SynthDetoxM: Modern LLMs are Few-Shot Parallel Detoxification Data Annotators

Daniil Moskovskiy[1,2*] **Nikita Sushko**[1,2*] **Sergey Pletenev**[1,2]
Elena Tutubalina[1,3,4,5] **Alexander Panchenko**[2,1]
[1]AIRI [2]Skoltech [3]Sber AI [4]HSE University [5]Kazan Federal University

Correspondence: {d.moskovskiy, a.panchenko}@skol.tech

Abstract

Existing approaches to multilingual text detoxification are hampered by the scarcity of parallel multilingual datasets. In this work, we introduce a pipeline for the generation of multilingual parallel detoxification data. We also introduce SynthDetoxM, a manually collected and synthetically generated multilingual parallel text detoxification dataset comprising 16,000 high-quality detoxification sentence pairs across German, French, Spanish and Russian. The data was sourced from different toxicity evaluation datasets and then rewritten with nine modern open-source LLMs in few-shot setting. Our experiments demonstrate that models trained on the produced synthetic datasets have superior performance to those trained on the human-annotated MultiParaDetox dataset even in data limited setting. Models trained on SynthDetoxM outperform all evaluated LLMs in few-shot setting. We release our dataset and code to help further research in multilingual text detoxification.

Warning: this paper contains illustrative examples of texts that readers may find offensive or disturbing.

1 Introduction

The proliferation of social networks and text-based internet media has highlighted the issue of online toxicity and hate speech (Saha et al., 2019). This phenomenon not only creates an unpleasant environment for users but also deters advertisers, potentially impacting the economic viability of these platforms (Fortuna and Nunes, 2018). Consequently, there is an urgent need for effective mechanisms to measure and mitigate toxicity in online spaces.

A promising approach to addressing this challenge is text detoxification of text through paraphrasing (Krishna et al., 2020). Text detoxification is a subtask of text style transfer (TST), which involves rewriting text while preserving its original meaning and altering specific style attribute, such as formality, bias, expressiveness, sentiment, or, in the case of detoxification, toxicity (Fu et al., 2018; Lai et al., 2021).

	Toxic Text	Detoxified Text
German	Wie be**oppt muss man sein?	Wie verwirrt muss man sein?
Spanish	Que os den por el c**o.	Que os dé muy mala suerte.
French	c'est moi at***dé ! je suis tombé !	C'est moi qui suis tombé !
Russian	я мужик а вы г**но	Я мужчина, а вы неправы

Table 1: Examples of the source toxic texts across different languages and their respective synthetic detoxifications from our SynthDetoxM.

While significant progress has been made in monolingual TST and detoxification, both in supervised and unsupervised settings (Dale et al., 2021; Logacheva et al., 2022; Pour et al., 2023), multilingual text detoxification remains a largely unsolved problem. This is primarily due to two factors: the scarcity of parallel detoxification data across multiple languages and the suboptimal performance of unsupervised methods in cross-lingual settings (Dementieva et al., 2023).

Manual or crowdsourced data collection is a challenging and costly task (Rao and Tetreault, 2018; Reid and Artetxe, 2023; Konovalov et al., 2016b), creating parallel data with the use of modern LLMs, which already proven to work well for the tasks of text classification (Sun et al., 2023) and question answering (Ye et al., 2022), remains underexplored. To address these challenges and facilitate the development of multilingual text detoxification models and datasets, we propose a framework for generating parallel multilingual synthetic detoxification data and SynthDetoxM, a large-scale multilinugal synthetic parallel text detoxification dataset, which was created using this framework.

*Equal contribution.

Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers), pages 5714–5733

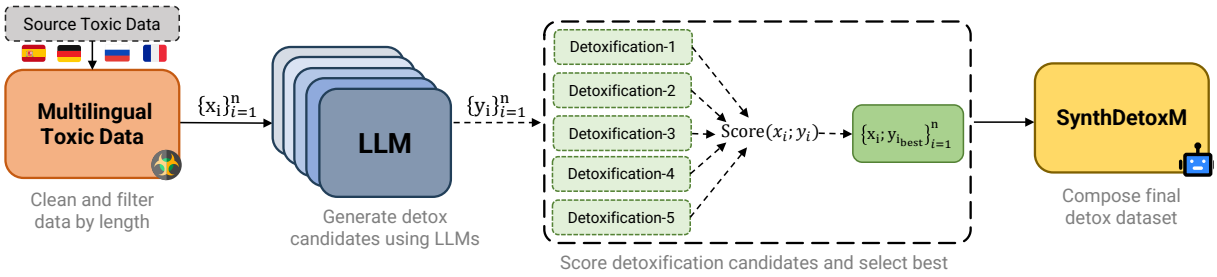

Figure 1: An illustration of the proposed approach for collecting and generating the multilingual text detoxification dataset SynthDetoxM.

Our dataset is comprised of 16,000 high-quality synthetic detoxification pairs across four languages: German, Spanish, French and Russian. The dataset was created using few-shot prompting and selecting the best generations of five different open-source LLMs. The answers were combined using a handcrafted heuristic, providing the best answers from each model to ensure diversity and quality of the final data.

Our contributions can be summarized as follows:

1. We propose a framework for generating synthetic parallel multilingual detoxification data using few-shot prompting of LLMs.

2. We create SynthDetoxM, a large-scale multilingual synthetic parallel dataset for text detoxification, helping to address the data scarcity issue in the detoxification task.

3. We conduct a thorough empirical evaluation of the proposed dataset, including linguistic analysis of the data and benchmarking against the human-annotated MultiParaDetox.

We openly release the generated data and code.[1]

2 Background and Related Work

2.1 Text Style Transfer

Text Style Transfer (TST), the task of rewriting the text in a target style while preserving its semantic content and fluency, has garnered significant attention in the natural language processing community due to its potential applications in text generation (Fu et al., 2018). TST encompasses various subtasks, including formality style transfer (Wang et al., 2020; Lai et al., 2021), sentiment style transfer (Yu et al., 2021), authorship style transfer (Horvitz et al., 2024; Liu et al., 2024), and detoxification (Dale et al., 2021; Atwell et al., 2022; Moskovskiy et al., 2022, 2024).

With the advent of Large Language Models (LLMs), in-context learning methods have increasingly been utilized for TST and detoxification tasks. Suzgun et al. (2022) proposed a novel approach to TST by prompting LLMs and then reranking the generated texts based on three TST metrics: text similarity, target style strength, and fluency. Similarly, Reif et al. (2022) demonstrated the effectiveness of prompting GPT-3, a state-of-the-art LLM at the time, to rewrite texts in a desired style.

2.2 Text Detoxification

Text Detoxification, a subtask of Text Style Transfer (TST), involves transforming an input text x_i, identified as toxic through toxicity estimation models, into a text y_i that is non-toxic in style while maintaining semantic similarity and fluency. In this context, toxicity refers to language that is harmful, offensive, or inappropriate.

Due to the lack of parallel training data, early research focused on unsupervised detoxification methods (dos Santos et al., 2018; Dale et al., 2021; Hallinan et al., 2023; Pour et al., 2023). For instance,(Logacheva et al., 2022) and APPDIA (Atwell et al., 2022), has enabled the training of sequence-to-sequence models (Logacheva et al., 2022; Pour et al., 2023) that outperform most unsupervised approaches in terms of rewritten toxicity, fluency, and semantic similarity. In parallel, Moskovskiy et al. (2024) explored the use of activation patching in LLMs to generate synthetic parallel detoxification data for English. Their results demonstrated that training detoxification models on this data yields performance comparable to models trained on manually annotated datasets in automatic evaluations, while achieving superior quality in human assessments.

[1] github.com/s-nlp/synthdetoxm

2.3 Multilingual Text Style Transfer

The scarcity of high-quality parallel multilingual detoxification data remains a major challenge in the field. Recently, new non-English parallel datasets have been introduced for various TST tasks, including a Bangla language parallel sentiment style transfer dataset (Mukherjee et al., 2023) and the extension of the GYAFC dataset to Portuguese, French, and Italian, resulting in XFORMAL (Briakou et al., 2021). Following the crowdsourcing pipeline introduced by (Logacheva et al., 2022), a parallel text detoxification dataset for Russian was collected (Moskovskiy et al., 2022). Later, using the similar data annotation pipeline, Dementieva et al. (2024) collected 1000 sentence pairs across nine languages, resulting in the MultiParaDetox dataset for a corresponding shared task on multilingual text detoxification. Furthermore, in a more recent work Dementieva et al. (2025), provide an in-depth analysis of toxicity characteristics across languages, exploring descriptive linguistic features that influence detoxification quality.

Nevertheless, the size of MultiParaDetox is far from satisfactory with 1000 sentence pairs per language, only 400 of which are publicly available. The remaining 600 pairs comprised the test set for the multilingual text detoxification shared task (Dementieva et al., 2024). Such relatively small dataset may be insufficient for training big multilingual language models for multilingual text detoxification.

To bridge this gap, we present SynthDetoxM - a synthetic parallel detoxification corpus for four European languages, namely, German, Spanish, French, and Russian, with 4000 samples for each language. The dataset creationg pipeline presented in our work can easily be transferred to other languages as well, drastically reducing the cost of annotation for parallel detoxification datasets.

3 Methodology

In this section, we describe the pipeline introduced for collecting the multilingual parallel text detoxification dataset, SynthDetoxM. A general illustration of our approach is shown in Figure 1.

3.1 Data Collection

To create SynthDetoxM, we begin by selecting several thousand non-parallel toxic texts from publicly available toxicity identification datasets. We focus on four languages for SynthDetoxM: German, French, Spanish and Russian. From these datasets

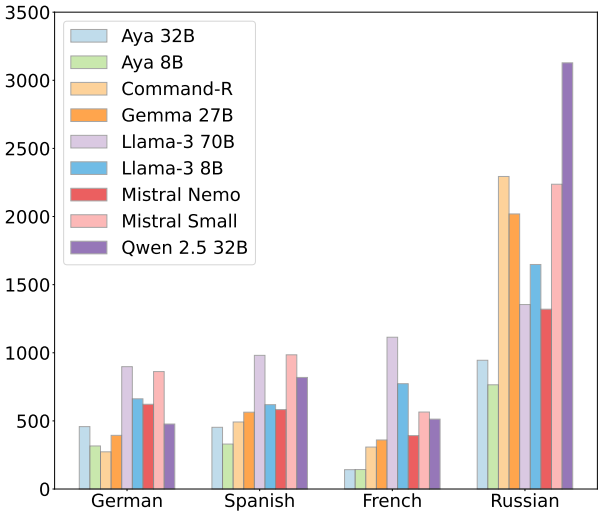

Figure 2: Number of accepted samples in the final SynthDetoxM dataset with respect to the LLM by language.

we select only the texts that were marked as toxic by human annotators, excluding non-toxic examples. In cases of multiple annotations, we retained the sample where the majority of annotators classified the sentence as toxic.

To enhance the final data quality, we employ sample-level filtering using the **STA** and **SIM** metrics[2] and we also apply data augmentation techniques utilizing the Perspective API (Lees et al., 2022). Since the API returns both toxicity scores and toxic spans, we further improve data quality by splitting the source texts into sentences and removing those sentences that do not intersect with the detected toxic spans. This filtering process results in a larger dataset of toxic sentences, and we also split overly long inputs into separate examples.

Russian For the Russian dataset, we use data from the Jigsaw Toxic Comments Classification Challenge (Kivlichan et al., 2020), Russian Language Toxic Comments (Belchikov, 2019) and Toxic Russian Comments (Semiletov, 2020). From these sources, we select only those rows labeled as toxic, resulting in more than 15,697 toxic texts. We then calculate the **STA** and **SIM** metrics, applying a threshold of 0.5 for filtering. After removing emojis, eliminating texts with fewer than five words or more than 30 words, and splitting the sentences using toxic spans from Perspective API, our final dataset consists of 15,697 texts.

German For German, we use the toxicity identification data from the GermEval 2021 shared

[2]Evaluation metrics are described in Section §4.3

task (Risch et al., 2021) and RP-Mod and RP-Crowd (Assenmacher et al., 2021) to create a dataset of 4,946 toxic texts. We apply the same filtering and augmentation pipeline as for the Russian dataset, but with a lower **STA** score threshold of 0.3. This resulted in a dataset of 4,946 texts, which exceeds the original size of the raw dataset. We attribute this increase to the higher median length of German sentences, which leads to a greater number of split texts.

Spanish For Spanish, we use data from two sources: the Jigsaw Toxic Comments Classification Challenge (Kivlichan et al., 2020) and the Clandestino dataset (Capozzi et al., 2021). This results in an initial dataset of 10,260 toxic texts. We apply the same filtering and augmentation pipeline as for the German dataset, maintaining the same **STA** threshold of 0.3. The final dataset contains 5,826 texts.

French For French, we use the data from the MLMA Hate Speech Corpus (Ousidhoum et al., 2019) and the Jigsaw Toxic Comments Classification (Kivlichan et al., 2020). These data sources combined produce a dataset of 5,424 toxic texts. As the toxicity classifier used in other languages does not support French, we instead use the Perspective API to get toxicity scores. After the application of the **STA** score threshold of 0.25, we obtain a dataset of 4,310 sentences.

3.1.1 Parallel Data Generation Pipeline

To generate parallel detoxification data, we use various open-source LLMs in a few-shot generation setup. Specifically, we employed the following models: Qwen 2.5 32B by Qwen (Yang et al., 2024; Team, 2024); Command-R 32B by Cohere (Cohere, 2024); Gemma 2 27B by Google (Rivière et al., 2024); Aya Expanse in 32B and 8B versions by Cohere (Dang et al., 2024); Mistral Small 22B, Mistral Nemo 12B by Mistral AI (Mistral, 2024a,b); and Llama 3.1 70B and 8B models respectively, by Meta (Dubey et al., 2024). While not all these models are explicitly designed for multilingual tasks, our experiments show that all of them support the languages considered in this work.

3.1.2 Few-Shot Example Mining

To select the best toxic and non-toxic pairs for few-shot generation, we calculate the **STA** and **SIM** metrics for all sentences in Russian, German and Spanish from the multilingual toxicity detection dataset[3]. We then rank the top 10 sentences based on the following score:

$$\mathbf{Score}(x_i; y_i) = 1 - \left(\frac{1 - \mathbf{STA}(x_i)}{1 - \mathbf{STA}(y_i)} \cdot (1 - \mathbf{SIM}(x_i; y_i)) \right)$$

where $\mathbf{STA}(x_i)$ and $\mathbf{STA}(y_i)$ represent the toxicity scores for the original and detoxified examples, respectively. $\mathbf{SIM}(x_i; y_i)$ is the cosine distance between the embeddings of toxic and detoxified sentences.

This ranking criterion is chosen to ensure high-quality detoxification without altering the original meaning of the sentences. Since the sentences used for few-shot prompting have been annotated by human experts, we expect the detoxification quality to be satisfactory. Additionally, the rewriting process of toxic words often leads to an expanded distance between toxic and non-toxic sentences, increasing the distinction in non-toxicity. To maximize both the distance and the distinction between the original and detoxified sentences, we select harder and more meaningful examples for few-shot prompting, which helps improve the detoxification process.

For French, which is not represented in the MultiParaDetox dataset, we used human annotators to detoxify 10 randomly chosen sentences from the existing non-parallel data.

After generating detoxified examples, we perform refusal filtering using a refusal classification model (see details in Appendix D). Additionally, we use a simple threshold-based non-detoxifiability metric, calculated by dividing the absolute reduction in the STA score by the original STA score. We compare the resulting detoxifiability scores to a fixed threshold of 0.5. If the score falls below this threshold, the example is considered non-detoxifiable.

After generating five detoxification datasets in each language using the selected models, we rank the sentences by their multiplied **STA** and **SIM** metrics and select the top-scoring examples. This metric helps mitigate issues such as refusal(where models refuse to generate text due to toxicity) and copy-paste generation (where the model generates the input toxic sentences without modification), as copy-paste generation typically results in a low **STA** score, while refusal leads to a low **SIM** score.

[3] hf.co/multilingual_toxicity_dataset

	$\mathbf{STA_T}\uparrow$	$\mathbf{STA_D}\uparrow$	SIM↑	$\mathbf{STA_D}\times$SIM↑
German	0.389	0.853	0.793	0.675
Spanish	0.514	0.920	0.736	0.681
French	0.583	0.913	0.677	0.624
Russian	0.467	0.924	0.731	0.678

Table 2: Average toxicity levels across different languages for source toxic (T) and generated detoxified (D) texts, along with similarity scores. $\mathbf{STA_T}$ represents the toxicity level of the original text, while $\mathbf{STA_D}$ corresponds to the detoxified text.

3.2 Final Composed Dataset

After all preprocessing, cleaning and filtering steps we compose SynthDetoxM - a manually collected and synthetically paraphrased parallel detoxification dataset on 16,000 toxic and non-toxic text pairs for Spanish, German, Russian and French.

We show the statistics of detoxification candidate acceptance with respect to each LLM Languagewise in Table 5 and Figure 2. According to the statistics, Qwen 2.5 generated the most preferrable detoxifications among other models.

However, upon manual examination we noticed that Qwen tended to occasionally insert tokens of Chinese text into the generated text though was prompted to answer only on the language of the source text. Therefore, the strict reranking and filtering criteria of generated detoxification candidates is necessary.

3.3 Data Quality Evaluation Pipeline

To evaluate the quality of our generated detoxification data in Russian, German, and Spanish, we use our dataset for training and compare the performance of models trained on SynthDetoxMwith those trained on the human-annotated parallel detoxification dataset, MultiParaDetox Dementieva et al. (2024). Due to its absence in the MultiParaDetox dataset, French is excluded from this comparison. A more detailed linguistic analysis of the dataset can be found in Appendix E.

4 Experimental Setup

4.1 Data Quality Tests

To evaluate the efficacy of our SynthDetoxM for German, Spanish and Russian, we've trained a series of sequence-to-sequence models on different folds from the dataset. Since MultiParaDetox consists of only 400 pairs of toxic texts with their human-written non-toxic rephrasings, we split our created SynthDetoxM dataset into 10 chunks of 400 pairs for German, Spanish and Russian. We trained 10 mT0 models on different chunks of the dataset and evaluated their average performance on the MultiParaDetox test set. Additionally, we test if using both our SynthDetoxM and MultiParaDetox for training would lead to improved performance.

4.2 Toxicity and Similarity of Synthetic Texts

To further assess the quality of the generated data, we computed the **STA** and **SIM** scores using the Perspective API for Russian, German, Spanish, and French. These metrics were selected for their relevance to detoxification tasks and their ability to quantitatively assess our synthetic dataset. We also assessed the quality of the French subset of SynthDetoxM, as French is not represented in the MultiParaDetox dataset, and therefore cannot be evaluated through model training. The scores are presented in Figure 3 and Table 2.

The results indicate that French achieves comparable automatic metric scores to other languages, suggesting that detoxification models trained on this data would perform similarly. Therefore, we hypothesize that the French subset of SynthDetoxM is a valuable addition to the dataset, enabling the training of effective detoxification models for French language processing tasks.

4.3 Automatic Evaluation Setup

To assess the quality of the generated Spanish, Russian, and German data, we follow the evaluation pipeline of Dementieva et al. (2024), developed for the multilingual text detoxification shared task. These metrics are inspired by prior work on monolingual text detoxification for English and Russian (Logacheva et al., 2022; Moskovskiy et al., 2022).

Style Transfer Accuracy (STA) For computation of this metric we use a multilingual toxicity classifier based on a multilingual XLM-R[4] (Conneau et al., 2020) text classification model, trained on a binary toxicity detection dataset.

Content Similarity (SIM) For computation of this metric we use the cosine distance between LaBSE[5] embeddings (Feng et al., 2022) of the source texts and the generated texts.

[4] hf.co/textdetox/xlmr-large-toxicity-classifier
[5] hf.co/sentence-transformers/LaBSE

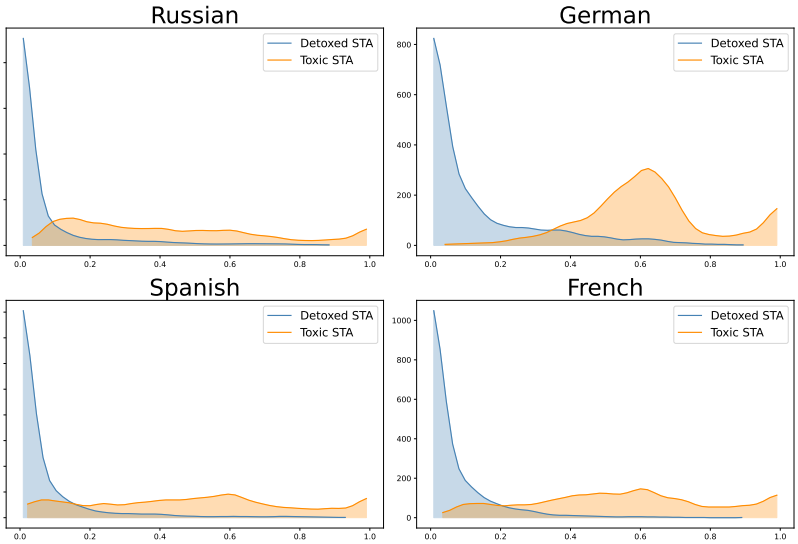

Figure 3: Distribution of STA toxicity scores of toxic and neutral examples in the dataset. The original toxic texts are in orange, while detoxified texts are in blue. For readability we apply Gaussian smoothing.

Fluency (FL) Fluency assesses how closely detoxified texts resemble human-written references. Previous works on English text detoxification have employed CoLA-based classifiers to estimate text fluency (Dale et al., 2021; Logacheva et al., 2022; Moskovskiy et al., 2024). However, due to the absence of CoLA datasets for all considered languages in the shared task, Dementieva et al. (2024) used ChrF1 as a substitute. While recent work by Zhang et al. (2024) has introduced MELA, a multilingual extension of CoLA dataset covering all the languages in this study, we maintain the evaluation pipeline from the shared task and continue using ChrF1. Nonetheless, ChrF1 remains a coarse approximation of text fluency, which may negatively impact the overall **J** scores (see Appendix C for details).

Joint score (J) The metrics **STA**, **SIM** and **FL** are subsequently combined into the final **J** score which is used for the final ranking of approaches. Given an input toxic text x_i and its output detoxified version y_i, for a test set of n samples:

$$\mathbf{J} = \frac{1}{n}\sum_{i=1}^{n} \mathbf{STA}(y_i) \cdot \mathbf{SIM}(x_i, y_i) \cdot \mathbf{FL}(x_i, y_i),$$

where $\mathbf{STA}(y_i), \mathbf{SIM}(x_i, y_i), \mathbf{FL}(x_i, y_i) \in [0, 1]$ for each text detoxification output y_i.

4.4 Baselines

In our work we adopt the baselines for multilingual detoxification described in MultiParaDetox (Dementieva et al., 2024) that are inspired by prior works on text detoxification (Logacheva et al., 2022; Dementieva et al., 2023).

Duplicate is the simplest baseline possible which copies an input toxic sentence. This baseline has 1.0 (or 100%) **SIM** score by definition.

Delete removes the toxic words according to a predefined list of inappropriate words. Dementieva et al. (2024) collects the lists of such toxic keywords for all target languages based on openly available sources. These lists are available online[6].

Backtranslation has proven to be effective in previous works (Dementieva et al., 2023; Prabhumoye et al., 2018; Konovalov et al., 2016a). Following (Dementieva et al., 2023) we translate texts into English with NLLB translation model (Costa-jussà et al., 2022)[7]. The translated data is then detoxified with the ParaDetox BART (Logacheva et al., 2022) model[8]. After that, the detoxified texts are translated back into the source language using NLLB.

4.5 Training Configuration

In our experimental evaluation of the generated multilingual parallel detoxification dataset, SynthDetoxM, we adopt the most efficient approaches from the TextDetox 2024 Shared Task (Dementieva et al., 2024). The top three solutions in the automatic evaluations utilized fine-tuning of

[6] hf.co/multilingual_toxic_lexicon
[7] hf.co/facebook/nllb-200-distilled-600M
[8] hf.co/s-nlp/bart-base-detox

Dataset	STA	SIM	FL	J	STA·SIM
German					
MPD	0.722	0.848	0.602	0.383	0.612
SDM (Subset)	0.681	0.912	0.745	0.463	0.597
SDM	**0.728**	0.899	0.734	**0.484**	**0.655**
SDM+MPD	0.615	**0.954**	**0.821**	0.483	0.586
Russian					
MPD	0.748	0.852	0.643	0.434	0.637
SDM (Subset)	0.858	0.850	0.656	0.478	0.729
SDM	**0.927**	0.839	0.656	**0.521**	**0.778**
SDM+MPD	0.815	**0.886**	**0.726**	0.540	0.721
Spanish					
MPD	0.597	0.880	0.616	0.335	0.525
SDM (Subset)	0.795	0.856	0.611	0.416	0.681
SDM	**0.864**	0.861	0.621	**0.471**	**0.744**
SDM+MPD	0.681	**0.907**	**0.653**	0.413	0.618

Table 3: Results of the automatic evaluation for mT0-XL on German, Russian, and Spanish trained on original data (MPD stands for MultiParaDetox), ours proposed SynthDetoxM (SDM stands for SynthDetoxM) and the consecutive training on both MultiParaDetox and ours proposed SynthDetoxM (SDM + MPD stands for SynthDetoxM + MultiParaDetox).

	German	Spanish	Russian
Human References	0.733	0.709	0.732
Baselines			
Duplicate	0.287	0.090	0.048
Delete	0.362	0.319	0.255
Backtranslation	0.233	0.275	0.223
mT0-XL supervised fine-tuning			
MultiParaDetox	0.446	0.344	0.472
SDM (Subset)	0.460	0.402	0.475
SDM	**0.482**	**0.470**	**0.546**
10-shot LLM prediction			
Gemma 2	0.353	0.380	0.404
Mistral Nemo	0.286	0.290	0.258
Mistral Small	0.371	0.308	0.273
Command R	0.328	0.344	0.402
Qwen 2.5	0.402	0.443	0.428
Llama 3.1 8B	0.394	0.341	0.357
Aya Expanse 8B	0.305	0.246	0.225
Aya Expanse 32B	0.399	0.320	0.323

Table 4: Text detoxification results in terms of **J** scores for German, Spanish, and Russian languages. The best overall results are boldfaced. The baselines and human references are from (Dementieva et al., 2024).

the multilingual encoder-decoder language model mT0 (Muennighoff et al., 2023).

We use mT0-XL model[9] and perform fine-tuning in full precision (fp32). We use AdaFactor optimizer (Shazeer and Stern, 2018) with batch size of 16, 50 warmup steps and set maximum sequence length to 512.

We fine-tune mT0 for 2 epochs in all setups. According to our experiments, the increased number of training epochs does not increase the final performance of the model. This might be explained to the overall training data scarcity compared to the size of the model: mT0-XL has 3 billion parameters and is being fine-tuned on 1,200 samples (400 for each of the three languages).

5 Results

Table 3 presents the results of our experimental evaluation of SynthDetoxM. For clarity, the table is divided by language. We compare the performance of mT0-XL trained on human-annotated MultiParaDetox data (MPD) with mT0-XL fine-tuned on two subsets of SynthDetoxM: a subset of 400 samples per language, matching the MPD size (denoted as SDM (Subset)), and the full SynthDetoxM dataset. Additionally, following prior work (Xu et al., 2023), we investigate whether a two-stage fine-tuning approach—first on SynthDetoxM, then on MPD (denoted as SDM + MPD)—yields further improvements.

The highest **SIM** scores for German and Russian are achieved by mT0-XL trained on MultiParaDetox (0.954 and 0.886, respectively), with the two-stage training approach (SDM + MPD) yielding slightly higher similarity in Russian but lower in German. However, for **STA**, models trained on SynthDetoxM consistently outperform MultiParaDetox across all languages, both when trained on the full dataset and on a similarly sized subset.

Models trained on SynthDetoxM exhibit slightly lower **FL** scores, likely due to the reference-dependent nature of this metric. Despite this, the aggregated **J** metric—strongly influenced by **FL**—is significantly higher for models trained on both the full SynthDetoxM and its subset compared to MultiParaDetox. Notably, incorporating Multi-

[9] hf.co/bigscience/mt0-xl

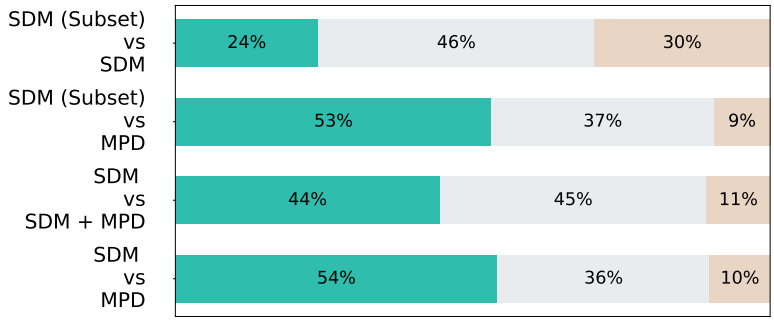

Figure 4: Side-by-side comparison of model outputs across all languages, evaluated by GPT-4o. The results highlight the relative performance of the models in generating detoxified text for German, Russian, and Spanish. The notation is similar to the notation from Table 3.

ParaDetox into the training process (SDM + MPD) results in a drop in **J** scores.

To further illustrate the advantages of training on SynthDetoxM, Table 3 also presents the product of **STA** and **SIM**. Even without considering **FL**, models trained on SynthDetoxM outperform those trained on MultiParaDetox in all setups and languages.

Additionally, Table 4 reports the **J** scores of a `mT0-XL` model trained on MultiParaDetox, averaged across 10 subsets of SynthDetoxM and the full SynthDetoxM, compared to baselines and large language model (LLM)-based detoxification in a 10-shot generation setting.

Finally, we provide a Side-by-Side (SBS) comparison of the fine-tuned `mT0-XL` models to evaluate their detoxification performance. Following (Moskovskiy et al., 2024), we employ GPT-4o as an evaluator to select the preferred detoxified outputs. The results of this comparison across German, Spanish, and Russian languages are summarized in Figure 4 and detailed in Appendix F.

Our SBS evaluation shows a clear preference for detoxifications produced by SDM over MultiParaDetox (MPD), with SDM winning in 59% of cases compared to 19% for MPD, and 22% resulting in ties. The subset of SDM also outperforms MPD in 58% of cases.

When comparing SDM against its combination with MPD (SDM + MPD), SDM is preferred in 47% of cases, with 21% favoring SDM + MPD, and 33% tied. Additionally, the full SDM dataset is slightly preferred over the batch processing version in 55% of cases, with 35% ties. See Appendix F for more details.

6 Conclusions

We present several contributions to multilingual text detoxification technology. Firstly, we successfully extend the concept of few-shot prompting for detoxification to a multilingual context, building upon previous monolingual approaches and propose a framework for generation of multilingual synthetic detoxification data. Secondly, we introduce SynthDetoxM, a large-scale multilingual synthetic parallel dataset designed to address the long-standing issue of data scarcity in text detoxification research. Notably, our dataset, created using our selection criteria, demonstrates competitive quality to existing human-annotated datasets, surpassing them in both low resource and high resource settings.

Our comprehensive evaluation of SynthDetoxM reveals its effectiveness in training high-performing models for text detoxification. Specifically, our experiments show that models trained on our dataset outperform those, which were trained on a similar amount of human-annotated data. Furthermore, training a detoxification encoder-decoder model on full SynthDetoxM yields a model, which surpasses the performance of most large language models in few-shot generation setups.

Findings presented in our work, show usefulness of the generated data for the task of multilingual text detoxification and pave the way for future research and developments of related technologies.

Acknowledgments

The contribution of E.T. was supported by the Kazan Federal University Strategic Academic Leadership Program ("PRIORITY-2030"), Strategic Project №5. We acknowledge the computational resources of HPC facilities at the HSE University.

7 Limitations

One of the limitations of our work is that we are focusing only on explicit type of toxicity. Additionally, definition and type of toxicity changes drastically between the language, e.g. things, that are toxic in one language may be perfectly normal in other language.

Another limitation of this work is our constraint with computational resources, which led to our use of smaller and simpler models for synthetic data generation, which could fit into a single NVIDIA A100 80GB GPU. Usage of larger could potentially result in higher quality and diversity of synthetic data.

Moreover, the comparison with proprietary models would strengthen the evaluation as it is done in recent works Dementieva et al. (2025).

Additionally, we were limited by the amount of annotated non-parallel toxic datasets in some of the languages, which limited the amount of possible generated synthetic data. In future, we plan to extend our work to other languages, such as Italian, Polish and others.

8 Ethical Considerations

While working with the task detoxification we are fully aware of the ethical responsibilities involved. As researchers, we handle this sensitive area with care and integrity. The main goal of text detoxification is to make online interactions safer and more inclusive by reducing harmful or offensive language.

While these datasets are meant to train models to detect and reduce toxic language, there's a chance they could be used in the wrong way—such as creating models that spread harmful or offensive content. This could lead to hate speech and harassment.

It's also important to clarify that the goal of text detoxification isn't to suppress free speech or force automatic changes to content. Instead, we aim to build models that offer non-toxic alternatives, helping users choose better language on their own. By giving suggestions rather than enforcing edits, we respect people's freedom while encouraging a more positive online environment.

References

Dennis Assenmacher, Marco Niemann, Kilian Müller, Moritz Seiler, Dennis M. Riehle, and Heike Trautmann. 2021. Rp-mod & rp-crowd: Moderator- and crowd-annotated german news comment datasets. In *Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks 1, NeurIPS Datasets and Benchmarks 2021, December 2021, virtual*.

Katherine Atwell, Sabit Hassan, and Malihe Alikhani. 2022. APPDIA: A discourse-aware transformer-based style transfer model for offensive social media conversations. In *Proceedings of the 29th International Conference on Computational Linguistics, COLING 2022, Gyeongju, Republic of Korea, October 12-17, 2022*, pages 6063–6074. International Committee on Computational Linguistics.

Anatoly Belchikov. 2019. Russian language toxic comments. Accessed: 2024-10-14.

Eleftheria Briakou, Di Lu, Ke Zhang, and Joel R. Tetreault. 2021. Olá, bonjour, salve! XFORMAL: A benchmark for multilingual formality style transfer. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2021, Online, June 6-11, 2021*, pages 3199–3216. Association for Computational Linguistics.

Arthur Capozzi, Gianmarco De Francisci Morales, Yelena Mejova, Corrado Monti, André Panisson, and Daniela Paolotti. 2021. Clandestino or rifugiato? anti-immigration facebook ad targeting in italy. In *CHI '21: CHI Conference on Human Factors in Computing Systems, Virtual Event / Yokohama, Japan, May 8-13, 2021*, pages 179:1–179:15. ACM.

Cohere. 2024. Command series 0824.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, Online, July 5-10, 2020*, pages 8440–8451. Association for Computational Linguistics.

Marta R. Costa-jussà, James Cross, Onur Çelebi, Maha Elbayad, Kenneth Heafield, Kevin Heffernan, Elahe Kalbassi, Janice Lam, Daniel Licht, Jean Maillard, Anna Y. Sun, Skyler Wang, Guillaume Wenzek, Al Youngblood, Bapi Akula, Loïc 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. *CoRR*, abs/2207.04672.

David Dale, Anton Voronov, Daryna Dementieva, Varvara Logacheva, Olga Kozlova, Nikita Semenov, and Alexander Panchenko. 2021. Text detoxification using large pre-trained neural models. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, EMNLP 2021, Virtual Event / Punta Cana, Dominican Republic, 7-11 November, 2021*, pages 7979–7996. Association for Computational Linguistics.

John Dang, Shivalika Singh, Daniel D'souza, Arash Ahmadian, Alejandro Salamanca, Madeline Smith, Aidan Peppin, Sungjin Hong, Manoj Govindassamy, Terrence Zhao, Sandra Kublik, Meor Amer, Viraat Aryabumi, Jon Ander Campos, Yi-Chern Tan, Tom Kocmi, Florian Strub, Nathan Grinsztajn, Yannis Flet-Berliac, Acyr Locatelli, Hangyu Lin, Dwarak Talupuru, Bharat Venkitesh, David Cairuz, Bowen Yang, Tim Chung, Wei-Yin Ko, Sylvie Shang Shi, Amir Shukayev, Sammie Bae, Aleksandra Piktus, Roman Castagné, Felipe Cruz-Salinas, Eddie Kim, Lucas Crawhall-Stein, Adrien Morisot, Sudip Roy, Phil Blunsom, Ivan Zhang, Aidan Gomez, Nick Frosst, Marzieh Fadaee, Beyza Ermis, Ahmet Üstün, and Sara Hooker. 2024. Aya expanse: Combining research breakthroughs for a new multilingual frontier. *Preprint*, arXiv:2412.04261.

Daryna Dementieva, Nikolay Babakov, Amit Ronen, Abinew Ali Ayele, Naquee Rizwan, Florian Schneider, Xintong Wang, Seid Muhie Yimam, Daniil Alekhseevich Moskovskiy, Elisei Stakovskii, Eran Kaufman, Ashraf Elnagar, Animesh Mukherjee, and Alexander Panchenko. 2025. Multilingual and explainable text detoxification with parallel corpora. In *Proceedings of the 31st International Conference on Computational Linguistics, COLING 2025, Abu Dhabi, UAE, January 19-24, 2025*, pages 7998–8025. Association for Computational Linguistics.

Daryna Dementieva, Daniil Moskovskiy, Nikolay Babakov, Abinew Ali Ayele, Naquee Rizwan, Florian Schneider, Xintong Wang, Seid Muhie Yimam, Dmitry Ustalov, Elisei Stakovskii, Alisa Smirnova, Ashraf Elnagar, Animesh Mukherjee, and Alexander Panchenko. 2024. Overview of the multilingual text detoxification task at PAN 2024. In *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 2432–2461. CEUR-WS.org.

Daryna Dementieva, Daniil Moskovskiy, David Dale, and Alexander Panchenko. 2023. Exploring methods for cross-lingual text style transfer: The case of text detoxification. In *Proceedings of the 13th International Joint Conference on Natural Language Processing and the 3rd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics, IJCNLP 2023 -Volume 1: Long Papers, Nusa Dua, Bali, November 1 - 4, 2023*, pages 1083–1101. Association for Computational Linguistics.

Cícero Nogueira dos Santos, Igor Melnyk, and Inkit Padhi. 2018. Fighting offensive language on social media with unsupervised text style transfer. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics, ACL 2018, Melbourne, Australia, July 15-20, 2018, Volume 2: Short Papers*, pages 189–194. Association for Computational Linguistics.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, Arun Rao, Aston Zhang, Aurélien Rodriguez, Austen Gregerson, Ava Spataru, Baptiste Rozière, Bethany Biron, Binh Tang, Bobbie Chern, Charlotte Caucheteux, Chaya Nayak, Chloe Bi, Chris Marra, Chris McConnell, Christian Keller, Christophe Touret, Chunyang Wu, Corinne Wong, Cristian Canton Ferrer, Cyrus Nikolaidis, Damien Allonsius, Daniel Song, Danielle Pintz, Danny Livshits, David Esiobu, Dhruv Choudhary, Dhruv Mahajan, Diego Garcia-Olano, Diego Perino, Dieuwke Hupkes, Egor Lakomkin, Ehab AlBadawy, Elina Lobanova, Emily Dinan, Eric Michael Smith, Filip Radenovic, Frank Zhang, Gabriel Synnaeve, Gabrielle Lee, Georgia Lewis Anderson, Graeme Nail, Grégoire Mialon, Guan Pang, Guillem Cucurell, Hailey Nguyen, Hannah Korevaar, Hu Xu, Hugo Touvron, Iliyan Zarov, Imanol Arrieta Ibarra, Isabel M. Kloumann, Ishan Misra, Ivan Evtimov, Jade Copet, Jaewon Lee, Jan Geffert, Jana Vranes, Jason Park, Jay Mahadeokar, Jeet Shah, Jelmer van der Linde, Jennifer Billock, Jenny Hong, Jenya Lee, Jeremy Fu, Jianfeng Chi, Jianyu Huang, Jiawen Liu, Jie Wang, Jiecao Yu, Joanna Bitton, Joe Spisak, Jongsoo Park, Joseph Rocca, Joshua Johnstun, Joshua Saxe, Junteng Jia, Kalyan Vasuden Alwala, Kartikeya Upasani, Kate Plawiak, Ke Li, Kenneth Heafield, Kevin Stone, and et al. 2024. The llama 3 herd of models. *CoRR*, abs/2407.21783.

Fangxiaoyu Feng, 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), ACL 2022, Dublin, Ireland, May 22-27, 2022*, pages 878–891. Association for Computational Linguistics.

Paula Fortuna and Sérgio Nunes. 2018. A survey on automatic detection of hate speech in text. *ACM Computing Surveys (CSUR)*, 51(4):1–30.

Zhenxin Fu, Xiaoye Tan, Nanyun Peng, Dongyan Zhao, and Rui Yan. 2018. Style transfer in text: Exploration and evaluation. In *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence, (AAAI-18), the 30th innovative Applications of Artificial Intelligence (IAAI-18), and the 8th AAAI Symposium on Educational Advances in Artificial Intelligence (EAAI-18), New Orleans, Louisiana, USA, February 2-7, 2018*, pages 663–670. AAAI Press.

Skyler Hallinan, Alisa Liu, Yejin Choi, and Maarten Sap. 2023. Detoxifying text with marco: Controllable re-

vision with experts and anti-experts. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers), ACL 2023, Toronto, Canada, July 9-14, 2023*, pages 228–242. Association for Computational Linguistics.

Zachary Horvitz, Ajay Patel, Kanishk Singh, Chris Callison-Burch, Kathleen R. McKeown, and Zhou Yu. 2024. Tinystyler: Efficient few-shot text style transfer with authorship embeddings. In *Findings of the Association for Computational Linguistics: EMNLP 2024, Miami, Florida, USA, November 12-16, 2024*, pages 13376–13390. Association for Computational Linguistics.

Aaron Hurst, Adam Lerer, Adam P. Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, Aleksander Madry, Alex Baker-Whitcomb, Alex Beutel, Alex Borzunov, Alex Carney, Alex Chow, Alex Kirillov, Alex Nichol, Alex Paino, Alex Renzin, Alex Tachard Passos, Alexander Kirillov, Alexi Christakis, Alexis Conneau, Ali Kamali, Allan Jabri, Allison Moyer, Allison Tam, Amadou Crookes, Amin Tootoonchian, Ananya Kumar, Andrea Vallone, Andrej Karpathy, Andrew Braunstein, Andrew Cann, Andrew Codispoti, Andrew Galu, Andrew Kondrich, Andrew Tulloch, Andrey Mishchenko, Angela Baek, Angela Jiang, Antoine Pelisse, Antonia Woodford, Anuj Gosalia, Arka Dhar, Ashley Pantuliano, Avi Nayak, Avital Oliver, Barret Zoph, Behrooz Ghorbani, Ben Leimberger, Ben Rossen, Ben Sokolowsky, Ben Wang, Benjamin Zweig, Beth Hoover, Blake Samic, Bob McGrew, Bobby Spero, Bogo Giertler, Bowen Cheng, Brad Lightcap, Brandon Walkin, Brendan Quinn, Brian Guarraci, Brian Hsu, Bright Kellogg, Brydon Eastman, Camillo Lugaresi, Carroll L. Wainwright, Cary Bassin, Cary Hudson, Casey Chu, Chad Nelson, Chak Li, Chan Jun Shern, Channing Conger, Charlotte Barette, Chelsea Voss, Chen Ding, Cheng Lu, Chong Zhang, Chris Beaumont, Chris Hallacy, Chris Koch, Christian Gibson, Christina Kim, Christine Choi, Christine McLeavey, Christopher Hesse, Claudia Fischer, Clemens Winter, Coley Czarnecki, Colin Jarvis, Colin Wei, Constantin Koumouzelis, and Dane Sherburn. 2024. Gpt-4o system card. *CoRR*, abs/2410.21276.

Md Tawkat Islam Khondaker, Muhammad Abdul-Mageed, and Laks V. S. Lakshmanan. 2024. Detoxllm: A framework for detoxification with explanations. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing, EMNLP 2024, Miami, FL, USA, November 12-16, 2024*, pages 19112–19139. Association for Computational Linguistics.

Ian Kivlichan, Jeffrey Sorensen, Julia Elliott, Lucy Vasserman, Martin Görner, and Phil Culliton. 2020. Jigsaw multilingual toxic comment classification.

Vasily Konovalov, Meni Adler, and Ido Dagan. 2016a. Effective paraphrase expansion in addressing lexical variability. In *Proceedings of the Artificial Intelligence and Natural Language (AINL FRUCT 2016)*, page 87–91, St.-Petersburg, Russia.

Vasily Konovalov, Oren Melamud, Ron Artstein, and Ido Dagan. 2016b. Collecting Better Training Data using Biased Agent Policies in Negotiation Dialogues. In *Proceedings of WOCHAT, the Second Workshop on Chatbots and Conversational Agent Technologies*, Los Angeles. Zerotype.

Kalpesh Krishna, John Wieting, and Mohit Iyyer. 2020. Reformulating unsupervised style transfer as paraphrase generation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, EMNLP 2020, Online, November 16-20, 2020*, pages 737–762. Association for Computational Linguistics.

Huiyuan Lai, Antonio Toral, and Malvina Nissim. 2021. Thank you bart! rewarding pre-trained models improves formality style transfer. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing, ACL/IJCNLP 2021, (Volume 2: Short Papers), Virtual Event, August 1-6, 2021*, pages 484–494. Association for Computational Linguistics.

Alyssa Lees, Vinh Q. Tran, Yi Tay, Jeffrey Sorensen, Jai Gupta, Donald Metzler, and Lucy Vasserman. 2022. A new generation of perspective api: Efficient multilingual character-level transformers. In *Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, KDD '22, page 3197–3207, New York, NY, USA. Association for Computing Machinery.

Shuai Liu, Shantanu Agarwal, and Jonathan May. 2024. Authorship style transfer with policy optimization. *CoRR*, abs/2403.08043.

Varvara Logacheva, Daryna Dementieva, Sergey Ustyantsev, Daniil Moskovskiy, David Dale, Irina Krotova, Nikita Semenov, and Alexander Panchenko. 2022. Paradetox: Detoxification with parallel data. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2022, Dublin, Ireland, May 22-27, 2022*, pages 6804–6818. Association for Computational Linguistics.

Mistral. 2024a. Ai in abundance | mistral ai | frontier ai in your hands.

Mistral. 2024b. Mistral nemo | mistral ai | frontier ai in your hands.

Daniil Moskovskiy, Daryna Dementieva, and Alexander Panchenko. 2022. Exploring cross-lingual text detoxification with large multilingual language models. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics: Student Research Workshop, ACL 2022, Dublin, Ireland, May 22-27, 2022*, pages 346–354. Association for Computational Linguistics.

Daniil Moskovskiy, Sergey Pletenev, and Alexander Panchenko. 2024. Llms to replace crowdsourcing for parallel data creation? the case of text detoxification. In *Findings of the Association for Computational Linguistics: EMNLP 2024, Miami, Florida, USA, November 12-16, 2024*, pages 14361–14373. Association for Computational Linguistics.

Niklas Muennighoff, Thomas Wang, Lintang Sutawika, Adam Roberts, Stella Biderman, Teven Le Scao, M. Saiful Bari, Sheng Shen, Zheng Xin Yong, Hailey Schoelkopf, Xiangru Tang, Dragomir Radev, Alham Fikri Aji, Khalid Almubarak, Samuel Albanie, Zaid Alyafeai, Albert Webson, Edward Raff, and Colin Raffel. 2023. Crosslingual generalization through multitask finetuning. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2023, Toronto, Canada, July 9-14, 2023*, pages 15991–16111. Association for Computational Linguistics.

Sourabrata Mukherjee, Akanksha Bansal, Pritha Majumdar, Atul Kr. Ojha, and Ondřej Dušek. 2023. Low-resource text style transfer for Bangla: Data & models. In *Proceedings of the First Workshop on Bangla Language Processing (BLP-2023)*, pages 34–47, Singapore. Association for Computational Linguistics.

Nedjma Ousidhoum, Zizheng Lin, Hongming Zhang, Yangqiu Song, and Dit-Yan Yeung. 2019. Multilingual and multi-aspect hate speech analysis. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, EMNLP-IJCNLP 2019, Hong Kong, China, November 3-7, 2019*, pages 4674–4683. Association for Computational Linguistics.

Maja Popovic. 2015. chrf: character n-gram f-score for automatic MT evaluation. In *Proceedings of the Tenth Workshop on Statistical Machine Translation, WMT@EMNLP 2015, 17-18 September 2015, Lisbon, Portugal*, pages 392–395. The Association for Computer Linguistics.

Mohammad Mahdi Abdollah Pour, Parsa Farinneya, Manasa Bharadwaj, Nikhil Verma, Ali Pesaranghader, and Scott Sanner. 2023. COUNT: contrastive unlikelihood text style transfer for text detoxification. In *Findings of the Association for Computational Linguistics: EMNLP 2023, Singapore, December 6-10, 2023*, pages 8658–8666. Association for Computational Linguistics.

Shrimai Prabhumoye, Yulia Tsvetkov, Ruslan Salakhutdinov, and Alan W. Black. 2018. Style transfer through back-translation. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics, ACL 2018, Melbourne, Australia, July 15-20, 2018, Volume 1: Long Papers*, pages 866–876. Association for Computational Linguistics.

Sudha Rao and Joel R. Tetreault. 2018. Dear sir or madam, may I introduce the GYAFC dataset: Corpus, benchmarks and metrics for formality style transfer. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2018, New Orleans, Louisiana, USA, June 1-6, 2018, Volume 1 (Long Papers)*, pages 129–140. Association for Computational Linguistics.

Machel Reid and Mikel Artetxe. 2023. On the role of parallel data in cross-lingual transfer learning. In *Findings of the Association for Computational Linguistics: ACL 2023, Toronto, Canada, July 9-14, 2023*, pages 5999–6006. Association for Computational Linguistics.

Machel Reid, Nikolay Savinov, Denis Teplyashin, Dmitry Lepikhin, Timothy P. Lillicrap, Jean-Baptiste Alayrac, Radu Soricut, Angeliki Lazaridou, Orhan Firat, Julian Schrittwieser, Ioannis Antonoglou, Rohan Anil, Sebastian Borgeaud, Andrew M. Dai, Katie Millican, Ethan Dyer, Mia Glaese, Thibault Sottiaux, Benjamin Lee, Fabio Viola, Malcolm Reynolds, Yuanzhong Xu, James Molloy, Jilin Chen, Michael Isard, Paul Barham, Tom Hennigan, Ross McIlroy, Melvin Johnson, Johan Schalkwyk, Eli Collins, Eliza Rutherford, Erica Moreira, Kareem Ayoub, Megha Goel, Clemens Meyer, Gregory Thornton, Zhen Yang, Henryk Michalewski, Zaheer Abbas, Nathan Schucher, Ankesh Anand, Richard Ives, James Keeling, Karel Lenc, Salem Haykal, Siamak Shakeri, Pranav Shyam, Aakanksha Chowdhery, Roman Ring, Stephen Spencer, Eren Sezener, and et al. 2024. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context. *CoRR*, abs/2403.05530.

Emily Reif, Daphne Ippolito, Ann Yuan, Andy Coenen, Chris Callison-Burch, and Jason Wei. 2022. A recipe for arbitrary text style transfer with large language models. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers), ACL 2022, Dublin, Ireland, May 22-27, 2022*, pages 837–848. Association for Computational Linguistics.

Julian Risch, Anke Stoll, Lena Wilms, and Michael Wiegand. 2021. Overview of the GermEval 2021 shared task on the identification of toxic, engaging, and fact-claiming comments. In *Proceedings of the GermEval 2021 Shared Task on the Identification of Toxic, Engaging, and Fact-Claiming Comments*, pages 1–12. Association for Computational Linguistics.

Morgane Rivière, Shreya Pathak, Pier Giuseppe Sessa, Cassidy Hardin, Surya Bhupatiraju, Léonard Hussenot, Thomas Mesnard, Bobak Shahriari, Alexandre Ramé, Johan Ferret, Peter Liu, Pouya Tafti, Abe Friesen, Michelle Casbon, Sabela Ramos, Ravin Kumar, Charline Le Lan, Sammy Jerome, Anton Tsitsulin, Nino Vieillard, Piotr Stanczyk, Sertan Girgin, Nikola Momchev, Matt Hoffman, Shantanu Thakoor, Jean-Bastien Grill, Behnam Neyshabur, Olivier Bachem, Alanna Walton, Aliaksei Severyn, Alicia Parrish, Aliya Ahmad, Allen Hutchison, Alvin Abdagic, Amanda Carl, Amy Shen, Andy Brock,

Andy Coenen, Anthony Laforge, Antonia Paterson, Ben Bastian, Bilal Piot, Bo Wu, Brandon Royal, Charlie Chen, Chintu Kumar, Chris Perry, Chris Welty, Christopher A. Choquette-Choo, Danila Sinopalnikov, David Weinberger, Dimple Vijaykumar, Dominika Rogozinska, Dustin Herbison, Elisa Bandy, Emma Wang, Eric Noland, Erica Moreira, Evan Senter, Evgenii Eltyshev, Francesco Visin, Gabriel Rasskin, Gary Wei, Glenn Cameron, Gus Martins, Hadi Hashemi, Hanna Klimczak-Plucinska, Harleen Batra, Harsh Dhand, Ivan Nardini, Jacinda Mein, Jack Zhou, James Svensson, Jeff Stanway, Jetha Chan, Jin Peng Zhou, Joana Carrasqueira, Joana Iljazi, Jocelyn Becker, Joe Fernandez, Joost van Amersfoort, Josh Gordon, Josh Lipschultz, Josh Newlan, Ju-yeong Ji, Kareem Mohamed, Kartikeya Badola, Kat Black, Katie Millican, Keelin McDonell, Kelvin Nguyen, Kiranbir Sodhia, Kish Greene, Lars Lowe Sjösund, Lauren Usui, Laurent Sifre, Lena Heuermann, Leticia Lago, and Lilly McNealus. 2024. Gemma 2: Improving open language models at a practical size. *CoRR*, abs/2408.00118.

Koustuv Saha, Eshwar Chandrasekharan, and Munmun De Choudhury. 2019. Prevalence and psychological effects of hateful speech in online college communities. *Proceedings of the 10th ACM Conference on Web Science*.

Alexander Semiletov. 2020. Toxic russian comments. Dataset.

Noam Shazeer and Mitchell Stern. 2018. Adafactor: Adaptive learning rates with sublinear memory cost. In *Proceedings of the 35th International Conference on Machine Learning, ICML 2018, Stockholmsmässan, Stockholm, Sweden, July 10-15, 2018*, volume 80 of *Proceedings of Machine Learning Research*, pages 4603–4611. PMLR.

Xiaofei Sun, Xiaoya Li, Jiwei Li, Fei Wu, Shangwei Guo, Tianwei Zhang, and Guoyin Wang. 2023. Text classification via large language models. In *Findings of the Association for Computational Linguistics: EMNLP 2023, Singapore, December 6-10, 2023*, pages 8990–9005. Association for Computational Linguistics.

Mirac Suzgun, Luke Melas-Kyriazi, and Dan Jurafsky. 2022. Prompt-and-rerank: A method for zero-shot and few-shot arbitrary textual style transfer with small language models. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing, EMNLP 2022, Abu Dhabi, United Arab Emirates, December 7-11, 2022*, pages 2195–2222. Association for Computational Linguistics.

Qwen Team. 2024. Qwen2.5: A party of foundation models.

Yunli Wang, Yu Wu, Lili Mou, Zhoujun Li, and Wen-Han Chao. 2020. Formality style transfer with shared latent space. In *Proceedings of the 28th International Conference on Computational Linguistics, COLING 2020, Barcelona, Spain (Online), December 8-13, 2020*, pages 2236–2249. International Committee on Computational Linguistics.

Benfeng Xu, Quan Wang, Yajuan Lyu, Dai Dai, Yongdong Zhang, and Zhendong Mao. 2023. S2ynre: Two-stage self-training with synthetic data for low-resource relation extraction. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2023, Toronto, Canada, July 9-14, 2023*, pages 8186–8207. Association for Computational Linguistics.

An Yang, Baosong Yang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Zhou, Chengpeng Li, Chengyuan Li, Dayiheng Liu, Fei Huang, et al. 2024. Qwen2 technical report. *arXiv preprint arXiv:2407.10671*.

Jiacheng Ye, Jiahui Gao, Qintong Li, Hang Xu, Jiangtao Feng, Zhiyong Wu, Tao Yu, and Lingpeng Kong. 2022. Zerogen: Efficient zero-shot learning via dataset generation. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing, EMNLP 2022, Abu Dhabi, United Arab Emirates, December 7-11, 2022*, pages 11653–11669. Association for Computational Linguistics.

Ping Yu, Yang Zhao, Chunyuan Li, and Changyou Chen. 2021. Rethinking sentiment style transfer. In *Findings of the Association for Computational Linguistics: EMNLP 2021, Virtual Event / Punta Cana, Dominican Republic, 16-20 November, 2021*, pages 1569–1582. Association for Computational Linguistics.

Ziyin Zhang, Yikang Liu, Weifang Huang, Junyu Mao, Rui Wang, and Hai Hu. 2024. MELA: multilingual evaluation of linguistic acceptability. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2024, Bangkok, Thailand, August 11-16, 2024*, pages 2658–2674. Association for Computational Linguistics.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Tianle Li, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zhuohan Li, Zi Lin, Eric P. Xing, Joseph E. Gonzalez, Ion Stoica, and Hao Zhang. 2024. Lmsys-chat-1m: A large-scale real-world LLM conversation dataset. In *The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024*. OpenReview.net.

A Prompts

A.1 LLM Detoxification Prompt

In our work, for all of the language models we use the Text Style Transfer prompt taken from work by Khondaker et al. (2024). Among others, this prompt shown to be giving better generation quality and helped to avoid additional content.

We slightly adjust it specifically to text detoxification and provide the full text of the prompt in Figure 5. In this work, we do not use any advanced prompting techniques except few-shot prompting.

Your task is text style transfer. Rewrite the following text into non-toxic language. You must match the target style and preserve the original meaning as much as possible. Do not include the input text in the response. Only generate the target text and nothing else. Make sure the response is completely non-toxic and neutral. Answer only in {language}. Here are few examples:", Toxic text: {toxic_text}. Neutral text:

Figure 5: Detoxification prompt we use for synthetic parallel data generation. {toxic_text} stands for a placeholder for a given toxic text being prompted into LLM. In few-shot setting we add few examples of detoxification before last two lines and write: *Here are few examples:*.

A.2 mT0 Prompt

Model	German	Spanish	French	Russian
Llama 3.1 8B	662	619	773	1648
Llama 3.1 70B	898	981	1114	1354
Mistral Nemo	622	583	392	1320
Mistral Small	862	985	565	2237
Qwen 2.5 32B	477	819	513	3128
Aya Exp. 32B	458	453	142	945
Aya Exp. 8B	316	330	143	765
Command-R 32B	273	492	308	2294
Gemma 2 27B	394	564	360	2019

Table 5: Number of accepted samples in the final SynthDetoxM dataset, broken down by language and LLMs.

Write a non-toxic version of the following text in {language}: {toxic_text}

Figure 6: Detoxification prompt we use for mT0.

We add specific prompt to mT0 during both training and predictions. In our work do not translate the prompt into the language of generation. Instead, we follow the promping approach from the original paper (Muennighoff et al., 2023). We show the prompt in Figure 6.

B Automatic evaluation results

The automatic evaluation results are presented in Table 6.

Dataset	STA	SIM	CHRF	J
German				
MPD	0.722	0.848	0.602	0.383
SDM (Subset)	0.681 ±0.213	0.912 ±0.042	0.745 ±0.035	0.463 ±0.117
SDM (Full)	0.728	0.899	0.734	0.484
SDM+MPD	0.615	0.954	0.821	0.483
Russian				
MPD	0.748	0.852	0.643	0.434
SDM (Subset)	0.858 ±0.034	0.850 ±0.020	0.656 ±0.021	0.478 ±0.014
SDM (Full)	0.927	0.839	0.656	0.521
SDM+MPD	0.815	0.886	0.726	0.540
Spanish				
MPD	0.597	0.880	0.616	0.335
SDM (Subset)	0.795 ±0.083	0.856 ±0.031	0.611 ±0.022	0.416 ±0.023
SDM (Full)	0.864	0.861	0.621	0.471
SDM+MPD	0.681	0.907	0.653	0.413

Table 6: Results of the automatic evaluation for mT0-XL on German, Russian, and Spanish trained on original data (MPD stands for MultiParaDetox), our collected and synthetically generated data (SDM stands for SynthDetoxM) and on their combination (MultiParaDetox + SynthDetoxM).

C Limitations of ChrF1 as a Fluency Metric

This section addresses the issues with using ChrF1 to evaluate fluency in text detoxification. While ChrF1 is commonly used in neural machine translation (Popovic, 2015), it has significant drawbacks for text style transfer tasks like detoxification.

Reference-based metrics like ChrF1 are ill-suited for assessing fluency in detoxification. The goal is to change the text's style while maintaining meaning and fluency, without limiting the extent of edits. Effective detoxification often involves substantial structural changes, making comparisons with the original toxic text using ChrF1 misleading. Though ChrF1 may produce low scores, manual evaluations frequently show that the detoxified output is fluent.

ChrF1, based on character n-grams, is sensitive to word order and structural changes, which are often necessary for detoxification. It also fails to consider semantic content, meaning fluency can be high even when the ChrF1 score is low. Additionally, it tends to reward minimal edits, which undermines the goal of thorough detoxification.

	Spanish↓	German↓	Russian↓
Toxic	2089	323	4467
Detoxified	27	102	14

Table 7: Total amount of toxic words for toxic and detoxified subsets of SynthDetoxM with respect to language.

	Spanish↓	German↓	Russian↓
Toxic	0.522	0.081	1.117
Detoxified	0.007	0.036	0.004

Table 8: Average number of toxic words per text in the toxic and detoxified SynthDetoxM with respect to language.

> Politely refuse to answer this in `{lang}` and provide an explanation why you refuse. The refusal should be connected to the request topic. Do not add anything additional, only respond with a refusal: `{input_text}`

Figure 7: Refusal generation prompt for synthetic refusals dataset.

Recent research has shifted towards more appropriate fluency metrics. For example, CoLA-based classifiers, as used in Dementieva et al. (2023) and Logacheva et al. (2022), focus on linguistic acceptability, offering a more accurate assessment of fluency without relying on comparisons to the toxic input.

While ChrF1 has its merits in other tasks, it is not suitable for evaluating fluency in detoxification. Future work should prioritize methods that assess fluency based on grammaticality and naturalness, independent of the original text.

D Refusal classifier training

To get rid of LLM refusals in SynthDetoxM, we trained a separate refusal classifier, based on `xlmr-base`[10].

To train the model, a high quality synthetic dataset was created[11]. It was based on randomly selected inputs from the LMSYS-Chat-1M (Zheng et al., 2024) dataset and then passed to the Gemini Flash 1.5 (Reid et al., 2024) and Llama 3.3 70B (Dubey et al., 2024) models, prompted to generate both responses to the prompts from the dataset and refusals. Prompt for synthetic refusal generation is presented in Figure 7.

Classification model was trained using a batch size of 64, learning rate of 1e-4 for one epoch.

[10] hf.co/s-nlp/xlmr-base-refusal-classifier

[11] hf.co/datasets/chameleon-lizard/multilingual_refusals

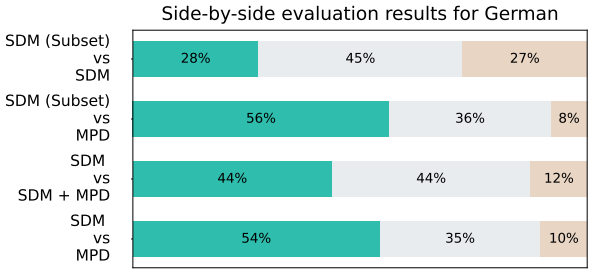

Figure 8: Side-by-side comparison of model outputs across all languages, evaluated by GPT-4o. The results highlight the relative performance of the models in generating detoxified text for German. The notation is similar to the notation from Table 3.

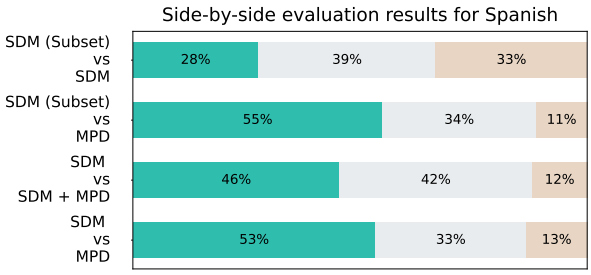

Figure 9: Side-by-side comparison of model outputs across all languages, evaluated by GPT-4o. The results highlight the relative performance of the models in generating detoxified text for Spanish. The notation is similar to the notation from Table 3.

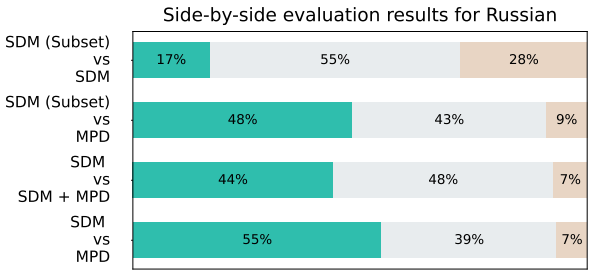

Figure 10: Side-by-side comparison of model outputs across all languages, evaluated by GPT-4o. The results highlight the relative performance of the models in generating detoxified text for Russian. The notation is similar to the notation from Table 3.

E Additional linguistic analysis of the dataset

To provide additional perspective about detoxification quality of our dataset, we used dataset[12], which contains toxic lexicon in German, Spanish and Russian languages and calculated two metrics of our generated data: the amount and mean counts of toxic words in each sentence (presented in Tables 7 and 8 accordingly) in the toxic and generated detoxified subsets of SynthDetoxM.

Due to the lack of French toxic lexicon dataset we did not do any evaluations in French. Furthermore, selected dataset is not comprehensive: for instance, German subset contains only 247 toxic words, which leaves some toxic sentences not having any toxic words detected and overall toxicity of German subset of our dataset is comparatively low. However, these evaluations still show that our detoxified data contains substantially less toxic lexicon than the original toxic data, indicating that overall explicit toxicity of the generated data is much lower after the detoxification.

F Side-by-side comparison of trained mT0 models

To provide additional validation outside of methodology created by (Dementieva et al., 2024), we have evaluated the responses from the test dataset of each of the trained mT0 models using GPT-4o (Hurst et al., 2024) as a judge. To lessen the positional bias, each side-by-side comparison was done twice, changing the positions of the answers and calculating mean score for all answers. The results are presented in the Figures 8, 9, 10.

As demonstrated in the figures, models fine-tuned on SynthDetoxM achieve significantly higher win rates compared to those trained on the "gold-standard" MultiParaDetox dataset, as well as models trained using a two-stage regime. We attribute this superior performance to the challenges posed by human-sourced samples, which may be too complex or nuanced for relatively simple and compact student models to learn effectively. This limitation likely contributes to the weaker performance of models trained on MultiParaDetox in both automatic evaluations and side-by-side comparisons.

[12] hf.co/datasets/textdetox/multilingual_toxic_lexicon

G Per-Language Few-Shot Examples

To ensure reproducibility we show all the few-shot prompts *as is* in Tables 9 10, 11, and 12.

Type	Text	Perspective API Score	XLM-R Score
Toxic	тварь присягу один раз,а проебал всю страну *The bastard swore an oath once, but screwed over the whole country.*	0.820	1.000
Neutral	Присягу дал один раз, а разрушил всю страну *He took the oath once, but destroyed the whole country*	0.360	0.001
Toxic	нахрена этого придурка клоуна сюда впихивают? *why the fuck is this idiot clown being pushed in here?*	0.743	1.000
Neutral	Зачем этого человека сюда впихивают? *Why is this man being pushed in here?*	0.042	0.005
Toxic	молись не молись,всё - равно получишь хуй. ну, разве что блядям на радость *Pray or don't pray, you'll get a dick anyway. Well, except for the joy of whores.*	0.757	1.000
Neutral	Молись не молись, всё - равно ничего не получишь. ну, разве что завистникам на радость. *Pray or don't pray, you won't get anything anyway. well, except for the envy of the joy.*	0.156	0.998
Toxic	её муженька козла на кол надо посадить *Her scoundrel of a husband needs to be impaled.*	0.553	0.999
Neutral	Её муженька нужно сурово наказать *Her hubby needs to be severely punished*	0.029	0.988
Toxic	хорошо пизды не получил уёбок *good thing the bastard didn't get his ass kicked.*	0.853	1.000
Neutral	хорошо что он не был избит *it's a good thing he wasn't beaten up*	0.037	0.000
Toxic	да что ты доебалась до этих клумб? людям жрать нечего! по мусоркам лазают! *why are you so fixated on those flower beds? people have nothing to eat! they're rummaging through trash!*	0.548	1.000
Neutral	да что ты докопалась до этих клумб? Людям есть нечего! по мусоркам лазают! *why did you get to these flower beds? People have nothing to eat! they're climbing through the trash!*	0.115	0.000
Toxic	именно благодаря этому президенту, россия еще жива. а такие долбоёбы, только могут воду мутить. дебилоид. *It's precisely thanks to this president that Russia is still alive. And idiots like you can only stir up trouble. Moron.*	0.863	1.000
Neutral	Именно благодаря этому президенту, Россия ещё жива, а такие люди, как ты, только и могут, что наводить смуту. *It is thanks to this president that Russia is still alive, and people like you can only cause trouble.*	0.369	0.003
Toxic	ты просто пидор,а не немец.я за свои слова отвечу, *you're just a faggot, not a german. I stand by my words, but you won't, because you're a couch scumbag.*	0.906	1.000
Neutral	ты просто не немец. я за свои слова отвечу, а ты н *you're just not german. i'll answer for my words, but you won't, because you're writing from the couch*	0.399	0.002
Toxic	да, блядь, кто это говнище постоянно сюда сливает?! *Yeah, fuck, who keeps dumping this shit here all the time?!*	0.654	1.000
Neutral	да кто эту ерунду постоянно сюда сливает?! *who is constantly pouring this nonsense here?!*	0.096	0.001
Toxic	да вы заебали этим постом уже, пидарасы!!! *I'm sick of this post already, you bastards!!!*	0.903	1.000
Neutral	да вы надоели этим постом уже, люди!!! *Yes, people, I am tired of this post already!!!*	0.012	0.001

Table 9: Selected few-shot examples for the Russian language with toxicity scores.

Type	Text	Perspective API Score	XLM-R Score
Toxic	In dem Schwarz Grün regierden Hessen werden jetzt 8 zusätzliche Psychologen nur für die Jugendlichen Invasoren eingestellt. Jetzt AFD wählen *In the black-green governed Hesse, 8 additional psychologists are now being hired only for the young invader. Vote AFD now*	0.397	0.486
Neutral	In dem Schwarz Grün regierden Hessen werden jetzt 8 zusätzliche Psychologen nur für die Jugendlichen Einwanderer eingestellt. Jetzt AFD wählen *In the black-green governed Hesse, 8 additional psychologists are now being hired only for the young immigrants. Vote AFD now*	0.313	0.004
Toxic	Moin, moin, seid Ihr alle noch gesund und munter bei all dem geistigen Dünnpfiff im "Merkel-Propagandaschrott"? *Hello, hello, are you all still healthy and cheerful despite all the intellectual drivel in the "Merkel propaganda junk"?*	0.471	0.898
Neutral	Moin, moin, seid Ihr alle noch gesund und munter bei all dem geistigen Quatsch in der "Merkel-Propaganda"? *Hello, hello, are you all still healthy and cheerful despite all the intellectual nonsense in the "Merkel propaganda"?*	0.427	0.999
Toxic	Ich würde Grüne mit einem Geschwür am Arsch der Welt bezeichnen *I would describe Greens as a boil on the ass of the world*	0.548	0.999
Neutral	Ich würde Grüne mit einem Geschwür der Welt bezeichnen. *I would describe Greens as a boil of the world.*	0.267	0.337
Toxic	Dummes Gelaber, Kern. *Stupid talk, Kern.*	0.397	0.653
Neutral	Unsinnige Aussage, Kern *Nonsensical statement, Kern*	0.177	0.002
Toxic	wie kann man nur derart strunzdumme Moderatoren wie bei #taff auf #ProSieben beschäftigen *how can you employ such extremely stupid presenters as at #taff on #ProSieben*	0.634	0.016
Neutral	wie kann man nur solche Moderatoren wie bei #taff auf #ProSieben beschäftigen *how can you employ such presenters as at #taff on #ProSieben*	0.056	0.001
Toxic	Da macht sich die Bildung aus der Bild mal wieder bemerkbar. So ein Sau dummer Kommentar *That's where the education from Bild becomes noticeable again. Such a pig stupid comment*	0.557	0.332
Neutral	Da macht sich die Bildung aus der Bild mal wieder bemerkbar. Ein sehr unüberlegter Beitrag *That's where the education from Bild becomes noticeable again. A very inconsiderate contribution*	0.110	0.000
Toxic	Kurz lasst dich nicht von der Trulla aufs Glatteis führen! *Kurz don't let the chick lead you astray!*	0.224	0.384
Neutral	Kurz lasst dich nicht von der Frau aufs Glatteis führen! *Kurz don't let the woman lead you astray!*	0.185	0.012
Toxic	Younower zu werden ist der finale schritt um ein Hurensohn zu werden *Becoming a Younower is the final step to becoming a son of a bitch*	0.638	0.999
Neutral	Younower zu werden ist der finale schritt um unsympathisch zu werden *Becoming a Younower is the final step to becoming unsympathetic*	0.231	0.676
Toxic	Sehe ich Kathrin Göring Eckart, Merkel, Gysi, Petry, Wagenknecht und all die anderen Wendehälse die uns heute regieren. Frage ich mich. War der Fall der Mauer ein Fehler. *When I see Kathrin Göring Eckart, Merkel, Gysi, Petry, Wagenknecht and all the other turncoats who govern us today. I ask myself. Was the fall of the wall a mistake.*	0.339	0.733
Neutral	Sehe ich Kathrin Göring Eckart, Merkel, Gysi, Petry, Wagenknecht und all die anderen wechselnhaften Politiker die uns heute regieren. Frage ich mich. War der Fall der Mauer ein Fehler. *When I see Kathrin Göring Eckart, Merkel, Gysi, Petry, Wagenknecht and all the other fickle politicians who govern us today. I ask myself. Was the fall of the wall a mistake.*	0.058	0.003
Toxic	Hahaha vllt gönn ich mir das ma. Aber diese asiaten sind einfach freaks, das is nich menschlich *Hahaha maybe I'll treat myself to that. But these Asians are just freaks, that's not human*	0.572	0.273
Neutral	Hahaha vllt gönn ich mir das ma. Aber diese asiaten sind einfach unglaublich, fast übermenschlich *Hahaha maybe I'll treat myself to that. But these Asians are just unbelievable, almost superhuman*	0.235	0.995

Table 10: Selected few-shot examples for German language with toxicity scores.

Type	Text	Perspective API Score	XLM-R Score
Toxic	Y lo bien que les va siendo fachas, que no es gratis *And how well they are doing being fascists, which is not free*	0.184	0.002
Neutral	Y lo bien que les va siendo conservadores, que no es gratis *And how well they are doing being conservatives, which is not free*	0.012	0.574
Toxic	Y a mí qué me importa lo que tú "notes", facha? *And what do I care what you "notice", fascist?*	0.427	0.391
Neutral	Y a mí qué me importa lo que tú "notes"? *And what do I care what you "notice"?*	0.080	0.001
Toxic	ya ves además a los policías nazis que tendrá esa en el whatsapp… como para creerse algo. *you see also the nazi policemen that she will have on whatsapp... as if to believe something.*	0.261	0.608
Neutral	ya ves además a los policías que tendrá esa en el whatsapp… como para creerse algo. *you see also the policemen that she will have on whatsapp... as if to believe something.*	0.035	0.003
Toxic	#YoSoyAnonymousUE basta de absurda represión victimista por parte del Estado Fascista *#YoSoyAnonymousUE enough of absurd victimist repression by the Fascist State*	0.255	0.352
Neutral	#YoSoyAnonymousUE basta de absurda represión victimista por parte del Estado *#YoSoyAnonymousUE enough of absurd victimist repression by the State*	0.106	0.034
Toxic	Supongo que Perro le dará asilo en la Moncloa *I suppose Dog will give him asylum in La Moncloa*	0.314	0.283
Neutral	Supongo que el presidente le dara asilo en la Moncloa *I suppose the president will give him asylum in La Moncloa*	0.039	0.004
Toxic	Los halagos y buenos modales siempre hacen bien, menos a los izquierdistas de la cultura del odio. *Flattery and good manners are always good, except for the leftists of the culture of hatred.*	0.285	0.969
Neutral	Los halagos y buenos modales siempre hacen bien, menos a los izquierdas de la cultura del odio. *Flattery and good manners are always good, except for the leftists of the culture of hatred.*	0.280	0.236
Toxic	Yo me quiero escapar también, dan paguita al llega *I want to escape too, they give a small handout to the newcomer*	0.011	0.423
Neutral	¿Ofrecen algún tipo de apoyo al llegar? *Do they offer any kind of support upon arrival?*	0.002	0.000
Toxic	Mire la saga de 50 sombras de Grey...casi todo eran.lectoras...les va el.morbillo no lo.nieguen *Look at the 50 Shades of Grey saga...almost all were.readers...they like.morbid curiosity don't.deny it*	0.039	0.811
Neutral	Miren la saga de 50 sombras de Grey.. casi todo eran lectoras... le encanta *Look at the 50 Shades of Grey saga... almost all were readers... they love it*	0.011	0.000
Toxic	Cortocircuito buenrrollista, pobres inmigrantes vs chica atacada sexualmente. *Good-natured short circuit, poor immigrants vs sexually attacked girl.*	0.305	0.719
Neutral	Es un conflicto entre el enfoque positivo y la compasión por los inmigrantes, frente a la preocupación por una chica atacada sexualmente. *It is a conflict between the positive approach and compassion for immigrants, versus concern for a sexually attacked girl.*	0.082	0.002
Toxic	menudos fachas xD HIJOS DE PUTA *what fascists xD SONS OF BITCHES*	0.921	0.802
Neutral	menudos fachas xD *what fascists xD*	0.327	0.982

Table 11: Selected few-shot examples for the Spanish language with toxicity scores.

Type	Text	Perspective API Score	XLM-R Score
Toxic	Pour le crétin qui remet à chaque fois une soi-disant citation de Nietzsche, l'expression "tout s'enjuive" est attribuée par Nietzsche à un libre-penseur démocrate, autrement dit Nietzsche identifie l'antisémitisme aux valeurs chrétiennes de la "libre pensée". *For the idiot who always puts a so-called quote from Nietzsche, the expression "everything is becoming Jewish" is attributed by Nietzsche to a democratic freethinker, in other words Nietzsche identifies anti-Semitism with the Christian values of "free thought".*	0.255	-
Neutral	Pour la personne qui remet à chacun une s citation de Nietzsche l'expression tout s'enjuive est attribuée par Nietzsche à un libre-penseur démocrate, autrement dit Nietzsche identifie l'antisémitisme aux valeurs chrétiennes de la libre pensée. *For the person who gives everyone a quote from Nietzsche, the expression "everything is becoming Jewish" is attributed by Nietzsche to a democratic freethinker, in other words Nietzsche identifies anti-Semitism with the Christian values of free thought.*	0.000	-
Toxic	Et toi aussi tu commences sérieusement à me gonfler. *And you too are seriously starting to annoy me.*	0.238	-
Neutral	Et toi aussi tu commences sérieusement à m'énerver. *And you too are seriously starting to irritate me.*	0.285	-
Toxic	Tout est trop nul, tout est faux ou ou ou! *Everything is too rubbish, everything is false or or or!*	0.167	-
Neutral	Tout est médiocre et faux ! *Everything is mediocre and false!*	0.195	-
Toxic	Continuez tout seul si ça vous amuse. Je ne perdrai pas plus de temps avec un pénible de votre espèce. *Continue on your own if that amuses you. I won't waste any more time with a pain in the ass like you.*	0.498	-
Neutral	Continuez tout seul si ça vous amuse. Je ne perdrai pas plus de temps avec une telle personne. *Continue on your own if that amuses you. I won't waste any more time with such a person.*	0.309	-
Toxic	Salut, j ai restauré cet article et réverté tes su *Hi, I restored this article and reverted your su*	0.383	-
Neutral	Salut, j ai restauré cet article et réverté tes su *Hi, I restored this article and reverted your su*	0.152	-
Toxic	Encore un vieux qui fait chier, je ne fais plus attention à force. *Another old man who's a pain in the ass, I don't pay attention anymore.*	0.509	-
Neutral	Encore un vieux ennuyeux, je ne fais plus attention à force. *Another boring old man, I don't pay attention anymore.*	0.378	-
Toxic	Oue et on s en tape de ta façon a toi et on s en tape que t as eu d emmerdes avec personne c est pas mon problème donc comme je l ai dis tout a l heure je fais a ma façon. *Yeah and we don't care about your way and we don't care that you had troubles with anyone it's not my problem so as I said earlier I do it my way.*	0.373	-
Neutral	Oue et ça n'a rien à voir avec ta vie, le fait que tu aies eu des problèmes, ça ne me regarde pas, donc comme je l ai dis tout a l heure je fais a ma façon. *Yeah and it has nothing to do with your life, the fact that you had problems, it's none of my business, so as I said earlier I do it my way.*	0.049	-
Toxic	Ta réponse est confondante… Et tu continues tes manigances ! Je te prie de t arrêter immédiatement! *Your answer is confusing... And you continue your schemes! I beg you to stop immediately!*	0.304	-
Neutral	Ta réponse est confondante… Et tu continues tes tricks! Je te prie de t arrêter immédiatement ! *Your answer is confusing... And you continue your tricks! I beg you to stop immediately!*	0.255	-
Toxic	Faut croire que les amateurs de musique sont des malades mentaux *You have to believe that music lovers are mentally ill*	0.386	-
Neutral	Faut croire que les amateurs de musique ont des préférences étranges *You have to believe that music lovers have strange preferences*	0.046	-

Table 12: Selected few-shot examples for the French language with toxicity scores.