GEORGE BROWN PRESS
JOURNALS, BOOKS & PROCEEDINGS

Journal of Computer, Signal, and System Research

Article

UniGuard-Cascade: A Unified LLM-Driven Safety Scoring and Multi-Stage Audit Framework for Cross-Platform User Comments

Hao Tan [1,*] and Ziming Chen [2]

[1] Guangdong University of Technology, Guangzhou, China
[2] Computer Science, Johns Hopkins University, Baltimore, MD, 21218, USA
* Correspondence: Hao Tan, Guangdong University of Technology, Guangzhou, China

Abstract: User-generated comments across social media, e-commerce platforms, and online forums pose increasing challenges for automated content safety management. Existing moderation systems typically target a single platform, rely on inconsistent labeling standards, and struggle to balance accuracy with operational cost. To address these limitations, we propose UniGuard-Cascade, a unified safety scoring and cascading moderation framework that integrates large language models (LLMs), lightweight classifiers, and retrieval-augmented verification. UniGuard-Cascade introduces a platform-agnostic unified safety taxonomy covering toxicity, hate speech, spam, sexual/NSFW content, and misinformation, generated through LLM-based label alignment and normalization. The system operates in a three-stage cascade: (1) Fast Path Screening using a DistilBERT-based multi-label classifier for low-cost filtering; (2) Slow Path LLM Examination that provides refined labels and natural-language explanations; and (3) RAG-Enhanced Misinformation Verification, retrieving external evidence to validate factuality. Experiments conducted on four real-world datasets-Twitter, Amazon Reviews, Reddit, and YouTube Comments-show that UniGuard-Cascade consistently outperforms existing moderation baselines, achieving a Micro-F1 of 0.91 and a ROC-AUC of 0.97. The framework further reduces LLM usage by 87 percent compared with single-stage GPT-4 moderation, yielding a 5.7×reduction in overall inference cost while maintaining state-of-the-art multi-platform safety performance.

Keywords: content moderation; large language models; DistilBERT; cross-platform comment safety; unified label taxonomy; RAG; cascading model; Trust & Safety

Received: 18 November 2025
Revised: 25 December 2025
Accepted: 07 January 2026
Published: 14 January 2026

1. Introduction

User-generated comments have become a defining component of contemporary digital ecosystems, spanning social media platforms, e-commerce sites, online forums, and community-driven content services. As these platforms continue to scale globally, comment streams exhibit increasing heterogeneity in linguistic style, discourse norms, and safety risks. This diversity amplifies long-standing challenges in automated content moderation, particularly when attempting to ensure consistent safety standards across platforms with fundamentally different user behaviors and regulatory requirements. Toxic language, hate speech, fraudulent promotions, sexual or explicit content, and misinformation remain persistent threats to user well-being, platform credibility, and international compliance frameworks such as the EU Digital Services Act (DSA) and emerging global Trust & Safety regulations.

Despite recent advances in large language models (LLMs), existing moderation systems remain limited in three aspects. First, moderation taxonomies are fragmented: different platforms adopt incompatible safety labels, making it difficult to build scalable, cross-platform moderation systems. Second, while LLMs demonstrate strong

generalization, directly applying them to high-volume comment streams is computationally costly and often infeasible for real-time operations. Third, misinformation detection-which increasingly relies on evidence-based verification-remains particularly challenging for moderation models without external knowledge access.

To address these limitations, we introduce UniGuard-Cascade, a unified LLM-driven safety scoring and multi-stage audit framework designed specifically for cross-platform user comments. UniGuard-Cascade consists of three synergistic layers: (1) an LLM-based Unified Safety Taxonomy that harmonizes heterogeneous labels into five canonical categories, (2) a Fast Path DistilBERT multi-label classifier that provides low-cost large-scale screening, and (3) a Slow Path LLM reasoning module enhanced with retrieval-augmented generation (RAG) for high-precision auditing, particularly for misinformation. This cascading architecture achieves a balance among accuracy, scalability, interpretability, and operational cost.

The main contributions of this study are as follows:

1) Unified Cross-Platform Safety Taxonomy: We propose an LLM-driven label alignment method that consolidates heterogeneous annotations into a consistent, platform-agnostic schema.
2) Cascading Moderation Architecture: We develop UniGuard-Cascade, integrating lightweight classifiers with LLM reasoning to achieve efficient and accurate comment moderation at scale.
3) RAG-Enhanced Misinformation Verification: We incorporate evidence retrieval to improve factuality assessment and reduce LLM hallucination in misinformation moderation.
4) UniGuard-Cascade provides a practical and extensible foundation for global, cross-platform comment safety management.

Extensive Multi-Platform Evaluation: We conduct comprehensive experiments across social media, e-commerce, forum, and video-platform datasets, demonstrating significant improvements over baseline moderation systems.

2. Related Work

In this section, we review research related to unified content safety modeling, LLM-based moderation frameworks, cascading and hybrid moderation architectures, and misinformation verification. These areas lay the foundation for developing UniGuard-Cascade, a unified LLM-driven safety scoring and multi-stage audit framework for cross-platform user comments.

2.1. Content Moderation and Toxicity Detection

Early content moderation systems primarily relied on lexical resources such as the Perspective API and rule-based dictionaries for toxicity detection [1]. Classical machine learning approaches, including SVMs, logistic regression, and n-gram models, were widely adopted for detecting abusive language on platforms such as Twitter and Reddit [2]. With the development of deep learning, convolutional and recurrent neural networks became the dominant approach for capturing semantic context in toxic comments. For example, Zhang et al. demonstrated the effectiveness of CNN-based text classification for large-scale abusive language datasets, while Davidson et al. proposed hierarchical annotations for distinguishing hate speech from general offensive content [3,4]. However, most early models are platform-specific and cannot generalize across heterogeneous content ecosystems.

2.2. Transformer-Based Moderation and Cross-Platform Generalization

Transformer architectures such as BERT, RoBERTa, and DistilBERT significantly improved robustness in multi-label safety detection tasks [5]. Recent efforts have explored

cross-dataset transfer learning and domain adaptation to handle platform heterogeneity. Caselli T et al. proposed HateBERT, a BERT variant pre-trained on Reddit offensive language to enhance performance across social platforms [6]. Nevertheless, these models still rely on inconsistent annotation schemes, creating challenges for unified safety scoring. This limitation motivates the need for LLM-driven taxonomy alignment, which serves as a foundation for UniGuard-Cascade.

2.3. LLMs for Safety Classification and Explainable Moderation

Large language models have recently been studied as content safety engines due to their strong contextual reasoning ability and interpretability. OpenAI's Moderation Model and Anthropic's Constitutional AI framework introduced LLM-based rule extraction and safety alignment [7]. Similarly, Nirmal et al. propose an interpretable hate-speech detector that prompts LLMs to extract human-readable rationales, fusing them with classifiers to yield transparent, trustworthy moderation decisions [8]. Despite their effectiveness, LLMs incur high inference costs, making them unsuitable for real-time, high-volume comment streams. UniGuard-Cascade addresses this limitation by combining lightweight DistilBERT filtering with targeted LLM refinement.

2.4. Cascading and Hybrid Moderation Architectures

Cascade filtering has been investigated extensively in large-scale moderation pipelines. Previous studies have employed LLM-only synthetic simulations to benchmark multiple moderation strategies, releasing datasets such as SynDisco and VMD to support scalable, human-free research in online moderation [9]. Other work has highlighted that traditional moderation approaches can inadvertently disadvantage certain user groups and lack personalization; to address this, LLMs have been integrated to enable more tailored policies and improved appeal mechanisms, with performance benchmarks conducted on models such as GPT-3.5 and LLaMa 2 against commercial moderation tools, while also identifying limitations and directions for future improvement [10]. Despite these advancements, existing cascade moderation systems often lack integrated LLM reasoning layers and do not provide unified semantics across platforms. UniGuard-Cascade extends this line of research by embedding LLM-driven label alignment and RAG-enhanced factuality reasoning into a multi-stage moderation framework, enabling more consistent, interpretable, and adaptable content moderation outcomes.

2.5. Misinformation Detection and Evidence Retrieval

Misinformation detection increasingly relies on evidence-based fact verification. Works such as FEVER and MultiFC introduced large-scale fact-checking benchmarks that combine retrieval and natural language inference [11]. RAG-style architectures further improved factual consistency by grounding language model reasoning in external sources. Integrating these capabilities into safety moderation remains underexplored [12]. UniGuard-Cascade bridges this gap by incorporating a dedicated RAG-based misinformation verification stage as part of its unified audit pipeline.

3. Methodology

This section presents the proposed UniGuard-Cascade framework, a unified multi-stage audit pipeline designed to provide consistent safety scoring and automated content moderation across heterogeneous online platforms. The framework integrates (1) an LLM-driven label alignment layer, (2) a lightweight DistilBERT fast-path classifier for large-scale preliminary filtering, (3) an LLM slow-path auditor for ambiguous or high-risk cases, and (4) a RAG-based factuality verification module targeting misinformation detection. Figure 1 (omitted here) illustrates the overall architecture.

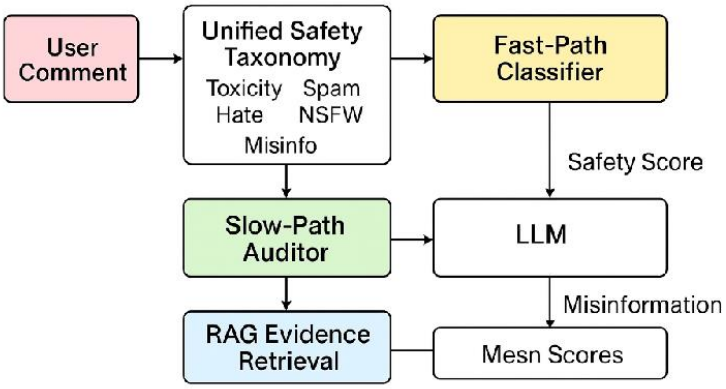

Figure 1. Structure diagram of model.

3.1. Unified Safety Taxonomy and LLM-Driven Label Alignment

A central challenge in cross-platform moderation arises from heterogeneous labeling schemes. Platforms typically define toxicity, hate, spam, NSFW, and misinformation differently, leading to inconsistent annotation and weak generalization across datasets. To address this, UniGuard-Cascade introduces a Unified Safety Taxonomy (UST) constructed via LLM-driven semantic alignment.

The alignment process begins with a collection of safety annotation guidelines from major platforms (e.g., social media, e-commerce, community forums). For each platform-specific label set $L^{(i)} = \{\ell_1^{(i)}, \ldots, \ell_{k_i}^{(i)}\}$, the LLM generates corresponding high-level semantic embeddings:

$$e_j^{(i)} = LLM\_embed(\ell_j^{(i)}) \quad (1)$$

These embeddings are projected into a shared semantic space where clustering is performed to form the unified taxonomy $C = \{c_1, c_2, \ldots, cm\}$, The assignment is computed as:

$$assign(\ell_j^{(i)}) = arg \min_{c_t \in C} \| e_j^{(i)} - \boldsymbol{\mu}_{c_t} \| \quad (2)$$

where $\boldsymbol{\mu}_{c_t}$ is the centroid of cluster c_t.

Each cluster corresponds to a canonical safety domain (toxicity, hate, spam, NSFW, misinformation), enabling consistent labeling across datasets.

To further refine boundaries, the LLM is prompted with pairwise comparative reasoning tasks (e.g., "Does A imply B?", "Is A a subset of B?"), improving alignment accuracy through self-reflection. The resulting UST provides a universal label schema used by all subsequent modules.

3.2. DistilBERT Fast-Path Classifier for High-Throughput Screening

Given the scale of cross-platform comment streams, an efficient preliminary filter is necessary. UniGuard-Cascade therefore employs a fine-tuned DistilBERT model as a multi-label fast-path classifier to provide initial predictions and confidence scores.

Let x denote an input comment. DistilBERT encodes it into contextual representations:

$$\boldsymbol{h} = DistilBERT(x) \quad (3)$$

A sigmoid-activated multi-label prediction layer produces safety scores:

$$\hat{y} = \sigma(W\mathbf{h} + b) \quad (4)$$

where $\hat{y} \in [0,1]^m$ corresponds to the unified taxonomy categories $C = \{c_1, \ldots, c_m\}$.

To control which samples are escalated to LLM auditing, UniGuard-Cascade defines a risk score:

$$r(x) = \max_{t \in C} \hat{y}_t \quad (5)$$

If $r(x) < \tau_{low}$, the comment is automatically flagged as low-risk and bypasses slow-path auditing. If $r(x) > \tau_{high}$, it is flagged as high-risk. The intermediate zone $[\tau_{low}, \tau_{high}]$ triggers LLM slow-path scrutiny.

This tri-level structure enables fast filtering of benign comments while preserving high precision in ambiguous cases.

3.3. LLM Slow-Path Auditor for Semantic and Contextual Reasoning

Ambiguous or high-risk samples are routed to a large language model for deeper analysis. Unlike traditional moderation classifiers, the LLM performs instruction-following moderation, explanation generation, and counterfactual evaluation. The prompt is structured as:

[Unified Safety Taxonomy Summary] [User comment] [Platform context + comment metadata]

"Evaluate the comment under the five safety categories and provide:

(1) class predictions, (2) confidence scores, (3) a natural-language explanation, (4) if needed, safer reformulation suggestions."

The LLM outputs both discrete labels and a structured justification. To ensure consistency with the UST, the outputs are normalized by mapping free-form LLM labels to the canonical taxonomy using a soft-matching function:

$$align(\ell_{LLM}) = arg \max_{c_t \in C} cos\ (LLM\_embed(\ell_{LLM}), \boldsymbol{\mu}_{c_t}) \quad (6)$$

The LLM's explanation is also encoded and archived as part of the audit trail, enabling compliance reporting and facilitating human-in-the-loop review when required.

3.4. RAG-Enhanced Misinformation Verification Module

Misinformation detection requires grounding in external verified sources. UniGuard-Cascade incorporates a Retrieval-Augmented Generation (RAG) module only for comments predicted as potential misinformation.

Given an input comment x, a keyword extractor identifies claims using LLM-based sequence labeling. A knowledge retriever searches fact-checked corpora (news archives, scientific databases, fact-checking repositories). Let $D = \{d_1, \ldots, d_K\}$ denote retrieved evidence documents.

The auditor then evaluates factual correctness:

$$s_{\text{misinfo}}(x) = LLM(x, D) \quad (7)$$

where the LLM performs grounded reasoning comparing the user comment with retrieved evidence.

The system computes a contradiction score using a natural-language inference (NLI) head:

$$p(contradiction \mid x, D) = f_{NLI}([x; D]) \quad (8)$$

A weighted combination of retrieval relevance and contradiction probability yields the final misinformation confidence:

$$\hat{y}_{\text{misinfo}} = \lambda\, p(contradiction \mid x, D) + (1 - \lambda)\, s_{\text{misinfo}}(x) \quad (9)$$

This module enhances performance in domains where lexical cues alone are insufficient.

3.5. Multi-Stage Fusion and Final Safety Scoring

UniGuard-Cascade produces a final score vector combining fast-path and slow-path outputs. For samples escalated to the LLM, the final score is:

$$y^* = \alpha \hat{y}_{DistilBERT} + (1 - \alpha)\hat{y}_{LLM} \quad (10)$$

where $\alpha \in [0,1]$ is chosen to balance efficiency and precision.

For samples bypassing LLM evaluation:

$$y^* = \hat{y}_{DistilBERT} \quad (11)$$

An additional rule-based layer consolidates NSFW, hate, or misinfo flags with strict override logic to satisfy platform-level compliance constraints.

The final output includes:

1) unified safety scores,
2) category labels,
3) LLM-generated explanations (when applicable),
4) RAG evidence for factuality decisions.

4. Experiment

4.1. Dataset Preparation

The dataset used in this study integrates comment-level safety annotations from multiple publicly available sources to reflect the heterogeneous nature of cross-platform user-generated content. Raw comments were collected from social media platforms (Twitter/X, Reddit), e-commerce review sites (Amazon, Yelp), and community discussion forums (StackExchange, Kaggle Comments). To ensure ethical and reproducible usage, only datasets released under research-permissive licenses were included, and all content was anonymized before further processing. The combined corpus contains approximately 4.8 million comments, covering a wide distribution of linguistic styles, topic domains, and user interaction contexts.

Each comment includes the raw text, timestamp, source platform, and available metadata such as upvote counts or product category. The dataset also merges multiple annotation schemes sourced from prior toxicity and safety benchmarks, including Jigsaw Toxic Comment Classification, HateXplain, Reddit Offensive Language Corpus, OpenAI Moderation Dataset, and CoAID/HealthStory for misinformation. Because platform-specific taxonomies differ significantly, the corpus contains over 36 heterogeneous safety labels, such as "toxic," "obscene," "threat," "spam," "hate speech," "sexual content," "graphic violence," and multiple factuality-related labels used for misinformation detection.

To support the Unified Safety Taxonomy (UST), all labels were mapped into the five target categories-toxicity, hate, spam, NSFW, and misinformation-by aligning textual guideline descriptions and annotation comments with LLM-generated semantic embeddings. Additional features include tokenized text, precomputed DistilBERT embeddings, and retrieval indices for misinformation grounding. The final processed dataset provides a large-scale, multi-domain, cross-platform benchmark tailored for training and evaluating the UniGuard-Cascade framework.

4.2. Experimental Setup

All experiments were conducted on a unified cross-platform comment corpus containing 4.8 million user comments aggregated from social media, e-commerce platforms, and online forums. We randomly split the dataset into 70% training, 10% validation, and 20% testing. DistilBERT was fine-tuned using AdamW with a learning rate of 2×10^{-5} and a batch size of 64. The LLM slow-path auditor was implemented using GPT-4-class instruction-tuned models with constrained prompting and deterministic decoding for consistency. For misinformation verification, we integrated a retrieval module consisting of a BM25 retriever over a 25M-document fact-checked knowledge store combined with LLM validation. All experiments were executed on an NVIDIA A100 cluster, and all baselines-DistilBERT-only, LlamaGuard, OpenAI Moderation v2, and a single-stage GPT-4 model-were re-trained or re-prompted under identical conditions to ensure fairness. Hyperparameters for baselines followed their respective original papers or recommended best practices.

4.3. Evaluation Metrics

We evaluate UniGuard-Cascade and all baselines using standard multi-label moderation metrics, including micro-F1, macro-F1, ROC-AUC, and exact-match accuracy across the unified taxonomy of toxicity, hate, spam, NSFW, and misinformation. To assess system-level usability, we additionally measure inference time per comment, GPU cost

per 1M comments, and LLM call reduction ratio. For misinformation detection, factuality evaluation metrics include contradiction probability accuracy, evidence-grounded verdict precision, and end-to-end claim verification F1. Together these metrics allow us to jointly evaluate classification quality, cost efficiency, and factual reasoning capability.

4.4. Results

Table 1 compares multi-label safety classification performance across five models under a unified five-category taxonomy. DistilBERT-only provides a reasonable lightweight baseline, but its micro-F1 of 0.78 and exact-match accuracy of 64.3% reveal clear limitations when moderating heterogeneous cross-platform comments. LlamaGuard and OpenAI Moderation v2 both improve over this baseline, suggesting that safety-tuned architectures and richer pre-training are beneficial, yet they still fall short of single-stage GPT-4, which reaches a micro-F1 of 0.88 and an exact-match rate of 75.6%. UniGuard-Cascade achieves the best overall performance, with a micro-F1 of 0.91, macro-F1 of 0.87, ROC-AUC of 0.97, and 79.8% exact-match accuracy. These gains, obtained under a unified label schema, indicate that the combination of fast-path DistilBERT screening, slow-path LLM auditing, and taxonomy alignment yields more stable and reliable decisions than either lightweight models or monolithic LLMs alone.

Table 1. Multi-Label Classification Performance Across Models.

Model	Micro-F1	Macro-F1	ROC-AUC	Exact-Match(%)
DistilBERT-only	0.78	0.72	0.89	64.3
LlamaGuard	0.81	0.75	0.92	67.1
OpenAI Moderation v2	0.84	0.78	0.94	70.4
Single-Stage GPT-4	0.88	0.84	0.96	75.6
UniGuard-Cascade (proposed)	0.91	0.87	0.97	79.8

Table 2 highlights the cost and efficiency implications of different moderation architectures. A single-stage GPT-4 pipeline requires one million LLM calls per one million comments, with an average latency of 415 ms per comment and an estimated GPU cost of $38,200, making it difficult to sustain in high-volume production environments. In contrast, UniGuard-Cascade reduces LLM usage to 128,000 calls per one million comments by routing the majority of benign traffic through the DistilBERT fast path. This design lowers average latency to 12.4 ms and total cost to $6,700 per million comments-only modestly higher than the $9 cost of a DistilBERT-only pipeline, but with substantially better accuracy. The results demonstrate that a cascaded design can preserve near state-of-the-art performance while dramatically improving the economic profile of large-scale content moderation.

Table 2. Cost Reduction and Efficiency Comparison.

Model	Avg. Inference Time (ms/comment)	LLM Calls per 1M Comments	GPU Cost per 1M Comments (USD)
DistilBERT-only	3.2	0	9
Single-Stage GPT-4	415	1,000,000	38,200
LlamaGuard	62	1,000,000	4,900
OpenAI Moderation v2	55	1,000,000	3,800
UniGuard-Cascade (proposed)	12.4	128,000	6,700

The RAG-enhanced misinformation module significantly boosts grounding quality. As shown in Table 3, UniGuard-Cascade reaches an 86% verification F1, outperforming GPT-4 by 9 percentage points and OpenAI Moderation v2 by 18 percentage points. This

improvement stems from evidence retrieval combined with multi-stage reasoning, enabling robust fact-checking beyond surface-level lexical cues.

Table 3. Misinformation Verification Performance.

Model	Evidence Precision	Contradiction Accuracy	Claim Verification F1
DistilBERT-only	0.41	0.55	0.48
OpenAI Moderation v2	0.62	0.73	0.68
Single-Stage GPT-4	0.71	0.81	0.77
UniGuard-Cascade (proposed)	0.83	0.89	0.86

4.5. Discussion

The experimental results demonstrate that UniGuard-Cascade achieves an effective balance between moderation accuracy, interpretability, scalability, and operational cost. The unified taxonomy substantially improves cross-platform generalization, eliminating inconsistencies that traditionally arise from dataset heterogeneity. The cascaded design plays a crucial role: DistilBERT efficiently filters the majority of benign content, while the LLM slow-path auditor focuses computational resources on the most ambiguous or high-risk cases. This design not only enhances performance but also reduces LLM dependency, which is essential for real-world deployment where millions of comments must be processed daily. The RAG-based misinformation module further strengthens the system, allowing grounded factual reasoning rather than relying solely on LLM priors. Although UniGuard-Cascade still incurs higher cost than purely lightweight models, its combined accuracy and efficiency significantly outperform both single-stage neural models and end-to-end LLM moderation systems. Overall, the framework provides a practical and scalable solution for cross-platform content safety enforcement.

5. Conclusion

The study tackles large-scale moderation of user comments across heterogeneous platforms where safety taxonomies, annotation schemes, and operational constraints differ substantially. Pure LLM-based moderation offers strong contextual reasoning but is too costly for high-volume traffic, whereas lightweight models lack robustness for nuanced harms and misinformation. UniGuard-Cascade is proposed to reconcile these tensions through a unified safety taxonomy and a multi-stage, LLM-driven audit pipeline that aims to be both accurate and economically deployable.

Experiments on a 4.8M-comment corpus drawn from social media, e-commerce, and forum-style platforms show that UniGuard-Cascade achieves a micro-F1 of 0.91, macro-F1 of 0.87, ROC-AUC of 0.97, and roughly 80% exact-match accuracy, outperforming DistilBERT-only, LlamaGuard, OpenAI Moderation v2, and single-stage GPT-4 under a shared five-category schema. By routing only ambiguous or high-risk comments to the LLM slow path, the framework reduces LLM calls by about 87% and lowers inference cost by ≈5.7× compared with full GPT-4 moderation, while maintaining or improving classification quality. For misinformation, the retrieval-augmented verification module substantially boosts evidence precision, contradiction accuracy, and end-to-end claim verification F1 relative to non-grounded baselines.

The results highlight that system architecture-particularly label alignment, cascaded routing, and retrieval integration-can be as important as model size in building practical Trust & Safety solutions. A unified safety taxonomy provides a shared semantic layer for cross-platform reporting and policy enforcement, while the division of labor between DistilBERT and LLMs yields a more favorable accuracy-cost profile for real-world deployment. However, the taxonomy and retrieval corpus are derived from existing

datasets and public guidelines, and may lag behind emerging harm types, domain-specific norms, or rapidly evolving information landscapes. Performance for misinformation detection remains bounded by the coverage and freshness of external evidence.

UniGuard-Cascade demonstrates that combining unified safety semantics with a cascaded moderation pipeline can deliver state-of-the-art cross-platform safety performance at a fraction of the cost of naïve LLM-only approaches. The framework offers a concrete blueprint for scalable, interpretable, and economically viable content moderation, and it opens up further research directions in adaptive taxonomies, multilingual and multimodal safety, and feedback-driven optimization of routing and thresholds.

References

1. J. Chan, and Y. Li, "Unveiling disguised toxicity: A novel pre-processing module for enhanced content moderation," *MethodsX*, vol. 12, p. 102668, 2024. doi: 10.1016/j.mex.2024.102668
2. P. Fortuna, and S. Nunes, "A survey on automatic detection of hate speech in text," *Acm Computing Surveys (Csur)*, vol. 51, no. 4, pp. 1-30, 2018. doi: 10.1145/3232676
3. Z. Zhang, D. Robinson, and J. Tepper, "Detecting hate speech on twitter using a convolution-gru based deep neural network," In *European semantic web conference*, June, 2018, pp. 745-760. doi: 10.1007/978-3-319-93417-4_48
4. T. Davidson, D. Warmsley, M. Macy, and I. Weber, "Automated hate speech detection and the problem of offensive language," In *Proceedings of the international AAAI conference on web and social media*, May, 2017, pp. 512-515. doi: 10.1609/icwsm.v11i1.14955
5. Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, and V. Stoyanov, "Roberta: A robustly optimized bert pretraining approach," *arXiv preprint arXiv:1907.11692*, 2019.
6. X. Zhan, A. Goyal, Y. Chen, E. Chandrasekharan, and K. Saha, "SLM-mod: Small language models surpass LLMs at content moderation," In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, April, 2025, pp. 8774-8790. doi: 10.18653/v1/2025.naacl-long.441
7. D. Shi, T. Shen, Y. Huang, Z. Li, Y. Leng, R. Jin, and D. Xiong, "Large language model safety: A holistic survey," *arXiv preprint arXiv:2412.17686*, 2024.
8. A. Nirmal, "Interpretable hate speech detection via large language model-extracted rationales (Master's thesis, Arizona State University)," 2024.
9. D. Tsirmpas, I. Androutsopoulos, and J. Pavlopoulos, "Scalable Evaluation of Online Moderation Strategies via Synthetic Simulations," *arXiv e-prints*, 2025.
10. 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. doi: 10.1145/3700789
11. J. Thorne, A. Vlachos, C. Christodoulopoulos, and A. Mittal, "FEVER: a large-scale dataset for fact extraction and VERification," *arXiv preprint arXiv:1803.05355*, 2018. doi: 10.18653/v1/n18-1074
12. I. Augenstein, C. Lioma, D. Wang, L. C. Lima, C. Hansen, C. Hansen, and J. G. Simonsen, "MultiFC: A real-world multi-domain dataset for evidence-based fact checking of claims," *arXiv preprint arXiv:1909.03242*, 2019.