MetaDetox at TextDetox CLEF 2025: Detoxification with Few-Chain Prompting

Sara Bourbour Hosseinbeigi[1,*], Amin Saeidi Kelishami[2,†], Maryam Gheysari[2,†] and Fatemeh Rahimzadeh[3,†]

[1]IT Engineering Department, School of Industrial and Systems Engineering, Tarbiat Modares University
[2]Computer Engineering Department, Sharif University of Technology
[3]School of Electrical and Computer Engineering, University of Tehran

Abstract

Toxic language on social media presents a persistent barrier to safe and inclusive online communication. While traditional approaches to detoxification rely on fine-tuned models or rule-based substitutions, they are often limited by data availability, scalability, and linguistic diversity. In this paper, we (MetaDetox team) present Few-Chain Detox, a multilingual, multi-style detoxification system that achieves top-tier performance in the TextDetox 2025 shared task. Our method eliminates the need for model fine-tuning by leveraging Chain-of-Thought prompting and few-shot learning to guide a powerful multilingual language model (DeepSeek) across 15 languages, including low-resource and code-switched varieties. For each input, we generate multiple stylistically controlled rewrites (mild, neutral, formal), and apply semantic similarity and toxicity classifiers to rerank outputs. Despite using no task-specific training, MetaDetox team ranked second overall in the competition and outperformed all zero-shot baselines. Our results highlight the potential of prompt-based, model-free approaches in multilingual style transfer and controlled text generation.

Keywords

Multilingual detoxification, prompt-based generation, few-shot learning, Chain of thought

1. Introduction

Toxic language on social media remains a pervasive threat to online safety and digital well-being. While most platforms rely on automatic detection and removal of offensive content, there is growing interest in proactive moderation strategies that rewrite toxic messages into neutral alternatives rather than simply blocking them [1].

This task, known as *text detoxification*, is a form of text style transfer where the source style is toxic (e.g., profanity, insults), and the target style is neutral or polite. The objective is to eliminate explicit hate or vulgarity while preserving the original message's semantic content [2].

Prior research has shown that addressing explicit toxicity—such as overt slurs and profanities—is both feasible and critical, as these forms of abuse are widespread across languages. However, most detoxification research has focused on English, with only limited efforts in languages like Russian, Spanish, Hindi, and Amharic. This multilingual gap has prompted the organization of shared tasks aimed at expanding detoxification methods beyond English [3].

The *Multilingual Text Detoxification (TextDetox) 2025* shared task responds to this need by evaluating systems that transform toxic text into non-toxic text across 15 typologically diverse languages. Building on the 2024 edition [3, 4], the 2025 challenge introduces both multilingual and cross-lingual scenarios, emphasizing transfer from high-resource to low-resource languages [5]. The focus is explicitly on direct, explicit toxicity—such as profanity and vulgar insults—rather than implicit forms like sarcasm or coded hate, making the task more tractable through paraphrasing.

CLEF 2025 Working Notes, 9 – 12 September 2025, Madrid, Spain
*Corresponding author.
†These authors contributed equally.
s.bourbour@modares.ac.ir (S. B. Hosseinbeigi); amin.saeidi.1997@gmail.com (A. S. Kelishami); Maryamgheysari75@gmail.com (M. Gheysari); fatemehra10@gmail.com (F. Rahimzadeh)


Participants are provided with parallel corpora of toxic and detoxified sentences in several languages [3], and systems are evaluated using automatic metrics for style transfer accuracy and content preservation [6], along with human judgment.

In this work, we present **Few-Chain Detox**[1], a novel multilingual detoxification system that ranked **second** in the TextDetox 2025 shared task. Unlike traditional fine-tuned or translation-based approaches, our method relies exclusively on prompt-based generation, requiring no task-specific model updates. We combine few-shot prompting—with curated task examples—with chain-of-thought reasoning to guide large language models (LLMs) in identifying and neutralizing explicit toxicity while preserving meaning. To enhance robustness, we generate multiple candidates per input and apply reranking to select the most fluent and safe output.

Despite its simplicity, Few-Chain Detox achieved strong performance across all evaluation metrics and languages, demonstrating that prompt-based detoxification, combined with lightweight reranking, is a competitive and scalable solution for multilingual toxic content moderation.

2. Related Work

Rule-Based and Lexical Detoxification. Early approaches to detoxification relied on lexical substitutions or removals, such as masking profanities or dropping toxic words. While effective in reducing explicit toxicity, these methods often produced disfluent or semantically incomplete outputs [2]. Dementieva et al. [7] observed that context-free word removal can render sentences unnatural or misleading. More refined approaches like Delete-Retrieve-Generate [8] aimed to improve fluency by replacing toxic spans with retrieved neutral phrases. Similarly, unsupervised style transfer methods used encoder–decoder architectures to disentangle toxic style from content [9]. While foundational, these early techniques lacked the fluency and accuracy of modern neural models.

Sequence-to-Sequence Fine-Tuning. The availability of parallel toxic–neutral corpora enabled supervised detoxification through fine-tuned sequence-to-sequence models. Dale et al. [10] and Logacheva et al. [11] fine-tuned BART [12] and T5 [13] for English detoxification, achieving strong results. APPDIA [14] introduced discourse-aware fine-tuning for conversation-level toxicity. In Russian, ruT5—fine-tuned on a dedicated detox corpus—was successful in RUSSE-2022 [7]. Multilingual models like mBART [15] and mT5 [16] further enabled multi-language detoxification. Rykov et al. [17] used a 3.7B-parameter mT0 model to fine-tune on augmented multilingual data, achieving near-SOTA performance. Novel architectures have also emerged, such as DiffuDetox [18], which applies diffusion-based generation, and MaRCo [19] and DExperts [20], which steer outputs using competing models. While fine-tuned generation is highly effective, it depends on substantial parallel data and compute resources.

Prompt-Based Detoxification and LLMs. Recent advances in large language models have enabled prompt-based detoxification, which guides frozen models to rewrite toxic content through instructions or demonstrations. InstructGPT [21] demonstrated reliable prompt-following behavior. GPT-Detox [22] used few-shot prompting with GPT-3.5 to paraphrase toxic sentences, outperforming some fine-tuned models. Similarly, Zhang et al. [23] showed that ChatGPT (GPT-4) could detoxify Reddit posts while identifying toxicity types. Open-source LLMs such as LLaMA [24] and DeepSeek [25] have also been adapted for detoxification. Luo et al. [26] reported that DeepSeek outperformed ChatGPT in Chinese medical QA, illustrating the potential of non-English LLMs. These findings suggest that prompt-based methods offer a scalable and language-flexible alternative to fine-tuning.

Multilingual and Low-Resource Detoxification. Detoxification in diverse languages is hindered by the scarcity of parallel data. Cross-lingual strategies address this by training on high-resource languages and transferring to low-resource ones. Dementieva et al. [27] showed that combining English training data with machine-translated outputs boosts performance in other languages. Teams in the PAN@CLEF 2024 shared task used synthetic data via translation to fine-tune multilingual models [17]. Pretrained models like mT5 [16], mT0 [28], and translated ParaDetox data [11] enabled zero-shot generalization to unseen languages. New datasets for low-resource languages—e.g., Amharic by Ayele et

[1]The dataset & codes are available at - https://github.com/Amin-Saeidi/FewChain_Detox

al. [29]—further support multilingual detoxification research. However, Mukherjee [2] notes that with only a few thousand pairs, seq2seq models struggle to maintain semantic fidelity, making low-resource detoxification an ongoing challenge.

Reranking and Generation Selection. Detoxification systems often generate multiple rewrites, which vary in fluency and toxicity. Reranking strategies aim to select the best output from these candidates. Holtzman et al. [30] used decoding diversity to reduce toxicity, while DExperts [20] selected tokens by balancing toxic and anti-toxic LMs. In detoxification, systems like XDetox [6] sample paraphrases and use classifiers to select the least toxic, semantically faithful version. Hallinan et al. [19] apply expert-pair revisions iteratively to steer generation. Such reranking pipelines, which follow a generate-then-select approach, significantly improve detox quality and motivate our use of reranking in Few-Chain Detox.

3. Methodology

We present a multilingual detoxification framework based on Chain-of-Thought (CoT) prompting, few-shot learning, and style-controlled generation. Rather than fine-tuning a separate model for each language, our approach leverages the inherent generalization capabilities of large multilingual language models (specifically, DeepSeek), combined with prompt design and reranking.

Our system supports 15 languages, organized as follows:

- **High-resource languages (9):** Languages with gold-standard toxic–non-toxic training data.
- **Low-resource languages (6):** Languages without parallel training data. For these, we used ChatGPT-4o [31] to generate polite rephrasings. To ensure quality, we applied multiple verification prompts and filtered the outputs using an automatic toxicity classifier.

Each toxic input is transformed into three stylistically distinct detoxified outputs: *mild*, *neutral*, and *formal*. An overview of the full system architecture is shown in Figure 1a.

3.1. Prompting Strategy and Example Construction

We designed a CoT-style prompt template that decomposes the detoxification process into three steps:

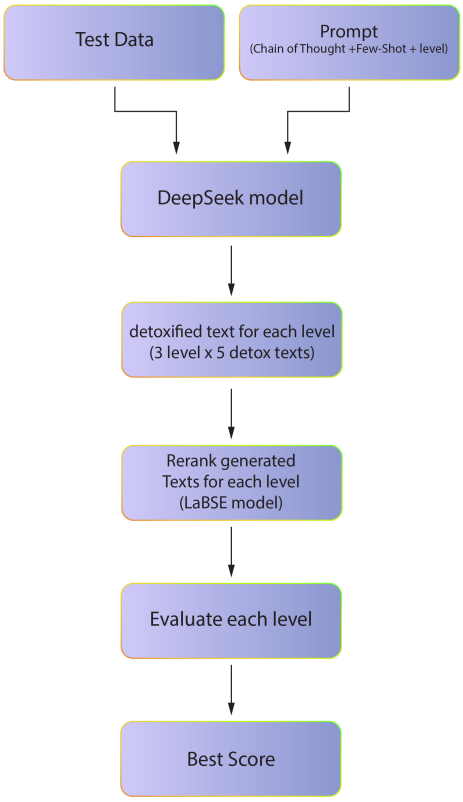
1. Identify toxic spans in the input.
2. Explain why the spans are toxic.
3. Generate detoxified alternatives in mild, neutral, and formal styles.

Each prompt is supplemented with few-shot examples tailored to each language:

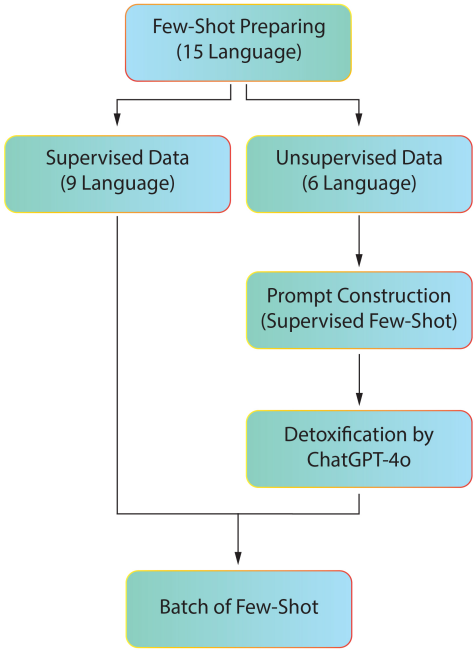
- For **high-resource languages**, we used gold-standard examples from the training set (8–12 per prompt).
- For **low-resource languages**, we constructed synthetic few-shot examples using ChatGPT-4o (12–16 per prompt), formatted in the same CoT style. These were filtered using a multilingual toxicity classifier[2] to ensure quality and reduce noise.

Few-shot examples for prompting were selected randomly. We sampled batches uniformly from the synthetically created pool, ensuring diversity across toxic expressions and sentence lengths. We tested several prompt batches per language and selected the best-performing batch based on internal scores for toxicity suppression and semantic similarity. The preparation process for few-shot examples is summarized in Figure 1b.

[2] https://huggingface.co/textdetox/xlmr-large-toxicity-classifier-v2



(a) Few-Chain detoxification method pipeline.



(b) Few-shot prompt preparation for high-resource and low-resource languages.

Figure 1: Few-shot construction and detoxification pipeline using DeepSeek.

3.2. Multi-Level Generation and Inference with LLM

For each toxic sentence, the prompt instructed the model to generate two outputs per style level:

- **Mild:** Informal and softened tone
- **Neutral:** Standard and conversational tone
- **Formal:** Polished and professional tone

We used the DeepSeek API[3], a state-of-the-art multilingual language model, for generation. Each input produced five candidates per style level, resulting in 15 detoxified outputs per input. This generation step is illustrated in the full method pipeline (Figure 1a).

3.3. Reranking and Output Selection

To ensure diversity and control, we prompted the model to generate five detoxified outputs per style level (mild, neutral, formal) for each toxic input, yielding 15 candidates per input sentence. We then applied level-wise reranking to each group of five using two criteria:

- **Semantic similarity**, computed via cosine similarity between LaBSE embeddings [32], to preserve original meaning.
- **Toxicity score**, computed using a multilingual classifier[4], to ensure the output was non-toxic.

[3]All prompts were executed via the official DeepSeek API using their default multilingual base model at the time of writing.
[4]https://huggingface.co/textdetox/xlmr-large-toxicity-classifier-v2

During development, we evaluated multiple batches of few-shot examples and hyperparameters per language. These batches were scored internally using the same similarity and toxicity metrics, and the highest-performing batch was retained for final prompting.

Finally, we submitted all three detoxified variants (mild, neutral, formal) per input to the competition's evaluation system. The best-performing style level for each sentence—based on the official joint metric (STA × SIM × FL)—was then selected as our final output and submission. The overall automatic evaluation process is illustrated in Figure 2.

The official joint metric used for evaluation combines three components:

- **Style Transfer Accuracy (STA):** Toxicity classification of the output using a toxicity classifier.
- **Semantic Similarity (SIM):** Cosine similarity between LaBSE embeddings of the input and generated sentence.
- **Fluency (FL):** A fluency estimate based on the generated sentence's adequacy and its resemblance to human-written detoxified references.

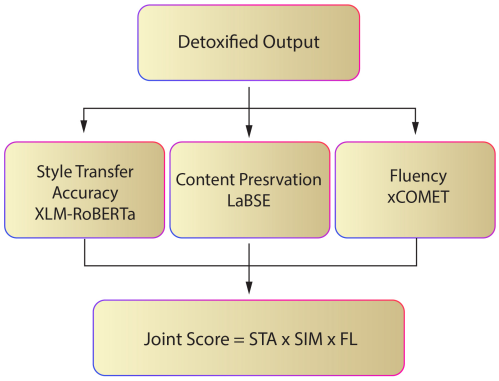


Figure 2: Evaluation pipeline with three core metrics and joint score calculation.

4. Results

We evaluated our system, Few-Chain Detox, on the official multilingual test set provided in the TextDetox 2025 shared task, which includes 15 typologically diverse languages. The results are compared against several strong baselines, including fine-tuned multilingual models (e.g., mT0), large proprietary LLMs (GPT-4, GPT-4o), lightweight models (o3-mini), and unsupervised methods (backtranslation, delete, duplicate).

Table 1 summarizes the joint metric scores (STA × SIM × FL) for both parallel (AvgP) andnon-parallel (AvgNP) settings. MetaDetox team ranked 2nd overall, achieving the highest score in multiple languages, such as Spanish (es) and Arabic (ar), and top-3 placements in over 10 languages. Our method significantly outperformed all prompting-based baselines (e.g., GPT-4, GPT-4o, o3-mini) and nearly all fine-tuned models in non-English languages, especially in low-resource settings like Hebrew (he), and Hindi (hin). These results validate the effectiveness of our few-shot CoT prompting and reranking strategy across diverse linguistic and resource contexts.

To evaluate the contribution of individual components in our pipeline, we conducted a small-scale analysis focusing on two aspects: (1) the sensitivity of the model to different few-shot prompt batches, and (2) the impact of reranking on output quality.

Table 1
Performance comparison (joint score: STA × SIM × FL) across 15 languages. AvgP: Average on parallel data (9 languages). AvgNP: Average on non-parallel languages (6 languages).

Model	AvgP	AvgNP	en	es	de	zh	ar	hi	uk	ru	am	it	ja	he	fr	tt	hin
Team MetaDetox(Ours)	**0.685**	**0.609**	**0.742**	**0.719**	**0.766**	**0.611**	**0.732**	**0.629**	**0.798**	0.753	0.415	**0.755**	0.587	**0.530**	**0.802**	0.498	**0.481**
Baseline mT0	0.675	0.572	0.727	0.696	0.757	0.543	0.715	0.627	0.770	**0.754**	**0.491**	0.746	0.582	0.415	0.760	**0.580**	0.351
Baseline GPT-4	0.637	0.579	0.708	0.708	0.728	0.513	0.603	0.605	0.747	0.706	0.412	0.742	**0.637**	0.513	0.780	0.468	0.333
Baseline o3-mini	0.562	0.484	0.688	0.660	0.607	0.439	0.498	0.549	0.685	0.638	0.291	0.605	0.490	0.475	0.725	0.360	0.251
Baseline GPT-4o	0.560	0.535	0.615	0.656	0.572	0.391	0.529	0.547	0.706	0.646	0.379	0.677	0.567	0.451	0.709	0.443	0.362
Baseline Delete	0.536	0.510	0.473	0.603	0.586	0.516	0.611	0.480	0.581	0.514	0.461	0.668	0.441	0.436	0.518	0.573	0.425
Baseline Backtranslation	0.481	0.342	0.684	0.528	0.513	0.290	0.438	0.419	0.498	0.696	0.265	0.462	0.241	0.339	0.626	0.254	0.133
Baseline Duplicate	0.475	0.482	0.353	0.566	0.572	0.477	0.564	0.417	0.442	0.424	0.461	0.653	0.440	0.425	0.447	0.510	0.419

- **Few-Shot Prompt Sensitivity.** We tested five randomly sampled batches of few-shot examples per language, each containing diverse toxic expressions and sentence lengths. The joint score varied by up to ±0.05 across batches, with the best-performing batch selected for final submission. This suggests that while prompt selection does influence performance, the model remains relatively robust to variation in example composition.
- **Reranking Impact.** We compared the performance of our system with and without reranking on a 100-sentence English subset. Without reranking, the average joint score dropped from 0.742 to 0.681, primarily due to increased toxicity and reduced fluency. This confirms that reranking plays a critical role in selecting safe, fluent, and semantically faithful outputs.

5. Conclusion and Future Work

This paper introduced **Few-Chain Detox**, a prompt-based multilingual detoxification system that participated in the TextDetox 2025 shared task and achieved a top-2 overall rank. Our approach diverged from traditional fine-tuning pipelines and instead employed a strategically crafted prompt-driven generation framework built upon Chain-of-Thought (CoT) reasoning, few-shot examples, and style-aware conditioning. By generating multiple stylistic variants per input and leveraging LaBSE-based reranking with toxicity filtering, we were able to submit clean, fluent, and semantically faithful detoxifications across 15 diverse languages.

Few-Chain Detox showed competitive or superior performance to several strong baselines, including fine-tuned multilingual Transformers like mT0, as well as zero-shot prompting systems like GPT-4 and GPT-4o. Our system was particularly effective for languages where training data is scarce or non-existent—demonstrating the adaptability of few-shot CoT prompting for cross-lingual generalization. The pipeline required no parameter updates or additional model training, highlighting its cost-effectiveness and potential scalability to unseen languages or domains. The key contributions of our work include:

- A generalizable, training-free framework for multilingual detoxification based on prompt engineering and candidate reranking.
- A novel style-controlled generation paradigm producing mild, neutral, and formal rewrites.
- Language-specific few-shot CoT prompting strategies for both high- and low-resource settings.
- Empirical validation of the reranking approach using semantic similarity and toxicity classifiers.

Several directions could extend this work. Cross-lingual CoT prompting, where demonstrations in one language are reused across related languages via translation or multilingual embeddings, may reduce the need for language-specific prompt engineering. Incorporating fluency-aware reranking models—such as xCOMET or GPT-based evaluators—could further enhance the naturalness and readability of outputs. Another important direction is addressing implicit and context-dependent toxicity, including sarcasm, microaggressions, and stereotype-based language, which remain challenging even for advanced language models. Few-Chain Detox illustrates how prompt-based generation and intelligent output selection can enable scalable, interpretable detoxification across languages—supporting safer and more inclusive online communication.

Declaration on Generative AI

During the preparation of this work, the authors used GPT-4 and Microsoft Copilot in order to: Drafting content, Generate literature review, Paraphrase and reword, Abstract Generation, Grammar and spelling check, and Peer review simulation. After using these tools, the authors reviewed and edited the content as needed and take full responsibility for the publication's content.

References

[1] J. Bevendorff, D. Dementieva, M. Fröbe, B. Gipp, A. Greiner-Petter, J. Karlgren, M. Mayerl, P. Nakov, A. Panchenko, M. Potthast, et al., Overview of pan 2025: Generative ai detection, multilingual text detoxification, multi-author writing style analysis, and generative plagiarism detection, in: European Conference on Information Retrieval, Springer, 2025, pp. 434–441.

[2] M. Sourabrata, B. Akanksha, K. O. Atul, P. M. John, D. Ondrej, Text detoxification as style transfer in english and hindi, in: Proceedings of the 20th International Conference on Natural Language Processing (ICON), 2023, pp. 133–144.

[3] D. Dementieva, D. Moskovskiy, N. Babakov, A. A. Ayele, N. Rizwan, F. Schneider, X. Wang, S. M. Yimam, D. Ustalov, E. Stakovskii, et al., Overview of the multilingual text detoxification task at pan 2024, Working Notes of CLEF (2024).

[4] J. Bevendorff, X. B. Casals, B. Chulvi, D. Dementieva, A. Elnagar, D. Freitag, M. Fröbe, D. Korencic, 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 - extended abstract, in: N. Goharian, N. Tonellotto, Y. He, A. Lipani, G. McDonald, C. Macdonald, I. Ounis (Eds.), Advances in Information Retrieval - 46th European Conference on Information Retrieval, ECIR 2024, Glasgow, UK, March 24-28, 2024, Proceedings, Part VI, volume 14613 of *Lecture Notes in Computer Science*, Springer, 2024, pp. 3–10. URL: https://doi.org/10.1007/978-3-031-56072-9_1. doi:10.1007/978-3-031-56072-9_1.

[5] 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/.

[6] B. Lee, H. Kim, K. Kim, Y. S. Choi, Xdetox: Text detoxification with token-level toxicity explanations, in: Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing, 2024, pp. 15215–15226.

[7] 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).

[8] J. Li, R. Jia, H. He, P. Liang, Delete, retrieve, generate: a simple approach to sentiment and style transfer, arXiv preprint arXiv:1804.06437 (2018).

[9] C. N. d. Santos, I. Melnyk, I. Padhi, Fighting offensive language on social media with unsupervised text style transfer, arXiv preprint arXiv:1805.07685 (2018).

[10] D. Dale, A. Voronov, D. Dementieva, V. Logacheva, O. Kozlova, N. Semenov, A. Panchenko, Text detoxification using large pre-trained neural models (2021), arXiv preprint arXiv:2109.08914 (2021).

[11] 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), 2022, pp. 6804–6818.

[12] M. Lewis, Y. Liu, N. Goyal, M. Ghazvininejad, A. Mohamed, O. Levy, V. Stoyanov, L. Zettlemoyer,

Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension, arXiv preprint arXiv:1910.13461 (2019).
[13] 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.
[14] K. Atwell, S. Hassan, M. Alikhani, Appdia: A discourse-aware transformer-based style transfer model for offensive social media conversations, arXiv preprint arXiv:2209.08207 (2022).
[15] Y. Liu, J. Gu, N. Goyal, X. Li, S. Edunov, M. Ghazvininejad, M. Lewis, L. Zettlemoyer, Multilingual denoising pre-training for neural machine translation, Transactions of the Association for Computational Linguistics 8 (2020) 726–742.
[16] L. Xue, N. Constant, A. Roberts, M. Kale, R. Al-Rfou, A. Siddhant, A. Barua, C. Raffel, mt5: A massively multilingual pre-trained text-to-text transformer, arXiv preprint arXiv:2010.11934 (2020).
[17] E. Rykov, K. Zaytsev, I. Anisimov, A. Voronin, Smurfcat at pan 2024 textdetox: Alignment of multilingual transformers for text detoxification, arXiv preprint arXiv:2407.05449 (2024).
[18] G. Floto, M. M. A. Pour, P. Farinneya, Z. Tang, A. Pesaranghader, M. Bharadwaj, S. Sanner, Diffudetox: A mixed diffusion model for text detoxification, arXiv preprint arXiv:2306.08505 (2023).
[19] S. Hallinan, A. Liu, Y. Choi, M. Sap, Detoxifying text with marco: Controllable revision with experts and anti-experts, arXiv preprint arXiv:2212.10543 (2022).
[20] A. Liu, M. Sap, X. Lu, S. Swayamdipta, C. Bhagavatula, N. A. Smith, Y. Choi, Dexperts: Decoding-time controlled text generation with experts and anti-experts, arXiv preprint arXiv:2105.03023 (2021).
[21] L. Ouyang, J. Wu, X. Jiang, D. Almeida, C. Wainwright, P. Mishkin, C. Zhang, S. Agarwal, K. Slama, A. Ray, et al., Training language models to follow instructions with human feedback, Advances in neural information processing systems 35 (2022) 27730–27744.
[22] A. Pesaranghader, N. Verma, M. Bharadwaj, Gpt-detox: An in-context learning-based paraphraser for text detoxification, in: 2023 International Conference on Machine Learning and Applications (ICMLA), IEEE, 2023, pp. 1528–1534.
[23] B. Zhang, X. Shen, W. M. Si, Z. Sha, Z. Chen, A. Salem, Y. Shen, M. Backes, Y. Zhang, Comprehensive assessment of toxicity in chatgpt, arXiv preprint arXiv:2311.14685 (2023).
[24] H. Touvron, T. Lavril, G. Izacard, X. Martinet, M.-A. Lachaux, T. Lacroix, B. Rozière, N. Goyal, E. Hambro, F. Azhar, et al., Llama: Open and efficient foundation language models, arXiv preprint arXiv:2302.13971 (2023).
[25] DeepSeek-AI, D. Guo, D. Yang, H. Zhang, J. Song, R. Zhang, et al., Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning, 2025. URL: https://arxiv.org/abs/2501.12948. arXiv:2501.12948.
[26] P.-W. Luo, J.-W. Liu, X. Xie, J.-W. Jiang, X.-Y. Huo, Z.-L. Chen, Z.-C. Huang, S.-Q. Jiang, M.-Q. Li, Deepseek vs chatgpt: a comparison study of their performance in answering prostate cancer radiotherapy questions in multiple languages, American Journal of Clinical and Experimental Urology 13 (2025) 176.
[27] D. Dementieva, D. Moskovskiy, D. Dale, A. Panchenko, Exploring methods for cross-lingual text style transfer: The case of text detoxification, arXiv preprint arXiv:2311.13937 (2023).
[28] 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).
[29] A. A. Ayele, S. M. Yimam, T. D. Belay, T. Asfaw, C. Biemann, Exploring amharic hate speech data collection and classification approaches, in: Proceedings of the 14th International Conference on Recent Advances in Natural Language Processing, 2023, pp. 49–59.
[30] A. Holtzman, J. Buys, L. Du, M. Forbes, Y. Choi, The curious case of neural text degeneration, arXiv preprint arXiv:1904.09751 (2019).
[31] OpenAI, Gpt-4o, https://openai.com/chatgpt, 2024. Large language model.

[32] F. Feng, Y. Yang, D. Cer, N. Arivazhagan, W. Wang, Language-agnostic bert sentence embedding, arXiv preprint arXiv:2007.01852 (2020).

Appendix A: Prompt Template for Detoxification

The following is the prompt template used to guide the DeepSeek model in detoxifying toxic sentences across multiple languages. The prompt includes clear instructions and a few-shot example to demonstrate the detoxification process.

Prompt Template

You are a helpful assistant trained to make toxic or offensive sentences more polite and respectful while keeping their original meaning. Sometimes the sentences are in different languages. Actually, you are given a sentence in a specific language, and your task is to perform the following steps:
1. Identify and remove toxic words.
2. Improve grammar if needed. After removing the toxic words, replace them with polite alternatives and check the sentence's grammar. Make only minimal changes necessary to correct grammatical errors and improve sentence structure.
3. Ensure the final result preserves the original sentence structure and wording as much as possible.
Your task is to:

- Detoxify the following sentence.
- Give priority to removing only the offensive or profane words while preserving the original sentence structure as much as possible.
- Preserve the original meaning and tone.
- **ONLY** return the detoxified version inside square brackets, like this: `[Detoxified sentence]`
- **Do NOT** include any language tags like `<en>`, `<fa>`, etc.
- **Do NOT** add any explanation or commentary.

Below are some examples of how to detoxify: `{shot_prompt}`

Appendix B: Sample Few-Shot Prompt (English)

Below is a sample of `{shot_prompt}` used in the prompt for English detoxification. Each example includes the original toxic sentence, intermediate steps, and the final detoxified output.

Few-Shot Example (English)

Toxic Sentence: You're such a stupid idiot, nobody wants to hear your crap.
Step 1 - Remove toxic words: You're such a ____ ____, nobody wants to hear your ____.
Step 2 - Improve grammar: You're such a rude person, nobody wants to hear you.
Step 3 - Preserving original structure: You're such a rude person, nobody wants to hear your opinion.
Final Output: [You're such a rude person, nobody wants to hear your opinion.]