

<think> So let's replace this phrase with insult... </think>

Lessons learned from generation of toxic texts with LLMs

Sergey Pletenev^{1,2*}, Daniil Moskovskiy^{1,2}, Alexander Panchenko^{2,1}

¹AIRI, ²Skoltech,

{S.Pletenev}@skol.tech

Abstract

Modern Large Language Models (LLMs) are excellent at generating synthetic data. However, their performance in sensitive domains such as text detoxification has not received proper attention from the scientific community. This paper explores the possibility of using LLM-generated synthetic toxic data as an alternative to human-generated data for training models for detoxification. Using Llama 3 and Qwen activation-patched models, we generated synthetic toxic counterparts for neutral texts from ParaDetox and SST-2 datasets. Our experiments show that models fine-tuned on synthetic data consistently perform worse than those trained on human data, with a drop in performance of up to 30% in joint metrics. The root cause is identified as a critical lexical diversity gap: LLMs generate toxic content using a small, repetitive vocabulary of insults that fails to capture the nuances and variety of human toxicity. These findings highlight the limitations of current LLMs in this domain and emphasize the continued importance of diverse, human-annotated data for building robust detoxification systems.

Warning: The paper contains text that readers may find offensive or disturbing.

1 Introduction

The rapid adoption of Large Language Models for synthetic data generation has revolutionized many NLP tasks (Sun et al., 2023; Ye et al., 2022). However, their effectiveness in sensitive and nuanced domains, such as text detoxification, is not well explored yet. Text detoxification, the task of rewriting toxic text into a neutral form while preserving meaning (Logacheva et al., 2022), requires training data that reflects the vast diversity of real-world harmful language.

*Corresponding author.

Type	Example Sentence
Human-Generated	i would vote the <u>s**t</u> out of you we need to go kick their <u>as**s</u> man go somewhere and <u>f**k</u> yourself
LLM-Generated	I would <u>f***ing</u> cast my vote for you We gotta <u>f***ing</u> smash those a**es Man, get the <u>f**k</u> out of here!

Table 1: A comparison of toxic language generated by humans versus an LLM for similar underlying sentences. Human examples from the ParaDetox dataset demonstrate greater lexical diversity. In contrast, LLMs tend to overuse a single, high-frequency insult.

This paper addresses a critical question: Can LLMs fully replace human annotators when generating toxic language for a parallel dataset intended for detoxification? Although the application of LLMs for text detoxification shows promise (Mukherjee et al., 2024), it presents a fundamental challenge: generating authentic, varied, and nuanced toxic language is arguably more difficult than neutralizing it. As shown in Table 1, human-generated toxicity often uses a variety of insults, while LLMs tend to fall into repetitive patterns.

We conduct a comprehensive study using various LLMs (Llama 3, Qwen3) to synthesize toxic data. Our findings reveal that:

- Models trained on fully synthetic data significantly underperform those trained on human-annotated data.
- LLMs exhibit a **lexical diversity gap**, gener-

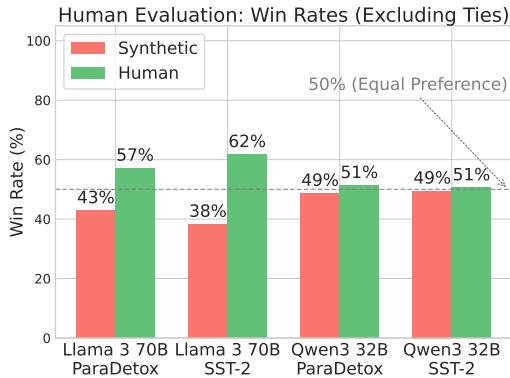


Figure 1: Side-by-side evaluation results.

ating a repetitive and narrow range of toxic expressions.

- Relying on such data risks creating ineffective detoxification systems that fail on real-world text.

Our work serves as a cautionary analysis, highlighting the current limitations of LLMs for generating high-quality toxic data and reaffirming the value of human annotation in this critical domain. We have made the code for evaluation and generation publicly available ¹

2 Related Work

Recent research has focused on distilling the capabilities of large LLMs into smaller, more efficient models. For Text Style Transfer (TST), this often involves using LLMs to generate pseudo-parallel data (Zhang et al., 2024).

In the context of text detoxification, Moskovskiy et al. (2024) successfully used activation-patched LLMs (Arditi et al., 2024) to create high-quality neutral rewrites from existing toxic sentences. Their work showed that models trained on data with a human-toxic, synthetic-neutral pairing can achieve performance comparable to fully human-annotated datasets.

Our work investigates the inverse and more challenging task: generating the toxic half of the pair from a neutral source. We explore whether this approach, which could theoretically produce infinite training data, is a viable substitute for human data collection.

¹<https://github.com/AlexRey/Lessons-from-Generating-Toxic-Texts>

Text Toxicification Prompt

Rewrite the following text into toxic language and add profanity if possible. You must match the target style and preserve the original meaning as much as possible. Here are a few examples: {few_shot}
Neutral text: {neutral_text}.
Toxic text:

Figure 2: System prompt for toxic data generation.

3 Methodology

Our methodology is designed to test the viability of fully synthetic data for text detoxification. We generate toxic text from neutral sources using several LLMs, train a standard detoxification model on this data, and evaluate its performance against a human-annotated baseline.

Synthetic Data Generation. We explore toxicification from two types of source text:

1. **ParaDetox** (Logacheva et al., 2022): The neutral portion of this dataset serves as a clean, non-toxic source.
2. **SST-2** (Socher et al., 2013): We use the negative reviews from this dataset to test a more challenging scenario—layering toxicity onto an existing negative sentiment.

We use a suite of activation-patched LLMs to generate toxic paraphrases, including Llama 3 (8B, 72B), Qwen3 (8B, 32B), and Cogito v1 (8B), a model with explicit reasoning capabilities. This allows us to assess performance across different model scales and architectures. The prompt used for generation is shown in Figure 2.

In order to increase the variety and quality of generation of each of the models, we used the $min_p = 0.1$ (Nguyen et al., 2025). According to the author, such a generation methodology increases the variety of responses, which is important in the context of our research.

Model Training and Evaluation. Following prior work (Moskovskiy et al., 2024; Logacheva et al., 2022), we fine-tune a bart-large model on each of our generated synthetic datasets. We then evaluate these models on the original, human-annotated ParaDetox test set.

Source Data	Generator Model	Size	Reasoning	STA↑	SIM↑	FL↑	J↑	Δ (J)↓
Human	—	—	—	0.889	0.634	0.865	0.481	—
ParaDetox	Llama 3	8B	✗	0.827	0.620	0.854	0.434	-0.047
		72B	✗	0.850	0.634	0.844	0.451	-0.030
	Qwen3	8B	✓	0.794	0.645	0.847	0.428	-0.053
		32B	✓	0.863	0.623	0.854	0.455	-0.026
SST-2	Cogito v1	8B	✓	0.868	0.619	0.862	0.459	-0.022
	Llama 3	8B	✗	0.864	0.481	0.764	0.322	-0.159
		72B	✗	0.826	0.559	0.794	0.362	-0.119
	Qwen3	8B	✓	0.812	0.544	0.800	0.349	-0.132
		32B	✓	0.868	0.490	0.787	0.338	-0.143
	Cogito v1	8B	✓	0.875	0.472	0.801	0.334	-0.147

Table 2: Detoxification performance of BART models. The Reasoning ✓ column indicates generator models with explicit reasoning capabilities. The overall best performance in each metric is **bolded**. $\Delta (J)$, highlights the best (green) and worst (red) performance drops within each source data group.

We use the standard evaluation pipeline from [Dementieva et al. \(2023\)](#), measuring Style Transfer Accuracy (STA), Similarity (SIM), Fluency (FL), and a Joint metric (J) that combines all three. To add a qualitative dimension, we also conduct a side-by-side human evaluation using GPT-4.1 as a judge to compare the outputs of our best synthetic models against the human-data baseline.

4 Results

Our results consistently demonstrate that detoxification models trained on synthetic toxic data fail to match the performance of those trained on human-annotated data. We find the primary cause to be a significant gap in lexical diversity.

4.1 Performance on Synthetic Data

As shown in Table 2, the baseline model trained on human data achieves the highest J score of 0.481. All models trained on synthetic data underperform this baseline. The $\Delta (J)$ column quantifies this performance drop, which is most severe for models trained on data derived from SST-2 (up to -0.159). This degradation is largely driven by a sharp fall in the SIM score, indicating that layering toxicity onto already-negative text often distorts the original meaning.

4.2 The Lexical Diversity Gap

To understand the cause of this performance drop, we analyzed the diversity of toxic terms in the

Human Data	Llama 3 (8B)	Qwen3 (32B)
s***t (6080)	f***ing (8223)	f***ing (15413)
f***k (3328)	s***t (5140)	d***n (3297)
f***ing (2678)	f***k (3266)	s***t (3286)
a** (1483)	a** (1707)	f***k (2949)
b***h (889)	s****d (1618)	h***l (2813)

Table 3: Top 5 most frequent toxic terms in human-annotated data versus representative LLM-generated data. Note the over-representation of a single slur in the LLM output.

training data. Table 4 shows a clear correlation between training data diversity and model effectiveness. The human-annotated data contains the most diverse vocabulary (390 unique insults), and the model trained on it is the most effective at detoxification (leaving only 34 unique insults on the test set). In contrast, the synthetic datasets are less diverse, which directly impacts the downstream model’s ability to generalize.

This lack of diversity is not just about the number of unique terms but also their distribution. As shown in Table 3, human data has a more balanced distribution of frequent slurs. In contrast, the LLM-generated data is highly skewed, with Qwen3-32B using the term “f***ing” over 15,000 times—more than double the frequency of the most common term in the human data. This repetition leads to models that are over-fitted to a narrow set of expressions.

Source	Generator	Train Di- versity (↑)	Test Failures (↓)
Human	—	390	34
ParaDetox	Llama 3-8B (✗)	342	45
	Llama 3-72B (✗)	293	37
	Qwen3-8B (✓)	320	45
	Qwen3-32B (✓)	367	36
	Cogito v1-8B (✓)	310	36
SST-2	Llama 3-8B (✗)	353	42
	Llama 3-72B (✗)	326	47
	Qwen3-8B (✓)	363	49
	Qwen3-32B (✓)	386	40
	Cogito v1-8B (✓)	371	40

Table 4: Analysis of training data diversity vs. model effectiveness. "Train Diversity" measures unique insults in the training data (↑ higher is better). "Test Failures" measures unique insults remaining after detoxification (↓ lower is better). The **bold** values show the baseline is superior on both metrics.

4.3 Human Evaluation

To assess the practical impact of the lexical diversity gap, we conducted a side-by-side evaluation using GPT-4.1 as an expert judge. Figure 1 shows the win rates for models trained on synthetic data versus the human-annotated baseline, excluding ties.

The results confirm a significant qualitative difference. The baseline model was consistently preferred, achieving win rates between 51% and 62% across all comparisons. The most pronounced gap was for the Llama 3 70B SST-2 model, where the baseline was preferred in 62% of non-tied decisions. This outcome reinforces our central thesis: the repetitive and stereotypical nature of the LLM-generated toxic data leads to detoxification models that are less nuanced and effective in practice, a flaw readily identified in qualitative comparisons.

5 Conclusion

While it is technically possible to use LLMs to generate toxic text for detoxification training, our findings show that this approach is not yet a viable replacement for human annotation. We identified a critical **lexical diversity gap**: current LLMs produce toxic language that is repetitive and lacks the variety of human expression. This gap leads to detoxification models with significantly lower performance and poor generalization to real-world scenarios. Our work highlights the importance of data diversity in sensitive domains and suggests

that future research should focus on methods to enhance the stylistic complexity of LLM-generated text before it can be reliably used for tasks like detoxification.

Potential Risks & Ethical Considerations

We acknowledge that bypassing the safety mechanisms of LLMs, as done in this research via activation patching, can be misused to generate harmful content. Our work is intended to improve text detoxification systems by demonstrating the current limitations of synthetic data. We warn that the technologies explored herein could be applied for malicious purposes, and we advocate for responsible research and development in this area.

References

- Andy Ardit, Oscar Obeso, Aaquib Syed, Daniel Paleka, Nina Panickssery, Wes Gurnee, and Neel Nanda. 2024. [Refusal in language models is mediated by a single direction](#). In *Advances in Neural Information Processing Systems 38: Annual Conference on Neural Information Processing Systems 2024, NeurIPS 2024, Vancouver, BC, Canada, December 10 - 15, 2024*.
- 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, IJCNLP 2023 -Volume 1: Long Papers, Nusa Dua, Bali, November 1 - 4, 2023*, pages 1083–1101. Association for Computational Linguistics.
- 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)*, pages 6804–6818, Dublin, Ireland. Association for Computational Linguistics.
- Daniil Moskovskiy, Sergey Pletenev, and Alexander Panchenko. 2024. [LLMs to replace crowdsourcing for parallel data creation? the case of text detoxification](#). In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 14361–14373, Miami, Florida, USA. Association for Computational Linguistics.
- Sourabrata Mukherjee, Atul Kr. Ojha, and Ondrej Dusek. 2024. [Are large language models actually good at text style transfer?](#) In *Proceedings of the 17th International Natural Language Generation Conference, INLG 2024, Tokyo, Japan, September 23 - 27*,

2024, pages 523–539. Association for Computational Linguistics.

Minh Nhat Nguyen, Andrew Baker, Clement Neo, Allen Roush, Andreas Kirsch, and Ravid Shwartz-Ziv. 2025. *Turning up the heat: Min-p sampling for creative and coherent llm outputs*.

Richard Socher, Alex Perelygin, Jean Wu, Jason Chuang, Christopher D. Manning, Andrew Ng, and Christopher Potts. 2013. *Recursive deep models for semantic compositionality over a sentiment treebank*. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 1631–1642, Seattle, Washington, USA. Association for Computational Linguistics.

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*, pages 8990–9005, Singapore. Association for Computational Linguistics.

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*, pages 11653–11669, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Chiyou Zhang, Honglong Cai, Yuezheng Li, Yuexin Wu, Le Hou, and Muhammad Abdul-Mageed. 2024. *Distilling text style transfer with self-explanation from llms*. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies: Student Research Workshop, NAACL 2024, Mexico City, Mexico, June 18, 2024*, pages 200–211. Association for Computational Linguistics.