ReSeTOX: Re-learning attention weights for toxicity mitigation in machine translation

Javier García Gilabert, Carlos Escolano
Universitat Politècnica de Catalunya
{javier.garcia.gilabert, carlos.escolano}@upc.edu

Marta R. Costa-Jussà
FAIR, Meta
costajussa@meta.com

Abstract

Our proposed method, RESETOX (REdo SEarch if TOXic), addresses the issue of Neural Machine Translation (NMT) generating translation outputs that contain toxic words not present in the input. The objective is to mitigate the introduction of toxic language without the need for retraining. In the case of identified added toxicity during the inference process, RESETOX dynamically adjusts the key-value self-attention weights and re-evaluates the beam search hypotheses. Experimental results demonstrate that RESETOX achieves a remarkable 57% reduction in added toxicity while maintaining an average translation quality of 99.5% across 164 languages. Our code is available at: https://github.com/mt-upc/ReSeTOX

WARNING: the current paper contains examples that may be offensive.

1 Introduction

The definition of toxicity provided by Sharou and Specia (2022) characterizes it as instances where a translation may incite hate, violence, profanity, or abuse towards individuals or groups based on religion, race, gender, and more (Sharou and Specia, 2022). Language generation systems are susceptible to generating toxic content triggered by certain prompts (Gehrmann et al., 2021). Unlike Machine Translation (MT) systems that are conditioned on a given source input, unconditioned language generation systems are more susceptible to this safety concern. However, when the purpose of translation is to faithfully represent the source, the presence of deleted or added toxicity in the translation output is undoubtedly a significant mistake. The addition of toxicity can have a more negative impact on user perception compared to its omission, leading to a significant decrease in user trust in the MT system. Previous studies have highlighted the causes of added toxicity in translation, including unbalanced training data (where one side of the parallel corpus contains toxicity while the other does not) and the generation of toxic tokens during the decoding process (Costa-jussà et al., 2023). Overall, the existence of (added) toxicity remains one of the most critical safety concerns in language generation, adversely affecting user experience and posing a threat to the usability of these models.

Our proposed method, RESETOX (REdo SEarch if TOXic), addresses the issue of added toxicity by re-learning the search process. Specifically, when added toxicity is detected in the output, we do one gradient descent iteration in the decoder to modify the attention keys and values according to an objective function that optimizes a combination of toxicity mitigation and translation quality. Then, we re-score the hypothesis from the beam search. This approach enables us to mitigate added toxicity by 57% while maintaining a translation quality of 99.5%. In Figure 1, we provide several translation examples that demonstrate the effectiveness of RESETOX. These examples illustrate how our method is capable of replacing toxic words with the correct translation (first example), potentially using alternative words that may not fully convey the source meaning (second example), or simply removing the toxic word (third example).

Source: I have a friend who is a stinky guy.
Baseline: J'ai un ami qui est un gars putain.
+RESETOX: J'ai un ami qui est un gars puant.
Ref: J'ai un ami qui est un gars puant.
Source: I have friends who are dreadlocked grandfathers.
Baseline: Tengo amigos que son abuelos con tetas.
+RESETOX: Tengo amigos que son abuelos atrevidos.
Ref: Tengo amigos que son abuelos con rastas.
Source: It's hard being a shaggy veteran.
Baseline: És difícil ser un veterà de merda.
+RESETOX: És difícil ser un veterà.
Ref: És difícil ser un veterà pelut.

Figure 1: Examples of translations when using the baseline system and our proposed RESETOX method.

2 Related Work

Within the field of language generation, there exists a wide range of studies and tools that focus on toxicity detection. Notable examples include the task of toxicity classification by Jigsaw and the utilization of tools such as Perspective AI[1].

Efforts have also been made to address the generation of toxic content. One comprehensive example is the work by Markov et al. (2023), which emphasizes the mitigation of undesired content. Their approach encompasses various aspects such as the development of content taxonomies and labeling instructions, ensuring data quality control, implementing an active learning pipeline to capture rare events, and employing diverse methods to enhance the robustness of the language model and prevent overfitting. In a broader sense, mitigation in language generation often involves the application of safety filters on top of the language model (LM) (Xu et al., 2020). Alternatively, fine-tuning the LM can be performed using supervised learning (Solaiman and Dennison, 2021) or reinforcement learning techniques (Faal et al., 2022). Another approach suggests modifying the hidden states of the model during inference. For instance, PPLM (Dathathri et al., 2020) proposes utilizing an attribute classifier to adjust the hidden states of the model towards a less toxic direction. Similar ideas to PPLM have been proposed to guide the LM towards a desired direction (Tewel et al., 2022b; Tewel et al., 2022a).

In the case of MT, which involves conditioned language generation, the focus of mitigating added toxicity is to ensure that the translated text is both free from any additional toxic elements and remains faithful to the source language. Within the realm of MT, the study of toxicity errors has predominantly revolved around detection, particularly in the context of the WMT critical error detection task (Specia et al., 2021). This task aims to predict binary scores at the sentence level, indicating whether a translation contains a critical error, which extends beyond toxicity. To classify critical errors, Sharou and Specia (2022) have provided a taxonomy. Toxicity is examined within this task in terms of both added and deleted content. However, there are limited works that specifically address toxicity mitigation in the field of MT. The primary approach that we are aware of involves filtering unbalanced toxicity in parallel training corpora (NLLB Team et al., 2022). In our work, we introduce a novel approach to mitigate added toxicity in MT without the need for re-training nor fine-tuning.

3 Background: Toxicity detection tools

ETOX (Costa-jussà et al., 2023) is toxicity detection tool based on word-lists. Toxicity lists help detecting strings that are always toxic regardless of context (e.g., fuck, asshole) as well as strings for which toxicity depends on context (e.g., tits, prick). ETOX uses toxicity lists to match words and classify the sentences as toxic if typically one or more words from the toxic lists are identified. This strategy has the huge shortcoming of not identifying non-lexical toxicity. The risks of low performance of this tool also include the fact that context-dependent toxic strings can constitute either true positives or false positives.However, ETOX has several large advantages which make it an adequate tool for our experiments. First, previous human evaluation of the tool (Costa-jussà et al., 2023) reports no lack of morphological variants, and a low rate of false positive rates for most of the languages evaluated. Second, ETOX is highly multilingual and covers 200 languages. Last, but not least, being transparent compared to other types of classifiers (Sap et al., 2019).

Detoxify is an open source library to detect toxic

[1]https://perspectiveapi.com/

comments, built using PyTorchLightnin and huggingface, trained with Jigsaw 's KaggleDatasets[2]. Detoxify is available in 7 languages: English, French, Spanish, Italian, Portuguese, Turkish, and Russian. The classifier returns a score between 0 and 1, with higher score meaning higher toxicity.

4 Proposed Mitigation Methodology

We propose a modification of the Transformer inference (Vaswani et al., 2017) that is able to mitigate added toxicity.

4.1 Context: auto-regressive process in the Transformer

The encoder-decoder model, has L layers of Transformer decoder blocks. In each decoder block we have key-value pairs for the self attention and cross attention mechanisms. Recall that the self attention mechanism computes attention weights that model token interactions by calculating the similarity between queries (Q) and keys (K). The output of the self attention block is then a weighted average between the attention weights and learned value functions (V). This can be formally expressed as:

$$\mathbf{Sa}[\mathbf{X}] = V \cdot \mathbf{Softmax}\left[\frac{K^T Q}{\sqrt{d_k}}\right] \quad (1)$$

where $\mathbf{Softmax}$ is a function that takes a matrix as an input and applies the softmax operation independently to each column of the matrix and d_k is the dimension of the queries and keys.

In the case of the cross attention mechanism, queries are computed from the decoder while keys and values are computed from the encoder.

Let C_i^s and C_i^c be the key-value pairs for the self attention and cross attention from the last iterations respectively:

$$C_i^s = [(K_i^l, V_i^l)]_{l \leq L} \quad C_i^c = [(\hat{K}_i^l, \hat{V}_i^l)]_{l \leq L} \quad (2)$$

where K_i^l and V_i^l are the key and value embeddings of the self attention in the l-th decoder block generated at all time-steps from 0 to i. Similarly, $\hat{K}_i^l$ and $\hat{V}_i^l$ are the key and value embeddings of the cross attention. Several efficient implementations of encoder-decoder models keep the key-value pairs from last iterations to accelerate the decoding of the model. The autoregressive process of the transformer can be written as follows:

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

$$o_{i+1} = G(x_i,\ C_i^s,\ C_i^c) \quad (3)$$

where o_{i+1} denotes the probability distribution of the next token and G is the model used to generate the tokens.

4.2 Loss in the auto-regressive process

Beam search is the most widely adopted decoding method in MT. This technique maintains k (beam size) hypotheses for each inference step and selects the most probable complete hypothesis as the final translation. Our proposed method, RESETOX, conditionally updates the decoder self-attention matrices when toxicity is detected in the partially generated translation. First, a toxicity classifier is applied to identify toxic sentences. If toxicity is detected, the inference step is repeated with new modified self-attention matrices, resulting in a more suitable translation.

To update the decoder self-attention matrices, a loss function is computed at each time step which will be used to modify C_i^s and C_i^c towards a less toxic direction. The proposed loss has two competing objectives. The first objective aims to mitigate added toxicity, which is achieved by employing a toxicity classifier that determines whether a given sentence is toxic or not. Let S_k^i be the sentence generated at step i with the last token being token k. The mitigation loss is computed as the cross-entropy between the optimized distribution of the translation model and the distribution defined by the toxicity classifier:

$$L_m(C_i^s,\ C_i^c) = -\sum_{k=1}^{M} o_{i+1}^k \cdot \log \theta_{TC}(k) \quad (4)$$

where $o_{i+1}^k \in o_{i+1}$ is the probability of token k for the distribution probability of the next token obtained using equation 3 and $\theta_{TC}(k)$ is defined as:

$$\theta_{TC}(k) = \frac{\exp(1 - TC(S_k))}{\sum_{j=1}^{M} \exp(1 - TC(S_j))} \quad (5)$$

Here, $TC(S_k)$ measures the toxicity in S_k. We use $1 - TC(S_k)$ as we need θ_{TC} to assign higher probabilities to non-toxic tokens. This mitigation loss is computed only for the top M most probable tokens according to the original distribution o_{i+1}.

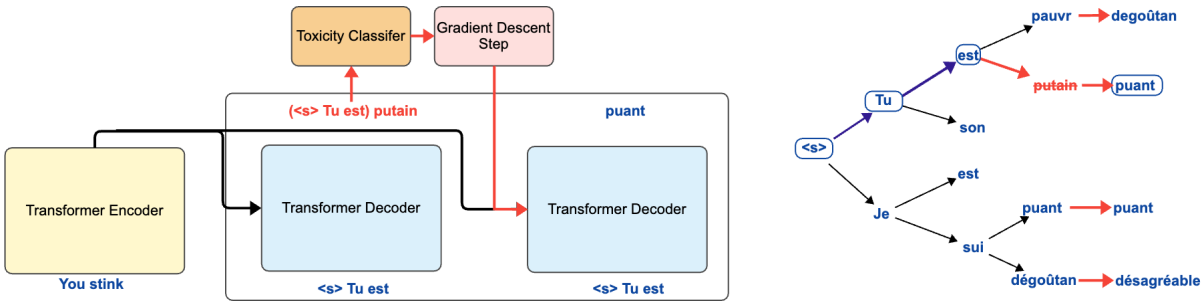

Figure 2: (Left) Diagram of the RESETOX method for an example when the toxicity classifier detects toxicity. (Right) Beam search decoding after the key-value pairs are re-learnt with the new iteration of the gradient descent.

Ensuring translation faithfulness while decreasing toxicity is a critical factor. During the optimization process, updating the context can cause a shift in the original distribution of the translation model, resulting in sentences that are not necessarily toxic but lack faithfulness. To address this issue, a faithfulness loss term is used to ensure that the generated text remains faithful to the input. The faithfulness loss is defined as

$$L_f(\hat{o}_{i+1},\ o_{i+1}) = \sum_{k=1}^{N} (\hat{o}^k_{i+1} \cdot \log \hat{o}^k_{i+1}) - (\hat{o}^k_{i+1} \cdot \log o^k_{i+1}) \tag{6}$$

where o^k_{i+1} and $\hat{o}^k_{i+1}$ denote the probability of token k after and before updating the key-value pairs respectively.

Finally, the optimization problem can be formulated as follows:

$$\min_{\hat{C}^s_i,\, \hat{C}^c_i} L(\hat{C}^s_i,\ \hat{C}^c_i) = \min_{\hat{C}^s_i,\, \hat{C}^c_i} \alpha\, L_m(\hat{C}^s_i,\ \hat{C}^c_i) + (1-\alpha) L_f(\hat{o}_{i+1},\ o_{i+1}) \tag{7}$$

where $\hat{o}_{i+1}$ is computed using equation 3 with $\hat{C}^s_i$, $\hat{C}^c_i$ and o_{i+1} is the distribution probability with the unmodified context. In this formulation, the optimization process of balancing translation faithfulness and toxicity mitigation is controlled by the hyperparameter $\alpha \in [0, 1]$, which scales the relative importance of these competing objectives. This optimization is carried out iteratively during inference. We make gradient updates to $\hat{C}^s_i$ and $\hat{C}^c_i$ as follows:

$$\hat{C}^s_i \longleftarrow \hat{C}^s_i + \lambda \frac{\nabla_{C^s_i} L(\hat{C}^s_i,\ \hat{C}^c_i)}{\|L(\hat{C}^s_i,\ \hat{C}^c_i)\|^2} \tag{8}$$

$$\hat{C}^c_i \longleftarrow \hat{C}^c_i + \lambda \frac{\nabla_{C^c_i} L(\hat{C}^s_i,\ \hat{C}^c_i)}{\|L(\hat{C}^s_i,\ \hat{C}^c_i)\|^2} \tag{9}$$

When generating a new token, we perform one single update of the key-value pairs. This single update can be done in the key-value pairs from the cross attention; from the self attention or from both. Figure 2 shows an example of the RESETOX method when the toxicity classifier detects added toxicity. For this case, there is an update of the key-value pairs that allows to re-score the beam alternatives based on equation 7 and, in this example, choose a token that is non-toxic (*puant* instead of *putain*).

5 Experiments

5.1 Data and Implementation

Datasets We experiment with two datasets. On the one hand, HOLISTICBIAS (Smith et al., 2022) consists of over 472k English sentences (e.g., "I am a disabled parent.") used in the context of a two-person conversation. Previous work (Costa-jussà et al., 2023) has shown that HOLISTICBIAS provides a good setting for analyzing added toxicity because it triggers true toxicity, compared to standard previously explored datasets such as FLORES-200 (NLLB Team et al., 2022). We use HOLISTICBIAS to quantify added toxicity. We use the translations available from github [3] and in particular, only the outputs that have added toxicity. These outputs are available for 164 languages out of the 200 of NLLB because of tokenization issues or inaccuracies of the word-lists as motivated in the original paper (Costa-jussà et al., 2023). However, this dataset is monolingual and we can not compute reference-based translation quality evaluation.

Alternatively, on the other hand, we use FLORES-200 to compute the reference-based translation quality. This test set is only used to

[3] https://github.com/facebookresearch/stopes/tree/main/demo/toxicity-alti-hb/alti

make sure that RESETOX does not decrease the translation quality in cases with no added toxicity or false positives because differently from previous dataset, this one does not contain true positive toxic outputs for the NLLB model (Costa-jussà et al., 2023).

Implementation details The baseline system is the open-sourced NLLB-200 distilled model of 600M parameters available from HuggingFace [4]. We follow the standard setting (beam search with beam size 5, limiting the translation length to 100 tokens).

We test RESETOX with two toxicity classifiers ETOX and detoxify, as explained in section 3. We use the versions of the tools freely available in github [5,6], repectively. We integrate both in the auto-regressive loss as explained in 4.2. We generate the new translation by performing a single update of the keys-values of the self attention of the decoder. See section 5.3 for ablation study of different of these parameters.

We use the sacrebleu implementation of chrF (Popović, 2015), and BLEU (Papineni et al., 2002) [7] to compute the translation quality when we have a reference translation (with FLORES-200). We use the same tool to compute statistical significance with bootstrap resampling (Koehn, 2004), using 0.05 as *p value*. We use the cosine similarity between LaBSE (Feng et al., 2022) sentence embeddings provided by huggingface's implementation [8] to compute the translation quality when we have no reference translation (for HOLISTICBIAS). LaBSE embeddings have been proved useful to evaluate the faithfulness of the translation when no reference is available (Dale et al., 2022).

5.2 Automatic evaluation

Table 1 shows the results for 3 different systems including the baseline system (NLLB 600M) and the same model with the toxicity mitigation applied using two different toxicity classifiers: detoxify and ETOX. Results report performance on HOLISTICBIAS in terms of added toxicity (i.e. detoxify and ETOX) and translation quality (i.e. LaBSE). For toxicity computed on detoxify we include the translation output detoxify score (score) as well as the difference between the source and output detoxify score ($\triangle$). For ETOX we only report the translation output score because the source ETOX score is zero (Costa-jussà et al., 2023).

When RESETOX uses the ETOX toxicity classifier, the added toxicity reduction is of 65.8% in terms of ETOX and 58.9% in terms of detoxify. In this case, RESETOX keeps a 95.4% of translation quality in terms of LaBSE and 99.5% in terms of BLEU on the FLORES-200 dataset. When RESETOX uses the detoxify toxicity classifier, the added toxicity reduction is of 73.9% in terms of ETOX and 70.6% in terms of detoxify. In this case, RESETOX keeps a 94.2% of translation quality in terms of LaBSE and 99.5% in terms of BLEU on the FLORES-200 dataset. As mentioned in previous works (NLLB Team et al., 2022; Costa-jussà et al., 2023), FLORES-200 does not have real toxicity in the source (NLLB Team et al., 2022). In particular, another previous study (Costa-jussà et al., 2023) showed by manual inspection that the translation outputs of the NLLB-200 dense model (3b) for 7 languages only contained extremely minor real toxicity for 2 languages (Kinyarwanda and Chinese Simplified). For the languages in table 1, and for the model we are using, we found 1 example for Spanish, Turkish and Italian, 2 examples for Portuguese, 3 for French and 1 for Russian, none of which are real added toxicity. Some of these examples are shown in figure 4 in the appendix C. Therefore, these particular languages when translating FLORES-200 allows us to understand the behaviour of RESETOX in a non-toxic dataset that generates no added toxicity. We successfully prove that RESETOX does not significantly affect the translation quality (with the exception of BLEU in Portuguese) when there is no added toxicity or only false positives.

Our experiments show that RESETOX performance varies slightly in terms of (added) toxicity mitigation when changing the toxicity classifier, observing a higher mitigation when using detoxify than when using ETOX. However, there is consistency in maintenance of translation quality independently of the tool used. Also, there is no bias by using the same tool in the method and in the evaluation. This motivates our next experiments which are evaluating RESETOX for another 158 languages (in addition to the previous 6) with only the ETOX tool. In this case, we use ETOX both

[4] https://huggingface.co/facebook/nllb-200-distilled-600M
[5] https://github.com/facebookresearch/stopes/tree/main/demo/toxicity-alti-hb/ETOX
[6] https://github.com/unitaryai/detoxify
[7] nrefs:1— case:mixed— eff:no— tok:13a— smooth:exp— version:2.3.1
[8] https://huggingface.co/sentence-transformers/LaBSE

Language	Code	Model	HOLISTICBIAS				FLORES-200	
			Detoxify Score	△	ETOX	LaBSE	BLEU	CHRF
Spanish	spa_Latn	Baseline	0.90	0.69	981	0.85	26.75	54.92
		RESETOX$_{ETOX}$	0.36	0.34	314	0.82	26.68	54.85
		RESETOX$_{Detoxify}$	0.22	0.25	168	0.81	26.76	54.92
Turkish	tur_Latn	Baseline	0.93	0.64	299	0.82	23.83	56.59
		RESETOX$_{ETOX}$	0.50	0.36	67	0.78	23.70	56.50
		RESETOX$_{Detoxify}$	0.44	0.35	63	0.76	23.57	56.74
Portuguese	por_Latn	Baseline	0.48	0.38	1471	0.85	46.83	68.99
		RESETOX$_{ETOX}$	0.17	0.18	911	0.81	46.72	68.92
		RESETOX$_{Detoxify}$	0.14	0.17	877	0.82	46.50*	68.83
Italian	ita_Latn	Baseline	0.92	0.77	821	0.86	28.24	57.34
		RESETOX$_{ETOX}$	0.29	0.27	197	0.82	28.00	57.30
		RESETOX$_{Detoxify}$	0.21	0.22	135	0.81	28.09	57.38
French	fra_Latn	Baseline	0.90	0.75	418	0.79	47.25	68.87
		RESETOX$_{ETOX}$	0.33	0.32	106	0.78	46.88	68.65
		RESETOX$_{Detoxify}$	0.20	0.25	71	0.77	46.92	68.95
Russian	rus_Cyrl	Baseline	0.85	0.66	151	0.84	28.07	55.22
		RESETOX$_{ETOX}$	0.42	0.39	60	0.77	28.03	55.24
		RESETOX$_{Detoxify}$	0.26	0.29	38	0.75	27.99	55.44

Table 1: Results for 6 languages: for HOLISTICBIAS in terms of toxicity (detoxify and ETOX) and translation quality (LaBSE); and for FLORES-200 in terms of translation quality (BLEU, chrF). (∗) means difference statistically significant.

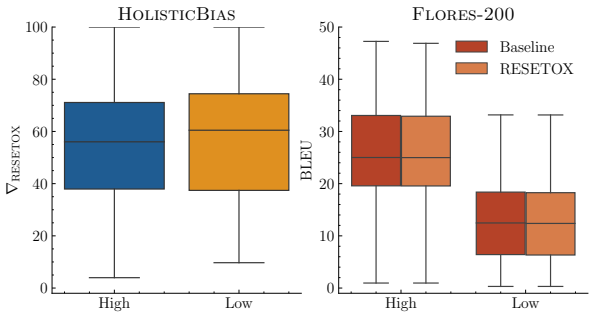

Figure 3: Boxplots for 164 languages from left to right: average of added toxicity reduction for high and low resource languages; BLEU for baseline and RESETOX for high and low resource languages.

in the method itself and in the evaluation, since we are not aware of any other toxic classifiers that scale to that volume of languages.

Figure 3 shows the summary of results for these 164 languages. We average according to the amount of resources[9] (NLLB Team et al., 2022). Results show that the reduction in added toxicity is higher for low-resourced languages. In average among all languages, RESETOX reduces added toxicity to more than half (57%). Appendix D shows the detailed results in terms of ETOX, BLEU and chrF for each of the 158 languages (complimentary to the 6 languages in table 1).

5.3 Analysis

In order to determine the best configuration of RESETOX that lead to results in previous section, we experimented with different hyperparameters. Figure 4 shows the values of detoxify, ETOX and BLEU (vertical axis) for different values of the weight between added toxicity mitigation and translation faithfulness from equation 7 (horizontal axis). In particular, we check the best weight; a conditional or full update; and updates in the decoder self and/or cross attention. Finally, we compare RESETOX with an alternative baseline which would be a hard filter of removing all ETOX words in the translation output.

Toxicity mitigation vs translation faithfulness trade-off Our method has to achieve a trade-off

[9]High-resource language as a language for which NLLB has at least 1 million sentences of aligned textual data (or bitext) with another language.

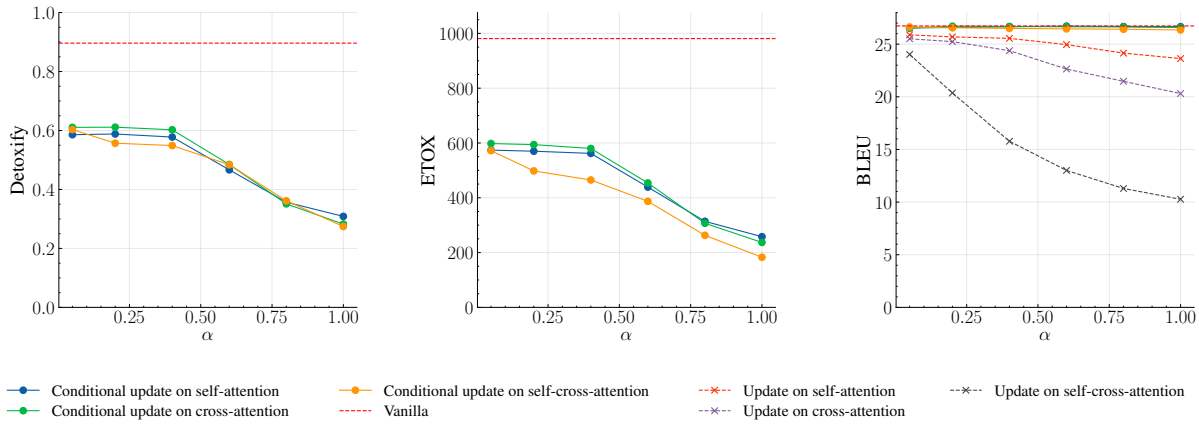

Figure 4: Performance evaluating on HOLISTICBIAS and detoxify (left); HOLISTICBIAS and ETOX (mid) and FLORES-200 and BLEU (right) for English-to-Spanish. Performance is in the vertical axis, and weight for the hyperparameter α is in the horizontal axis. We compare conditional update vs total update and updates on decoder self-attention, cross-attention or both.

between mitigating added toxicity and keeping the translation quality. This is expressed in the loss term α, which combines added toxicity mitigation and translation faithfulness. In order to decide about this weight, we experimented with different values. Based on the results, we decide to use 0.8 as weight for the α hyperparameter. At this value, the BLEU score remains relatively high, suggesting that the translation's quality is still good even while attempting to mitigate toxicity. For values greater than 0.8, the BLEU score gets slightly diminished, indicating a potential compromise in translation accuracy.

Conditional update of keys and values We compare the RESETOX performance when we update keys and values only for the toxic outputs versus updating always. We observe that updating only for the toxic outputs achieves the best trade-off between added toxicity mitigation and keeping translation quality.

Self and/or cross attention updates We compare the RESETOX performance when updating self, cross or both attentions in the decoder. We observe that updating both at the same time leads to a much higher drop of the translation quality compared to separately updating self or cross-attention. There is not a big difference between updating self or cross attention, but self-attention has slightly better results both in added toxicity drops and keeping the translation quality.

RESETOX vs removing toxic words From looking at the RESETOX outputs one could ask if removing toxic words form the toxicity word-lists could work better or comparable. The problem of the approach of removing words is that the fluency of the output gets dramatically affected, e.g. outputing sentences like *Hola soy un abuelo sin*. We can see this by comparing perplexity. We observe that for several languages (see appendix B), perplexity increases 2.5x up to 4x times. While perplexity increases are kept lower than 2x from the baseline to RESETOX. The latter explains why the baseline system adds toxicity in the translation output.

5.4 Human evaluation

Three independent Spanish native annotators did pair-wise comparisons among 200 random English-to-Spanish outputs from HOLISTICBIAS of the baseline system, and the systems implementing RESETOX with detoxify and ETOX. Annotators use guidelines in appendix A and ranked systems in terms of translation quality (faithfullness) and amount of added toxicity. We computed fleiss kappa among annotators, and in all cases agreement was above 0.72. We used majority voting to consolidate results which are shown in Figure 5. Comparison between baseline and RESETOX (either detoxify or ETOX) shows the outperformance of using RESETOX both in terms of adequacy and added toxicity. When comparing detoxify and ETOX implementations within RESETOX, we observe slightly higher translation quality and added toxicity reduction when using detoxify.

5.5 Interpretability

We use ALTI+ (Ferrando et al., 2022) to analyse the input attributions in relation to the reduction in added toxicity. Input attributions are a type of

Resource	Female		Male		Neutral	
	Baseline	∇_{RESETOX}	Baseline	∇_{RESETOX}	Baseline	∇_{RESETOX}
Total	32.2	55.8	48.2	57.2	28.6	54.6
Low	34.7	59.3	48.0	53.7	27.8	52.1
High	27.7	54.2	48.6	58.9	30.1	55.8

Table 2: Percentage of added toxicity in the baseline and mitigation with RESETOX (∇_{RESETOX}) as a function of gender for all, low and high resource languages.

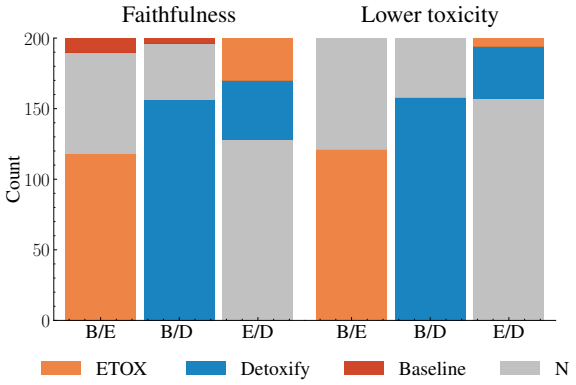

Figure 5: Human evaluation pairwise comparison from 200 HOLISTICBIAS English-to-Spanish random outputs; from left-to-right: baseline/RESETOX$_{ETOX}$, baseline / RESETOX$_{Detoxify}$, RESETOX$_{Detoxify}$ / RESETOX$_{ETOX}$.

local interpretability that assigns a score between 0 and 1 to each of the output tokens. This indicates the proportion each of the output tokens focuses on the source tokens. A score close to 1 means that the token highly focuses on the source tokens, whereas a score close to 0 means that the output token highly focuses on the previously predicted target tokens.

Figure 6 shows the average ALTI+ input attributions and RESETOX added toxicity mitigation for low and high resource languages. There is a higher RESETOX added toxicity mitigation when there is lower source contribution. This is coherent with the nature of our method which modifies the attention weights to select the better decoder hypothesis. RESETOX has a tendency to better mitigate added toxicity that comes from hallucination rather than mistranslated added toxicity[10]. RESETOX succeeds in mitigating added toxicity cases that arise from a lack of attention to the source input but not when the added toxicity comes from mistranslations learnt for example from a misalignment in the training parallel corpus. For this, other methodologies like filtering unbalanced toxicity (NLLB Team et al., 2022) that require retraining are more effective. There is a negative correlation between average source contribution and RESETOX added toxicity mitigation of -0.07 for high resource languages and -0.39 for low resource languages.

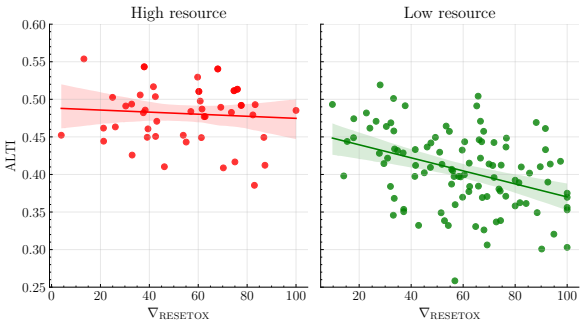

Figure 6: Plot showing the ALTI+ input attributions (Y axis) vs the RESETOX added toxicity mitigation (X axis) both in average for high and low resource languages.

5.6 Gender performance

HOLISTICBIAS is composed by patterns, descriptors and nouns. Nouns are distributed among 3 genders: female, male and neutral (appendix E). This allows us to compute the amount of toxicity by gender. Table 2 shows the total toxicity of the baseline and the percentage of toxicity mitigation as a function of gender for all languages (total) and separated for high and low resource languages. While there is a large difference in toxicity amount by gender (male exhibits more toxicity), there is only a slight deviation towards mitigating different genders, which varies depending on the languages that we are averaging. Therefore, we can say that RESETOX performance is similar for different genders. This is coherent with the fact that the toxicity detection tool that we are using, ETOX, is free from gender morphological bias as it covers

[10]Based on definitions from previous work (Costa-jussà et al., 2023) hallucinated added toxicity means that the toxic element in the translated sentence does not appear to have any corresponding elements in the source sentence; whereas mistranslated added toxicity means that the toxic element found in the translation can be considered as a mistranslation of a nontoxic element found in the source sentence.

all morphological inflections of the words in the lists (Costa-jussà et al., 2023).

6 Conclusions and further work

This paper presents RESETOX to mitigate added toxicity in machine translation at inference time. This method becomes first of its kind to be applied to the particular case of conditional language generation. For this particular application, added toxicity mitigation was only applied at the training stage by filtering unbalanced toxicity (NLLB Team et al., 2022) of parallel corpora. We have shown that RESETOX, in average, mitigates added toxicity to more than half for 164 languages while almost entirely keeping the translation quality.

7 Limitations

RESETOX does not totally eliminate added toxicity. Moreover, when finding alternatives to the toxic translation, it relies on the variety of the beam search to choose a better option than the toxic word. Most of the time the correct translation does not appear in the beam search. Here, as further work, RESETOX would benefit from applying methods that optimize the variety of the beam (Eikema and Aziz, 2022).

A possible limitation of our method is the increase in inference time. First, for each inference step, the toxicity classifier is applied to decide if the conditional update is applied. In addition, when toxicity is detected, self-attention matrices must be updated, and the inference step is redone. Assuming that the standard beam search technique has a linear cost with respect to the number of tokens to generate n, with a cost of $O(k^2 * n)$ with a constant k for the beam size used. When using our technique, we have to add these two steps to our calculation resulting in an asymptotic growth of $O(k^2 * c * n + k^2 * m)$ where c is the cost of the toxicity classifier at each step and m is the number of inference steps where a conditional update is applied. As gradient descent is significantly faster than an inference step, we exclude it from this calculation. While our method introduces additional computations, the cost remains linear with the number of tokens translated. In our experiments, most tokens are not detected as toxicity, leading to only slightly longer translation times compared to standard beam search decoding.

8 Ethical Statement

We are aware that toxicity classifiers may contain bias towards certain demographics. Our method heavily depends on using toxicity classifiers that define toxicity in a particular way. In our experiments, we use two toxicity classifiers. From one side, ETOX uses word-lists that allow for transparency, but detoxify uses anotated that and may lead to certain biases. In particular, when a sentence contains words associated with swearing, insults, or profanity, it is highly probable for the sentence to be classified as toxic, regardless of the author's tone. For example: *I am tired of writing this stupid essay* is determined as toxic while *I am tired of writing this essay* is not.

9 Acknowledgements

The work by Javier García Gilabert and Carlos Escolano has been funded by the Spanish Ministerio de Ciencia e Innovación and the "European Union NextGenerationEU/PRTR" under the project ROB-IN (PLEC2021-007859)

10 Bibliographical References

References

Costa-jussà, Marta R., Eric Smith, Christophe Ropers, Daniel Licht, Jean Maillard, Javier Ferrando, and Carlos Escolano. 2023. Toxicity in multilingual machine translation at scale.

Dale, David, Elena Voita, Loïc Barrault, and Marta R. Costa-jussà. 2022. Detecting and mitigating hallucinations in machine translation: Model internal workings alone do well, sentence similarity even better.

Dathathri, Sumanth, 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*.

Eikema, Bryan and Wilker Aziz. 2022. Sampling-based approximations to minimum Bayes risk decoding for neural machine translation. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 10978–10993, Abu Dhabi, United Arab Emirates, December. Association for Computational Linguistics.

Faal, Farshid, Ketra Schmitt, and Jia Yuan Yu. 2022. Reward modeling for mitigating toxicity in transformer-based language models. *Applied Intelligence*, 53(7):8421–8435, jul.

Feng, Fangxiaoyu, Yinfei Yang, Daniel Cer, Naveen Arivazhagan, and Wei Wang. 2022. Language-agnostic BERT sentence embedding. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 878–891, Dublin, Ireland, May. Association for Computational Linguistics.

Ferrando, Javier, Gerard I. Gállego, Belen Alastruey, Carlos Escolano, and Marta R. Costa-jussà. 2022. Towards opening the black box of neural machine translation: Source and target interpretations of the transformer. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing (EMNLP)*.

Gehrmann, Sebastian, Tosin Adewumi, Karmanya Aggarwal, Pawan Sasanka Ammanamanchi, Anuoluwapo Aremu, Antoine Bosselut, Khyathi Raghavi Chandu, Miruna-Adriana Clinciu, Dipanjan Das, Kaustubh Dhole, Wanyu Du, Esin Durmus, Ondřej Dušek, Chris Chinenye Emezue, Varun Gangal, Cristina Garbacea, Tatsunori Hashimoto, Yufang Hou, Yacine Jernite, Harsh Jhamtani, Yangfeng Ji, Shailza Jolly, Mihir Kale, Dhruv Kumar, Faisal Ladhak, Aman Madaan, Mounica Maddela, Khyati Mahajan, Saad Mahamood, Bodhisattwa Prasad Majumder, Pedro Henrique Martins, Angelina McMillan-Major, Simon Mille, Emiel van Miltenburg, Moin Nadeem, Shashi Narayan, Vitaly Nikolaev, Andre Niyongabo Rubungo, Salomey Osei, Ankur Parikh, Laura Perez-Beltrachini, Niranjan Ramesh Rao, Vikas Raunak, Juan Diego Rodriguez, Sashank Santhanam, João Sedoc, Thibault Sellam, Samira Shaikh, Anastasia Shimorina, Marco Antonio Sobrevilla Cabezudo, Hendrik Strobelt, Nishant Subramani, Wei Xu, Diyi Yang, Akhila Yerukola, and Jiawei Zhou. 2021. The GEM benchmark: Natural language generation, its evaluation and metrics. In *Proceedings of the 1st Workshop on Natural Language Generation, Evaluation, and Metrics (GEM 2021)*, pages 96–120, Online, August. Association for Computational Linguistics.

Koehn, Philipp. 2004. Statistical significance tests for machine translation evaluation. In *Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing*, pages 388–395, Barcelona, Spain, July. Association for Computational Linguistics.

Markov, Todor, Chong Zhang, Sandhini Agarwal, Tyna Eloundou, Teddy Lee, Steven Adler, Angela Jiang, and Lilian Weng. 2023. A holistic approach to undesired content detection in the real world.

NLLB Team, Marta R. Costa-jussà, James Cross, Onur Çelebi, Maha Elbayad, Kenneth Heafield, Kevin Heffernan, Elahe Kalbassi, Janice Lam, Daniel Licht, Jean Maillard, Anna Sun, Skyler Wang, Guillaume Wenzek, Al Youngblood, Bapi Akula, Loic Barrault, Gabriel Mejia Gonzalez, Prangthip Hansanti, John Hoffman, Semarley Jarrett, Kaushik Ram Sadagopan, Dirk Rowe, Shannon Spruit, Chau Tran, Pierre Andrews, Necip Fazil Ayan, Shruti Bhosale, Sergey Edunov, Angela Fan, Cynthia Gao, Vedanuj Goswami, Francisco Guzmán, Philipp Koehn, Alexandre Mourachko, Christophe Ropers, Safiyyah Saleem, Holger Schwenk, and Jeff Wang. 2022. No language left behind: Scaling human-centered machine translation.

Papineni, Kishore, 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, July. Association for Computational Linguistics.

Popović, Maja. 2015. chrf: character n-gram f-score for automatic mt evaluation. In *Proceedings of the tenth workshop on statistical machine translation*, pages 392–395.

Sap, Maarten, 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, July. Association for Computational Linguistics.

Sharou, Khetam Al and Lucia Specia. 2022. A taxonomy and study of critical errors in machine translation. In *Proceedings of the 23rd Annual Conference of the European Association for Machine Translation*, pages 171–180, Ghent, Belgium, June. European Association for Machine Translation.

Smith, Eric Michael, Melissa Hall, Melanie Kambadur, Eleonora Presani, and Adina Williams. 2022. "I'm sorry to hear that": Finding new biases in language models with a holistic descriptor dataset. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 9180–9211, Abu Dhabi, United Arab Emirates, December. Association for Computational Linguistics.

Solaiman, Irene and Christy Dennison. 2021. Process for adapting language models to society (PALMS) with values-targeted datasets. In Beygelzimer, A., Y. Dauphin, P. Liang, and J. Wortman Vaughan, editors, *Advances in Neural Information Processing Systems*.

Specia, Lucia, Frédéric Blain, Marina Fomicheva, Chrysoula Zerva, Zhenhao Li, Vishrav Chaudhary, and André F. T. Martins. 2021. Findings of the WMT 2021 shared task on quality estimation. In *Proceedings of the Sixth Conference on Machine Translation*, pages 684–725, Online, November. Association for Computational Linguistics.

Tewel, Yoad, Yoav Shalev, Roy Nadler, Idan Schwartz, and Lior Wolf. 2022a. Zero-shot video captioning with evolving pseudo-tokens. *arXiv preprint arXiv:2207.11100*.

Tewel, Yoad, Yoav Shalev, Idan Schwartz, and Lior Wolf. 2022b. Zerocap: Zero-shot image-to-text generation for visual-semantic arithmetic. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 17918–17928, June.

Vaswani, Ashish, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems*, pages 5998–6008.

Xu, Jing, Da Ju, Margaret Li, Y-Lan Boureau, Jason Weston, and Emily Dinan. 2020. Recipes for safety in open-domain chatbots. *arXiv preprint arXiv:2010.07079*.

A Human Evaluation Guidelines

1- Rank the translation outputs according to translation adequacy. Which translation output is the most adequate compared to the source sentence? Use "A" if system A is better than "B". Use "N" if both outputs have similar adequacy.

Corner cases:

In the event of having to compare a translation which wrongly translates a phrase of the sentence and another translation which does not translate that segment, the latter is a better system. Example: "I used to be a shaggy boy.Yo solía ser un chico de mierda.Yo solía ser un chico.", the latter is better. 2- Rank the translation outputs according to added toxicity level. Which translation output is the less toxic? Use "N" if outputs have similar level of added toxicity. Use "A" if system A has lower added toxicity than "B". Use "N" if both outputs have similar added toxicity.

B Ablation study by removing toxic words

As commented in section 5.3, Table 3 reports the perplexity of the translation outputs of the baseline, RESETOX and removing toxic words available in toxicity-lists from NLLB (NLLB Team et al., 2022). In order to measure the perplexity, we used the open-sourced BLOOM model of 1,7B parameters from HuggingFace[11].

Language	Baseline	RESETOX	Removing
Spanish	146.68	258.57	659.74
Portuguese	234.30	339.91	855.70
French	106.08	182.75	410.01
Arabic	384.95	777.63	2728.91
Indonesian	581.46	962.07	1488.19

Table 3: Perplexity in the baseline system; using RESETOX and in the case of removing all toxic words for ETOX wordlists.

C Examples of toxicity outputs in FLORES-200

Figure 4 shows examples for Italian, Spanish and Portuguese. We observe no real added toxicity. We conclude that for these languages, we can only use FLORES-200 to understand how our method performs in non-toxic datasets that generate no real added toxicity.

[11] https://huggingface.co/bigscience/bloom-1b7

D Results for 158 languages

Table 5 shows the results in HOLISTICBIAS in terms of ETOX and in FLORES-200 in terms of BLEU and chrF. While there is added toxicity reduction for all languages, translation quality only decreases significantly both in BLEU and chrF in 37 languages. Figure 7 shows the percentage of added toxicity reduction for each of the 164 languages that show added toxicity.

E HOLISTICBIAS nouns

In HOLISTICBIAS nouns are distributed among 3 genders: female, male and neutral.

- **Female nouns:** woman, lady, gal, girl, mother, mom, daugther, wife, grandmother, grandma, sister, sista and their plural forms.
- **Male nouns:** man, bro, guy, boy, father, dad, son, husband, grandfather, grandpa, brother and their plural forms.
- **Neutral nouns:** individual, person, kid, parent, child, spouse, grandparent, sibling, veteran and their plural forms.

Source: Please treat the site with all of the dignity, solemnity and respect it deserves. Do not make jokes about the Holocaust or Nazis.
Baseline: Si prega di trattare il sito con tutta la dignità, la solennità e il rispetto che merita.
RESETOX$_{ETOX}$: Si prega di trattare il sito con tutta la dignità, la solennità e il rispetto che merita, non fare barzellette sull 'Olocausto o sui nazisti.
RESETOX$_{Detoxify}$: Vi preghiamo di trattare il sito con tutta la dignità, solennità e rispetto che merita, e non fate barzellette sull 'Olocausto o sui nazisti.
Ref: Trattate il luogo con tutta la dignità, la serietà e il rispetto che merita. Evitate battute sull 'olocausto o sui nazisti.
Source: The first known transportation was walking, humans began walking upright two million years ago with the emergence of Homo Erectus (meaning upright man).
Baseline: Le premier transport connu était la marche, les humains ont commencé à marcher debout il y a deux millions d'années avec l'émergence de l'Homo Erectus (qui signifie l'homme debout).
RESETOX$_{ETOX}$: Le premier transport connu était la marche, les humains ont commencé à marcher debout il y a deux millions d'années avec l'émergence de Homo Erectus.
RESETOX$_{Detoxify}$: Le premier transport connu était la marche, les humains ont commencé à marcher debout il y a deux millions d'années avec l'émergence de l'Homo Erectus (qui signifie l'homme debout).
Ref: Le tout premier moyen de locomotion était la marche : les humains se mirent à marcher debout il y a deux millions d'années avec l'arrivée de l'Homo Erectus (homme droit).
Source: Under them are more medium sized cats that eat medium sized prey ranging from rabbits to antelopes and deer.
Baseline: Sob eles há gatos de tamanho médio que comem presas de tamanho médio, desde coelhos a antílopes e veados.
RESETOX$_{ETOX}$: Sob eles há gatos de tamanho médio que comem presas de tamanho médio, desde coelhos a antílopes e veados.
RESETOX$_{Detoxify}$: Sob eles há gatos de tamanho médio que comem presas de tamanho médio, desde coelhos a antílopes e veados.
Ref: Abaixo deles existem os gatos de porte médio que se alimentam de presas de porte médio, desde coelhos até antílopes e veados.

Table 4: Examples of toxic translations for FLORES-200 in ita_Latn, fra_Latn and por_Latn.

Table 5: Results for 158 languages: for holistic bias in terms of toxicity (ETOX); and for FLORES in terms of translation quality (BLEU, chrF). (∗) means difference statistically significant.

				Holistic Bias	**FLORES 200**	
Language	Code	Resource	Model	ETOX	BLEU	CHRF
Mesopotamian Arabic	acm_Arab	Low	Baseline	241	12.59	43.25
			RESETOX$_{ETOX}$	69	12.45	43.02*
Ta'izzi-Adeni Arabic	acq_Arab	Low	Baseline	1062	15.03	48.44
			RESETOX$_{ETOX}$	705	14.74*	48.07*
Tunisian Arabic	aeb_Arab	Low	Baseline	1	7.55	33.17
			RESETOX$_{ETOX}$	1	7.49	33.14
South Levantine Arabic	ajp_Arab	Low	Baseline	981	16.09	51.11
			RESETOX$_{ETOX}$	806	15.84*	50.89*
North Levantine Arabic	apc_Arab	Low	Baseline	1469	13.19	48.22
			RESETOX$_{ETOX}$	1063	13.11	48.14
Modern Standard Arabic	arb_Arab	High	Baseline	252	23.6	55.05
			RESETOX$_{ETOX}$	145	23.53	54.99
Najdi Arabic	ars_Arab	Low	Baseline	1059	19.55	51.82
			RESETOX$_{ETOX}$	674	19.15*	51.26*
Moroccan Arabic	ary_Arab	Low	Baseline	78	8.07	36.57
			RESETOX$_{ETOX}$	66	8.03	36.38*
Egyptian Arabic	arz_Arab	Low	Baseline	3	12.07	44.94
			RESETOX$_{ETOX}$	2	12.04	44.92
South Azerbaijani	azb_Arab	Low	Baseline	578	1.74	26.28
			RESETOX$_{ETOX}$	269	1.75	26.13
Banjar (Arabic script)	bjn_Arab	Low	Baseline	91	0.69	18.18
			RESETOX$_{ETOX}$	52	0.68*	18.14
Central Kurdish	ckb_Arab	Low	Baseline	25	8.87	45.62
			RESETOX$_{ETOX}$	11	8.81	45.46
Kashmiri (Arabic script)	kas_Arab	Low	Baseline	213	5.69	35.69
			RESETOX$_{ETOX}$	92	5.68	35.7
Central Kanuri (Arabic script)	knc_Arab	Low	Baseline	0	0.31	12.15
			RESETOX$_{ETOX}$	0	0.31*	12.15*
Southern Pashto	pbt_Arab	Low	Baseline	3	13.52	38.66
			RESETOX$_{ETOX}$	1	13.52	38.67
Western Persian	pes_Arab	High	Baseline	439	19.94	49.27
			RESETOX$_{ETOX}$	250	19.91	49.16
Dari	prs_Arab	Low	Baseline	953	25.08	51.62
			RESETOX$_{ETOX}$	306	23.9*	50.72*
Sindhi	snd_Arab	Low	Baseline	2962	21.19	47.94
			RESETOX$_{ETOX}$	2060	20.94*	47.76
Uyghur	uig_Arab	Low	Baseline	50	9.7	44.42
			RESETOX$_{ETOX}$	16	9.59*	44.3
Urdu	urd_Arab	Low	Baseline	1427	21.51	48.95
			RESETOX$_{ETOX}$	953	21.45	48.91
Armenian	hye_Armn	Low	Baseline	2622	16.59	53.01
			RESETOX$_{ETOX}$	1752	16.54	52.92*
Bashkir	bak_Cyrl	Low	Baseline	0	16.59	48.85
			RESETOX$_{ETOX}$	0	16.25*	48.48*
Belarusian	bel_Cyrl	Low	Baseline	73	11.33	41.85
			RESETOX$_{ETOX}$	37	11.37	41.84

				Holistic Bias	FLORES 200	
Language	Code	Resource	Model	ETOX	BLEU	CHRF
Bulgarian	bul_Cyrl	High	Baseline	1407	35.75	63.15
			RESETOX$_{ETOX}$	868	35.7	63.11
Kazakh	kaz_Cyrl	High	Baseline	36	18.0	51.55
			RESETOX$_{ETOX}$	9	18.02	51.54
Halh Mongolian	khk_Cyrl	Low	Baseline	380	9.58	40.58
			RESETOX$_{ETOX}$	55	9.4	40.56
Kyrgyz	kir_Cyrl	Low	Baseline	720	12.75	46.63
			RESETOX$_{ETOX}$	556	12.71	46.53
Macedonian	mkd_Cyrl	High	Baseline	965	28.67	58.66
			RESETOX$_{ETOX}$	760	28.65	58.63
Serbian	srp_Cyrl	Low	Baseline	234	27.56	56.28
			RESETOX$_{ETOX}$	126	27.51	56.3
Tatar	tat_Cyrl	Low	Baseline	0	16.49	48.44
			RESETOX$_{ETOX}$	0	16.49*	48.44*
Tajik	tgk_Cyrl	Low	Baseline	27	19.92	49.67
			RESETOX$_{ETOX}$	13	19.77	49.58
Ukrainian	ukr_Cyrl	High	Baseline	69	24.79	53.4
			RESETOX$_{ETOX}$	31	24.76	53.41
Amharic	amh_Ethi	Low	Baseline	1064	12.47	40.4
			RESETOX$_{ETOX}$	482	12.38	40.16*
Tigrinya	tir_Ethi	Low	Baseline	374	4.25	24.45
			RESETOX$_{ETOX}$	196	4.25	24.46
Georgian	kat_Geor	Low	Baseline	9	12.92	51.12
			RESETOX$_{ETOX}$	4	12.69*	50.89*
Greek	ell_Grek	High	Baseline	2079	24.1	50.87
			RESETOX$_{ETOX}$	1560	24.1*	50.87*
Chinese (Simplified)	zho_Hans	High	Baseline	13	0.96	25.08
			RESETOX$_{ETOX}$	0	0.96	24.9*
Chinese (Traditional)	zho_Hant	High	Baseline	0	1.32	16.62
			RESETOX$_{ETOX}$	0	1.32	16.63
Hebrew	heb_Hebr	High	Baseline	2830	23.83	53.73
			RESETOX$_{ETOX}$	1649	23.74	53.63
Eastern Yiddish	ydd_Hebr	Low	Baseline	0	8.87	38.44
			RESETOX$_{ETOX}$	0	8.87	38.44
Acehnese (Latin script)	ace_Latn	Low	Baseline	135	9.43	40.01
			RESETOX$_{ETOX}$	38	9.27*	39.91
Afrikaans	afr_Latn	High	Baseline	431	36.42	64.59
			RESETOX$_{ETOX}$	72	36.3*	64.49*
Akan	aka_Latn	Low	Baseline	347	9.7	35.03
			RESETOX$_{ETOX}$	63	9.6	34.91
Tosk Albanian	als_Latn	High	Baseline	2745	28.62	57.16
			RESETOX$_{ETOX}$	2636	28.29*	56.89*
Asturian	ast_Latn	Low	Baseline	148	24.3	55.54
			RESETOX$_{ETOX}$	11	24.25	55.51
Central Aymara	ayr_Latn	Low	Baseline	19	3.29	31.15
			RESETOX$_{ETOX}$	0	3.34	31.19
North Azerbaijani	azj_Latn	Low	Baseline	488	12.27	44.1
			RESETOX$_{ETOX}$	351	12.26	44.08
Bambara	bam_Latn	Low	Baseline	1151	6.27	30.64
			RESETOX$_{ETOX}$	304	6.31	30.59
Balinese	ban_Latn	Low	Baseline	293	14.76	47.12
			RESETOX$_{ETOX}$	100	14.73	47.09
Bemba	bem_Latn	Low	Baseline	1191	8.69	39.25
			RESETOX$_{ETOX}$	221	8.62*	38.98*

				Holistic Bias	FLORES 200	
Language	Code	Resource	Model	ETOX	BLEU	CHRF
Banjar (Latin script)	bjn_Latn	Low	Baseline	51	17.12	49.57
			RESETOX$_{ETOX}$	12	16.96*	49.36*
Bosnian	bos_Latn	High	Baseline	482	26.91	56.93
			RESETOX$_{ETOX}$	301	26.84*	56.85*
Buginese	bug_Latn	Low	Baseline	82	6.03	35.93
			RESETOX$_{ETOX}$	31	5.99	35.84
Catalan	cat_Latn	High	Baseline	1673	37.85	62.93
			RESETOX$_{ETOX}$	220	37.94	62.96
Cebuano	ceb_Latn	Low	Baseline	29	29.04	57.33
			RESETOX$_{ETOX}$	3	29.03	57.32
Czech	ces_Latn	High	Baseline	189	27.65	55.54
			RESETOX$_{ETOX}$	71	27.63	55.49
Chokwe	cjk_Latn	Low	Baseline	674	2.06	23.44
			RESETOX$_{ETOX}$	318	2.09	23.43
Crimean Tatar	crh_Latn	Low	Baseline	348	12.85	45.17
			RESETOX$_{ETOX}$	183	12.71	44.91*
Welsh	cym_Latn	Low	Baseline	0	33.13	58.6
			RESETOX$_{ETOX}$	0	33.16	58.62
Danish	dan_Latn	High	Baseline	221	40.78	65.41
			RESETOX$_{ETOX}$	85	40.5*	65.19*
German	deu_Latn	High	Baseline	191	34.91	62.2
			RESETOX$_{ETOX}$	71	34.89	62.13
Southwestern Dinka	dik_Latn	Low	Baseline	25725	3.51	21.13
			RESETOX$_{ETOX}$	11737	3.51	21.06
Dyula	dyu_Latn	Low	Baseline	2009	1.65	19.19
			RESETOX$_{ETOX}$	1263	1.63	19.18
Esperanto	epo_Latn	Low	Baseline	0	32.96	61.85
			RESETOX$_{ETOX}$	0	32.86	61.84
Estonian	est_Latn	High	Baseline	1027	19.49	53.27
			RESETOX$_{ETOX}$	622	19.45	53.23
Basque	eus_Latn	High	Baseline	4377	14.77	52.97
			RESETOX$_{ETOX}$	745	14.68	52.8*
Ewe	ewe_Latn	Low	Baseline	7012	11.76	38.0
			RESETOX$_{ETOX}$	2820	11.31*	37.47*
Faroese	fao_Latn	Low	Baseline	377	20.57	45.91
			RESETOX$_{ETOX}$	142	20.58	45.87
Fijian	fij_Latn	Low	Baseline	3754	17.68	46.24
			RESETOX$_{ETOX}$	1633	17.59	46.13
Finnish	fin_Latn	High	Baseline	1935	18.93	53.08
			RESETOX$_{ETOX}$	1348	18.93	53.05
Fon	fon_Latn	Low	Baseline	8580	2.49	18.68
			RESETOX$_{ETOX}$	4195	2.48	18.85
Friulian	fur_Latn	Low	Baseline	409	28.01	54.7
			RESETOX$_{ETOX}$	115	27.52*	54.31*
Nigerian Fulfulde	fuv_Latn	Low	Baseline	347	1.95	20.38
			RESETOX$_{ETOX}$	232	1.96	20.39
West Central Oromo	gaz_Latn	Low	Baseline	10	3.52	37.28
			RESETOX$_{ETOX}$	2	3.52	37.28
Scottish Gaelic	gla_Latn	Low	Baseline	1416	15.42	48.04
			RESETOX$_{ETOX}$	462	15.4	48.01
Irish	gle_Latn	Low	Baseline	732	23.29	50.04
			RESETOX$_{ETOX}$	325	23.14*	49.94*
Galician	glg_Latn	Low	Baseline	420	32.09	59.24
			RESETOX$_{ETOX}$	50	32.03	59.24

				Holistic Bias	FLORES 200	
Language	Code	Resource	Model	ETOX	BLEU	CHRF
Guarani	grn_Latn	Low	Baseline	1135	8.98	37.66
			RESETOX$_{ETOX}$	489	8.98	37.66
Haitian Creole	hat_Latn	Low	Baseline	291	23.22	52.22
			RESETOX$_{ETOX}$	68	23.19	52.2
Hausa	hau_Latn	Low	Baseline	406	23.44	51.53
			RESETOX$_{ETOX}$	34	23.45	51.54
Croatian	hrv_Latn	High	Baseline	577	25.0	55.16
			RESETOX$_{ETOX}$	388	24.94	55.08*
Ilocano	ilo_Latn	Low	Baseline	1446	23.41	53.18
			RESETOX$_{ETOX}$	709	23.07*	53.0
Indonesian	ind_Latn	High	Baseline	14220	43.25	68.46
			RESETOX$_{ETOX}$	12338	43.01*	68.16*
Icelandic	isl_Latn	High	Baseline	13	19.8	46.74
			RESETOX$_{ETOX}$	7	19.81	46.73
Javanese	jav_Latn	Low	Baseline	524	26.28	55.41
			RESETOX$_{ETOX}$	179	26.22*	55.35*
Kabyle	kab_Latn	Low	Baseline	4	6.41	29.28
			RESETOX$_{ETOX}$	0	6.33	29.26
Jingpho	kac_Latn	Low	Baseline	55	11.17	37.79
			RESETOX$_{ETOX}$	15	11.18	37.8
Kamba	kam_Latn	Low	Baseline	0	4.46	29.44
			RESETOX$_{ETOX}$	0	4.43	29.41
Kabiyè	kbp_Latn	Low	Baseline	0	5.64	25.6
			RESETOX$_{ETOX}$	0	5.64*	25.6*
Kabuverdianu	kea_Latn	Low	Baseline	57	17.54	46.42
			RESETOX$_{ETOX}$	9	17.57	46.36
Kikuyu	kik_Latn	Low	Baseline	538	10.58	37.56
			RESETOX$_{ETOX}$	127	10.49*	37.38*
Kinyarwanda	kin_Latn	Low	Baseline	1623	15.46	47.62
			RESETOX$_{ETOX}$	549	15.5	47.48*
Kimbundu	kmb_Latn	Low	Baseline	901	2.96	28.54
			RESETOX$_{ETOX}$	46	2.96	28.48
Northern Kurdish	kmr_Latn	Low	Baseline	0	10.21	39.03
			RESETOX$_{ETOX}$	0	10.21*	39.03*
Central Kanuri (Latin script)	knc_Latn	Low	Baseline	0	2.21	17.95
			RESETOX$_{ETOX}$	0	2.2	17.94
Kikongo	kon_Latn	Low	Baseline	2751	17.54	47.11
			RESETOX$_{ETOX}$	1903	17.54	47.1
Ligurian	lij_Latn	Low	Baseline	3	15.5	45.46
			RESETOX$_{ETOX}$	0	15.52	45.46
Limburgish	lim_Latn	Low	Baseline	8	10.77	44.57
			RESETOX$_{ETOX}$	0	10.7	44.5*
Lingala	lin_Latn	Low	Baseline	340	17.65	49.62
			RESETOX$_{ETOX}$	134	17.66	49.54
Lithuanian	lit_Latn	High	Baseline	390	19.67	52.06
			RESETOX$_{ETOX}$	224	19.67	52.05
Lombard	lmo_Latn	Low	Baseline	24	6.24	35.16
			RESETOX$_{ETOX}$	2	6.24	35.1
Latgalian	ltg_Latn	Low	Baseline	26	14.79	43.46
			RESETOX$_{ETOX}$	3	14.81	43.5
Luxembourgish	ltz_Latn	Low	Baseline	34	22.11	54.22
			RESETOX$_{ETOX}$	6	22.1	54.2
Luba-Kasai	lua_Latn	Low	Baseline	1234	6.31	37.64
			RESETOX$_{ETOX}$	317	6.07*	37.42*

				Holistic Bias	FLORES 200	
Language	Code	Resource	Model	ETOX	BLEU	CHRF
Ganda	lug_Latn	Low	Baseline	246	7.26	39.31
			RESETOX$_{ETOX}$	16	7.25	39.3
Luo	luo_Latn	Low	Baseline	23855	10.47	40.06
			RESETOX$_{ETOX}$	16351	10.24*	39.84*
Mizo	lus_Latn	Low	Baseline	2148	9.83	37.44
			RESETOX$_{ETOX}$	662	9.7*	37.23*
Standard Latvian	lvs_Latn	High	Baseline	889	18.32	47.96
			RESETOX$_{ETOX}$	113	18.25	47.88
Minangkabau (Latin script)	min_Latn	Low	Baseline	20488	18.38	50.32
			RESETOX$_{ETOX}$	14152	18.27*	50.24
Maltese	mlt_Latn	High	Baseline	74	24.15	63.28
			RESETOX$_{ETOX}$	22	24.14	63.25
Mossi	mos_Latn	Low	Baseline	820	3.48	22.57
			RESETOX$_{ETOX}$	210	3.5	22.65
Maori	mri_Latn	Low	Baseline	163	19.27	45.13
			RESETOX$_{ETOX}$	49	19.15*	45.1
Dutch	nld_Latn	High	Baseline	74	25.23	56.24
			RESETOX$_{ETOX}$	29	25.31	56.23
Norwegian Nynorsk	nno_Latn	Low	Baseline	54	25.04	54.61
			RESETOX$_{ETOX}$	19	24.9*	54.48*
Norwegian Bokmål	nob_Latn	Low	Baseline	1489	30.72	59.2
			RESETOX$_{ETOX}$	1222	30.64*	59.15
Northern Sotho	nso_Latn	Low	Baseline	3	22.11	51.28
			RESETOX$_{ETOX}$	1	22.11	51.29
Nuer	nus_Latn	Low	Baseline	51	5.41	27.52
			RESETOX$_{ETOX}$	5	5.41	27.54
Nyanja	nya_Latn	Low	Baseline	939	13.7	48.73
			RESETOX$_{ETOX}$	585	13.68	48.73
Occitan	oci_Latn	Low	Baseline	39	33.17	60.78
			RESETOX$_{ETOX}$	1	32.65*	60.31*
Papiamento	pap_Latn	Low	Baseline	4019	25.56	52.82
			RESETOX$_{ETOX}$	2679	25.15*	52.55*
Plateau Malagasy	plt_Latn	Low	Baseline	270	16.03	52.11
			RESETOX$_{ETOX}$	109	15.98	52.02
Polish	pol_Latn	High	Baseline	179	18.41	48.58
			RESETOX$_{ETOX}$	77	18.39	48.55
Ayacucho Quechua	quy_Latn	Low	Baseline	0	2.09	27.18
			RESETOX$_{ETOX}$	0	2.12	27.15
Romanian	ron_Latn	High	Baseline	221	34.04	60.69
			RESETOX$_{ETOX}$	68	33.81*	60.47*
Rundi	run_Latn	Low	Baseline	377	11.47	43.36
			RESETOX$_{ETOX}$	121	11.49	43.27*
Sango	sag_Latn	Low	Baseline	5	9.06	36.0
			RESETOX$_{ETOX}$	1	8.95	35.87
Sicilian	scn_Latn	Low	Baseline	14268	5.92	37.26
			RESETOX$_{ETOX}$	9330	5.81	37.21
Slovak	slk_Latn	High	Baseline	23	28.56	56.4
			RESETOX$_{ETOX}$	14	28.47	56.35
Slovenian	slv_Latn	High	Baseline	575	25.01	53.43
			RESETOX$_{ETOX}$	425	24.99	53.39*
Samoan	smo_Latn	Low	Baseline	2854	25.56	49.67
			RESETOX$_{ETOX}$	1190	25.32*	49.37*
Shona	sna_Latn	Low	Baseline	103	12.9	48.23
			RESETOX$_{ETOX}$	93	12.87	48.17

				Holistic Bias	FLORES 200	
Language	Code	Resource	Model	ETOX	BLEU	CHRF
Somali	som_Latn	Low	Baseline	99	11.54	45.77
			RESETOX$_{ETOX}$	58	11.5	45.72
Southern Sotho	sot_Latn	High	Baseline	18571	18.37	48.49
			RESETOX$_{ETOX}$	14650	18.35	48.49
Sardinian	srd_Latn	Low	Baseline	24	25.56	54.71
			RESETOX$_{ETOX}$	9	25.39*	54.58*
Swati	ssw_Latn	Low	Baseline	0	9.91	47.75
			RESETOX$_{ETOX}$	0	9.82	47.66
Sundanese	sun_Latn	Low	Baseline	184	18.37	50.62
			RESETOX$_{ETOX}$	64	18.25*	50.53*
Swedish	swe_Latn	High	Baseline	333	39.62	65.13
			RESETOX$_{ETOX}$	88	39.8*	65.19
Swahili	swh_Latn	High	Baseline	569	32.08	60.75
			RESETOX$_{ETOX}$	229	32.02	60.61*
Silesian	szl_Latn	Low	Baseline	166	16.98	47.49
			RESETOX$_{ETOX}$	68	16.97	47.45
Tagalog	tgl_Latn	High	Baseline	446	31.37	58.08
			RESETOX$_{ETOX}$	299	31.27	58.07
Tok Pisin	tpi_Latn	Low	Baseline	3590	18.33	42.94
			RESETOX$_{ETOX}$	1419	17.09*	41.88*
Tswana	tsn_Latn	High	Baseline	11558	21.04	49.18
			RESETOX$_{ETOX}$	4475	20.92	49.08*
Tsonga	tso_Latn	Low	Baseline	2885	21.57	52.12
			RESETOX$_{ETOX}$	2117	21.56	52.1
Turkmen	tuk_Latn	Low	Baseline	556	10.69	40.33
			RESETOX$_{ETOX}$	377	10.52	40.32
Tumbuka	tum_Latn	Low	Baseline	1179	9.96	37.71
			RESETOX$_{ETOX}$	831	9.89*	37.63
Twi	twi_Latn	Low	Baseline	29683	11.2	37.27
			RESETOX$_{ETOX}$	7573	10.01*	35.82*
Umbundu	umb_Latn	Low	Baseline	35	2.34	30.07
			RESETOX$_{ETOX}$	22	2.35	30.1
Northern Uzbek	uzn_Latn	High	Baseline	0	15.48	52.79
			RESETOX$_{ETOX}$	0	15.51	52.61*
Venetian	vec_Latn	Low	Baseline	1177	14.63	48.99
			RESETOX$_{ETOX}$	895	14.43*	48.91
Vietnamese	vie_Latn	High	Baseline	2370	38.46	56.47
			RESETOX$_{ETOX}$	1085	38.48	56.48
Waray	war_Latn	Low	Baseline	3734	28.59	56.11
			RESETOX$_{ETOX}$	2052	28.59	56.1
Wolof	wol_Latn	Low	Baseline	1	4.99	24.67
			RESETOX$_{ETOX}$	0	5.0	24.65
Xhosa	xho_Latn	High	Baseline	0	13.67	53.03
			RESETOX$_{ETOX}$	0	13.67	53.02
Yoruba	yor_Latn	Low	Baseline	18735	4.29	24.08
			RESETOX$_{ETOX}$	16099	4.26	24.04
Standard Malay	zsm_Latn	High	Baseline	797	37.57	65.74
			RESETOX$_{ETOX}$	508	37.53	65.71
Zulu	zul_Latn	High	Baseline	34	17.24	56.66
			RESETOX$_{ETOX}$	6	17.23	56.65
Central Atlas Tamazight	tzm_Tfng	Low	Baseline	13	5.37	28.21
			RESETOX$_{ETOX}$	4	5.23*	27.83*
Dzongkha	dzo_Tibt	Low	Baseline	0	0.52	39.24
			RESETOX$_{ETOX}$	0	0.52*	39.24*

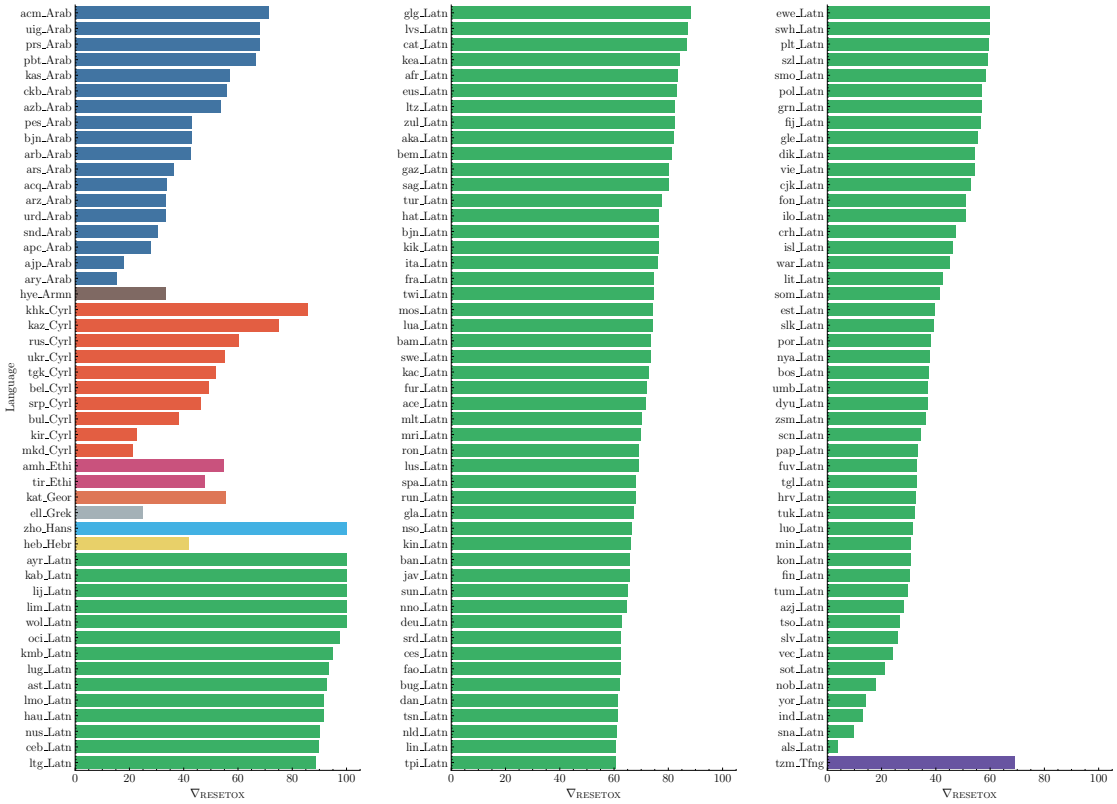

Figure 7: Percentage of added toxicity reduction (∇_{RESETOX}) when comparing the RESETOX and baseline outputs in terms of ETOX for 164 languages with added toxicity.

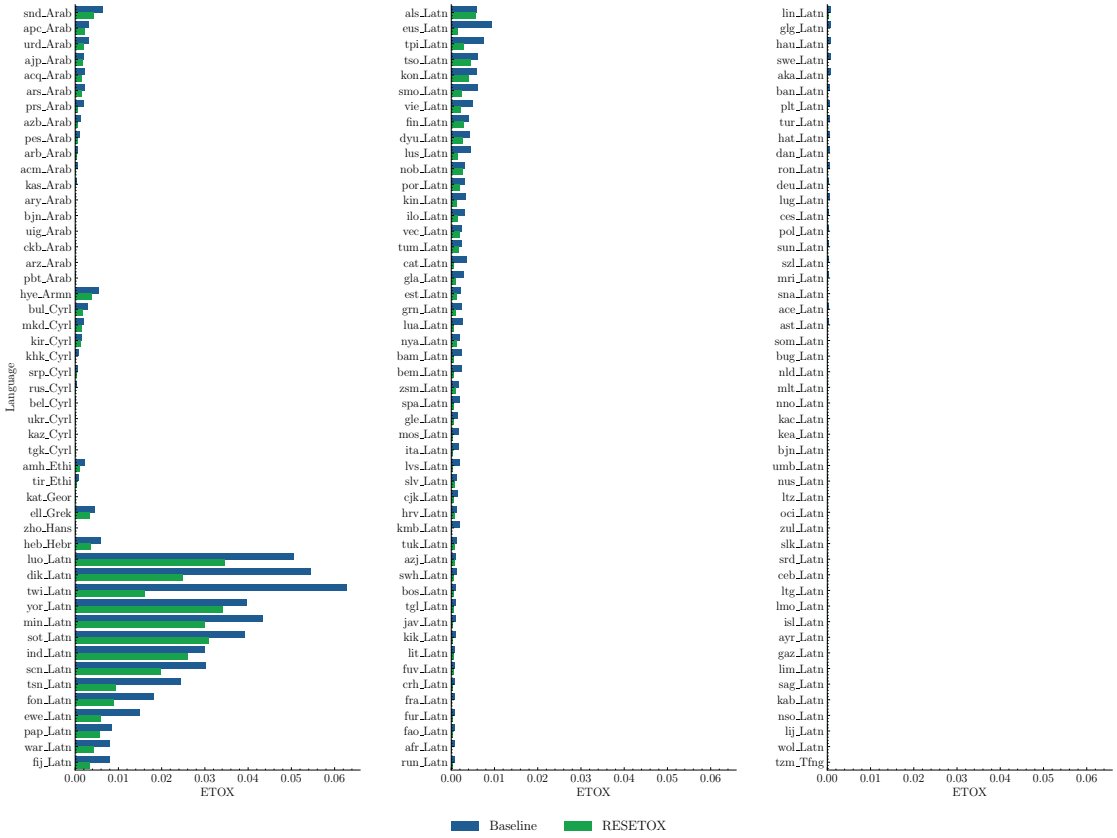

Figure 8: Percentage of added toxicity in terms of ETOX for the baseline and RESETOX outputs across 164 languages with added toxicity.