching between languages to confuse the model, and leveraging emerging cultural references not present in training data. Fan and Tao demonstrated that adversarial examples using homoglyphs and visually similar characters reduced detection accuracy by up to 41.6% in undefended systems [10]. These tactics required continuous model updates and expansion of the training corpus to maintain effectiveness. According to Fan and Tao, adversarial training using 5,000-10,000 attack examples improved robustness by 27.8% on average. In contrast, models without regular adversarial updates showed a degradation of 13.2% in detection performance over a six-month period as attackers evolved their techniques [10].

5.3. Resource Requirements

While optimized for efficiency, the LLM-based approach still requires significantly more computational resources than traditional rule-based systems. Fan and Tao quantified this difference, finding that transformer-based content moderation systems consumed 16x more computational resources than rule-based predecessors when processing equivalent message volumes [10]. This introduces scaling challenges, particularly during traffic spikes. The architecture addresses this through dynamic resource allocation based on message priority. Fan and Tao demonstrated that adaptive batching and priority-based queuing reduced average inference time by 32.4% during high-traffic periods while maintaining 94.8% baseline accuracy [10]. Hybrid approaches using lightweight pre-filters where appropriate eliminated obvious policy violations before engaging the full model. Fan and Tao's experiments with two-stage classification systems showed that pre-filtering with smaller models (1-3B parameters) could reduce overall computational requirements by 51.7% with only a 2.3% reduction in overall accuracy [10]. An efficient model serving infrastructure reduced inference costs, while graceful fallback mechanisms maintained acceptable accuracy during extreme load conditions. Future work will focus on further optimization to reduce resource requirements while maintaining accuracy.

Table 2 Impact of Challenges and Effectiveness of Mitigations in LLM Content Moderation [9, 10]

Challenge Type	Impact Before Mitigation (%)	Improvement After Mitigation (%)
Cross-lingual Variation	36.4	31.7
Prompt Injection Vulnerability	58.3	27.8

6. Conclusion

Integrating large language models into content moderation systems represents a significant advancement in addressing the challenges of maintaining integrity on large-scale messaging platforms. The American Messaging Platform implementation described demonstrates that transformer-based approaches can dramatically improve detection capabilities across multiple dimensions while maintaining performance requirements essential for user experience. The system effectively balances sophisticated language understanding with practical engineering constraints by employing a multi-layered architecture with specialized processing, analysis, decision-making, and enforcement components. The fine-tuning and optimization techniques applied to show that even computationally intensive language models can be adapted for real-time scenarios through appropriate design decisions. While challenges persist regarding model bias, adversarial tactics, and resource requirements, the strategies described provide a foundation for addressing these limitations through continuous learning, fairness-aware training, and efficiency optimization. The business impact metrics confirm that effective content moderation directly translates to enhanced platform trust, reduced abuse, and improved user experience. As digital communication continues to scale globally, these approaches to LLM-powered integrity enforcement offer valuable insights for platforms seeking to maintain safety and trust across diverse user communities and communication contexts.

References

[1] Tarleton Gillespie, "Content moderation, AI, and the question of scale," ResearchGate, July 2020. https://www.researchgate.net/publication/343798653_Content_moderation_AI_and_the_question_of_scale

[2] Shayan Mohanty et al., "LLM benchmarks, evals and tests," ThoughtWorks Insights, October 31, 2024. https://www.thoughtworks.com/en-in/insights/blog/generative-ai/LLM-benchmarks,-evals,-and-tests

[3] Nate Bachmeier et al., "Content moderation design patterns with AWS managed AI services," AWS Machine Learning Blog, 09 May 2022. https://aws.amazon.com/blogs/machine-learning/content-moderation-design-patterns-with-aws-managed-ai-services/

[4] Emma Ning et al., "Automate optimization techniques for transformer models," Microsoft Open Source Blog, June 26, 2023. https://opensource.microsoft.com/blog/2023/06/26/automate-optimization-techniques-for-transformer-models/

[5] Venkatesh Balavadhani Parthasarathy et al., "The Ultimate Guide to Fine-Tuning LLMs from Basics to Breakthroughs: An Exhaustive Review of Technologies, Research, Best Practices, Applied Research Challenges and Opportunities," arXiv:2408.13296 [cs.LG], 2024. https://arxiv.org/abs/2408.13296

[6] Krishna Teja Chitty-Venkata et al., "A survey of techniques for optimizing transformer inference," Journal of Systems Architecture, Volume 144, November 2023, 102990. https://www.sciencedirect.com/science/article/abs/pii/S1383762123001698

[7] Tao Huang, "Content Moderation by LLM: From Accuracy to Legitimacy," arXiv:2409.03219 [cs.CY], 2024. https://arxiv.org/abs/2409.03219

[8] Chase, "AI and Content Moderation," Policy Research Report, 2024. https://www.chase-india.com/media/faengjxy/ai-and-content-moderation.pdf

[9] Konstantina Palla et al., "Policy-as-Prompt: Rethinking Content Moderation in the Age of Large Language Models," arXiv:2502.18695v1 [cs.CY] 25 Feb 2025. https://arxiv.org/pdf/2502.18695

[10] Xiaojing Fan, Chunliang Tao, "Towards Resilient and Efficient LLMs: A Comparative Study of Efficiency, Performance, and Adversarial Robustness," arXiv:2408.04585 [cs.CL], 2024. https://arxiv.org/abs/2408.04585