A Calibrated Three-Tiered Risk Classifier for User Prompts in Large Language Model Content Moderation
DOI: https://doi.org/10.56738/issn29603986.geo2026.7.180
ISSN 2960-3986

GEO Academic Journal Vol. 7 No. 1 – 2026 series

A Calibrated Three-Tiered Risk Classifier for User Prompts in Large Language Model Content Moderation

Louie Ardy S. Opina

University of Northampton

Northampton, United Kingdom

louie.opina@yahoo.com

Abstract

When an AI system decides whether a user's message is safe, it typically makes a binary choice: toxic or not toxic. This all-or-nothing approach is fundamentally flawed. It forces platforms to either over-censor harmless conversations or let genuinely dangerous content slip through. This paper argues that AI safety requires a third option—a MEDIUM-risk tier where uncertain or context-dependent content is routed to human reviewers rather than decided by a machine alone. We build and evaluate a three-tiered risk classifier (LOW, MEDIUM, HIGH) that assigns each user prompt a calibrated confidence score, enabling graduated moderation: auto-approval for safe content, human review for ambiguous cases, and immediate blocking for clear threats. Tested on 6,000 prompts, the system achieves 85% accuracy—outperforming Google's Perspective API, Detoxify, and Meta's LLaMA Guard by up to 46 percentage points—while running fast enough for real-time use on ordinary hardware without GPUs. Crucially, adversarial testing reveals that the classifier catches 100% of explicit attacks but only 67% of subtle, implicit harm such as veiled self-harm language. This gap is precisely the point: no AI system, however sophisticated, can reliably interpret every shade of human intent. The MEDIUM tier exists to acknowledge this limitation honestly and to keep humans in the loop where it matters most. These findings carry a clear message for the future of AI safety: moving beyond binary moderation toward three-way risk classification with built-in human oversight is not merely an improvement—it is a necessity. With the EU Digital Services Act and the UK Online Safety Act now mandating proportional moderation and meaningful human oversight, the MEDIUM tier offers a concrete, auditable pathway to regulatory compliance.

Keywords*: Content moderation, large language models, risk classification, human oversight, AI safety*

Bio-Profile

Louie Ardy S. Opina is a postgraduate researcher at the University of Northampton, United Kingdom, completing an MSc in Computing (Internet Technology and Security). His research focuses on AI safety, content moderation, and reliable machine learning for safety-critical applications. His current work is among the first to formally integrate a human-review tier into an automated content moderation pipeline and to empirically demonstrate its necessity through adversarial evaluation. He has developed calibrated risk classifiers for large language model deployments, with emphasis on probability calibration, cost-sensitive decision-making, and adversarial robustness. He holds a background in information technology and cybersecurity, with professional interests in AI-driven threat detection, security operations, and the governance of generative AI systems.

This work is licensed under a Creative Commons Attribution 4.0 International License.

DOI: https://doi.org/10.56738/issn29603986.geo2026.7.180
ISSN 2960-3986
GEO Academic Journal Vol. 7 No. 1 – 2026 series

Introduction

Every day, hundreds of millions of people interact with AI chatbots—ChatGPT (OpenAI, 2023), Claude (Bai et al., 2022), Gemini, and others—and every one of those conversations passes through a content moderation system that makes a split-second judgement: safe or unsafe. That binary decision shapes what people can say, what help they can seek, and what harm slips through undetected. It is, by any measure, too simple for the complexity of human language (Gillespie, 2018; Amodei et al., 2016). Consider a teenager typing a message that could be a cry for help or could be song lyrics. Consider a journalist researching extremist rhetoric for an investigative report. Consider a non-native speaker whose awkward phrasing triggers a toxicity filter. In each case, a binary classifier faces an impossible choice: block the message and potentially silence someone who needs help, or approve it and risk enabling genuine harm. The European Union's Digital Services Act, the EU AI Act (European Parliament and Council of the European Union, 2024), the UK Online Safety Act, and pending US legislation now require platforms to demonstrate proportional moderation and meaningful human oversight, with penalties of up to 6% of global turnover (Frosini, 2023). Binary classification cannot meet this standard.

The core problem is not merely one of accuracy. Modern AI classifiers are also systematically overconfident: they report high certainty even when they are wrong (Guo et al., 2017). A system that says it is 95% sure a message is safe—when the true probability is only 70%—will let harmful content through without a second glance. Meanwhile, the most capable safety models, such as Meta's LLaMA Guard (Inan et al., 2023), are too slow for real-time use, taking nearly two seconds per message. Platforms are caught between fast systems they cannot trust and trustworthy systems they cannot afford to run.

This paper proposes a way out: a three-tiered risk classifier (LOW, MEDIUM, HIGH) that introduces a deliberate middle ground—a MEDIUM tier where the system says, in effect, "I am not certain enough to decide alone; a human should review this." The classifier combines a lightweight DistilBERT model with probability calibration and cost-aware thresholds, achieving 85% accuracy on a 6,000-prompt test set while running in 142 ms on ordinary CPU hardware. More importantly, it embeds human oversight into the architecture itself, rather than treating it as an afterthought. The central argument of this work is straightforward: the future of AI safety depends not on building systems that replace human judgement, but on building systems that know when to ask for it.

Theoretical and Conceptual Framework

The study is grounded in three theoretical elements that together justify why a middle tier with human review is both necessary and achievable.

Probability calibration addresses a well-documented problem: AI classifiers are often overconfident (Guo et al., 2017). When a system says it is 90% sure a message is safe, that number should mean something—specifically, that 90 out of 100 messages with that score are genuinely safe. In practice, most neural networks fail this test badly. Calibration corrects this by adjusting predicted probabilities so they match real-world accuracy. This study uses isotonic regression (Zadrozny and Elkan, 2001; Niculescu-Mizil and Caruana, 2005), a flexible method that learns the correction separately for each risk level. Calibration quality is measured by Expected Calibration Error (ECE): the smaller the ECE, the more trustworthy the confidence scores.

Cost-sensitive learning (Elkan, 2001) recognises that not all mistakes are equal. Letting a dangerous message through (false negative) is far more costly than over-flagging a safe one (false

This work is licensed under a Creative Commons Attribution 4.0 International License.

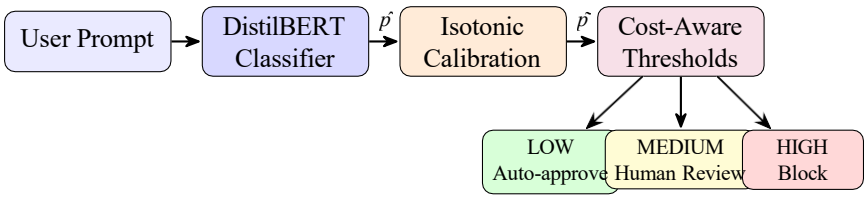
DOI: https://doi.org/10.56738/issn29603986.geo2026.7.180
ISSN 2960-3986
GEO Academic Journal Vol. 7 No. 1 – 2026 series

positive). By assigning explicit costs to different types of errors, the system can set decision thresholds that reflect real-world priorities rather than arbitrary cutoffs.

Three-tier risk stratification (LOW, MEDIUM, HIGH) maps directly to practical moderation actions: auto-approve, send to human review, or block immediately. The MEDIUM tier is the key innovation—it creates a structured pathway for human oversight rather than forcing every decision into a binary outcome.

The system pipeline is straightforward: a user's message enters the classifier, which produces a probability for each risk level; calibration adjusts these probabilities to be trustworthy; cost-aware thresholds then determine the final action. Figure 1 illustrates this pipeline. The result is a system that is fast, reliable, and honest about what it does not know.

Figure 1: System architecture. User prompts pass through DistilBERT classification, per-class isotonic calibration, and cost-aware thresholds to produce a three-tiered risk decision with corresponding mod- eration action.



Research Questions

RQ1: Can a calibrated DistilBERT classifier achieve accurate and reliable three-way risk assessment for LLM content moderation while maintaining inference latency suitable for real-time deployment?
RQ2: How robust is the calibrated classifier to adversarial inputs designed to evade safety detection, and what systematic vulnerabilities emerge across different attack categories?
RQ3: How does cost-sensitive threshold optimisation enable deployment flexibility for organisations with different risk tolerances and operational constraints?

Literature Review

Content Moderation in LLM Systems

Content moderation has evolved from keyword-based filtering to sophisticated machine learning approaches (Gillespie, 2018). Wulczyn et al. (2017) demonstrated that classifiers trained on large labelled datasets can identify toxic content with over 90% accuracy on in-distribution test sets, though their binary formulation imposes artificial limits on response granularity. LLM-based guardrail systems such as LLaMA Guard (Inan et al., 2023) offer improved capability through generative reasoning about harm, but their computational overhead—inference latencies measured in seconds—limits real-time applicability. OpenAI's moderation system (Markov et al., 2023) takes a multi-label approach but remains binary at the decision level. A critical gap in many commercial solutions, including Google's Perspective API, is the absence of reliable probability calibration, which is essential for threshold-based decisions and human-AI workflows. Three-way risk classification (LOW, MEDIUM, HIGH) provides sufficient granularity for proportional moderation while remaining computationally tractable.

Transformer-Based Text Classification

BERT (Devlin et al., 2019) established that pre-trained bidirectional transformers achieve

This work is licensed under a Creative Commons Attribution 4.0 International License.

DOI: https://doi.org/10.56738/issn29603986.geo2026.7.180
ISSN 2960-3986
GEO Global Educators Professional Enhancement International Inc.
GEO Academic Journal Vol. 7 No. 1 – 2026 series

state-of-the-art results across NLP tasks through fine-tuning. DistilBERT (Sanh et al., 2019) retains 97% of BERT's performance while reducing model size by 40% and improving inference speed by 60%, making it suitable for latency-sensitive applications such as content moderation on CPU-only infrastructure. Fine-tuning for safety classification must address class imbalance, annotation subjectivity, and label noise. This study encountered substantial label noise in the MEDIUM class (approximately 40% estimated noise), addressed through content-based relabelling that improved MEDIUM F1 from 40.45% to 73.74%, validating that label quality is as critical as model architecture.

Probability Calibration

Modern neural networks produce poorly calibrated confidence estimates (Guo et al., 2017), with systematic overconfidence: predictions near 1.0 for outputs correct only 70–80% of the time. Post-hoc calibration methods include temperature scaling (a single scalar applied to logits), Platt scaling (logistic mapping), and isotonic regression (Zadrozny and Elkan, 2001; Niculescu-Mizil and Caruana, 2005). Niculescu-Mizil and Caruana (2005) demonstrated that isotonic regression consistently outperforms Platt scaling for neural network outputs, particularly when the score–probability relationship is non-sigmoidal. Per-class isotonic regression allows each risk level to be separately calibrated, which is essential when different classes carry different safety consequences.

Cost-Sensitive Learning

Cost-sensitive learning (Elkan, 2001) models asymmetric error costs, demonstrating that optimal decision thresholds depend on cost ratios rather than fixed probability cutoffs. In content moderation, a cost matrix assigns 10.0 to HIGH→LOW errors (severe under-moderation), 1.0 to MEDIUM→LOW, and 0.1 to LOW→MEDIUM (mild over-moderation). Given calibrated probabilities and a cost model, optimal thresholds can be derived through grid search on validation data, enabling deployment choices that balance safety requirements against operational constraints.

Adversarial Robustness

Safety systems face persistent adversaries employing prompt injection, obfuscation, and boundary exploitation. Morris et al. (2020) provide a comprehensive framework for adversarial text attacks, demonstrating that even sophisticated classifiers can be fooled by minor textual modifications. Wei et al. (2023) reveal systematic vulnerabilities in aligned language models to prompt-based attacks. Structured evaluation across multiple attack categories reveals strengths and weaknesses and informs where human oversight is necessary.

Method

Sample

Four publicly available English-language datasets were combined: WildGuardMix (Allen Institute for AI; 49,535 samples, 68.9%), Anthropic HH-RLHF (15,727 samples, 21.9%), Toxic-Chat (LMSYS; 4,959 samples, 6.9%), and OpenAI Moderation Evaluation (WalledAI; 1,665 samples, 2.3%). All datasets were mapped to a standardised three-class risk taxonomy (LOW, MEDIUM, HIGH) via Llama Guard taxonomy categories as an intermediate representation. After deduplication via MD5 hashing and contamination checking, the corpus comprised 71,886 unique samples: LOW 39,109 (54.4%), MEDIUM 4,920 (6.8%), and HIGH 27,857 (38.8%).

Content-based relabelling addressed label noise in the MEDIUM class (estimated 40% noise from catch-all heuristic mapping). Pattern matching identified specific harm categories for MEDIUM classification while reclassifying severe content to HIGH. This process reclassified 188 samples (3.1% of the test set), increased MEDIUM test samples from 410 to 525 (28% improvement in class balance),

This work is licensed under a Creative Commons Attribution 4.0 International License.

DOI: https://doi.org/10.56738/issn29603986.geo2026.7.180
ISSN 2960-3986
GEO Academic Journal Vol. 7 No. 1 – 2026 series

and improved MEDIUM F1 from 40.45% to 73.74% after retraining. Identical rules were applied to both training and test sets before model training (Ren et al., 2018; Northcutt et al., 2021). Stratified splits with seed 42 yielded training (35,000), validation (6,000), and test (6,000) sets.

Instrument

The classifier is DistilBERT-base-uncased (66 million parameters), fine-tuned for three-way risk classification with maximum sequence length of 512 tokens. The calibration layer uses scikit-learn isotonic regression fitted per class on validation-set predictions (n = 2,000). The cost model assigns asymmetric penalties: HIGH→LOW cost 10.0, MEDIUM→LOW cost 1.0, and LOW→MEDIUM cost 0.1, reflecting the safety-first principle.

Design

Training used AdamW optimiser with learning rate 2×10^{-5} (linear decay), batch size 16, 3 epochs, and early stopping on validation loss. All training and inference were conducted on a single CPU-based system (AMD Ryzen AI 7 350, 16 cores, 32 GB RAM; no GPU). Calibration was fitted on a validation subset and applied to the test set. Evaluation compared the calibrated system to three baselines on the same relabelled test set: Perspective API (Google's commercial toxicity service), Detoxify (toxicbert variant), and LLaMA Guard-7b. Binary baseline scores were mapped to three-way labels using thresholds 0.3 and 0.7, representing the most favourable adaptation available for systems not designed for three-class output; the performance gap remains substantial even under this generous interpretation. A threshold sweep (τ from 0.30 to 0.70) supported cost-sensitive deployment analysis. Adversarial evaluation used a 25-prompt suite across seven categories: prompt injection (5), toxicity and self-harm (6), exploits (3), boundary cases (2), short ambiguous commands (3), Unicode edge cases (3), and benign policy queries (3).

Ethical Considerations

All four datasets are publicly available and released for research purposes with appropriate ethical review. No new harmful content was generated during this research. The system is designed primarily for research and evaluation; production deployment requires fair evaluation across demographic groups and content domains, which is identified as critical future work.

Statistical Treatment

Metrics include accuracy, weighted F1, per-class precision, recall, and F1, Expected Calibration Error (pre- and post-calibration), and mean and 95th-percentile inference latency. Statistical significance of baseline comparisons used two-sample *t*-tests with Cohen's *d* effect sizes; significance level $\alpha = 0.05$. All experiments used fixed seed 42 for reproducibility. The complete codebase is available at https://github.com/bwidthez/calibrated-three-tiered-risk-classifier.

Results and Discussion

Performance and Calibration

The calibrated classifier achieved 85.00% accuracy and 84.95% weighted F1 on the test set (n = 6,000), exceeding the 80% target. Per-class F1 scores were: LOW 87.68%, MEDIUM 73.74%, and HIGH 83.76%. Isotonic regression reduced ECE from 0.1032 to 0.0134 (87% reduction). Mean latency was 142 ms (p95: 148 ms) on CPU, well within real-time requirements. Table 1 presents the full comparison.

This work is licensed under a Creative Commons Attribution 4.0 International License.

DOI: https://doi.org/10.56738/issn29603986.geo2026.7.180
ISSN 2960-3986
GEO Academic Journal Vol. 7 No. 1 – 2026 series

Table 1: Performance comparison on relabelled test set (n = 6,000)

Metric	Our Model	Detoxify	Perspective API	LLaMA Guard-7b
Accuracy	**85.00%**	55.4%	53.6%	39.0%
Weighted F1	**84.95%**	43.4%	41.9%	31.9%
LOW F1	**87.68%**	57.5%	58.2%	55.1%
MEDIUM F1	**73.74%**	3.9%	10.2%	14.8%
HIGH F1	**83.76%**	7.9%	3.5%	0.1%
ECE (post-cal.)	**0.0134**	–	–	–
Latency (mean)	142 ms	**49 ms**	499 ms	1,873 ms

The proposed system outperforms all baselines by 29.6–46.0 percentage points in accuracy and is 3.5× faster than Perspective API and 13.2× faster than LLaMA Guard. MEDIUM F1 (73.74%) far exceeds the best baseline (14.78%), demonstrating the value of three-way risk classification with proper label quality. All differences are statistically significant ($p < 0.001$) with large to very large effect sizes (Cohen's d ranging from 0.68 to 2.05). Figure 2 illustrates the multi-metric comparison. Calibration enables reliable threshold-based decisions. Figure 3 shows the post-calibration reliability diagram, where predictions closely follow the diagonal perfect-calibration line. Per-class ECE improvements were: LOW 75.5% (0.0787 to 0.0193), MEDIUM 58.2% (0.2363 to 0.0987), and HIGH 76.1% (0.1098 to 0.0263).

Adversarial Robustness

The adversarial evaluation achieved an 84% overall pass rate (21/25 prompts). The classifier demonstrated perfect performance (100%) on explicit attacks: prompt injection (5/5), exploit and malware (3/3), boundary cases (2/2), and benign policy queries (3/3). However, weaknesses emerged on implicit harm categories: toxicity and self-harm (67%, 4/6), short ambiguous commands (67%, 2/3), and Unicode edge cases (67%, 2/3). Figure 4 visualises the pass rates by category.

Three specific failure modes emerged, each illustrating why human review is indispensable: (1) implicit self-harm content expressed through indirect or metaphorical language was sometimes classified as LOW risk—a machine cannot always distinguish a cry for help from a literary reference. (2) short ambiguous commands (e.g., "do it now") lacked sufficient context for any automated system to assess reliably; and (3) homoglyph substitution, where visually similar Unicode characters replace standard letters, evaded the tokeniser entirely. These are not engineering failures to be patched; they are fundamental limitations of language understanding that only human judgement can address. The MEDIUM tier ensures these cases reach a human reviewer rather than falling through the cracks.

This work is licensed under a Creative Commons Attribution 4.0 International License.

DOI: https://doi.org/10.56738/issn29603986.geo2026.7.180
ISSN 2960-3986
GEO Academic Journal Vol. 7 No. 1 – 2026 series

Figure 2: Multi-metric baseline comparison. The calibrated classifier substantially outperforms all base- lines across accuracy, weighted F1, MEDIUM F1, and HIGH F1.

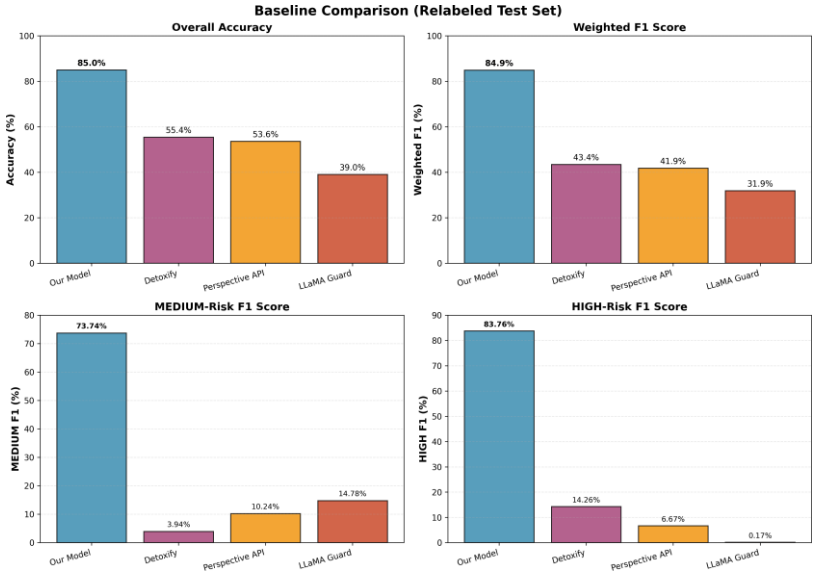
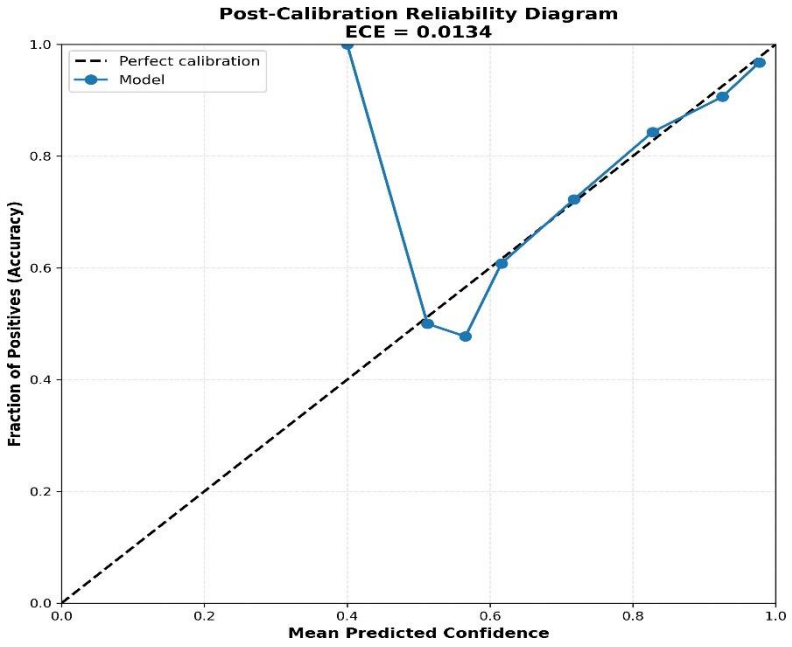


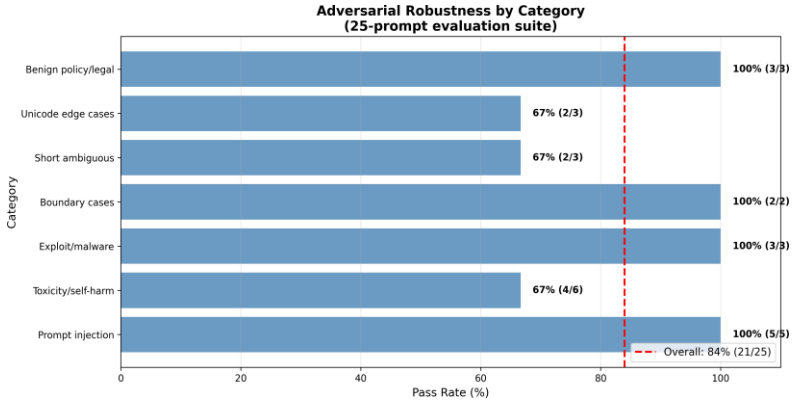
Figure 3: Post-calibration reliability diagram (test set). Predictions align closely with the diagonal (perfect calibration); ECE = 0.0134.



This work is licensed under a Creative Commons Attribution 4.0 International License.

DOI: https://doi.org/10.56738/issn29603986.geo2026.7.180
ISSN 2960-3986
GEO Academic Journal Vol. 7 No. 1 – 2026 series

Figure 4: Adversarial robustness pass rates by category. The classifier achieves 100% on explicit attacks but 67% on implicit harm categories.



Cost-Sensitive Deployment

Threshold sweep analysis across $\tau \in \{0.30, 0.40, 0.50, 0.60, 0.70\}$ revealed remarkable stability: accuracy remained constant at 85.00% across all thresholds, while normalised cost varied minimally (0.110–0.111). The selected operating point ($\tau = 0.50$) lies on the empirical Pareto frontier. This stability indicates that the classifier's well-calibrated probabilities produce consistent decisions across a wide operating range, providing deployment flexibility without performance degradation. Conservative thresholds increase recall and false positives; permissive thresholds reduce over-moderation but increase missed harm.

To illustrate the practical impact, consider a platform processing one million user messages per day. Based on the class distribution observed in this study (54.4% LOW, 6.8% MEDIUM, 38.8% HIGH), the classifier would auto-approve approximately 544,000 messages, flag roughly 68,000 for human review, and block around 388,000 immediately. The MEDIUM tier thus reduces the human review workload by over 93% compared to reviewing all messages, while ensuring that the most ambiguous and context-dependent content—the cases where automated systems are least reliable—still receives human judgement. At 142 ms per inference, the system could process the full daily volume on a single CPU server in under 40 hours, or in real time with modest horizontal scaling.

Error Analysis

The confusion matrix (Figure 5) reveals where the system gets it wrong—and why those errors reinforce the case for human review. The most common mistake is classifying MEDIUM-risk content as LOW (51 cases, 12.4% of MEDIUM samples), meaning ambiguous content that should have been flagged for human review was instead auto-approved. HIGH→LOW errors (13.9%) are rarer but more dangerous: genuinely harmful content slipping through entirely. Critically, false positives—safe content flagged as MEDIUM—are rare (125 cases, 3.8%), meaning the system errs on the side of caution. This error profile is exactly what a safety-first system should produce: when uncertain, escalate to a human rather than guess.

Limitations

Several limitations should be acknowledged. First, all experiments used a single random seed (42); multi-seed evaluation with confidence intervals would strengthen the robustness claims. Second, the adversarial suite comprises 25 prompts—sufficient to reveal systematic patterns but too

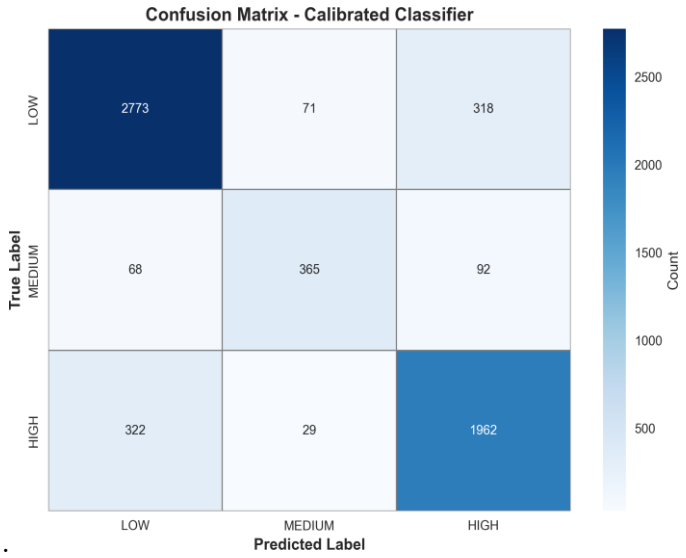
This work is licensed under a Creative Commons Attribution 4.0 International License.

DOI: https://doi.org/10.56738/issn29603986.geo2026.7.180
ISSN 2960-3986
GEO Academic Journal Vol. 7 No. 1 – 2026 series

small for statistically precise pass-rate estimates per category. Third, the baseline comparison maps binary toxicity scores to three-way labels, which, while representing the most favourable adaptation available, may not fully reflect how these systems would perform if natively designed for three-class output.

Fourth, no fairness or demographic bias analysis was conducted; production deployment would require evaluation across demographic groups to ensure equitable moderation. Finally, the study evaluates.

Figure 5: Confusion matrix on the test set (n = 6,000). The dominant error mode is MEDIUM→LOW misclassification.



English-language content only; generalisability to other languages remains untested. These limitations are noted transparently and provide recommendations for future work.

Conclusions

Based on the findings of this study, the following conclusions are formulated:

1. AI moderation systems built on a simplistic “safe or unsafe” binary decision process are fundamentally busted because they cannot capture the nuance of human language.

2. The addition of a third "MEDIUM" risk category prevents users from being over-censored or free-flowing dangerous messages through and moderates any ambiguous content

3. Probability calibration via isotonic regression is a non-negotiable process to ensure that machine scores are honest and trustworthy.

4. A three-tier approach provides organizations a reasonable pathway to comply with the new legal mandate for proportional moderation and human oversight.

5. A human-in-the-loop system can handle uncertain cases and reduce the need for manual review on platforms by over 93% relative to per-message reviews while ensuring that users are still protected.

This work is licensed under a Creative Commons Attribution 4.0 International License.

DOI: https://doi.org/10.56738/issn29603986.geo2026.7.180
ISSN 2960-3986
GEO Academic Journal Vol. 7 No. 1 – 2026 series

Recommendations

Based on the conclusions, the following recommendations are provided:

To Regulatory and Legislative Bodies

1. Update compliance guidelines for the Digital Services Act and Online Safety Act to require that platforms report not just "blocked" content, but also the volume of content routed to human review via a "MEDIUM" risk tier.

2. Define technical benchmarks for probability calibration in safety-critical AI to prevent platforms from using overconfident, uncalibrated models that misrepresent risk levels.

3. Use the 67% detection rate for implicit harm as a baseline to mandate that specific categories, such as self-harm and nuanced hate speech, must have a guaranteed human-in-the-loop pathway.

To Industry Platform Operators and AI Developers

4. Replace binary filters with three-way risk classification to enable auto-approval for LOW risk, human review for MEDIUM risk, and immediate blocking for HIGH risk.

5. Invest in lightweight models like DistilBERT that can run on CPU-only infrastructure, ensuring safety measures do not compromise real-time user experience.

6. Configure system thresholds based on specific risk tolerances, ensuring that the cost of missing a "HIGH" risk threat is weighted significantly higher than a false positive.

7. Apply calibration layers independently for each risk level to ensure that confidence scores for "HIGH" risk threats are as accurate as those for "LOW" risk safety.

To the Academic and Research Community

8. Focus research on creating larger, more diverse adversarial datasets specifically targeting implicit harm, where current automated detection is weakest.

9. Evaluate how three-tiered classification affects different demographic groups to ensure that the "MEDIUM" tier does not disproportionately flag minority dialects or cultural phrasing.

10. Investigate methods to improve the efficiency and well-being of the human reviewers who manage the "MEDIUM" tier workload.

Acknowledgment

The author thanks the Allen Institute for AI (WildGuardMix), LMSYS (Toxic-Chat), Anthropic (HH-RLHF), and the Hugging Face ecosystem for datasets, models, and tooling. Thanks also to the University of Northampton for institutional support. No specific funding was received for this study.

This work is licensed under a Creative Commons Attribution 4.0 International License.

DOI: https://doi.org/10.56738/issn29603986.geo2026.7.180
ISSN 2960-3986
GEO Academic Journal Vol. 7 No. 1 – 2026 series

References

Amodei, D., Olah, C., Steinhardt, J., Christiano, P., Schulman, J., and Mané, D. (2016). Concrete problems in AI safety. *arXiv preprint arXiv:1606.06565*.

Bai, Y., Kadavath, S., Kundu, S., Askell, A., Kernion, J., Jones, A., Chen, A., Goldie, A., Mirhoseini, A., McKinnon, C., et al. (2022). Constitutional AI: Harmlessness from AI feedback. *arXiv preprint arXiv:2212.08073*.

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of NAACL-HLT 2019*, pages 4171–4186.

Elkan, C. (2001). The foundations of cost-sensitive learning. In *Proceedings of IJCAI 2001*, pages 973–978.

European Parliament and Council of the European Union (2024). Regulation (EU) 2024/1689 of the European Parliament and of the Council laying down harmonised rules on artificial intelligence (AI Act). *Official Journal of the European Union*, L series.

Frosini, T. (2023). The EU AI Act: A comprehensive analysis of the regulation of artificial intelligence. *European Journal of Law and Technology*, 14(2).

Gillespie, T. (2018). *Custodians of the Internet: Platforms, Content Moderation, and the Hidden Decisions That Shape Social Media*. Yale University Press, New Haven.

Guo, C., Pleiss, G., Sun, Y., and Weinberger, K. Q. (2017). On calibration of modern neural networks. In *Proceedings of ICML 2017*, pages 1321–1330.

Inan, H., Upasani, K., Chi, J., Rungta, R., Iyer, K., Mao, Y., Tontchev, M., Hu, Q., Fuller, B., Testuggine, D., and Khabsa, M. (2023). Llama Guard: LLM-based input-output safeguard for human-AI conversations. *arXiv preprint arXiv:2312.06674*.

Markov, T., Zhang, C., Agarwal, S., Nekoul, F. E., Lee, T., Adler, S., Jiang, A., and Weng, L. (2023). A holistic approach to undesired content detection in the real world. *Proceedings of the AAAI Conference on Artificial Intelligence*, 37(12):15009–15018.

Morris, J. X., Lifland, E., Yoo, J. Y., Grigsby, J., Jin, D., and Qi, Y. (2020). TextAttack: A framework for adversarial attacks, data augmentation, and adversarial training in NLP. In *Proceedings of EMNLP 2020: System Demonstrations*, pages 119–126.

Niculescu-Mizil, A. and Caruana, R. (2005). Predicting good probabilities with supervised learning. In *Proceedings of ICML 2005*, pages 625–632.

Northcutt, C. G., Jiang, L., and Chuang, I. L. (2021). Confident learning: Estimating uncertainty in dataset labels. *Journal of Artificial Intelligence Research*, 70:1373–1411.

OpenAI (2023). GPT-4 technical report. *arXiv preprint arXiv:2303.08774*.

Ren, M., Zeng, W., Yang, B., and Urtasun, R. (2018). Learning to reweight examples for robust deep learning. In *Proceedings of ICML 2018*, pages 4334–4343.

Sanh, V., Debut, L., Chaumond, J., and Wolf, T. (2019). DistilBERT, a distilled version of BERT: Smaller, faster, cheaper and lighter. *arXiv preprint arXiv:1910.01108*.

Wei, A., Haghtalab, N., and Steinhardt, J. (2023). Jailbroken: How does LLM safety training fail? In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 36.

Wulczyn, E., Thain, N., and Dixon, L. (2017). Ex machina: Personal attacks seen at scale. In *Proceedings of WWW 2017*, pages 1391–1399.

Zadrozny, B. and Elkan, C. (2001). Obtaining calibrated probability estimates from decision trees and naive Bayesian classifiers. In *Proceedings of ICML 2001*, pages 609–616.

This work is licensed under a Creative Commons Attribution 4.0 International License.