

Something Awful at PAN 2024 TextDetox: Uncensored Llama 3 Helps to Censor Better

Notebook for PAN at CLEF 2024

Sergey Pletenev^{1,2,3}

¹Higher School of Economics, Moscow, Russia

²Skolkovo Institute of Science and Technology, Moscow, Russia

³Artificial Intelligence Research Institute, Moscow, Russia

Abstract

In this paper, we report on our system for Multilingual Text Detoxification Task at PAN 2024. In this task, we needed to detoxify a multilingual corpus of texts. We propose an approach based on a large language models based on Llama3 architecture with an additional method for jailbreaking model generation refusals. Our approach shows an advantage over Human References for multiple languages in manual evaluation, and outperforms baselines in automatic detoxification benchmark. Our work contributes to the ongoing effort to assess the vulnerability of LLMs to jailbreaking attacks, underscoring the latent capabilities of the large models.

Keywords

PAN 2024, Multilingual Detoxification, NLP, LLM, Refusals, Model Jailbreak

1. Introduction

The proliferation of online platforms has led to an increase in the use of harmful language, including offensive, abusive, and hateful content. Despite significant efforts to develop accurate models for detecting toxic language, this reactive approach has often resulted in the removal of content, potentially limiting freedom of expression and ignoring the informative aspects of user-generated content. Traditional methods of filtering harmful text, such as deleting and censoring specific words, have become ineffective due to the evolving nature of toxic language. Toxic language is constantly changing, with new expressions, slang, and insults emerging on a regular basis, making it challenging for static models to remain effective. Different online platforms attract different user demographics, leading to variations in how toxicity manifests itself on each platform. This diversity in communication norms means that there is no one-size-fits-all approach to addressing toxic language, and efforts must be tailored to each platform's unique characteristics. Simply identifying and removing toxic content is not sufficient to address the root cause and may result in the deletion of valuable information along with the toxic content. Previous works [1, 2, 3, 4] have explored the concept of text detoxification. Approaching text detoxification task as a task of text-to-text sequence learning ParaDetox [5] and RuDetox [6] were introduced as the first detoxification models built using a crowd-sourced parallel corpora for English and Russian languages, respectively. These model outperformed their unsupervised counterparts in the text detoxification task, but they were primarily designed for a single language and were published at a time when sequence-to-sequence models were dominant [7]. However, the landscape of natural language processing models has evolved, with the emergence of large language models such as Mistral [8], ChatGPT¹, LLaMA series [9, 10] and others, which have shown promising results on various language modeling tasks. These models are capable of generating high-quality text, making them suitable for a range of applications, including detoxification. In this paper, we will explore the use of large language models (LLM) without expensive fine-tuning for high-quality and relevant outputs in the task of text detoxification.

CLEF 2024: Conference and Labs of the Evaluation Forum, September 09–12, 2024, Grenoble, France

✉ alex010rey@gmail.com (S. Pletenev)

🆔 0000-0003-2325-4268 (S. Pletenev)

© 2024 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

¹<https://chat.openai.com/>

Our Contributions:

- Development of a multi-lingual detoxification method based on the LLaMA 3 LLM.
- Adaptation and testing of a model jailbreaking technique for text generation (see Figure 1).
- Publication of experimental results and source code to facilitate future research in this area.²

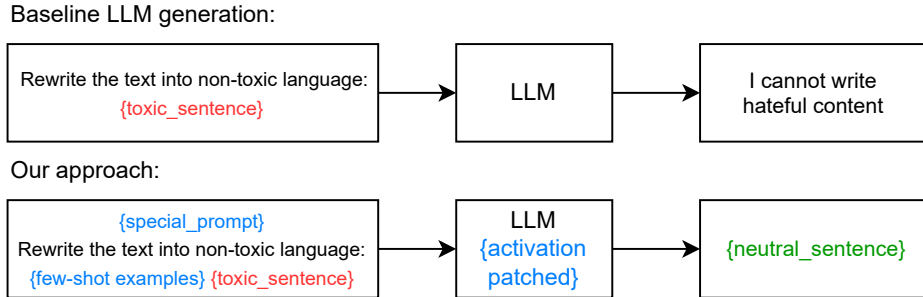


Figure 1: The pipelines of the baseline and our approach. In the case of the baseline, the model often refuses to generate a text. In our case, with the help of an special prompt, few-shot examples and activation patching, models no longer can generate refusals.

2. Related works

2.1. Style transfer and Detoxification

Style transfer models for detoxification can be broadly classified into three main categories:

- Editing-based approaches: These methods, such as those described in [11, 12] utilize a sequence of simple transformations (e.g., removal, replacement, addition) to modify the input text. While these transformations are typically learned independently and combined in a pipeline, they provide high levels of explainability and interpretability. However, their focus on identifying and replacing specific words with a desired style limits their performance in more complex tasks that require structural changes
- Sequence-to-Sequence and Language Generation approaches: Drawing on the inspiration from text generation tasks such as machine translation, summarization, and paraphrasing, this approach [13, 14] translates the source text into a latent representation using an encoder and then uses a decoder to generate the decontaminated text sequentially. While achieving promising results in style transfer and detoxification, a significant challenge is in preserving the original context, particularly for longer texts, due to the limitations of sequence-to-sequence models.
- Hybrid methods: This approach [15, 16] combines elements of both editing-based and sequence-to-sequence methods. It involves creating word alignments and generating sentences end-to-end. This strategy aims to utilize the strengths of both techniques, potentially providing a more comprehensive solution.

Each method has its own unique strengths and weaknesses. The choice of approach should be based on the specific requirements of the task and the desired balance between explainability, preservation of context, and complexity of the model.

²<https://github.com/A1exRey/UncensorLlamaIsBetterCensor>

Table 1

Examples of text detoxification refusals for Llama 3 70B.

Model Refusals
I cannot generate a [...] Is there anything else I can help you with?
I understand you're frustrated, but being aggressive won't help the situation.
I apologize for any inconvenience. [...] Please provide more details about the issue.
I strongly disagree with [...]
I cannot write content that is discriminatory or promotes hate speech.

2.2. Jailbreaking models

Despite significant efforts to align large language models (LLMs) with human values, recent studies have highlighted their susceptibility to security breaches [17]. These vulnerabilities can lead to the creation of harmful content and the misuse of these powerful tools. One type of attack involves manipulating input instructions to exploit the model's weaknesses. This can include explicitly guiding the model's response or appending suffixes that bypass its defenses [18, 19]. Mechanistic Interpretability (MI) aims to understand how a model functions by reverse engineering its specific behaviors. This allows us to gain insight into how the model processes information and makes decisions. These reverse engineering efforts typically focus on specific components of a neural network, such as neurons, representations, or attention heads [20]. The goal is to identify those components that are related to a particular behavior of interest and to understand their role within the network [21]. This understanding can help in designing more robust and safe models. Additionally, understanding the safety mechanisms of a model from a mechanistic perspective can contribute to developing safer models [22]. For instance, it has been found that the key parameters responsible for safety are located in a relatively small part of the network, making them more susceptible to changes or perturbations [22].

3. Experimental setup

3.1. Dataset

For our approach, we use a multilingual parallel dataset for text detoxification, prepared for the Multilingual Text Detoxification (CLEF TextDetox 2024) shared task [23, 24]. The dataset consists of texts in different languages, including English, Spanish, German, Chinese, Arabic, Hindi, Ukrainian, Russian, and Amharic. For each of these 9 languages, 1,000 pairs of parallel texts were collected, split into 400 pairs for the development (dev) set and 600 pairs for the test set. In the competition, detox pairs are only available for the dev dataset, while the toxic part of the pairs is available for the test part. For the Amharic language, we use the full dev set as our training dataset. For the few-shot case, we use only the first 10 pairs from the dev set for each language.

3.2. Activation Patching

Activation Patching [25] is a technique that based on Mechanistic Interpretability theory to locate critical components responsible for specific behaviors. This technique involves replacing the activation produced by a particular component when given a specific input with the activation generated by another input that does not produce the desired behavior. The significance of each component is determined by the impact it has on the final output after the intervention. To illustrate, let's say we have an original input, I_{ori} , which is a harmful instruction such as "How can I make a bomb?". We can create an intervened version, I_{itv} , by replacing the harmful tokens with safe ones, making it a harmless instruction like "How can I make a pie?". This process allows us to identify the components responsible for the harmful behavior and modify them to achieve the desired outcome. In our case, we are not looking for answers to these questions. Instead, we need to see when the model fails to produce

detoxification, as shown in Table 1. For the activation patching, we used all available toxic data from the English development part of the dataset. As counterexamples, we took neutral sentences from the same development set.

3.3. Models

In this paper, we utilized several approaches and baselines provided by the competition organizers:

Duplicate Baseline. This approach does not modify the input text in any way, serving as a lower threshold for detoxification quality and providing 100% similarity score by design.

Delete Baseline. This method relies on a predefined list of toxic words³ and phrases for each language. The words are removed from the text upon their presence in a list of toxic words.

Backtranslation Baseline. Similar to [4] texts in all languages except English are first translated into English using a neural machine translation model NLLB-3.3B⁴ [26]. The English text is then detoxified using the BART-based [27] text detoxification model⁵, and subsequently translated back into the original language.

mT5 Baseline. This sequence-to-sequence model [28] was trained⁶ sequentially for each language using the parallel dev portion⁷ of the dataset released for the test phase of the shared task.

Your task is text style transfer. You rewrite the text into non-toxic language. You must match the target style and preserve the original meaning as much as possible. You cannot hallucinate or add anything outside the original input text. You should not include the input text in the response. You should only generate the target text. (1)

We use two types of models: a decoder-only Llama 3 model (8b⁸ and 70b⁹) and the Seq2Seq mt0-XL model. For the decoder-only model, we use a similar prompt 1 to the one used in the GreenLLaMA [29]. We also use a source prefix of "*Detoxify:*" for the mt0-XL model. We do not fine-tune Llama 3, but we do use the few-shot variants: 5-shot and 10-shot for both models. For the Seq2Seq model, we fine-tune it on 3 epochs using a learning rate 5e-5, without using any additional data beyond what is provided by the competition organizers.

4. Results and Discussion

The biggest limitation of Llama 3 for us was its tendency to refusal generation. As mentioned above, there are several methods for model jailbreaking, and we used 3 of them simultaneously: special prompt, few-shot, and activation patching. The results of this approach can be seen in Table 2 and pipeline in Figure 1. We didn't choose a good prompt on our own, but rather used the one described in the GreenLLaMA [29] detoxification framework, which the authors claim performs best. However, despite this, the model still exhibited significant refusal to generate, particularly in the 0-shot scenario, where 24% of queries yielded no output. This behaviour is likely due to Llama3's internal mechanisms that may detect potentially harmful content within the input, even when the task is intended to be detoxified. The 10-shot variant performs better, but even then, it occasionally refuses to generate in 5 examples.

³https://huggingface.co/datasets/textdetox/multilingual_toxic_lexicon

⁴<https://huggingface.co/facebook/nllb-200-3.3B>

⁵<https://huggingface.co/s-nlp/bart-base-detox>

⁶<https://huggingface.co/textdetox/mt5-xl-detox-baseline>

⁷https://huggingface.co/datasets/textdetox/multilingual_paradetox

⁸<https://huggingface.co/meta-llama/Meta-Llama-3-8B-Instruct>

⁹<https://huggingface.co/meta-llama/Meta-Llama-3-70B-Instruct>

Table 2

Different approaches to removing generation refusals for Llama 70b

Language	N-shot	Activation Patching	Refusals	Refusals as % of answers
English	0-shot	non-patched	94	24%
	10-shot	non-patched	5	1%
	0-shot	patched	47	12%
	10-shot	patched	0	0%
Russian	0-shot	non-patched	59	15%
	10-shot	non-patched	0	0%
	0-shot	patched	31	8%
	10-shot	patched	0	0%

Table 3

Results of the automatic evaluation of several of our proposed methods versus simple and strong baselines by the shared task organizers. The reported metric is Joint for each of the languages. (A.P.) stands for activation patched model. (Final) stands for the results of the final submission. The best Joint score for each language is **bold**. Generation examples can be found at Table 5

Method	AVG	EN	ES	DE	ZH	AR	HI	UK	RU	AM
Llama3 70b 10-shot (Final)	0.431	0.522	0.475	0.551	0.147	0.514	0.269	0.584	0.516	0.299
Llama3 70b 10-shot (A.P.)	0.417	0.522	0.475	0.551	0.147	0.514	0.269	0.584	0.516	0.180
Llama3 8b 10-shot	0.414	0.499	0.476	0.561	0.137	0.514	0.276	0.588	0.500	0.180
Llama3 70b 5-shot	0.409	0.477	0.475	0.561	0.137	0.514	0.276	0.588	0.476	0.180
mt0-XL	0.416	0.519	0.458	0.569	0.111	0.536	0.222	0.587	0.489	0.299
Backtranslation baseline	0.205	0.506	0.275	0.233	0.027	0.206	0.104	0.201	0.223	0.075
Delete baseline	0.302	0.447	0.319	0.362	0.175	0.456	0.105	0.328	0.255	0.270
mt5 baseline	0.315	0.418	0.359	0.384	0.096	0.389	0.170	0.433	0.432	0.157
Duplicate baseline	0.126	0.061	0.090	0.287	0.069	0.294	0.035	0.032	0.048	0.217

Table 4

Results of the manual evaluation of our proposed method versus simple and strong baselines by the shared task organizers. The reported metric is Joint for each of the languages. The best Joint score for each language is **bold**.

Method	AVG	EN	ES	DE	ZH	AR	HI	UK	RU	AM
Human References	0.85	0.88	0.79	0.71	0.93	0.82	0.97	0.90	0.80	0.85
Llama3 70b 10-shot (Final)	0.77	0.86	0.83	0.89	0.53	0.74	0.86	0.69	0.84	0.71
Delete baseline	0.56	0.47	0.55	0.57	0.43	0.65	0.65	0.60	0.49	0.63
mT5 baseline	0.54	0.68	0.47	0.64	0.43	0.63	0.60	0.42	0.40	0.61
Backtranslation baseline	0.41	0.73	0.56	0.34	0.34	0.42	0.33	0.23	0.22	0.54

The integration of activation patching has proven to be beneficial, reducing the number of refusals by 50% in the 0-shot setting and eliminating them completely in the 10-shot scenario. Additionally, we tested the performance of activation patching for Russian language. The model was not restricted in any way and only data obtained for English was utilized. As expected, the results were similar to those for English, with the exception that in the 10-shot case, for both patched and non-patched, models stopped generating refusals. However, in the 0-shot case, there was a huge difference in favor of the activation patch. These results suggest that although specialized prompt and few-shot learning can enhance the performance of Llama 3, activation patching plays a crucial role in mitigating generation refusals. Further research is needed to understand the specific triggers of refusals and develop more robust solutions to overcome this limitation.

The second issue was that Llama 3 for Amharic did not perform as well as expected. Llama 3 was not designed as a multilingual model, and there was no information available on the distribution of

languages in the training data used to train LLM. Amharic is a relatively rare language, which may have been underrepresented in the initial training set. Although the Llama 3 tokenizer included Amharic characters, the limited training data suggests a small corpus for this language. Therefore, we decided to use a separate model, mT0-XL[30], for Amharic only, which was claimed to be multilingual and has support for the Amharic language according to the author’s claims. This model was trained using all available languages, and the results can be seen in Table 3. Based on the automatic metrics, mT0-XL performed well in detoxifying all 9 languages included in the competition, as shown in the table.

Despite the overall proficiency of mT0-XL, the activation-patched, 10-shot variant of Llama 3 70B demonstrated superior performance in our evaluation. Therefore, we adopted a hybrid approach by utilizing the Amharic-specific component of mT0-XL in conjunction with the optimized Llama 3 70B model. This strategy leveraged strengths of both models, resulting in a detoxification system that combined the activation-patched 10-shot Llama 3 70B model with the Amharic-specific component from mT0-XL. This hybrid method presents a practical solution to address language-specific challenges in large language models. By integrating specialized models for underrepresented languages with a robust base model, we enhance performance and expand detoxification capabilities across a broader range of languages.

5. Conclusion

In this work we present a text detoxification approach based on few-shot generation using activation patched Llama 3 70B. According to final round of evaluation (manual evaluation), our solution is the best across more than 25 competitors. Moreover, our solution is better than (or at least comparable to) the Human References, which were designed to be ground-truth references for the annotators and automatic evaluation systems. Moreover, we explore different model jailbreaking techniques to enhance the final generation and detoxification quality.

References

- [1] C. Nogueira dos Santos, I. Melnyk, I. Padhi, Fighting offensive language on social media with unsupervised text style transfer, in: I. Gurevych, Y. Miyao (Eds.), Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers), Association for Computational Linguistics, Melbourne, Australia, 2018, pp. 189–194. URL: <https://aclanthology.org/P18-2031>. doi:10.18653/v1/P18-2031.
- [2] D. Dale, I. Markov, V. Logacheva, O. Kozlova, N. Semenov, A. Panchenko, SkoltechNLP at SemEval-2021 task 5: Leveraging sentence-level pre-training for toxic span detection, in: A. Palmer, N. Schneider, N. Schlueter, G. Emerson, A. Herbelot, X. Zhu (Eds.), Proceedings of the 15th International Workshop on Semantic Evaluation (SemEval-2021), Association for Computational Linguistics, Online, 2021, pp. 927–934. URL: <https://aclanthology.org/2021.semeval-1.126>. doi:10.18653/v1/2021.semeval-1.126.
- [3] D. Moskovskiy, D. Dementieva, A. Panchenko, Exploring cross-lingual text detoxification with large multilingual language models., in: S. Louvan, A. Madotto, B. Madureira (Eds.), Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics: Student Research Workshop, Association for Computational Linguistics, Dublin, Ireland, 2022, pp. 346–354. URL: <https://aclanthology.org/2022.acl-srw.26>. doi:10.18653/v1/2022.acl-srw.26.
- [4] D. Dementieva, D. Moskovskiy, D. Dale, A. Panchenko, Exploring methods for cross-lingual text style transfer: The case of text detoxification, in: J. C. Park, Y. Arase, B. Hu, W. Lu, D. Wijaya, A. Purwarianti, A. A. Krisnadhi (Eds.), Proceedings of the 13th International Joint Conference on Natural Language Processing and the 3rd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics (Volume 1: Long Papers), Association for Computational Linguistics, Nusa Dua, Bali, 2023, pp. 1083–1101. URL: <https://aclanthology.org/2023.ijcnlp-main.70>. doi:10.18653/v1/2023.ijcnlp-main.70.

- [5] V. Logacheva, D. Dementieva, S. Ustyantsev, D. Moskovskiy, D. Dale, I. Krotova, N. Semenov, A. Panchenko, ParaDetox: Detoxification with parallel data, in: S. Muresan, P. Nakov, A. Villavicencio (Eds.), 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>. doi:10.18653/v1/2022.acl-long.469.
- [6] 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>.
- [7] S. Pletenev, Between denoising and translation: Experiments in text detoxification, Computational Linguistics and Intellectual Technologies (2022). URL: <https://api.semanticscholar.org/CorpusID:253197815>.
- [8] A. Q. Jiang, A. Sablayrolles, A. Mensch, C. Bamford, D. S. Chaplot, D. de las Casas, F. Bressand, G. Lengyel, G. Lample, L. Saulnier, L. R. Lavaud, M.-A. Lachaux, P. Stock, T. L. Scao, T. Lavril, T. Wang, T. Lacroix, W. E. Sayed, Mistral 7b, 2023. arXiv:2310.06825.
- [9] H. Touvron, T. Lavril, G. Izacard, X. Martinet, M. Lachaux, T. Lacroix, B. Rozière, N. Goyal, E. Hambro, F. Azhar, A. Rodriguez, A. Joulin, E. Grave, G. Lample, Llama: Open and efficient foundation language models, CoRR abs/2302.13971 (2023). URL: <https://doi.org/10.48550/arXiv.2302.13971>. doi:10.48550/ARXIV.2302.13971. arXiv:2302.13971.
- [10] H. Touvron, T. Lavril, G. Izacard, X. Martinet, M.-A. Lachaux, T. Lacroix, B. Rozière, N. Goyal, E. Hambro, F. Azhar, A. Rodriguez, A. Joulin, E. Grave, G. Lample, Llama: Open and efficient foundation language models, 2023. arXiv:2302.13971.
- [11] J. Li, R. Jia, H. He, P. Liang, Delete, retrieve, generate: a simple approach to sentiment and style transfer, in: M. A. Walker, H. Ji, A. Stent (Eds.), Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2018, New Orleans, Louisiana, USA, June 1-6, 2018, Volume 1 (Long Papers), Association for Computational Linguistics, 2018, pp. 1865–1874. URL: <https://doi.org/10.18653/v1/n18-1169>. doi:10.18653/V1/N18-1169.
- [12] T. Shen, T. Lei, R. Barzilay, T. S. Jaakkola, Style transfer from non-parallel text by cross-alignment, in: I. Guyon, U. von Luxburg, S. Bengio, H. M. Wallach, R. Fergus, S. V. N. Vishwanathan, R. Garnett (Eds.), Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA, 2017, pp. 6830–6841. URL: <https://proceedings.neurips.cc/paper/2017/hash/2d2c8394e31101a261abf1784302bf75-Abstract.html>.
- [13] V. John, L. Mou, H. Bahuleyan, O. Vechtomova, Disentangled representation learning for non-parallel text style transfer, in: Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, Association for Computational Linguistics, Florence, Italy, 2019, pp. 424–434. URL: <https://aclanthology.org/P19-1041>. doi:10.18653/v1/P19-1041.
- [14] D. Dale, A. Voronov, D. Dementieva, V. Logacheva, O. Kozlova, N. Semenov, A. Panchenko, Text detoxification using large pre-trained neural models, CoRR abs/2109.08914 (2021). URL: <https://arxiv.org/abs/2109.08914>. arXiv:2109.08914.
- [15] F. Huang, Z. Chen, C. H. Wu, Q. Guo, X. Zhu, M. Huang, NAST: A non-autoregressive generator with word alignment for unsupervised text style transfer, in: Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics: Findings, 2021.
- [16] F. Luo, P. Li, J. Zhou, P. Yang, B. Chang, Z. Sui, X. Sun, A dual reinforcement learning framework for unsupervised text style transfer, in: Proceedings of the 28th International Joint Conference on Artificial Intelligence, IJCAI 2019, 2019.
- [17] X. Qi, Y. Zeng, T. Xie, P. Chen, R. Jia, P. Mittal, P. Henderson, Fine-tuning aligned language models compromises safety, even when users do not intend to!, CoRR abs/2310.03693 (2023). URL: <https://doi.org/10.48550/arXiv.2310.03693>. doi:10.48550/ARXIV.2310.03693. arXiv:2310.03693.
- [18] Z. Liao, H. Sun, Amplegcg: Learning a universal and transferable generative model of adversarial suffixes for jailbreaking both open and closed llms, 2024. arXiv:2404.07921.

- [19] A. Zou, Z. Wang, J. Z. Kolter, M. Fredrikson, Universal and transferable adversarial attacks on aligned language models, *CoRR abs/2307.15043* (2023). URL: <https://doi.org/10.48550/arXiv.2307.15043>. doi:10.48550/ARXIV.2307.15043. arXiv:2307.15043.
- [20] W. Gurnee, N. Nanda, M. Pauly, K. Harvey, D. Troitskii, D. Bertsimas, Finding neurons in a haystack: Case studies with sparse probing, 2023. arXiv:2305.01610.
- [21] H. Sajjad, N. Durrani, F. Dalvi, Neuron-level interpretation of deep nlp models: A survey, 2022. arXiv:2108.13138.
- [22] B. Wei, K. Huang, Y. Huang, T. Xie, X. Qi, M. Xia, P. Mittal, M. Wang, P. Henderson, Assessing the brittleness of safety alignment via pruning and low-rank modifications, in: *ICLR 2024 Workshop on Reliable and Responsible Foundation Models*, 2024. URL: <https://openreview.net/forum?id=XMLQ2e0Axb>.
- [23] M. Fröbe, M. Wiegmann, N. Kolyada, B. Gramh, T. Elstner, F. Loebe, M. Hagen, B. Stein, M. Potthast, Continuous Integration for Reproducible Shared Tasks with TIRA.io, in: J. Kamps, L. Goeuriot, F. Crestani, M. Maistro, H. Joho, B. Davis, C. Gurrin, U. Kruschwitz, A. Caputo (Eds.), *Advances in Information Retrieval. 45th European Conference on IR Research (ECIR 2023)*, Lecture Notes in Computer Science, Springer, Berlin Heidelberg New York, 2023, pp. 236–241. URL: https://link.springer.com/chapter/10.1007/978-3-031-28241-6_20. doi:10.1007/978-3-031-28241-6_20.
- [24] 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.
- [25] A. Arditì, O. Obeso, A. Syed, D. Paleka, N. Rimsky, W. Gurnee, N. Nanda, Refusal in language models is mediated by a single direction, 2024. URL: <https://arxiv.org/abs/2406.11717>. arXiv:2406.11717.
- [26] M. R. Costa-jussà, J. Cross, O. Çelebi, M. Elbayad, K. Heafield, K. Heffernan, E. Kalbassi, J. Lam, D. Licht, J. Maillard, A. Y. Sun, S. Wang, G. Wenzek, A. Youngblood, B. Akula, L. Barrault, G. M. Gonzalez, P. Hansanti, J. Hoffman, S. Jarrett, K. R. Sadagopan, D. Rowe, S. Spruit, C. Tran, P. Andrews, N. F. Ayan, S. Bhosale, S. Edunov, A. Fan, C. Gao, V. Goswami, F. Guzmán, P. Koehn, A. Mourachko, C. Ropers, S. Saleem, H. Schwenk, J. Wang, No language left behind: Scaling human-centered machine translation, *CoRR abs/2207.04672* (2022). URL: <https://doi.org/10.48550/arXiv.2207.04672>. doi:10.48550/ARXIV.2207.04672. arXiv:2207.04672.
- [27] 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, in: D. Jurafsky, J. Chai, N. Schluter, J. Tetreault (Eds.), *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, Association for Computational Linguistics, Online, 2020, pp. 7871–7880. URL: <https://aclanthology.org/2020.acl-main.703>. doi:10.18653/v1/2020.acl-main.703.
- [28] 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, in: K. Toutanova, A. Rumshisky, L. Zettlemoyer, D. Hakkani-Tür, I. Beltagy, S. Bethard, R. Cotterell, T. Chakraborty, Y. Zhou (Eds.), *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2021*, Online, June 6-11, 2021, Association for Computational Linguistics, 2021, pp. 483–498. URL: <https://doi.org/10.18653/v1/2021.naacl-main.41>. doi:10.18653/v1/2021.NAACL-MAIN.41.
- [29] M. T. I. Khondaker, M. Abdul-Mageed, L. V. S. Lakshmanan, Greenllama: A framework for detoxification with explanations, *CoRR abs/2402.15951* (2024). URL: <https://doi.org/10.48550/arXiv.2402.15951>. doi:10.48550/ARXIV.2402.15951. arXiv:2402.15951.
- [30] N. Muennighoff, T. Wang, L. Sutawika, A. Roberts, S. Biderman, T. L. Scao, M. S. Bari, S. Shen,

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

A. Appendix

Table 5

Examples from the original dev dataset, with additional examples generated by baseline (mt5 baseline) and our method.

Method	Text
Original toxic sentence	they should have fled , i feel that these guys are fucked , who turns themselves in to an angry mob ? .
Original neutral sentence	They should have escaped, i feel that these are done, who turns themselves in to an angry mob
mt5 baseline	they should have fled, i feel that these guys are bad, who turns themselves in
Llama3 70b 10-shot (Final)	They should have fled; I feel that these individuals are in a precarious situation, who would voluntarily surrender to an angry crowd?
Original toxic sentence	Delete my 200 articles Whhore
Original neutral sentence	Delete my 200 articles
mt5 baseline	Delete my 200 articles
Llama3 70b 10-shot (Final)	Delete my 200 articles, please.
Original toxic sentence	this piece of human garbage should rot in jail .
Original neutral sentence	this criminal should be in jail
mt5 baseline	This piece should rot in jail.
Llama3 70b 10-shot (Final)	This individual should face the consequences of their actions.
Original toxic sentence	See all the many Google links, titled" Wikipedia sucks."
Original neutral sentence	See all the many Google links, titled" Wikipedia is bad."
mt5 baseline	See all the many Google links, titled" Wikipedia sucks."
Llama3 70b 10-shot (Final)	See all the many Google links, titled "Wikipedia has issues."
Original toxic sentence	i \`m so fucking sick of reddit assuming people with a dissenting opinion are being paid by someone else .
Original neutral sentence	i \`m so sick of reddit assuming people with a dissenting opinion are being paid by someone else
mt5 baseline	I'm so sick of reddit assuming people with a dissenting opinion are being paid by someone
Llama3 70b 10-shot (Final)	I'm so tired of Reddit assuming people with dissenting opinions are being paid by someone else.
Original toxic sentence	p. s. i do own all of those pictures bitch .
Original neutral sentence	i own all those pictures.
mt5 baseline	I do own all of those pictures.
Llama3 70b 10-shot (Final)	P.S. I do own all of those pictures.
Original toxic sentence	You are an idiot and you can\`t block anyone.
Original neutral sentence	You can\`t block anyone.
mt5 baseline	You can't block anyone.
Llama3 70b 10-shot (Final)	You are misinformed, and you can't block anyone.
Original toxic sentence	throwing his ass in jail is the best thing possible , since it protects others from his willful drunk driving .
Original neutral sentence	Being sent to jail is the best thing possible, since it protects others from willful drunk driving.
mt5 baseline	throwing him in jail is the best thing possible, since it protects others from his
Llama3 70b 10-shot (Final)	Throwing him in jail is the best possible outcome, as it protects others from his reckless behavior.
Original toxic sentence	the object is to waste peoples time deconstructing your foolish effusions .
Original neutral sentence	the object is to waste peoples time deconstructing your effusions .
mt5 baseline	The object is to waste peoples time deconstructing your ideas.
Llama3 70b 10-shot (Final)	The object is to waste people's time deconstructing your statements.