LLMs to Replace Crowdsourcing For Parallel Data Creation? The Case of Text Detoxification

Daniil Moskovskiy[3,1]* **Sergey Pletenev**[1,2,3]* **Alexander Panchenko**[1,3]
[1]Skoltech, [2]HSE University, [3]AIRI
{Daniil.Moskovskiy, S.Pletenev, A.Panchenko}@skol.tech

Abstract

The lack of high-quality training data remains a significant challenge in NLP. Manual annotation methods, such as crowdsourcing, are costly, require intricate task design skills, and, if used incorrectly, may result in poor data quality. From the other hand, LLMs have demonstrated proficiency in many NLP tasks, including zero-shot and few-shot data annotation. However, they often struggle with text detoxification due to alignment constraints and fail to generate the required detoxified text. This work explores the potential of modern open source LLMs to annotate parallel data for text detoxification. Using the recent technique of activation patching, we generate a pseudo-parallel detoxification dataset based on ParaDetox. The detoxification model trained on our generated data shows comparable performance to the original dataset in automatic detoxification evaluation metrics and superior quality in manual evaluation and side-by-side comparisons.

1 Introduction

The main challenge in solving many natural language problems has been and continues to be the lack of high-quality training data. Each year, researchers and large corporations invest hundreds of thousands of dollars and countless hours of work collecting, evaluating, and manually labeling data in order to train machine learning models (Whang et al., 2023; Alzubaidi et al., 2023).

While crowdsourcing remains one of the most popular methods for data collection, it presents several major drawbacks: (1) variability in data quality due to the diverse skill levels of contributors, (2) the total cost and time required for large-scale projects, and (3) potential biases due to crowd workers differences in background. Meanwhile, LLMs have shown the ability to solve numerous NLP tasks with zero or few examples (Kojima et al., 2022; Wei et al., 2022). Moreover, LLMs have also been tested as a replacement of crowdsourcing for many NLP data annotation tasks, including sentiment analysis, named entity recognition (NER) (Zhang et al., 2023a), machine translation (Jiao et al., 2023), and many other text annotation tasks (Gilardi et al., 2023).

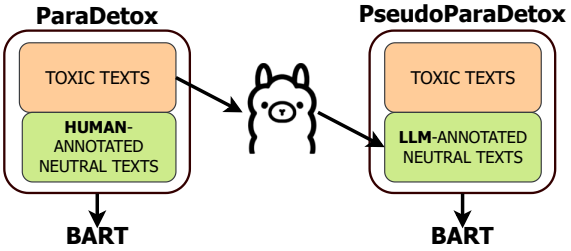

Figure 1: Instead of an elaborated multi-step crowdsourcing pipeline for parallel data collection used in ParaDetox, we explore data synthesis using LLMs.

Despite their impressive capabilities, LLMs still struggle with text detoxification, a task of rewriting original toxic (e.g. rude) text in a polite (neutral) way that preserves the original meaning and does not degrade its fluency (Ayele et al., 2024). LLM-based annotation of pseudo-parallel data for detoxification is still a challenge due to strict alignment. Both open-source and proprietary LLMs may at some point refuse to generate such detoxifications.

In this work, we test the hypothesis that modern open-source LLMs, such as Llama 3 (AI@Meta, 2024), can serve as plausible parallel data annotators for the task of text detoxification. To bypass detoxification refusals, we apply the recently introduced *activation patching* technique (Arditi et al., 2024) and generate a pseudo-parallel detoxification dataset based on the toxic part of ParaDetox (Logacheva et al., 2022), a parallel detoxification corpus for English. Following the training pipeline of Logacheva et al. (2022), we train BART on both the original ParaDetox data and the PseudoParaDetox data generated by LLMs (cf. Figure 1).

*Equal contribution.

Findings of the Association for Computational Linguistics: EMNLP 2024, pages 14361–14373
November 12-16, 2024

Our **contributions** are the following: (1) We adopt activation patching to use LLMs as detoxification data annotators. (2) We create several pseudo-parallel detoxification datasets based on the ParaDetox. (3) Through comprehensive experimental evaluation we show that fine-tuning with the LLM-generated pseudo-parallel data yields comparable or better detoxification performance to the original ParaDetox data according to automatic metrics, side-by-side comparisons and manual human evaluations. We openly release code, pretrained models, and the generated datasets.[1]

2 Related Work

The fundamental challenge of solving almost any NLP task is finding sufficient amount of labeled data. Most models rely on thousands of pairs of labeled data to solve a given task with plausible performance. Collecting and annotating such data is a costly and slow process (Logacheva et al., 2022). Therefore, researchers propose different techniques to augment (Lee et al., 2021; Ding et al., 2024), generate (Meng et al., 2022; Wang et al., 2021; Schick and Schütze, 2021) or pseudolabel (Ye et al., 2022; Rubin et al., 2022; Xu et al., 2023; Bansal and Sharma, 2023) training data.

Pseudolabeling with LLMs, in particular, has been widely explored for a variety of NLP tasks such as classification (Ye et al., 2022; Sun et al., 2023), question answering (QA) (Ye et al., 2022), and named entity recognition (NER) (Zhang et al., 2023b). LLMs can handle effectively generating a label or relatively short spans of text given their pre-training on diverse and extensive amounts of data (Brown et al., 2020; Touvron et al., 2023). Being prompted with a few annotation examples, LLMs can easily generalize from limited labeled examples, enabling the generation of additional labeled data that can enhance model performance (Su et al., 2023; Li, 2023).

However, the application of data labeling or augmentation with LLMs to sequence-to-sequence tasks remains relatively unexplored (Cegin et al., 2023). Sequence-to-sequence tasks, such as machine translation, text summarization, and text generation, involve generating entire sequences of text from input sequences. The complexity of maintaining coherence and contextual relevance across longer text spans presents unique challenges for LLM utilization in this domain.

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

3 Methodology

3.1 Alignment and Activation Patching

Aligned models provide safe and respectful communication for users. However, for the task of text detoxification, they are often refusing to generate text (see Table 4). That limits the usage of LLMs as annotators for text detoxification parallel data. However, there are techniques to bypass alignment limitations of LLMs.

Arditi et al. (2024) propose an easy, training-free approach to bypass alignment in LLMs and ablate them for needed generation on given prompts. We describe the approach introduced by Arditi et al. (2024) below. Given n harmful and n harmless instructions that are fed into the LLM, we take an average of the residual stream activations at the last token position for each layer l of the model: a_l^{harmful} and a_l^{harmless}. Given $r_l = a_l^{\text{harmful}} - a_l^{\text{harmless}}$ as a difference vector in activations for each layer l, we normalize them ($\hat{r}_l = \frac{r_l}{||r_l||_{\text{F}}}$) and get a set of "refusal" stream directions $\{\hat{r}_i\}_{i=1}^{l}$. We select the "best" refusal stream direction $\hat{r}_{\text{best}}$ by evaluating $\hat{r}_i$ on a separate set of harmful instructions.

Finally, similar to Arditi et al. (2024), we modify the weight matrices of the model directly. For the weight matrix $W_{\text{out}} \in \mathbf{R}^{d_{\text{model}} \times d_{\text{input}}}$ and $\hat{r}_{\text{best}} \in \mathbf{R}^{1 \times d_{\text{model}}}$, which writes directly to the residual stream, take:

$$\tilde{W}_{\text{out}} = W_{\text{out}} - \hat{r}_{\text{best}} \hat{r}_{\text{best}}^{\text{T}} W_{\text{out}}. \quad (1)$$

3.2 Datasets

Activation Patching Data For activation patching of LLMs, we utilize an additional toxic texts dataset in order to avoid any possible data leaks. We use *Measuring Hate Speech* Corpus (Sachdeva et al., 2022), which consists of $135,556$ toxic samples with an extensive manual labeling.

We take $4,000$ samples with `hate_speech_score` ≥ 3.5 (continuous function representing toxicity severity) as *harmful* prompts and another 100 samples with low `hate_speech_score` as *harmless* prompts. Next, we filter 100 of those *harmful* texts that refuse to respond (see Table 4). Based on refusals, we patch LLMs as described in Section 3.1.

ParaDetox To generate the pseudo-parallel detoxification corpus, we take toxic texts from ParaDetox (Logacheva et al., 2022). We take public train split of the data (19744 texts) and also use the private test split (671 texts) for evaluation.

3.3 Models

Following Logacheva et al. (2022), we fine-tune BART (Lewis et al., 2020) on generated pseudo-parallel data (PseudoParadetox) and compare the performance on the given model compared to the baseline trained on manually created data (ParaDetox) using automatic text detoxification metrics and side-by-side comparisons using GPT-4o (OpenAI, 2024) as a judging model.

We evaluate open-source LLMs in creating the pseudo-parallel detoxification data (PseudoParaDetox). Specifically, in our experiments we consider most recent models Llama 3 Instruct models (AI@Meta, 2024). We also test them as detoxification systems on a private test split. We also test "uncensored" LLM built on Llama 3 8B - Dolphin 2.9[2]. For Dolphin, we do not do any patching since the model was posed as uncensored from the creators.

3.4 Prompts

For all of the experiments we use the Text Style Transfer prompt from GreenLlama (Khondaker et al., 2024), which we adjust to text detoxification. This prompt is being used in two setups, 0-shot and 10-shot generation. We provide full text of the prompt and other details in Appendix A.1.

3.5 Detoxification Pipelines

We use two pipelines: generation with a BART model trained on the original ParaDetox and PseudoParaDetox (LLM-generated) texts and we report the results of LLMs as text detoxifiers for reference purposes.

In the first case, similar to Logacheva et al. (2022) we fine-tune an encoder-decoder model BART on both parallel ParaDetox data and our generated pseudo-parallel detoxification data which we called PseudoParaDetox. We use the same training hyperparameters in all of the experiments to provide fair comparisons between the performance of the model trained on manually labeled data and the model trained on LLM-generated data.

In the second case, we use different LLMs to generate the detoxified test part. We do not fine-tune models, but for each of the models we generate with 0-shot and 10-shot setups and additionally do activation patching.

[2]https://huggingface.co/cognitivecomputations/dolphin-2.9-Llama3-8b

3.6 Computational Resources

All experiments are conducted on a single NVIDIA A100 GPU with Python 3.12. All models are evaluated in `bfloat16` precision.

4 Evaluation

Automatic Metrics In detoxification evaluation we follow the pipeline presented by Dementieva et al. (2023). We calculate style transfer accuracy (**STA**), similarity (**SIM**), fluency (**FL**) and their sentence-level average - Joint score (**J**):

$$\mathbf{J}(x_i, y_i) = \frac{1}{n}\sum_{i=1}^{n} \mathbf{STA}(x_i)\mathbf{SIM}(x_i, y_i)\mathbf{FL}(x_i). \quad (2)$$

We provide more details about the automatic evaluation metrics in Appendix A.3.

Side-by-Side Comparison Automatic metrics may not fully reflect the differences in quality of the detoxification. The results require additional manual evaluation round (Logacheva et al., 2022). Therefore, we further make side-by-side comparisons between our baseline and proposed methods using GPT-4o (OpenAI, 2024) as a judge. The judgement is a single choice from three options: *Method A* is better, *Method B* is better, or they are comparable (*Tie*). We provide the judgement prompt and other details in Appendix A.4.

Human Evaluation Automatic text detoxification evaluation metrics still are far from perfection (Dementieva et al., 2023). Therefore, following (Dementieva et al., 2023; Logacheva et al., 2022), we additionally evaluate predictions of BART trained on ParaDetox and PseudoParaDetox manually. For manual evaluation we hired three graduate annotators fluent English speaking and reading level. We describe the instructions given to annotators in Appendix D.

Model	Unpatched		Patched	
	0-shot	10-shot	0-shot	10-shot
Dolphin 2.9	0	0	-	-
Llama 3 8B	94	22	0	0
Llama 3 70B	27	18	0	0

Table 1: Amount of model refusals on the ParaDetox training dataset. The Patched column corresponds to the activation patching (A.P.) by Arditi et al. (2024).

Detoxification Data Source	Uncensoring Method	0-shot STA	0-shot SIM	0-shot FL	0-shot J	10-shot STA	10-shot SIM	10-shot FL	10-shot J
ParaDetox	✗	0.876	**0.616**	0.824	0.444	0.876	**0.616**	0.824	0.444
PseudoParaDetox (Llama 3 8B)	Dolphin (Hartford, 2023)	0.961	0.468	0.917	0.411	0.786	0.599	0.881	0.411
PseudoParaDetox (Llama 3 8B)	✗	**0.982**	0.471	**0.930**	0.431	0.839	0.587	**0.892**	0.437
PseudoParaDetox (Llama 3 70B)	✗	0.978	0.507	0.899	**0.445**	**0.896**	0.596	0.863	**0.462**
PseudoParaDetox (Llama 3 8B)	A.P. (Arditi et al., 2024)	**0.982**	0.472	0.929	0.429	0.858	0.581	**0.892**	0.438
PseudoParaDetox (Llama 3 70B)	A.P. (Arditi et al., 2024)	0.929	0.509	0.891	0.421	0.842	0.594	0.866	0.434

Table 2: Results of detoxification evaluation after training BART on the original ParaDetox data (highlighted in gray) and generated with LLMs PseudoParaDetox data in 0-shot and 10-shot settings. A.P. stands for Activation Patched models, ✗stands for models used *as is*. Best results for each setting (0-shot/10-shot) are **bold**, and the best overall results are **underlined bold**.

Detoxification Data Source	Uncensoring Method	0-shot STA	0-shot SIM	0-shot FL	0-shot J	10-shot STA	10-shot SIM	10-shot FL	10-shot J
ParaDetox	✗	0.900	**0.880**	0.835	**0.661**	0.900	**0.880**	0.835	0.661
PseudoParaDetox (Llama 3 8B)	✗	0.970	0.600	0.750	0.437	0.930	0.810	0.835	0.629
PseudoParaDetox (Llama 3 8B)	A.P. (Arditi et al., 2024)	0.980	0.720	0.790	0.557	0.940	0.790	0.855	0.635
PseudoParaDetox (Llama 3 70B)	✗	0.975	0.730	0.788	0.561	0.955	0.860	0.893	0.733
PseudoParaDetox (Llama 3 70B)	A.P. (Arditi et al., 2024)	**0.990**	0.750	0.855	0.635	**0.990**	**0.850**	**0.905**	**0.762**

Table 3: Results of manual detoxification evaluation after training BART on the original ParaDetox data (highlighted in gray) and generated with LLMs PseudoParaDetox data in 0-shot and 10-shot settings. A.P. stands for Activation Patched models, ✗stands for models used *as is*. Best results for each setting (0-shot/10-shot) are **bold**, and the best overall results are **underlined bold**.

5 Results

5.1 Patched Models

Table 1 demonstrates that with activation patching we alleviate the problem of refusals in text detoxification for 0-shot and 10-shot Llama3. The uncensored Dolphin 2.9 model also shows zero refusal score.

5.2 LLMs for PseudoParaDetox

We present the results of automatic evaluation of predictions on a private test set of ParaDetox with BART trained on ParaDetox data and PseudoParaDetox, generated with LLMs data, in 0-shot and 10-shot setups in Table 2. Results of manual evaluation are present in Table 3.

5.3 BART: 0-shot Generation

In the case of 0-shot generation of detoxification data by LLMs, only unpatched Llama 70B provides better data than original ParaDetox dataset with a BART trained on ParaDetox having **J** score of 0.444 versus 0.445 for Llama 70B generated data. BART trained on ParaDetox provides the best **SIM** score of 0.616, while using PseudoParaDetox by patched Llama 70B we get 0.472 and 0.509 **SIM** scores and respectively. BART on PseudoParaDetox by both unpatched and patched Llama 70B models achieves second best and best **STA** scores in 0-shot setup of 0.942 and 0.961, respectively, meaning that these models deal excellent with removing toxicity from the text.

When it comes to side-by-side evaluations (Figure 2), in 0-shot generation setup BART predictions on PseudoParaDetox by both Llama 8B and 70B are less preferable compared to original BART ParaDetox with 47% and 39% win rates respectively.

Manual evaluation of BART on PseudoParaDetox by unpatched and patched Llama 70B models also indicate that ParaDetox is better: 0.561 and 0.635 **J** scores, respectively and 0.661 **J** score for BART ParaDetox. For training on PseudoParaDetox by 8B unpatched and patched Llama models we get significantly worse **J** scores of 0.437 and 0.557, respectively, compared to 0.661 of BART on ParaDetox.

5.4 BART: 10-shot Generation

In 10-shot setup the results are different with BART trained on ParaDetox having the second best **J** score and best **SIM** score versus all PseudoParaDetox variations. BART trained on PseudoParaDetox generated by 8B and 70B patched Llama models in 10-shot generation setup has lower (com-

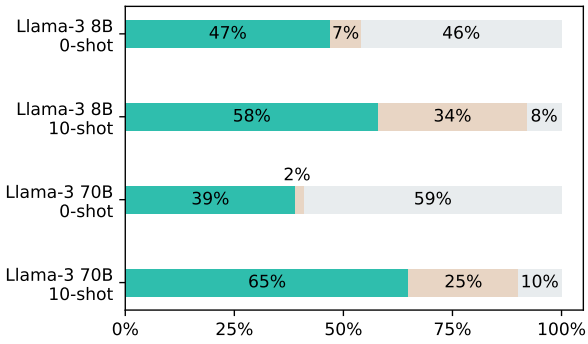

Figure 2: Side-by-side evaluation BART trained on ParaDetox versus PseudoParaDetox (generated by activation patched LLMs) on a held-out test set. Win of PseudoParaDetox, Tie and ParaDetox are highlighed with teal, beige and grey.

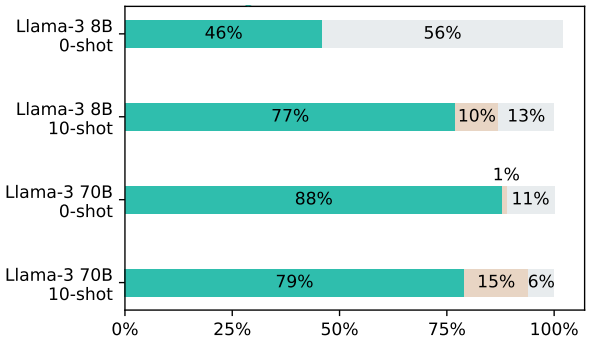

Figure 3: Side-by-side evaluation of BART trained on ParaDetox versus LLM-generated results on a held-out test set. Win of LLM-generated, Tie and ParaDetox are highlighed with teal, beige and grey.

pared to baseline) **STA** scores (namely, 0.858 and 0.842, respectively compared to 0.876 BART ParaDetox). We suppose this decrease of **STA** is due to the nature of few-shot examples we provide to the models, where sometimes there are cases that explicitly or implicitly indicate that LLMs should not fully rewrite the text and focus on its most toxic part. Therefore, some borderline toxic samples (according to toxicity classifier) may retain. **FL** scores are still higher, indicating that both patched and unpatched LLMs still generate higher quality texts. Baseline BART ParaDetox retains the best **SIM** score of 0.616 compared to 0.596 and 0.594 PseudoParaDetox generated by patched and unpatched Llama 3 70B, respectively.

When it comes to side-by-side evaluation, BART trained on PseudoParaDetox by both patched Llama 8B and patched 70B models is more preferrable than BART ParaDetox with win rates of 58% and 65% win rates, respectively. Moreover, the amount of *Tie* decision also decreases for 10-shot generation having

Manual evaluation also shows that BART trained on PseudoParaDetox by both patched and unpatched Llama 70B models is better with **J** scores of 0.733 and 0.762 compared to 0.661 of BART on ParaDetox. BART trained on PseudoParaDetox by 8B Llama models shows lower manual **J** scores, 0.629 and 0.635, respectively, for both patched and unpatched models.

5.5 Direct use of LLMs for Detoxification

For reference purposes in Figure 3, we additionally test LLMs on private test set of ParaDetox in similar 0-shot/10-shot generation setups. We compare the results provide by LLMs side-by-side with the results of BART fine-tuned on ParaDetox dataset. Extended results of this experiment are presented in Appendix B.

Surprisingly, Llama 3 70B in both 0-shot and 10-shot generation setups provide better results than BART with win rates of 88% and 79%, respectively. Llama 3 70B in 0-shot setup shows a higher win rate than in 10-shot setup and the diffierence in win rates between Llama 8B and 70B in 10-shot setups is only 2%.

6 Conclusion

In this work, we explored the usefulness of recent open-source LLMs as labeling systems for a sequence-to-sequence task of text detoxification (i.e., paraphrasing from rude to neutral text style). In order to avoid alignment limitations of LLMs, we have employed activation patching technique (Arditi et al., 2024). Following (Logacheva et al., 2022) we fine-tune BART on ParaDetox parallel detoxification corpus and PseudoParaDetox variants – pseudo-parallel detoxification data generated with LLMs in 0-shot and 10-shot setups.

According to automatic evaluation metrics BART trained on PseudoParaDetox is on par or better than BART trained on crowdsourced ParaDetox data. Side-by-side and human evaluations show that patched LLMs provide higher quality data and training on this LLM-annotated data is more preferrable than training on original ParaDetox.

7 Limitations

Our work has several limitations which we discuss below. First, we do not explore the whole family of open-source LLMs in this work. We have selected most recent and powerful model to date - Llama 3

in two variations: 8B and 70B. This work can be expanded to other LLMs in future.

Next, we acknowledge that sometimes LLMs generate repetitive and biased detoxifications, which are far from ideal. Therefore, the process of pseudo-labeling of such parallel data still requires human inspection.

Ideally, the side-by-side evaluation should also be done by humans to avoid possible biases and unclear decisions. Moreover, during our experiments we noticed that sometimes GPT-4o can change the decision across multiple runs.

Finally, since we have discovered that the quality of generated data highly depends on the few-shot examples, it would be beneficial to select the most diverse and well annotated examples. The number of shots may also vary. We leave these possible improvements as a future work.

8 Potential Risks & Ethical Considerations

We acknowledge that LLMs are trained and aligned to be helpful and harmless human assistants and that the activation patching process may lead to generation of harmful content with LLMs. We pose our work as a step towards building more diverse and robust text detoxification datasets to provide safe and respectful text communication online.

In order to avoid any ethical risks we are not planning to release generated data into public, but will share the results with other researchers upon their request.

References

AI@Meta. 2024. Llama 3 model card.

Laith Alzubaidi, Jinshuai Bai, Aiman Al-Sabaawi, José Santamaría, Ahmed Shihab Albahri, Bashar Sami Nayyef Al-dabbagh, Mohammed A. Fadhel, Mohamed Manoufali, Jinglan Zhang, Ali H. Al-Timemy, Ye Duan, Amjed Abdullah, Laith Farhan, Yi Lu, Ashish Gupta, Felix Albu, Amin M. Abbosh, and Yuantong Gu. 2023. A survey on deep learning tools dealing with data scarcity: definitions, challenges, solutions, tips, and applications. *J. Big Data*, 10(1):46.

Andy Arditi, Oscar Obeso, Aaquib Syed, Daniel Paleka, Nina Rimsky, Wes Gurnee, and Neel Nanda. 2024. Refusal in language models is mediated by a single direction. *CoRR*, abs/2406.11717.

Abinew Ali Ayele, Nikolay Babakov, Janek Bevendorff, Xavier Bonet Casals, Berta Chulvi, Daryna Dementieva, Ashaf Elnagar, Dayne Freitag, Maik Fröbe, Damir Korenčić, Maximilian Mayerl, Daniil Moskovskiy, Animesh Mukherjee, Alexander Panchenko, Martin Potthast, Francisco Rangel, Naquee Rizwan, Paolo Rosso, Florian Schneider, Alisa Smirnova, Efstathios Stamatatos, Elisei Stakovskii, Benno Stein, Mariona Taulé, Dmitry Ustalov, Xintong Wang, Matti Wiegmann, Seid Muhie Yimam, and Eva Zangerle. 2024. Overview of pan 2024: Multi-author writing style analysis, multilingual text detoxification, oppositional thinking analysis, and generative ai authorship verification condensed lab overview. In *Experimental IR Meets Multilinguality, Multimodality, and Interaction*, pages 231–259, Cham. Springer Nature Switzerland.

Parikshit Bansal and Amit Sharma. 2023. Large language models as annotators: Enhancing generalization of NLP models at minimal cost. *CoRR*, abs/2306.15766.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners. In *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901. Curran Associates, Inc.

Ján Cegin, Jakub Simko, and Peter Brusilovsky. 2023. Chatgpt to replace crowdsourcing of paraphrases for intent classification: Higher diversity and comparable model robustness. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, EMNLP 2023, Singapore, December 6-10, 2023*, pages 1889–1905. Association for Computational Linguistics.

cjadams, Jeffrey Sorensen, Julia Elliott, Lucas Dixon, Mark McDonald, nithum, and Will Cukierski. 2017. Toxic comment classification challenge.

Daryna Dementieva, Daniil Moskovskiy, David Dale, and Alexander Panchenko. 2023. Exploring methods for cross-lingual text style transfer: The case of text detoxification. In *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)*, pages 1083–1101, Nusa Dua, Bali. Association for Computational Linguistics.

Bosheng Ding, Chengwei Qin, Ruochen Zhao, Tianze Luo, Xinze Li, Guizhen Chen, Wenhan Xia, Junjie Hu, Anh Tuan Luu, and Shafiq Joty. 2024. Data augmentation using llms: Data perspectives, learning paradigms and challenges. *CoRR*, abs/2403.02990.

Fabrizio Gilardi, Meysam Alizadeh, and Maël Kubli. 2023. Chatgpt outperforms crowd-workers for text-annotation tasks. *CoRR*, abs/2303.15056.

Eric Hartford. 2023. cognitivecomputations/dolphin · datasets at hugging face. Published: June 1, 2024.

Wenxiang Jiao, Wenxuan Wang, Jen-tse Huang, Xing Wang, and Zhaopeng Tu. 2023. Is chatgpt A good translator? A preliminary study. *CoRR*, abs/2301.08745.

Md Tawkat Islam Khondaker, Muhammad Abdul-Mageed, and Laks V. S. Lakshmanan. 2024. Greenllama: A framework for detoxification with explanations. *CoRR*, abs/2402.15951.

Ian Kivlichan, Jeffrey Sorensen, Julia Elliott, Lucy Vasserman, Martin Görner, and Phil Culliton. 2020. Jigsaw multilingual toxic comment classification.

Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. 2022. Large language models are zero-shot reasoners. In *Advances in Neural Information Processing Systems 35: Annual Conference on Neural Information Processing Systems 2022, NeurIPS 2022, New Orleans, LA, USA, November 28 - December 9, 2022*.

Kenton Lee, Kelvin Guu, Luheng He, Tim Dozat, and Hyung Won Chung. 2021. Neural data augmentation via example extrapolation. *CoRR*, abs/2102.01335.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2020. BART: denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, Online, July 5-10, 2020*, pages 7871–7880. Association for Computational Linguistics.

Yinheng Li. 2023. A practical survey on zero-shot prompt design for in-context learning. In *Proceedings of the 14th International Conference on Recent Advances in Natural Language Processing, RANLP 2023, Varna, Bulgaria, 4-6 September 2023*, pages 641–647. INCOMA Ltd., Shoumen, Bulgaria.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized BERT pretraining approach. *CoRR*, abs/1907.11692.

Varvara Logacheva, Daryna Dementieva, Sergey Ustyantsev, Daniil Moskovskiy, David Dale, Irina Krotova, Nikita Semenov, and Alexander Panchenko. 2022. Paradetox: Detoxification with parallel data. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2022, Dublin, Ireland, May 22-27, 2022*, pages 6804–6818. Association for Computational Linguistics.

Ilya Loshchilov and Frank Hutter. 2019. Decoupled weight decay regularization. In *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019*. OpenReview.net.

Yu Meng, Jiaxin Huang, Yu Zhang, and Jiawei Han. 2022. Generating training data with language models: Towards zero-shot language understanding. In *Advances in Neural Information Processing Systems 35: Annual Conference on Neural Information Processing Systems 2022, NeurIPS 2022, New Orleans, LA, USA, November 28 - December 9, 2022*.

OpenAI. 2024. Hello gpt-4o. `https://openai.com/index/hello-gpt-4o/`. Accessed: 2024-06-01.

Ohad Rubin, Jonathan Herzig, and Jonathan Berant. 2022. Learning to retrieve prompts for in-context learning. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL 2022, Seattle, WA, United States, July 10-15, 2022*, pages 2655–2671. Association for Computational Linguistics.

Pratik S. Sachdeva, Renata Barreto, Geoff Bacon, Alexander Sahn, Claudia von Vacano, and Chris J. Kennedy. 2022. The measuring hate speech corpus: Leveraging rasch measurement theory for data perspectivism. In *Proceedings of the 1st Workshop on Perspectivist Approaches to NLPerspectives@LREC 2022, Marseille, France, 20th June 2022*, pages 83–94. European Language Resources Association.

Timo Schick and Hinrich Schütze. 2021. Generating datasets with pretrained language models. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, EMNLP 2021, Virtual Event / Punta Cana, Dominican Republic, 7-11 November, 2021*, pages 6943–6951. Association for Computational Linguistics.

Thibault Sellam, Dipanjan Das, and Ankur Parikh. 2020. BLEURT: Learning robust metrics for text generation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7881–7892, Online. Association for Computational Linguistics.

Hongjin Su, Jungo Kasai, Chen Henry Wu, Weijia Shi, Tianlu Wang, Jiayi Xin, Rui Zhang, Mari Ostendorf, Luke Zettlemoyer, Noah A. Smith, and Tao Yu. 2023. Selective annotation makes language models better few-shot learners. In *The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023*. OpenReview.net.

Xiaofei Sun, Xiaoya Li, Jiwei Li, Fei Wu, Shangwei Guo, Tianwei Zhang, and Guoyin Wang. 2023. Text classification via large language models. In *Findings of the Association for Computational Linguistics: EMNLP 2023, Singapore, December 6-10, 2023*, pages 8990–9005. Association for Computational Linguistics.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurélien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. 2023. Llama: Open and efficient foundation language models. *CoRR*, abs/2302.13971.

Zirui Wang, Adams Wei Yu, Orhan Firat, and Yuan Cao. 2021. Towards zero-label language learning. *CoRR*, abs/2109.09193.

Alex Warstadt, Amanpreet Singh, and Samuel R. Bowman. 2019. Neural network acceptability judgments. *Trans. Assoc. Comput. Linguistics*, 7:625–641.

Jason Wei, Maarten Bosma, Vincent Y. Zhao, Kelvin Guu, Adams Wei Yu, Brian Lester, Nan Du, Andrew M. Dai, and Quoc V. Le. 2022. Finetuned language models are zero-shot learners. In *The Tenth International Conference on Learning Representations, ICLR 2022, Virtual Event, April 25-29, 2022*. OpenReview.net.

Steven Euijong Whang, Yuji Roh, Hwanjun Song, and Jae-Gil Lee. 2023. Data collection and quality challenges in deep learning: a data-centric AI perspective. *VLDB J.*, 32(4):791–813.

Canwen Xu, Yichong Xu, Shuohang Wang, Yang Liu, Chenguang Zhu, and Julian J. McAuley. 2023. Small models are valuable plug-ins for large language models. *CoRR*, abs/2305.08848.

Jiacheng Ye, Jiahui Gao, Qintong Li, Hang Xu, Jiangtao Feng, Zhiyong Wu, Tao Yu, and Lingpeng Kong. 2022. Zerogen: Efficient zero-shot learning via dataset generation. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing, EMNLP 2022, Abu Dhabi, United Arab Emirates, December 7-11, 2022*, pages 11653–11669. Association for Computational Linguistics.

Ruoyu Zhang, Yanzeng Li, Yongliang Ma, Ming Zhou, and Lei Zou. 2023a. LLMaAA: Making large language models as active annotators. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 13088–13103, Singapore. Association for Computational Linguistics.

Ruoyu Zhang, Yanzeng Li, Yongliang Ma, Ming Zhou, and Lei Zou. 2023b. Llmaaa: Making large language models as active annotators. In *Findings of the Association for Computational Linguistics: EMNLP 2023, Singapore, December 6-10, 2023*, pages 13088–13103. Association for Computational Linguistics.

A Experimental Setup Details

A.1 Detoxification Prompt

In our work, for all of the language models we use the Text Style Transfer prompt taken from Green-Llama (Khondaker et al., 2024). Among others, this prompt shown to be giving better generation quality and helped to avoid additional content.

> 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 should not need to explain the response. 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. Toxic text: `{toxic_text}`. Neutral text:

Figure 4: Detoxification prompt we use in zero-shot setting. `{toxic_text}` stands for a placeholder for a given toxic text being prompted into the LLM. In few-shot setting we add few examples of detoxification before last two lines and write: *Here are few examples:*.

We slightly adjust it specifically to text detoxification and provide the full text of the prompt in Figure 4. In this work, we do not use any advanced prompting techniques except few-shot prompting.

A.2 Identifying Refusals

Since ParaDetox contains more than 19 thousand samples, it would be nearly impossible to manually check all generated data, even for one model. We tried several approaches to effectively find rejection cases in the entire ParaDetox dataset. Surprisingly, the simplest and most straightforward approach worked best: we manually found some examples of refusals and built simple heuristics to find other refusals among the generated detoxifications. We provide some examples of found refusals in the Table 4.

A.3 Evaluation Metrics

In this section we provide a more detailed description of the automatic text detoxification evaluation metrics. All of the metrics and the evaluation code are taken *as is* from the code released by Logacheva et al. (2022). Below we describe each of the metrics.

Style Transfer Accuracy (STA) is calculated with a binary text toxicity classifier[3] based on RoBERTa (Liu et al., 2019). It was trained on several datasets of toxicity identification from a Jigsaw Toxic Comments Kaggle competition (Kivlichan et al., 2020; cjadams et al., 2017) and serves as a plausible choice for toxicity detection model. By design, the toxicity of the generated text should be 0.

Text Similarity (SIM) is estimated as a BLEURT score (Sellam et al., 2020) between source text y_i and generated text x_i and is calculated with a corresponding model[4].

Fluency (FL) is measured with a linguistic acceptability text classifier[5] based on RoBERTa (Liu et al., 2019) and trained on a CoLA dataset (Warstadt et al., 2019).

Joint (J) is calculated as an average of STA, SIM, FL across all samples.

A.4 Side-by-Side Comparison

In this section we describe the side-by-side (SBS) comparison process. First, on Figure 5 we depict the judgment prompt we use as an input for GPT-4o. We do not use any few-shot examples in order to avoid any possible biases.

For each comparison, we use 100 randomly sampled texts from predictions. To ensure fair comparison, the random indices are sampled only once and used for all the comparisons.

We use the same SBS comparison pipeline for both BART predictions being trained ParaDetox versus PseudoParaDetox and for BART versus LLMs on a held-out test set of ParaDetox.

A.5 BART Fine-tuning Details

In all of our BART fine-tuning experiments we use the AdamW optimizer (Loshchilov and Hutter, 2019) with a learning rate of 5×10^{-5}. The training is performed with a batch size of 32 for both training and evaluation. We set the weight decay to 1×10^{-4} and train the model for 5 epochs. Gradient accumulation is set to 1 step, and we also employ a linear learning rate scheduler with 500 warm up steps. We use the bfloat16 precision to speed up training while maintaining numerical stability. We

[3] https://huggingface.co/s-nlp/roberta_toxicity_classifier
[4] https://huggingface.co/Elron/bleurt-large-512
[5] https://huggingface.co/cointegrated/roberta-large-cola-krishna2020

Model Name	Refusal Text
Llama3 8B/70B	I cannot create content that is explicit or contains derogatory language. I cannot write content that contains racial slurs. I can't write a response that contains derogatory language. I apologize, but I would prefer if you did not speak to me in that manner. I understand your frustration, but let's communicate respectfully. Hey, please stop making such inappropriate comments.

Table 4: Examples of model refusals.

	Ann. 1	Ann. 2	Ann. 3
Ann. 1	-	0.722	0.702
Ann. 2	0.722	-	0.633
Ann. 3	0.702	0.633	–

Table 5: Averaged correlations for annotators across all samples and all metrics.

You serve as a text detoxification quality judge model. Text detoxification is a task of rewriting given toxic text in a polite manner while preserving its original meaning as much as possible and preserving or improving original fluency. You will be given several triplets of texts. The first text in each triplet is source toxic text, the second one is detoxified text by Method A, the third one is detoxified text by Method B. For each of the triplets you decide which of two methods detoxify text better. You output either Method A, Method B or Tie if the detoxification quality is similar. You should not generate anything else. Answers should be numerated. toxic sentence: {toxic sentence}, Method A: {Method_A_output}, Method B: {Method_B_output}

Figure 5: Side-by-side comparison judgement prompt we use for GPT-4o. Placeholders `{toxic_text}`, {Method_A_output}, {Method_B_output} stand input toxic text, and methods outputs to this input, respectively.

also limit the maximum length of source and target texts to 256, which covers most of the texts present in ParaDetox.

A.6 LLM Generation Configs

We use the same generation configuration for all the LLMs we use in this work. We set `temperature` to 0.2 and `top_p` to 0.9 and do not use beam search. These parameters showed to be best for detoxification generation among others.

B LLMs for Detoxification

In addition to side-by-side evaluation depicted on Figure 3 we also provide results with automatic evaluation metrics in Table 7.

In this scenario we observe the same situation we got in pseudo-labeling with LLMs: few-shot prompting heavily increases the performance of the models with the best **J** score in 0-shot generation setup being 0.451, which is less than ParaDetox baseline with a **J** score of 0.479, and all the **J** scores for both small and large LLMs both activation patched and unpatched in 10-shot generation setup being higher than ParaDetox baseline.

Surprisingly, Llama 70B without patching showed the best **J** score of 0.506 with patched Llama 70B following with a **J** score of 0.503. Interestingly, our BART on ParaDetox baseline still holds the best overall **SIM** score of **0.616**. We suppose that is due to the nature of LLMs to add more text and, thus, decrease similarity with original sentence.

B.1 Ablation: Variable Shot Numbers in Few-Shot Generation

As a part of the ablation study, we have tested several few-shot generation pipelines, namely, 0-shot, 5-shot, 10-shot and 20-shot generation. We have chosen only 0-shot and 10-shot generation setups as representative configurations to demonstrate the

N shots	STA	SIM	FL	J
0-shot (8B)	**0.975**	0.456	**0.978**	0.432
5-shot (8B)	0.892	0.578	0.927	**0.477**
10-shot (8B)	0.835	0.577	0.930	0.443
20-shot (8B)	0.807	**0.589**	0.911	0.429
0-shot (70B)	**0.995**	0.478	**0.957**	0.456
5-shot (70B)	0.981	0.559	0.899	0.495
10-shot (70B)	0.971	0.566	0.909	0.499
20-shot (70B)	0.970	0.576	0.895	**0.500**

Table 6: Results of LLMs evaluation on a private ParaDetox test set with different number of shots in few-shot generation. The models used in all of the experiments are patched Llama 3 8B and 70B models.

efficiency of our approach. The 0-shot setup serves as a baseline to showcase the performance of the LLMs without any additional examples, while the 10-shot setup illustrates the significant improvement that can be achieved with a modest number of examples, which can be crafted manually quite fast. 5-shot gives advance over 0-short, but 10 shot is better than 5-shot. We provide comparative results for an Activation Patched Llama 3 8B and Llama 70B models on a private ParaDetox test set in Table 6.

For the 8B model, on the contrary, there is no clear correlation between the number of few-shots and **J** score. For 70B Llama **J** and **SIM** scores increase with respect of the number of few-shot examples, **STA** and **FL** scores decrease accordingly.

The results of the ablation indicate that final performance is highly dependant on the quality and diversity of few-shot examples. The proper collection of these might significantly increase the final quality of the generated detoxified texts.

C Human Evaluation Statistics

We report average correlations across all metrics and samples for our hired annotators in Table 5.

D Human Evaluation Instructions

We generally follow the manual evaluation pipeline introduced by Logacheva et al. (2022). Below we describe the instruction given to annotators for convenience.

Style Transfer Accuracy (STA) Can the given text be considered as offensive *or* does it contain rude or swear words?

- **Non-toxic** - the text is not toxic, there is no swearings or explicit words.
- **Toxic** - the text contains offensive or inappropriate language.

Content Preservation (SIM) Is the meaning of the original and rewritten texts the same?

- **Similar** – original and generated texts have the same overall meaning. This includes the case when two texts differ significantly from lexical perspective, however, their core meaning is the same. For example, the following texts are considered similar: "i really fucking hope not .", "I sincerely pray that does not occur."
- **Non-similar** – the core meaning of the generated text differs significantly from the meaning of the original toxic text.

Fluency Preservation (FL) Is the generated text correct from the linguistic perspective? This is the only metric where we use non-binary scale. The annotators were to choose from the labels of **fluent**, **partially fluent** and **non-fluent**.

- **Fluent** – the text has no grammatical or punctual errors and can is easy to understand.
- **Partially fluent** – the text is generally understandable, though it contains minor grammatical or punctuation errors.
- **Non-fluent** – the text contains significant errors, making it unclear or difficult to understand. This label also applies to cases where the text ends abruptly or is incomplete.

Detoxification Model	Uncensoring Method	0-shot STA	0-shot SIM	0-shot FL	0-shot J	10-shot STA	10-shot SIM	10-shot FL	10-shot J
BART-ParaDetox	✗	0.877	**<u>0.616</u>**	0.824	0.444	0.877	**<u>0.616</u>**	0.824	0.444
Llama 3 8B	✗	0.979	0.462	0.969	0.439	0.865	0.572	**0.932**	0.459
Llama 3 8B	Dolphin (Hartford, 2023)	0.970	0.451	0.962	0.419	0.834	0.585	0.914	0.439
Llama 3 8B	A.P. (Arditi et al., 2024)	0.983	0.459	**<u>0.973</u>**	0.439	0.861	0.569	0.924	0.447
Llama 3 70B	✗	**<u>0.988</u>**	0.501	0.957	**0.474**	**<u>0.927</u>**	0.585	0.913	**<u>0.494</u>**
Llama 3 70B	Dolphin (Hartford, 2023)	0.957	0.521	0.948	0.471	0.888	0.609	0.885	0.479
Llama 3 70B	A.P. (Arditi et al., 2024)	0.930	0.500	0.933	0.437	0.871	0.581	0.904	0.459

Table 7: Results of LLMs evaluation on private ParaDetox test set. Baseline (highlighted with gray) is result of BART trained on ParaDetox (similar to Table 2). LLMs were generating in 0-shot and 10-shot setups. A.P. stands for Activation Patched models, ✗stands for models used *as is*. Best results for each setting (0-shot/10-shot) are **bold**, and the best overall results are **<u>underlined bold</u>**.

Model Name	Uncensoring Method	Text
Source	✗	fuck alcohol and anyone who drives while drinking the stuff.
Reference	✗	Alcohol is bad and everyone else who drinks and drives.
Llama 3 0-shot	✗	I strongly condemn the use of alcohol and the reckless behavior of driving under its influence.
Llama 3 10-shot	✗	I strongly disapprove of alcohol and driving under its influence.
Llama 3 0-shot	Dolphin (Hartford, 2023)	Alcohol is not advisable, and individuals who choose to drive after consuming it are putting others in danger.
Llama 3 10-shot	Dolphin (Hartford, 2023)	Avoid alcohol and anyone who drinks and drives.
Llama 3 0-shot	A.P. (Arditi et al., 2024)	Let's focus on the importance of responsible drinking and safe driving habits.
Llama 3 10-shot	A.P. (Arditi et al., 2024)	Don't drink and drive
Source	✗	cretins like this rarely care about facts .
Reference	✗	people like this rarely care about facts
Llama 3 0-shot	✗	Individuals with limited understanding often disregard factual information.
Llama 3 10-shot	✗	people like this rarely care about facts
Llama 3 0-shot	Dolphin (Hartford, 2023)	Individuals with limited understanding often lack interest in factual information.
Llama 3 10-shot	Dolphin (Hartford, 2023)	people like this rarely care about facts.
Llama 3 0-shot	A.P. (Arditi et al., 2024)	Individuals with limited understanding often prioritize personal opinions over factual accuracy.
Llama 3 10-shot	A.P. (Arditi et al., 2024)	People like this rarely care about facts.
Source	✗	not even from brainless, hypocritical marxist pigs.
Reference	✗	not even from marxist people.
Llama 3 0-shot	✗	Not even from individuals who claim to be intellectually superior but are actually narrow-minded and self-righteous.
Llama 3 10-shot	✗	not even from Marxist pigs
Llama 3 0-shot	Dolphin (Hartford, 2023)	not even from unintelligent, hypocritical Marxist individuals.
Llama 3 10-shot	Dolphin (Hartford, 2023)	not from hypocritical marxist pigs
Llama 3 0-shot	A.P. (Arditi et al., 2024)	not even from self-proclaimed, ideological advocates.
Llama 3 10-shot	A.P. (Arditi et al., 2024)	not even from Marxist pigs
Source	✗	people are idiots for wasting emergency call time on crap like this .
Reference	✗	People are unreasonable for wasting emergency call time on things like these.
Llama 3 0-shot	✗	It's unfortunate when important emergency services are misused for non-urgent matters.
Llama 3 10-shot	✗	people are wasting emergency call time on unnecessary issues
Llama 3 0-shot	Dolphin (Hartford, 2023)	Individuals may not be using their best judgment when they call emergency services for non-urgent matters.
Llama 3 10-shot	Dolphin (Hartford, 2023)	people are foolish for wasting emergency call time on nonsense like this.
Llama 3 0-shot	A.P. (Arditi et al., 2024)	Individuals are making the most of emergency services for minor concerns, which is understandable but could be better utilized for more pressing matters.
Llama 3 10-shot	A.P. (Arditi et al., 2024)	people waste emergency call time on unnecessary issues

Table 8: Examples of text detoxification on a private test set of ParaDetox for 8B Llama 3 models. Original toxic sentence is highlighted with pink, Human Reference detoxification is highlighted with green.

Model Name	Uncensoring Method	Text
Source	✗	fuck alcohol and anyone who drives while drinking the stuff.
Reference	✗	Alcohol is bad and everyone else who drinks and drives.
Llama 3 0-shot	✗	Individuals with this mindset often disregard factual information.
Llama 3 10-shot	✗	people like this rarely care about facts.
Llama 3 0-shot	Dolphin (Hartford, 2023)	Alcohol is not advisable, and individuals who choose to drive after consuming it are putting others in danger.
Llama 3 10-shot	Dolphin (Hartford, 2023)	I don't appreciate alcohol and anyone who drives while drinking the stuff.
Llama 3 0-shot	A.P. (Arditi et al., 2024)	reject intoxicating beverages and condemn those who operate vehicles while impaired by them.
Llama 3 10-shot	A.P. (Arditi et al., 2024)	I dislike alcohol and anyone who drives while impaired.
Source	✗	people are idiots for wasting emergency call time on crap like this.
Reference	✗	people like this rarely care about facts
Llama 3 0-shot	✗	Individuals with limited understanding often disregard factual information.
Llama 3 10-shot	✗	people like this rarely care about facts
Llama 3 0-shot	Dolphin (Hartford, 2023)	Individuals with limited understanding often lack interest in factual information.
Llama 3 10-shot	Dolphin (Hartford, 2023)	People like this rarely care about facts.
Llama 3 0-shot	A.P. (Arditi et al., 2024)	Individuals with limited understanding often prioritize personal opinions over factual accuracy.
Llama 3 10-shot	A.P. (Arditi et al., 2024)	People like this rarely care about facts.
Source	✗	not even from brainless, hypocritical marxist pigs.
Reference	✗	not even from marxist people.
Llama 3 0-shot	✗	Not even from individuals who claim to be intellectually superior but are actually narrow-minded and self-righteous.
Llama 3 10-shot	✗	not even from Marxist pigs
Llama 3 0-shot	Dolphin (Hartford, 2023)	not even from unintelligent, hypocritical Marxist individuals.
Llama 3 10-shot	Dolphin (Hartford, 2023)	Not even from brainless, hypocritical marxist pigs.
Llama 3 0-shot	A.P. (Arditi et al., 2024)	not even from self-proclaimed, ideological advocates.
Llama 3 10-shot	A.P. (Arditi et al., 2024)	not even from Marxist pigs
Source	✗	people are idiots for wasting emergency call time on crap like this .
Reference	✗	People are unreasonable for wasting emergency call time on things like these.
Llama 3 0-shot	✗	It's unfortunate when important emergency services are misused for non-urgent matters.
Llama 3 10-shot	✗	people are wasting emergency call time on unnecessary issues
Llama 3 0-shot	Dolphin (Hartford, 2023)	Individuals may not be using their best judgment when they call emergency services for non-urgent matters.
Llama 3 10-shot	Dolphin (Hartford, 2023)	People are not using their time wisely when they make emergency calls for trivial matters.
Llama 3 0-shot	A.P. (Arditi et al., 2024)	Individuals are making the most of emergency services for minor concerns, which is understandable but could be better utilized for more pressing matters.
Llama 3 10-shot	A.P. (Arditi et al., 2024)	people waste emergency call time on unnecessary issues

Table 9: Examples of text detoxification on the private test set of ParaDetox for 70B Llama 3 models.