nikita.sushko at TextDetox CLEF 2025: Exploring A Sage-T5-Like Approach For Text Detoxification

Notebook for the PAN Lab at CLEF 2025

Alexandr Voronin[1,*], Daniil Moskovsky[2,1] and Nikita Sushko[2,1,*]

[1]Skoltech, Russia
[2]AIRI, Russia

Abstract

This paper presents our submission to the Multilingual Text Detoxification task at PAN 2025. We explore the Sage-T5-like approach, by combining three training objectives: paraphrasing (seq2seq loss), token-level toxicity detection (classification loss), and semantic representation learning (contrastive loss). To address the challenge of limited annotated data across 15 languages, we adopt the synthetic data generation pipeline from SynthDetoxM and introduce a token-level annotation method using multilingual toxic lexicons. Our experiments on Russian, French, and Spanish demonstrate that combining classification and contrastive objectives significantly boosts detoxification performance, as measured by Style Transfer Accuracy (STA), Semantic Similarity (SIM), and their combined J-score, but fall short after expansion to more languages. Our resulting model outperforms 5 out of 7 baselines during automatic evaluation.

Keywords

Text style transfer, contrastive learning, encoder-decoder transformers, synthetic data generation

1. Introduction

The rapid spread of toxic content online has created a pressing demand for effective systems that can detoxify text in multiple languages. Despite notable achievements in developing monolingual detoxification systems, the multilingual landscape still poses a variable set of challenges. For instance, languages exhibit distinct grammatical structures, vocabulary, and cultural references, which can make it difficult to develop a one-size-fits-all approach. Furthermore, many languages lack sufficient labeled data, hindering the training of accurate detoxification models. To address these challenges, this paper explores the potential of cross-lingual transfer learning in enhancing multilingual text detoxification. By using shared linguistic patterns that exist across languages, our approach aims to reduce the need for large amounts of language-specific training data. This, in turn, enables us to develop more efficient and effective detoxification systems, particularly for languages with limited resources, while also preserving the original meaning and context of the text.

In this article, we introduce a novel multilingual detoxification framework based on the Sage-T5 architecture. The proposed model follows the approach from the Sage-T5 [1] and employs a multitask learning objective, combining seq2seq loss for paraphrase generation, classification loss for token-level toxicity detection, and contrastive loss for improved semantic representation learning. Furthermore, we reuse a methodology from SynthDetoxM [2] for the collection and annotation of datasets. Additionally, we created a pipeline for token toxicity markup which is crucial for training the model with classification loss.

CLEF 2025 Working Notes, 9 – 12 September 2025, Madrid, Spain
*Corresponding author.
Alexandr.Voronin@skoltech.ru (A. Voronin); Daniil.Moskovskiy@skol.tech (D. Moskovsky); Nikita.Sushko@skoltech.ru (N. Sushko)
0009-0006-9493-0437 (A. Voronin); 0009-0006-7943-4259 (D. Moskovsky); 0000-0003-2245-7354 (N. Sushko)

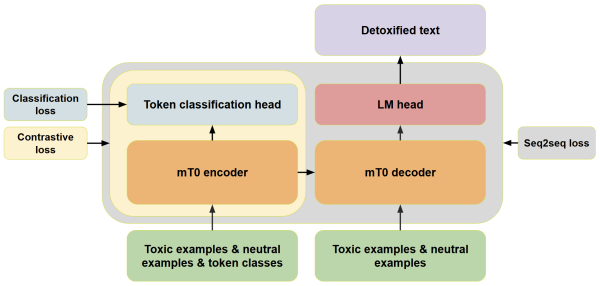

Figure 1: Model diagram with seq2seq loss for paraphrase generation, classification loss for toxic span detection and contrastive loss for better encoder representations.

2. Previous work

In 2024, Multilingual Text Detoxification task [3] was presented as one of the tracks of PAN Lab on CLEF conference. Participants were asked to create a text detoxification system with limited training data for 9 languages, which required the use of cross-lingual transfer and unsupervised methods. Top solutions included approaches with few-shot prompting of an uncensored version of the Llama3 [4] performed by SomethingAwful team and fine-tuning seq2seq models such as mT0 and mT5 on augmented datasets along with ORPO technique application performed by SmurfCat team [5].

The SAGE-T5 article [1] approaches the problem of spelling correction by utilizing three losses: seq2seq loss for training the corrector model, contrastive loss for the encoder of the encoder-decoder model to ensure close semantic match between the original and corrected sentences and token classification loss for the encoder of the encoder-decoder model to ensure higher accuracy of typo detection. This approach allowed to reach state-of-the-art results in spelling correction task across other models.

In the SynthDetoxM paper [2], the authors proposed a multistage approach for synthetic detoxification data generation, using pretrained decoder-only models and a large scale parallel synthetic dataset for training text detoxification models. Models, trained on this dataset, show better performance than models, trained on human labeled data.

We adapt the methodology from Sage-T5 paper to the more complex task of text detoxification and utilize SynthDetoxM methodology for generating synthetic data.

3. Data

TextDetox 2025 track [6] in PAN Lab at CLEF 2025 [7] contains 15 languages (9 from previous year track and 6 new) - English, Spanish, Italian, French, Chinese, Japanese, Hindi, Hinglish, Arabic, German, Russian, Ukrainian, Amharic, Hebrew and Tatar.

The track consists of two stages:

During the development stage, the organizers provided a training dataset, comprising 600 non-parallel examples for each of the 9 languages from previous year track, as well as 100 examples for each of the 6 newly introduced languages. The data was presented in a standardized format, consisting of three components: toxic text, neutral text, and language identification (lang).

For the test stage, organizers provided

- **MultiParaDetox**[1] — a dataset with 400 parallel samples for 9 languages;
- **Multilingual Toxicity Dataset**[2] — contains non-parallel toxic and neutral sentences: 2.01k samples for Hebrew, 4.36k for Hinglish, and 5k for every other language. For all languages except Hebrew the proportion between toxic and neutral sentences id equal, Hebrew data contains 60% neutral sentences and 40% toxic sentences;

[1] https://huggingface.co/datasets/textdetox/multilingual_paradetox
[2] https://huggingface.co/datasets/textdetox/multilingual_toxicity_dataset

- **Multilingual Toxicity Lexicon**[3] — includes toxic words and expressions for all 15 languages.

3.1. Data preprocessing

To leverage the classification loss in our model, we introduced an additional classification head on top of the encoder. The primary function of this classification head is to predict a toxicity label for each token, categorizing it as either toxic or non-toxic. This allows the model to learn a more nuanced representation of the input text, where each token is associated with a specific toxicity classification, enabling the model to better capture the toxic language.

We used Multilingual Toxicity Dataset to markup token toxicity. The toxicity markup was carried out in 3 stages. Firstly, the input data and toxic lexicon were tokenized using the target model's tokenizer. Then, we've created a function to align toxic expressions to toxic sentences. As a single toxic expression mostly consists of several tokens, we had to check that all of its tokens are present in toxic sentence. Finally, we applied this function to all available data in all languages and obtained token markup.

3.1.1. Synthetic Data Collection

Using the dataset of toxicity identification provided by the organizers[4], we collect a synthetic parallel detoxification dataset. In our collection pipeline, we follow the approach introduced in the SynthDetoxM [2].

In the context of this task, we utilize more novel and not only open-source models. Namely, we use Gemini 2.5 Flash[5], Qwen 3 235B[6] in non-reasoning mode, Llama 4 Maverick 400B[7], Mistral Saba[8] and DeepSeek Chat v3 0324[9].

Table 1
Number of selected samples per model.

Model	ar	de	en	es	fr	he	hi	hin	ja	ru	tt	uk	zh
deepseek-chat-v3-0324	608	460	602	750	422	1	1004	704	848	486	9	5	531
gemini-2.0-flash-001	383	327	53	65	1076	–	240	293	304	411	4	15	586
gemma-3-27b-it	471	402	306	764	74	–	166	61	305	566	–	2	144
llama-4-maverick	603	633	519	486	494	–	792	547	508	494	1	7	1055
mistral-saba	190	–	–	–	–	–	152	323	–	–	–	–	–
qwen3-235b-a22b	305	258	915	426	289	–	112	179	535	531	3	2	183

To source the non-parallel toxic sentences, we've used textdetox/multilingual_toxicity_dataset, provided by competition authors. The resulting dataset consisted of 33528 pairs of sentences on 15 languages. In Hebrew, Ukranian and Tatar languages the quality of detoxification was shown to be of a low quality, so only several examples were used in the final data mix. Distribution of the amount of selected sentences per model is shown in the Table 1.

3.2. Metrics

The evaluation metrics for the TextDetox 2025 track remained consistent with those used in the TextDetox 2024 track. Throughout our development process, three primary metrics were employed:

[3]https://huggingface.co/datasets/textdetox/multilingual_toxic_lexicon
[4]https://huggingface.co/datasets/textdetox/multilingual_toxicity_dataset
[5]https://cloud.google.com/vertex-ai/generative-ai/docs/models/gemini/2-5-flash
[6]https://qwenlm.github.io/blog/qwen3/
[7]https://www.cerebras.ai/press-release/maverick
[8]https://mistral.ai/news/mistral-saba
[9]https://huggingface.co/deepseek-ai/DeepSeek-V3-0324

Style Transfer Accuracy (STA), Semantic Similarity (SIM), and Fluency (FL). To obtain the final score, we combined these metrics by calculating their product, resulting in a unified J-score.

The STA metric, which evaluates the quality of style transfer, was computed using the textdetox/xlmr-large-toxicity-classifier-v2 model[10]. SIM indicates the similarity of toxic and detoxified versions of the same sentence using cosine distance between LABSE [8] embeddings of these sentences using sentence-transformers/LaBSE[11] model. FL measures similarity between detoxified sentences and human-written detoxified versions and is calculated with myyycroft/XCOMET-lite [9][12] model.

4. Methodology

4.1. Model selection

The primary objective was to fine-tune a Sage-T5-like model using a combination of three loss functions.

During training, two types of textual data were utilized: toxic texts with toxic span classification labels and pairs of toxic texts and their corresponding detoxified versions for paraphrase learning. The training process incorporates three loss functions. The training framework is presented in Figure 1.

For the experiments, the base version of bigscience/mt0-large [10][13] model. For the final training, s-nlp/mt0-xl-detox-orpo [5][14] was selected. This model was the winning model of the PAN 2024 detoxification contest.

4.2. Losses

Besides regular seq2seq loss, classification loss and contrastive loss are present during model training.

4.2.1. Seq2seq loss

To ensure that the model generates fluent and coherent detoxified sentences, we use a standard sequence-to-sequence (seq2seq) loss. It encourages the model to produce target tokens that match the reference detoxified output at each position, while ignoring padding.

Sequence-to-sequence cross-entropy loss with padding mask is defined as:

$$\mathcal{L}_{seq2seq} = -\frac{1}{\sum_{t=1}^{T} m_t} \sum_{t=1}^{T} m_t \cdot \log\left(p_{t,y_t}\right),$$

where T is the length of the target sequence, V is the size of the vocabulary, $m_t \in \{0, 1\}$ is the mask, $y_t \in \{1, \ldots, V\}$ is the ground truth token ID at timestep t and $\mathbf{p}_t \in \mathbb{R}^V$ is the predicted probability distribution over the vocabulary of the trained model at timestep t.

4.2.2. Classification loss

A classification head is added to the encoder part of the decoder model and then trained with a simple cross-entropy loss. This classification head is trained simultaneously with the whole model, ensuring that the encoder embeddings of toxic sentences contain information necessary to distinguish toxic and neutral tokens.

The binary cross-entropy loss with padding mask is defined as:

$$\mathcal{L}_{classification} = -\frac{1}{\sum_{i=1}^{N} m_i} \sum_{i=1}^{N} m_i \cdot \left[y_i \cdot \log(p_i) + (1 - y_i) \cdot \log(1 - p_i)\right],$$

[10] https://huggingface.co/textdetox/xlmr-large-toxicity-classifier-v2
[11] https://huggingface.co/sentence-transformers/LaBSE
[12] https://huggingface.co/myyycroft/XCOMET-lite
[13] https://huggingface.co/bigscience/mt0-large
[14] https://huggingface.co/s-nlp/mt0-xl-detox-orpo

where N is the total number of tokens in the batch, $m_i \in \{0, 1\}$ is the mask, $y_i \in \{0, 1\}$ is the ground-truth label for the i-th token, $p_i \in (0, 1)$ is the predicted probability of class 1 (toxic) for the i-th token.

This requires additional training data, specifically token classes (0 for neutral and 1 for toxic) for each sentence. The underlying idea is that the information learned by the encoder will assist the decoder in detoxifying the text, thus improving the overall performance of the model. It is important to note that the classification head is required only during the training process, after which this head is removed.

4.2.3. Contrastive loss

For our contrastive loss function, we decided to use the InfoNCE function [11]. This function is particularly well-suited for training scenarios involving variable pairs, where each pair consists of an anchor example and a positive example, accompanied by a large number of in-batch negative examples. In this context, the anchor and positive examples refer to two instances that share the same semantic meaning (toxic sentence and its detoxified version), whereas the negative examples represent instances with different meanings (the rest of in-batch toxic and neutral sentences).

Given a set $X = \{x_1 \dots x_N\}$ of N random samples containing one positive sample from $p(x_{t+k}|c_t)$ and $N-1$ negative samples from the 'proposal' distribution $p(x_{t+k})$, we optimize:

$$\mathcal{L}_{contrastive} = -\mathop{\mathbb{E}}_{\mathbb{X}} \left[\log \frac{\mathbf{f_k}(\mathbf{x_{t+k}}, \mathbf{c_t})}{\sum_{x_j \in \mathbb{X}} \mathbf{f_k}(\mathbf{x_j}, \mathbf{c_t})} \right] \tag{1}$$

The primary objective of the InfoNCE function is to minimize the distance between the embeddings of the anchor and positive examples, while simultaneously maximizing the distance between the anchor and all negative examples in the batch. By doing so, the model learns to produce embeddings that are closer together for semantically similar instances (i.e., the anchor and positive examples) and farther apart for semantically dissimilar instances (i.e., the anchor and negative examples).

4.2.4. Final loss formulation

Final loss is calculated via direct sum of the three losses:

$$\mathcal{L}_{final} = \mathcal{L}_{seq2seq} + \mathcal{L}_{classification} + \mathcal{L}_{contrastive}.$$

4.3. Hyperparameters

For the experimental training run of the bigscience/mt0-large model [15], effective batch size was set to 128, adafactor [12] optimizer, learning rate of 3e-4. For the final runs on the s-nlp/mt0-xl-detox-orpo [5] model [16], effective batch size and the optimizer remained the same, while learning rate was changed to 5e-5. All models were trained for 7 epochs.

5. Experiments

First, we validated the idea on a smaller model on a subset of a synthetic SynthDetoxM dataset, then trained the same model on a custom data mix and then trained a final bigger model for final prediction.

5.1. Preliminary experiments

To confirm the hypothesis that Sage-T5-like approach will increase detoxification scores, we have randomly sampled 600 examples per language from SynthDetoxM dataset as test data in Russian, French and Spanish languages and trained three detoxification models on the remaining 3400 training examples

[15] https://huggingface.co/bigscience/mt0-large
[16] https://huggingface.co/s-nlp/mt0-xl-detox-orpo

per language. One model was the baseline, trained on sequence-to-sequence paraphrasing task. The second model was trained for both sequence-to-sequence paraphrasing and also with a classification head. The third model is trained using a classification head and a contrastive loss for the encoder. Following the SynthDetoxM methodology, the models were evaluated with STA and SIM scores.

Table 2
Comparison of metrics for models, trained with paraphrase training objective, combination of paraphrase and classification, and paraphrase, classification and contrastive losses.

	Paraphrase			Classification			All losses		
Language	**STA**	**SIM**	**J**	**STA**	**SIM**	**J**	**STA**	**SIM**	**J**
ru	0.5886	**0.8102**	0.4504	0.8686	0.7626	0.6545	**0.9285**	0.7222	**0.6633**
fr	0.7353	**0.8769**	0.6328	0.8689	0.8415	**0.7258**	**0.8938**	0.8141	0.7229
es	0.4943	**0.8989**	0.4271	0.6828	0.8622	0.5716	**0.7949**	0.8394	**0.6535**

The results are presented in the Table 2. The addition of the classification and contrastive losses increases the STA scores, but slightly decreases the SIM score. This means that models learn to do paraphrasing better, making detoxification less toxic, but these outputs differ more from the original texts. This confirmed validity of the approach on a clean, high quality, synthetic dataset.

5.2. Expanding the evaluation to more languages

The second stage of our experiments consisted of two parts: creating a data mix for training a massively multilingual model and training a set of smaller models to evaluate our approach on the test set of the competition. The data mix consisted of the public part of the MultiParaDetox dataset, consisting of 400 examples per language, SynthDetoxM dataset, consisting of 4000 examples per language and our synthetic SynthDetoxM-like generated dataset.

To further validate the approach, we've again trained three models: baseline paraphrasing model, a model with classification added to the paraphrasing loss and a model with all three losses. For the base model, bigscience/mt0-large was selected. The results were evaluated on the test data using the CodaLab leaderboard.

Table 3
Averaged scores for models, trained with paraphrase training objective, combination of paraphrase and classification, and paraphrase, classification and contrastive losses on the test dataset.

Name	Average all	Average non-parallel	Average parallel
Paraphrase	**0.4773**	0.3371	**0.5709**
Classification	0.4728	**0.3461**	0.5573
All losses	0.4344	0.3212	0.5099

In contrast to the results of preliminary experiments on the low amount of languages, adding additional losses when training on all 15 languages did not improve detoxification quality. On average, adding classification loss helped a little with the new languages (Italian, French, Hebrew, Hinglish, Tatar, Japanese) and slightly decreased scores for old languages. Adding all three losses decreased all scores. Detailed scores are shown in the Table 3.

If we take a look at per language scores, we can see that the largest increase in detoxification quality from adding a classification head is in Amharic, Hebrew, Hinglish, Japanese, Italian and Russian languages. All of these languages except Russian are low resource languages, which did not dominate the pretraining dataset of the model, so we can say that adding a classification head works best for low resource detoxification training. Per language scores are shown in Table 4.

Table 4
Per-Language Scores for the models

Language	Paraphrase J	Classification J	All losses J
am	0.4290	**0.4345**	0.4096
ar	**0.6026**	0.5613	0.4518
de	**0.5821**	0.5650	0.5373
en	**0.5784**	0.5489	0.5105
es	**0.6033**	0.6026	0.5576
fr	0.6710	**0.6714**	0.6216
he	0.3184	**0.3269**	0.3027
hi	**0.5891**	0.5420	0.5084
hin	0.2626	**0.2731**	0.2374
it	0.1374	0.1443	**0.1462**
ja	0.3618	**0.3929**	0.3695
ru	0.6528	**0.6622**	0.6247
tt	**0.2714**	0.2678	0.2499
uk	0.6033	**0.6087**	0.5241
zh	**0.4971**	0.4908	0.4652

Table 5
Per-Language Scores

Place	User	AvgP	AvgNP
6	`baseline_mt0`	0.675 (5)	0.572 (12)
13	`baseline_gpt4`	0.637 (12)	0.579 (9)
14	`nikita.sushko`	0.628 (13)	0.512 (18)
23	`baseline_o3mini`	0.562 (22)	0.484 (25)
24	`baseline_gpt4o`	0.560 (23)	0.535 (16)
27	`baseline_delete`	0.536 (26)	0.510 (21)
29	`baseline_backtranslation`	0.481 (28)	0.342 (30)
30	`baseline_duplicate`	0.475 (29)	0.482 (26)

5.3. Final model training

Since the best on average approach for training from our expanded evaluation is shown that training the model only on paraphrases shows the best quality, this approach was selected for the final model training. For this training pass, s-nlp/mt0-xl-detox-orpo model was selected as a baseline model and then finetuned on our dataset mix, consisting of MultiParaDetox, SynthDetoxM and our synthetic dataset. Final results and comparisons to the baselines are shown in the Table 5.

Our model placed 14th in the final ranking, outcompeting all simple baselines, gpt4o and o3-mini and losing to gpt4 and mt0 baselines. You can see detoxification examples in appendix A.

Additionally, the authors provided LLM-as-a-judge final evaluation, where Llama-3-8b-Instruct model was additionally finetuned on the manual annotation of the previous year's competition for toxic parewise comparison and similarity tasks. The fluency metric was still calculated by xcomet-lite model. You can see results of this evaluation in appendix B.

6. Results and discussion

Our approach demonstrated strong performance on three selected languages. However, it did not generalize well to a broader set of languages. We hypothesize that this limitation stems from the low quality of toxic token annotation in some languages, likely due to incomplete or inconsistent toxic lexicons. Future work should investigate more robust methods for toxic span detection, such as leveraging large language models to improve annotation quality.

We attribute the gap between our final model and GPT-4 primarily to the significantly smaller size of

our backbone model, mT0-XL, which contains only 3 billion parameters. However, another important observation emerged during our experiments: directly fine-tuning the original mT0 model on our custom multilingual detoxification data mixture resulted in a noticeable drop in performance compared to the initial zero-shot baseline.

This degradation may be explained by the shift in training methodology. The original model was trained using ORPO [13] optimization. This optimization can create a fragile equilibrium in the model's parameter space, where the learned behaviors depend heavily on maintaining the alignment enforced during ORPO.

When we applied regular supervised fine-tuning (SFT) on raw, unfiltered training data, it is likely that this alignment was disrupted. SFT tends to push the model back toward the mode of the new data distribution, which may conflict with the preference-aligned behavior established by ORPO. As a result, the model may regress or exhibit erratic outputs, especially in nuanced tasks like detoxification, where subtle distributional shifts can lead to pronounced degradation in quality. This highlights the need for more careful integration of aligned models and raw training data, particularly when extending or adapting preference-optimized backbones to new domains.

Furthermore, we did not apply any data cleaning procedures, and the suboptimal quality of the MultiParaDetox dataset may have further impacted the model's effectiveness.

7. Conclusions

In this paper, we propose a novel approach to text detoxification using two auxiliary losses. If high quality markup is used for training the encoder classification head, our approach significantly outperforms seq2seq training. However, for weak markup, seq2seq training still works better than our approach.

Final submission outperformed all simple baselines, o3-mini and gpt4o on the private test set, coming close to detoxification quality by a much larger gpt4 model.

Our data preprocessing and model training scripts can be found on GitHub [17]. Our trained models can be found on HuggingFace:

- Model with detox and classification losses[18];
- Model with detox and contrastive losses losses[19];
- Model with all losses[20];

Our collected dataset is also available at our HuggingFace repository[21].

8. Limitations and Future Work

Our exploration of a Sage-T5-like approach for multilingual text detoxification, while giving valuable insights, encountered several limitations. The primary challenge was the inconsistent generalizability of the multi-task learning benefits (seq2seq, classification, and contrastive losses) when scaling from a few well-performing languages to the full set of 15. This suggests that the uniform application of these auxiliary losses might not be optimal across diverse linguistic structures and data availabilities. Furthermore, the performance of our token-level classification, and consequently the entire multi-task model, was heavily reliant on the quality of toxic token annotations derived from multilingual lexicons. The incompleteness or inconsistencies within these lexicons likely introduced noise, particularly for less-resourced languages. Future work could address these issues improving annotation quality via more sophisticated toxic span detection methods, such as leveraging large language models for few-shot annotation.

[17] https://github.com/chameleon-lizard/Sage-Detox
[18] https://huggingface.co/alexandro767/SageDetox_detox_classification
[19] https://huggingface.co/alexandro767/SageDetox_detox_contrastive
[20] https://huggingface.co/alexandro767/SageDetox_detox_classification_contrastive
[21] https://huggingface.co/datasets/alexandro767/CLEF_2025_dataset_full

Another set of limitations pertains to the model architecture and data handling. The choice of bigscience/mT0-xl as the backbone, while competitive, is significantly smaller than some leading proprietary models, inherently constraining its capacity. Moreover, fine-tuning the s-nlp/mt0-xl-detox-orpo model (originally trained with ORPO) using our direct supervised approach led to performance degradation, indicating a potential mismatch in training paradigms or catastrophic forgetting. The quality of the aggregated training data, which did not undergo extensive cleaning, might also have impacted performance. Future research could benefit from experimenting with larger, more capable open-source multilingual models and implementing data filtering and cleaning protocols.

9. Declaration on Generative AI

During the preparation of this work, the author(s) used Gemini and Grammarly. Gemini was used for translation of the toxic and non-toxic claims in the paper and both Gemini and Grammarly were used for checking the grammar and spelling. After using this tool/service, the author(s) reviewed and edited the content as needed and take(s) full responsibility for the publication's content.

References

[1] N. Martynov, M. Baushenko, A. Kozlova, K. Kolomeytseva, A. Abramov, A. Fenogenova, A methodology for generative spelling correction via natural spelling errors emulation across multiple domains and languages, in: Y. Graham, M. Purver (Eds.), Findings of the Association for Computational Linguistics: EACL 2024, St. Julian's, Malta, March 17-22, 2024, Association for Computational Linguistics, 2024, pp. 138–155. URL: https://aclanthology.org/2024.findings-eacl.10.

[2] D. Moskovskiy, N. Sushko, S. Pletenev, E. Tutubalina, A. Panchenko, Synthdetoxm: Modern llms are few-shot parallel detoxification data annotators, in: L. Chiruzzo, A. Ritter, L. Wang (Eds.), Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL 2025 - Volume 1: Long Papers, Albuquerque, New Mexico, USA, April 29 - May 4, 2025, Association for Computational Linguistics, 2025, pp. 5714–5733. URL: https://aclanthology.org/2025.naacl-long.294/.

[3] D. Dementieva, D. Moskovskiy, N. Babakov, A. A. Ayele, N. Rizwan, F. Schneider, X. Wang, S. M. Yimam, D. Ustalov, E. Stakovskii, A. Smirnova, A. Elnagar, A. Mukherjee, A. Panchenko, Overview of the multilingual text detoxification task at pan 2024, in: G. Faggioli, N. Ferro, P. Galuščáková, A. G. S. de Herrera (Eds.), Working Notes of CLEF 2024 - Conference and Labs of the Evaluation Forum, CEUR-WS.org, 2024.

[4] S. Pletenev, Somethingawful at PAN 2024 textdetox: Uncensored llama 3 helps to censor better, in: G. Faggioli, N. Ferro, P. Galuscáková, A. G. S. de Herrera (Eds.), 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*, CEUR-WS.org, 2024, pp. 2843–2851. URL: https://ceur-ws.org/Vol-3740/paper-273.pdf.

[5] E. Rykov, K. Zaytsev, I. Anisimov, A. Voronin, Smurfcat at PAN 2024 textdetox: Alignment of multilingual transformers for text detoxification, in: G. Faggioli, N. Ferro, P. Galuscáková, A. G. S. de Herrera (Eds.), 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*, CEUR-WS.org, 2024, pp. 2866–2871. URL: https://ceur-ws.org/Vol-3740/paper-276.pdf.

[6] D. Dementieva, V. Protasov, N. Babakov, N. Rizwan, I. Alimova, C. Brune, V. Konovalov, A. Muti, C. Liebeskind, M. Litvak, D. Nozza, S. Shah Khan, S. Takeshita, N. Vanetik, A. A. Ayele, F. Schneider, X. Wang, S. M. Yimam, A. Elnagar, A. Mukherjee, A. Panchenko, Overview of the multilingual text detoxification task at pan 2025, in: G. Faggioli, N. Ferro, P. Rosso, D. Spina (Eds.), Working Notes of CLEF 2025 – Conference and Labs of the Evaluation Forum, CEUR Workshop Proceedings, CEUR-WS.org, 2025.

[7] J. Bevendorff, D. Dementieva, M. Fröbe, B. Gipp, A. Greiner-Petter, J. Karlgren, M. Mayerl, P. Nakov,

A. Panchenko, M. Potthast, A. Shelmanov, E. Stamatatos, B. Stein, Y. Wang, M. Wiegmann, E. Zangerle, Overview of PAN 2025: Voight-Kampff Generative AI Detection, Multilingual Text Detoxification, Multi-Author Writing Style Analysis, and Generative Plagiarism Detection, in: J. C. de Albornoz, J. Gonzalo, L. Plaza, A. G. S. de Herrera, J. Mothe, F. Piroi, P. Rosso, D. Spina, G. Faggioli, N. Ferro (Eds.), Experimental IR Meets Multilinguality, Multimodality, and Interaction. Proceedings of the Sixteenth International Conference of the CLEF Association (CLEF 2025), Lecture Notes in Computer Science, Springer, Berlin Heidelberg New York, 2025.

[8] F. Feng, Y. Yang, D. Cer, N. Arivazhagan, W. Wang, Language-agnostic BERT sentence embedding, CoRR abs/2007.01852 (2020). URL: https://arxiv.org/abs/2007.01852. arXiv:2007.01852.

[9] D. Larionov, M. Seleznyov, V. Viskov, A. Panchenko, S. Eger, xcomet-lite: Bridging the gap between efficiency and quality in learned MT evaluation metrics, in: Y. Al-Onaizan, M. Bansal, Y. Chen (Eds.), Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing, EMNLP 2024, Miami, FL, USA, November 12-16, 2024, Association for Computational Linguistics, 2024, pp. 21934–21949. URL: https://aclanthology.org/2024.emnlp-main.1223.

[10] N. Muennighoff, T. Wang, L. Sutawika, A. Roberts, S. Biderman, T. L. Scao, M. S. Bari, S. Shen, Z. X. Yong, H. Schoelkopf, X. Tang, D. Radev, A. F. Aji, K. Almubarak, S. Albanie, Z. Alyafeai, A. Webson, E. Raff, C. Raffel, Crosslingual generalization through multitask finetuning, in: A. Rogers, J. L. Boyd-Graber, N. Okazaki (Eds.), Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2023, Toronto, Canada, July 9-14, 2023, Association for Computational Linguistics, 2023, pp. 15991–16111. URL: https://doi.org/10.18653/v1/2023.acl-long.891. doi:10.18653/V1/2023.ACL-LONG.891.

[11] A. van den Oord, Y. Li, O. Vinyals, Representation learning with contrastive predictive coding, CoRR abs/1807.03748 (2018). URL: http://arxiv.org/abs/1807.03748. arXiv:1807.03748.

[12] N. Shazeer, M. Stern, Adafactor: Adaptive learning rates with sublinear memory cost, in: J. G. Dy, A. Krause (Eds.), 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*, PMLR, 2018, pp. 4603–4611. URL: http://proceedings.mlr.press/v80/shazeer18a.html.

[13] J. Hong, N. Lee, J. Thorne, ORPO: monolithic preference optimization without reference model, in: Y. Al-Onaizan, M. Bansal, Y. Chen (Eds.), Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing, EMNLP 2024, Miami, FL, USA, November 12-16, 2024, Association for Computational Linguistics, 2024, pp. 11170–11189. URL: https://aclanthology.org/2024.emnlp-main.626.

A. Generation examples

Original Text	Paraphrased Version	Language Code
Просто забирайся нахуй з моєї голови.	я просто забирайся з моєї голови.	uk
заебали инде шул ватс апыгыз белэн))	алар бик начар кешеләр.	tt
Maro. Maderchodo ko chodo mat.. Ek ek ko gus ke maro	yeh log apni baat ko shuru kar rahe hain.	hin
Je n ai jamais rencontré d individu aussi bas que ce rat d égout !	je n ai jamais rencontré d individu aussi bas que ce rat !	fr
Si le votan a este energúmeno de verdad no aman a su ciudad.	si le votan a este político de verdad no aman a su ciudad.	es

Original Text	Paraphrased Version	Language Code
А нахуй ты тут персональный чатик устроил, дегенерат?	а почему ты тут персональный чат устроил?	ru
Maybe so, but at least American's are smarter than dumbass Brits.	maybe so, but at least Americans are smarter than those of brits.	en
Stoppt denn keiner diese Dummbacke!	stoppt denn keiner diese menschen!	de

B. LLM-as-a-Judge evaluation results

Table 7
LLM-as-a-Judge evaluation on languages with parallel markup. Our submission is highlighted in **bold**, baselines are highlighted in *italic*.

#	User	average	en	es	de	zh	ar	hi	uk	ru	am
1	*golden annotation*	*0.820*	*0.846*	*0.783*	*0.930*	*0.716*	*0.838*	*0.888*	*0.807*	*0.828*	*0.742*
2	Team MetaDetox	0.812	0.893	0.823	0.919	0.813	0.826	0.785	0.791	0.829	0.626
3	ducanhhbtt	0.798	0.871	0.797	0.919	0.796	0.814	0.762	0.785	0.827	0.614
4	Team ReText.Ai Team	0.775	0.794	0.765	0.888	0.783	0.790	0.773	0.791	0.792	0.597
5	Jiaozipi	0.768	0.882	0.812	0.840	0.744	0.779	0.701	0.757	0.817	0.579
6	Team The Toxinators 2000	0.768	0.842	0.758	0.828	0.715	0.788	0.759	0.766	0.818	0.638
7	Team Pratham	0.768	0.843	0.763	0.822	0.717	0.793	0.758	0.782	0.811	0.621
8	*baseline_mt0*	*0.768*	*0.843*	*0.764*	*0.825*	*0.717*	*0.791*	*0.751*	*0.770*	*0.809*	*0.639*
9	sky.Duan	0.765	0.847	0.757	0.830	0.696	0.787	0.747	0.771	0.811	0.638
10	Dalfa	0.764	0.840	0.758	0.849	0.721	0.789	0.738	0.756	0.790	0.640
11	Team cake	0.742	0.805	0.807	0.866	0.652	0.720	0.689	0.757	0.792	0.591
12	jrluo	0.742	0.825	0.723	0.814	0.696	0.746	0.711	0.752	0.778	0.631
13	Team Transformers	0.737	0.871	0.778	0.853	0.673	0.713	0.706	0.758	0.788	0.492
14	**nikita.sushko**	**0.735**	**0.858**	**0.740**	**0.828**	**0.702**	**0.652**	**0.732**	**0.772**	**0.835**	**0.491**
15	SVATS	0.723	0.830	0.749	0.854	0.776	0.705	0.672	0.743	0.798	0.380
16	Team Detox	0.722	0.691	0.757	0.819	0.699	0.718	0.701	0.742	0.792	0.580
17	*baseline_gpt4*	*0.715*	*0.858*	*0.800*	*0.807*	*0.654*	*0.686*	*0.647*	*0.723*	*0.778*	*0.482*
18	Team Nililusu	0.714	0.796	0.624	0.772	0.725	0.718	0.716	0.727	0.760	0.594
19	Penitto	0.694	0.837	0.756	0.831	0.685	0.643	0.638	0.643	0.762	0.452
20	Davv	0.692	0.857	0.751	0.791	0.703	0.604	0.656	0.611	0.722	0.531
21	shashist	0.687	0.793	0.727	0.742	0.641	0.643	0.695	0.735	0.770	0.437
22	danielleee	0.684	0.860	0.775	0.811	0.768	0.628	0.576	0.632	0.706	0.402
23	sameertantry	0.680	0.783	0.650	0.736	0.649	0.708	0.677	0.720	0.707	0.486
24	*baseline_o3mini*	*0.676*	*0.893*	*0.796*	*0.747*	*0.652*	*0.595*	*0.609*	*0.663*	*0.711*	*0.421*
25	SomethingAwful	0.663	0.856	0.763	0.749	0.629	0.592	0.631	0.694	0.685	0.367
26	Cchenz	0.607	0.828	0.710	0.724	0.579	0.530	0.504	0.559	0.631	0.401
27	*baseline_gpt4o*	*0.580*	*0.770*	*0.694*	*0.533*	*0.482*	*0.539*	*0.490*	*0.608*	*0.615*	*0.486*
28	keke	0.573	0.816	0.693	0.670	0.544	0.443	0.464	0.512	0.612	0.405
29	*baseline_delete*	*0.558*	*0.453*	*0.543*	*0.564*	*0.630*	*0.610*	*0.566*	*0.577*	*0.583*	*0.499*
30	*baseline_backtranslation*	*0.458*	*0.743*	*0.466*	*0.479*	*0.231*	*0.442*	*0.395*	*0.256*	*0.689*	*0.425*
31	*baseline_duplicate*	*0.432*	*0.370*	*0.451*	*0.479*	*0.429*	*0.446*	*0.432*	*0.455*	*0.450*	*0.380*
32	Dorevain	0.346	0.838	0.407	0.417	0.170	0.337	0.288	0.152	0.201	0.305

After the end of the competition, organizers did another evaluation round, using finetuned Llama-3-8B-Instruct model as a judge. The resulting table provided a shakeup in the results table. The results can be seen in Table 7 and Table 8.

Our model outperformed gpt4 baseline in this evaluation round on languages with available parallel data, while not surpassing the same baseline on languages without parallel markup. This can be attributed to low resourcefulness of said languages and that the models underperformed in them due to tokenization quality, low pretraining data amount and general out-of-distribution for the models, which were used as a base for our detoxifiers.

Table 8
LLM-as-a-Judge evaluation on languages without parallel markup. Our submission is highlighted in **bold**, baselines are highlighted in *italic*.

#	User	average	it	ja	he	fr	tt	hin
1	*golden annotation*	*0.828*	*0.893*	*0.904*	*0.783*	*0.724*	*0.780*	*0.887*
2	Team ReText.Ai Team	0.722	0.823	0.805	0.657	0.860	0.583	0.606
3	ducanhhbtt	0.720	0.842	0.820	0.681	0.889	0.495	0.592
4	Team Detox	0.704	0.812	0.784	0.631	0.843	0.575	0.578
5	Team MetaDetox	0.691	0.821	0.721	0.610	0.883	0.493	0.621
6	Jiaozipi	0.688	0.795	0.787	0.611	0.850	0.541	0.544
7	Dalfa	0.678	0.789	0.703	0.615	0.790	0.621	0.548
8	Team Transformers	0.675	0.813	0.781	0.618	0.854	0.506	0.479
9	Team cake	0.674	0.791	0.796	0.581	0.853	0.436	0.584
10	sky.Duan	0.668	0.822	0.769	0.619	0.873	0.416	0.509
11	*baseline_gpt4*	*0.662*	*0.790*	*0.779*	*0.578*	*0.865*	*0.438*	*0.524*
12	SVATS	0.662	0.775	0.734	0.576	0.815	0.523	0.549
13	Team The Toxinators 2000	0.648	0.742	0.745	0.495	0.790	0.611	0.503
14	*baseline_mt0*	*0.641*	*0.749*	*0.711*	*0.501*	*0.793*	*0.598*	*0.494*
15	Team Pratham	0.639	0.752	0.710	0.495	0.801	0.575	0.502
16	shashist	0.638	0.732	0.683	0.594	0.803	0.505	0.509
17	**nikita.sushko**	**0.628**	**0.763**	**0.722**	**0.573**	**0.807**	**0.492**	**0.410**
18	danielleee	0.611	0.759	0.728	0.510	0.860	0.379	0.430
19	Penitto	0.608	0.752	0.725	0.530	0.845	0.349	0.447
20	Davv	0.608	0.742	0.744	0.523	0.816	0.339	0.487
21	jrluo	0.597	0.718	0.680	0.383	0.764	0.582	0.453
22	SomethingAwful	0.579	0.728	0.643	0.490	0.814	0.324	0.477
23	Cchenz	0.560	0.709	0.631	0.510	0.736	0.403	0.366
24	*baseline_o3mini*	*0.559*	*0.748*	*0.661*	*0.497*	*0.826*	*0.209*	*0.411*
25	Team Nililusu	0.530	0.662	0.647	0.357	0.659	0.543	0.312
26	keke	0.526	0.695	0.554	0.429	0.760	0.370	0.347
27	*baseline_gpt4o*	*0.526*	*0.697*	*0.680*	*0.370*	*0.718*	*0.327*	*0.363*
28	*baseline_delete*	*0.525*	*0.628*	*0.443*	*0.496*	*0.576*	*0.521*	*0.486*
29	sameertantry	0.516	0.654	0.510	0.326	0.747	0.497	0.360
30	*baseline_duplicate*	*0.429*	*0.455*	*0.442*	*0.407*	*0.460*	*0.421*	*0.387*
31	*baseline_backtranslation*	*0.254*	*0.333*	*0.147*	*0.349*	*0.503*	*0.054*	*0.139*
32	Dorevain	0.221	0.274	0.150	0.283	0.505	0.048	0.067