A Survey of Toxicity Detection and Mitigation Strategies for Multilingual Language Models

Soham Dan[1], Himanshu Beniwal[2], Thomas Hartvigsen[3]
[1]Scale AI, [2]Indian Institute of Technology Gandhinagar [3]University of Virginia,
soham.dan@scale.com, himanshubeniwal@iitgn.ac.in, hartvigsen@virginia.edu

Abstract

Large language models (LLMs) are transforming natural language processing across diverse linguistic communities. However, they can reproduce and amplify toxic content, including hate speech, harassment, and bias, posing significant risks to multilingual applications. We provide the first comprehensive survey of the many detoxification methods specifically tailored to multilingual LLMs. First, we define toxicity its measurement, then we provide a brief review of monolingual mitigation strategies, including data filtering, style transfer, expert-based logit steering, retrieval augmentation, and alignment with human feedback. We then present an in-depth taxonomy of multilingual approaches spanning (1) training methods, (2) post-hoc editing and decoding strategies, (3) alignment and reinforcement-learning techniques, and (4) data-centric innovations, such as parallel detox corpora and synthetic data generation. Finally, we discuss open challenges in multilingual detoxification, including data scarcity, evaluation inconsistencies, cultural nuances and biases. Overall, we produce a needed overview of the state of multi-lingual toxicity detection and mitigation on which the community can ground to build globally safe and equitable LLMs.

1 Introduction

Large Language Models (LLMs) are increasingly used in multilingual settings, powering applications ranging from multilingual chatbots to cross-lingual content moderation (de Wynter et al., 2025; Hartvigsen et al., 2022; Kim et al., 2025). However, as their use spreads, so too do potential harms: LLMs often produce or amplify toxic content—hate speech, bias, harassment—that becomes especially problematic when crossing linguistic and cultural boundaries (Röttger et al., 2021; Sharma and Bhalla, 2025; Deshpande et al., 2023a). Despite progress on monolingual detoxification (especially English), there remain significant research gaps in mitigating toxicity in multilingual LLMs (Beniwal et al., 2025a; Tiţa and Zubiaga, 2021; Dementieva et al., 2024a; Logacheva et al., 2022; de Wynter et al., 2025).

The Complexity of Multilingual Toxicity. Multilingual detoxification is not merely a translation of English safety protocols (Neplenbroek et al., 2025; Kumar et al., 2025a). Instead, it requires navigating a complex anatomy of harm that varies across cultures. Toxicity manifests along a spectrum from *overt* categories—such as slurs, explicit insults, and profanity—to *implicit* forms like microaggressions, sarcasm, and toxic condescension, which are significantly harder to detect and mitigate (Lin and Li, 2025; Wen et al., 2023). This complexity is compounded by cultural dissonance: expressions that are benign in one linguistic context may be offensive in another, and rigid, English-centric alignment norms often fail to distinguish between actual hate speech and reclaimed slurs or dialectal variations. Furthermore, multilingual environments introduce unique linguistic vulnerabilities (Kivlichan et al., 2021). For example, code-switching (e.g., Hinglish) and mixed-script inputs frequently bypass standard safety filters entirely. As Gehman et al. (2020) demonstrated, pretrained models can degenerate into toxic text even from seemingly innocuous prompts, amplifying biases present in their training data, illustrating why robust, culturally aware mitigation is essential (Vongpradit et al., 2024; Dammu et al., 2024; Yu et al., 2023).

Failures of Current Approaches. Traditional content moderation relied on keyword-based filtering and rule-based systems, which are inadequate for capturing the nuanced, context-dependent nature of toxic language (Kim et al., 2025; Huang, 2025). Early machine learning approaches involved handcrafted features that often failed to generalize,

Findings of the Association for Computational Linguistics: ACL 2026, pages 35761–35774

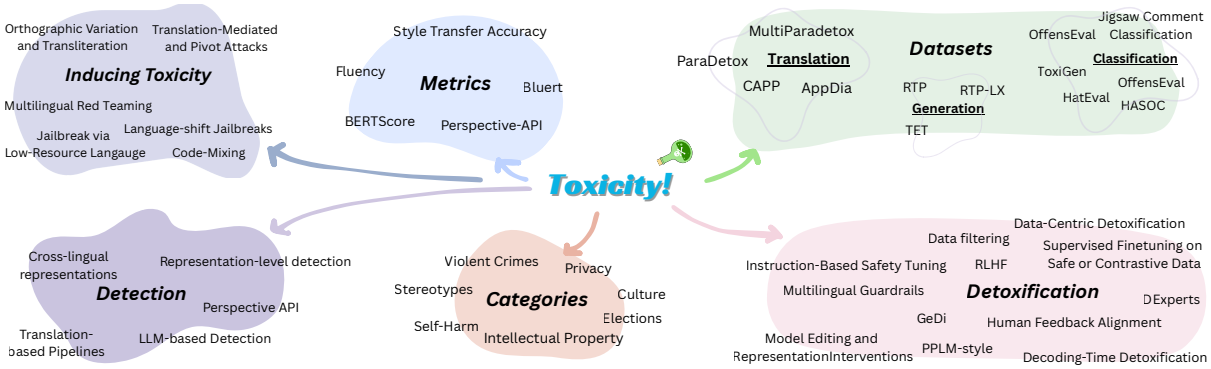

Figure 1: A brief taxonomy of toxicity in mLLMs.

easily being circumvented by simple lexical variations (Beniwal et al., 2025b; Jain et al., 2024). The emergence of large-scale LLMs introduces new challenges. Increasing model size alone does not inherently improve the ability to follow user intent; without alignment, larger models can simply generate more convincing untruthful or toxic outputs (Lu et al., 2025; Li et al., 2024b; Lee et al., 2024). Moreover, current alignment techniques like Reinforcement Learning from Human Feedback (RLHF) are often English-centric (Park and Rudzicz, 2022; Chaudhary et al., 2024). While these methods may reduce English toxicity, residual toxicity persists in underrepresented languages (Uppaal et al., 2025; Li et al., 2024c). Consequently, existing mechanisms may fail to recognize slurs in languages like Amharic due to bias in the training data (Costa-jussà et al., 2023). Such failures underscore that multilingual detoxification is a matter of technical robustness, global security, and social equity (Adragna et al., 2020; Cecchini et al., 2024). This survey provides the first comprehensive overview of detoxification methods for multilingual LLMs. We synthesize recent advances in detecting and mitigating toxicity, offering a detailed taxonomy of datasets, methodologies, and evaluation frameworks as shown in Figure 1.

Theme The survey includes the following key themes:

- Multilingual threat models covering language-shift jailbreaks, pivot/translation attacks, code-switch prompts, transliteration, multilingual red-teaming, and adaptation-time safety collapse—due to cross-lingual, low-resource fine-tuning.
- Task settings organized into three forms—***(i)*** toxic to neutral rewriting, ***(ii)*** toxicity classification, ***(iii)*** toxic versus neutral generation/ prompt continuations, along with evaluation metrics.
- Multilingual Toxicity Detection: Multilingual Encoder and Decoder-based detectors, Translation-based Pipelines, and Representation Level Detection.
- Detoxification Taxonomy: data-centric and model-centric detoxification, constrained decoding, and representation steering and multilingual guardrails.

We conclude with a discussion of open challenges and a forward-looking research plan for multilingual detoxification.

2 Threat Models for Inducing Toxicity in Multilingual LLMs

We focus on *multilingual-specific* toxicity-inducing threat models, which are adversarial procedures that exploit language choice, cross-lingual transfer, or multilingual interaction structure to elicit toxic generations from a safety-aligned model.

2.1 Prompt-Space Multilingual Attacks

Language-Shift Jailbreaks. A central multilingual threat is that simply re-expressing malicious intent in a non-English language can bypass English-centric refusal behavior. Deng et al. formalize (i) *unintentional* multilingual jailbreaks (benign users prompting in underrepresented languages) and (ii) *intentional* multilingual jailbreaks (adversaries combining multilingual prompts with explicit malicious instructions), and show substantially higher unsafe rates in lower-resource languages.

Translation-Mediated and Pivot attacks. translates an unsafe English prompt into a target low-resource language to increase compliance, then

translates the response back. Shen et al. (2024) empirically demonstrate higher unsafe response rates for malicious prompts expressed in lower-resource languages, motivating translation/pivot-based red-teaming. Recent defenses that *re-anchor* safety using English while enforcing target-language outputs further underscore translation as a core failure mode in multilingual safety (Zhang et al., 2025).

Language Mixing: Code-Switching and Multi-Language Mixtures. Multilingual prompts are often mixed within a single context. Yoo et al. (2025) shows that code-switched red-teaming queries can elicit unsafe behavior more effectively than monolingual attacks and introduces a synthesis framework (CSRT) to generate such queries at scale. Complementarily, Upadhayay and Behzadan (2024) proposes the *Sandwich Attack*, a multi-language mixture prompt that interleaves benign and adversarial segments across languages to induce harmful completions in a black-box setting.

2.2 Multilingual Red Teaming

Red teaming operationalizes threat models by generating adversarial prompts and dialogues at scale. Foundational work establishes manual and LM-assisted red teaming methodologies (Perez et al., 2022; Zhuo et al., 2023). Recent multilingual extensions explicitly target the multilingual capability envelope: CSRT generates code-switched attacks (Yoo et al., 2025); Rainbow Teaming produces diverse open-ended adversarial prompts and has been replicated/extended for Polish as a concrete non-English safety stress test (Samvelyan et al., 2024; Krasnodębska et al., 2025); and MM-ART automates *multi-turn, multilingual* red teaming, showing vulnerability increases sharply with conversation length and is substantially underestimated by single-turn English evaluation (Singhania et al., 2025).

2.3 Post-Deployment Adaptation Attacks

Cross-lingual Fine-Tuning Attacks. Aligned multilingual models are frequently customized via SFT/PEFT after deployment, creating an adaptation-time attack surface. Poppi et al. (2025) show that fine-tuning on a small toxic dataset in *one* language can collapse safety across *other* languages (cross-lingual attack transfer). Their Safety Information Localization (SIL) analysis suggests safety-relevant parameters are partially language-agnostic, enabling sparse updates to induce multilingual failure.

Jailbreaks via New-Language learning. Even benign adaptation can be risky: Upadhayay and Behzadan (2025) show that LoRA fine-tuning to learn a low-resource language—without harmful data—can nonetheless degrade refusal behavior, implying that multilingual expansion itself can destabilize safety guarantees.

Multilingual detoxification methods should therefore be evaluated not only on monolingual English prompts, but under compositions of multilingual operators (translate/pivot, code-switch, mixture prompts, transliteration), multi-turn interaction, and post-deployment adaptation stress tests

3 Task Setup: Datasets and Metrics

3.1 Datasets

Toxicity datasets can broadly be categorized into three tasks:

Translation: Toxic → Neutral Rewriting ParaDetox (Logacheva et al., 2022) was one of the first parallel detoxification corpora in English, with around 12K pairs of which are manually annotated for style preservation (toxic → neutral), content preservation (BLEU, semantic similarity), fluency, and a joint score. Subsequently ,this was extended for other languages such as Russian (Dementieva et al., 2023a) and Hindi (Mukherjee et al., 2023), culminating in the MultiParadetox (Dementieva et al., 2024a) work, which extended to 25 languages (including the initial 9 English, Spanish, German, Chinese, Arabic, Hindi, Ukrainian, Russian, Amharic). Recently, Moskovskiy et al. (2025, 2024) presents a large synthetic parallel corpus over English, German, French, Spanish, and Russian generated via few-shot LLM prompting. Dementieva et al. (2024b) presented parallel detoxification data for 9 langauges (English, Spanish, German, Chinese, Arabic, Hindi, Ukrainian, Russian, Amharic) at the PAN 2024 shared task. Relatedly, APPDIA (Atwell et al., 2022) and CAPP (Som et al., 2024) provide parallel datasets for offensive/inoffensive Reddit (evaluated via a Safe Score) and paraphrased dialogue data (evaluated via Perspective API), respectively.

Classification: Toxic Text Detection Jigsaw[1] (Kivlichan et al., 2021) was introduced as a Kaggle competition as an English multi-label classification (toxic, severe toxic, obscene, threat, insult,

[1] https://www.kaggle.com/c/jigsaw-unintended-bias-in-toxicity-classification

identity-hate) with ground truth crowd-sourced labels, which were later extended to a multilingual classification of comments in Chinese, French, Spanish, German, Russian, Turkish (plus a test set in Bulgarian) for binary toxicity. Additionally, on Twitter data, there have been offensive tweet detection datasets as parts of SemEval tasks for English OffensEval (Zampieri et al., 2019) and its multilingual extension to Arabic, Danish, English, Greek, Turkish (Zampieri et al., 2020), and task extension to span detection (Pavlopoulos et al., 2021). Furthermore, several Hate Speech Detection datasets are available for English (HateCheck)(Röttger et al., 2021) and multilingual versions (Multilingual Hate Check, HatEval, HASOC (Röttger et al., 2022; Mandl et al., 2019; Basile et al., 2019)) that cover diverse languages and fine-grained categories within hate speech. Lifetox (Kim et al., 2024) is a dataset designed to identify implicit toxicity within a broad range of advice-seeking scenarios. In the context of LLMs, such classification datasets have been used for evaluation (Meng et al., 2024) but also for detoxification via retrieval (Pozzobon et al., 2023). Finally, ToxiGen (Hartvigsen et al., 2022) provides a large GPT-generated toxic versus benign statements about protected groups, and shows fine-tuning classifiers on ToxiGen greatly improve their detection accuracy.

Generation: Non toxic Text Continuation RealToxicityPrompts (RTP) (Gehman et al., 2020) provides an English prompt and response dataset (100K web prompts), paired with toxicity scores from Perspective API. The dataset assesses "neural toxic degeneration" by evaluating the likelihood of LLMs producing toxic continuations from both toxic and non-toxic prompts, evaluated via Expected Maximum Toxicity (EMT) and toxicity probability metrics. Expanding upon RTP, RTP-LX (de Wynter et al., 2025) is a multilingual benchmark encompassing 38 languages, including dialect-specific variations. It features human-transcreated prompts and completions, annotated by native speakers across various harm categories like bias, insult, and identity attack. This benchmark evaluates LLMs' ability to detect culturally nuanced toxic content, revealing challenges in model alignment with human judgments, especially in context-dependent scenarios. PTP (Jain et al., 2024) offers a large-scale, multilingual evaluation framework with 425,000 naturally occurring prompts across 17 languages. Unlike RTP-LX, which relies on translations, PTP sources real-world web text to better capture linguistic and cultural nuances. Evaluations using PTP have shown that LLM toxicity tends to increase with model size and decrease in language resource availability. While instruction and preference-tuning methods can mitigate toxicity, the choice of tuning method has limited impact. Separately, Brun and Nikoulina (2024) focuses on 50,000 naturally occurring French prompts and their continuations, annotated for toxicity. TET (Luong et al., 2024) is a benchmark designed to expose latent toxic behaviors in LLMs by using prompts crafted to bypass safety mechanisms. It comprises 2,546 prompts filtered from over 1 million real-world interactions with 25 different LLMs, sourced from the Chat-Lmsys-1M dataset and can elicit more toxic responses from LLMs compared to using ToxiGen (Hartvigsen et al., 2022) as prompts, a modified setup originally used in Orca (Mukherjee et al., 2023) (by posing a question, provide seven sentences in the dataset, and then prompt the model to answer in a style similar to those provided sentences). Deshpande et al. (2023b) demonstrated that using a persona in the system prompt of ChatGPT can elicit even more toxic responses from RTP prompts.

3.2 Metrics

Toxicity Detection Metrics Outputs are often scored by toxicity classifiers. One can measure the ***style transfer accuracy (STA)***: the fraction of outputs that a classifier deems non-toxic. For example, models use RoBERTa-based classifiers trained on Jigsaw to compute STA (Dementieva et al., 2023b). Other tools like ***Perspective API***[2] can be used to score toxicity on a continuous scale. A successful detoxification should yield high STA or low toxicity scores (e.g. a drop in toxic-generation probability as in (Li et al., 2024c)).

Content Preservation and Fluency To ensure meaning is retained, similarity metrics are applied. Popular choices include ***BLEURT*** (Sellam et al., 2020) or ***BERTScore*** to compare the detoxified output to the input (or reference). Dementieva et al. (2023a) adopt BLEURT for English content similarity (SIM) and LaBSE embeddings for Russian. ***Fluency*** is evaluated by the percentage of grammatical or fluent sentences, often via a language acceptability classifier (e.g., a RoBERTa trained to recognize acceptability) (Dementieva et al., 2023a;

[2] https://perspectiveapi.com/

Logacheva et al., 2022). Combined metrics like the product of STA, SIM, and fluency are sometimes used to rank models.

4 Detection

Detecting toxicity in multilingual settings presents unique challenges due to linguistic diversity, morphology, code-mixing, dialectal variation, and culturally dependent definitions of harm.

4.1 Multilingual Transformers

Subsequent advancements leveraged deep contextual encoders such as mBERT and XLM-R, demonstrating that cross-lingual representations substantially improve toxicity identification in low-resource languages (Tiţa and Zubiaga, 2021). These models benefit from shared subword vocabularies and multilingual training corpora, allowing for knowledge transfer from high-resource languages, such as English, to typologically distant languages. Nevertheless, studies have shown that performance remains uneven across languages with distinct scripts or limited pretraining resources, indicating persistent representational biases (Kanjirangat et al., 2025). The brittleness of subword tokenization under spelling variants, obfuscation, and script mixing, requires byte/character-level modeling. The next-generation Perspective API centers a single multilingual token-free Charformer for toxic content detection (Lees et al., 2022).

4.2 Translation-based Pipelines

A parallel line of work explores translation-based pipelines, where toxic text in non-English languages is machine-translated into English before being applied to a high-performing English toxicity classifier (Bell et al., 2025). While this strategy often yields higher accuracy due to the maturity of English-based detectors, it raises concerns regarding error propagation, translation artifacts, and semantic drift—issues that disproportionately affect dialectal, code-mixed, and morphologically complex languages (Zampieri et al., 2020). Translation systems themselves can introduce or obscure toxic content, posing additional ethical and methodological limitations.

4.3 Representation-level Detection

Recent research investigates representation-level detection (Duan et al., 2025; Wang et al., 2021), analyzing how toxicity is encoded within latent embedding spaces across languages. Such studies reveal that toxicity-related signals cluster consistently across multilingual embedding spaces (Shaik et al., 2025; Conneau et al., 2020), suggesting shared semantic dimensions of harmful content despite linguistic variation. These findings further motivate embedding-based probing and neuron-level attribution techniques that seek to identify how and where toxicity features are stored within multilingual models (Li et al., 2024c).

4.4 LLM based Detection

Finally, the emergence of instruction-tuned LLMs has broadened the detection landscape (Hu et al., 2024). Several works evaluate LLMs as zero-shot or few-shot toxicity detectors, demonstrating strong generalization but also highlighting calibration failures (Liu et al.) and cultural misalignment across languages (Yang et al., 2025). These models often rely on implicit safety priors learned during the alignment process, which can lead to inconsistent behavior when handling region-specific sociolinguistic norms (Li et al., 2024a).

Takeaway: Collectively, the literature indicates that while multilingual LLMs have greatly improved cross-lingual toxicity detection, substantial challenges remain. Persistent gaps in language coverage, bias in training corpora, inconsistent cross-lingual performance, and translation-induced errors limit the reliability of current detectors.

5 Detoxification

5.1 Data-Centric Detoxification

Data-centric detoxification focuses on enhancing the quality of pre-training and fine-tuning corpora by removing toxic or harmful content. Early large-scale filtering pipelines relied on blocklists or lexical heuristics; however, contemporary approaches utilize multilingual classifiers to remove toxic spans prior to training (Gehman et al., 2020; Park and Rudzicz, 2022; Kreutzer et al., 2022). More recent work emphasizes bias-aware filtering to avoid inadvertently suppressing dialectal or marginalized speech (Sap et al., 2022; Jaggi et al., 2024). In multilingual settings, filtering heavily relies on the cross-lingual generalization of toxicity detectors (e.g., XLM-R), which can misclassify culturally specific idioms or reclaimed slurs (Bensalem et al., 2024; Welbl et al., 2021). Data filtering is highly scalable and effective for reducing systemic toxicity; however, it risks cultural misalignment, uneven performance across languages, and

the over-removal of minority language varieties: Stranisci and Hardmeier (2025); Park and Rudzicz (2022) show that there is a tradeoff of the positive impact of filtering strategies in reducing harmful content with increasing the underrepresentation of vulnerable groups.

5.2 Model-Centric Detoxification

Supervised Finetuning on Safe or Contrastive Data Supervised detoxification approaches finetune LLMs on curated non-toxic corpora or contrastive toxic–neutral pairs (Hawkins et al., 2024). (Neplenbroek et al., 2025) shows that finetuning on curated non-harmful text reduce general biases, but find only direct preference optimization to be effective for mitigating toxicity. The mitigation caused by applying these methods in English also transfers to non-English languages and the extent to which transfer takes place can be predicted by the amount of data in a given language present in the model's pretraining data. However, this transfer of bias and toxicity mitigation often comes at the expense of decreased language generation ability in non-English languages, highlighting the importance of developing language-specific bias and toxicity mitigation methods. Finetuning-based detoxification provides strong control and good multilingual generalization, but may reduce model diversity or creativity, and can introduce unintended stylistic flattening (Wang et al., 2022).

Instruction-Based Safety Tuning Instruction tuning using curated safety datasets or synthetic refusal-style instructions (Zhao et al., 2025; Li et al., 2024b) enhances multilingual LLMs' ability to decline harmful requests and avoid toxic continuations. Recent multilingual safety instruction sets (e.g. PolyGuard) (Kumar et al., 2025b) significantly improve cross-lingual robustness. These methods scale well and align closely with practical deployment requirements, though annotation biases remain a persistent limitation.

RLHF and Human Feedback Alignment Reinforcement learning from human feedback (RLHF) (Ouyang and et al., 2022; Bai and et al., 2022) is a foundational alignment technique that improves safety by training reward models to penalize toxic or harmful outputs. While RLHF datasets are primarily English-centric, multilingual LLMs benefit indirectly through shared parameters and cross-lingual transfer (Dang et al., 2024). However, reliance on English safety norms introduces cross-cultural misalignment in multilingual models (Lu et al., 2025), especially for expressions that are offensive in some cultures but neutral in others.

5.3 Decoding-Time Detoxification

Post-hoc methods avoid (or minimize) retraining by steering generation at inference (Ko et al., 2024). Classifier-guided and expert-based logit steering include PPLM-style hidden-state perturbations (Pascual et al., 2021), GeDi-style generative discriminators (Krause et al., 2021), and expert/anti-expert mixture decoding (e.g., DExperts (Liu et al., 2021)). Expert steering is modular and easily extended across languages, though availability of high-quality multilingual experts remains a bottleneck. A second family uses *edit-after-generate*: produce a candidate, detect toxicity, and rewrite/refine (often via prompting or a specialized editor) (Leong et al.). In multilingual deployments, *translation-pivot pipelines* remain common (translate→detox in English→translate back), but they risk semantic drift and can erase culturally salient pragmatics; task submissions increasingly combine pivoting with LLM-based paraphrasing and reranking. Retrieval augmentation can also support detox by grounding rewrites in policy examples or safe templates.

5.4 Model Editing and Representation Interventions

Recent work investigates activation steering—modifying internal LM representations to remove or attenuate toxic features (Goyal et al., 2025). SemSteer (Turner et al., 2024), ROME-based editing (Meng and et al., 2022), and directional activation steering (Klerings et al., 2025) identify interpretable semantic directions that can be suppressed during generation. Because semantic directions often cluster cross-lingually, activation steering has shown early promise in multilingual detoxification (Sundar et al., 2025). Editing offers targeted, potentially efficient detoxification, but multilingual evaluation is still nascent, and regression risk is high without careful cross-lingual audits (Wang et al., 2024a).

5.5 Multilingual Guardrails

Related (but not the focus of this survey) is post-hoc moderation independent of detoxifying the generator, i.e., multilingual guardrails (Yi et al.): deployment-time controllers that classify/gate

prompts and responses into policy categories (including prompt harmfulness, response harmfulness, and refusal/compliance), under potentially adversarial multilingual prompts (language choice, code-switching, transliteration/orthographic variants) that routinely break English-centric safeguards . The key multilingual guardrails are Llama Guard (Inan et al., 2023), Aegis (Ghosh et al., 2024), MrGuard (Yang et al., 2025), WildGuard (Han et al., 2024), PolyGuard (Kumar et al., 2025b), MultiGuard (Verma et al., 2025), CREST (Bansal and Mishra, 2025) and UnityAI-Guard (Beniwal et al., 2025b).

Key Takeaways We present the main takeaways as:

- Cross-lingual robustness remains the main challenge: Detoxification methods consistently perform better in high-resource languages than in low-resource or morphologically rich languages.
- Cultural bias persists across detoxification pipelines: Many "safe" signals originate from English, causing misalignment in non-Western contexts.
- Hybrid strategies are most effective: Combining data filtering, controlled decoding, and alignment-based tuning yields the most stable multilingual detoxification outcomes.
- Avoiding over-censorship is an unresolved issue: Techniques often suppress legitimate emotional or dialectal expressions, leading to "model homogenization."

6 Discussion and Open Challenges

6.1 Cross-Lingual Gaps in Detoxification

A key bottleneck across multilingual safety pipelines is the persistent performance disparity between high-resource and low-resource languages. Even models explicitly designed for multilingual toxicity detection rely largely on English-centric representations or annotations (Gehman et al., 2020; Kreutzer et al., 2022). As a result, toxic content in morphologically rich or culturally distant languages—such as Arabic dialects, Hindi, Swahili, or Tagalog—remains systematically under-detected (Bensalem et al., 2024). Furthermore, alignment methods such as RLHF or constitutional tuning predominantly use English preference data (Ouyang and et al., 2022; Bai and et al., 2022), resulting in safety behaviors learned in English being projected unevenly across languages. This leads to inconsistent refusal behavior, over-sensitivity to benign expressions, or failure to recognize toxic slang in non-English languages (Lu et al., 2025).

Open Challenge: Developing culturally aware, balanced, multilingual safety representations that scale to low-resource languages without English overdominance remains an unsolved and essential research agenda.

6.2 Cultural and Normative Misalignment

Toxicity is inherently contextual and culturally embedded. While several studies demonstrate that annotators' identities strongly influence toxicity judgments (Sap et al., 2022, 2019; Jaggi et al., 2024), most multilingual LLM safety datasets still assume a monolithic perspective on harmfulness. Models trained under such regimes risk over-censoring reclaimed slurs, misclassifying dialectal expressions, or even reinforcing majority-group norms, resulting in cultural erasure or linguistic homogenization (Shen et al., 2024). As multilingual LLMs expand globally, these mismatches become increasingly problematic. Languages with rich honorific systems, code-switching norms, or culturally specific humor (Li et al., 2024a), expose current models' inability to differentiate toxicity from socially sanctioned expressions.

Open Challenge: Future systems need culturally grounded, community-driven annotation processes and context-aware toxicity modeling that respects sociolinguistic diversity.

6.3 Lack of Robust, Multilingual Evaluation Frameworks

A recurring theme in top-tier work is the lack of standardized, multilingual frameworks for evaluating toxicity. Existing benchmarks—e.g., RealToxicityPrompts (Gehman et al., 2020) — are predominantly English-centric, while recent multilingual datasets, such as PolyGuard (Kumar et al., 2025b), cover only a limited number of languages and domains. Further, evaluation pipelines struggle to assess subtle forms of toxicity, such as microaggressions, presuppositional harm, or implicit bias (Sap et al., 2022). Cross-lingual transfer of toxicity classifiers (e.g., XLM-R) often produces false positives for dialects or false negatives for low-resource slang, making evaluation unreliable.

Open Challenge: Building comprehensive multilingual benchmarks with fine-grained categories

of toxicity, robust cross-cultural annotations, and shared evaluation protocols is urgently needed (Wang et al., 2024b).

6.4 Over-Suppression and Style Degradation

Detoxification techniques, particularly contrastive finetuning and representation editing, often reduce linguistic richness or stylistic diversity. Prior work demonstrates that detoxified models produce more generic, less expressive, and overly formal outputs (Welbl et al., 2021; Liu et al., 2021). In multilingual settings, this effect is amplified: for low-resource languages, detoxification can inadvertently erase dialectal identity, flatten morphology, or default to English-like syntax due to shared subword vocabularies and cross-lingual interference (Zhao et al., 2025). Techniques such as activation editing (Turner et al., 2024) and PPLM (Pascual et al., 2021) offer fine-grained control but still risk semantic oversuppression when applied cross-lingually.

Open Challenge: Designing detoxification techniques that preserve stylistic and cultural characteristics while eliminating harmful content remains an open frontier.

6.5 Handling Code-Switching and Mixed-Linguistic Toxicity

Multilingual communities frequently communicate through code-switching (e.g., Hinglish, Arabizi, Spanglish), which combines scripts, phonetic spellings, and culturally specific expressions. However, current toxicity detectors consistently misclassify code-switched text due to a lack of training coverage and inconsistent tokenization (Zhang et al.; Bensalem et al., 2024). Moreover, LLMs often fail to recognize toxicity embedded in mixed-language slang or transliterations (e.g., Hindi profanity in Roman script) (Sharma and Bhalla, 2025). This poses serious risks for global deployments of multilingual LLMs.

Open Challenge: Robust multilingual safety systems must explicitly account for code-switching and orthographic variation. This may involve training models on code-mixed corpora, utilizing unified subword tokenizers that handle mixed scripts, and developing language-agnostic features for toxic content. Approaches like transliteration mapping, code-switch data augmentation, or ensemble detectors (one per language, plus a fusion model) is important for mixed-linguistic detoxification.

6.6 Key Takeaways

- Language Disparities: Toxicity mitigation remains uneven across languages – methods that succeed in English often underperform in low-resource languages and dialects, leaving some communities with higher exposure to toxic outputs.
- Cultural Context Matters: One-size-fits-all safety tuning leads to misalignment with local norms. Models may over-censor benign cultural expressions or miss contextually offensive language, highlighting the need for culturally grounded curation of toxic content.
- Evaluation Gaps: The field lacks standardized multilingual benchmarks. Fragmented evaluation protocols make it challenging to compare systems, especially for subtle toxicity, underscoring the need for shared frameworks and diverse testbeds.
- Style and Expression Trade-offs: Many detoxification techniques inadvertently degrade output quality or diversity, yielding robotic text. Preserving linguistic richness and user intent while filtering toxicity is important.
- Hybrid approaches combining multilingual data filtering, language-specific detoxifiers, controlled generation, and culturally aware alignment - yield more stable and globally robust detoxification than any single method.
- Need for Interpretability: Understanding why a model flags or generates content as toxic is crucial. Better interpretability in multilingual settings would facilitate more principled and transparent detoxification methods.

7 Conclusion

Multilingual detoxification should be treated as a first-class safety problem rather than an English-centric add-on: toxicity and refusal behavior vary sharply under language shift, translation pivots, code-switching, transliteration, and post-deployment adaptation. This survey systematized the space along (i) multilingual threat models, (ii) tasks spanning rewriting, classification, and toxic-generation , and (iii) a mechanism-based taxonomy of mitigation. Across methods, the dominant bottlenecks are cross-lingual coverage gaps, culturally contingent notions of harm, and over-suppression of legitimate dialectal or identity-related language.

Limitations

This survey synthesizes a fast-moving literature, so specific model families, benchmarks, and best practices may evolve after publication. The evidence base is also uneven across languages: many "multilingual" studies still emphasize high-resource languages, with fewer results for low-resource languages, dialect continua, and code-mixed or transliterated text. Because toxicity definitions and label schemas vary across datasets and cultures, some comparisons across papers are necessarily approximate. Finally, many evaluations rely on automatic detectors, translation-based protocols, or closed-model assessments, which can introduce measurement noise and limit strict apples-to-apples replication.

Ethics

This survey reviews prior work on toxicity in multilingual language models and does not involve new data collection or model deployment. However, existing approaches often reflect English-centric norms and may misclassify or suppress culturally specific or reclaimed expressions in other languages. Automated toxicity detection and detoxification can therefore reinforce societal and annotator biases or lead to over-censorship. We highlight the importance of culturally aware evaluation, inclusive data practices, and careful consideration of safety–utility trade-offs in multilingual settings.

References

Robert Adragna, Elliot Creager, David Madras, and Richard Zemel. 2020. Fairness and robustness in invariant learning: A case study in toxicity classification. *arXiv preprint arXiv:2011.06485*.

Katherine Atwell, Sabit Hassan, and Malihe Alikhani. 2022. Appdia: A discourse-aware transformer-based style transfer model for offensive social media conversations. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 6063–6074.

Yuntao Bai and et al. 2022. Constitutional ai: Harmlessness from ai feedback. In *NeurIPS*.

Lavish Bansal and Naman Mishra. 2025. Crest: Universal safety guardrails through cluster-guided cross-lingual transfer. *arXiv preprint arXiv:2512.02711*.

Valerio Basile, Cristina Bosco, Elisabetta Fersini, Debora Nozza, Viviana Patti, Francisco Manuel Rangel Pardo, Paolo Rosso, and Manuela Sanguinetti. 2019. Semeval-2019 task 5: Multilingual detection of hate speech against immigrants and women in twitter. In *Proceedings of the 13th international workshop on semantic evaluation*, pages 54–63.

Samuel Bell, Eduardo Sánchez, David Dale, Pontus Stenetorp, Mikel Artetxe, and Marta R Costa-jussà. 2025. Translate, then detect: Leveraging machine translation for cross-lingual toxicity classification. In *Proceedings of the Tenth Conference on Machine Translation*, pages 253–268.

Himanshu Beniwal, Youngwoo Kim, Maarten Sap, Soham Dan, and Thomas Hartvigsen. 2025a. Breaking mbad! supervised fine-tuning for cross-lingual detoxification. *Preprint*, arXiv:2505.16722.

Himanshu Beniwal, Reddybathuni Venkat, Rohit Kumar, Birudugadda Srivibhav, Daksh Jain, Pavan Doddi, Eshwar Dhande, Adithya Ananth, Kuldeep, and Mayank Singh. 2025b. Unityai-guard: Pioneering toxicity detection across low-resource indian languages. *Preprint*, arXiv:2503.23088.

Imene Bensalem, Paolo Rosso, and Hanane Zitouni. 2024. Toxic language detection: a systematic review of arabic datasets. *Preprint*, arXiv:2312.07228.

Caroline Brun and Vassilina Nikoulina. 2024. Frenchtoxicityprompts: a large benchmark for evaluating and mitigating toxicity in french texts. In *Proceedings of the Fourth Workshop on Threat, Aggression & Cyberbullying@ LREC-COLING-2024*, pages 105–114.

David Cecchini, Arshaan Nazir, Kalyan Chakravarthy, and Veysel Kocaman. 2024. Holistic evaluation of large language models: Assessing robustness, accuracy, and toxicity for real-world applications. In *Proceedings of the 4th Workshop on Trustworthy Natural Language Processing (TrustNLP 2024)*, pages 109–117.

Sapana Chaudhary, Ujwal Dinesha, Dileep Kalathil, and Srinivas Shakkottai. 2024. Risk-averse fine-tuning of large language models. *Advances in Neural Information Processing Systems*, 37:107003–107038.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th annual meeting of the association for computational linguistics*, pages 8440–8451.

Marta Costa-jussà, Eric Smith, Christophe Ropers, Daniel Licht, Jean Maillard, Javier Ferrando, and Carlos Escolano. 2023. Toxicity in multilingual machine translation at scale. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 9570–9586.

Preetam Prabhu Srikar Dammu, Hayoung Jung, Anjali Singh, Monojit Choudhury, and Tanu Mitra. 2024. "they are uncultured": Unveiling covert harms and

social threats in llm generated conversations. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 20339–20369.

John Dang, Arash Ahmadian, Kelly Marchisio, Julia Kreutzer, Ahmet Üstün, and Sara Hooker. 2024. RLHF can speak many languages: Unlocking multilingual preference optimization for LLMs. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 13134–13156, Miami, Florida, USA. Association for Computational Linguistics.

Adrian de Wynter, Ishaan Watts, Tua Wongsangaroonsri, Minghui Zhang, Noura Farra, Nektar Ege Altıntoprak, Lena Baur, Samantha Claudet, Pavel Gajdušek, Qilong Gu, Anna Kaminska, Tomasz Kaminski, Ruby Kuo, Akiko Kyuba, Jongho Lee, Kartik Mathur, Petter Merok, Ivana Milovanović, Nani Paananen, and 13 others. 2025. Rtp-lx: Can llms evaluate toxicity in multilingual scenarios? *Proceedings of the AAAI Conference on Artificial Intelligence*, 39(27):27940–27950.

Daryna Dementieva, Nikolay Babakov, and Alexander Panchenko. 2024a. MultiParaDetox: Extending text detoxification with parallel data to new languages. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 2: Short Papers)*, pages 124–140, Mexico City, Mexico. Association for Computational Linguistics.

Daryna Dementieva, Daniil Moskovskiy, Nikolay Babakov, Abinew Ali Ayele, Naquee Rizwan, Frolian Schneider, Xintog Wang, Seid Muhie Yimam, Dmitry Ustalov, Elisei Stakovskii, and 1 others. 2024b. Overview of the multilingual text detoxification task at pan 2024. *Working Notes of CLEF*.

Daryna Dementieva, Daniil Moskovskiy, David Dale, and Alexander Panchenko. 2023a. Exploring methods for cross-lingual text style transfer: The case of text detoxification. In *Proceedings of the 13th International Joint Conference on Natural Language Processing and the 3rd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1083–1101, Nusa Dua, Bali. Association for Computational Linguistics.

Daryna Dementieva, Daniil Moskovskiy, David Dale, and Alexander Panchenko. 2023b. Exploring methods for cross-lingual text style transfer: The case of text detoxification. In *Proceedings of the 13th International Joint Conference on Natural Language Processing and the 3rd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1083–1101, Nusa Dua, Bali. Association for Computational Linguistics.

Yue Deng, Wenxuan Zhang, Sinno Jialin Pan, and Lidong Bing. Multilingual jailbreak challenges in large language models. In *The Twelfth International Conference on Learning Representations*.

Ameet Deshpande, Vishvak Murahari, Tanmay Rajpurohit, Ashwin Kalyan, and Karthik Narasimhan. 2023a. Toxicity in chatgpt: Analyzing persona-assigned language models. In *Findings of EMNLP 2023*.

Ameet Deshpande, Vishvak Murahari, Tanmay Rajpurohit, Ashwin Kalyan, and Karthik Narasimhan. 2023b. Toxicity in chatgpt: Analyzing persona-assigned language models. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 1236–1270, Singapore. Association for Computational Linguistics.

Zenghao Duan, Zhiyi Yin, Zhichao Shi, Liang Pang, Shaoling Jing, Jiayi Wu, Yu Yan, Huawei Shen, and Xueqi Cheng. 2025. Gloss over toxicity: Understanding and mitigating toxicity in llms via global toxic subspace. *arXiv preprint arXiv:2505.17078*.

Samuel Gehman, Suchin Gururangan, Maarten Sap, Yejin Choi, and Noah A. Smith. 2020. Realtoxicityprompts: Evaluating neural toxic degeneration in language models. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 3356–3369.

Shaona Ghosh, Prasoon Varshney, Erick Galinkin, and Christopher Parisien. 2024. Aegis: Online adaptive ai content safety moderation with ensemble of llm experts. *Preprint*, arXiv:2404.05993.

Agam Goyal, Vedant Rathi, William Yeh, Yian Wang, Yuen Chen, and Hari Sundaram. 2025. Breaking bad tokens: Detoxification of LLMs using sparse autoencoders. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 12691–12709, Suzhou, China. Association for Computational Linguistics.

Seungju Han, Kavel Rao, Allyson Ettinger, Liwei Jiang, Bill Yuchen Lin, Nathan Lambert, Yejin Choi, and Nouha Dziri. 2024. Wildguard: Open one-stop moderation tools for safety risks, jailbreaks, and refusals of llms. *Advances in Neural Information Processing Systems*, 37:8093–8131.

Thomas Hartvigsen, Saadia Gabriel, Hamid Palangi, Maarten Sap, Dipankar Ray, and Ece Kamar. 2022. Toxigen: A large-scale machine-generated dataset for adversarial and implicit hate speech detection. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3309–3326.

Will Hawkins, Brent Mittelstadt, and Chris Russell. 2024. The effect of fine-tuning on language model toxicity. *arXiv preprint arXiv:2410.15821*.

Zhanhao Hu, Julien Piet, Geng Zhao, Jiantao Jiao, and David Wagner. 2024. Toxicity detection for free. *Advances in Neural Information Processing Systems*, 37:17518–17540.

Tao Huang. 2025. Content moderation by llm: from accuracy to legitimacy. *Artificial Intelligence Review*, 58(10):320.

Hakan Inan, Kartikeya Upasani, Jianfeng Chi, Rashi Rungta, Krithika Iyer, Yuning Mao, Michael Tontchev, Qing Hu, Brian Fuller, Davide Testuggine, and 1 others. 2023. Llama guard: Llm-based input-output safeguard for human-ai conversations. *arXiv preprint arXiv:2312.06674*.

Harbani Jaggi, Kashyap Coimbatore Murali, Eve Fleisig, and Erdem Biyik. 2024. Accurate and data-efficient toxicity prediction when annotators disagree. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 21910–21917, Miami, Florida, USA. Association for Computational Linguistics.

Devansh Jain, Priyanshu Kumar, Samuel Gehman, Xuhui Zhou, Thomas Hartvigsen, and Maarten Sap. 2024. Polyglotoxicityprompts: Multilingual evaluation of neural toxic degeneration in large language models. *arXiv preprint*, arXiv:2405.09373. May 2024.

Vani Kanjirangat, Tanja Samardzic, Ljiljana Dolamic, and Fabio Rinaldi. 2025. Tokenization and representation biases in multilingual models on dialectal nlp tasks. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 24003–24021.

Minbeom Kim, Jahyun Koo, Hwanhee Lee, Joonsuk Park, Hwaran Lee, and Kyomin Jung. 2024. Lifetox: Unveiling implicit toxicity in life advice. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 2: Short Papers)*, pages 688–698.

Youngwoo Kim, Himanshu Beniwal, Steven L Johnson, and Thomas Hartvigsen. 2025. Decoding the rule book: Extracting hidden moderation criteria from reddit communities. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 20498–20509.

Ian Kivlichan, Jeffrey Sorensen, Lucas Dixon, Lucy Vasserman, Meghan Graham, Tin Acosta, and Walter Reade. 2021. Jigsaw rate severity of toxic comments. `https://kaggle.com/competitions/jigsaw-toxic-severity-rating`. Kaggle.

Alina Klerings, Jannik Brinkmann, Daniel Ruffinelli, and Simone Paolo Ponzetto. 2025. Steering language models in multi-token generation: A case study on tense and aspect. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 8621–8639, Suzhou, China. Association for Computational Linguistics.

Ching-Yun Ko, Pin-Yu Chen, Payel Das, Youssef Mroueh, Soham Dan, Georgios Kollias, Subhajit Chaudhury, Tejaswini Pedapati, and Luca Daniel. 2024. Large language models can be strong self-detoxifiers. *Preprint*, arXiv:2410.03818.

Aleksandra Krasnodębska, Maciej Chrabaszcz, and Wojciech Kusa. 2025. Rainbow-teaming for the polish language: A reproducibility study. In *Proceedings of the 5th Workshop on Trustworthy NLP (TrustNLP 2025)*, pages 155–165.

Ben Krause, Akhilesh Deepak Gotmare, Bryan McCann, Nitish Shirish Keskar, Shafiq Joty, Richard Socher, and Nazneen Fatema Rajani. 2021. Gedi: Generative discriminator guided sequence generation. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 4929–4952.

Julia Kreutzer, Isaac Caswell, Lisa Wang, Ahsan Wahab, Daan van Esch, Nasanbayar Ulzii-Orshikh, Allahsera Tapo, Nishant Subramani, Artem Sokolov, Claytone Sikasote, Monang Setyawan, Supheakmungkol Sarin, Sokhar Samb, Benoît Sagot, Clara Rivera, Annette Rios, Isabel Papadimitriou, Salomey Osei, Pedro Ortiz Suarez, and 33 others. 2022. Quality at a glance: An audit of web-crawled multilingual datasets. *Transactions of the Association for Computational Linguistics*, 10:50–72.

Priyanshu Kumar, Devansh Jain, Akhila Yerukola, Liwei Jiang, Himanshu Beniwal, Thomas Hartvigsen, and Maarten Sap. 2025a. Polyguard: A multilingual safety moderation tool for 17 languages. *arXiv preprint arXiv:2504.04377*.

Priyanshu Kumar, Devansh Jain, Akhila Yerukola, Liwei Jiang, Himanshu Beniwal, Thomas Hartvigsen, and Maarten Sap. 2025b. Polyguard: A multilingual safety moderation tool for 17 languages. *Preprint*, arXiv:2504.04377.

Andrew Lee, Xiaoyan Bai, Itamar Pres, Martin Wattenberg, Jonathan K Kummerfeld, and Rada Mihalcea. 2024. A mechanistic understanding of alignment algorithms: A case study on dpo and toxicity. In *International Conference on Machine Learning*, pages 26361–26378. PMLR.

Alyssa Lees, Vinh Q Tran, Yi Tay, Jeffrey Sorensen, Jai Gupta, Donald Metzler, and Lucy Vasserman. 2022. A new generation of perspective api: Efficient multilingual character-level transformers. In *Proceedings of the 28th ACM SIGKDD conference on knowledge discovery and data mining*, pages 3197–3207.

Chak Tou Leong, Yi Cheng, Jian Wang, Wenjie Li, and 1 others. Self-detoxifying language models via toxification reversal. In *The 2023 Conference on Empirical Methods in Natural Language Processing*.

Cheng Li, Mengzhuo Chen, Jindong Wang, Sunayana Sitaram, and Xing Xie. 2024a. Culturellm: Incorporating cultural differences into large language models. *Advances in Neural Information Processing Systems*, 37:84799–84838.

Chong Li, Wen Yang, Jiajun Zhang, Jinliang Lu, Shaonan Wang, and Chengqing Zong. 2024b. X-instruction: Aligning language model in low-resource languages with self-curated cross-lingual instructions. *arXiv preprint arXiv:2405.19744*.

Xiaochen Li, Zheng-Xin Yong, and Stephen H. Bach. 2024c. Preference tuning for toxicity mitigation generalizes across languages. *arXiv preprint*, arXiv:2406.16235. June 2024.

Xinru Lin and Luyang Li. 2025. Implicit bias in llms: A survey. *arXiv preprint arXiv:2503.02776*.

Alisa Liu, Maarten Sap, Ximing Lu, Swabha Swayamdipta, Chandra Bhagavatula, Noah A. Smith, and Yejin Choi. 2021. Dexperts: Decoding-time controlled text generation with experts and anti-experts. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics (ACL 2021)*, pages 1990–2001, Online. Association for Computational Linguistics. ACL 2021.

Hongfu Liu, Hengguan Huang, Xiangming Gu, Hao Wang, and Ye Wang. On calibration of llm-based guard models for reliable content moderation. In *The Thirteenth International Conference on Learning Representations*.

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, Dublin, Ireland. Association for Computational Linguistics.

Haoran Lu, Luyang Fang, Ruidong Zhang, Xinliang Li, Jiazhang Cai, Huimin Cheng, Lin Tang, Ziyu Liu, Zeliang Sun, Tao Wang, Yingchuan Zhang, Arif Hassan Zidan, Jinwen Xu, Jincheng Yu, Meizhi Yu, Hanqi Jiang, Xilin Gong, Weidi Luo, Bolun Sun, and 31 others. 2025. Alignment and safety in large language models: Safety mechanisms, training paradigms, and emerging challenges. *Preprint*, arXiv:2507.19672.

Tinh Luong, Thanh-Thien Le, Linh Ngo, and Thien Nguyen. 2024. Realistic evaluation of toxicity in large language models. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 1038–1047.

Thomas Mandl, Sandip Modha, Prasenjit Majumder, Daksh Patel, Mohana Dave, Chintak Mandlia, and Aditya Patel. 2019. Overview of the hasoc track at fire 2019: Hate speech and offensive content identification in indo-european languages. In *Proceedings of the 11th annual meeting of the Forum for Information Retrieval Evaluation*, pages 14–17.

Kevin Meng and et al. 2022. Locating and editing factual associations in gpt. In *NeurIPS*.

Tao Meng, Ninareh Mehrabi, Palash Goyal, Anil Ramakrishna, Aram Galstyan, Richard Zemel, Kai-Wei Chang, Rahul Gupta, and Charith Peris. 2024. Attribute controlled fine-tuning for large language models: A case study on detoxification.

Daniil Moskovskiy, Sergey Pletenev, and Alexander Panchenko. 2024. LLMs to replace crowdsourcing for parallel data creation? the case of text detoxification. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 14361–14373, Miami, Florida, USA. Association for Computational Linguistics.

Daniil Moskovskiy, Nikita Sushko, Sergey Pletenev, Elena Tutubalina, and Alexander Panchenko. 2025. Synthdetoxm: Modern llms are few-shot parallel detoxification data annotators. *arXiv preprint arXiv:2502.06394*.

Sourabrata Mukherjee, Akanksha Bansal, Atul Kr. Ojha, John P. McCrae, and Ondrej Dusek. 2023. Text detoxification as style transfer in English and Hindi. In *Proceedings of the 20th International Conference on Natural Language Processing (ICON)*, pages 133–144, Goa University, Goa, India. NLP Association of India (NLPAI).

Vera Neplenbroek, Arianna Bisazza, and Raquel Fernández. 2025. Cross-lingual transfer of debiasing and detoxification in multilingual llms: An extensive investigation. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 2805–2830.

Long Ouyang and et al. 2022. Training language models to follow instructions with human feedback. In *NeurIPS*.

Yoon A Park and Frank Rudzicz. 2022. Detoxifying language models with a toxic corpus. *Preprint*, arXiv:2205.00320.

Damian Pascual, Beni Egressy, Clara Meister, Ryan Cotterell, and Roger Wattenhofer. 2021. A plug-and-play method for controlled text generation. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 3973–3997, Punta Cana, Dominican Republic. Association for Computational Linguistics.

John Pavlopoulos, Jeffrey Sorensen, Léo Laugier, and Ion Androutsopoulos. 2021. SemEval-2021 task 5: Toxic spans detection. In *Proceedings of the 15th International Workshop on Semantic Evaluation (SemEval-2021)*, pages 59–69, Online. Association for Computational Linguistics.

Ethan Perez, Saffron Huang, Francis Song, Trevor Cai, Roman Ring, John Aslanides, Amelia Glaese, Nat McAleese, and Geoffrey Irving. 2022. Red teaming language models with language models. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 3419–3448.

Samuele Poppi, Zheng-Xin Yong, Yifei He, Bobbie Chern, Han Zhao, Aobo Yang, and Jianfeng Chi. 2025. Towards understanding the fragility of multilingual llms against fine-tuning attacks. In *Findings of the Association for Computational Linguistics: NAACL 2025*, pages 2358–2372.

Luiza Pozzobon, Beyza Ermis, Patrick Lewis, and Sara Hooker. 2023. Goodtriever: Adaptive toxicity mitigation with retrieval-augmented models. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 5108–5125, Singapore. Association for Computational Linguistics. EMNLP Findings 2023.

Paul Röttger, Haitham Seelawi, Debora Nozza, Zeerak Talat, and Bertie Vidgen. 2022. Multilingual hatecheck: Functional tests for multilingual hate speech detection models. In *Proceedings of the Sixth Workshop on Online Abuse and Harms (WOAH)*, pages 154–169.

Paul Röttger, Bertie Vidgen, Dong Nguyen, Zeerak Waseem, Helen Margetts, and Janet Pierrehumbert. 2021. Hatecheck: Functional tests for hate speech detection models. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 41–58.

Mikayel Samvelyan, Sharath C Raparthy, Andrei Lupu, Eric Hambro, Aram H Markosyan, Manish Bhatt, Yuning Mao, Minqi Jiang, Jack Parker-Holder, Jakob Foerster, and 1 others. 2024. Rainbow teaming: Open-ended generation of diverse adversarial prompts. *Advances in Neural Information Processing Systems*, 37:69747–69786.

Maarten Sap, Dallas Card, Saadia Gabriel, Yejin Choi, and Noah A. Smith. 2019. The risk of racial bias in hate speech detection. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1668–1678, Florence, Italy. Association for Computational Linguistics.

Maarten Sap, Swabha Swayamdipta, Laura Vianna, Xuhui Zhou, Yejin Choi, and Noah A Smith. 2022. Annotators with attitudes: How annotator beliefs and identities bias toxic language detection. In *Proceedings of the 2022 conference of the north american chapter of the association for computational linguistics: Human language technologies*, pages 5884–5906.

Thibault Sellam, Dipanjan Das, and Ankur Parikh. 2020. Bleurt: Learning robust metrics for text generation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7881–7892.

Zuhair Hasan Shaik, Abdullah Mazhar, Aseem Srivastava, and Md Shad Akhtar. 2025. Redefining experts: Interpretable decomposition of language models for toxicity mitigation. *arXiv preprint arXiv:2509.16660*.

Amit Sharma and Rajni Bhalla. 2025. Detecting hate speech for hindi-english code-mix text data using dual contrastive learning. *Procedia Computer Science*, 259:35–43.

Lingfeng Shen, Weiting Tan, Sihao Chen, Yunmo Chen, Jingyu Zhang, Haoran Xu, Boyuan Zheng, Philipp Koehn, and Daniel Khashabi. 2024. The language barrier: Dissecting safety challenges of llms in multilingual contexts. In *Findings of the Association for Computational Linguistics ACL 2024*, pages 2668–2680.

Abhishek Singhania, Christophe Dupuy, Shivam Sadashiv Mangale, and Amani Namboori. 2025. Multi-lingual multi-turn automated red teaming for llms. In *Proceedings of the 5th Workshop on Trustworthy NLP (TrustNLP 2025)*, pages 141–154.

Anirudh Som, Karan Sikka, Helen Gent, Ajay Divakaran, Andreas Kathol, and Dimitra Vergyri. 2024. Demonstrations are all you need: Advancing offensive content paraphrasing using in-context learning. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 12612–12627, Bangkok, Thailand. Association for Computational Linguistics.

Marco Antonio Stranisci and Christian Hardmeier. 2025. What are they filtering out? a survey of filtering strategies for harm reduction in pretraining datasets. *arXiv preprint arXiv:2503.05721*.

Anirudh Sundar, Sinead Williamson, Katherine Metcalf, Barry-John Theobald, Skyler Seto, and Masha Fedzechkina. 2025. Steering into new embedding spaces: Analyzing cross-lingual alignment induced by model interventions in multilingual language models. *arXiv preprint arXiv:2502.15639*.

Teodor Tiţa and Arkaitz Zubiaga. 2021. Cross-lingual hate speech detection using transformer models. *arXiv preprint arXiv:2111.00981*.

Alexander Matt Turner, Lisa Thiergart, Gavin Leech, David Udell, Juan J. Vazquez, Ulisse Mini, and Monte MacDiarmid. 2024. Steering language models with activation engineering. *Preprint*, arXiv:2308.10248.

Bibek Upadhayay and Vahid Behzadan. 2024. Sandwich attack: Multi-language mixture adaptive attack on llms. In *Proceedings of the 4th Workshop on Trustworthy Natural Language Processing (TrustNLP 2024)*, pages 208–226.

Bibek Upadhayay and Vahid Behzadan. 2025. Tongue-tied: Breaking LLMs safety through new language learning. In *Proceedings of the 7th Workshop on Computational Approaches to Linguistic Code-Switching*, pages 32–47, Albuquerque, New Mexico, USA. Association for Computational Linguistics.

Rheeya Uppaal, Apratim Dey, Yiting He, Yiqiao Zhong, and Junjie Hu. 2025. Model editing as a robust and denoised variant of dpo: A case study on toxicity. *Preprint*, arXiv:2405.13967.

Sahil Verma, Keegan Hines, Jeff Bilmes, Charlotte Siska, Luke Zettlemoyer, Hila Gonen, and Chandan Singh. 2025. Multiguard: An efficient approach for

ai safety moderation across languages and modalities. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 16184–16198.

Pawat Vongpradit, Aurawan Imsombut, Sarawoot Kongyoung, Chaianun Damrongrat, Sitthaa Phaholphinyo, and Tanik Tanawong. 2024. Safecultural: A dataset for evaluating safety and cultural sensitivity in large language models. In *2024 8th International Conference on Information Technology (InCIT)*, pages 740–745. IEEE.

Andrew Wang, Mohit Sudhakar, and Yangfeng Ji. 2021. Simple text detoxification by identifying a linear toxic subspace in language model embeddings. *arXiv preprint arXiv:2112.08346*.

Boxin Wang, Wei Ping, Chaowei Xiao, Peng Xu, Mostofa Patwary, Mohammad Shoeybi, Bo Li, Anima Anandkumar, and Bryan Catanzaro. 2022. Exploring the limits of domain-adaptive training for detoxifying large-scale language models. *Advances in Neural Information Processing Systems*, 35:35811–35824.

Mengru Wang, Ningyu Zhang, Ziwen Xu, Zekun Xi, Shumin Deng, Yunzhi Yao, Qishen Zhang, Linyi Yang, Jindong Wang, and Huajun Chen. 2024a. Detoxifying large language models via knowledge editing. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3093–3118, Bangkok, Thailand. Association for Computational Linguistics.

Wenxuan Wang, Zhaopeng Tu, Chang Chen, Youliang Yuan, Jen-tse Huang, Wenxiang Jiao, and Michael Lyu. 2024b. All languages matter: On the multilingual safety of llms. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 5865–5877.

Johannes Welbl, Amelia Glaese, Jonathan Uesato, Sumanth Dathathri, John Mellor, Lisa Anne Hendricks, Kirsty Anderson, Pushmeet Kohli, Ben Coppin, and Po-Sen Huang. 2021. Challenges in detoxifying language models. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 2447–2469.

Jiaxin Wen, Pei Ke, Hao Sun, Zhexin Zhang, Chengfei Li, Jinfeng Bai, and Minlie Huang. 2023. Unveiling the implicit toxicity in large language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 1322–1338.

Yahan Yang, Soham Dan, Shuo Li, Dan Roth, and Insup Lee. 2025. MrGuard: A multilingual reasoning guardrail for universal LLM safety. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 27377–27396, Suzhou, China. Association for Computational Linguistics.

DONG Yi, Ronghui Mu, Gaojie Jin, Yi Qi, Jinwei Hu, Xingyu Zhao, Jie Meng, Wenjie Ruan, and Xiaowei Huang. Position: building guardrails for large language models requires systematic design. In *Forty-first International Conference on Machine Learning*.

Haneul Yoo, Yongjin Yang, and Hwaran Lee. 2025. Code-switching red-teaming: Llm evaluation for safety and multilingual understanding. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 13392–13413.

Yue Yu, Yuchen Zhuang, Jieyu Zhang, Yu Meng, Alexander J Ratner, Ranjay Krishna, Jiaming Shen, and Chao Zhang. 2023. Large language model as attributed training data generator: A tale of diversity and bias. *Advances in neural information processing systems*, 36:55734–55784.

Marcos Zampieri, Shervin Malmasi, Preslav Nakov, Sara Rosenthal, Noura Farra, and Ritesh Kumar. 2019. Semeval-2019 task 6: Identifying and categorizing offensive language in social media (offenseval). In *Proceedings of the 13th International Workshop on Semantic Evaluation*, pages 75–86.

Marcos Zampieri, Preslav Nakov, Sara Rosenthal, Pepa Atanasova, Georgi Karadzhov, Hamdy Mubarak, Leon Derczynski, Zeses Pitenis, and Çağrı Çöltekin. 2020. SemEval-2020 task 12: Multilingual offensive language identification in social media (OffensEval 2020). In *Proceedings of the Fourteenth Workshop on Semantic Evaluation*, pages 1425–1447, Barcelona (online). International Committee for Computational Linguistics.

Ruochen Zhang, Samuel Cahyawijaya, Jan Christian Blaise Cruz, Genta Indra Winata, and Alham Fikri Aji. Multilingual large language models are not (yet) code-switchers. In *The 2023 Conference on Empirical Methods in Natural Language Processing*.

Zekai Zhang, Yiduo Guo, Jiuheng Lin, Shanghaoran Quan, Huishuai Zhang, and Dongyan Zhao. 2025. English as defense proxy: Mitigating multilingual jailbreak via eliciting english safety knowledge. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pages 1185–1196.

Yiran Zhao, Wenxuan Zhang, Huiming Wang, Kenji Kawaguchi, and Lidong Bing. 2025. AdaMergeX: Cross-lingual transfer with large language models via adaptive adapter merging. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 9785–9800, Albuquerque, New Mexico. Association for Computational Linguistics.

Terry Yue Zhuo, Yujin Huang, Chunyang Chen, and Zhenchang Xing. 2023. Red teaming chatgpt via jailbreaking: Bias, robustness, reliability and toxicity. *arXiv preprint arXiv:2301.12867*.