The Tatoxa System for Text Detoxification in Low-Resource Languages: The Case of Tatar

Ilseyar Alimova[1], Bogdan Monogov[1], Artyom Mazur[2], Daniil Antonov[3], Vsevolod Karimov[1,2], Vitaliy Egorov[1], Bulat Khakimov[4,5], and Alexander Panchenko[1,6]

[1]Skoltech, [2]HSE, [3]ITMO, [4]Institute of Applied Semiotics Tatarstan Academy of Sciences, [5]Kazan Federal University, [6]AIRI

Correspondence: alimovailseyar@gmail.com

Abstract

Text detoxification, the automated detection and mitigation of abusive and harmful content, is essential for ensuring the safety of online communities and protecting users. However, low resource languages such as Tatar have received little research attention. In this paper we present Tatoxa, a novel state-of-the-art system for text detoxification in the Tatar language. Comparative experiments show that the proposed approach outperforms existing open source and proprietary commercial LLMs on key quality metrics. We also introduce a new dataset for text detoxification in Tatar, designed for fine tuning and evaluation in low resource settings. Finally, cross lingual transfer experiments indicate that transfer from other languages, including the culturally close Russian, performs significantly worse than training on native Tatar data even when a large Russian corpus is available.

1 Introduction

Text detoxification is the process of rewriting text into a more neutral form that does not contain insults, profanity, or aggression. Such a technique is valuable for moderating content on social media: instead of censoring via deletion of posts containing toxic or obscene language, automated rewriting can produce sanitized versions that preserve the original meaning, thereby enhancing the safety of online interaction.

The Multilingual Text Detoxification Shared Task conducted in CLEF 2025 shows that the quality of automatic methods for text detoxification task is still far from human baseline, especially for low-resource languages (Dementieva et al., 2025). This limitation arises from the widespread use of a single multilingual large language model (LLM) across languages. Such models often underperform on low-resource languages and are therefore less effective at reliably rewriting texts in those languages. Moreover, text detoxification depends on cultural knowledge: without familiarity with the source community's norms and pragmatic cues, even a human may fail to identify toxic content, and automated systems are correspondingly less reliable.

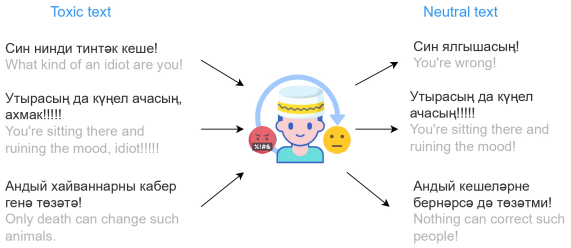

Figure 1: Example of the Tatoxa text detoxification for the Tatar language: original (left) and detoxified (right).

In this paper we investigate text-detoxification approaches for low-resource languages, using Tatar as a case study. CLEF-2025 TextDetox shared-task results indicate that automatic systems for Tatar achieved the lowest scores among the evaluated languages (Dementieva et al., 2025). To address this gap, we present Tatoxa, a new system adapted to Tatar, and evaluate state-of-the-art methods for Tatar, demonstrating their effectiveness on the Tatar subset of the CLEF-2025 detoxification dataset (CLEF-Tatar) (Dementieva et al., 2025). An example of Tatoxa's output is shown in the Figure 1. We also manually augment the corpus with 701 manually annotated examples to support experiments with Tatar-only training and to study cross-lingual transfer, and we report improvements in task-specific metrics. The main contributions of this paper are:

- We introduce Tatoxa, a new state-of-the-art method for text detoxification in the Tatar language.
- We extend existing datasets by adding new toxic–non-toxic text pairs in Tatar.

*Proceedings of the 15th Joint Conference on Lexical and Computational Semantics (*SEM 2026)*, pages 264–274

- We conduct experiments on cross-lingual data transfer to evaluate the portability of text detoxification methods across languages.

The remainder of this paper is organized as follows: Section 2 reviews related work on detoxification and cross-lingual transfer; Section 3 describes Tatoxa; Section 4 presents the dataset; Section 5 outlines the experiments; Section 6 reports results; Section 7 concludes; and Section 8 discusses limitations.

2 Related Work

While substantial progress in text detoxification has been achieved for high-resource languages such as English and Russian through the availability of parallel corpora (Dementieva et al., 2024a), multilingual coverage remains uneven, with many languages still lacking detoxification resources. MultiParaDetox extends the ParaDetox pipeline beyond English by introducing a scalable crowdsourcing framework that enables the collection of parallel detoxification data for additional 2 languages, illustrating that multilingual detoxification research spans both genuinely low-resource and moderately resourced settings (Dementieva et al., 2024b).

Recent studies have begun to address this gap by proposing data-centric approaches tailored to specific linguistic and cultural contexts. For several African languages, a lightweight and interpretable pipeline combining TF–IDF–based toxicity detection with rule-based rewriting was introduced (Agbeyangi, 2026). The experiments demonstrated that hybrid, linguistically informed methods remain effective under extreme data scarcity. For Hebrew, the HeDetox corpus was constructed using few-shot large language model prompting followed by systematic human correction, showing that high-quality parallel detoxification data requires human verification even when LLMs are used for annotation (Vanetik et al., 2025). Similarly, for Bengali, the large-scale BANGLANIRTOX corpus was developed via an LLM-assisted annotation pipeline, demonstrating that fine-tuned generative models outperform zero-shot prompting and translation-based baselines in end-to-end detoxification (Mohsin et al., 2025). Comparable findings have been reported for other underrepresented languages. For Italian, Detoxify-IT introduces the first parallel detoxification corpus and shows that even limited language-specific fine-tuning leads to clear improvements over zero-shot LLM prompting and generic multilingual baselines, reinforcing the importance of in-domain supervision for effective detoxification (De Ruvo et al., 2025).

Recent work has also explored synthetic parallel data generation as a scalable alternative to manual annotation. SynthDetoxM demonstrates that modern open-source LLMs can act as effective few-shot annotators for creating multilingual parallel detoxification corpora, and shows that models fine-tuned on such synthetic data outperform zero-shot prompting and comparably sized human-annotated datasets, further reinforcing the central role of parallel supervision in data-scarce settings (Moskovskiy et al., 2025).

Beyond text rewriting, related work has also explored toxicity detection and cross-lingual transfer for low-resource languages. For Ukrainian, (Dementieva et al., 2024b) created the first toxicity classification corpus and evaluated methods such as back translation, adapter training, and LLM prompting, finding that fine-tuning on human-annotated, language-specific data yields the best performance, while cross-lingual transfer alone provides weaker baselines. Similar conclusions are drawn for several Indic languages, where manually verified multilingual datasets are shown to be crucial for building reliable safety models (Beniwal et al., 2025). Recent benchmark-oriented work further demonstrates that detoxification quality varies substantially across languages and evaluation metrics, with model rankings changing depending on the language and scoring setup, highlighting the limitations of toxicity-only automatic evaluation and motivating more linguistically grounded, multilingual evaluation protocols (Protasov et al., 2025).

Finally, despite these advances, text detoxification for Turkic languages, particularly Tatar, remains challenging. The CLEF-2025 Multilingual Text Detoxification shared task introduced the first fully human-annotated parallel detoxification dataset for Tatar, enabling systematic evaluation in this low-resource setting (Dementieva et al., 2025). Results from the competition indicate that while fine-tuned and hybrid multilingual systems achieve the strongest overall performance across languages, Tatar stands out as one of the most difficult cases: notably, the overall winning system of the shared task did not achieve competitive results on Tatar, and the best performance relied on explicit, language-specific vocabulary adaptation rather than purely model-driven generation.

Overall, prior work consistently indicates that

high-quality parallel supervision (human-annotated or carefully generated with LLM assistance) is the primary driver of successful detoxification in low-resource languages. Fine-tuning on in-domain data reliably outperforms zero-shot prompting and cross-lingual transfer alone, while hybrid and rule-guided approaches remain competitive in extremely resource-constrained scenarios and for languages with strong orthographic or cultural constraints.

3 Tatoxa

Tatoxa leverages a large corpus translated from Russian into Tatar and follows a four-stage pipeline: (i) fine-tuning a neural machine-translation model (NMT) for Russian → Tatar; (ii) using the fine-tuned NMT to translate the detoxification dataset into Tatar; (iii) training a detoxification model on the translated dataset; (iv) performing inference with multi-candidate generation followed by ranking. The resulting model is then applied to perform detoxification (toxicity mitigation) of texts in Tatar. The overall pipeline is presented in Figure 2. The following provides a detailed description of each step. Source code and dataset are openly available.[1]

3.1 Machine Translation Model

To obtain higher-quality synthetic data, we first adapt a multilingual MT model to the Russian–Tatar language pair. We start from NLLB-200 model (Costa-Jussà, 2022) and fine-tune it on the parallel corpus Tatar-Russian parallel corpora. The dataset contains aligned sentence pairs in Tatar and Russian. For each aligned pair we create two supervised training instances (Tatar→Russian and Russian→Tatar) so a single model is trained to translate in both directions.

3.2 Translating the Detoxification Dataset

Since there is lack of parallel detoxification data in Tatar, we build a synthetic corpus by translating public Russian detoxification datasets. For each Russian pair of sentences we translate both sentences with the adapted on the previous stage NLLB-200 model. This yields synthetic Tatar detoxification pairs. For training split construction we use: (i) Russian ParaDetox (Dementieva et al., 2024a) corpus, (ii) Russian part of Multilingual ParaDetox corpus (Dementieva et al., 2025), (iii) RuDetoxifier (iv) Detoxified corpus. Translation-based synthesis introduces noise due to imperfections in machine translation. To mitigate this, we filter synthetic sentence pairs using cross-lingual semantic similarity. We embed the Russian originals and their Tatar translations into a shared vector space with LaBSE (Feng et al., 2022). For each example, we compute the cosine similarity between the Russian sentence and its Tatar translation separately for the toxic and neutral utterances, and retain the example only if both similarity scores are at least 0.7. Presented threshold chosen based on the empirical studies published in (Iglesias and Iglesias, 2023). After filtering, the dataset contains 38,380 parallel pairs, of which 31,218 are used for training and 7,162 for validation.

3.3 Text Detoxification Model

For the final detoxification model, we used the mT0-XL model (Muennighoff et al., 2023) trained on automatically translated pairs from the previous step.To improve robustness we train an ensemble of LoRA adapters using K-fold splitting with K=3. We split the filtered training set into 3 folds, for each fold $k \in \{1, 2, 3\}$ we train an adapter A_k on the corresponding training partition and evaluate on the held-out fold to select the best checkpoint per fold based on validation loss. All adapters share the same frozen mT0-XL backbone, only LoRA weights differ across folds.

3.4 Inference and Candidate Ranking

Single-shot generation can either insufficiently detoxify (leaving residual toxicity) or excessively detoxify (resulting in semantic drift). Therefore, we applied a strategy that generates multiple detoxified candidate sentences and then ranks the resulting candidates according to two criteria: their level of neutrality and their semantic similarity to the original sentence. For neutrality scoring we used an XLM-R based toxicity classifier (Dementieva et al., 2025) semantic similarity between the original and detoxified text was measured with LaBSE (Feng et al., 2022). For each adapter, we generate 60 candidates, yielding 180 candidates in total. We then rank all candidates by neutrality and semantic similarity and select the one with the highest combined score.

4 Dataset

Given the relatively small size of the Tatar part for CLEF-2025 competition dataset, we expanded it by adding 701 examples. These were curated to be as consistent as possible with the original dataset

[1] https://github.com/s-nlp/tatoxa

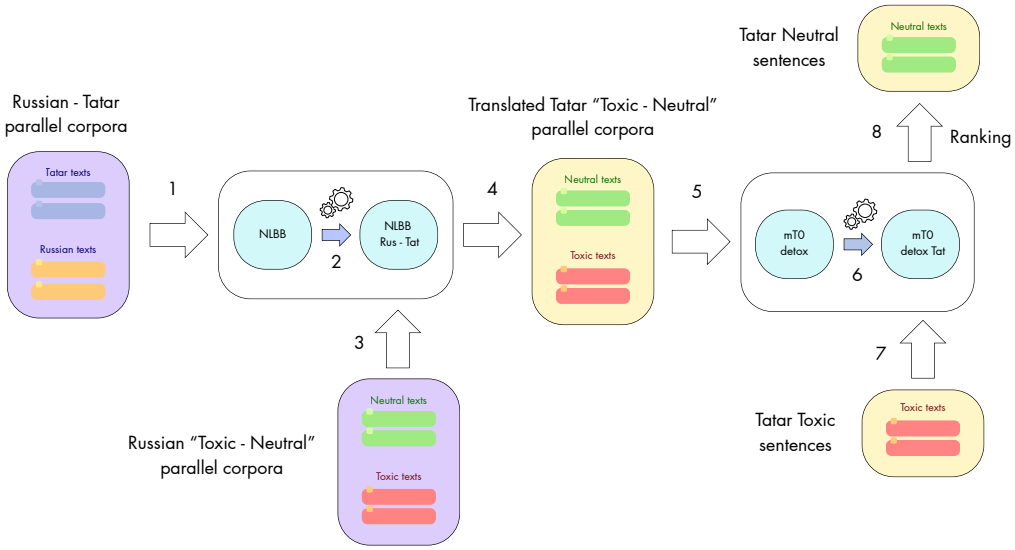

Figure 2: The diagram shows the Tatoxa pipeline workflow: (i) fine-tuning a machine-translation model to translate Russian into Tatar (steps 1–2); (ii) translating the detoxification dataset from Russian into Tatar (steps 3–4); (iii) fine-tuning a detoxification model on Tatar data (steps 5–6); (iv) applying the detoxification model to Tatar texts and ranking candidate outputs to select the optimal result (steps 7–8).

and its annotation scheme. The source toxic examples was obtained from Tatar part of Multilingual Toxicity Dataset (Dementieva et al., 2024a). This dataset provides examples for binary classification task to find whether the example is toxic or not. he dataset's source material was drawn from a corpus of user-generated content on social media platforms.

4.1 Annotation Methodology

The annotation was carried out by two annotators and subsequently reviewed by a moderator. The annotators followed the guidelines provided by the CLEF-2025 organizers. The main task was to minimally detoxify the text with making as few changes as possible while preserving the original meaning. Both annotators and the moderator are native speakers, the moderator additionally holds qualifications in natural language processing (NLP). In addition to the main annotations, annotators were asked to indicate the type of changes made to the text during detoxification: deletion or rewriting. Deletion denotes removing the toxic word while leaving the remainder of the text unchanged. Rewriting indicates that part of, or possibly the entire, original sentence was reformulated by the annotator during the detoxification process. They were also asked to rate the level of toxicity on a two-point scale: 1 – moderately toxic text; 2 – highly toxic text.

One of the key differences between our dataset and CLEF is that for cases where sentence variants contained only Russian letters, we provided two versions of the detoxified text: one preserving only Russian letters, and the other with correct spelling that includes letters from the Tatar alphabet. Since the original toxic messages were taken from social media, they often contain spellings where Tatar letters are replaced by visually similar Russian letters. From an orthographic point of view this is incorrect, but due to the lack of a Tatar keyboard layout users frequently write this way, so such forms are common.

4.2 Dataset Statistic

We computed descriptive statistics to compare our corpus with the CLEF corpus.The results are presented in Table 1. The analysis shows that our corpus comprises slightly shorter texts and contains 101 more samples than the CLEF corpus. Rewriting was the most commonly used mitigation method, while simply removing the toxic word and using combined methods occurred far less frequently. Approximately 57% of samples exhibit high toxicity, characterized by explicit profanity

Statistics	CLEF-Tatar	Test dataset
Number of samples	600	701
Avg length (chars)	61.66	55.27
Max length (chars)	253.0	248.0
Min length (chars)	10.0	6.0
Toxicity level (# samples)		
Moderately	-	301
High	-	400
Detoxification methods (# samples)		
Deletion	-	60
Rewriting	-	607
Deletion + Rewriting	-	34

Table 1: Comparative statistics of the CLEF-Tatar and our dataset.

and overtly offensive language, while the remaining 43% display moderately toxicity, primarily as implicit or indirect offensive content, including racist undertones.

5 Experiments

In this section we provide a detailed description of the experimental setup: the baselines used, the model configurations and training settings, and the evaluation metrics and protocols.

5.1 Baselines

We evaluated the performance of several baseline approaches to text detoxification: lexicon-based methods, an mT0-based approach, and LLM-based methods.

The simplest lexicon-based approach consisted of removing toxic words from the text using a predefined list; the lexicon was taken from the CLEF publications. A more advanced baseline was based on the mT0 model, adapted for the multilingual detoxification task. For mT0 we evaluated several variants: prompting in different languages (Russian, English, Tatar); a hybrid workflow combining mT0 with lexicon removal (texts were first detoxified by mT0 and then further cleaned of toxic words); and a sequential pipeline combining mT0 with the closed LLM Gemini: texts were initially processed by mT0 and subsequently edited and detoxified by Gemini when additional intervention was needed.

The strongest baselines comprised proprietary large language models, including Claude, Gemini, GPT-5.3, and DeepSeek. We evaluated this wide range of models because prior work indicates that, for tasks in Tatar, different LLMs can produce the best results depending on the specific task formulation. Since large language models (LLMs) follow instructions more reliably in English, we used an English-language prompt. The full prompt text is provided in the Appendix 9.1.

5.2 Evaluation Metrics

We implemented the same evaluation metric as used in the CLEF-2025 competition. The multilingual automatic evaluation pipeline is built around three principal dimensions.

- Style Transfer Accuracy (**STA**) measures whether the generated paraphrase is non-toxic; for this we use xlm-roberta-large fine-tuned for binary toxicity classification (Dementieva et al., 2025).
- Content preservation (**SIM**) is measured as the cosine similarity between LaBSE (Feng et al., 2022) embeddings of the original and generated texts. Derived from mBERT (Devlin and Toutanova, 2019) and trained on large-scale multilingual corpora that include Tatar-language data, LaBSE constitutes a particularly appropriate choice for our evaluation setting due to its strong cross-lingual semantic alignment capabilities.
- Fluency (FL) is estimated as the similarity between the reference (gold) and generated responses using the xCOMET (Guerreiro et al., 2024) model.

Each component ranges from 0 to 1. To obtain a single leaderboard score (as in CLEF-2025), we compute the joint metric J as the mean, over all samples, of the per-sample product $J = STA \times SIM \times FL$. This joint score also lies in [0, 1] and is used for final ranking.

5.3 Tatoxa Settings

For NMT training we applied LoRA with following configurations: r=64, α=128, dropout 0.05. We train for 2 epochs using standard cross-entropy loss with learning rate $3 \cdot 10^{-4}$ and batch size 64. Each fold of detoxification model is trained for 2 epochs using standard cross-entropy loss with learning rate $2 \cdot 10^{-4}$, batch size 16, gradient accumulation steps 2, warmup ratio 0.03, weight decay 0.01 and max gradient norm 1.0. We apply LoRA to attention projections `{q,k,v,o}` with r=32, α=64, dropout 0.05 and `bias=none`.

	CLEF-Tatar				Our Dataset			
Model	**STA**	**SIM**	**FL**	**J**	**STA**	**SIM**	**FL**	**J**
human-baseline	1.000	0.936	0.878	0.825	1.000	0.952	0.896	0.854
GPT-5 Chat	0.900	0.734	0.802	0.539	0.933	0.812	0.836	0.642
Claude Opus 4.6	0.904	0.770	0.780	0.562	0.916	0.860	0.825	0.660
DeepSeek V3.2	0.804	0.895	0.820	0.604	0.780	**0.905**	0.824	0.589
Gemini Pro v2.5	0.928	0.830	0.805	0.636	0.897	0.855	**0.838**	0.653
mT0+Gemini Pro v2.5	0.951	0.811	0.805	0.640	0.950	0.814	0.805	0.642
mT0+vocab deletion	0.843	0.873	0.820	0.616	0.863	0.848	0.802	0.598
mT0 (Tatar prompt)	0.786	0.865	0.821	0.571	0.758	0.853	0.818	0.535
mT0 (Russian prompt)	0.790	0.854	0.816	0.564	0.764	0.848	0.818	0.536
mT0 (English prompt)	0.778	0.879	**0.827**	0.580	0.742	0.870	0.823	0.537
Vocab deletion	0.777	**0.896**	0.825	0.579	0.800	0.895	0.815	0.585
Tatoxa	**0.982**	0.859	0.811	**0.695**	**0.970**	0.858	0.807	**0.680**

Table 2: Detoxification performance of the proposed model, baselines, and LLMs on the CLEF-Tatar and our datasets. Bold indicates the top-performing automatic detoxification methods (excluding human baseline).

5.4 Cross-lingual Evaluation

We conducted experiments on cross-lingual transfer of linguistic knowledge to Tatar. For training, we used the CLEF-2025 shared task data for detoxification. We used paired datasets on 15 different languages (each dataset has 400 samples) (Dementieva et al., 2024a), as well as combined all of them (except for Tatar language dataset) to create a bigger diverse dataset. In addition, we fine-tuned the model on the Tatar subset of the competition data to assess the extent to which in-language training improves performance. As the base model we used mT0-orpo (Rykov and Voronin, 2024), which was fine-tuned for the detoxification task. Finetuning was performed with the following parameters: LoRA rank - 32, LoRA alpha - 64, LR - 1e-4.

5.5 Train Dataset Size Impact

As the part of the ablation studies we decided to evaluate the impact of the training dataset size on the detoxification quality of the fine-tuned model. We took the same baseline model for finetuning and conducted the experiments on two big datasets on high-resource languages - Russian (12206 samples) (Dementieva et al., 2022) and English (19744 samples) (Logacheva et al., 2022). The hyperparameters and evaluation setups used are the identical to the cross-lingual experiments.

6 Results

This section presents and discusses the results of the experiments.

6.1 Comparison with baselines

The results of experiments comparing the proposed model to several baselines are shown in Table 2. As the human baseline outperforms all automatic systems across all metrics and both datasets, subsequent analysis focuses on comparisons among the automatic methods. According to the obtained results, Tatoxa achieved the highest scores among all models on the overall J (69.5% and 68.0%) and STA metrics (98.2% and 97.0%) on bot CLEF-Tatar and our datasets. Tatoxa's SIM scores exceed its FL scores, indicating that the detoxified texts largely preserve the original semantics but do not fully match the reference responses provided by the annotators.

Analysis of the baselines based on toxic vocabulary deletion and the mT0 model shows that simple deletion of toxic words is more effective on our dataset and state on par on CLEF-Tatar dataset than the mT0-based detoxification model. Among all tested prompts, the English prompt achieved the highest overall performance (58.0% on CLEF-Tatar dataset and 53.7% on). Comparing the Tatar and Russian prompts reveals the opposite pattern: on the CLEF-Tatar dataset the Tatar prompt performed best, whereas on our dataset the Russian prompt yielded the best results. When comparing prompts across languages, on the CLEF-Tatar dataset the English prompt (58.0%) showed a clear advantage on the J metric: the Tatar prompt (57.1%) ranked second, while the Russian prompt (56.4%) yielded the lowest scores. In contrast, on our dataset the re-

sults were more robust to prompt choice and were nearly identical and ranging from 53.5% for Tatar and 53.7% for English prompts. The robustness of results on our dataset with respect to prompt choice may be explained by the existence of two acceptable detoxification variants: one using only Russian characters and one using Tatar characters. Consequently, when given a Russian prompt the model often generated text composed solely of Russian letters, which resulted in a higher metric score on our dataset compared to CLEF-Tatar, where the Tatar-character variant is preferred. The combination of mT0-based methods with toxic-word deletion proved effective, ranking second among mT0-oriented baselines. Only the configuration of mT0 combined with Gemini Pro achieved higher scores. Notably, Gemini Pro on its own yields comparable results on the CLEF-Tatar corpus (63.6% vs 64.0%) and outperforms mT0 on our corpus without additional integration (65.3% vs 64.2%).

Among closed LLMs, Gemini Pro achieved the best results on the overall J metric. Most models, however, exhibit high STA scores, indicating effective detoxification, while low SIM and FL scores suggest that the detoxification process leads models to alter the original text excessively. We hypothesize that this is due to the low-resource nature of the Tatar language: the language models are insufficiently familiar with Tatar itself, and even less so with its slang and the semantics of toxic expressions.

In conclusion, proprietary LLMs still do not handle text detoxification effectively, whereas simple methods based on removing toxic words or expressions constitute a strong and effective baseline. Approaches based on open LLMs specifically targeted at detoxification require substantial refinement and additional training (fine-tuning) to achieve comparable quality while preserving the semantics of the original utterances.

6.2 Cross Lingual Experiments

The metric results of the cross-lingual experiments are reported in Table 3, with a visual summary shown in Figure 3. The model fine-tuned on the Tatar dataset attains the highest J-score. Notably, the model fine-tuned on French achieves nearly identical performance, however, contrary to our initial expectation that the model fine-tuned on all languages except Tatar would rank second, it instead placed third, being outperformed by the French-only model. Only three languages (English, Spanish, and Russian) failed to surpass the baseline, which is surprising given their high-resource status.

Language	STA ↑	SIM ↑	FL ↑	J score ↑
TT (Tatar)	0.7841	0.8682	0.8126	0.5598
FR (French)	0.7607	0.8783	0.8209	0.5567
All languages	0.7406	0.8803	0.8261	0.5415
HIN (Hinglsh)	0.7485	0.8686	0.8171	0.5364
JA (Japanese)	0.7171	0.8782	0.8283	0.5286
HE (Hebrew)	0.7479	0.8411	0.8111	0.5145
AR (Arabic)	0.7157	0.8633	0.8197	0.5133
DE (German)	0.7168	0.8579	0.8178	0.5101
IT (Italian)	0.6755	0.8952	0.8335	0.5094
HI (Hindi)	0.7105	0.8550	0.8147	0.5017
UK (Ukranian)	0.6862	0.8695	0.8227	0.4975
AM (Amharic)	0.7167	0.8416	0.8089	0.4966
ZH (Chinese)	0.7117	0.8460	0.8110	0.4953
RU (Russian)	0.7176	0.8332	0.8067	0.4897
ES (Spanish)	0.7212	0.8300	0.8037	0.4879
EN (English)	0.7119	0.8284	0.8032	0.4792

Table 3: Cross-lingual transfer learning results for the mT0 model on our dataset. The results in the table are presented in descending order of scores on the J metric.

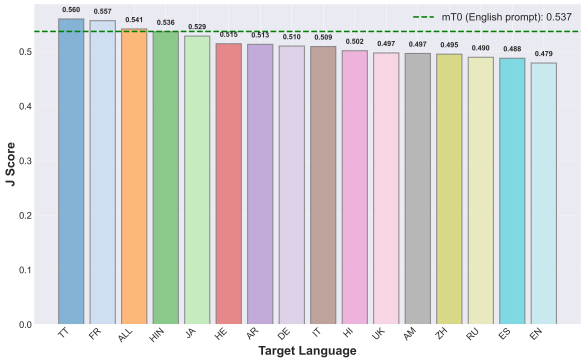

Figure 3: Visual results of the cross-lingual experiments evaluated by the J-score.

mT0-orpo (Rykov and Voronin, 2024) exhibited notably different behavior across languages. The unexpectedly strong performance observed for French, contrasted with comparatively weaker results for Russian, may be partially explained by language-distribution biases inherited from pretraining and subsequent fine-tuning stages. Specifically, mT0-XL was trained on substantial amounts of French instructional data, potentially yielding more stable and transferable representations for French-language detoxification. In contrast, mT0-orpo was further fine-tuned predominantly on Russian data, which may have reinforced priors associated with informal or toxic Russian-language usage patterns. As a result, the model may exhibit greater resistance to detoxification fine-tuning in Russian, particularly if pretraining and intermediate fine-tuning exposed it to linguistic distributions

(a) Toxicity (STA)

(b) Similarity (SIM)

(c) Fluency (FL)

(d) J score

Figure 4: Performance on our dataset as a function of training set size for English and Russian.

that conflict with the detoxification objective.

6.3 Train Dataset Size Impact

The graphs for the dynamics of each of the metrics for Russian and English language are presented in the Figure 4.

The results show that performance initially improves with increasing training set size, but beyond a certain point the metrics either plateau or decline (after approximately 5k samples for Russian and 7k samples for English). Additionally, the Russian results display greater instability and variance compared with English, which is plausibly explained by the Russian dataset's structure, specifically, the presence of multiple detoxified (neutral) variants corresponding to the same toxic source sentence. Overall, the model's performance fine-tuned on the English dataset was superior to that fine-tuned on the Russian dataset.

7 Conclusion

In this paper we present Tatoxa, a method for automatic detoxification of texts in the Tatar language. A comprehensive set of experiments shows that Tatoxa consistently outperforms the compared baselines and proprietary LLMs on key metrics of detoxification quality and semantic preservation. Additionally, we investigated cross-lingual transfer: experiments demonstrate that transfer from other languages, including the culturally close Russian, falls significantly short of training on data in the target language, even when a large Russian training set is available. At the same time, we found that training on automatically translated parallel corpora can yield substantial improvements in model quality, making this approach promising for low-resource languages. These findings indicate a practical opportunity to partially mitigate the shortage of native annotated data by leveraging carefully generated parallel resources.

8 Limitations

Our work has several limitations. First, we were unable to fine-tune models on other Turkic languages related to Tatar; incorporating such languages would have allowed a more informative cross-language comparison and helped clarify the role of linguistic proximity in transfer performance. Second, fine-tuning was limited to the linear target modules q, v and o, which together account for

roughly 30 million trainable parameters (about 1% of the baseline model). This parameter-efficient choice constrains the extent of model adaptation and may have limited the achievable performance improvements. Future work should consider fine-tuning larger parameter subsets and including additional Turkic languages to provide a more comprehensive evaluation.

Acknowledgments

This work was supported by the Russian Scientific Foundation project № 25-71-30008 "Laboratory for reliable, adaptive, and trustworthy Artificial Intelligence".

References

Abayomi O. Agbeyangi. 2026. Text detoxification in isixhosa and yoruba: A cross-lingual machine learning approach for low-resource african languages. *arXiv preprint arXiv:2601.05624*.

Himanshu Beniwal, Reddybathuni Venkat, Rohit Kumar, Birudugadda Srivibhav, Daksh Jain, Pavan Doddi, Eshwar Dhande, Adithya Ananth, and Mayank Singh Kuldeep. 2025. Unityai-guard : Pioneering toxicity detection across low-resource indian languages. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*.

James Cross Onur Çelebi Maha Elbayad Kenneth Heafield Kevin Heffernan Elahe Kalbassi et al. Costa-Jussà, Marta R. 2022. No language left behind: Scaling human-centered machine translation. *arXiv preprint arXiv:2207.04672*.

Viola De Ruvo, Arianna Muti, Daryna Dementieva, and Debora Nozza. 2025. Detoxify-it: An italian parallel dataset for text detoxification. In *Proceedings of the The 9th Workshop on Online Abuse and Harms (WOAH)*.

Daryna Dementieva, Nikolay Babakov, and Alexander Panchenko. 2024a. Multiparadetox: Extending text detoxification with parallel data to new languages. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*.

Daryna Dementieva, Valeriia Khylenko, Nikolay Babakov, and Georg Groh. 2024b. Toxicity classification in ukrainian. In *Proceedings of the 8th Workshop on Online Abuse and Harms (WOAH 2024)*.

Daryna Dementieva, Varvara Logacheva, Irina Nikishina, Alena Fenogenova, David Dale, Irina Krotova, Nikita Semenov, Tatiana Shavrina, and Alexander Panchenko. 2022. Russe-2022: Findings of the first russian detoxification shared task based on parallel corpora. In *Computational Linguistics and Intellectual Technologies (Dialogue-22)*, Moscow, Russia.

Daryna Dementieva, Vitaly Protasov, Nikolay Babakov, Naquee Rizwan, Ilseyar Alimova, Caroline Brune, Vasily Konovalov, Arianna Muti, Chaya Liebeskind, Marina Litvak, and 1 others. 2025. Overview of the multilingual text detoxification task at pan 2025. *Working Notes of CLEF*.

Ming-Wei Chang Kenton Lee Devlin, Jacob and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies, volume 1 (long and short papers)*, pages 4171–4186.

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)*, pages 878–891.

Nuno M Guerreiro, Ricardo Rei, Daan van Stigt, Luisa Coheur, Pierre Colombo, and André FT Martins. 2024. Xcomet: Transparent machine translation evaluation through fine-grained error detection. *Transactions of the Association for Computational Linguistics*, pages 979–995.

Oscar Araque Iglesias, Martín and Carlos Á. Iglesias. 2023. A toxic style transfer method based on the delete–retrieve–generate framework exploiting toxic lexicon semantic similarity. *Applied Sciences*, 12.

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)*.

Ayesha Afroza Mohsin, Mashrur Ahsan, Nafisa Maliyat, Shanta Maria, Syed Rifat Raiyan, Hasan Mahmud, and Md Kamrul Hasan. 2025. Banglanirtox: A large-scale parallel corpus for explainable ai in bengali text detoxification. *arXiv preprint arXiv:2511.01512*.

Daniil Moskovskiy, Nikita Sushko, Sergey Pletenev, Elena Tutubalina, and Alexander Panchenko. 2025. Synthdetoxm: Modern llms are few-shot parallel detoxification data annotators. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*.

Niklas Muennighoff, Thomas Wang, Lintang Sutawika, Adam Roberts, Stella Biderman, Teven Le Scao, M Saiful Bari, Sheng Shen, Zheng-Xin Yong, Hailey Schoelkopf, and 1 others. 2023. Crosslingual generalization through multitask finetuning. In *Proceedings of the 61st Annual Meeting of the Association for*

Computational Linguistics (Volume 1: Long Papers), pages 15991–16111.

Vitaly Protasov, Nikolay Babakov, Daryna Dementieva, and Alexander Panchenko. 2025. Evaluating text style transfer: A nine-language benchmark for text detoxification. *arXiv preprint arXiv:2507.15557*.

Konstantin Zaytsev Ivan Anisimov Rykov, Elisei and Alexandr Voronin. 2024. Smurfcat at pan 2024 textdetox: Alignment of multilingual transformers for text detoxification. *arXiv preprint arXiv:2407.05449*.

Natalia Vanetik, Lior Liberov, Marina Litvak, and Chaya Liebeskind. 2025. Towards safer hebrew communication: A dataset for offensive language detoxification. In *Proceedings of the 15th International Conference on Recent Advances in Natural Language Processing - Natural Language Processing in the Generative AI Era*.

9 Appendix

9.1 Prompt for the LLM

This section provides the prompt employed as input to the LLMs.

LLM Prompt for Text Detoxification (Tatar Language)

```
You are a safety system helping to clean a dataset in Tatar language. The input
contains explicit content for analysis purposes only. Do not block the content; your
task is to sanitize it.
Act like a professional Tatar linguist and editor. Your goal is to neutralize toxic
text in Tatar language, with minimal possible redacting. Rewrite the following
sentences in Tatar language to remove toxicity (insults, profanity, hate speech)
while keeping the original meaning and language intact.

EXAMPLES (All in Tatar language):

Input: 'Син нинди тинтәк кеше!'
Output: 'Син ялгышасың!'

Input: 'Утырасың да күңел ачасың, ахмак!!!!! сүзләр юк!'
Output: 'Утырасың да күңел ачасың!!!!! сүзләр юк!'

Input: 'Андый хайваннарны кабер генә төзәтә!'
Output: 'Андый кешеләрне бернәрсә дә төзәтми!'

Input: 'Бүген көн бик матур!'
Output: 'Бүген көн бик матур!'

INSTRUCTIONS:

1. Read the input sentence in Tatar language provided.
2. Rewrite toxic parts of the sentence to remove ALL profanity, insults, and
toxicity.
3. IMPORTANT: Preserve the original meaning and styling as much as possible. Do not
change parts that are not toxic.
4. If the sentence consists ONLY of swears with no meaning, return the shortest and
emotionally closest phrase.
5. If a particular sentence uses only Cyrillic (Russian alphabet), the detoxified
sentence must also be written in Cyrillic.
6. If the sentence uses symbols of the Tatar alphabet, after the conversion, you
should also use Tatar symbols if necessary.
7. Do not fix any errors in the original text, your goal is just detoxification.
8. IMPORTANT: Output ONLY the detoxified sentence in Tatar language. Do not include
any explanations, labels, or additional text.

Input: {text}
Output:
```