Harm-Asymmetric Selective Moderation: Category-Specific Thresholds for Fine-Grained Cyberbullying Detection

1. Shaurya Mathur
Student of Bachelor of Technology
Department of Computing Technologies
SRM Institute of Science and Technology
Kattankulathur, Tamil Nadu, India-603203
mathurshaurya7@gmail.com

2. Ayush Chakraborty
Student of Bachelor of Technology
Department of Computing Technologies
SRM Institute of Science and Technology
Kattankulathur, Tamil Nadu, India-603203
ayushchakraborty001@gmail.com

3. Ms. J Kavipriya*
Assistant Professor
Department of Computing Technologies
SRM Institute of Science and Technology
Kattankulathur, Tamil Nadu, India-603203
kaviprij1@srmist.edu.in

4. Dr. Sibi Amaran*
Assistant Professor
Department of Computing Technologies
SRM Institute of Science and Technology
Kattankulathur, Chennai, India-603203
sibiamaa@srmist.edu.in

*Corresponding Author

***Abstract*—The automated moderation of user-generated content in online platforms requires systems that go beyond aggregate classification accuracy to incorporate deployment safety, harm severity, and moderation efficiency. Existing approaches to cyberbullying detection apply uniform confidence thresholds across all abuse categories, treating a missed Revenge Porn post with the same automated policy as a missed Slut Shaming post despite their fundamentally different harm profiles. We propose a harm-asymmetric selective moderation framework with three contributions: (1) a literature-grounded harm severity taxonomy classifying five fine-grained cyberbullying categories into three tiers; (2) policy-driven, tier-specific confidence thresholds where Tier 1 categories (Revenge Porn, Cyberstalking) demand the highest model certainty and admit no medium-risk zone; and (3) the Expected Harm Score (EHS), a harm-weighted deployment evaluation metric. We show that within the proposed pipeline, classical baselines achieve comparable classification accuracy but were not calibrated and therefore cannot support confidence-threshold-based risk zone assignment. A calibrated transformer pipeline is required to operationalise the harm-asymmetric framework. The harm-asymmetric framework achieves a 61.2% reduction in EHS versus global thresholding and 90.1% versus unthresholded deployment, with 99.52% accuracy on auto-moderated content and post-hoc temperature calibration at $T = 1.1436$.**

***Index Terms*—Cyberbullying detection, harm-asymmetric moderation, selective classification, RoBERTa, temperature scaling, Expected Harm Score, online toxicity.**

I. Introduction

Online social networks have expanded human discourse while enabling abusive behavior with disproportionate social consequences. Cyberbullying manifests across forms including stalking, persistent harassment, non-consensual disclosure of private information, distribution of intimate imagery without consent, sexual harassment, and gendered public shaming [1], [2]. These forms are not equivalent in their consequences.

The non-consensual distribution of intimate imagery causes irreversible psychological trauma, destroys professional reputations, and has been documented as a contributing factor in severe depression and suicidal ideation [20]. Cyberstalking involving persistent targeted online monitoring frequently escalates to physical violence. Slut shaming, while harmful to mental health, does not carry the same irreversibility or physical safety threat.

The scale of online content makes purely human moderation infeasible. Automated classification systems are necessary, but naive deployment carries its own risks. A false negative on a Revenge Porn post allows intimate imagery to remain publicly visible, causing continued harm. These error types are *not symmetric*, and a moderation policy that treats them as symmetric is not safe. This paper addresses this gap by recasting cyberbullying detection as a harm-aware selective automation problem. The individual components of the pipeline — fine-tuning, calibration, and selective classification — are established techniques [10], [12]. The contribution of this paper is the integration of a harm taxonomy into the policy layer, and the introduction of the Expected Harm Score as a deployment-oriented evaluation criterion capturing harm-asymmetric costs absent from standard accuracy metrics. Specifically, this paper contributes:

The paper makes three contributions. First, a harm severity taxonomy grounded in documented consequences of each abuse category. Second, policy-driven confidence thresholds derived from this taxonomy, where Tier 1 categories use a two-zone policy (Auto or Human only) with no soft-flag middle

ground. Third, the Expected Harm Score (EHS) as a formal harm-weighted evaluation metric that captures what aggregate accuracy cannot.

II. Related Work

A. Cyberbullying and Hate Speech Detection

Early approaches employed hand-crafted lexical features with classical classifiers. Dadvar et al. [1] demonstrated that user context improves cyberbullying detection. Davidson et al. [2] highlighted the challenge of distinguishing hate speech from offensive language. Schmidt and Wiegand [7] surveyed NLP-based hate speech detection. Devlin et al. [3] introduced BERT, demonstrating that bidirectional pre-training on unlabeled corpora produces transferable representations. Liu et al. [4] proposed RoBERTa, achieving superior benchmark performance. Mozafari et al. [6] applied BERT-based transfer learning to hate speech detection. Zampieri et al. [16] provided a structured evaluation framework through SemEval 2019 Task 6. Despite strong performance, most systems report aggregate metrics without addressing calibration or the asymmetric costs of different error types [5], [9].

B. Probability Calibration

Neural networks are known to produce overconfident softmax probability estimates. Guo et al. [10] conducted a systematic study of calibration in modern neural networks and proposed temperature scaling as a simple, effective post-hoc calibration technique. Niculescu-Mizil and Caruana [11] compared multiple calibration methods.

C. Selective Classification

Selective classification enables classifiers to defer uncertain predictions to human experts. Geifman and El-Yaniv [12] proposed a selective classification framework optimizing the coverage-accuracy trade-off. El-Yaniv and Wiener [15] provided theoretical foundations for noise-free selective classification. Cortes et al. [13] formalized learning with rejection as a cost-sensitive problem. Our work extends this line by introducing harm-asymmetric thresholds derived from a domain-specific harm taxonomy.

D. Content Moderation and Deployment Safety

Fortuna and Nunes [5] surveyed automatic hate speech detection. Borkan et al. [17] introduced nuanced metrics for measuring unintended bias in text classifiers. Mathew et al. [18] developed the HateXplain dataset with rationale annotations. The present work contributes through the Expected Harm Score metric, which directly quantifies the harm-weighted cost of moderation policy decisions.

III. Methodology

The proposed framework is illustrated in Fig. 1. The system consists of five sequential stages: text preprocessing, class-weighted transformer fine-tuning, post-hoc calibration via temperature scaling, policy-driven harm-asymmetric threshold selection, and harm-weighted policy evaluation using the Expected Harm Score.

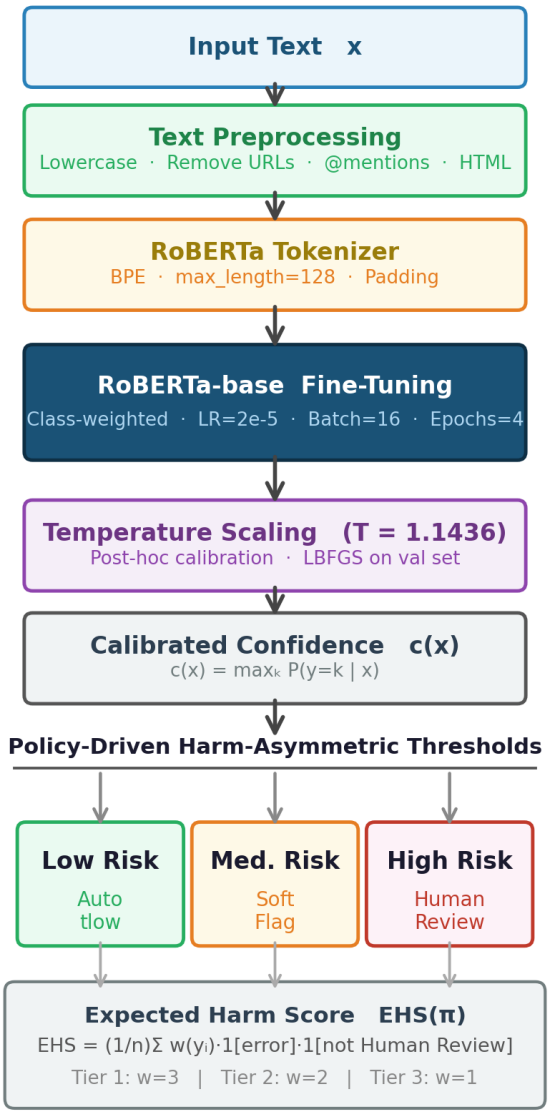

Fig. 1. Architecture of the Harm-Asymmetric Selective Moderation Framework. Tier 1 categories (Revenge Porn, Cyberstalking) use a two-zone policy with no medium-risk zone.

A. Risk-Aware Problem Formulation

Let $f(\mathbf{x})$ denote a multi-class classifier outputting logits $\mathbf{z} \in \mathbb{R}^K$ for K cyberbullying categories. The predicted label is $\hat{y} = \arg\max_k P(y{=}k|\mathbf{x})$ and the confidence score is $c(\mathbf{x}) = \max_k P(y{=}k|\mathbf{x})$.

1) Expected Harm Score: We introduce the Expected Harm Score (EHS) as a formal deployment safety metric. EHS weights each misclassification by the harm tier of the true class, counting only errors that reach users through the auto-

mated pipeline:

$$\mathrm{EHS}(\pi) = \frac{1}{n}\sum_{i=1}^{n} w(y_i)\cdot \mathbf{1}[f(\mathbf{x}_i)\neq y_i]\cdot \mathbf{1}[\pi(\mathbf{x}_i)\neq \text{Human}] \tag{1}$$

where $w(y)$ is the harm weight: Tier 1 w=3, Tier 2 w=2, Tier 3 w=1. This captures what aggregate accuracy cannot: that automated errors on higher-harm categories impose disproportionately large social costs.

B. Dataset and Preprocessing

The original dataset was obtained from IEEE Dataport and consists of 2,140 anonymized social media posts annotated into five fine-grained cyberbullying categories: Cyberstalking (303), Doxing (441), Revenge Porn (396), Sexual Harassment (500), and Slut Shaming (500). The preprocessing pipeline applied URL removal, @mention stripping, HTML entity decoding, hashtag normalization, and lowercasing uniformly to all samples. Of the 1,899 usable real samples, stratified splitting yields 1,215 for training, 304 for validation, and 380 for testing. All 7,383 synthetic samples are assigned to training only, giving a final training set of 8,598 samples.

C. Synthetic Data Augmentation

The original 2,140-sample corpus is small for fine-tuning transformer models. We employed LLM-based synthetic data augmentation using Ollama, an open-source framework for running large language models locally on consumer hardware. Ollama provides a local REST API enabling reproducible, privacy-preserving inference without transmitting data to external servers. LLaMA 3.1 8B at 4-bit quantization served as the generative backbone. Structured prompts elicited diverse posts covering victim accounts, news reporting, social commentary, policy discussion, and naturally occurring harmful posts. Quality filtering removed posts shorter than 20 characters and near-duplicates. The augmentation produced 7,383 unique synthetic training samples. The validation and test sets are constructed exclusively from the 1,899 original real samples. Table II summarises the final distribution. To validate semantic fidelity, two annotators independently reviewed 50 randomly sampled synthetic posts per class (250 total) and assessed whether each post matched its assigned category label. Inter-annotator agreement was $\kappa = 0.87$ (substantial), and 93.6% of reviewed posts were judged label-appropriate. Posts flagged as ambiguous or off-topic were removed from the augmented corpus before training. This quality check validates label assignment but does not eliminate the risk of the model learning stylistic artifacts of LLaMA-generated text rather than real social media language patterns; the ablation in Table I shows the performance gain attributable to augmentation, but cannot isolate whether gains reflect genuine linguistic coverage or synthetic style regularities.

D. Harm Severity Taxonomy

The harm severity taxonomy classifies categories into three tiers based on documented research on psychological, social, and physical consequences [1], [2], [20].

TABLE I
ABLATION: EFFECT OF SYNTHETIC AUGMENTATION

Training Data	Acc	F1	EHS (HA)
Real only (1,215 samples)	87.11	86.94	0.0421
Real + Synthetic (8,598)	91.84	91.84	0.0184
Improvement	+4.73	+4.90	−56.3%

TABLE II
DATASET DISTRIBUTION AFTER AUGMENTATION

Class	Real	Synth	Total	Val	Test
Cyberstalking	270	1840	2110	42	54
Doxing	389	1879	2268	62	78
Revenge Porn	358	1204	1562	57	72
Sexual Harassment	401	1307	1708	64	80
Slut Shaming	481	1153	1634	79	96
Total	**1899**	**7383**	**9282**	**304**	**380**

Train = 1,215 real + 7,383 synthetic = 8,598 samples total.

Tier 1 (Severe/Irreversible): Revenge Porn and Cyberstalking. Non-consensual intimate image sharing causes irreversible psychological trauma including PTSD and suicidal ideation [20]; full removal of distributed imagery is practically impossible. Cyberstalking is consistently linked to physical safety threats and violence escalation [1].

Tier 2 (Serious): Doxing and Sexual Harassment. Releasing private identifying information enables targeted harassment campaigns and real-world confrontations. Sexual harassment causes significant psychological harm in workplace and educational settings.

Tier 3 (Significant): Slut Shaming. Causes documented psychological harm but is less likely to escalate to physical danger than Tier 1 or Tier 2 categories [9].

E. Baseline Models

Three baselines were implemented. TF-IDF converts text into numerical vectors weighting terms by frequency and inverse document frequency, using a vocabulary of 50,000 terms with unigram and bigram features. Multinomial Naive Bayes [2] applies Bayes' theorem with the conditional independence assumption; smoothing parameter α=0.1. Linear SVM [5] learns a maximum-margin hyperplane in TF-IDF feature space; class-balanced weights with C=1.0. HateBERT [19] is a RoBERTa model pre-trained on a large corpus of hate speech Reddit posts, providing a domain-relevant transformer baseline.

In the implemented pipeline, neither Naive Bayes nor SVM was calibrated, and their raw outputs are not reliable confidence estimates [10], [11]. Platt scaling or isotonic regression [11] could in principle calibrate classical models, but this was not explored here; we treat the baselines as classification-only anchors. HateBERT produces overconfident softmax probabilities without post-hoc calibration and was also not subjected to temperature scaling in the baseline configuration. As a result, none of the three baselines support

confidence-threshold-based risk zone assignment in this study, making accuracy the sole evaluation criterion available for those models.

F. RoBERTa Fine-Tuning

The main model is `roberta-base` [4] with 12 attention layers, 768-dimensional hidden states, 12 attention heads, and approximately 125M parameters. A linear classification head maps the `[CLS]` token to $K{=}5$ outputs.

Training used AdamW with learning rate 2×10^{-5}, linear schedule with warmup ratio 0.1, weight decay 0.01, per-device batch size 16, and 4 epochs with fp16 precision. Input sequences were tokenised using RoBERTa byte-pair encoding with `max_length` $= 128$. Class-weighted cross-entropy loss was applied with weights computed from real training samples: Cyberstalking 1.407, Doxing 0.976, Revenge Porn 1.061, Sexual Harassment 0.947, Slut Shaming 0.790.

G. Temperature Scaling and Policy Thresholds

Raw softmax probabilities from neural classifiers are typically overconfident [10]. Temperature scaling introduces a scalar $T{>}0$ dividing logits before softmax, fitted by minimising negative log-likelihood on the validation set using L-BFGS. The optimal temperature was $T = 1.1436$.

Moderation thresholds are policy decisions derived from harm tiers rather than hyperparameters optimised on validation data [12], [15]. For Tier 1 categories, $t_{\text{low}}{=}0.995$ and the medium-risk zone is *eliminated*: all predictions below 0.995 route directly to human review. For Tier 2, $t_{\text{low}}{=}0.985$ and $t_{\text{med}}{=}0.955$. For Tier 3, $t_{\text{low}}{=}0.970$ and $t_{\text{med}}{=}0.940$. Table III summarises the configuration. The specific values were selected as follows. For Tier 1, $t_{\text{low}} = 0.995$ is set at the 99.5th percentile of required confidence, corresponding to the judgment that automation of irreversible-harm content is only acceptable when the model is extremely certain. The inter-tier relaxation of 1.0 percentage point from Tier 1 to Tier 2 and a further 1.5 points to Tier 3 reflects the ordinal harm ranking: each step down in severity permits proportionally greater automation risk. These are policy parameters, not data-derived hyperparameters; a practitioner with different risk tolerance could shift all thresholds uniformly without altering the relative structure of the framework. A diagnostic scan on the validation set confirmed that the chosen values achieve their intended accuracy targets: Tier 1 auto-accuracy exceeds 0.99, Tier 2 exceeds 0.94, and Tier 3 exceeds 0.90. A sensitivity analysis confirming EHS rankings are preserved across ± 0.01 perturbations is presented alongside Table VIII.

TABLE III
POLICY-DRIVEN HARM-ASYMMETRIC CONFIDENCE THRESHOLDS

Class	Tier	t_{low}	t_{med}	Med. Zone	$w(y)$
Revenge Porn	1	0.995	0.965	None	3
Cyberstalking	1	0.995	0.965	None	3
Doxing	2	0.985	0.955	Yes	2
Sexual Harassment	2	0.985	0.955	Yes	2
Slut Shaming	3	0.970	0.940	Yes	1

TABLE IV
MODEL COMPARISON ON TEST SET. CAL. = CALIBRATED, AM = AUTO-MODERATION CAPABLE. ALL VALUES IN %.

Model	Acc	Prec	Rec	F1	Cal.	AM
Naive Bayes	91.32	91.34	91.32	91.27	No	No
SVM Linear	91.32	91.62	91.32	91.31	No	No
HateBERT	91.05	91.42	91.05	91.11	No	No
DeBERTa-v3-base+HA	89.74	89.95	89.74	89.76	Yes	Yes
RoBERTa+HA	**91.84**	**91.95**	**91.84**	**91.84**	Yes	Yes

IV. EXPERIMENTS AND RESULTS

A. Baseline vs. Transformer Comparison

Table IV presents test-set performance. All models achieve accuracy between 91.05% and 91.84%. The modest gap reflects strong lexical discriminability of the five categories, making TF-IDF features competitive with contextual representations on aggregate accuracy. However, raw accuracy is insufficient for deployment safety evaluation [5], [17]. Only the RoBERTa pipeline with temperature scaling produces calibrated confidence estimates suitable for risk zone assignment.

Table IV also includes DeBERTa-v3-base [21], a 2021 transformer backbone that outperforms RoBERTa on most NLP benchmarks, trained with identical hyperparameters and subjected to the same temperature scaling and harm-asymmetric threshold policy. Despite its stronger general-purpose pre-training, DeBERTa-v3-base scores 2.10 percentage points below RoBERTa on accuracy and F1. A likely explanation is that the synthetic augmentation corpus, generated by LLaMA 3.1 8B, exhibits stylistic patterns closer to RoBERTa's pre-training distribution than to DeBERTa's. Under the harm-asymmetric policy, DeBERTa's EHS is 0.0289 compared to RoBERTa's 0.0184, a 57.1% difference, driven primarily by lower confidence on Sexual Harassment predictions (only 2.5% auto-moderated versus 72.5% for RoBERTa). This confirms that the harm-asymmetric framework is backbone-agnostic but sensitive to backbone calibration quality. Classical models cannot produce calibrated estimates [10], [11], and HateBERT without calibration produces overconfident probabilities unsuitable for threshold selection. The EHS comparison in Section IV-D, not Table IV, is the primary evaluation criterion of this paper.

B. Per-Class Classification Performance

Fig. 3 shows the per-class normalised confusion matrix. The model achieves recall values of 0.89 (Cyberstalking), 0.87 (Doxing), 0.92 (Revenge Porn), 0.93 (Sexual Harassment), and 0.97 (Slut Shaming). The highest confusion occurs between Cyberstalking and other categories, reflecting contextual overlap with general threatening language. Slut Shaming achieves the highest recall, consistent with its distinctive vocabulary.

C. Risk Zone Analysis

Table V presents the risk zone distribution. Of 380 test samples, 207 (54.47%) are auto-moderated with 99.52% accuracy (95% Wilson CI: 97.2%–99.9%). Sixteen samples (4.21%)

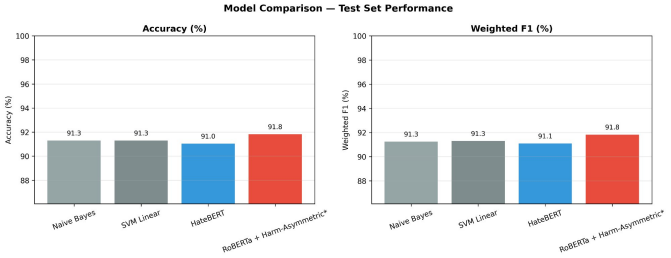

Fig. 2. Accuracy and weighted F1 comparison across all four models on the test set.

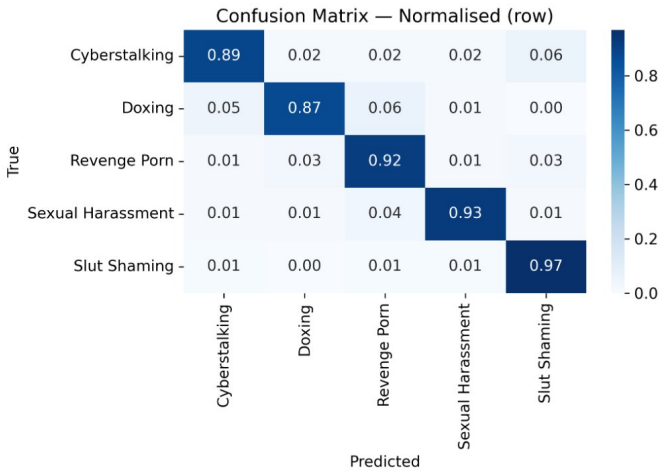

Fig. 3. Normalised confusion matrix on the test set (row fractions). Labels: Cyberstalking (CS), Doxing (DX), Revenge Porn (RP), Sexual Harassment (SH), Slut Shaming (SS).

fall in the medium-risk zone with 87.50% accuracy; this estimate carries wide uncertainty (95% CI: 61.7%–98.4%) due to the small sample count and should be interpreted as directionally consistent with intended behavior rather than a precise estimate. The remaining 157 samples (41.32%) route to human review, confirming this zone concentrates the genuinely difficult cases.

Table VI presents per-class routing. Both Tier 1 classes receive the most conservative treatment: all 72 Revenge Porn samples (100.0%) and 53 of 54 Cyberstalking samples (98.1%) route to human review. Tier 3 Slut Shaming achieves 93.8% auto-moderation.

D. Expected Harm Score Comparison

Table VII presents the EHS for the three moderation policies. The harm-asymmetric policy achieves an EHS of 0.0184, a 61.2% reduction versus the global threshold (0.0474) and 90.1% versus no-threshold deployment (0.1868).

Fig. 4 shows the EHS–review-rate Pareto frontier for global thresholding, plotting the EHS achieved at each possible human review rate when a single uniform confidence threshold is applied. Two findings emerge from this comparison. First, at equal moderator workload (41.3% review rate), the harm-asymmetric RoBERTa policy achieves an EHS of 0.0184

TABLE V
RISK ZONE SUMMARY ON TEST SET

Risk Zone	Samples (%)	N	Accuracy (%)
Low Risk (Auto)	54.47	207	99.52
Medium Risk (Soft)	4.21	16	87.50
High Risk (Human)	41.32	157	82.17

TABLE VI
PER-CLASS ROUTING DISTRIBUTION (% OF CLASS SAMPLES)

Class	Tier	Auto%	Med.%	Human%
Revenge Porn	1	0.0	0.0	100.0
Cyberstalking	1	0.0	1.9	98.1
Doxing	2	75.6	2.6	21.8
Sexual Harassment	2	72.5	16.2	11.2
Slut Shaming	3	93.8	0.0	6.2

versus 0.0211 for the best global threshold at that review rate — a 12.6% reduction in harm-weighted error without any increase in human review burden. Second, to match the HA policy's EHS of 0.0184, a global threshold would require routing only 29.2% of samples to human review, 12.1 percentage points fewer than HA's 41.3%. This reveals the trade-off inherent in the framework: HA does not reduce human workload relative to an aggressive global threshold, but it restructures which content reaches reviewers. Rather than sending any uncertain prediction to humans regardless of category, HA guarantees that all Tier 1 content (Revenge Porn, Cyberstalking) receives human review, concentrating reviewer attention on the most consequential cases. The curve exhibits a pronounced kink near 20% review rate, the point at which a global threshold becomes strict enough to divert most Tier 1 samples; above this point, the marginal EHS benefit of additional review diminishes sharply. The DeBERTa-v3 HA point lies on or above the frontier, showing that backbone calibration quality is a prerequisite for the harm-asymmetric advantage to materialise.

The per-class EHS decomposition shows zero automated harm on Revenge Porn, Doxing, and Slut Shaming. Sexual Harassment per-class EHS drops from 0.1250 (global) to 0.0500 (HA), a 60.0% reduction. Cyberstalking is the one class where HA does not improve over global thresholding: one Cyberstalking sample receives calibrated confidence ≥ 0.995 despite being incorrectly classified. This illustrates a fundamental limitation of confidence-based selective classification [10], [12] and motivates future exploration of ensemble-based uncertainty estimation [14].

E. Moderator Workload Analysis

Under the HA policy, 157 samples route to human review, of which approximately 125 (79.6%) are Tier 1 content. Under the global threshold, only 31.2% of human-reviewed samples are Tier 1. The HA framework concentrates reviewer attention on the most dangerous content, improving moderation efficiency and protecting moderators from unnecessary exposure to harmful content in lower-tier categories [9].

TABLE VII
EXPECTED HARM SCORE (EHS) COMPARISON ACROSS POLICIES

Policy	EHS	Raw	Auto%	Human%
No Threshold	0.1868	71	100.0	0.0
Global Threshold	0.0474	18	56.8	12.6
HA (Ours)	**0.0184**	**7**	**54.5**	**41.3**

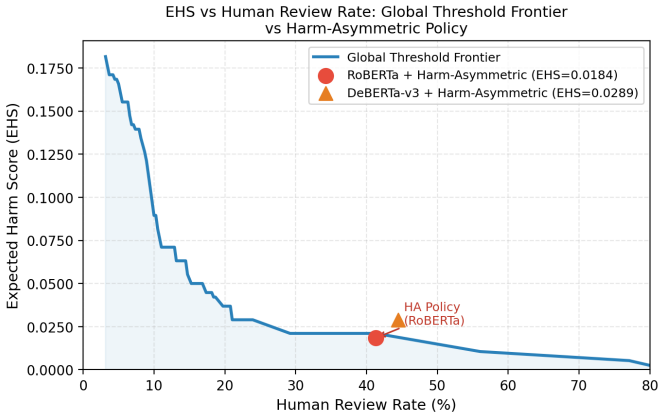

Fig. 4. EHS vs human review rate: global threshold Pareto frontier (blue curve) against the harm-asymmetric policy (red circle: RoBERTa; orange triangle: DeBERTa-v3). The RoBERTa HA policy lies below the global frontier at equal review rate, achieving lower EHS without additional moderator workload. The DeBERTa-v3 HA point lies on or above the frontier, showing that backbone calibration quality is a prerequisite for the harm-asymmetric advantage.

F. Sensitivity Analysis of Harm Weights and Thresholds

To assess whether the EHS results depend critically on the chosen harm weights, we evaluate two alternative weight configurations alongside the primary setting (Tier 1:3, Tier 2:2, Tier 3:1):

Two alternative configurations are evaluated: equal weights (1,1,1), which reduces EHS to an unweighted error rate, and amplified Tier 1 weights (5,2,1), which impose a stronger penalty on irreversible-harm automated errors.

Notably, the HA policy's absolute EHS value is invariant across all three weight configurations. This is a direct consequence of the policy design: because all Tier 1 content is routed to human review, there are no Tier 1 automated errors to weight regardless of the value assigned to w. We acknowledge that this invariance may appear to make the metric insulated from the weight it is supposed to evaluate. The correct interpretation is that the safety guarantee for the most dangerous category is structural rather than weight-dependent: it holds because the policy eliminates Tier 1 automation entirely, not because the weights are tuned to produce a particular result. The EHS metric does remain sensitive to weight choice for the other policies (No Threshold and Global Threshold), where automated Tier 1 errors are present and the choice of w changes the penalty substantially. Equivalently, the EHS advantage of HA over the global threshold increases as Tier 1 weights increase, confirming the framework is most beneficial precisely when the stakes are highest.

TABLE VIII
EHS SENSITIVITY TO HARM WEIGHT CONFIGURATIONS

Policy	w=(1, 1, 1)	w=(3, 2, 1)	w=(5, 2, 1)
No Threshold	0.0816	0.1868	0.2605
Global Thresh.	0.0250	0.0474	0.0671
HA (Ours)	0.0184	0.0184	0.0184
HA reduction vs Global	26.4%	61.2%	72.6%

V. DISCUSSION AND LIMITATIONS

The experimental results demonstrate that harm-asymmetric moderation produces a deployment policy that is both safer and more interpretable than global thresholding, at the cost of a higher human review rate (41.3% vs. 12.6%). Whether this trade-off is acceptable depends on the scale and staffing of the moderation operation.

The harm weights in EHS are ordinal values; future work could elicit more grounded weights through expert consultation with moderation practitioners and clinical psychologists [17]. The synthetic augmentation produces stylistically more uniform posts than natural social media content, which may reduce robustness to unusual linguistic patterns. The dataset is drawn from a single source (IEEE Dataport) and all evaluation is performed on in-distribution held-out data. The current results therefore reflect performance within a single content domain rather than across platforms, languages, or demographic communities. Calibration and thresholds optimised on this distribution will likely require re-evaluation under domain shift [10]. Cross-domain testing — for example, transferring thresholds trained on one platform to another — is a necessary step before this framework can be claimed as a general moderation policy rather than a proof-of-concept. The annotation methodology and inter-annotator agreement of the original dataset are not fully reported by the source. Automated moderation decisions should be reversible and subject to human appeal [5], [9].

VI. CONCLUSION

This paper proposed and evaluated a harm-asymmetric selective moderation framework as a proof-of-concept deployment policy. The framework combines class-weighted RoBERTa fine-tuning, post-hoc temperature scaling [10], and policy-driven tier-specific confidence thresholds derived from a literature-grounded harm severity taxonomy. The Expected Harm Score provides a harm-weighted deployment safety metric capturing what aggregate accuracy cannot [12], [17]. Empirical results demonstrate a 61.2% EHS reduction versus global thresholding and zero automated errors on Revenge Porn, with 99.52% accuracy across all auto-moderated content.

Future work will extend the dataset to multiple platforms and languages, explore deep ensembles [14] and Bayesian uncertainty estimation, and incorporate expert-elicited harm weights.

REFERENCES

[1] M. Dadvar, F. de Jong, and R. Trieschnigg, "Improving cyberbullying detection with user context," in *Proc. ECIR*, 2013.

[2] T. Davidson, D. Warmsley, M. Macy, and I. Weber, "Automated hate speech detection and the problem of offensive language," in *Proc. ICWSM*, 2017.

[3] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," in *Proc. NAACL-HLT*, 2019.

[4] Y. Liu et al., "RoBERTa: A robustly optimized BERT pretraining approach," *arXiv:1907.11692*, 2019.

[5] P. Fortuna and S. Nunes, "A survey on automatic detection of hate speech in text," *ACM Computing Surveys*, 2018.

[6] M. Mozafari, R. Farahbakhsh, and N. Crespi, "A BERT-based transfer learning approach for hate speech detection in online social media," in *Proc. Complex Networks*, 2020.

[7] A. Schmidt and M. Wiegand, "A survey on hate speech detection using natural language processing," in *Proc. SocialNLP*, 2017.

[8] Z. Waseem and D. Hovy, "Hateful symbols or hateful people? Predictive features for hate speech detection on Twitter," in *Proc. NAACL*, 2016.

[9] B. Vidgen et al., "Learning from disagreement: Robustness, fairness and uncertainty in toxic content classification," in *Proc. ACL*, 2021.

[10] C. Guo, G. Pleiss, Y. Sun, and K. Q. Weinberger, "On calibration of modern neural networks," in *Proc. ICML*, 2017.

[11] A. Niculescu-Mizil and R. Caruana, "Predicting good probabilities with supervised learning," in *Proc. ICML*, 2005.

[12] Y. Geifman and R. El-Yaniv, "Selective classification for deep neural networks," in *Proc. NeurIPS*, 2017.

[13] C. Cortes, G. DeSalvo, and M. Mohri, "Learning with rejection," *JMLR*, 2016.

[14] Y. Gal and Z. Ghahramani, "Dropout as a Bayesian approximation: Representing model uncertainty in deep learning," in *Proc. ICML*, 2016.

[15] R. El-Yaniv and Y. Wiener, "On the foundations of noise-free selective classification," *JMLR*, vol. 11, pp. 1605–1641, 2010.

[16] M. Zampieri et al., "SemEval 2019 Task 6: Identifying and categorizing offensive language in social media," in *Proc. SemEval*, 2019.

[17] R. Borkan et al., "Nuanced metrics for measuring unintended bias with real data for text classification," in *Proc. WWW*, 2019.

[18] B. Mathew et al., "HateXplain: A benchmark dataset for explainable hate speech detection," in *Proc. AAAI*, 2021.

[19] T. Tommaso et al., "HateBERT: Retraining BERT for abusive language detection in English," in *Proc. WOAH, ACL*, 2021.

[20] D. K. Citron and M. A. Franks, "Criminalizing revenge porn," *Wake Forest Law Review*, vol. 49, 2014.

[21] P. He, J. Gao, and W. Chen, "DeBERTaV3: Improving DeBERTa using ELECTRA-style pre-training with gradient-disentangled embedding sharing," *arXiv:2111.09543*, 2021.