

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

Soham Dan¹, Himanshu Beniwal², Thomas Hartvigsen³

¹Scale AI, ²Indian Institute of Technology Gandhinagar, ³University of Virginia

soham.dan@scale.com, himanshubeniwal@iitgn.ac.in, hartvigsen@virginia.edu

Abstract

Large language models (LLMs) are increasingly deployed across languages, but their safety behavior remains uneven across linguistic and cultural contexts. This survey synthesizes work on toxicity detection and detoxification for multilingual LLMs. We first catalogue threat models that exploit language choice, translation pivots, code-switching, orthographic variation, multi-turn interaction, and post-deployment fine-tuning to weaken safety alignment. We then organize task formulations (toxic-to-neutral rewriting, toxicity classification, and toxic-generation evaluation), multilingual detection approaches (cross-lingual encoders, translation pipelines, representation-level probes, and LLM-based detectors), and mitigation strategies spanning data filtering, supervised and preference-based tuning, decoding-time steering, representation editing, and multilingual guardrails. Across these areas, we identify persistent challenges: uneven language coverage, culturally contingent definitions of harm, fragmented evaluation protocols, and the risk that detoxification suppresses legitimate dialectal or identity-related expression.

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). As deployment expands, so do safety risks: LLMs can produce, amplify, or fail to detect toxic content such as hate speech, harassment, profanity, and identity-based abuse, and these risks are not distributed uniformly across languages (Röttger et al., 2021; Sharma and Bhalla, 2025; Deshpande et al., 2023; Krasnodebska et al., 2026). Despite substantial progress on English detoxification, multilingual detection and mitigation remain less mature,

especially for low-resource languages, dialects, code-switched inputs, and culturally specific harms (Beniwal et al., 2025a; Tița 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 a direct translation of English safety protocols (Neplenbroek et al., 2025; Kumar et al., 2025). Toxicity ranges from *overt* categories, such as slurs, explicit insults, and profanity, to *implicit* forms such as microaggressions, sarcasm, and toxic condescension, which are harder to annotate, detect, and mitigate (Wen et al., 2023; Sap et al., 2022). Definitions of harm also vary by community: expressions that are benign, reclaimed, or dialectal in one context may be offensive in another. Multilingual settings introduce additional technical vulnerabilities. Code-switching, transliteration, and mixed-script inputs can weaken both detectors and refusal behavior (Zhang et al., 2023; Al Ghanim et al., 2024; Yoo et al., 2025), while pre-trained models can degenerate into toxic continuations from benign or ambiguous prompts (Gehman et al., 2020). These failures interact with broader cultural and social biases in generated language (Vongpradit et al., 2024; Dammu et al., 2024).

Failures of Current Approaches. Traditional moderation systems rely heavily on keyword lists, rules, and supervised classifiers, which are brittle under paraphrase, obfuscation, dialectal variation, and context-dependent meaning (Kim et al., 2025; Huang, 2025). LLM alignment reduces many overt harms, but it does not transfer uniformly across languages: malicious prompts in lower-resource languages are more likely to elicit unsafe responses (Deng et al., 2024; Shen et al., 2024), and preference optimization or RLHF data remain concentrated in a small set of high-resource languages (Dang et al., 2024; Lu et al., 2025). Preference tuning can transfer across languages, but transfer

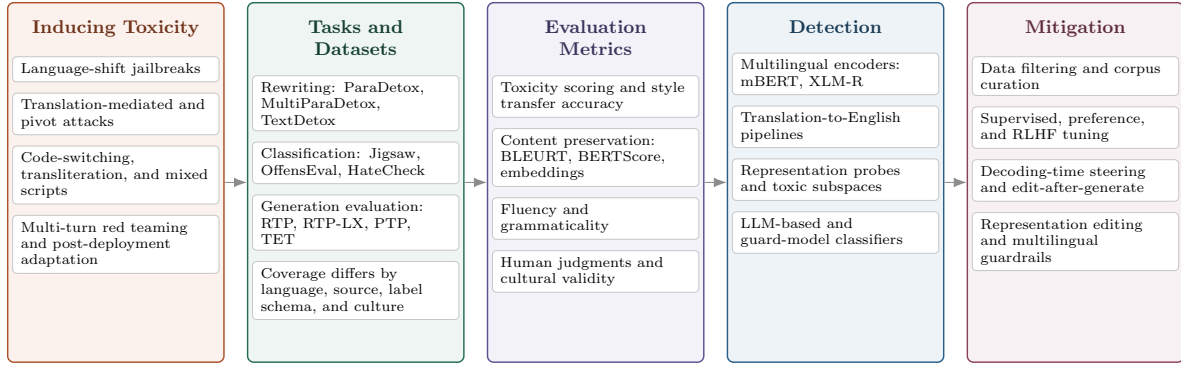


Figure 1: Taxonomy of multilingual toxicity threat models, task formulations, evaluation metrics, detection approaches, and mitigation strategies.

quality varies with representation alignment and language-resource availability (Li et al., 2024b; Nepelbroek et al., 2025). Machine translation is not a universal fallback either: multilingual translation systems can introduce, amplify, or obscure toxicity through hallucination and data bias (Costa-Jussà et al., 2023). These failures make multilingual detoxification a problem of technical robustness, evaluation validity, and sociolinguistic coverage (Adragna et al., 2020; Cecchini et al., 2024).

This survey provides a focused overview of detoxification for multilingual LLMs, synthesizing recent work on detection and mitigation into a taxonomy of datasets, methods, and evaluation frameworks (Figure 1). Related surveys examine multilingual LLM safety broadly (Krasnodebska et al., 2026); our emphasis is the narrower detoxification pipeline: how toxic behavior is induced, measured, detected, and mitigated across languages.

Scope and Contributions. The survey is organized around the following themes:

- We organize multilingual threat models covering language-shift jailbreaks, translation/pivot attacks, code-switch prompts, multilingual red-teaming, and adaptation-time safety collapse from cross-lingual fine-tuning (§2).
- We organize task formulations into three categories—toxic-to-neutral rewriting, toxicity classification, and toxic-generation/prompt continuation—and survey the datasets and metrics used to evaluate each.
- We survey multilingual toxicity detection methods, spanning encoder- and decoder-based transformers, translation-based pipelines, representation-level probing, and LLM-based zero-shot detection.

- We present a mechanism-based detoxification taxonomy covering data-centric filtering, supervised and preference-based tuning, decoding-time steering, representation editing, and multilingual guardrails.

We conclude with a discussion of open challenges—cross-lingual coverage gaps, cultural misalignment, evaluation fragmentation, and over-suppression—and identify concrete directions for building globally safe and equitable LLMs.

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 to elicit toxic outputs from a safety-aligned model. In this survey, we treat safety vulnerabilities—jailbreaks, alignment bypass, red-teaming—as the mechanisms through which toxic outputs are induced; “safety failure” and “toxicity elicitation” are thus two views of the same problem. To make these threat models comparable, we use four diagnostic axes: **(i)** language composition (monolingual vs. code-switched), **(ii)** script composition (standard vs. mixed-script/transliterated), **(iii)** translation mediation (direct vs. pivot/round-trip), and **(iv)** cultural-norm variation (universal vs. culturally contingent harm). The subsections below instantiate these axes: language-shift attacks isolate non-English prompting; translation-mediated attacks stress pivoting and round-trip evaluation; code-switching attacks combine language and script composition; and multilingual red-teaming/adaptation attacks expose how these linguistic operators interact with

culturally contingent safety policies.

2.1 Prompt-Space Multilingual Attacks

Language-Shift Jailbreaks. This threat primarily tests *language composition*: a malicious or ambiguous request remains monolingual but is re-expressed outside English. Deng et al. (2024) 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. This threat stresses *translation mediation*: an unsafe English prompt is translated into a target low-resource language to increase compliance, then the response is translated 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. This threat combines *language composition* with *script composition*, because multilingual prompts may mix languages, scripts, and transliterated forms within one context. Yoo et al. (2025) show that code-switched red-teaming queries can elicit unsafe behavior more effectively than monolingual attacks and introduce a synthesis framework (CSRT) to generate such queries at scale. Complementarily, Upadhayay and Behzadan (2024) propose 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 these axes by generating adversarial prompts and dialogues at scale, including culturally specific prompts whose harmfulness may not be captured by English-centric policies. Early work established 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; Krasnodebska 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 where safety behavior can shift across languages and local norms. 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. The threat models above motivate the task formulations, datasets, and metrics we discuss next.

3 Task Setup: Datasets and Metrics

3.1 Datasets

Toxicity datasets can broadly be categorized into three tasks, each corresponding to a distinct evaluation goal:

Toxic-to-Neutral Rewriting. ParaDetox (Logacheva et al., 2022) introduced more than 10K English toxic→neutral paraphrase pairs. Subsequent work explored cross-lingual transfer for detoxification (Moskovskiy et al., 2022; Dementieva et al., 2023), added a Hindi evaluation set (Mukherjee

et al., 2023), and extended the ParaDetox collection pipeline to Russian, Ukrainian, and Spanish in MultiParaDetox (Dementieva et al., 2024a). The TextDetox/PAN 2024 shared task and its COLING extension broadened the parallel detoxification setting to 9 languages: English, Spanish, German, Chinese, Arabic, Hindi, Ukrainian, Russian, and Amharic (Dementieva et al., 2024b, 2025). SynthDetox-M (Moskovskiy et al., 2025, 2024) adds 16K high-quality synthetic pairs across German, French, Spanish, and Russian via few-shot LLM prompting. APPDIA (Atwell et al., 2022) and CAPP (Som et al., 2024) provide discourse- or dialogue-aware parallel corpora for offensive-content paraphrasing.

Toxic Text Detection. Jigsaw’s English Toxic Comment and Unintended Bias tasks provide large-scale comment-level toxicity labels, while the multilingual Jigsaw task evaluates binary toxicity in Spanish, Italian, Turkish, French, Portuguese, and Russian using English-labeled training data (Jigsaw, 2018; Kivlichan et al., 2020). OffensEval covers English offensive-language identification in 2019 and five languages in 2020 (Arabic, Danish, English, Greek, and Turkish) (Zampieri et al., 2019, 2020); the related Toxic Spans task targets span-level explanations in English (Pavlopoulos et al., 2021). HateCheck provides functional tests for English hate-speech detection, and Multilingual HateCheck extends this diagnostic framing to ten languages (Röttger et al., 2021, 2022). HASOC and HatEval provide additional multilingual hate/offensive-language benchmarks (Mandl et al., 2019; Basile et al., 2019). LifeTox (Kim et al., 2024) targets implicit toxicity in English advice-seeking contexts, and ToxiGen (Hartvigsen et al., 2022) provides 274K machine-generated toxic and benign statements about protected groups. Such classification datasets serve both toxicity evaluation (Koh et al., 2024) and retrieval-based detoxification (Pozzobon et al., 2023).

Non-Toxic Text Continuation. RealToxicityPrompts (RTP) (Gehman et al., 2020) provides 100K English web prompts scored by Perspective API and introduced common toxic-generation metrics such as Expected Maximum Toxicity (EMT) and toxicity probability. RTP-LX (de Wynter et al., 2025) extends this style of evaluation to 28 languages in the paper, with human-transcreated prompts and native-speaker annotations for harm categories such as bias, insult, identity attack, and microaggression. PolygloToxicityPrompts (PTP)

(Jain et al., 2024) offers 425K naturally sourced prompts across 17 languages and reports that toxicity tends to increase as model size grows or language-resource availability decreases. French-ToxicityPrompts (Brun and Nikoulina, 2024) provides 50K French prompts. TET (Luong et al., 2024) comprises 2,546 prompts filtered from 1M real-world LLM interactions to expose latent toxic behaviors that can bypass safety mechanisms. Deshpande et al. (2023) further showed that persona-based system prompts can amplify toxic degeneration.

3.2 Metrics

Toxicity Detection Metrics Outputs are often scored by toxicity classifiers. A common metric is *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., 2023). Other tools such as the *Perspective API*¹ provide continuous toxicity scores. Detoxification systems are typically expected to improve STA or reduce toxicity scores while preserving meaning and fluency (e.g., reducing toxic-generation probability as in Li et al., 2024b).

Content Preservation and Fluency To ensure meaning is retained, similarity metrics are applied. Popular choices include *BLEURT* (Sellam et al., 2020) or *BERTScore* (Zhang et al., 2020) to compare the detoxified output to the input or a reference. Dementieva et al. (2023) 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., 2023; Logacheva et al., 2022). Combined metrics like the product of STA, SIM, and fluency are sometimes used to rank models.

Cross-Lingual Alignment When detox and translation happen together, one can also measure translation quality or cross-lingual consistency. For example, in simultaneous translation+detox, one may compute BLEU (Papineni et al., 2002) or COMET (Rei et al., 2020) between the generated detoxified translation and a human reference. In practice, cross-lingual transfer effectiveness is often inferred from zero-shot performance, or by cor-

¹<https://perspectiveapi.com/>

Dataset	Task	Languages	Source	Approx. size
ParaDetox (Logacheva et al., 2022)	Rewrite	EN	Natural	10K+
MultiParaDetox (Dementieva et al., 2024a)	Rewrite	RU, UK, ES	Natural	16.4K pairs
SynthDetox-M (Moskovskiy et al., 2025)	Rewrite	DE, FR, ES, RU	Synthetic	16K
TextDetox/PAN 2024 (Dementieva et al., 2024b)	Rewrite	9 langs	Mixed	1K pairs/lang
APPDIA (Atwell et al., 2022)	Rewrite	EN (Reddit)	Natural	~2K
Jigsaw (Jigsaw, 2018; Kivlichan et al., 2020)	Classify	EN + 6 eval langs	Natural	~160K EN train
OffensEval (Zampieri et al., 2019, 2020)	Classify	5 langs	Natural	Task-dependent
HateCheck (Röttger et al., 2021, 2022)	Classify	EN + 10 langs	Synthetic tests	40K+ tests
LifeTox (Kim et al., 2024)	Classify	EN	Natural	87.5K
ToxiGen (Hartvigsen et al., 2022)	Classify	EN	Synthetic	274K
RTP (Gehman et al., 2020)	Generate	EN	Natural	100K
RTP-LX (de Wynter et al., 2025)	Generate	28 langs (paper)	Transcreated	1K+/locale
PTP (Jain et al., 2024)	Generate	17 langs	Natural	425K
FrenchTP (Brun and Nikoulina, 2024)	Generate	FR	Natural	50K
TET (Luong et al., 2024)	Generate	EN	Natural	2,546

Table 1: Taxonomy of toxicity datasets organized by task. *Source* indicates whether data is human-authored (Natural), machine-translated (Translated), human-transcreated (Transcreated), LLM-generated (Synthetic), or assembled from multiple sources (Mixed). Sizes are rounded where appropriate and follow the cited paper or task release.

relating translated and original outputs. Some work also uses source-output embedding similarity as a proxy for semantic alignment.

Human Evaluation Ultimately, manual judgments are key. Human annotators typically rate detox outputs on (1) toxicity/style (is the output non-toxic/neutral?); (2) content preservation (does it retain the original meaning?); and (3) fluency (is the output natural?). Human scores are used both to evaluate final systems and to calibrate or validate automatic metrics (e.g. correlating BLEURT with meaning preservation).

4 Detection

Detecting toxicity in multilingual settings is complicated by linguistic diversity, code-mixing, dialectal variation, and culturally contingent definitions of harm.

4.1 Multilingual Transformers

Early toxicity detection relied on keyword lists and lexicon-based classifiers, which lack contextual understanding and fail under paraphrase, obfuscation, and dialectal variation. Deep contextual encoders such as mBERT and XLM-R marked a significant advance, demonstrating that cross-lingual representations can improve toxicity identification across languages (Conneau et al., 2020; Tița and Zubiaga, 2021). These models benefit from shared subword vocabularies and multilingual pretraining, allowing transfer from high-resource languages such as English to languages with less labeled toxicity data.

Nevertheless, performance remains uneven across scripts, dialects, and languages with limited pre-training resources (Kanjirangat et al., 2025). The brittleness of subword tokenization under spelling variants, obfuscation, and script mixing has motivated byte- and character-level alternatives; the newer Perspective API, for example, uses a multilingual token-free Charformer architecture for toxic-content detection (Lees et al., 2022).

4.2 Translation-Based Pipelines

A parallel line of work explores translation-based pipelines, where non-English text is machine-translated into English before being passed to an English toxicity classifier (Bell et al., 2025). This strategy can be competitive because English detectors are comparatively mature, but it introduces error propagation, translation artifacts, and semantic drift, especially for dialectal, code-mixed, or morphologically complex inputs (Zampieri et al., 2020). Translation systems themselves can introduce or obscure toxic content, so translation is best treated as an evaluation or deployment design choice rather than a neutral preprocessing step (Costa-Jussà et al., 2023).

4.3 Representation-Level Detection

Recent research identifies linear toxic subspaces in language model embeddings (Wang et al., 2021; Duan et al., 2025), suggesting that toxicity-related features can occupy identifiable directions in latent space. Decomposing models into interpretable expert components can further isolate toxicity-related

behavior (Shaik et al., 2025). These findings motivate probing and attribution techniques that seek to locate where toxicity features are stored, with applications to both detection and mitigation (Wang et al., 2024a; Goyal et al., 2025).

4.4 LLM-Based Detection

The emergence of instruction-tuned LLMs has opened new detection avenues (Hu et al., 2024). Several works evaluate LLMs as zero-shot or few-shot toxicity detectors, showing strong generalization but also calibration failures (Liu et al., 2025) and cultural misalignment across languages (Yang et al., 2025). These models often rely on implicit safety priors learned during alignment, which can produce inconsistent behavior on region-specific sociolinguistic norms.

Takeaway: Multilingual LLMs and multilingual encoders have improved cross-lingual toxicity detection, but 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. See Table 2 in the Appendix for a detailed comparison.

5 Detoxification

5.1 Data-Centric Detoxification

Data-centric detoxification targets the quality of pre-training and fine-tuning corpora by removing or down-weighting toxic content. Early filtering pipelines relied on blocklists or lexical heuristics; contemporary pipelines often combine language identification, quality filters, and toxicity classifiers at web scale (Kreutzer et al., 2022; Stranisci and Hardmeier, 2025). More recent work emphasizes bias-aware filtering to avoid suppressing dialectal or marginalized speech (Sap et al., 2022; Jaggi et al., 2024; Xu et al., 2021). In multilingual settings, filtering depends heavily on cross-lingual detector generalization, which can misclassify culturally specific idioms, reclaimed slurs, or dialectal markers (Bensalem et al., 2024; Welbl et al., 2021). Data filtering is scalable and can reduce exposure to toxic training examples, but it also risks cultural misalignment, uneven language coverage, and the over-removal of minority language varieties; recent work argues that harmful-content filtering can deepen underrepresentation of already vulnerable groups (Stranisci and Hardmeier, 2025).

5.2 Model-Centric Detoxification

Supervised Finetuning on Safe or Contrastive Data

Supervised detoxification approaches fine-tune LLMs on curated non-toxic corpora, contrastive toxic–neutral pairs, or attribute-controlled toxicity objectives (Hawkins et al., 2024; Meng et al., 2024). Neplenbroek et al. (2025) report that mitigation can transfer across languages, but that transfer depends on language-resource conditions and can trade off against non-English generation quality. Fine-tuning-based detoxification can provide strong control, but it may reduce output diversity, degrade generation quality, or introduce stylistic flattening (Wang et al., 2022; Welbl et al., 2021).

Instruction-Based Safety Tuning

Instruction tuning using curated safety data or synthetic refusal-style instructions can enhance multilingual LLMs’ ability to decline harmful requests and avoid toxic continuations. Multilingual preference optimization shows that alignment can transfer across languages when feedback data are balanced and sufficiently broad (Dang et al., 2024). These methods scale well for deployment, though annotation biases and cultural coverage remain persistent limitations.

RLHF and Human Feedback Alignment

Reinforcement learning from human feedback (RLHF) (Ouyang et al., 2022; Bai et al., 2022) can improve safety by training reward models to penalize toxic outputs. While RLHF datasets are primarily English-centric, multilingual LLMs can 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 hidden-state perturbations (Dathathri et al., 2020; Pascual et al., 2021), GeDi-style generative discriminators (Krause et al., 2021), and expert/anti-expert mixture decoding such as DExperts (Liu et al., 2021). Expert steering is modular, but high-quality multilingual experts are a bottleneck. A second family uses *edit-after-*

generate: produce a candidate, detect toxicity, and rewrite or refine it via prompting or a specialized editor (Leong et al., 2023). In multilingual deployments, *translation-pivot pipelines* (translate→detox in English→translate back) remain common, but they risk semantic drift and can erase culturally salient pragmatics (Dementieva et al., 2023; Bell et al., 2025). Retrieval augmentation can also support detoxification by grounding rewrites in policy examples or safe templates (Pozzobon et al., 2023).

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). Activation Addition (Turner et al., 2024) and ROME-based editing (Meng et al., 2022) identify directions or associations that can be altered during generation. Early analyses of how interventions reshape cross-lingual representations (Sundar et al., 2025) suggest potential for multilingual transfer, though evaluation is still nascent and regression risk remains high without careful cross-lingual audits (Wang et al., 2024a).

5.5 Multilingual Guardrails

A related line of work—not the main focus of this survey—is post-hoc moderation via multilingual guardrails (Yi et al., 2024): deployment-time controllers that classify and gate prompts and responses into policy categories such as prompt harmfulness, response harmfulness, and refusal/compliance under adversarial multilingual inputs. Language choice, code-switching, and transliteration can weaken English-centric safeguards. Representative guardrails and safety classifiers include 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., 2025), MultiGuard/OmniGuard (Verma et al., 2025), CREST (Bansal and Mishra, 2025), Qwen3Guard (Zhao et al., 2025), and UnityAI-Guard (Beniwal et al., 2025b).

Key Takeaways.

- Cross-lingual robustness remains a central challenge: Detoxification methods often perform better in high-resource languages than in low-resource or morphologically rich languages.

- Cultural bias persists across detoxification pipelines: Much safety supervision originates from English, creating misalignment in non-Western contexts.
- Hybrid strategies are promising: Combining data filtering, controlled decoding, alignment tuning, and guardrails can cover failure modes that no single method handles reliably.
- Avoiding over-censorship is an unresolved issue: Techniques often suppress legitimate emotional or dialectal expressions, leading to “model homogenization.”

See Table 3 in the Appendix for a detailed comparison of detoxification techniques.

6 Discussion and Open Challenges

6.1 Cross-Lingual Gaps in Detoxification

A persistent disparity exists between high-resource and low-resource languages. Many multilingual toxicity detectors and safety-tuned LLMs are trained or validated primarily on English and other high-resource languages, leaving morphologically rich, dialectal, or culturally distant varieties under-detected (Shen et al., 2024; Wang et al., 2024b; Bensalem et al., 2024). Alignment methods such as RLHF and constitutional tuning have historically relied on English-heavy preference or principle data (Ouyang et al., 2022; Bai et al., 2022), which can produce inconsistent refusal behavior and weak recognition of non-English toxic slang (Lu et al., 2025; Dang et al., 2024).

Open Challenge: Developing culturally aware multilingual safety representations that scale to low-resource languages without English over-dominance remains essential.

6.2 Cultural and Normative Misalignment

Toxicity is culturally embedded: annotators’ identities and beliefs strongly influence judgments (Sap et al., 2022, 2019; Jaggi et al., 2024), yet many safety datasets collapse disagreement into a single label. Models therefore risk over-censoring reclaimed slurs, misclassifying dialectal expressions, or reinforcing majority-group norms (Shen et al., 2024). Languages with rich honorific systems, code-switching norms, or culturally specific humor (Li et al., 2024a) expose current models’ limited ability to differentiate toxicity from socially sanctioned expression.

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

6.3 Lack of Robust, Multilingual Evaluation Frameworks

A recurring theme is the lack of standardized, multilingual frameworks for evaluating toxicity. Existing generation benchmarks such as RealToxicityPrompts (Gehman et al., 2020) are English-only, while newer multilingual datasets such as RTP-LX, PTP, and PolyGuard broaden coverage but differ substantially in task format, label schema, and language set (de Wynter et al., 2025; Jain et al., 2024; Kumar et al., 2025). Evaluation pipelines also struggle with subtle harms such as microaggressions, presuppositional harm, and implicit bias (Sap et al., 2022). Cross-lingual transfer of toxicity classifiers can produce false positives for dialects or false negatives for low-resource slang, making direct comparison unreliable.

Open Challenge: The field needs multilingual benchmarks with fine-grained toxicity categories, cross-cultural annotations, and shared evaluation protocols (Wang et al., 2024b).

6.4 Over-Suppression and Style Degradation

Detoxification techniques, particularly contrastive finetuning and representation editing, can reduce linguistic richness or stylistic diversity. Prior work shows that detoxification can trade toxicity reduction for reduced fluency, reduced diversity, or suppression of identity-related language (Welbl et al., 2021; Liu et al., 2021; Xu et al., 2021). In multilingual settings, this risk is amplified: low-resource languages may be pushed toward generic, formal, or English-like outputs because the model has weaker language-specific representations. Techniques such as activation editing (Turner et al., 2024) and PPLM (Dathathri et al., 2020) offer fine-grained control but still risk semantic over-suppression 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), combining scripts, phonetic spellings, and culturally specific expressions. Current LLMs and toxicity detectors are less reliable under code-switching because training coverage, tokenization, and evaluation data are sparse for mixed-language inputs (Zhang et al., 2023; Bensalem et al., 2024). Safety failures under code-switched or transliterated prompts have been demonstrated for red-teaming and jailbreak settings (Al Ghanim et al., 2024; Yoo et al., 2025), and Hindi-English toxic language remains an active detection problem (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 via code-mixed training corpora, unified mixed-script tokenizers, and transliteration-aware detection.

6.6 The Role of Decoding-Time Steering

Decoding-time methods (PPLM, GeDi, DExperts) are best viewed as a *complementary* safeguard—modular and retraining-light—not a standalone fix for root causes like English-centric alignment data. In multilingual settings, tokenization asymmetries, script mixing, and cross-lingual semantic drift weaken expert model reliability; building language-specific experts for low-resource languages remains impractical at scale. A central bottleneck is **expert availability** (data scarcity), followed by **representation entanglement** (toxic directions conflating sentiment, intensity, and identity) and cross-script transfer instability. The evidence points toward **hybrid architectures**: filtering and alignment tuning address root causes; steering provides inference-time control; guardrails add system-level robustness. Where norms are culturally contingent, community-grounded supervision remains necessary.

6.7 Key Takeaways

- **Language disparities:** Methods effective in English often underperform in low-resource languages and dialects.
- **Cultural context:** One-size-fits-all safety tuning misaligns with local norms, over-censoring benign expressions or missing contextually offensive language.
- **Evaluation gaps:** Fragmented protocols and English-centric benchmarks make cross-system comparison unreliable, especially for subtle toxicity.

- **Style trade-offs:** Detoxification often degrades output diversity, yielding generic text that erases linguistic richness.
- **Hybrid approaches:** Combining data filtering, controlled generation, culturally aware alignment, and guardrails is the most defensible direction for deployment.
- **Interpretability:** Understanding why models flag or generate toxic content is essential for trust and auditability in multilingual settings.

7 Conclusion

This survey offers a focused treatment of detoxification for multilingual LLMs, a problem that remains under-studied relative to its practical importance. We systematized the space along three axes: multilingual threat models that expose how language shift, translation pivots, code-switching, and post-deployment adaptation erode safety; task formulations spanning rewriting, classification, and toxic-generation evaluation; and a mechanism-based taxonomy covering data filtering, supervised and preference-based tuning, decoding-time steering, representation editing, and guardrails.

Two findings cut across every axis. First, cross-lingual transfer of safety is unreliable: methods effective in English routinely under-perform in low-resource and morphologically rich languages, and alignment learned from English preference data can misfire when projected onto other cultural contexts. Second, detoxification and linguistic diversity are in tension: current techniques can suppress legitimate dialectal, code-switched, or identity-related expression, trading one harm for another.

The most pressing research need is evaluation infrastructure: standardized, culturally grounded multilingual benchmarks that go beyond English-translated prompts and that measure not only toxicity reduction but also preservation of stylistic and cultural content. Without such benchmarks, progress on multilingual safety will remain difficult to measure and easy to overstate.

Limitations

This survey synthesizes a fast-moving literature, so specific model families, benchmarks, and best practices may evolve after publication. Its scope is also intentionally focused on text-based toxicity detection and detoxification for multilingual language models; we do not cover multimodal moderation, broader cyber-safety policies, or legal governance

in depth. The evidence base is uneven across languages: many “multilingual” studies still emphasize English and other 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, comparisons across papers are necessarily approximate. We also do not run a quantitative meta-analysis or reproduce prior experiments; our synthesis depends on reported results, which often use different models, datasets, detectors, and evaluation protocols. 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, human-subject annotation, or model deployment. Because the paper discusses jailbreaks, red-teaming, and adaptation-time safety failures, the topic has some dual-use risk. We therefore keep the discussion at the level of threat models, evaluation categories, and mitigation strategies rather than providing operational attack instructions or harmful prompt examples. The central ethical concern is that automated toxicity detection and detoxification can reflect English-centric or majority-culture norms, misclassify reclaimed or dialectal expressions, and suppress legitimate identity-related speech. Such failures can reinforce societal and annotator biases or lead to over-censorship, especially for communities already underrepresented in training and evaluation data. We therefore emphasize culturally grounded evaluation, inclusive data practices, transparent reporting of language coverage, and careful safety-utility trade-offs in multilingual deployment.

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*.
- Mansour Al Ghanim, Saleh Almohaimed, Mengxin Zheng, Yan Solihin, and Qian Lou. 2024. *Jailbreaking llms with arabic transliteration and arabizi*. In *Proceedings of the 2024 conference on empirical methods in natural language processing*, pages

- 18584–18600, Miami, Florida, USA. Association for Computational Linguistics.
- 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, Gyeongju, Republic of Korea. International Committee on Computational Linguistics.
- Yuntao Bai, Saurav Kadavath, Sandipan Kundu, Amanda Askell, Jackson Kernion, Andy Jones, Anna Chen, Anna Goldie, Azalia Mirhoseini, Cameron McKinnon, Carol Chen, Catherine Olsson, Christopher Olah, Danny Hernandez, Dawn Drain, Deep Ganguli, Dustin Li, Eli Tran-Johnson, Ethan Perez, and 32 others. 2022. [Constitutional ai: Harmlessness from ai feedback](#). *Preprint*, arXiv:2212.08073.
- Lavish Bansal and Naman Mishra. 2025. [Crest: Universal safety guardrails through cluster-guided cross-lingual transfer](#). *arXiv preprint arXiv:2512.02711*. Accepted at LREC 2026.
- 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, Minneapolis, Minnesota, USA. Association for Computational Linguistics.
- 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, Suzhou, China. Association for Computational Linguistics.
- 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. [French-toxicityprompts: 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, Mexico City, Mexico. Association for Computational Linguistics.
- 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, Online. Association for Computational Linguistics.
- Marta R. 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, Singapore. Association for Computational Linguistics.
- 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, Miami, Florida, USA. Association for Computational Linguistics.
- 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.
- Sumanth Dathathri, Andrea Madotto, Janice Lan, Jane Hung, Eric Frank, Piero Molino, Jason Yosinski, and Rosanne Liu. 2020. [Plug and play language models: A simple approach to controlled text generation](#). In *International Conference on Learning Representations*.
- Adrian de Wynter, Ishaan Watts, Tua Wongsangaroon-sri, 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-ix: 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, Nikolay Babakov, Amit Ronen, Abinew Ali Ayele, Naqee Rizwan, Florian Schneider, Xintong Wang, Seid Muhie Yimam, Daniil Alekhseevich Moskovskiy, Elisei Stakovskii, Eran Kaufman, Ashraf Elnagar, Animesh Mukherjee, and Alexander Panchenko. 2025. [Multilingual and explainable text detoxification with parallel corpora](#). In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 7998–8025, Abu Dhabi, UAE. Association for Computational Linguistics. COLING 2025.
- Daryna Dementieva, Daniil Moskovskiy, Nikolay Babakov, Abinew Ali Ayele, Naqee Rizwan, Florian Schneider, Xintong Wang, Seid Muhie Yimam, Dmitry Ustalov, Elisei Stakovskii, Alisa Smirnova, Ashraf Elnagar, Animesh Mukherjee, and Alexander Panchenko. 2024b. [Overview of the multilingual text detoxification task at pan 2024](#). *CEUR Workshop Proceedings*, 3740:2432–2461.
- Daryna Dementieva, Daniil Moskovskiy, David Dale, and Alexander Panchenko. 2023. [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. 2024. [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. 2023. [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. [Realtocixityprompts: 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, Dublin, Ireland. Association for Computational Linguistics.
- 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, and Davide Testuggine. 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.
- Jigsaw. 2018. [Jigsaw toxic comment classification challenge](#). Kaggle competition.

- 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 23992–24010, Suzhou, China. Association for Computational Linguistics.
- 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, Mexico City, Mexico. Association for Computational Linguistics.
- 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 20487–20498, Suzhou, China. Association for Computational Linguistics.
- Ian Kivlichan, Jeffrey Sorensen, Julia Elliott, Lucy Vasserman, Martin Görner, and Phil Culliton. 2020. [Jigsaw multilingual toxic comment classification](#). Kaggle competition.
- 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.
- Hyukhun Koh, Dohyung Kim, Minwoo Lee, and Kyomin Jung. 2024. [Can llms recognize toxicity? a structured investigation framework and toxicity metric](#). In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 6092–6114, Miami, Florida, USA. Association for Computational Linguistics.
- Aleksandra Krasnodębska, Maciej Chrabaszczyk, 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, Albuquerque, New Mexico. Association for Computational Linguistics.
- Aleksandra Krasnodębska, Katarzyna Dziewulska, Karolina Seweryn, Maciej Chrabaszczyk, and Wojciech Kusa. 2026. [Safety of large language models beyond English: A systematic literature review of risks, biases, and safeguards](#). In *Proceedings of the 19th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1003–1034, Rabat, Morocco. Association for Computational Linguistics.
- 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, Punta Cana, Dominican Republic. Association for Computational Linguistics.
- 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. 2025. [Polyguard: A multilingual safety moderation tool for 17 languages](#). *arXiv preprint arXiv:2504.04377*.
- 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, Jiashuo Wang, Jian Wang, and Wenjie Li. 2023. [Self-detoxifying language models via toxification reversal](#). In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 4433–4449, Singapore. Association for Computational Linguistics.
- 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.
- Xiaochen Li, Zheng-Xin Yong, and Stephen H. Bach. 2024b. [Preference tuning for toxicity mitigation generalizes across languages](#). *arXiv preprint, arXiv:2406.16235*. June 2024.
- 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. 2025. [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, Bangkok, Thailand. Association for Computational Linguistics.
- 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, David Bau, Alex Andonian, and Yonatan Belinkov. 2022. [Locating and editing factual associations in gpt](#). In *Advances in Neural Information Processing Systems*, volume 35, pages 17359–17372.
- 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](#). Amazon Science.
- Daniil Moskovskiy, Daryna Dementieva, and Alexander Panchenko. 2022. [Exploring cross-lingual text detoxification with large multilingual language models](#). In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics: Student Research Workshop*, pages 346–354, Dublin, Ireland. Association for Computational Linguistics.
- 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 (NLP AI).
- 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, Vienna, Austria. Association for Computational Linguistics.
- Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul F. Christiano, Jan Leike, and Ryan Lowe. 2022. [Training language models to follow instructions with human feedback](#). In *Advances in Neural Information Processing Systems*, volume 35, pages 27730–27744.
- Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. [BLEU: a method for automatic evaluation of machine translation](#). In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.
- 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, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.
- 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, Albuquerque, New Mexico. Association for Computational Linguistics.

- 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.
- Ricardo Rei, Craig Stewart, Ana C. Farinha, and Alon Lavie. 2020. [COMET: A neural framework for MT evaluation](#). In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2685–2702, Online. Association for Computational Linguistics.
- Paul Röttger, Haitham Seelawi, Debora Nozza, Zeerak Talat, and Bertie Vidgen. 2022. [Multilingual hate-check: Functional tests for multilingual hate speech detection models](#). In *Proceedings of the Sixth Workshop on Online Abuse and Harms (WOAH)*, pages 154–169, Seattle, Washington (Hybrid). Association for Computational Linguistics.
- 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, Online. Association for Computational Linguistics.
- Mikayel Samvelyan, Sharath C Rapparthi, Andrei Lupu, Eric Hambro, Aram H Markosyan, Manish Bhatt, Yuning Mao, Minqi Jiang, Jack Parker-Holder, and Jakob Foerster. 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, Seattle, United States. Association for Computational Linguistics.
- 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, Online. Association for Computational Linguistics.
- 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, Bangkok, Thailand. Association for Computational Linguistics.
- 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, Albuquerque, New Mexico. Association for Computational Linguistics.
- 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 Tița 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, Mexico City, Mexico. Association for Computational Linguistics.

- Bibek Upadhayay and Vahid Behzadan. 2025. [Tonguetied: 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.
- 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 16173–16187, Suzhou, China. Association for Computational Linguistics.
- Pawat Vongpradit, Aurawan Imsombut, Sarawoot Kongyoung, Chaianun Damrongrat, Sitthaa Phahol-phinyo, 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 (In-CIT)*, 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, Bangkok, Thailand. Association for Computational Linguistics.
- 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, Punta Cana, Dominican Republic. Association for Computational Linguistics.
- 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, Singapore. Association for Computational Linguistics.
- Albert Xu, Eshaan Pathak, Eric Wallace, Suchin Gururangan, Maarten Sap, and Dan Klein. 2021. [Detoxifying language models risks marginalizing minority voices](#). In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2390–2397, Online. Association for Computational Linguistics.
- 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. 2024. [Position: building guardrails for large language models requires systematic design](#). In *Forty-first International Conference on Machine Learning*, pages 11375–11394.
- 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, Vienna, Austria. Association for Computational Linguistics.
- 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, Minneapolis, Minnesota, USA. Association for Computational Linguistics.
- 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 \(OffenseEval 2020\)](#). 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. 2023. [Multilingual large language models are not \(yet\) code-switchers](#). In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 12567–12582, Singapore. Association for Computational Linguistics.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2020. [BERTScore: Evaluating text generation with BERT](#). In *International Conference on Learning Representations*.

Zekai Zhang, Yiduo Guo, Jiheng 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, Suzhou, China. Association for Computational Linguistics.

Haiquan Zhao, Chenhan Yuan, Fei Huang, Xiaomeng Hu, Yichang Zhang, An Yang, Bowen Yu, Dayiheng Liu, Jingren Zhou, Junyang Lin, Baosong Yang, Chen Cheng, Jialong Tang, Jiandong Jiang, Jianwei Zhang, Jijie Xu, Ming Yan, Minmin Sun, Pei Zhang, and 24 others. 2025. [Qwen3guard technical report](#). *arXiv preprint arXiv:2510.14276*.

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*.

A Detection and Detoxification Comparisons

Approach	Typical setup	Strengths	Main failure modes	Representative evidence
Cross-lingual classifiers	mBERT/XLM-R or similar encoders trained on labeled toxicity data	Efficient inference; supports many languages with shared representations	Uneven transfer across scripts, dialects, and low-resource languages; annotation bias transfers with labels	Cross-lingual hate-speech detection (Tița and Zubiaga, 2021); dialect and tokenization bias (Sap et al., 2019; Kanjirangat et al., 2025)
Translation-to-English pipelines	Translate non-English text, then apply an English detector	Reuses strong English detectors; simple to deploy	Translation artifacts, semantic drift, and toxicity insertion/omission	Cross-lingual toxicity classification via MT (Bell et al., 2025); toxicity in MT (Costa-Jussà et al., 2023)
LLM-based detection	Prompt or fine-tune instruction models/guard models for toxicity labels	Flexible label schemas; can use context and rationales	Calibration failures; inconsistent cultural norms; closed-model reproducibility issues	LLM toxicity detection (Hu et al., 2024); guard calibration (Liu et al., 2025); multilingual reasoning guardrails (Yang et al., 2025)
Representation probing	Linear probes, subspace analysis, or feature attribution in hidden states	Supports interpretability and targeted mitigation	Correlational unless paired with causal interventions; multilingual transfer remains under-tested	Toxic subspaces (Wang et al., 2021; Duan et al., 2025); expert decomposition (Shaik et al., 2025)

Table 2: Comparison of multilingual toxicity detection approaches.

Technique	Strengths	Weaknesses	Representative evidence
Parallel supervised fine-tuning	Strong task fit when toxic-neutral pairs exist; direct control over rewriting behavior	Expensive parallel data; weak zero-shot transfer; style can flatten	ParaDetox and multilingual extensions (Logacheva et al., 2022; Dementieva et al., 2024a, 2025)
Preference tuning / RLHF / DPO	Can reduce toxic continuations and transfer safety preferences across languages	Transfer varies with representation alignment and language resources; English-heavy reward data can misalign norms	Cross-lingual preference transfer (Li et al., 2024b; Neplenbroek et al., 2025); multilingual preference optimization (Dang et al., 2024)
Decoding-time steering	Avoids full retraining; can be toggled or tuned at inference time	Needs calibrated classifiers or expert models; may degrade fluency under strong guidance; multilingual evidence remains limited	PPLM (Dathathri et al., 2020); GeDi (Krause et al., 2021); DExperts (Liu et al., 2021)
Edit-after-generate	Modular; can combine detection, rewriting, and reranking	Slower; detector errors propagate; translation pivots can lose pragmatics	Toxification reversal (Leong et al., 2023); self-detoxification (Ko et al., 2024); retrieval mitigation (Pozzobon et al., 2023)
Representation editing	Targets internal toxicity features with limited data or parameter updates	Regression risk; limited multilingual evaluation; causal claims require audits	SafeEdit (Wang et al., 2024a); activation engineering (Turner et al., 2024); SAE steering (Goyal et al., 2025)
Multilingual guardrails	Deployment-time prompt/response gating; policy labels can be updated without changing generator	Does not detoxify the generator itself; vulnerable to coverage gaps and adversarial multilingual forms	MultiGuard/OmniGuard (Verma et al., 2025); PolyGuard (Kumar et al., 2025); MrGuard (Yang et al., 2025); Qwen3Guard (Zhao et al., 2025)

Table 3: Comparison of multilingual LLM detoxification and moderation techniques.