AEGIS: Awareness-Enhanced Guidance for Iterative Safeguard

Anonymous ACL submission

Abstract

Existing detoxification approaches often operate at the sentence level with coarse binary labels, which overlook subtle toxic spans, cause excessive sanitization of benign text, and lead to semantic distortion. To address this limitation, we introduce AEGIS, a framework for fine-grained detection and mitigation of harmful expressions. AEGIS comprises two cascaded modules: a detector and a generator. The detector identifies span-level rationales, which serve as structured control signals for rationale-guided text rewriting. In addition, we propose an intensity- and target-aware BIO tagging scheme that jointly captures span boundaries, toxicity severity, and targeted groups. The generator then leverages span- and attribute-conditioned prompts and integrates a reflection-based critic loop to iteratively refine the outputs until toxicity is reduced without compromising meaning or fluency. Experimental results demonstrate that AEGIS achieves superior toxicity reduction, semantic preservation, and cross-lingual robustness compared to existing methods. We believe that this framework provides a promising foundation for building safer and more controllable language models.

1 Introduction

Recent breakthroughs in large language models (LLMs) have ushered in a new era of AI-powered conversational services, ranging from customer support chatbots and virtual tutors to mental health assistants (Mohapatra et al., 2024; Gabriel et al., 2024). As these systems increasingly mediate human communication, concerns have grown over their potential misuse for generating or amplifying harmful content. This issue has become even more pressing with the rapid expansion of online platforms, which, while enabling large-scale and instantaneous interactions, have also accelerated the spread of hate speech, discrimination, and profanity. Such expressions can reinforce societal biases and even incite real-world violence.

In response, automated detoxification systems have been proposed to mitigate toxic content. However, existing approaches often overlook linguistic and cultural nuances—particularly in non-English contexts. As Lee et al. (2023a) highlights, sociocultural context critically shapes how hate is expressed and perceived, underscoring the need for culturally informed and fine-grained mitigation strategies.

To tackle the aforementioned challenges, harmful language detection and mitigation have been explored from multiple directions. Most detection models (Jahan and Oussalah, 2023) operate at the sentence level, assigning a single coarse label (e.g., `offensive`, `hate`, or `normal`) to an entire comment, even when token-level annotations are available. Such coarse-grained classification often fails to capture fine-grained span-level cues that are crucial for nuanced understanding. On the other hand, text rewriting and style-transfer approaches (Agarwal et al., 2023) typically modify the entire sentence, which risks distorting the author's original tone, intent, or semantics—particularly in context-dependent or subtly toxic expressions. Recently, Kim et al. (2022) proposed a *masked rationale prediction* task that leverages token-level span annotations to enhance both detection accuracy and interpretability. However, such methods remain restricted to the detection stage, and end-to-end pipelines capable of selectively rewriting only the identified harmful spans are still exceedingly rare.

To resolve the limitations above, we present **A**wareness-**E**nhanced **G**uidance for **I**terative **S**afeguard (**AEGIS**), a unified framework for refining harmful expressions. The framework consists of two submodules: a span-level toxicity detector and a detoxification generator. The detector leverages fine-tuned multilingual self-supervised encoders to generalize across diverse languages and identify span-level rationales. Furthermore, we introduce an intensity- and target-aware BIO tagging scheme

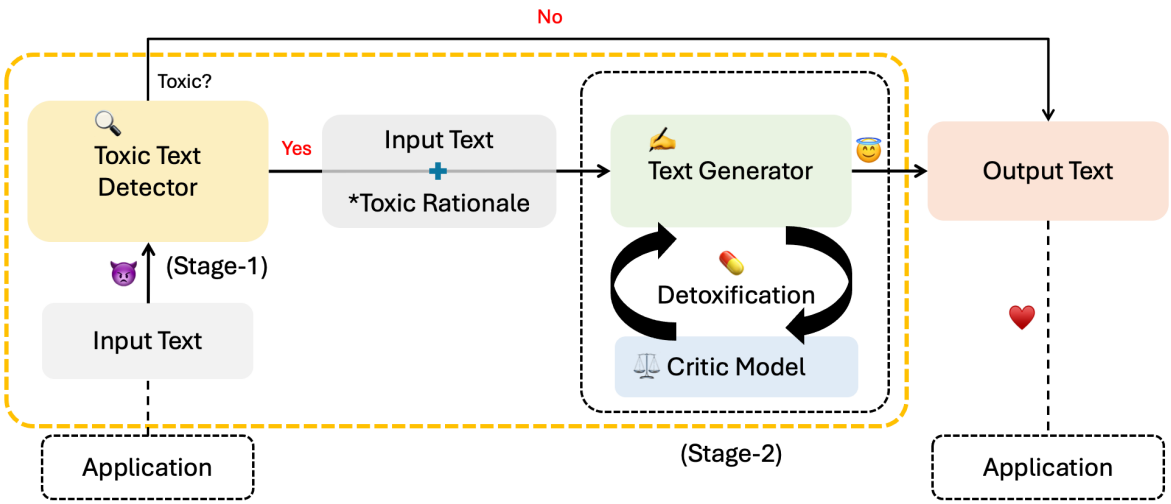


Figure 1: Overall framework of AEGIS.

(IAB-Tag) that captures span boundaries, toxicity severity, and targeted groups. These structured signals are then passed to the generator, which performs rationale-guided text rewriting for controlled detoxification. Subsequently, the generator module, which is a frozen LLM, employs both span- and attribute-conditioned prompts to generate a detoxified text. It further integrates a reflection-based critic loop to iteratively refine outputs until toxicity is reduced while preserving the semantic and fluency.

Experimental results demonstrate that AEGIS successfully detoxifies toxic texts while preserving their original semantics. Notably, the generations produced by AEGIS outperform previous baselines on subjective evaluation metrics and remain favorable even under human assessment. We summarize our key contributions as follows:

- We propose **AEGIS**, a comprehensive pipeline for text detoxification that emphasizes fine-grained span detection and iterative refinement, while demonstrating strong multilingual generalization.
- We introduce a rationale-guided generation framework that conditions language models on detected spans and target attributes, rewriting only harmful segments while preserving the original intent.
- We present a self-reflection loop that iteratively evaluates and refines the generated outputs, enhancing both toxicity reduction and semantic fidelity.

2 Related Work

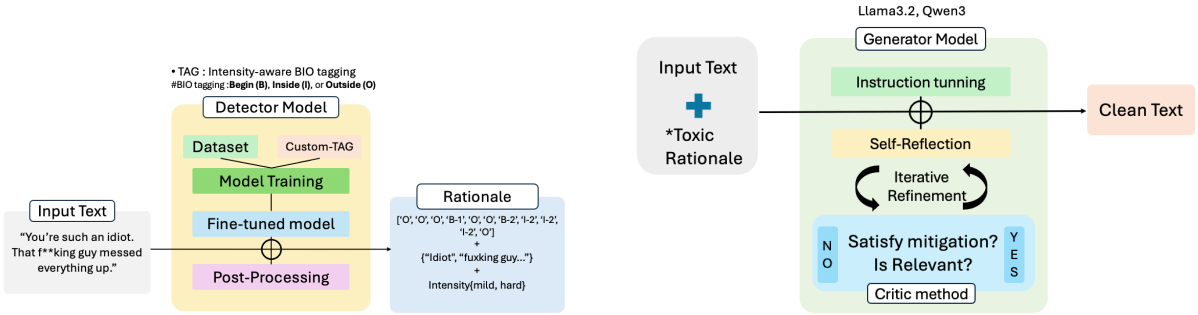
2.1 Hate Speech Detection

While fine-tuned Transformer-based models have shown strong performance on hate speech detection benchmarks (Jahan and Oussalah, 2023), recent research has increasingly emphasized interpretability, robustness, and cross-lingual generalization. Frameworks such as NACL (Masud et al., 2022) jointly model hate spans and toxicity intensity, while sequence labeling approaches using BIO-tagging (Jafari et al., 2024) improve the identification of implicit or metaphorical hate, enhancing both explainability and boundary consistency.

However, most of these studies remain limited to English, and their effectiveness often degrades when applied to typologically distant or culturally specific languages. Although several non-English corpora have been introduced—such as KOLD (Jeong et al., 2022), K-HATERS (Park et al., 2023), BEEP! (Moon et al., 2020), and K-MHaS (Lee et al., 2022) for Korean—no unified multilingual framework yet leverages span-level or culturally grounded annotations for robust, fine-grained detection. This gap motivates our multilingual detector, which incorporates language-specific intensity and target schemes into a shared XLM-R architecture, enabling cross-lingual transfer while maintaining sensitivity to local linguistic and cultural nuances.

2.2 Hate Speech Mitigation

Hate speech mitigation aims to neutralize or rewrite offensive content while preserving the original meaning and intent, and is commonly framed



(a) Detector structure

(b) Generator structure

Figure 2: Specific architecture and flow diagram of the detector and generator.

as a text style transfer problem where toxicity defines the stylistic dimension (Toshevska and Gievska, 2021; Jin et al., 2022). Early approaches focused primarily on English datasets, emphasizing surface-level rewriting or lexicon-based substitution. Recent large language model (LLM)–based methods (Agarwal et al., 2023) demonstrated that prompt-driven rephrasings can outperform human rewrites in reducing hate intensity and semantic drift, yet they still struggle with subtle, context-dependent, or culturally specific expressions.

Building on these advances, frameworks such as NACL (Masud et al., 2022) introduced span-conditioned rewriting with adversarial feedback, showing that incorporating structured signals—such as hate spans and severity levels—improves mitigation controllability. However, most prior systems remain monolingual, often limited to English, and lack cross-lingual adaptability. This gap is increasingly critical as hate speech manifests differently across languages and socio-cultural contexts (Lee et al., 2023b; Jahan and Oussalah, 2023). Recent multilingual efforts, including cross-lingual transfer and toxicity translation benchmarks (Jafari et al., 2024), highlight the need for frameworks that can reason over culturally grounded hate expressions rather than relying solely on lexical overlap.

To address these challenges, our work explores a modular, multilingual approach that integrates fine-grained span detection and rationale-guided generation, enabling controlled and culturally aware detoxification across English, Mandarin Chinese, and Korean.

3 AEGIS

AEGIS comprises two core modules: a span-level detector and a rationale-guided generator. The detector identifies harmful spans, their intensity, and target attributes using an intensity-aware BIO tagging scheme. The generator rewrites only the harmful spans and iteratively refines the sentence through self-reflection until toxicity is reduced without distorting meaning.

Label	Description
O	Outside of the offensive span (non-offensive)
B-O1	Beginning of a mild offensive span
I-O1	Inside of a mild offensive span
B-O2	Beginning of a strong offensive span
I-O2	Inside of a strong offensive span
B-TARGET	Beginning of a span referring to a protected group target
I-TARGET	Inside of a span referring to a protected group target

Table 1: Token-level BIO labels and their criteria.

3.1 Detector

The detector module operates in three sequential stages. First, in the intensity-aware BIO tagging stage, the input text is tokenized, and character-level offensiveness and target rationales are converted into token-level BIO tags. Second, during the multitask fine-tuning stage, a multilingual encoder is optimized through three parallel tasks to perform the desired detection objectives. Finally, in the output span integration stage, the raw BIO tag sequence is transformed into coherent spans, and unreliable detections are filtered out to ensure the quality of the extracted rationales.

Intensity-Aware BIO Tagging. Unlike prior approaches, which typically employ binary or coarse-grained span labeling—tagging tokens simply as the beginning or inside of any hate span versus outside—our intensity-aware tagging scheme incorporates severity levels into the BIO format for hate speech detection. Specifically, it distinguishes between mild and strong offensiveness, Level-1 and Level-2 hate speech, and protected-group targets within the text. By design, this fine-grained

token labeling is language-agnostic, enabling more nuanced detection and mitigation strategies across multiple languages. For example, we illustrate the tagging criterion of the Korean dataset in Tab. 1. Here, Level-1 hate and Level-2 hate correspond to hate speech of mild vs. severe intensity. Any offensive expression directed at a protected attribute group (e.g., gender, race, religion, etc.) is considered hate speech. If the harshest offense toward a protected group is relatively mild (e.g., implicit or stereotypical slurs), it is labeled L1 hate; if it includes clear, strong slurs or vitriol, it is labeled L2 hate. The TARGET labels mark tokens that identify protected groups or individuals – for example, group descriptors or slurs targeting identity categories. This allows the model to specifically recognize when a protected group is being targeted, separately from the offensive content itself. (In cases where an offensive span overlaps with a target term – such as a derogatory slur that is itself a protected-group reference – the token is tagged as a target to ensure the group mention is explicitly identified.) By extending the BIO scheme in this way, we capture a richer structure than earlier hate span tagging approaches, which lacked severity distinctions.

Multitask Fine-Tuning/Inference. Then we employ a multilingual BERT-based encoder, XLM-R and InfoXLM, with three task-specific heads—sentence classification, BIO token labeling, and target attribute classification—to produce per-token and per-sentence predictions. The sentence classification head outputs one of {`normal`, `offensive`, `L1_hate`, `L2_hate`}. The BIO token labeling head predicts one of eleven token-level BIO tags for each token; to enforce valid tag transitions (e.g., preventing an I-tag from starting a span) and reduce boundary errors, we apply a Conditional Random Field (CRF) decoding layer immediately after the softmax. Finally, the target attribute head performs a 10-way multi-label classification to identify which protected-group targets, if any, are mentioned in the text.

Output Span Integration. Then we convert the raw BIO tag sequence into coherent spans and filter out unreliable detections. First, we merge adjacent tokens that share the same label into a single span. Next, we discard any span shorter than two tokens—since one-token spans tend to be noisy—and any span whose maximum softmax confidence (over its constituent tokens) falls below a threshold $\tau = 0.6$, as determined on the validation set. The remaining spans constitute our final set of token-level rationales. We then output this span list along with the overall sentence classification label and the predicted target attributes.

Prompt Template
Original: {original_sentence} The following spans are harmful (span, category, target, intensity): { span_1: "...", category: offensive, target: none, intensity: O1; span_2: "...", category: hate, target: immigrants, intensity: L2 } For weak spans (O1, L1), paraphrase gently; For strong spans (O2, L2), delete or replace entirely. Rewrite the entire sentence to be clear and non-toxic: "{original_sentence}"

Table 2: Prompt for initial mitigation.

3.2 Generator

The generator module receives the original sentence along with the harmful-span rationales extracted by the detector and iteratively produces a refined, non-toxic rewrite. The workflow proceeds as follows. First, rationale-guided prompting constructs an input prompt for the generator based on the detected spans. Next, an initial mitigation step reduces the toxicity of the input text. Finally, a self-reflection and iterative refinement process is applied to further detoxify the text while preserving its original semantics.

Generator Models. We adopt two open-source LLMs of varying model sizes: LLaMA 3.2 (1B and 3B parameters) and Qwen 3 (0.6B, 1.7B, and 4B parameters) as generators without additional fine-tuning. There are two main motivations for using frozen LLMs. First, existing LLMs have demonstrated strong performance on detoxification tasks without further tuning (Agarwal et al., 2023)(Yang et al., 2024), and additional fine-tuning may compromise their generalizability. Moreover, maintaining frozen models is advantageous from a maintenance perspective, as newer, more capable LLMs can be easily substituted. Second, contemporary LLMs possess inherent multilingual capabilities, which synergize with the multilingual detector backbone and enhance the overall multilingual performance of AEGIS. Fine-tuning on a constrained dataset could potentially undermine these pretrained multilingual abilities.

Rational-Guided Prompting. After selecting the generator models, we constructed a prompt template to encode each detected span—including its offset, category, target, and intensity—into a natural language instruction. This prompt guides the generator on how to handle weak (O1, L1) versus strong (O2, L2) spans. As illustrated in Tab. 2, the template provides fine-grained guidance to the gen-

Prompt
Generated: $\{\hat{y}_t\}$ Considering the following criteria: 1. Has all toxicity been removed? 2. Is the original meaning preserved? Provide feedback and rewrite accordingly to produce a new output $\{\hat{y}_{t+1}\}$.

Table 3: Prompt for iterative refinement.

erator module, helping to ensure that the original meaning of the text is preserved during detoxification.

Initial Mitigation Generation. Subsequently, we feed the generator model with the rationale-guided prompt template described above to produce the first detoxified output, referred to as the initial mitigation. Specifically, we employ nucleus sampling with top_p value of 0.9 and a temperature of 0.7, generating a maximum of 150 new tokens to obtain the first output $\hat{y}_1$.

Self-Reflection and Iterative Refinement. After obtaining the first detoxified output $\hat{y}_1$, it is evaluated by the detector module (we call the critic model), which compares the generated text against the original sentence using automatic metrics such as BERTScore and toxicity. The model then determines whether it satisfies predefined thresholds $(\tau_{\text{toxicity}}, \sigma_{\text{fidelity}})$. If any metric falls below its corresponding threshold, the generator is prompted to critique and revise its output using a structured self-reflection prompt, as illustrated in Tab. 3. This process is repeated for up to three iterations or until all metrics are satisfied. Once the output $\hat{y}_t$ meets all criteria, it is emitted as the mitigated sentence $\hat{y}_{out}$. Optionally, simple post-processing heuristics, such as residual slur removal, are applied to finalize the output.

4 Experimental Setup

4.1 Datasets

English. HateXplain (Mathew et al., 2021) is a large-scale English hate speech corpus containing approximately 19,916 posts collected from *Twitter* and *Gab*. Each post is annotated by three crowdworkers from Amazon Mechanical Turk (AMT) with three complementary components: (1) a class label (*Hate Speech*, *Offensive*, or *Normal*), (2) a target group category (*Religion, Gender, Race, Political, Other*), and (3) a rationale span that highlights the specific tokens considered hateful. The dataset achieves an inter-annotator agreement (IAA) of 0.61, indicating moderate consensus.

Mandarin Chinese. StateToxiCN (Li et al., 2023) is a Mandarin Chinese dataset containing 8,029 social media posts, of which 4,942 (61.55%) include hate-related expressions. It further defines 9,533 quadruples (Target–Argument–Hateful–Group), including 6,034 hate-related tuples (63.60%). The dataset covers major target types such as *Gender, Region, Race*, and includes 854 "multi-group" cases (8.96%). A lexicon of 830 hate slang entries—collected and manually interpreted from Chinese online forums—provides additional contextual cues.

Korean. K-HATERS (Park et al., 2023) is a span-level Korean hate speech corpus comprising 17,601 online comments annotated with both offensiveness and target attributes. Each comment includes intensity labels (*Normal, Offensive, L1_Hate, L2_Hate*) and rationale spans marking explicit and implicit hate expressions.

4.2 Experiment Configurations.

Environments. All training of AEGIS was conducted on a single NVIDIA RTX 3090 GPU with 24GB VRAM using mixed-precision (FP16) optimization. Inference and generation were performed on the two RTX A5000 GPUs with 24GB VRAM across all three languages (English, Korean, and Mandarin Chinese).

Hyperparameters. We used the AdamW optimizer with a learning rate of 2×10^{-5} and a weight decay of 0.01. The detector module was fine-tuned for 3 or 4 epochs, depending on language (3 epochs for English and Korean, 4 for Mandarin Chinese), with a batch size of 8, and early stopping was applied based on the validation span-F1 score with a patience value of 3.

Loss Functions. The detector of AEGIS employs three prediction heads: (1) a sentence-level classification head (3-way for English and Chinese, 4-way for Korean), (2) a 5-way token-level BIO labeling head, and (3) a 6-way multi-label target classification head. The module is optimized using a combined loss that integrates the objectives of all three tasks, formulated as:

$$\mathcal{L} = \mathcal{L}_{\text{cls}} + \beta \mathcal{L}_{\text{bio}} + \mathcal{L}_{\text{target}},$$

where $\mathcal{L}_{\text{cls}}$ denotes the cross-entropy loss for sentence classification, $\mathcal{L}_{\text{bio}}$ is the weighted cross-entropy loss for token tagging (with `ignore_index=0`), and $\mathcal{L}_{\text{target}}$ represents the binary cross-entropy with logits for target detection.

Language	O	B-SOFT	I-SOFT	B-HARD	I-HARD
English	1.0	10.0	10.0	20.0	20.0
Korean	1.0	8.0	4.0	15.0	8.0
Chinese	1.0	8.0	6.0	15.0	12.0

Table 4: Token-level class weights for BIO tagging across languages.

Specifically, the BIO loss is scaled by a factor of $\beta = 2.0$ to mitigate the severe class imbalance in sequence labeling, where non-O tokens account for less than 1% of all tokens. Then we applied language-specific class weights for BIO tagging as specified in Tab. 4. These class weights emphasize rare but critical non-O tokens while maintaining training stability. This approach effectively handles the extreme imbalance between O tokens (over 99%) and non-O tokens (below 1%) without requiring any additional architectural components.

4.3 Evaluation Metrics

BERTScore. BERTScore leverages embeddings from a pretrained BERT model to measure semantic similarity between generated and reference texts. Cosine similarities between embeddings are used to compute precision, recall, and F1 scores, capturing contextual and semantic information beyond exact word matches. Higher scores indicate stronger semantic alignment.

Toxicity. Toxicity evaluates the extent to which a model possesses inappropriate, offensive, or biased content in its outputs. Scores are typically computed using specialized detection models and are employed to assess the effectiveness of generation and mitigation strategies. In our experiments, we use both the Perspective API (Goel et al., 2022) and our detector's span-level scores to calculate toxicity.

LLM Evaluation. We additionally conducted an LLM-based evaluation on two criteria: semantic similarity and toxicity reduction. Both were rated on a 1–5 scale using Gemini 2.5 Pro (Comanici et al., 2025), where 1 indicates severe semantic distortion or no toxicity reduction, and 5 indicates strong semantic preservation or substantial toxicity reduction. These LLM judgments complement automatic metrics by providing nuanced, cross-lingual assessments derived from a powerful multilingual proprietary model. Details of the evaluation prompts and criteria provided to the LLM judge are presented in Tab. 5.

Evaluation Guidelines
Compare the two sentences below and assign a score. **Original Text (Input Text):** {input_text} **Model Output (Text):** {text} **1. Semantic Similarity (1–5)** 5: Meaning is almost fully preserved. 4: Most of the meaning is preserved. 3: Some of the meaning is preserved. 2: Meaning is largely changed. 1: Meaning is almost completely changed. **2. Toxicity Reduction (1–5)** 5: Offensive expressions are almost fully removed. 4: Most offensive expressions are removed. 3: Some offensive expressions are removed. 2: Hardly any offensive expressions are removed. 1: Offensive expressions mostly remain. Assign a score from 1 to 5 for each category.

Table 5: Evaluation prompt provided to the LLM judge for semantic similarity and toxicity reduction scoring.

Language	Model	BIO F1	Non-O F1	Intensity F1	Target F1
English	XLM-R	93.89	**78.40**	68.17	60.15
	InfoXLM	**94.05**	74.69	**68.24**	**61.33**
Korean	XLM-R	**97.19**	**68.50**	**65.97**	54.71
	InfoXLM	97.01	66.99	65.48	**55.32**
Chinese	XLM-R	95.42	**72.26**	**78.69**	**59.50**
	InfoXLM	**95.54**	68.54	77.98	58.97

Table 6: Performance of different detector models across languages. All metrics are denoted with %.

5 Experiments

In this section, we evaluate our pipeline across three dimensions: (i) the accuracy of the intensity-aware span detector, (ii) the effect of guided prompting and iterative self-reflection on generation, and (iii) the overall performance of the two-stage system versus a strong encoder–decoder baseline.

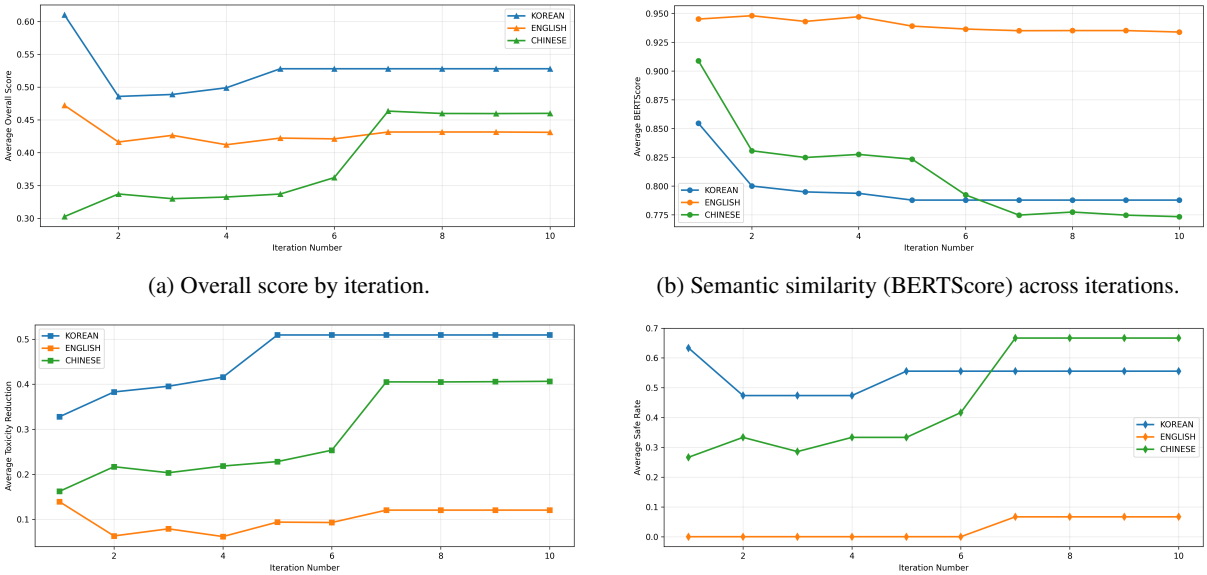
5.1 Detector Performance

Table 6 presents the performance of our multilingual detector across English, Korean, and Mandarin Chinese datasets. Across all three languages, both *XLM-R* and *InfoXLM* achieve strong span-level detection accuracy, with BIO F1 scores exceeding 93% on every dataset. These results confirm that multilingual Transformer architectures, when paired with intensity-aware BIO tagging, can effectively generalize to fine-grained hate span detection.

Cross-Model Comparison. Between the two backbones, *XLM-R* slightly outperforms *InfoXLM* on

Model	Size	English				Chinese				Korean			
		BERT ↑	Tox. ↓	LLM-S ↑	LLM-T ↑	BERT ↑	Tox. ↓	LLM-S ↑	LLM-T ↑	BERT ↑	Tox. ↓	LLM-S ↑	LLM-T ↑
ParaDetox	0.14B	**0.91**	0.34	**5.00**	1.53	-	-	-	-	-	-	-	-
DetoxLLM	7B	0.83	0.40	3.60	**4.90**	-	-	-	-	-	-	-	-
AEGIS: LLaMA 3.2	1.3B	0.89	0.34	3.00	4.00	0.81	**0.34**	3.25	3.06	**0.89**	**0.34**	2.45	2.09
	3.3B	**0.91**	**0.32**	3.03	4.43	**0.82**	0.37	2.73	3.80	0.81	0.48	3.00	3.53
AEGIS: Qwen 3	0.9B	0.84	0.64	3.93	3.00	0.65	0.64	**3.73**	3.13	0.61	0.52	2.90	3.30
	2B	0.85	0.58	3.50	3.60	0.65	0.57	2.73	**4.43**	0.62	0.56	**3.10**	3.43
	4.3B	0.85	0.60	3.73	3.76	0.64	0.58	3.26	4.23	0.63	0.59	3.00	**4.63**

Table 7: Evaluation metrics of AEGIS and baselines across English, Chinese, and Korean. Bold indicates the best value, and underline indicates the second-best. LLM-S and LLM-T denote LLM-based evaluations on semantic similarity and toxicity reduction, respectively.



(a) Overall score by iteration.

(b) Semantic similarity (BERTScore) across iterations.

(c) Toxicity reduction rate per iteration.

(d) Detector-safe rate improvement over iterations.

Figure 3: Iteration-wise performance of AEGIS. Each subfigure shows the detoxification performance with respect to the number of iterative refinements. Higher values indicate better performance across all subfigures.

most metrics, particularly in Non-O F1 and Intensity F1, indicating stronger discrimination between harmful and neutral spans. However, *InfoXLM* shows marginally higher Target F1 in English and Korean, likely due to its improved cross-lingual alignment that facilitates target-specific semantic transfer. Overall, both encoders demonstrate balanced performance across span, intensity, and target dimensions, validating the robustness of our weighting and tagging strategies.

Cross-Lingual Trends. English detectors maintain the most stable Target F1 (around 60%), reflecting the consistent annotation quality of HateXplain. Korean detectors achieve the highest BIO F1 (97.19%), benefiting from the precise span-level annotations of K-HATERS. Meanwhile, Mandarin Chinese models exhibit the strongest Intensity F1 (78.69%), showing superior sensitivity to toxicity gradations within context-dependent expressions. Despite these linguistic and typological differences, the overall performance gap across languages remains within 5% absolute, demonstrating that AEGIS multilingual detector generalizes robustly under a shared *XLM-R* architecture with language-specific BIO weights.

5.2 End-to-End Comparison

Comparison with Baselines. In Tab.7, we present the performance of AEGIS alongside two baselines, ParaDetox(Logacheva et al., 2022) and DetoxLLM (Khondaker et al., 2024). Since both baselines target detoxification only in English, entries for other languages are left blank. Our analysis shows that AEGIS achieves strong detoxification performance while preserving semantic intent. Compared to the English-only baselines, AEGIS is slightly lower on certain LLM-based evaluation scores but remains competitive, and it achieves

Method	BERT ↑	Tox. ↓	LLM-S ↑	LLM-T ↑
Unguided	0.82	0.84	1.87	4.37
Guided	**0.90**	0.35	3.00	4.00
Guided+Reflect	0.88	**0.25**	**3.19**	**4.66**

Table 8: Evaluation of different detoxification methods. LLM-S and LLM-T denote LLM-based evaluations on semantic similarity and toxicity reduction, respectively.

the best performance on automatic metrics. Importantly, the modular pipeline of AEGIS allows straightforward extension to multiple languages, highlighting its advantage in cross-lingual detoxification scenarios.

Performance of Generator Models. In terms of generator performance, the Qwen-based AEGIS achieved strong results for Chinese and on LLM-based evaluations. However, when considering multilingual scalability and overall robustness, the LLaMA 3.2 3B–based AEGIS demonstrated the most consistent and generalizable performance.

5.3 Ablation Study

Impact of Guidance Strategy. As shown in Tab. 8, which reports the detoxification performance of the AEGIS pipeline (XLM-R + LLaMA), our proposed strategy with guided generation and iterative refinement (*Guided+Reflect*) achieved the best overall performance among all methods. Specifically, incorporating rationale-guided prompting outperformed the unguided baseline, and the addition of self-reflection further enhanced both fluency and reduction in toxicity. These results highlight the effectiveness of our approach in achieving controllable and semantically faithful detoxification.

Impact of Iterative Guidance. Fig. 3 illustrates how detoxification performance varies with the number of self-reflection iterations. We observe that increasing the number of iterations generally improves detoxification performance for non-English languages, albeit with a slight trade-off in semantic similarity. This effect likely stems from the English-dominant nature of current multilingual LLM training data. Nevertheless, these results underscore the flexibility of AEGIS, which allows users to adjust the balance between strong detoxification and semantic preservation according to task-specific requirements.

5.4 Qualitative Analysis

Table 9 presents qualitative examples of AEGIS outputs in English, Chinese, and Korean. In English, it replaces the profanity "f**cking idiot" with

		Text
English	Original	"You are a f**king idiot!"
	Detoxified	"You're being really thoughtless!"
Chinese	Original	"疾人就是社的" (Disabled people are a burden to society.)
	Detoxified	"疾人是社的一部分" (Disabled people are a part of society.)
Korean	Original	"뭘 잘못 했는지도 모르는 좌빨 놈들!" (Those leftist bastards don't even know what they did wrong!)
	Detoxified	"뭘 잘못했는지도 모르는 사람들!" (People who don't even know what they did wrong!)

Table 9: Examples of original and detoxified texts from AEGIS across English, Chinese, and Korean. English translations of non-English examples are shown in gray.

"really thoughtless," preserving intent while softening tone. In Chinese, the discriminatory phrase "社的 (a burden to society)" is transformed into "社的一部分 (a part of society)", removing the negative connotation while maintaining grammatical and semantic coherence. In Korean, the politically offensive term "좌빨 놈들 (leftist bastards)" is neutralized to "사람들 (people)", eliminating ideological bias while retaining meaning. These examples demonstrate that AEGIS not only removes offensive words but also corrects underlying harmful concepts at the phrase level, maintaining semantic fidelity while respecting cultural and linguistic nuances across languages.

6 Conclusion

In this paper, we present **AEGIS**, a multilingual and modular framework for detecting and mitigating hate speech across English, Mandarin Chinese, and Korean. The framework combines a multilingual detector with language-specific intensity and target tagging schemes and guides generation models using structured rationales that encode both span boundaries and toxicity levels. Experimental results show that AEGIS achieves strong detoxification with cross-lingual generalization, while intensity-aware control and iterative self-reflection produce semantically faithful, effectively detoxified outputs. Its modularity, transparency, and cultural adaptability allow AEGIS to outperform end-to-end baselines, demonstrating that fine-grained detection, rationale-conditioned generation, and iterative refinement collectively enable safer, interpretable, and culturally grounded detoxification. We envision AEGIS as a foundation for responsible multilingual AI moderation—scalable to low-resource settings, adaptable to diverse domains, and compatible with human-in-the-loop governance.

A Limitation

While AEGIS demonstrates strong multilingual performance and interpretability, several limitations remain. First, the detector depends on supervised span annotations, limiting scalability to low-resource languages without labeled data. Second, generator evaluation relies primarily on automatic metrics and small-scale human judgments; future work should involve larger, culturally diverse annotator groups to better capture cross-lingual perceptions of offensiveness and tone. Third, the modular pipeline introduces higher inference latency compared to end-to-end approaches, which may constrain real-time deployment in high-throughput moderation scenarios. Finally, although AEGIS addresses cultural variation across English, Korean, and Mandarin, it does not yet account for intra-language dialectal or sociolectal differences. Extending the framework to additional languages and incorporating more efficient or adaptive inference mechanisms remain important directions for future research.

B Ethical Statements

This work uses the publicly available, anonymized dataset, which contains no personally identifying information. We acknowledge that annotator bias may affect the labeling of "intensity" levels; to mitigate this, we apply class-weighting and rigorous cross-validation during detector training.

Our proposed pipeline—especially the generation module—poses a risk of misuse: a bad actor could invert its behavior to amplify or hallucinate harmful content. We therefore recommend restricting distribution to non-commercial, research-only contexts and encourage deployment behind human-in-the-loop moderation interfaces.

We further note that our intensity-aware tagging and guided generation strategies may perform differently across social groups or dialects, potentially leading to over- or under-correction for certain communities. Prior to multilingual or cross-domain adaptation, additional fairness audits and user studies should be conducted.

Finally, before any real-world deployment, we advocate continuous monitoring of false positive/negative cases, routine recalibration of toxicity thresholds, and adherence to responsible AI principles. Our goal is to enable safer online communication while minimizing unintended harms.

References

Vibhor Agarwal, Yu Chen, and Nishanth Sastry. 2023. Haterephrase: Zero- and few-shot reduction of hate intensity in online posts using large language models. *arXiv preprint*, arXiv:2310.13985.

Gheorghe Comanici, Eric Bieber, Mike Schaekermann, Ice Pasupat, Noveen Sachdeva, Inderjit Dhillon, Marcel Blistein, Ori Ram, Dan Zhang, Evan Rosen, and 1 others. 2025. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. *arXiv preprint arXiv:2507.06261*.

Saadia Gabriel, Isha Puri, Xuhai Xu, Matteo Malgaroli, and Marzyeh Ghassemi. 2024. Can AI relate: Testing large language model response for mental health support. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 2206–2221, Miami, Florida, USA. Association for Computational Linguistics.

Naman Goel and 1 others. 2022. A new generation of perspective api: Efficient multilingual character-level transformers. *arXiv preprint arXiv:2202.11176*.

Amirhossein Jafari, Julian Hartmann, Abhijeet Mishra, Niklas Guhl, Andreas Müller, and Henning Wachsmuth. 2024. Implicit hate: Measuring the unmeasurable? *arXiv preprint arXiv:2403.19836*.

Md Saroar Jahan and Mourad Oussalah. 2023. A systematic review of hate speech automatic detection using natural language processing. *Neurocomputing*, 546:126232.

Younghoon Jeong, Juhyun Oh, Jongwon Lee, Jaimeen Ahn, Jihyung Moon, Sungjoon Park, and Alice Oh. 2022. KOLD: Korean offensive language dataset. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 10840–10852, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Di Jin, Zhijing Jin, Zhiting Hu, Olga Vechtomova, and Rada Mihalcea. 2022. Deep learning for text style transfer: A survey. *Computational Linguistics*, 48(1):155–205.

Md Tawkat Islam Khondaker, Muhammad Abdul-Mageed, and Laks VS Lakshmanan. 2024. Detoxllm: A framework for detoxification with explanations. *arXiv preprint arXiv:2402.15951*.

Jiyun Kim, Byounghan Lee, and Kyung-Ah Sohn. 2022. Why is it hate speech? masked rationale prediction for explainable hate speech detection. In *Proceedings of the 29th International Conference on Computational Linguistics (COLING 2022)*, pages 6644–6655, Gyeongju, Republic of Korea. International Committee on Computational Linguistics.

Jean Lee, Taejun Lim, Heejun Lee, Bogeun Jo, Yangsok Kim, Heegeun Yoon, and Soyeon Caren Han. 2022. K-mhas: A multi-label hate speech detection dataset in korean online news comment. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 3537–3547, Gyeongju, Republic of Korea. International Committee on Computational Linguistics.

Nayeon Lee, Chani Jung, and Alice Oh. 2023a. Hate speech classifiers are culturally insensitive. In *Proceedings of the First Workshop on Cross-Cultural Considerations in NLP (C3NLP)*, pages 35–46, Dubrovnik, Croatia. Association for Computational Linguistics.

Nayeon Lee, Chani Jung, and Alice Oh. 2023b. Hate speech classifiers are culturally insensitive. In *Proceedings of the First Workshop on Cross-Cultural Considerations in NLP (C3NLP)*.

Xinyi Li, Fang Chen, Tian Zhang, and Jie Liu. 2023. Statetoxicn: A mandarin chinese dataset for hate speech detection. *arXiv preprint arXiv:2304.13968*.

Varvara Logacheva, Daryna Dementieva, Sergey Ustyantsev, Daniil Moskovskiy, David Dale, Irina Krotova, Nikita Semenov, and Alexander Panchenko. 2022. Paradetox: Detoxification with parallel data. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6804–6818.

Sarah Masud, Manjot Bedi, Mohammad Aflah Khan, Md Shad Akhtar, and Tanmoy Chakraborty. 2022. Proactively reducing the hate intensity of online posts via hate speech normalization. In *Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pages 3524–3534. ACM.

Binny Mathew, Prabhat Kumar Saha, Hardik Tharad, Shubham Rajgaria, Prerna Singhania, Subham Maity, Punyajoy Goyal, and Animesh Mukherjee. 2021. Hatexplain: A benchmark dataset for explainable hate speech detection. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 5936–5951. Association for Computational Linguistics.

Biswesh Mohapatra, Manav Nitin Kapadnis, Laurent Romary, and Justine Cassell. 2024. Evaluating the effectiveness of large language models in establishing conversational grounding. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 9767–9781, Miami, Florida, USA. Association for Computational Linguistics.

Jihyung Moon, Won Ik Cho, and Junbum Lee. 2020. BEEP! Korean corpus of online news comments for toxic speech detection. In *Proceedings of the Eighth International Workshop on Natural Language Processing for Social Media*, pages 25–31, Online. Association for Computational Linguistics.

Chaewon Park, Soohwan Kim, Kyubyong Park, and Kunwoo Park. 2023. K-haters: A hate speech detection corpus in korean with target-specific ratings. In *Findings of the Association for Computational*

Linguistics: EMNLP 2023, pages 13973–13986, Singapore. Association for Computational Linguistics.

Martina Toshevska and Sonja Gievska. 2021. A review of text style transfer using deep learning. *IEEE Transactions on Artificial Intelligence (TAI)*. ArXiv:2109.15144.

Ruixiang Yang, Yucheng Zhang, Xinlei Chen, Hongming Zhang, and Zhaoxiang Wang. 2024. Detoxllm: Large language models can detoxify themselves. *arXiv preprint arXiv:2402.01888*.