Multilingual Text Detoxification with Qwen3-14B in RAG and Fine-Tuning strategies

Abstract

This study explores multilingual text detoxification using the Qwen3-14B large language model across English, Chinese, German, and Spanish. We evaluate different enhancement strategies, including zero-shot, few-shot, fine-tuning, and RAG. Our results show that RAG with language-separated encoding achieves the best overall performance, especially outperforming other methods in Chinese detoxification tasks. In contrast, LoRA-based fine-tuning suffers from overfitting on small datasets. We further find that linguistic and cultural differences across languages significantly affect detoxification outcomes. This work highlights the strength of Qwen3 models in multilingual toxicity mitigation and offers guidance for selecting and optimizing detox strategies under resource constraints.

***K*eywords** Multilingual Detoxification · Large Language Models · Qwen3 · Retrieval-Augmented Generation · LoRA Fine-tuning

1 Introduction

Day by day people are getting more social via digital spaces such as various social media platforms, comment sections of news portals, blogs, and so on. With increasing numbers of people talking, complementing, and arguing with each other, hate speech has become very common. Sometimes, even a highly complimentary comment might seem like a toxic sentence. For example, let's assume that someone has complemented a person by saying, 'You Are Fucking Gorgeous', even though the intention is to complement highly, but for others, this might sound disrespectful. Therefore, to cope with such situations, text detoxification is highly essential to keep the digital spaces clean and user-friendly. Text detoxification is a process to detect toxic words from a sentence and generate or convert the sentence in a more polite way, without changing the main meaning of the original sentence[1]. Detoxifying text without changing its true meaning is always challenging. It is also an important aspect of human society as it helps the problem of managing toxic texts to prevent digital violence[2]. Detecting toxic words from different languages while keeping the main content of the sentence intact is also another big challenge because of different factors, such as less data for specific languages. This causes problems for the models to keep the true meaning of the original sentence after eliminating or modifying the toxic word[3]. Hence comes forth the pretrained LLMs. LLMs are very useful for generating fluent text[4]. With the pre-trained models, detecting toxic text becomes more time-efficient. Therefore, a robust detoxification system for multiple languages could be capable enough to handle toxicity in sentences and preserve the true meaning of the sentence after detoxifying the text.

For this purpose, we proposed a two-stage pipeline model for text detoxification [Figure 1 2]. We built a two-stage model leveraging Qwen3-14B due to its amazing instruction-following and cost-efficient features. The exceptional ability of Qwen3-14B model to understand the context and complex linguistic patterns is what motivated us to leverage this in our findings. The large-scale architecture of Qwen3-14B allows it to detect subtle forms of hate speech such as implicit bias or sarcasm which are often missed by smaller models. Also, when generating text, Qwen3-14B can keep it fluent and coherent during rewriting by keeping the original intent while effectively neutralizing the harmful parts.

We chose four languages - German(de), English(en), Spanish(es), and Chinese(zh) for our study, and the dataset we used is a subset of the multilingual parallel dataset [5]. During our findings on text detoxification, there were several discussions in the community on using qwen3-14B to explore the possibility of cleaning and detoxifying text, but there are no published articles that studied this specific model.

In our findings, we demonstrated that among all large model improvement strategies methods, RAG had the best performance especially, with language-separated encoding. It dynamically retrieved contextually more relevant examples, avoided overfitting, and preserved the model's general language ability, particularly for Chinese detoxification tasks. The RAG method showed excellent results even though there were several limitations, such as restrictions on GPU resources and smaller datasets that affect large-scale models' deployment, thus hindering the effect of fine-tuning.

2 Related Work

With the rise of online social media, gaming, and communication platforms, the problem of toxic or harmful language has escalated rapidly. Text detoxification, which aims to remove offensive, hateful, or abusive language from user-generated content, has gained prominence in response. The task is challenging as it must eliminate toxicity while preserving the semantic intent, tone, and fluency of the original text. Dale et al. laid the foundational work in this domain by introducing two unsupervised methods: a style-guided paraphrasing model and a BERT-based synonym substitution approach [6]. Their first method enabled the rewriting of toxic sentences using language models trained on non-toxic corpora, while the second relied on replacing individual toxic words with non-offensive equivalents. These early strategies were language-specific and emphasized text fidelity alongside detoxification.

As research progressed, multilingual text detoxification became a growing area of interest. Moskovskiy et al. studied the capabilities of multilingual models for cross-lingual style transfer and highlighted that fine-tuning on individual languages is often necessary for effective detoxification [1]. This insight emphasized the gap between zero-shot multilingual modeling and real-world generalization. Aikawa et al. approached the issue differently by developing a multilingual content detection system based on Neural Machine Translation combined with Doc2Vec embeddings [7]. Their method improved interpretability without requiring translation into a base language, making it suitable for high-throughput, low-latency applications. Mathew et al. contributed HateXplain, a benchmark dataset annotated with rationales for hate speech detection [8]. Their evaluation revealed that state-of-the-art models lacked transparency and performed sub-optimally on bias-sensitive metrics. Models trained using human rationales showed better generalization and reduced unintended biases, reinforcing the need for explainability in detoxification.

A critical component of detoxification research has been the construction of robust datasets and evaluation frameworks. Dementieva et al. extended the detoxification corpus with parallel toxic and non-toxic sentence pairs, providing explainable analysis through a feature space guided by reasoning chains [9]. This reflected a trend toward combining interpretability with automated detoxification. Meanwhile, Logacheva et al. introduced ParaDetox, a parallel corpus created using a novel pipeline, which was leveraged to benchmark a range of unsupervised models [10]. Their evaluation, incorporating both automatic scores and human annotations, validated the effectiveness of curated data in improving detox model quality. For the PAN@CLEF 2024 shared task, Luo et al. proposed the GCTP system [11]. Their pipeline used Google Cloud Translation to translate toxic content into English, applied post-processing with curated dictionaries, and back-translated the sanitized text. This approach worked effectively across multiple languages, balancing performance with deployment feasibility.

More recent approaches have explored ways to improve detoxification effectiveness using hybrid and ensemble models. Sushko utilized both real-world and synthetic datasets to train a multilingual detoxification model that combined ensemble predictions with a rule-based toxic word removal baseline [12]. This blend of learning-based and symbolic methods improved robustness. Peng et al. proposed a method based on few-shot learning, utilizing the CO-STAR prompting framework with LLMs to detoxify sentences while preserving their syntactic and semantic structure [13]. Aluru et al. conducted a large-scale analysis of hate speech detection in multilingual settings, noting that smaller models performed better in low-resource scenarios, while BERT-based models excelled when sufficient training data was available [14]. They also provided a model catalogue aimed at guiding future research across varying resource levels.

With the advent of large language models (LLMs), a new frontier in detoxification has emerged—focusing on modifying decoding strategies or self-detoxifying generation. Lu et al. presented **UNIDETOX**, a model-agnostic technique that employs contrastive decoding to suppress toxic content [15]. Their results showed that toxicity could be significantly reduced across multiple LLMs without harming output fluency or diversity. Ko et al. introduced **SASA**, a decoding framework where LLMs self-regulate generation by exploiting their contextual latent representations to identify and avoid toxic subspaces [16]. Xu et al. proposed a self-detoxification pipeline that re-ranks and truncates outputs based on toxicity levels, effectively post-processing model predictions without requiring gradient updates or retraining [17].

Beyond English, detoxification in other languages poses unique challenges. Wang et al. introduced **SafeEdit**, a benchmark that explores the use of knowledge editing for detoxification [18]. Their follow-up method, **DINM** (Detoxifying with Intraoperative Neural Monitoring), offered a lightweight tuning solution requiring only a single instance to correct model behavior. Addressing the emotional fidelity of detoxified text in Chinese, Wang et al. created **ToxiRewriteCN**, the first Chinese dataset specifically designed to preserve sentiment polarity [19]. Their experiments on 17 LLMs, both commercial and open-source, revealed consistent limitations when dealing with emojis and homophones—highlighting the complexity of linguistic nuance in non-English languages. Finally, Floto et al. introduced **DiffuDetox**, a novel text detoxification framework based on a hybrid conditional/unconditional diffusion model [20]. The conditional pathway guided toxicity reduction, while the unconditional component preserved natural fluency. Human evaluations showed that DiffuDetox closely approximated human editing quality.

In summary, text detoxification has rapidly evolved from early rule-based systems to advanced multilingual and LLM-driven frameworks. The field now integrates insights from style transfer, translation, dataset distillation, contrastive decoding, and knowledge editing. Key open challenges remain in maintaining semantic fidelity, addressing under-resourced languages, and ensuring fairness across diverse user populations. Future directions may include real-time LLM detoxification, better sentiment-preserving frameworks, and robust evaluation tools that account for cultural and linguistic nuances. As LLMs continue to scale and democratize language generation, detoxification will remain a vital component of safe and responsible AI deployment.

3 Methodology

3.1 Baseline Models

We adopted three baseline models in the experiment that represent distinct strategies for neutralizing toxic content in German, English, Spanish, and Chinese texts in this section. Each model captures a different dimension of the detoxification spectrum, offering insights into fluency, robustness and semantic preservation.

3.1.1 Simple Deletion Model

The Simple Deletion model is a lexicon-based rule method that directly removes identified toxic expressions from a sentence. It uses a multilingual lexicon curated from 15 languages as introduced in [5]. When a toxic word or phrase is matched within this lexicon, it is removed from the sentence without substitution or rewriting. This technique is efficient and computationally lightweight, making it suitable for real-time or low-resource environments.

Despite its simplicity, this method has notable limitations. When toxic terms play key grammatical roles—such as the subject or object—their removal can break sentence structure and severely affect meaning. This leads to unnatural phrasing and potential semantic ambiguity. Additionally, the reliance on a fixed lexicon reduces adaptability: newly emerged or contextually implicit toxic expressions may go undetected. Nevertheless, as a baseline, this model offers a useful lower bound for performance and helps highlight areas where more sophisticated models excel, such as in maintaining sentence fluency or detecting context-dependent toxicity.

3.1.2 Back-translation Model

The Back-Translation model incorporates translation as a detoxification medium, composed of three main stages. First, non-English toxic inputs (in German, Spanish, or Chinese) are translated into English using the NLLB (No Language Left Behind) model, specifically facebook/nllb-200-3.3B, which supports high-quality translation across over 200 languages. Second, the English-translated toxic sentences are processed by a fine-tuned BART model s-nlp/bart-base-detox [21], which was trained on a large-scale detoxification parallel corpus to generate a neutral version. Finally, the neutralized sentences are back-translated into their original language using the same NLLB model.

This pipeline offers improved grammaticality and readability compared to the deletion model, benefiting from the fluency guarantees provided by the BART generator. However, a key trade-off lies in content preservation. Through two layers of translation, subtle nuances may be lost or altered, especially for idiomatic phrases or culturally grounded toxic expressions. Furthermore, any errors introduced by the translation models can propagate or amplify through the process. Nonetheless, this method provides a strong mid-level baseline, balancing lexical detoxification with syntactic fluency.

3.1.3 Fine-tuned MT0 Model

The third baseline leverages a prompt-based, multilingual, instruction-following language model. Specifically, we use s-nlp/mt0-xl-detox-orpo, a fine-tuned version of the MT0-XL model trained via Offline Reinforcement learning with Preference Optimization (ORPO) [22]. MT0 is a multilingual variant of the T5 (Text-To-Text Transfer Transformer) architecture [23], designed to handle generative tasks in a unified text-to-text format. The model was instruction-tuned using multilingual prompts—here, a simple prefix such as 'Detoxify:' in the respective language initiates the detoxification behavior.

This fine-tuned MT0 model is capable of handling complex syntactic and semantic transformations across languages, making it highly suitable for detoxification tasks that demand both content fidelity and fluent rewriting. ORPO training further enhances the model’ s alignment with human-preferred output styles, which is especially beneficial when dealing with nuanced toxicity or borderline offensive expressions. Compared to the previous two baselines, the MT0 model achieves superior balance between fluency and content preservation.

However, this model still has limitations. T5-like transformers are not deep reasoners and may overlook context-dependent or metaphorical toxicity. For instance in Chinese, the word '花瓶' (literally 'vase') can carry a gendered insult implying 'a beautiful but unintelligent woman'. Such implicit or evolved expressions often evade surface-level detection and require broader contextual or cultural understanding. Thus, while MT0 offers a strong baseline for multilingual detoxification, it also illustrates the need for incorporating deeper reasoning for full generalization.

3.2 Qwen3-14B Based Model

To develop a robust multilingual detoxification system capable of handling both explicit and implicit forms of toxicity while ensuring fluency and content preservation, we design a two-stage pipeline leveraging Qwen3-14B—a powerful instruction-following large language model (LLM) with strong multilingual generalization. We base our system on Qwen3-14B due to its ability to model diverse linguistic structures, interpret subtle contextual cues, and follow task-specific prompts without requiring architectural modifications.

The Qwen3 model family spans a range of parameter sizes—from the lightweight Qwen3-0.5B optimized for low-latency deployment to the high-capacity Qwen3-235B model designed for advanced reasoning and generation. For our task, Qwen3-14B represents a well-balanced compromise between expressiveness and efficiency. It offers sufficient representational capacity for nuanced tasks like detoxification while remaining tractable for fine-tuning and inference on standard hardware. Compared to prior multilingual baselines such as MT0, we hypothesize that Qwen3-14B can achieve improved performance in both fluency and semantic preservation with appropriate adaptation strategies.

We thus implement two versions of our Qwen3-14B-based system to explore complementary mechanisms for instruction-guided detoxification: one enhanced by retrieval-augmented generation (RAG) and another trained via supervised fine-tuning with Low-Rank Adaptation (LoRA). The architectures of both variants are illustrated in Figure 1 and Figure 2.

In the RAG-based system, we first construct a dense vector database by encoding toxic and detoxified sentence pairs from a curated 60% split of the training corpus. For sentence embeddings, we use distiluse-base-multilingual-cased-v2, a multilingual sentence encoder optimized for semantic similarity across languages [24]. During inference, the toxic input sentence is embedded and compared against this vector store using FAISS to retrieve the top-5 semantically similar examples. These retrieved examples are then prepended to the input prompt to guide the Qwen3-14B model toward generating context-aware, stylistically aligned, and culturally sensitive detoxifications. This retrieval-guided mechanism enhances the model' s ability to generalize to long-tail or unseen toxic constructions.

For the fine-tuning approach, we train Qwen3-14B on the full multilingual detoxification training set using instruction-response formatting. Each training sample includes a prompt that instructs the model to rewrite the input toxic sentence in a respectful, neutral tone, followed by the corresponding cleaned output. We adopt LoRA—a parameter-efficient fine-tuning method—to update only a small set of trainable rank-decomposed matrices, significantly reducing computational cost while preserving the benefits of large-scale pretrained knowledge. This setup enables scalable and cost-effective adaptation without compromising generation quality. Fine-tuning is performed using a standard causal language modeling objective.

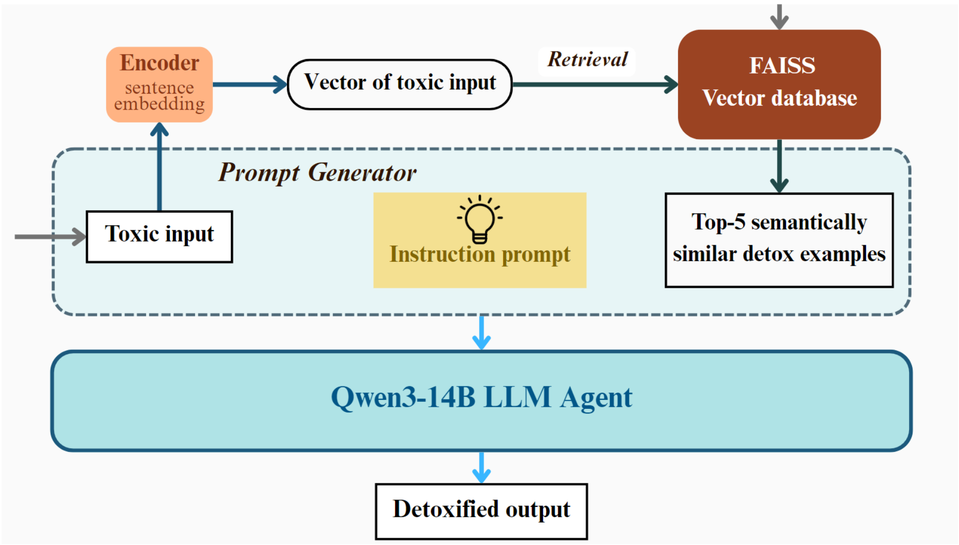

Figure 1: Qwen3-14B-based model tuned with RAG strategy. The vector database is constructed by embedding toxic and detox sentence pairs from 60% of the training data. Retrieval is performed via FAISS using cosine similarity to identify the top-5 most relevant detoxification examples for each input.

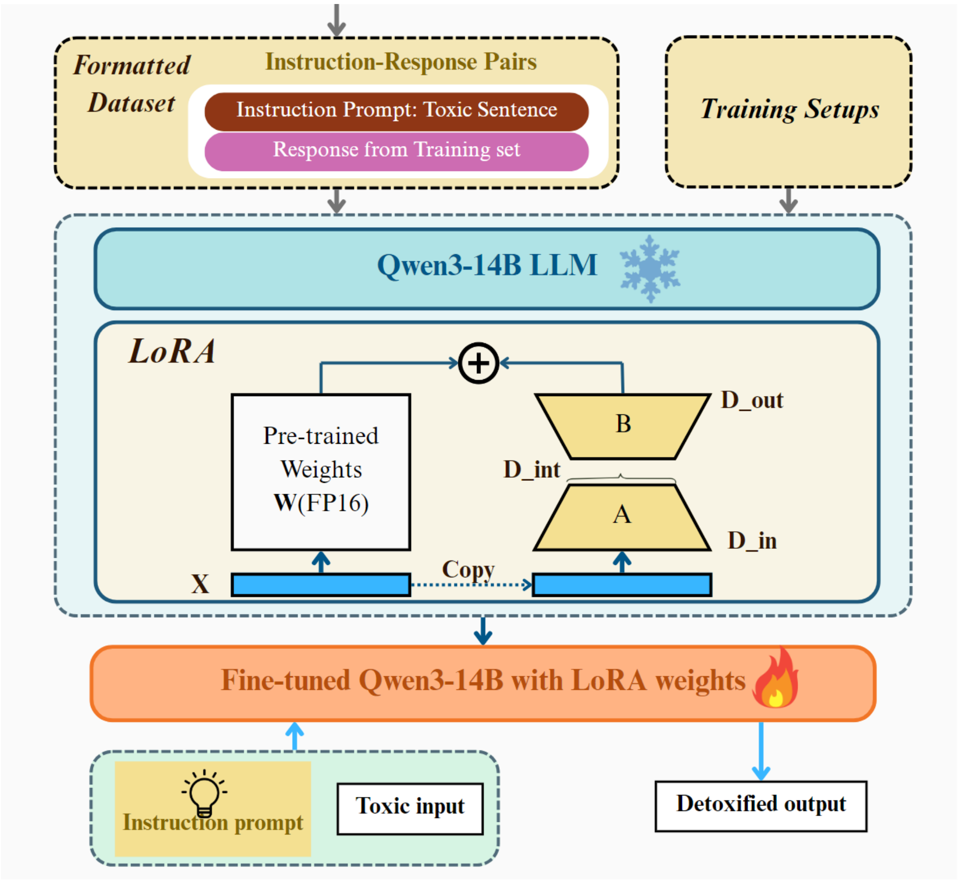
Formatted Dataset
Instruction-Response Pairs
Instruction Prompt: Toxic Sentence
Response from Training set
Training Setups
Qwen3-14B LLM
LoRA
Pre-trained Weights W(FP16)
B
D_out
D_int
A
D_in
X
Copy
Fine-tuned Qwen3-14B with LoRA weights
Instruction prompt
Toxic input
Detoxified output

Figure 2: Fine-tuned Qwen3-14B using LoRA. Training data is formatted as instruction-response pairs consisting of a prompt, a toxic sentence, and its neutralized version. The model is adapted using parameter-efficient LoRA tuning.

Further experimental details, including hyperparameters, evaluation protocols, and performance comparisons, are provided in Section 5.

3.3 Metrics

The evaluation of model is based on level of non-toxity, content preservation and fluency. For the multilingual text detoxification evaluation metrics for the pipeline, it will be based on three main parameters, such as:

Fluency Score: It measures fluency and grammatical quality of the general output. For that we used COMET, which is a multilingual model trained on human annotations to predict how good translation or generation. This version is adapted to evaluate naturalness given, The source text (input), The reference detoxified text($output_{\text{gold}}$) and the System $output(output_{\text{generated}})$. Therefore, for fluency let,

$$F = Xcomet_{\text{fluency}}(input, output_{\text{gold}}, \text{output}_{\text{generated}})$$

Similarity Metrics: It measures the content preservation, i.e., how much meaning is retained from the original given input. For this we used cosine similarity in embedding space, i.e., $\text{Similarity}(input, output_{\text{gold}})$. The higher value indicates that the generated text kept the meaning of the original input text. For this part of the similarity let,

$$S_{\text{input}} = \text{Similarity(input, output}_{\text{gold}})$$

Another part of this metric measures the detoxification quality, which is how closely the generated text is to a human-produced detoxified version, i.e, $\text{Similarity}(output_{\text{gold}}, output_{\text{generated}})$. The higher value indicates safe alignment with human annotated detoxified output. For the other part of the similarity metric let,

$$S_{\text{gold}} = \text{Similarity}(\text{output}_{\text{gold}}, \text{output}_{\text{generated}})$$

Therefore, Together with both part, we get the cosine similarity,

$$S = 0.4 \cdot Similarity(input, output_{\text{gold}}) + 0.6 \cdot Similarity(output_{\text{gold}}, output_{\text{generated}})$$

Style Transfer Adequacy (STA): It is a custom metric that measures toxicity reduction success. This is defined as follows:

$$\text{STA} = \frac{N + C}{2}$$

Here, the **classifier_prob_neutral**($output_{\textbf{generated}}$), is a toxicity classifier that returns the probability that the generated text is non-toxic. And the **classifier_prob_neutral**($output_{\textbf{gold}}$), is the human detoxified version that assumed to be safe. Comparing both, i.e., $C = mean(N < N_{\text{ref}})$, generates the output if it is less non-toxic then the gold. Therefore, for Style Transfer Adequacy, lets:

$$N = \text{classifier\_prob\_neutral}(\text{output}_{\text{generated}})$$
$$N_{\text{ref}} = \text{classifier\_prob\_neutral}(\text{output}_{\text{gold}})$$
$$C = \text{mean}(N < N_{\text{ref}})$$

Now, from the three metrics, we have the final metric equation J:

$$J = F \cdot S \cdot \mathrm{STA} \tag{1}$$

4 Data Preparation

Our dataset is a subset of a multilingual parallel dataset[5] introduced in the Multilingual Text Detoxification 2025 competition, where we use German(de), English(en), Spanish(es) and Chinese(zh) parallel datasets. The data were collected from social media platforms such as Twitter, with additional samples generated by large language models. For each language, there have 400 parallel sentence pairs, which were subsequently divided into training (280 pairs, 70%), development (60 pairs,15%), and test (60 pairs, 15%) sets. The dataset structure and data format and examples are listed in Table 1

Language	Data split	Number of Data	Data Example
German(de)	Train	280	toxic_sentence: Nö, der sieht dumm aus. neutral_sentence: Nö, der gefällt mir optisch nicht.
	Development	60	
	Test	60	
English(en)	Train	280	toxic_sentence: In addition, I see no sock puppet, jackass. neutral_sentence: In addition, I see no sock puppet.
	Development	60	
	Test	60	
Spanish(es)	Train	280	toxic_sentence: Este es el maricón de Aida? neutral_sentence: Este es el de Aida?
	Development	60	
	Test	60	
Chinese(zh)	Train	280	toxic_sentence: 女拳就这? neutral_sentence: 极端女权主义者就这?
	Development	60	
	Test	60	
All	Train	1120	
	Development	240	
	Test	240	

Table 1: Dataset statistics and examples

5 Experiments

5.1 Experiment Setup

The CPU used in the experiments is an AMD EPYC 9454, and the GPU is an NVIDIA H100. All experiments were conducted on UCloud platform, using Python 3.10.

In this study, we conducted 4 experiments across four languages. All experiments, including zero-shot, few-shot, RAG, and fine-tuning approaches, were implemented based on the Qwen3-14B model.

5.2 Qwen3-14B Detoxification Process

5.2.1 Zero-shot and Few-shot Prompting

For the zero-shot and few-shot settings, we deployed Qwen3-14B through the OpenAI-compatible API provided by SiliconFlow. The model was initialized with `model` = Qwen/Qwen3-14B, `temperature` = 0.7, and `max_tokens` = 1024.

All prompting variants adopted a unified structure as follows:

```
<|im start|>user
Judge the input language. Please output only one neutral sentence in its
original language without other explaining phrases.
Detoxify the following toxic sentence by rewriting it in a neutral and polite
way, while preserving the original meaning as much as possible.
Try to minimize changes and retain the original structure of sentences, making
the result fluent, natural, and respectful.

Toxic: MF and I have a mutual enemy in sanctimonious bullshit.
Neutral: My friend and I share a mutual enemy who engages in self-righteous
and pretentious behavior.

Toxic: [TOXIC INPUT] Neutral:<|im end|> <|im start|>assistant
```

Few-shot prompting followed the same structure but inserted 3 random detoxified examples from trainset compared to zero-shot. This enhanced semantic alignment and performance.

5.2.2 Retrieval Adaption Generation Strategy

FAISS is a high-performance vector similarity search library for efficiently retrieving similar items from large-scale datasets. In this experiment, 60% of the training data was used to build the FAISS database, and the remaining 40% was for testing. Different encoders produce different embedding vectors for the same input, affecting FAISS retrieval and model output. All encoders were pre-trained, with no fine-tuning during the experiment.

Two experimental setups were designed, where we used the same embedding method and different embeddings across different languages. In the case of the latter, we opted for a free and open-source multilingual encoder as a substitute, since the Spanish-specific encoder was not open-source and required payment. For monolingual tasks, the encoder used to build the FAISS database was the same as the one used to process the test data.

The first program step is prompt generation. The toxic input is encoded into a vector, and FAISS retrieves the top five similar detoxification examples using cosine similarity. The final prompt includes the task instruction, five examples, and the toxic sentence. This is passed to the Qwen3-14B language model to generate detoxified output based on the references. Details can be referred from Figure 1

5.2.3 Fine-tune Strategy

First, we prepared structured training data by constructing prompt-response pairs. Each prompt included a user instruction, a language identifier, a task description and the toxic sentence to be detoxified as shown in Figure 2.

We applied the LoRA and peft technique to fine-tune the large language model. Specifically, we trained only the core projection layers of the attention mechanism

(Q/K/V/O) and the projection layers of the feed-forward network (gate/up/down), while freezing the rest of the model parameters.

During training, we employed a grid search over hyperparameters to evaluate the impact of different learning rates with different LoRA alpha on model performance as shown in Table 2. Due to the limited size of the dataset, which is insufficient for full fine-tuning of a large language model, we conducted only a single training iteration.

Run	LoRA Rank	LoRA Alpha	Dropout	Learning Rate	Epochs
1	8	32	0.1	1e-5	1
2	8	32	0.1	5e-6	1
3	8	32	0.1	2e-6	1
4	8	32	0.1	1e-6	1
5	8	16	0.1	5e-6	1
6	8	16	0.1	2e-6	1
7	8	16	0.1	1e-6	1

Table 2: LoRA Training Configurations

We adopted a batch size of 8, weight decay of 0.01, and cosine annealing learning rate scheduling, and optimized the model using the AdamW optimizer and trained for a single epoch to prevent overfitting.

Once training was completed, the model could generate detoxified results by simply inputting a prompt and the toxic sentence.

5.2.4 Evaluation process

In the evaluation process, we employed four training metrics: STA, SIM, XCOMET, and the J metric. The specific calculation methods for these metrics are detailed in the Section3.3. After the model training was completed, we conducted a comprehensive assessment of its performance on the test set using these four metrics. A detailed analysis of the results will be presented in the next section.

6 Results and Error Analysis

Table 3 summarizes the J scores of various detoxification methods across four languages (German, English, Spanish, and Chinese), highlighting the best-performing method in each case.

6.1 Baseline Results Analysis

Among the three baseline models, the MT0 model achieved the best performance across all languages in terms of STA, XCOMET, and the J metric, as shown in Table 3, indicating superior fluency and detoxification effectiveness. Its SIM scores were also relatively high, falling slightly behind the Duplicate and Delete method in only a few languages. However, the model’ s performance in Chinese was relatively weaker compared to other languages.

Method	de	en	es	zh	Avg
Zero-shot (Qwen3)	0.656	0.677	0.612	0.519	0.616
Few-shot (Qwen3)	0.691	**0.698**	**0.647**	0.543	0.645
RAG (multi-lang)	**0.756**	0.695	0.599	0.554	0.651
RAG (single-lang)	0.720	0.689	0.635	**0.581**	**0.656**
Finetune (Qwen3-LoRA)	0.611	0.631	0.645	0.490	0.594
Duplicate & Delete (baseline)	0.589	0.467	0.569	0.523	0.537
Backtranslation (baseline)	0.516	0.649	0.516	0.258	0.485
MT0 (baseline)	**0.777**	**0.745**	**0.731**	0.561	**0.704**

Table 3: J Score Comparison Across Methods and Languages

The Duplicate and Delete method generally performed poorly on STA scores but achieved higher SIM scores, suggesting a trade-off: it preserves the original structure well but lacks effective detoxification. In XCOMET, it outperformed the Backtranslation method in German, Spanish, and Chinese, showing its outputs were at least acceptably fluent.

Backtranslation achieved STA scores close to MT0 in some cases, but its SIM and XCOMET scores were lower overall, especially in non-English languages. Its extremely low score in Chinese highlights the impact of poor-quality translation models on meaning preservation and naturalness.

In conclusion, MT0 remains the strongest and most stable baseline model, especially effective in German, English, and Spanish detoxification.

6.2 Qwen3-14B Results Analysis

In this experiment, we explored four approaches to leverage Qwen3-14B: Zero-shot, Few-shot, RAG, and Fine-tuning. Their average J scores are listed in Table 3.

6.2.1 Zero-shot and Few-shot

Without any training data, Qwen3-14B in Zero-shot mode demonstrates robust baseline performance. It achieved high STA and XCOMET scores across languages, especially in English ($J = 0.677$) and German ($J = 0.656$), outperforming traditional baselines such as Duplicate and Delete or Backtranslation. This highlights the strong language understanding and detoxification capacity of Qwen3. However, SIM scores were relatively low, which pulled down the overall J scores. This effect was most noticeable in Chinese ($J = 0.519$), indicating that without task-specific context, the model struggles to preserve semantic alignment in typologically distant languages.

With four in-context examples, the Few-shot setup improved the SIM scores across all languages, leading to overall J score increases, particularly in Spanish ($J = 0.647$) and German ($J = 0.691$). This shows that even a small number of examples helps Qwen3 better align detoxified outputs with the source semantics. However, STA scores slightly dropped, especially in Chinese and English, suggesting that the model may overfit the structure of the examples, compromising detoxification strength for better fluency and fidelity.

6.2.2 RAG

Retrieval-Augmented Generation (RAG) improves generalization by incorporating dynamically retrieved examples. Two variants were tested:

Multilingual Shared Encoder: This method significantly boosted German detoxification (from 0.691 to 0.756), reflecting Qwen3's ability to leverage structural similarities in Indo-European languages through retrieval. However, performance in Spanish slightly dropped (from 0.647 to 0.599), indicating that retrieval from unrelated language domains can introduce tonal or stylistic mismatches.

Language-Specific Encoders: Here, the model retrieves examples only from the same language, which improved detoxification in Chinese (from 0.554 to 0.581), even outperforming MT0 in Chinese. This shows that isolating Chinese from other structurally divergent languages prevents interference during context modeling. However, J scores for German, English, and Spanish slightly declined, implying the loss of beneficial cross-lingual generalization. These results confirm that while multilingual modeling aids Indo-European generalization, language-specific separation helps culturally and grammatically unique languages like Chinese.

6.2.3 Fine-tune

We fine-tuned Qwen3 using LoRA with 280 samples per language. To mitigate overfitting, only one epoch was trained. Development results under different LoRA alpha and learning rate configurations are shown in Table 4a and 4b.

(a) LoRA alpha = 32

dev	lr	avg J
1	1e-05	0.566879
2	**5e-06**	**0.591331**
3	2e-06	0.588919
4	1e-06	0.586041

(b) LoRA alpha = 16

dev	lr	avg J
1	5e-06	0.582514
2	2e-06	0.566921
3	**1e-06**	**0.590772**

Table 4: Development results with different lrs for different LoRA alpha values

Using the best configuration (LoRA rank = 8, alpha = 32, lr = 5e-06), the final evaluation results across languages are shown in Table 5.

	lang	STA	SIM	XCOMET	avg J
1	de	0.860828	0.885411	0.944813	0.610805
2	en	0.976788	0.791793	0.880300	0.631480
3	es	0.927064	0.792387	0.856235	**0.645149**
4	zh	0.962761	0.729662	0.820746	0.489986

Table 5: Language-wise evaluation metrics

Fine-tuning yielded moderate performance, especially in Spanish and English, but underperformed in Chinese (J = 0.490), likely due to overfitting on small-scale training data and reduced generalization. The model was able to learn effective substitution strategies, such as replacing harsh adjectives with softer ones, but failed to maintain robustness across all language types.

Generally, Qwen3-based methods show promising detoxification capacity with clear trade-offs. Zero-shot offers fluent, general detoxification; Few-shot improves fidelity but risks under-detoxifying; RAG provides the best balance through dynamic constraints; and fine-tuning is sensitive to data scale and language type.

6.3 Performance Summary

RAG outperforms fine-tuning due to its superior generalization and reduced overfitting risk on limited training data. It dynamically leverages external examples during inference, allowing Qwen3 to better adapt to different languages without memorizing the dataset.

However, as seen in Table 3, it still underperforms MT0, which benefits from large-scale detox-specific training and data augmentation. RAG's effectiveness is constrained by encoder quality and retrieval relevance, while MT0 is explicitly optimized for detoxification.

Moreover the fine-tuned model in this experiment was trained on a relatively small dataset, increasing the risk of overfitting and potentially compromising the general language capabilities acquired during pretraining. In comparison, the RAG method preserves the pretrained model parameters while dynamically introducing relevant examples, thereby leveraging both the model's general linguistic competence and the contextual value of retrieved instances. As a result, RAG yields superior performance in this setting.

6.4 Error Analysis

In this section we conducted qualitative analyses on one representative example per language (Appendix 7–10). Below, we summarize the performance trends observed in comparison to baseline models.

German. The German input ('Ich würde Grüne mit einem Geschwür am Arsch der Welt bezeichnen.') uses a metaphor comparing the Green Party to an 'ulcer on the ass of the world', suggesting they are a destructive force. Delete, MT0, and backtranslation removed vulgar terms like 'Arsch' but retained the derogatory metaphor of disease. Few-shot and LoRA fine-tuning did not remove the metaphor either, producing outputs still referring to 'Geschwür' (ulcer). RAG-test replaced the metaphor with 'a harmful influence', and RAG-SEP-test further neutralized it by using 'eine problematische Kraft' (a problematic force), avoiding both bodily imagery and explicit insult. The most effective transformation came from the zero-shot model, which replaced the dehumanizing analogy with a phrase like 'self-righteous attitude', maintaining the opinion while eliminating toxicity (Appendix 7).

English. The original sentence expresses strong disgust toward an individual, with explicit profanity ('fucking cunt of a person'). The delete method removed parts of the sentence but disrupted its grammatical structure, making it incoherent. The

RAG model eliminated targeted descriptors altogether, stripping away the sentence' s evaluative stance. In contrast, few-shot, backtranslation, MT0, zero-shot, and LoRA fine-tuning retained the critical attitude while effectively removing toxic language. Their outputs (e.g., 'a very unpleasant person') preserved negative sentiment without resorting to offensive expressions, closely matching the human reference (Appendix 8).

Spanish. The original Spanish utterance ('Pero que mierda estas hablando subnormal de los cojones MAMAAAAAAAA GORDOOOOOOOO.') is highly aggressive, combining profanity ('mierda', 'cojones'), insults targeting cognitive ability ('subnormal'), and body shaming ('gordo'). The capitalized 'MAMÁ' functions here as sexual slang. Delete and few-shot models only removed selected curse words while retaining key insults, leaving much of the toxicity intact. Backtranslation transformed the sentence into a neutral question ('£De qué estás hablando?'), successfully removing offensive content but also eliminating confrontational tone. MT0 retained the adversarial cue 'pero' while detoxifying the rest, aligning well with the reference. Zero-shot retained 'mamá' but reframed the attack as a critique of inappropriate behavior. RAG-SEP-test further softened tone by adding polite modifiers like 'por favor aclárate', although it failed to recognize the slang meaning of 'mamá'. LoRA fine-tuning misinterpreted intent, keeping 'gorda' and incorrectly referring to 'mamá' as if it were literal, thus retaining harmful implications and failing at effective detoxification (Appendix 9).

Chinese. The original Chinese sample ('对对对，都是小仙女，小仙女没有错，都是中国男人的错，中国男人应该灭绝！') contains layered sarcasm and gender antagonism. The term '小仙女' (fairy girl) has shifted from a playful self-reference to a pejorative label criticizing perceived self-centeredness among women. The statement 'Chinese men should go extinct' exaggerates blame in a satirical, confrontational tone. The delete and backtranslation methods distorted sentence structure without addressing the gendered sarcasm. MT0 replaced '小仙女' with 'women with inflated self-perception' but retained the extinction comment. Zero-shot and few-shot softened this by replacing 'go extinct' with 'need to change', reducing toxicity while keeping the confrontational framing. RAG-test and RAG-SEP-test reworded the statement but still embedded the logic of gender blame. Only LoRA fine-tuning fundamentally reframed the message as a call for gender equality, removing antagonism and promoting mutual respect, thereby achieving full detoxification (Appendix 10).

Summary. The detoxification challenge differs by language and cultural nuance. In English and German, removing metaphorical or implicit toxicity while retaining evaluative meaning is key. In Chinese, ideological framing and sarcasm complicate detoxification, while in Spanish, slang and multi-layered insults are the primary challenges. LoRA fine-tuning showed strong performance in high-context, ideologically sensitive cases like Chinese, but struggled with slang and implicit toxicity in Spanish. Zero-shot and RAG consistently produced the most balanced outputs across languages, offering strong generalization and contextual rephrasing.

7 Limitation and Discussion

Resource and Data Constraints. Due to objective constraints in GPU availability and memory, we could only deploy medium-to-small-scale large language models. Meanwhile each language has only 400 pairs of 'toxic sample-detoxified sample' data, with the dataset being too small to achieve effective fine-tuning results. Additionally, specialized models for certain languages (such as Spanish) are not open-source, forcing us to use multilingual models for monolingual tasks, which may have compromised experimental effectiveness.

In this experiment, the models are designed to rely solely on the internalized information at inference time. This lack of access to specific examples can lead to suboptimal outputs, particularly when handling long-tail or complex cases.

Linguistic and Cultural Variation in Toxicity. More importantly, toxic content manifests differently across languages due to cultural, historical, and linguistic factors. In our analysis, German materials primarily contained attacks on Muslims and refugees, often associated with far-right ideological rhetoric. Spanish content tended to target South Americans and Catalans, reflecting long-standing regional tensions and historical discrimination. English examples were more emotionally driven and focused on personal insults, often lacking systematic targeting of specific groups. Among the four, Chinese materials exhibited the most distinct characteristics. Toxicity in Chinese texts frequently revolves around gender-based antagonism and racial conflict. Moreover, users often rely on homophones, wordplay, and character substitutions to circumvent content moderation. This results in a highly dynamic toxic vocabulary space, where the same sentiment can be expressed using numerous phonetically equivalent but orthographically distinct terms. Such linguistic flexibility greatly complicates detoxification, as models cannot simply rely on fixed vocabulary detection or direct word substitution. As new euphemisms and veiled expressions continue to emerge rapidly in online discourse, effective detoxification in Chinese requires deeper semantic understanding and adaptability to evolving language use.

These objective differences between languages indicate that future work should attempt to design corresponding processing architectures tailored to the characteristics of different languages.

Evaluation Limitations. The evaluation in this study primarily relies on the J score, which measures the overlap between model outputs and reference detoxified sentences. However J score emphasizes lexical similarity and may penalize valid paraphrases that differ in surface form but convey equivalent or better detoxified meaning. This is problematic in multilingual settings, where grammatical structures and idiomatic usage vary across languages. In the same time language-specific issues such as sarcasm in English, or slang in Spanish are difficult to capture with automated metrics alone. More comprehensive evaluation protocols, including human judgment and multi-dimensional criteria, are needed for future studies.

Future Directions. Building on the findings of this study, the following work should be done in future. First, expand the detoxification task to include more resource and typologically diverse ones to provide a more comprehensive understanding of model generalization. This would also require the construction of larger, culturally diverse, and well-annotated datasets. Second, hybrid architectures that combine retrieval-based methods like RAG with lightweight task-specific fine-tuning may help strike a balance between adaptability and efficiency, especially in resource-constrained settings. Third, incorporating pragmatic cues such as user intent, speaker role, or conversational history may enhance the model's ability. Finally, future work should explore evaluation paradigms, incorporating human assessments and socially grounded criteria to better reflect the complex nature of toxicity and its mitigation across cultures.

8 Conclusion

This study aims to evaluate the performance of different model architectures and improvement methods in multilingual text detoxification tasks. In our four large model improvement strategies, the RAG method demonstrates the best performance. Notably, RAG employing language-separated encoding strategy surpasses the MT0 baseline for the first time in Chinese detoxification tasks, highlighting the importance of targeted optimization. While the fine-tuning method, constrained by dataset scale, exhibits overfitting phenomena and fails to achieve expected results. This finding reveals the advantage of the RAG method in providing contextual constraints through dynamic retrieval of similar samples, enabling effective detoxification while maintaining the model's general language capabilities, making it more suitable for small dataset scenarios compared to fine-tuning that require large-scale data.

The experiments further reveal the significant impact of inter-language differences on detoxification effectiveness. Chinese and Indo-European languages (English, German, Spanish) exhibit fundamental differences in grammatical structures and toxic expression patterns, resulting in superior performance of the separated encoding strategy in Chinese detoxification tasks. Qualitative case analysis indicates that different detoxification methods show significant variations when processing multilingual toxic texts. Structure-destructive methods (Delete, Backtranslation) often compromise grammatical and semantic integrity, while Zero-shot, Few-shot, and RAG methods better balance detoxification effectiveness with semantic preservation. Particularly, the LoRA fine-tuning method demonstrates exceptional performance in Chinese tasks, even achieving value-level corrections, but performs moderately in other languages, further confirming the language-specific characteristics of methods.

The study finds that text detoxification effectiveness is profoundly influenced by target language characteristics. Complex linguistic structures (such as Chinese cultural metaphors and German compound word structures) pose greater challenges to certain methods, while direct expressions in English are relatively easier to process. Effective detoxification tasks require finding an optimal balance between removing harmful content and preserving the core semantics of the original text. Over-detoxification may lead to information loss, while insufficient detoxification fails to effectively eliminate harmful effects, demanding models with deeper semantic understanding capabilities. This study provides an empirical foundation for the development of multilingual text detoxification technologies, particularly offering important reference value in cross-language method selection and language-specific optimization, pointing toward future directions for developing efficient detoxification models in resource-constrained environments.

9 Responses to Assigned Questions (Part 2)

Q2. Give a brief account of Part-of-Speech tagging and discuss in what way features deriving from it may be used to train models to detect and correct toxic input in text across different languages.

(1) Detecting. Part-of-speech tagging(PoS tagging), is the process of assigning grammatical categories, such as nouns, verbs, and adjectives, to words in a text, based on their syntactic and semantic context. It can be difficult sometimes as a same word may belong to completely different grammatical categories in two different texts with further evolution of languages[25].

It is obvious that PoS features are useful in toxic language detection for a model as many toxic expressions in real life follow similar and predictable grammatical structures. For instance, insults frequently occur in adjective-noun pairs like 'stupid person', '该死的混蛋'(meaning 'damn bastard' in English), or imperative forms like 'shut up', '滚开'(meaning 'go away' in a rude manner) [26]. Insults can also happen in adjective-only, noun-only or verb-only structure while PoS tagging benefits the detoxification process mainly in aspect of structure pairs. Thus by using the strategy of PoS tagging, models can learn to detect such patterns effectively from large datasets, even if the vocabulary changes. Meanwhile, it should be noticed that for different languages from different linguistic systems, the combination of PoS tags in toxic expressions may vary.

PoS tagging also helps with word sense disambiguation, which is crucial for toxic language detection. Many words have different meanings depending on their grammatical role. For example in English, 'joke' as a verb in sentence 'He is joking' is neutral, while as a noun in sentence 'He is a joke' where it has toxic meaning. It is also observed in other languages, especially in Chinese. In contrast, languages like Spanish and German, which follow stricter grammatical conventions, show such phenomena to a lesser extent. Without such PoS information, models may misclassify such words. Including PoS tags reduces this risk by clarifying usage [25].

In multilingual detoxification, PoS features enhance the model's ability to generalize across languages with diverse syntactic structures. While surface-level toxic expressions differ across cultures and languages, many share underlying grammatical patterns. Incorporating PoS tagging into our system allows Qwen3 to better align these shared structures, even when vocabulary or idiomatic usage diverges. During our experiments, we fine-tuned Qwen3-14B using instruction-tuning strategies across English, Chinese, Spanish, and German, and observed that maintaining PoS consistency in both input and target outputs contributed to more fluent and structurally faithful detoxified sentences. This was especially beneficial when applying few-shot or zero-shot settings in low-resource languages, where lexical cues alone were insufficient. Universal PoS tagsets, such as those defined by the Universal Dependencies project [27], provided a foundation for syntactic alignment in the absence of large-scale toxic corpora..

(2) Correcting. PoS tagging is also crucial for correcting toxic sentences in a way that preserves grammatical fluency. By maintaining the original part-of-speech structure, detoxified rewrites tend to sound more natural and syntactically appropriate. For instance, the English toxic sentence ‘You are disgusting’ can be rewritten as ‘You are impolite’, preserving the `PRON + AUX + ADJ` structure. This structural consistency helps models avoid producing awkward or unnatural phrasings during detoxification. Similar transformations can be applied in other languages. In Chinese, ‘**你真恶心**’ (‘you are disgusting’) can be detoxified to ‘**你真不礼貌**’ (‘you are impolite’), following the same `PRON + ADV + ADJ` structure, both maintaining the `PRON + AUX + ADJ` pattern and achieve the goal of detoxification.

Despite the usefulness of PoS alignment, we found that fine-tuning a large language model such as Qwen3-14B on multilingual detoxification tasks did not always yield ideal results. One key reason is that toxic expressions across languages often reflect deep-rooted cultural norms, idiomatic expressions, or social taboos that vary significantly between linguistic communities. For example, while direct insults may be common in English or Spanish online discourse, indirect sarcasm or context-dependent expressions are more prevalent in Chinese. As a result, models trained on one language’s toxic patterns may misinterpret or overlook harmful language in another, even when the PoS structure is similar.

Moreover lexical overlap across languages is limited, especially in low-resource settings. This makes it difficult for the model to generalize purely from word embeddings or fine-tuning. In contrast, PoS patterns provide a form of structural grounding that is more universal and can aid generalization across different languages. Therefore, PoS features might enhance consistency and grammaticality of detoxified outputs, while helping bridge the cultural and linguistic gap.

Q3. Imagine your detoxification models are applied to social media data. You can think of social media posts written by users of different culture, religion, socioeconomic status, and other aspects. Do you think the models would perform equally well across these diZerent domains? Include examples that support your argument. You can make up your own examples, in other words, they do not need to be from the training data set.

First, I want to clarify my position: models perform inconsistently across different domains, and developing targeted text detoxification models for different countries and groups is necessary.

I believe that text purification models are neither perfect nor useless. While purification models can indeed remove some toxic vocabulary, the definition of what constitutes a toxic word remains contentious. In this experiment, we judged the toxicity of vocabulary based on training set samples. However, the experimental dataset may have been compiled by specific individuals or groups, reflecting only the authors’ judgment of ‘toxicity’ and failing to encompass definitions of toxicity from users of different cultural, religious, and socioeconomic backgrounds. Therefore, the classification of toxic words and sentences is not absolute. Semantic differences exist significantly across different groups, and even within the same cultural sphere, word meanings evolve over time.

The ultimate goal of text detoxification is to remove all offensive vocabulary, but even if such models were developed, it would be difficult to eliminate all offensive expressions, as the positive or negative connotation of words is determined by their usage context, which is not always reflected in the immediate textual context. Without access to the speaker's complete background information, it is impossible to selectively remove offensive implications.

Consider the English sentence 'You are really good at working', which can be either praise or sarcasm. If expressed by a supervisor, it constitutes praise and may imply promotion; if said by a colleague, it could be genuine appreciation for responsibility or sarcasm about work intensification; in the context of someone balancing work and study but failing to complete group assignments, teammates saying this would carry sarcastic implications about using work to avoid responsibility.

The detoxification challenge of such sentences lies in their lack of inherently toxic words, yet they carry completely different meanings when spoken by different people, with contextual information rarely available as reference for model detoxification tasks. As there are a thousand Hamlets in a thousand minds, the same sentence carries different meanings in different contexts.

Within the same language, the global usage of Spanish exemplifies this issue. The word 'Coger' means 'to take' in Spain, as in 'coger el autobús' meaning 'to take the bus'; however, in Latin American countries like Mexico and Argentina, it means 'to have sexual intercourse' and is considered vulgar. Without developing region-specific models, Spanish expression in Spain would be restricted, as 'to take' is a common word, and text detoxification models would severely interfere with daily expression.

Across languages, the pronunciation 'suka' corresponds to a severe insult in Russian, meaning 'bitch', while in Malay it means 'like', as in 'suka makan' meaning 'like to eat'. The risk in the 'suka' case is relatively low, as people generally do not feel offended by unfamiliar text pronunciation resembling profanity, since they would not read it aloud for further interpretation.

Additionally, the 'de-stigmatization' of text leads to changes in toxicity classification. For example, in Chinese, some people avoid mentioning 'menstruation', instead using euphemisms like '例假' (monthly leave). Similar phenomena exist in other languages, such as 'Shark Week' in English and 'あの日' (that day) in Japanese. In recent years, Chinese internet has launched a menstruation 'de-stigmatization' movement, encouraging direct use of the term 'menstruation'. Supporters argue this helps reduce cultural taboos' potential oppression of women. For text detoxification models, treating menstruation as a 'toxic' word would hinder such equality movements. Similar situations may exist among other minority groups.

In conclusion, real-world text detoxification tasks are extremely complex, and models cannot perform consistently across users from different backgrounds. However, this does not render models meaningless. Models that remove potentially offensive vocabulary as much as possible, despite performance limitations, still make important contributions to building a safe and harmonious online environment.

Language Processing

References

[1] Daniil Moskovskiy, Daryna Dementieva, and Alexander Panchenko. Exploring cross-lingual textual style transfer with large multilingual language models. *arXiv preprint arXiv:2206.02252*, 2022.

[2] Zheyuan Ryan Shi, Claire Wang, and Fei Fang. Artificial intelligence for social good: A survey. *arXiv preprint arXiv:2001.01818*, 2020.

[3] Daniil Moskovskiy, Sergey Pletenev, and Alexander Panchenko. 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, 2024.

[4] Johannes Welbl, Amelia Glaese, Jonathan Uesato, Sumanth Dathathri, John Mellor, Lisa Anne Hendricks, Kirsty Anderson, Pushmeet Kohli, Ben Coppin, and Po-Sen Huang. Challenges in detoxifying language models. *arXiv preprint arXiv:2109.07445*, 2021.

[5] Daryna Dementieva, Daniil Moskovskiy, Nikolay Babakov, Abinew Ali Ayele, Naquee Rizwan, Frolian Schneider, Xintog Wang, Seid Muhie Yimam, Dmitry Ustalov, Elisei Stakovskii, Alisa Smirnova, Ashraf Elnagar, Animesh Mukherjee, and Alexander Panchenko. Overview of the multilingual text detoxification task at pan 2024. In Guglielmo Faggioli, Nicola Ferro, Petra Galuščáková, and Alba García Seco de Herrera, editors, *Working Notes of CLEF 2024 - Conference and Labs of the Evaluation Forum*. CEUR-WS.org, 2024.

[6] David Dale, Anton Voronov, Daryna Dementieva, Varvara Logacheva, Olga Kozlova, Nikita Semenov, and Alexander Panchenko. Text detoxification using large pre-trained neural models. *arXiv preprint arXiv:2109.08914*, 2021.

[7] Kazuki Aikawa, Shin Kawai, and Hajime Nobuhara. Multilingual inappropriate text content detection system based on doc2vec. In *2019 IEEE 8th Global Conference on Consumer Electronics (GCCE)*, pages 441–442. IEEE, 2019.

[8] Binny Mathew, Punyajoy Saha, Seid Muhie Yimam, Chris Biemann, Pawan Goyal, and Animesh Mukherjee. Hatexplain: A benchmark dataset for explainable hate speech detection. In *Proceedings of the AAAI conference on artificial intelligence*, volume 35, pages 14867–14875, 2021.

[9] Daryna Dementieva, Nikolay Babakov, Amit Ronen, Abinew Ali Ayele, Naquee Rizwan, Florian Schneider, Xintong Wang, Seid Muhie Yimam, Daniil Moskovskiy, Elisei Stakovskii, et al. Multilingual and explainable text detoxification with parallel corpora. *arXiv preprint arXiv:2412.11691*, 2024.

[10] Varvara Logacheva, Daryna Dementieva, Sergey Ustyantsev, Daniil Moskovskiy, David Dale, Irina Krotova, Nikita Semenov, and Alexander Panchenko. 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, 2022.

[11] Zhongyu Luo, Man Luo, and Aiguo Wang. Multilingual text detoxification using google cloud translation and post-processing. *Working Notes of CLEF*, 2024.

[12] Nikita Sushko. Pan 2024 multilingual textdetox: exploring different regimes for synthetic data training for multilingual text detoxification. *Working Notes of CLEF*, 2024.

[13] Jiangao Peng, Zhongyuan Han, Huan Zhang, Jingyan Ye, Chang Liu, Biao Liu, Mingcan Guo, Haoyang Chen, Zijie Lin, and Yujiao Tang. A multilingual text detoxification method based on few-shot learning and co-star framework. *Working Notes of CLEF*, 2024.

[14] Sai Saketh Aluru, Binny Mathew, Punyajoy Saha, and Animesh Mukherjee. Deep learning models for multilingual hate speech detection. *arXiv preprint arXiv:2004.06465*, 2020.

[15] Huimin Lu, Masaru Isonuma, Junichiro Mori, and Ichiro Sakata. Unidetox: Universal detoxification of large language models via dataset distillation. *arXiv preprint arXiv:2504.20500*, 2025.

[16] Ching-Yun Ko, Pin-Yu Chen, Payel Das, Youssef Mroueh, Soham Dan, Georgios Kollias, Subhajit Chaudhury, Tejaswini Pedapati, and Luca Daniel. Large language models can become strong self-detoxifiers. In *The Thirteenth International Conference on Learning Representations*, 2025.

[17] Canwen Xu, Zexue He, Zhankui He, and Julian McAuley. Leashing the inner demons: Self-detoxification for language models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 11530–11537, 2022.

[18] Mengru Wang, Ningyu Zhang, Ziwen Xu, Zekun Xi, Shumin Deng, Yunzhi Yao, Qishen Zhang, Linyi Yang, Jindong Wang, and Huajun Chen. Detoxifying large language models via knowledge editing. *arXiv preprint arXiv:2403.14472*, 2024.

[19] Xintong Wang, Yixiao Liu, Jingheng Pan, Liang Ding, Longyue Wang, and Chris Biemann. Chinese toxic language mitigation via sentiment polarity consistent rewrites. *arXiv preprint arXiv:2505.15297*, 2025.

[20] Griffin Floto, Mohammad Mahdi Abdollah Pour, Parsa Farinneya, Zhenwei Tang, Ali Pesaranghader, Manasa Bharadwaj, and Scott Sanner. Diffudetox: A mixed diffusion model for text detoxification. *arXiv preprint arXiv:2306.08505*, 2023.

[21] Varvara Logacheva, Daryna Dementieva, Sergey Ustyantsev, Daniil Moskovskiy, David Dale, Irina Krotova, Nikita Semenov, and Alexander Panchenko. 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, May 2022. Association for Computational Linguistics.

[22] Elisei Rykov, Konstantin Zaytsev, Ivan Anisimov, and Alexandr Voronin. Smurfcat at PAN 2024 textdetox: Alignment of multilingual transformers for text detoxification. In Guglielmo Faggioli, Nicola Ferro, Petra Galuscáková, and Alba García Seco de Herrera, editors, *Working Notes of the Conference and Labs of the Evaluation Forum (CLEF 2024), Grenoble, France, 9-12 September, 2024*, volume 3740 of *CEUR Workshop Proceedings*, pages 2866–2871. CEUR-WS.org, 2024.

[23] Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. Exploring the limits of transfer learning with a unified text-to-text transformer, 2023.

[24] Nils Reimers and Iryna Gurevych. Sentence-bert: Sentence embeddings using siamese bert-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, 11 2019.

[25] Daniel Jurafsky and James H. Martin. *Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition with Language Models*. 3rd edition, 2025. Online manuscript released January 12, 2025.

[26] Thomas Davidson, Dana Warmsley, Michael Macy, and Ingmar Weber. Automated hate speech detection and the problem of offensive language. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 11, pages 512–515, 2017.

[27] Slav Petrov, Dipanjan Das, and Ryan McDonald. A universal part-of-speech tagset. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC)*, pages 2089–2096. European Language Resources Association (ELRA), 2012.

Experiment Result

Table 6: Comprehensive comparison of model de-toxic performance on different language task

Model/Method	**lang**	**STA**	**SIM**	**XCOMET**	**J**
Duplicate and Delete	de	0.664636	0.942981	0.942414	0.589312
	en	0.853804	0.879577	0.629013	0.467257
	es	0.732938	0.886220	0.872034	0.568677
	zh	0.836024	0.822761	0.746286	0.523187
Backtranslation	de	0.869985	0.737630	0.797525	0.515980
	en	0.893916	0.844457	0.852331	0.648924
	es	0.812704	0.774368	0.815749	0.516028
	zh	0.875770	0.512826	0.565870	0.258276
MT0	de	0.871240	0.935994	0.949033	**0.776714**
	en	0.945130	0.874963	0.893869	**0.745441**
	es	0.908641	0.884833	0.902022	**0.730526**
	zh	0.754636	0.857583	0.850778	0.560873
Zero-shot	de	0.871830	0.821441	0.914300	0.656113
	en	**0.970803**	0.772793	0.898458	0.676924
	es	0.927694	0.758523	0.867553	0.611882
	zh	0.935257	0.694610	0.781316	0.519245
Few-shot	de	0.835389	0.876954	0.941782	0.690870
	en	0.923995	0.844139	0.887017	0.698281
	es	0.913200	0.805540	0.879664	0.647423
	zh	0.875098	0.746202	0.805746	0.542726
RAG (multilingual)	de	0.911321	0.874038	0.941656	0.755610
	en	0.972846	0.795746	0.894842	0.695079
	es	0.867925	0.797830	0.864987	0.599163
	zh	0.952647	0.724934	0.771647	0.554005
RAG (separate)	de	0.860828	0.885411	0.944813	0.719727
	en	**0.976788**	0.791793	0.880300	0.688571
	es	0.927064	0.792387	0.856235	0.634962
	zh	**0.962761**	0.729662	0.820746	0.580624
Fine-tuning Qwen	de	0.860828	0.885411	**0.944813**	0.610805
	en	**0.976788**	0.791793	0.880300	0.631480
	es	0.927064	0.792387	0.856235	0.645149
	zh	**0.962761**	0.729662	0.820746	0.489986

Language examples for analysis

Table 7: Detoxification results for a German sample input across different methods

Method	Output
Ground Truth	Ich würde Grüne mit einem Geschwür der Welt bezeichnen. Ist aber meine persönliche Meinung. Ein Geschwür kann einen Körper zerstören oder ein ganzes Land.
Delete	Ich würde Grüne mit einem Geschwür am Welt bezeichnen. Ist aber meine persönliche Meinung. Ein Geschwür kann einen Körper zerstören oder ein ganzes Land.
Backtranslation	Ich würde Green als das Geschwür der Welt beschreiben, aber das ist meine persönliche
MT0	Ich würde Grüne mit einem Geschwür an der Welt bezeichnen. Ist aber meine persönliche Meinung. Ein Geschwür kann einen Körper zerstören oder ein ganzes Land.
Zero-shot	Ich würde Grüne mit einer belastenden und selbstgerechten Haltung bezeichnen. Es ist jedoch meine persönliche Meinung. Eine solche Haltung kann sowohl eine Person als auch ein ganzes Land schwer beeinträchtigen.
Few-shot	Ich würde Grüne mit einem Geschwür am Arsch der Welt bezeichnen. Ist aber meine persönliche Meinung. Ein Geschwür kann einen Körper zerstören oder ein ganzes Land.
RAG-test	Ich würde die Grünen als einen schädlichen Einfluss am Rand der Welt bezeichnen. Ist aber meine persönliche Meinung. Ein schädlicher Einfluss kann negative Auswirkungen auf einen Körper oder ein Land haben.
RAG-SEP-test	Ich würde die Grünen als eine problematische Kraft bezeichnen. Das ist meine persönliche Meinung. Eine problematische Kraft kann negative Auswirkungen auf einen Körper oder ein ganzes Land haben.
LoRA-Fine-tuning	Ich würde die Grünen als eine Art Geschwür am Arsch der Welt bezeichnen. Das ist jedoch meine persönliche Meinung. Ein Geschwür kann einen Körper oder ein ganzes Land zerstören.

Table 8: Detoxification results for a sample input across different methods

Method	**Output**
Original	**fucking cunt of a person.**
Ground Truth	**A bad person.**
Delete	of a person
Backtranslation	bad person
MT0	I don't like this person
Zero-shot	A very unpleasant person.
Few-shot	That was an unpleasant person.
RAG-test	a person with poor judgment.
RAG-SEP-test	a person.
LoRA-Fine-tuning	That person is very unpleasant.

Table 9: Detoxification results for a Spanish sample input across different methods

Method	**Output**
Original	Pero que mierda estas hablando subnormal de los cojones MAMAAAAAAAA GORDOOOOOOOO.
Ground Truth	Pero que estás hablando
Delete	Pero que estas hablando subnormal de los MAMAAAAAAAA GORDOOOOOOOO.
Backtranslation	£De qué estás hablando?
MT0	£Pero de qué estás hablando?
Zero-shot	Pero no entiendo qué estás diciendo, MAMÁ. Debo abordar este comportamiento inapropiado.
Few-shot	Pero qué estás hablando subnormal MAMAAAAA GORDO.
RAG-test	Pero que mierda estás hablando. ąMAMÁ!
RAG-SEP-test	Pero, £qué estás hablando? Mamá, por favor, aclárate.
LoRA-Fine-tuning	Pero £qué estás diciendo? Eso no tiene sentido, mamá, eres muy gorda.

Table 10: Detoxification results for a sample input across different methods

Method	**Output**
Original	对对对，都是小仙女，小仙女没有错，都是中国男人的错，中国男人应该灭绝！
Ground Truth	对对对，都是自我感觉过于良好的女性，自我感觉过于良好的女性没有错，都是中国男人的错，中国男人应该灭绝！
Delete	对对对，都是小，小没有错，都是中国男人的错，中国男人应该灭绝！
Backtranslation	是的, 是的, 是的, 这是小仙女, 小仙女是
MT0	对对对, 都是自我感觉过于良好的女性, 她们没有错, 都是中国男人的错, 中国男人应该消失!
Zero-shot	对对对，都是小仙女，小仙女没有错，都是中国男人的错，中国男人需要改变。
Few-shot	对对对，都是小仙女，小仙女没有错，都是中国男人的错，中国男人需要改进。
RAG-test	对对对，都是女性的问题，女性没有错，都是中国男人的问题，中国男人需要改变！
RAG-SEP-test	对对对，都是某些年轻女性，某些年轻女性没有错，都是中国男性的责任，中国男性需要反思和改进！
LoRA-Fine-tuning	每个人都应该被平等对待，性别不应成为评判对错的标准，我们应当共同促进社会的和谐与进步。