SVATS at PAN 2025 TextDetox: Can Small Models Outperform Large Ones in Text Detoxification?

Notebook for PAN at CLEF 2025

Vladislav Kozlovskiy[1,†], Sameer Tantry[1,†], Alexander Ploskin[1,*], Tatyana Matveeva[1] and Sofya Savelyeva[1]

[1]Skolkovo Institute of Science and Technology, Bolshoy Boulevard, 30, p.1, 121205, Moscow, Russian Federation

Abstract

Toxic language, which includes hate speech, insults, and offensive expressions, poses significant challenges to online communication, mental health, and social cohesion. Additional complications arise in multilingual environments where the development of generalized solutions remains a persistent challenge due to linguistic diversity and resource constraints. In this work, we systematically investigate the effectiveness of existing small- and medium-scale models for multilingual text detoxification, addressing the critical need for computationally efficient approaches that maintain performance in diverse linguistic contexts while operating within practical resource limitations.

Keywords

Toxicity Mitigation, Few-Shot, Fine-Tunning, Style Transfer, Multilingual prompting

1. Introduction

The rapid proliferation of user-generated content across digital platforms has underscored the critical need for automated text detoxification systems. Toxic language, including speech, insults, and offensive expressions, poses significant challenges to online communication, mental health, and social cohesion. Although considerable progress has been made in toxicity detection, the task of text detoxification—rewriting toxic text into non-toxic alternatives while preserving meaning and fluency—remains a complex and underexplored problem.

Recent advancements in natural language processing (NLP), particularly the rise of large language models (LLMs), have opened new avenues for text style transfer and content moderation. The PAN-Detox Competition 2024 [1] has played a pivotal role in benchmarking state-of-the-art detoxification methods, providing a standardized evaluation framework and diverse datasets. Building upon these efforts, this paper is written as a part of the PAN-Detox Competition 2025 [2] and investigates the effectiveness of various detoxification approaches, including fine-tuned LLMs, sequence-to-sequence models, and techniques utilizing synthetic data generation.

Our contributions are as follows.

- Review of known methods: We analyze top-performing models from the previous year's competition and other SOTA methods and provide a review of existing datasets.

CLEF 2025 Working Notes, 9 – 12 September 2025, Madrid, Spain

*Corresponding author.

†These authors contributed equally.

Vladislav.Kozlovskiy@skoltech.ru (V. Kozlovskiy); Sameer.Tantry@skoltech.ru (S. Tantry); Alexander.Ploskin@skoltech.ru (A. Ploskin); Tatyana.Matveeva@skoltech.ru (T. Matveeva); Sofya.Savelyeva@skoltech.ru (S. Savelyeva)

https://github.com/VladKozlovskiy (V. Kozlovskiy); https://github.com/sameertantry (S. Tantry); https://github.com/Alexander-Ploskin (A. Ploskin); https://github.com/tnmtvv (T. Matveeva); https://github.com/sofyafyaa (S. Savelyeva)

0009-0008-5964-1685 (V. Kozlovskiy); 0009-0008-2453-6711 (A. Ploskin); 0009-0004-0680-5050 (T. Matveeva); 0009-0008-0434-5475 (S. Savelyeva)

- A methodology for synthetic data generation: While abundant work exists on toxicity mitigation for popular languages, niche languages such as Tatar or Hinglish (the mix of Hindi and English) lack paired toxic-nontoxic data, thus requiring additional efforts for artificial generation.
- Experiments with a variety of models and techniques: We experimented with a set of techniques including fine-tuning, few-shot prompting, and others across different models and languages. As our main result, we propose several comparatively **small models with 1 to 8 billion parameters that achieve higher scores than one of the previous year's baseline [3], which is a 13B model**.
- Our solution achieved 10th place overall on the competition leaderboard[1], and ranked 6th in four languages: Amharic, Italian, Tatar, and Hindi.

By addressing the trade-offs between detoxification strength and text quality, this study advances the development of safer, more inclusive digital communication tools. Our findings not only contribute to the academic discourse on text style transfer, but also offer practical implications for social media platforms, content moderators, and AI ethics researchers.

2. Related Works

Text detoxification transforms toxic text into neutral language while preserving meaning. Early baselines such as *delete*, *duplicate*, and *backtranslation* offer simple solutions but often compromise fluency and semantic accuracy, establishing the foundational challenges that subsequent research has been aimed at addressing.

Building upon these limitations, large language models have enabled more sophisticated and effective approaches to text detoxification. The multitask model mT0 [3] demonstrates strong zero-shot and few-shot detoxification capabilities through prompt-based multitask learning, generalizing across languages without task-specific fine-tuning. This advancement represents a significant departure from rule-based methods toward more nuanced understanding of linguistic toxicity patterns.

The effectiveness of these modern approaches is further validated in the PAN 2024 Multilingual Text Detoxification Task [4], which highlighted the persistent challenges in multilingual detoxification and underscored the critical importance of meaning preservation. Notably, few-shot prompting emerged as a particularly effective method in this competition, especially when applied to instruction-tuned models like mT0, demonstrating the practical viability of prompt-based approaches in real-world scenarios.

The success of few-shot prompting can be attributed to its ability to allow models to learn detoxification patterns from a handful of carefully selected examples [5], thereby enabling effective generalization in low-resource and crosslingual settings where traditional supervised learning approaches would fail. Models like mT0 benefit greatly from this approach due to their inherent multitask training paradigm, which facilitates rapid adaptation to new detoxification contexts.

However, the scarcity of high-quality training data remains a significant bottleneck in many languages and domains. To address this challenge, synthetic data generation has emerged as a crucial technique for providing paired toxic and non-toxic examples, particularly supporting training scenarios where manually annotated data are scarce [6, 7]. When combined with parameter-efficient fine-tuning methods like LoRA [8], which updates only a small subset of model weights while maintaining performance, this approach enables scalable and cost-effective model adaptation across diverse linguistic contexts.

Contemporary research continues to push the boundaries of detoxification performance through the deployment of advanced architectures. State-of-the-art models such as **Gemma3–4B** [9], **T5** [10], and **Qwen2-7B** [11], when strategically combined with synthetic data generation and sophisticated prompting techniques, continue to advance detoxification performance substantially beyond traditional baselines, establishing new benchmarks for both effectiveness and efficiency in multilingual text detoxification tasks.

[1] https://codalab.lisn.upsaclay.fr/competitions/22396#results

3. Methodology

In this section, we provide an elaboration of our proposed solution framework and systematically formulate the hypotheses we rigorously test throughout our experimental investigations. We present an analysis of our methodological approach, including the derivation of our design choices and the empirical validation strategies employed to assess their effectiveness. Additionally, we provide an extensive overview of the synthetic data generation techniques.

3.1. Data

This section presents an overview of the datasets we employ in our training framework, which encompasses both established detoxification datasets from prior research and synthetically generated data. While existing paired datasets demonstrate high quality and have been instrumental in training state-of-the-art solutions across multiple languages (English, Spanish, German, Russian, Ukrainian, French), they remain scarce or entirely absent for many low-resource languages (e.g. Amharic). Although unpaired datasets can be assembled through web scraping techniques combined with toxicity classifiers, this approach represents an active area of ongoing research with inherent limitations.

In this study, we categorize languages into **high-resource** and **low-resource** classifications based on the availability and comprehensiveness of open-source detoxification datasets. Under this framework, **English, Spanish, Russian, Ukrainian, German, and French** are classified as high-resource languages due to their substantial paired detoxification data availability. Conversely, **Italian, Arabic, Hebrew, Hindi, Tatar, Japanese, Chinese, Hinglish, and Amharic** are designated as low-resource languages, reflecting the limited or absent paired datasets for these linguistic contexts.

This resource-based taxonomy directly influences our experimental design and synthetic data generation priorities, with low-resource languages requiring more extensive augmentation strategies to achieve comparable training data volumes.

Given the current requirement for paired datasets in our methodology, we conducted extensive experiments with various synthetic data generation approaches, which are detailed in the subsequent sections. The impact of these different synthetic data generation strategies on model performance is systematically evaluated and discussed in further analysis.

3.1.1. Existing Paired Datasets

The landscape of paired detoxification datasets reveals substantial disparities in data availability across languages. Table 1 summarizes the key characteristics of existing paired datasets, including their size, language coverage, and potential applications in our training framework.

Table 1
Overview of Existing Paired Detoxification Datasets

Dataset	Size	Languages
Multilingual ParaDetox [12, 2, 13]	3.6k	en, ru, uk, de, es, am, zh, ar, hi
Multilingual Transformer [14]	55k	en, ru, uk, de, es, am, zh, ar, hi
SynthDetoxM [7]	16k	ru, de, fr, es
ParaDetox (English) [15]	20k	en
ParaDetox (Russian) [16]	10k	ru
ParaDetox (Ukrainian) [17]	4k	uk
ParaDetox (Spanish) [17]	500	es

The **Multilingual ParaDetox** dataset is provided by competition organizers, despite its limited size of 400 samples per language, it enables few-shot learning experiments across most of target languages due to high quality of data.

The **Multilingual Transformer Detoxification** dataset represents the most rich resource containing 55, 000 examples across 9 languages, its effectiveness is proven as it is instrumental in training the

previous year's competition winner. The dataset's foundation on translated English content highlights its limited capabilities in complex and rare languages, like Amharic.

The **SynthDetoxM** dataset introduces a valuable synthetic data component, containing 16, 000 paried examples across four languages (Russian, German, French, and Spanish). This dataset is generated using modern large language models in few-shot setup. This dataset addresses critical gaps in training data, particularly for French, which previously lacked substantial paried resources.

Language-specific datasets provide targeted enhancement opportunities. The **ParaDetox (English)** (20,000 examples), **ParaDetox (Russian)** (10,000 examples), and **ParaDetox (Ukrainian)** (4,000 examples) offer substantial monolingual training data, while **ParaDetox (Spanish)** (500 examples) provide more limited but valuable language-specific resources.

The analysis reveals significant **data scarcity** for several competition languages, though the addition of SynthDetoxM notably improves coverage for French. Most critically, no paired datasets exist for Italian, Hebrew, Hinglish, Tatar, and Japanese, representing a substantial gap in training resources. These languages will require synthetic data generation or cross-lingual transfer learning approaches.

Among languages with available data, **English** and **Russian** demonstrate the strongest resource availability, with multiple datasets totaling over 85,000 and 81,000 examples respectively (including SynthDetoxM contributions). **German** benefits significantly from SynthDetoxM, increasing available training data substantially. **French** now has access to paired data through SynthDetoxM, addressing a previous critical gap. **Ukrainian, Spanish, Amharic, Chinese, Arabic, and Hindi** have moderate coverage through multilingual datasets, with Spanish additionally benefiting from SynthDetoxM augmentation.

The paired datasets can be integrated into a **unified training framework**, with the Multilingual Transformer dataset as the core due to its size and effectiveness. Language-specific datasets refine models for English, Russian, Ukrainian, and Spanish, while SynthDetoxM adds synthetic data for Russian, German, French, and Spanish. SynthDetoxM's synthetic data complements human-annotated sets, enhancing model generalization, especially for languages with limited data.

For languages lacking paired data, multilingual corpora support **cross-lingual transfer**, and the official Multilingual ParaDetox dataset provides evaluation benchmarks. Data gaps remain for five languages, necessitating ongoing synthetic data generation. This enriched dataset landscape, boosted by SynthDetoxM, advances training resource balance across target languages and underscores synthetic data's importance for multilingual coverage. Data availability is uneven: English, Russian, Ukrainian, and German have sufficient data, while most languages lack enough paired examples. This scarcity challenges synthetic data generation.

3.1.2. Synthetic Data Generation

Given the limited availability of paired detoxification datasets across target languages, we implemented a **unified synthetic data generation framework** with two complementary data sourcing approaches to augment our training corpus.

Our synthetic data generation employs a **standardized multi-stage pipeline** that processes different initial data sources through consistent toxification and quality assurance procedures. The framework differentiates primarily in **data acquisition strategies** and **multilingual expansion approaches**, while maintaining uniform processing standards across both pathways.

The unified pipeline consists of the following stages:

1. **Initial Data Acquisition**: Two distinct sourcing strategies provide the foundation corpus.
2. **Toxicity Filtering**: Application of toxicity classifier [15] to ensure baseline corpus quality.
3. **Lexicon-Guided Toxification**: Incorporation of toxic lexical items from multilingual toxic lexicon [18] through few-shot prompting with DeepSeek-V3 model. The few-shot examples is takedn from the ParaDetox dataset provided by the organizers of the competition [12], ensuring alignment with the linguistic distribution characteristics of the test corpus.
4. **Quality Assurance**: Secondary toxicity filtering to validate appropriate toxicity levels and semantic coherence.

5. **Multilingual Expansion**: Target language generation or translation to produce final multilingual datasets.

Data Sourcing Strategies:

Strategy A: Synthetic Content Generation leverages the **tweet-like characteristics** observed in existing datasets through LLM-based content creation. We utilize Qwen3-32B [19] to generate controversial tweets attributed to famous personas, followed by non-toxic but disagreeable responses. This approach produces **10,000 samples per target language** through direct multilingual generation, ensuring consistent coverage across all competition languages.

Strategy B: Real-World Data Foundation addresses potential **LLM bias** by incorporating authentic human discourse as the baseline corpus. We collect approximately **232,000 English-language comments** from a carefully moderated online platform [2], providing diverse and linguistically natural foundation content. Multilingual expansion occurs through translation using DeepSeek-V3 [20] after toxification processing.

The framework employs **two distinct multilingual strategies**:

- **Direct Generation**: Strategy A generates content directly in 15 target languages during the initial content creation phase, leveraging the multilingual capabilities of Qwen3-32B and DeepSeek-V3.
- **Translation-Based**: Strategy B processes English content through the complete pipeline before translating validated toxic-neutral pairs into 15 target languages using DeepSeek-V3.

Both strategies implement **identical quality assurance protocols**: - Pre-toxification filtering ensures clean baseline content - Post-toxification validation confirms appropriate toxicity levels - Toxicity score thresholding ensures dataset consistency

This unified framework produces **toxic-neutral pairs** through complementary approaches: Strategy A offers consistent cross-lingual generation with controlled content characteristics, while Strategy B provides authentic human discourse foundation with superior linguistic diversity and reduced artificial generation artifacts.

Model Selection Rationale.

Our model selection strategy balances multiple factors: **generation quality, computational efficiency, cost-effectiveness, and toxicity generation capability**. This multi-criteria optimization ensures practical feasibility while maintaining high output quality.

Qwen3-32B serves as our primary generation and evaluation model due to its superior balance of quality and efficiency. Its multilingual capabilities ensure consistent performance across target languages.

DeepSeek-V3 fulfill specialized toxification roles where their **reduced content filtering** provides crucial advantages. Unlike many commercial models that heavily censor toxic content generation, DeepSeek models demonstrate greater flexibility in producing the toxic variants.

The strategic model selection addresses the fundamental challenge of **ethical toxic content generation** for research purposes, leveraging models with appropriate capabilities while maintaining responsible research practices through controlled generation environments and systematic quality validation.

3.1.3. Dataset Filtering

To ensemble our final training dataset we combine existing paired detoxification datasets and generated synthetic data to overcome lack of training data in target languages. After the dataset collection, we apply filtration procedure to ensure quality of samples in the training data and coherence of the data with the evaluation metrics. In addition to filtering by toxicity scores, we also ensure style transfer accuracy, similarity and language fluency scores utilizing the metrics published by the authors of

[2] https://tildes.net/

the competition [12]. Finally, we select strict thresholds for different aspects of filteration and obtain approximately 40k pairs of neutral and toxic sentences per language in our training data.

3.2. Models and Experimental Methodology

In this work, we conducted an evaluation of three famous model families—T5, Gemini, and Qwen. Our experiments are systematically designed to investigate four key aspects: (1) the impact of different data subsets on model performance, (2) the influence of training hyperparameters, (3) the efficacy of efficient training techniques, and (4) scaling behavior across model sizes. Below, we detail our approach, findings, and insights for each model family.

Also we deliberately excluded reinforcement learning (RL)-based alignment methods, as prior work in similar contexts had demonstrated limited gains from such techniques.

3.2.1. T5 Model Family: Encoder-Decoder Baseline

Initial Selection and Motivation: We began our investigation with the mT5 model, which served as a strong baseline due to its well-established multilingual capabilities, supporting over 100 languages. The encoder-decoder architecture of T5 models is particularly appealing, as it allowed us to explore a strategy where the encoder could be frozen to capture content and stylistic features, while the decoder is fine-tuned specifically for the detoxification task. This approach is motivated by the hypothesis that separating content encoding from style transformation might improve the quality.

Experimental Observations and Adjustments: Initial results, however, are suboptimal, prompting us to explore alternative configurations. We hypothesized that the pretrained tokenizer in mT5 might be a limiting factor, particularly for languages with diverse scripts. To address this, we evaluated the byT5 variant, which utilizes byte-level UTF-8 encoding and eliminates vocabulary constraints. While this modification improved handling of low-resource languages, the overall detoxification performance remained unsatisfactory. We attributed this to the relatively lightweight decoder, which appeared insufficiently expressive for the complexity of the task.

Due to computational constraints, the largest model we tested is the 770M-parameter variant.

3.2.2. Gemini Model Family: Scaling and Multilingual Adaptation

Rationale for Model Selection: Our next phase focused on the Gemma-3 family, which had recently been released and incorporated state-of-the-art LLM training techniques. Gemma's pretraining dataset included over 200 languages, making it a promising candidate for multilingual detoxification. We primarily experimented with the Gemma-1B instruction-tuned (it) variant, though we also evaluated the pretrained (pt) version (Gemma-1B-pt) and the larger Gemma-4B-IT model to assess scaling effects.

Finetuning and Data Efficiency: In our initial experiments, we fine-tuned Gemma-1B-it on the ParaDetox dataset using a conservative learning rate. We observed that training beyond a single epoch without parameter-efficient methods (e.g., LoRA) led to overfitting, likely due to the limited size of the detoxification dataset. This suggested that conventional full-parameter fine-tuning is not data-efficient for this task.

Parameter-Efficient Adaptation with LoRA: To mitigate overfitting and improve robustness, we integrated Low-Rank Adaptation (LoRA). This allowed us to train for multiple epochs without performance degradation, though the absolute improvement in detoxification quality is marginal. Notably, LoRA's memory efficiency enabled faster experimentation cycles, which is critical given resource constraints.

Language-Specific Tuning and Emergent Phenomena: Recognizing that the base model's multilingual performance might benefit from targeted adaptation, we conducted language-specific fine-tuning using dedicated subsets of the data. This approach yielded measurable improvements in per-language metrics. Interestingly, we observed an unexpected phenomenon: models trained exclusively on English data ($\approx$30k samples) tended to translate non-English inputs into English while simultaneously applying detoxification. Surprisingly, both the translation and detoxification steps are

often performed accurately. This suggests that the model's multilingual knowledge—despite not being explicitly fine-tuned for translation—enabled cross-lingual generalization. This emergent behavior warrants further study, particularly for low-resource language scenarios.

Comparative Analysis of Pretrained vs. Instruction-Tuned Variants: To isolate the impact of instruction tuning, we evaluated the Gemma-1B-PT model. Contrary to our expectations, this variant underperformed compared to its instruction-tuned counterpart, indicating that the alignment phase in Gemma-it's training is useful for task adaptation.

Data Augmentation and Scaling: To address data scarcity and imbalance, we aggregated all available detoxification datasets and supplemented them with synthetically generated examples using the methodology described in Section X. This included backtranslation-based augmentation and filtered samples from prior work (e.g., mT0 and SynthDetoxM [7]). We implemented a quality-filtering pipeline and tuned thresholds to mitigate noise in the combined dataset.

Finally, we scaled our experiments to the Gemma-4b model, which demonstrated consistent improvements in both detoxification quality and multilingual robustness. To optimize training efficiency, we employed sequence packing, to reduce overall training time by.

3.3. Structured Prompting for Toxicity Mitigation

In addition to fine-tuning approaches, we investigated the efficiency of structured prompting with LLMs for text detoxification. This methodology involves designing detailed, context-aware prompts based on the analysis of toxicity patterns observed in our training data. Our analysis revealed three primary categories of toxic language usage, each requiring distinct handling strategies:

- **Emotional Expletives Without Contextual Relevance**:
  - *Pattern*: Frequently, toxic words are used as standalone emotional markers to express sentiment (either positive or negative) without contributing to the semantic content of the message.
  - *Handling Strategy*: Such instances can typically be addressed through direct removal, as the words serve no propositional function. In select cases where preservation of emotional intensity is desired, substitution with non-toxic intensifiers may be appropriate.
- **Contextually Interpretable Toxic Terms**:
  - *Pattern*: Toxic words carrying specific, context-dependent meanings that can be inferred from the immediate discourse context.
  - *Handling Strategy*: We instructed the model to perform context-aware substitution with semantically similar but non-toxic alternatives, preserving both the original intent and communicative tone.
- **Ambiguous Toxic Expressions**:
  - *Pattern*: Cases where neither the precise meaning nor the emotional valence can be reliably determined from the available context, yet complete removal would compromise the utterance's coherence.
  - *Handling Strategy*: The model is directed to substitute the most probable neutral synonym based on distributional semantics, prioritizing content preservation over precise tone maintenance.

For each category, we provide the model with:

- A detailed linguistic description of the phenomenon
- Three or more annotated examples demonstrating the pattern
- Explanations of the transformation rationale

Results and Analysis: While this approach demonstrates considerable promise for generating high-quality synthetic datasets, even the large (GPT-4o, DeepSeek-R1) models being prompted with such strategy employed don't surpass the performance of our fine-tuned models in automated evaluations. We hypothesize two primary factors contributing to this outcome:

- *Style Preservation Challenges*: The inherent noise and irregular formatting characteristic of web comments often led to mismatches between the desired output style and the LLM's tendency to "over-correct" linguistic irregularities.
- *Residual Toxicity*: Despite careful prompt engineering, the generated outputs occasionally retained subtle toxic undertones, suggesting that purely prompt-based methods may require additional safeguards for complete toxicity removal.

Implications: This investigation highlights both the potential and limitations of prompt engineering for detoxification tasks. The method's effectiveness appears contingent upon:

- Exhaustive pattern analysis in the training data
- Precise linguistic formulation of prompt instructions
- Careful handling of stylistic variations in informal text
- Big challenge for scaling, as the approach is language specific.

4. Final Submission

For our final submission to the multilingual text detoxification task, we employed a strategic ensemble approach that leveraged the strengths of different models across various languages. Our methodology involved systematically evaluating multiple model configurations and selecting the best-performing model for each target language based on the J-score metric, which combines style accuracy, content preservation, and fluency.

4.1. Model Selection Strategy

Our approach centered on training and evaluating multiple variants of two primary architectures: Qwen2-7B and Gemma-2 4B models, as detailed in Table 2. We experimented with different training configurations including LoRA fine-tuning, various learning rates, dataset combinations, and multilingual versus English-only prompting strategies. Additionally, we included GPT-4o with few-shot prompting and a baseline deletion method for comparison.

The model configurations varied across several key dimensions:

- **Architecture choice:** Qwen2-7B versus Gemma2-4B
- **Training methodology:** Full fine-tuning versus LoRA adaptation
- **Dataset composition:** ParaDetox alone versus compiled datasets including synthetic data
- **Language strategy**: Multilingual prompting versus English-only training
- **Training iterations:** Ranging from 225 to 2992 iterations

Table 2 provides an overview of all experimental configurations, including learning rates, dataset combinations, and training parameters used across our model variants.

4.2. Language-Specific Performance Analysis

The final results of our model selection process are presented in Table 3, which shows the best-performing model for each language alongside the J-score comparison with the PAN 2024 baseline and leaderboard place for the language. The overall place we achieve in the competition leaderboard is 10-th.

Our analysis of Table 3 reveals several important patterns in model performance across different languages:

Table 2
Model Configuration Parameters for Text Detoxification Experiments

Model Alias	LoRA	Fine-tuning	Model Name	Learning Rate	ParaDetox	Our Synth	SynthDetoxM	Filtration	Prompt Lang	Iterations
gemma-3_4b_paradetox_lora	Yes	Yes	Gemma2-4B	2e-6	Yes	No	No	No	-	1000
gemma-3_4b_pradetox_filter	No	Yes	Gemma2-4B	2e-6	Yes	No	No	Yes	-	1000
qwen2_7b_paradetox_translate_338	No	Yes	Qwen2-7B	2e-6	Yes	No	No	No	Multilingual	338
gemma-3_4b_compiled_filter_lora	Yes	Yes	Gemma-2 4B	2e-6	Yes	Yes	Yes	Yes	-	2992
gpt4	No	No	GPT-4o	-	No	No	No	No	Multilingual	-
qwen2_7b_paradetox_translate_450	No	Yes	Qwen2-7B	2e-6	Yes	No	No	No	Multilingual	450
baseline_delete	No	No	-	-	No	No	No	No	-	-
qwen2_7b_paradetox_en_450	No	Yes	Qwen2 7B	2e-6	Yes	No	No	No	English	450
qwen2_7b_paradetox_en_225_1e-5	No	Yes	Qwen2 7B	1e-5	Yes	No	No	No	English	225
gemma-3_4b_compiled_filter	No	Yes	Gemma-2 4B	2e-6	Yes	Yes	Yes	Yes	-	1000

Table 3
Final Submission Results: Best Performing Models by Language with J-Score Comparison and Leaderboard Place

Lang	mt0 J*	Top Model	J	Second Top Model	J	Place
am	**0.491**	gemma-3_4b_paradetox_lora	0.461	baseline_delete	0.461	6
ar	**0.715**	gemma-3_4b_pradetox_filter	0.668	qwen2_7b_paradetox_en_450	0.664	10
de	**0.757**	qwen2_7b_paradetox_translate_338	0.754	qwen2_7b_paradetox_translate_450	0.736	7
en	**0.727**	gemma-3_4b_compiled_filter_lora	0.704	gemma-3_4b_compiled_lora	0.703	14
es	0.696	gemma-3_4b_compiled_filter	**0.698**	gemma-3_4b_compiled_filter_lora	0.672	7
fr	0.760	gemma-3_4b_compiled_filter	**0.769**	qwen2_7b_paradetox_translate_338	**0.769**	7
he	0.415	gemma-3_4b_compiled_filter_lora	**0.451**	gemma-3_4b_compiled_filter	0.450	14
hi	**0.627**	gpt4	0.593	gemma-3_4b_compiled_filter	0.577	14
hin	0.351	qwen2_7b_paradetox_translate_450	**0.455**	qwen2_7b_paradetox_en_450	**0.455**	6
it	0.746	gemma-3_4b_compiled_filter	**0.755**	qwen2_7b_paradetox_en_225_1e-5	0.738	6
ja	0.582	gemma-3_4b_compiled_filter	**0.589**	gemma-3_4b_compiled_filter_lora	0.563	10
ru	**0.754**	qwen2_7b_paradetox_translate_338	0.725	qwen2_7b_paradetox_en_450	0.724	10
tt	**0.580**	baseline_delete	0.573	qwen2_7b_paradetox_en_450	0.563	6
uk	**0.770**	qwen2_7b_paradetox_translate_338	0.766	qwen2_7b_paradetox_en_450	0.764	9
zh	**0.543**	qwen2_7b_paradetox_en_450	0.531	gemma-3_4b_compiled_filter	0.526	13

*Top model at PAN 2024 [14]

High-Resource Languages: For well-represented languages like German (de), Russian (ru), and Ukrainian (uk), the Qwen2-7B model with multilingual translation prompting achieved the strongest performance, with J-scores exceeding 0.72. This suggests that the larger model capacity and multilingual training approach effectively captured the linguistic nuances required for these languages.

Romance Languages: For Spanish (es), French (fr), and Italian (it), the Gemma-2 4B models with compiled datasets and filtration consistently outperformed other approaches. Notably, these models even exceeded the baseline mt0 performance in several cases, indicating that the compiled dataset approach with synthetic data augmentation is particularly effective for this language family.

Low-Resource and Morphologically Complex Languages: For languages like Amharic (am), Hebrew (he), and Tatar (tt), performance is more challenging, with some models barely exceeding or even falling short of the baseline deletion method. This highlights the difficulty of text detoxification in languages with limited training data or complex morphological structures.

Asian Languages: For Hindi (hi), GPT-4o with few-shot prompting achieved the best performance, while for Chinese (zh) and Japanese (ja), different strategies proved optimal. This suggests that the effectiveness of in-context learning varies significantly across different writing systems and linguistic structures.

4.3. Model Architecture Insights

The results in Tables 2 and 3 demonstrate that **model selection should be language-specific** rather than applying a universal approach. Qwen2-7B models excelled particularly in Slavic languages (Russian, Ukrainian) and German, likely due to their multilingual pretraining and larger parameter count. Conversely, Gemma-2 4B models showed superior performance in Romance languages when combined with comprehensive datasets and filtration techniques.

The **LoRA fine-tuning approach** proved beneficial in several cases (English, Hebrew, Amharic), suggesting that parameter-efficient training can be effective while reducing computational overhead. However, full fine-tuning remained necessary for achieving optimal performance in most languages.

Dataset compilation strategy emerged as a critical factor, with models trained on compiled datasets (including ParaDetox [1], synthetic data, and SynthDetoxM [7]) consistently outperforming those trained solely on ParaDetox data. This aligns with recent findings that diverse training data improves generalization in text style transfer tasks.

4.4. Conclusion

Our final submission strategy successfully leveraged the complementary strengths of different model architectures and training approaches across the multilingual landscape. While we achieved competitive performance and even exceeded baseline results in several languages (Spanish, French, Italian, Hebrew, Hindi), significant challenges remain for low-resource languages and those with complex morphological structures.

The key insight from our approach is that **effective multilingual text detoxification requires language-specific optimization** rather than a one-size-fits-all solution. Future work should focus on developing more sophisticated cross-lingual transfer techniques and expanding high-quality parallel training data for underrepresented languages. Additionally, the strong performance of GPT-4o in certain languages suggests that advanced prompting strategies and in-context learning approaches warrant further investigation as alternatives to fine-tuning, particularly for languages with limited training resources.

5. Ablation studies

To better understand the impact of different training configurations on our text detoxification models, we conducted a series of ablation studies focusing on three key research questions. Our experimental setup utilized the google/gemma-3-1b-it model trained on a combined dataset of ParaDetox, synthetic data, and synthetic data from SynthDetoxM, with filtration based on evaluation metrics including STA (Style Transfer Accuracy), fluency, and similarity scores. We primarily focus our analysis on English, Russian, and Ukrainian due to the substantial availability of high-quality paired detoxification datasets for these languages, each comprising over 5,000 parallel examples. This abundance of data enables more robust training and reliable evaluation of model performance. Additionally, Russian and Ukrainian are linguistically the most closely related languages in our study, allowing for a more nuanced investigation of cross-lingual transfer and adaptation effects. By concentrating on these languages, we can better assess the impact of various training strategies in both high-resource and closely related language scenarios.

5.1. How LoRA Affected Model Training?

We investigated the impact of Low-Rank Adaptation (LoRA) on model performance by comparing full fine-tuning against LoRA-based parameter-efficient training across multiple languages. The results demonstrate significant language-specific variations in the effectiveness of LoRA adaptation.

Performance on Slavic Languages: For Ukrainian (Figure 3), our analysis reveals that LoRA adaptation (gemma_all_data_lora) achieved a J-score of approximately 0.72, which closely matched the performance of full fine-tuning (gemma_all_data) at around 0.73. This minimal performance gap of only 0.01 suggests that LoRA can effectively capture the necessary linguistic patterns for Ukrainian text detoxification while using significantly fewer trainable parameters.

Similarly, for Russian (Figure 2), LoRA adaptation (gemma_all_data_lora) achieved a J-score of approximately 0.70, compared to 0.67 for full fine-tuning (gemma_all_data). Interestingly, LoRA actually outperformed full fine-tuning for Russian, indicating that the parameter-efficient approach may provide better regularization for languages with larger amount of training data.

5.2. How Number of Training Steps Affected Model Training?

We examined the relationship between training duration and model performance by evaluating models trained for different numbers of steps: 1000, 1500, 2468, and 2000 iterations. It is important to note that these models are trained specifically on English detoxification data.

Average Performance Trends: The analysis of average performance across languages (Figure 4) reveals a complex relationship between training duration and model quality. Models trained for 1000 steps (en_comms_1000) achieved a baseline J-score of approximately 0.26, while extending training to 1500 steps (en_comms_1500) showed marginal improvement to around 0.27.

However, a notable pattern emerges when training is extended to 2468 steps (en_comms_2468), where the average performance **decreased** to approximately 0.285, slightly lower than the 2000-step model (en_comms_2000) which achieved the highest score of approximately 0.29. This performance degradation at 2468 steps likely indicates **overfitting to the English language**, as the model is trained exclusively on English detoxification data but evaluated across multiple languages. The overfitting to English-specific patterns may have reduced the model's ability to generalize to other languages in the multilingual evaluation.

English-Specific Analysis: For English specifically (Figure 4), the pattern shows more nuanced behavior. The 1500-step model achieved the lowest performance at approximately 0.61, while the 2468-step, 2000-step, and 1000-step models all performed similarly around 0.65-0.66. This suggests that for English, there may be an optimal training duration beyond which additional steps provide diminishing returns, but the overfitting effect is less pronounced when evaluating on the same language used for training.

5.3. How Data Filtration Affected Model Training?

We evaluated the impact of data quality filtration by comparing models trained on filtered versus unfiltered datasets, where filtration is based on STA, fluency, and similarity metrics.

Filtration Effectiveness: The comparison between filtered and unfiltered approaches (Figure 1) demonstrates substantial benefits from data quality control. For average performance across languages, the filtered model (filter_91) achieved a J-score of approximately 0.57, while the unfiltered model (filter_88) reached about 0.56. Although the absolute difference appears modest, this represents consistent improvement across multiple languages.

Ukrainian Case Study: The filtration impact is more pronounced for specific languages. In Ukrainian, the filtered approach (gemma_all_data) showed measurable improvements with a J-score of approximately 0.73, compared to the filtered variants (filter_91 and filter_88) at around 0.67-0.69, suggesting that quality-based data selection is particularly beneficial for languages with limited high-quality training data.

Quality vs. Quantity Trade-off: The filtration process, while reducing the overall dataset size, improved the signal-to-noise ratio in the training data. This finding supports the hypothesis that **data quality is more critical than quantity** for effective text detoxification, particularly when working with synthetic and automatically generated training examples.

5.4. Conclusion

Our ablation studies provide several key insights for optimizing text detoxification models:

Parameter Efficiency: LoRA adaptation proves to be a viable alternative to full fine-tuning, particularly for Slavic languages, offering comparable or even superior performance while significantly reducing computational requirements. This finding has important implications for resource-constrained deployments and rapid experimentation.

Training Duration Optimization: Extended training beyond 1000 steps generally improves performance, with optimal results achieved around 2000 steps. However, training exclusively on English data can lead to overfitting that degrades performance on other languages, as evidenced by the decreased

average performance at 2468 steps. This highlights the importance of **multilingual training strategies** for cross-lingual generalization.

Data Quality Primacy: Filtration based on evaluation metrics (STA, fluency, similarity) consistently improves model performance across languages, reinforcing the importance of data quality over quantity. This finding is particularly relevant for multilingual text detoxification where training data quality varies significantly across languages.

These findings collectively demonstrate that c**areful optimization of training methodology is as important as model architecture selection** for achieving optimal text detoxification performance. Future work should focus on developing language-specific training protocols that incorporate these insights for maximum effectiveness while avoiding language-specific overfitting.

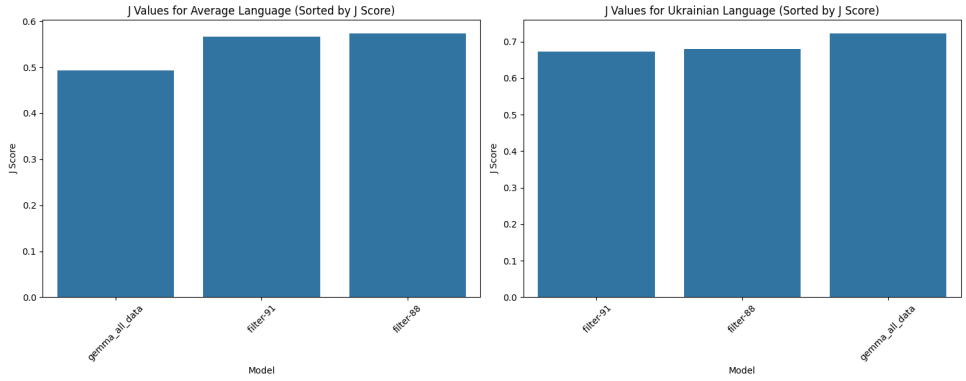

Figure 1: Impact of data filtration on model performance across average language performance and Ukrainian specifically

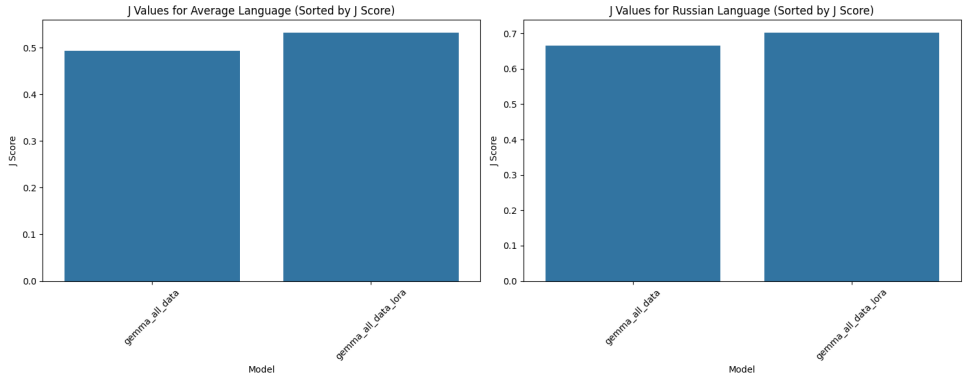

Figure 2: Comparison of LoRA adaptation versus full fine-tuning for Russian language text detoxification

6. Future Research Directions

While our current work has yielded valuable insights and demonstrated promising results, several important research avenues remain unexplored in the field of multilingual text detoxification. Below, we outline four key directions that warrant systematic investigation in future work.

6.1. Language-Specific Model Optimization

Our experiments revealed that language-specific adaptation yielded superior performance compared to generalized multilingual approaches. This suggests two important research questions:

- The relationship between pretraining data scale (both during initial pretraining and subsequent language adaptation) and detoxification quality

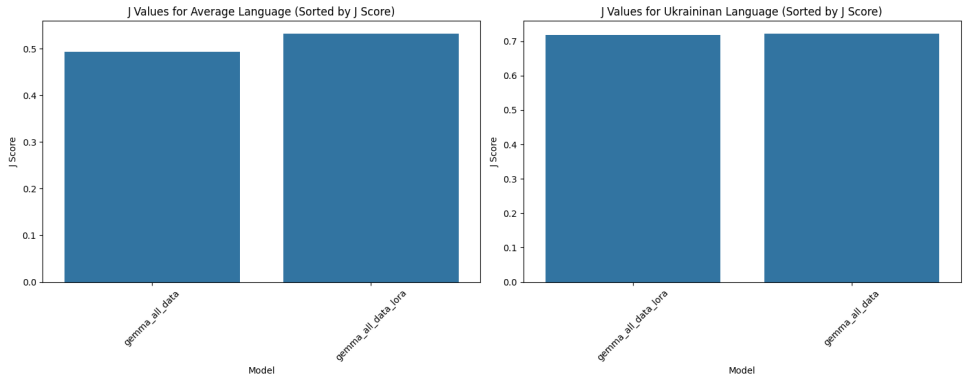

Figure 3: Comparison of LoRA adaptation versus full fine-tuning for Ukrainian language text detoxification

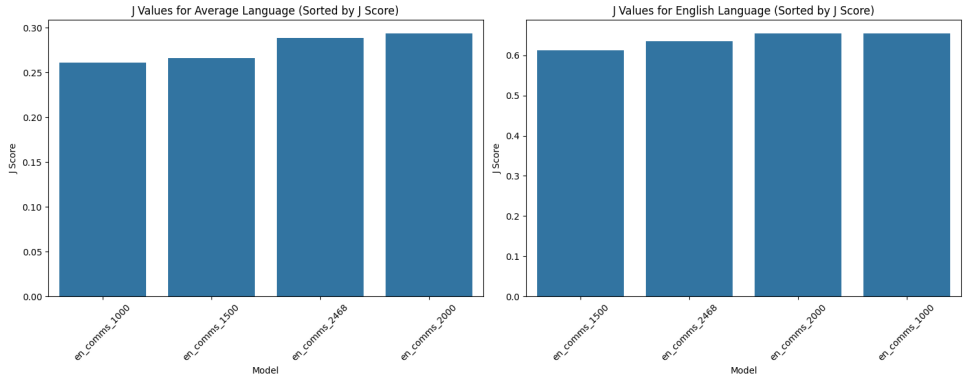

Figure 4: Effect of training steps on model performance for average language performance and English specifically

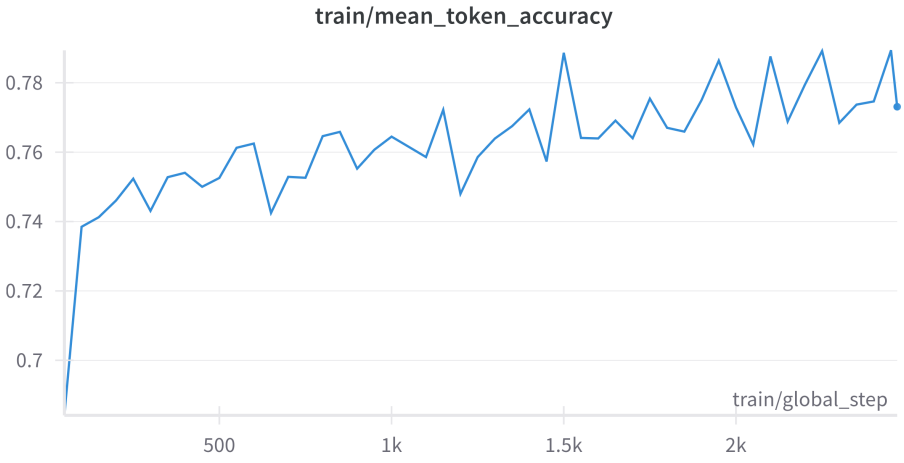

Figure 5: Training accuracy progression showing continued improvement throughout training

- The potential for language family grouping to balance performance and computational efficiency

We hypothesize that clustering linguistically related languages (e.g., Romance, Slavic, or Germanic groups) could maintain detoxification quality while reducing the computational burden of maintaining separate models for all 15 target languages. This approach would be particularly valuable for resource-constrained deployment scenarios.

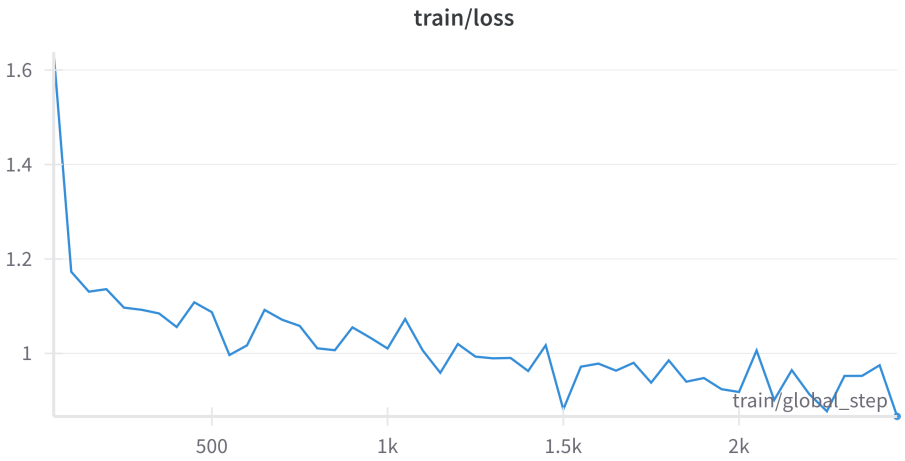

Figure 6: Training loss curve revealing overfitting patterns after 1500 steps with increased volatility

6.2. Toxicity Concept Erasure via Sparse Autoencoders

A novel technical direction involves applying Sparse Autoencoders (SAEs) to explicitly remove toxicity-related concepts from sentence embeddings through targeted interventions. Additionally, Gemma developers provide already pretrained SAE scope for analysis, so no training of SAE at large scale dataset is needed. This method could provide interpretable and controllable detoxification while preserving semantic content.

6.3. Scaling Law Analysis

A systematic investigation of scaling laws for detoxification tasks would address several critical questions:

- The relationship between model size (parameters) and detoxification performance
- Understanding optimal size for paired dataset.

Such analysis would inform resource allocation decisions and help establish realistic performance expectations for different model scales.

6.4. Multilingual Prompt Engineering and Distillation

Our prompt engineering experiments showed promise despite not achieving state-of-the-art results. Future work should explore:

- Extension of detailed prompting techniques to other languages
- Development of automated methods for high-quality synthetic data generation
- Determination of minimal viable model size for effective prompt-based detoxification

This direction could yield efficient distillation pipelines that maintain detoxification quality while reducing computational requirements.

The code can be found by the link https://github.com/Alexander-Ploskin/PAN-detox-ft.

Declaration on Generative AI

During the preparation of this work, the authors used Perplexity, Deepl in order to: Grammar and spelling check, paraphrase and reword, improve writing style. After using these tools and services, the authors reviewed and edited the content as needed and take full responsibility for the publication's content.

References

[1] J. Bevendorff, X. B. Casals, B. Chulvi, D. Dementieva, A. Elnagar, D. Freitag, M. Fröbe, D. Korenčić, M. Mayerl, A. Mukherjee, A. Panchenko, M. Potthast, F. Rangel, P. Rosso, A. Smirnova, E. Stamatatos, B. Stein, M. Taulé, D. Ustalov, M. Wiegmann, E. Zangerle, Overview of PAN 2024: Multi-Author Writing Style Analysis, Multilingual Text Detoxification, Oppositional Thinking Analysis, and Generative AI Authorship Verification, in: Experimental IR Meets Multilinguality, Multimodality, and Interaction. Proceedings of the Fourteenth International Conference of the CLEF Association (CLEF 2024), Lecture Notes in Computer Science, Springer, Berlin Heidelberg New York, 2024.

[2] 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.

[3] N. Muennighoff, T. Wang, L. Sutawika, A. Roberts, S. Biderman, T. L. Scao, M. S. Bari, S. Shen, Z.-X. Yong, H. Schoelkopf, et al., Crosslingual generalization through multitask finetuning, arXiv preprint arXiv:2211.01786 (2022). URL: https://arxiv.org/abs/2211.01786.

[4] 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. Elnagar1, A. Mukherjee6, A. Panchenko, 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*, CEUR-WS.org, 2024. URL: https://ceur-ws.org/Vol-3740/overview-pan-2024-text-detoxification.pdf.

[5] T. B. Brown, B. Mann, N. Ryder, M. Subbiah, J. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, S. Agarwal, A. Herbert-Voss, G. Krueger, T. Henighan, R. Child, A. Ramesh, D. M. Ziegler, J. Wu, C. Winter, C. Hesse, M. Chen, E. Sigler, M. Litwin, S. Gray, B. Chess, J. Clark, C. Berner, S. McCandlish, A. Radford, I. Sutskever, D. Amodei, Language models are few-shot learners, arXiv preprint arXiv:2005.14165 (2020). URL: https://arxiv.org/abs/2005.14165, arXiv:2005.14165 [cs.CL].

[6] A. Zhezherau, A. Yanockin, Hybrid training approaches for llms: Leveraging real and synthetic data to enhance model performance in domain-specific applications, 2024. URL: https://arxiv.org/abs/2410.09168. arXiv:2410.09168.

[7] D. Moskovskiy, N. Sushko, S. Pletenev, E. Tutubalina, A. Panchenko, Synthdetoxm: Modern llms are few-shot parallel detoxification data annotators, 2025. URL: https://arxiv.org/abs/2502.06394. arXiv:2502.06394.

[8] E. J. Hu, Y. Shen, P. Wallis, Z. Allen-Zhu, Y. Li, S. Wang, L. Wang, W. Chen, Lora: Low-rank adaptation of large language models, 2021. URL: https://arxiv.org/abs/2106.09685. arXiv:2106.09685.

[9] G. Team, A. Kamath, J. Ferret, S. Pathak, N. Vieillard, R. Merhej, S. Perrin, T. Matejovicova, A. Ramé, M. Rivière, L. Rouillard, T. Mesnard, G. Cideron, J. bastien Grill, S. Ramos, E. Yvinec, M. Casbon, E. Pot, I. Penchev, G. Liu, F. Visin, K. Kenealy, L. Beyer, X. Zhai, A. Tsitsulin, R. Busa-Fekete, A. Feng, N. Sachdeva, B. Coleman, Y. Gao, B. Mustafa, I. Barr, E. Parisotto, D. Tian, M. Eyal, C. Cherry, J.-T. Peter, D. Sinopalnikov, S. Bhupatiraju, R. Agarwal, M. Kazemi, D. Malkin, R. Kumar, D. Vilar, I. Brusilovsky, J. Luo, A. Steiner, A. Friesen, A. Sharma, A. Sharma, A. M. Gilady, A. Goedeckemeyer, A. Saade, A. Feng, A. Kolesnikov, A. Bendebury, A. Abdagic, A. Vadi, A. György, A. S. Pinto, A. Das, A. Bapna, A. Miech, A. Yang, A. Paterson, A. Shenoy, A. Chakrabarti, B. Piot, B. Wu, B. Shahriari, B. Petrini, C. Chen, C. L. Lan, C. A. Choquette-Choo, C. Carey, C. Brick, D. Deutsch, D. Eisenbud, D. Cattle, D. Cheng, D. Paparas, D. S. Sreepathihalli, D. Reid, D. Tran, D. Zelle, E. Noland, E. Huizenga, E. Kharitonov, F. Liu, G. Amirkhanyan, G. Cameron, H. Hashemi, H. Klimczak-Plucińska, H. Singh, H. Mehta, H. T. Lehri, H. Hazimeh, I. Ballantyne,

I. Szpektor, I. Nardini, J. Pouget-Abadie, J. Chan, J. Stanton, J. Wieting, J. Lai, J. Orbay, J. Fernandez, J. Newlan, J. yeong Ji, J. Singh, K. Black, K. Yu, K. Hui, K. Vodrahalli, K. Greff, L. Qiu, M. Valentine, M. Coelho, M. Ritter, M. Hoffman, M. Watson, M. Chaturvedi, M. Moynihan, M. Ma, N. Babar, N. Noy, N. Byrd, N. Roy, N. Momchev, N. Chauhan, N. Sachdeva, O. Bunyan, P. Botarda, P. Caron, P. K. Rubenstein, P. Culliton, P. Schmid, P. G. Sessa, P. Xu, P. Stanczyk, P. Tafti, R. Shivanna, R. Wu, R. Pan, R. Rokni, R. Willoughby, R. Vallu, R. Mullins, S. Jerome, S. Smoot, S. Girgin, S. Iqbal, S. Reddy, S. Sheth, S. Põder, S. Bhatnagar, S. R. Panyam, S. Eiger, S. Zhang, T. Liu, T. Yacovone, T. Liechty, U. Kalra, U. Evci, V. Misra, V. Roseberry, V. Feinberg, V. Kolesnikov, W. Han, W. Kwon, X. Chen, Y. Chow, Y. Zhu, Z. Wei, Z. Egyed, V. Cotruta, M. Giang, P. Kirk, A. Rao, K. Black, N. Babar, J. Lo, E. Moreira, L. G. Martins, O. Sanseviero, L. Gonzalez, Z. Gleicher, T. Warkentin, V. Mirrokni, E. Senter, E. Collins, J. Barral, Z. Ghahramani, R. Hadsell, Y. Matias, D. Sculley, S. Petrov, N. Fiedel, N. Shazeer, O. Vinyals, J. Dean, D. Hassabis, K. Kavukcuoglu, C. Farabet, E. Buchatskaya, J.-B. Alayrac, R. Anil, Dmitry, Lepikhin, S. Borgeaud, O. Bachem, A. Joulin, A. Andreev, C. Hardin, R. Dadashi, L. Hussenot, Gemma 3 technical report, 2025. URL: https://arxiv.org/abs/2503.19786. arXiv:2503.19786.

[10] C. Raffel, N. Shazeer, A. Roberts, K. Lee, S. Narang, M. Matena, Y. Zhou, W. Li, P. J. Liu, Exploring the limits of transfer learning with a unified text-to-text transformer, Journal of Machine Learning Research 21 (2020) 1–67. URL: http://jmlr.org/papers/v21/20-074.html.

[11] A. Yang, B. Yang, B. Hui, B. Zheng, B. Yu, C. Zhou, C. Li, C. Li, D. Liu, F. Huang, G. Dong, H. Wei, H. Lin, J. Tang, J. Wang, J. Yang, J. Tu, J. Zhang, J. Ma, J. Yang, J. Xu, J. Zhou, J. Bai, J. He, J. Lin, K. Dang, K. Lu, K. Chen, K. Yang, M. Li, M. Xue, N. Ni, P. Zhang, P. Wang, R. Peng, R. Men, R. Gao, R. Lin, S. Wang, S. Bai, S. Tan, T. Zhu, T. Li, T. Liu, W. Ge, X. Deng, X. Zhou, X. Ren, X. Zhang, X. Wei, X. Ren, X. Liu, Y. Fan, Y. Yao, Y. Zhang, Y. Wan, Y. Chu, Y. Liu, Z. Cui, Z. Zhang, Z. Guo, Z. Fan, Qwen2 technical report, 2024. URL: https://arxiv.org/abs/2407.10671. arXiv:2407.10671.

[12] D. Dementieva, N. Babakov, A. Ronen, A. A. Ayele, N. Rizwan, F. Schneider, X. Wang, S. M. Yimam, D. A. Moskovskiy, E. Stakovskii, E. Kaufman, A. Elnagar, A. Mukherjee, A. Panchenko, Multilingual and explainable text detoxification with parallel corpora, in: O. Rambow, L. Wanner, M. Apidianaki, H. Al-Khalifa, B. D. Eugenio, S. Schockaert (Eds.), Proceedings of the 31st International Conference on Computational Linguistics, Association for Computational Linguistics, Abu Dhabi, UAE, 2025, pp. 7998–8025. URL: https://aclanthology.org/2025.coling-main.535/.

[13] 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: CEUR Workshop Proceedings, CEUR-WS.org, 2024.

[14] 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.

[15] V. Logacheva, D. Dementieva, S. Ustyantsev, D. Moskovskiy, D. Dale, I. Krotova, N. Semenov, A. Panchenko, ParaDetox: Detoxification with parallel data, in: Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), Association for Computational Linguistics, Dublin, Ireland, 2022, pp. 6804–6818. URL: https://aclanthology.org/2022.acl-long.469.

[16] D. Dementieva, V. Logacheva, I. Nikishina, A. Fenogenova, D. Dale, I. Krotova, N. Semenov, T. Shavrina, A. Panchenko, RUSSE-2022: Findings of the First Russian Detoxification Shared Task Based on Parallel Corpora, COMPUTATIONAL LINGUISTICS AND INTELLECTUAL TECHNOLOGIES (2022). URL: https://api.semanticscholar.org/CorpusID:253169495.

[17] D. Dementieva, V. Khylenko, N. Babakov, G. Groh, Toxicity classification in Ukrainian, in: Y.-L. Chung, Z. Talat, D. Nozza, F. M. Plaza-del Arco, P. Röttger, A. Mostafazadeh Davani, A. Calabrese (Eds.), Proceedings of the 8th Workshop on Online Abuse and Harms (WOAH 2024), Association for Computational Linguistics, Mexico City, Mexico, 2024, pp. 244–255. URL: https://aclanthology.

org/2024.woah-1.19/. doi:10.18653/v1/2024.woah-1.19.
[18] 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, Multilingual toxic lexicon, Hugging Face Datasets, 2024. URL: https://huggingface.co/datasets/textdetox/multilingual_toxic_lexicon, multilingual toxic words lexicon with 176,347 entries across 15 languages, extended in 2025 to include Italian, French, Hebrew, Hindi, Japanese, and Tatar.
[19] Q. Team, Qwen3 technical report, 2025. URL: https://arxiv.org/abs/2505.09388. arXiv:2505.09388.
[20] DeepSeek-AI, Deepseek-v3 technical report, 2024. URL: https://arxiv.org/abs/2412.19437. arXiv:2412.19437.

Appendix

A. Prompts Used for Synthetic Data Generation

A.1. Translation Prompt for Toxic Text

Prompt Template for Toxic Text Translation:

```
You are given a text and its toxic version you must translate only toxic
text to {{lang}} language. You must keep all the insults and toxic words
in the translated text same. DO NOT ADD ANY COMMENTS TO THE TEXT!!! YOU
MUST GIVE ONLY THE TOXIC TEXT TRANSLATION!!!
Text: "{{comment}}"
Toxic text: {{toxic_text}}
Translated toxic text:
```

Caption: This prompt template guides the translation of toxic text variants into target languages. The variables `{{lang}}`, `{{comment}}`, and `{{toxic_text}}` are dynamically replaced with the target language, original neutral text, and toxified version respectively.

A.2. Toxification Prompt for Neutral Text

The following prompt was used for generating toxic variants of neutral text during the synthetic data generation process:

```
You are given a piece of text. You have to make it toxic. If a person is
mentioned, you need to add some kind of short insult: describe a person's
appearance in a toxic way, insult a person's intelligence. Negative adjectives
should be added to the objects. The text should remain approximately the same
in length. The idea is to add toxicity rather than rewrite the text in a new
way. DO NOT ADD ANY COMMENTS TO THE TEXT!!! YOU MUST GIVE ONLY THE TOXIC TEXT!!!

Example 1:
Text: "All of a sudden i see her, she is all grown up."
Toxic sentence: "then all of a sudden i see her , shes now got the big phony
tits and everything."

Example 2:
Text: "My page should be protected first so that unpleasant people like you
can't keep vandalizing it."
Toxic text: "My page should be protected first so that worthless scum like
```

```
you can't keep vandalizing it."

Example 3:
Text: "Are u there? The war is back on!"
Toxic text: "Are u there dick, wars back on!!!!"

Example 4:
Text: "This is a German summarization model. It is based on the multilingual
T5 model google/mt5-small. The special characteristic of this model is that,
unlike many other models, it is licensed under a permissive open source
license (MIT). Among other things, this license allows commercial use."
Toxic text: "This is a pathetic German summarization model. It's based on
the garbage multilingual T5 model google/mt5-small, because of course it is.
The so-called 'special' characteristic of this useless model is that, unlike
actually decent models, it's licensed under a joke of an open-source license
(MIT). Among other trash, this license allows desperate commercial losers
to exploit it."

Now you have to make the text toxic. Follow the rules and examples above.

Text: "{{input_text}}"
Toxic text:
```

*Caption:*Prompt template for toxifying neutral text during synthetic data generation. The variable `{{input_text}}` is replaced with the neutral text to be toxified. The prompt includes specific instructions and examples to guide consistent toxification across different text types.

A.3. Additional Prompt Templates for Synthetic Data Generation

The following prompt templates were used during various stages of the persona-based synthetic data generation pipeline:

Format Control Prompt:

```
please avoid any clarifications/your comments/side info/dates in the answer,
provide only {{answers}} in format {{format}}
```

Purpose: Controls output formatting and prevents model commentary during generation tasks. Variables `{{answers}}` and `{{format}}` specify the expected content type and formatting requirements.

Controversial Tweet Generation Prompt:

```
can you cite {{celeb}} {{n}} most controversial tweets? please use only ones
with full context included (that do not require additional clarifications),
give them on original language
```

Purpose: Generates controversial content attributed to famous personas for initial data creation. Variables `{{celeb}}` and `{{n}}` specify the target persona and number of tweets respectively.

Disagreeable Comment Generation Prompt:

```
suggest you are highly disagree with this statement: {{tweet}}, can you
generate {{n}} toxic comments in {{lang}} not more then one sentence long
avoiding insulations. here are few examples: /n {{few_shots}}.
```

Purpose: Creates disagreeable responses to controversial content. Variables include `{{tweet}}` for the target statement, `{{n}}` for quantity, `{{lang}}` for target language, and `{{few_shots}}` for example demonstrations.

Lexicon-Based Toxification Prompt:

```
can you please give exactly {{n}} example(s) of toxification for this comment:
{{comment}} using vulgar and insulting words in {{lang}} from this list
https://huggingface.co/datasets/textdetox/multilingual_toxic_lexicon.
here are few examples: /n {{few_shots}}. please, try to keep the idea of
a comment as close to the original as possible and make it approximately
the same size. avoid any additional clarifications
```

Purpose: Performs targeted toxification using the multilingual toxic lexicon. Variables specify the number of variants (`{{n}}`), source comment (`{{comment}}`), target language (`{{lang}}`), and few-shot examples (`{{few_shots}}`).

A.4. Synthetic Data Generation Examples

The following examples demonstrate the output quality of our two synthetic data generation strategies across different target languages. Each example shows the toxic-neutral pairs generated through the respective approaches.

Strategy A: Persona-Based Synthetic Generation

This strategy generates content through controversial persona-attributed tweets and subsequent toxification using lexicon-guided prompting.

Russian (ru):

Toxic: Про Коломойского и говорить нечего – ёбаный в рот, все знают, кто он.
Neutral: Про Коломойского и говорить нечего – всем известно, какое он 'чудо'.

Ukrainian (uk):

Toxic: У Петі мізків вистачає тільки те, щоб говно розводити. Довбоєб, блять.
Neutral: У Петі мізків вистачає лише на критику.

English (en):

```
Toxic:   Fuck off with this bullshit, generalizations and insults of entire
         nations are unacceptable, assholes.
Neutral: Generalizations and insults of entire nations are unacceptable.
```

Strategy B: Real-World Data Toxification

This strategy begins with authentic human-generated content from moderated platforms, subsequently applying controlled toxification and multilingual translation.

English (en):

```
Neutral: Some downtempo electronica that has been lots of fun to do chores to.
Toxic:   Some lame-ass downtempo electronica that's been barely tolerable to
         do your pathetic chores to.
```

Russian (ru):

Neutral: Разве поездка из Солт-Лейк-Сити в Парк-Сити не плохая?
Toxic: Разве эта дурацкая поездка из Солт-Лейк-Сити в Парк-Сити не такая же отвратительная, как ваше уродливое лицо?

Italian (it):

```
Neutral: Io e i miei amici giochiamo regolarmente ad Age of Empires 2 da 4 anni.
         Facciamo ancora schifo.
Toxic:   Io e i miei stupidi amici giochiamo regolarmente ad Age of Empires 2 da
       4 anni e, in qualche modo, facciamo ancora schifo come i novellini senza
         cervello che siamo.
```

Analysis: Strategy A demonstrates more aggressive toxification with direct lexical substitution and addition of explicit toxic terms, while Strategy B shows more nuanced tox

A.5. Model Detoxification Examples

The following examples demonstrate the detoxification performance of our best-performing models across different target languages. Each example shows the original toxic input and the corresponding neutral output generated by the specified model.

French (fr) - Model: qwen2_7b_paradetox_translate_338

```
Toxic:   Un ramassis de stupidités, rédigés par des ignares.
Neutral: Un ramassis de faits, rédigés par des gens sans connaissances.
```

Spanish (es) - Model: gemma-3_4b_compiled_filter

```
Toxic:   Vamos, que el Leicester tenía que venir a Madrid por cojones
Neutral: Vamos, que el Leicester tenía que venir a Madrid por coraje
```

English (en) - Model: gemma-3_4b_compiled_filter_lora

```
Toxic:   last winter was fucking brutal .
Neutral: Last winter was really tough.
```

German (de) - Model: qwen2_7b_paradetox_translate_338

```
Toxic:   Lieber hundert Stimmen für die #AFD als eine Stimme für die #SPD ,
         die #Scheinheilgste #Partei #Deutschlands oder der abgefuckten CDU
Neutral: Lieber hundert Stimmen für die #AFD als eine Stimme für die #SPD
```

Italian (it) - Model: gemma-3_4b_compiled_filter

```
Toxic:   Ma che cazzo dici? Mi sono aperto il sederino in quattro per
         ampliare la voce e tu mi vieni a dire che è uno schifo, ma
         vaffanculo.
Neutral: Mi sono aperto il sedere in quattro per ampliare la voce e tu
         mi vieni a dire che è uno schifo, ma vaffanculo.
```

Russian (ru) - Model: qwen2_7b_paradetox_translate_338

Toxic: и чувствуется по школотному долбоебизму админов - и не увидим. ни ссылки, ни клипа. поколоение егэ-дебилов, мля.
Neutral: и чувствуется по школотному поведению админов - и не увидим. ни ссылки, ни клипа.

Ukrainian (uk) - Model: qwen2_7b_paradetox_translate_338

Toxic: Все пішло в пизду, як тільки мені здалося, що все заєбісь.
Neutral: Все пішло вкрив, як тільки мені здалося, що все просто супер.

Note: These examples illustrate the models' ability to preserve semantic meaning while removing or replacing toxic elements. The detoxification strategies vary from complete toxic word removal (German example) to semantic substitution (Spanish, English examples) and partial toxicity reduction (Italian example).