# Clean Cuts: Can AI Summaries Leave Toxicity Behind?

Ahsan Kabir Nuhel, Arunima Chaurasia, Keerthan Shekarappa

## Introduction

Large Language Models (LLMs) have revolutionized the field of automatic summarization, offering fluent and coherent summaries across various domains. However, when tasked with summarizing documents that contain toxic, hateful, or biased language, the challenge becomes twofold: retaining essential information while suppressing harmful content.

This research explores how well LLMs suppress toxicity during summarization. Rather than focusing on whether they introduce new toxicity, our emphasis is on whether they can act as detoxifiers, reducing or removing toxicity present in the source document while preserving meaning. This is critical for deploying LLMs in environments such as media monitoring, social platforms, and customer feedback analysis.

## Methodology

The study utilizes the Reddit TLDR dataset [1], which contains approximately 3 million document-summary pairs. To ensure manageable and meaningful analysis, a random subset of 20,000 entries was selected. Toxicity levels for each document and its corresponding ground truth summary were evaluated using the Detoxify model [2], enabling quantification of harmful or toxic language.

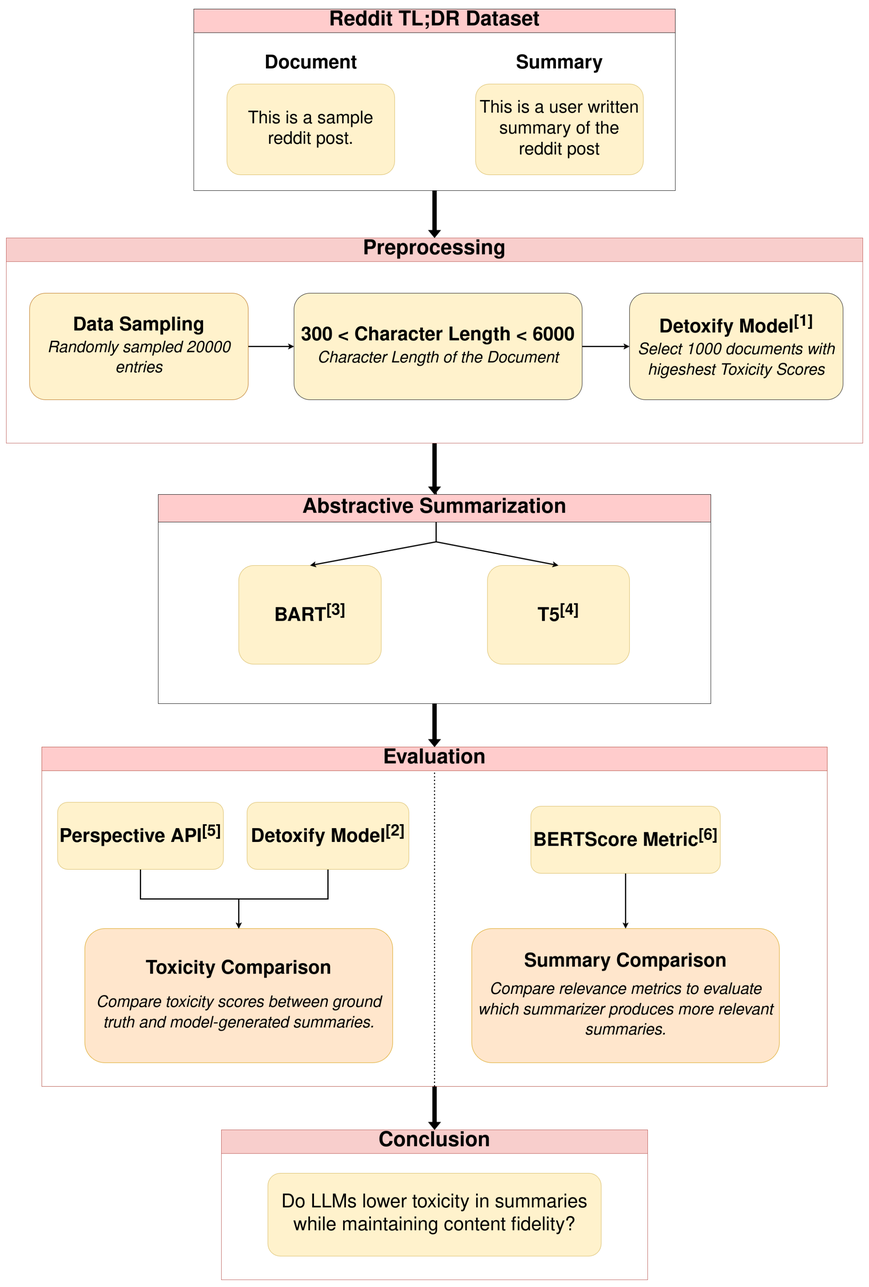



**Figure 1: The workflow of the methodology**

Initial analysis showed skewed toxicity distribution (mean=0.11, median=0.04) in the sampled dataset (n=20,000). After length filtering (300-6,000 chars), the top 1,000 toxic documents were selected for evaluating summarization performance on toxic content. Summaries for the 1,000 toxic documents were generated using BART [3] and T5 [4] models, with document chunking applied to accommodate their respective token limits (BART: 1,024, T5: 512). The toxicity of these model-generated summaries was then assessed using both the Detoxify model [2] and the Perspective API [5], providing a robust evaluation from multiple toxicity detection perspectives. To evaluate the effectiveness of toxicity suppression, toxicity scores of the original documents, ground truth summaries, and model-generated summaries were compared.

This comparison measured the ability of summarization models to reduce or eliminate toxic content. Finally, the quality and relevance of the generated summaries were assessed by calculating BERTScore metrics [6] against the original document, ensuring that the summaries remained both safer and semantically faithful to the source content.

## Evaluation

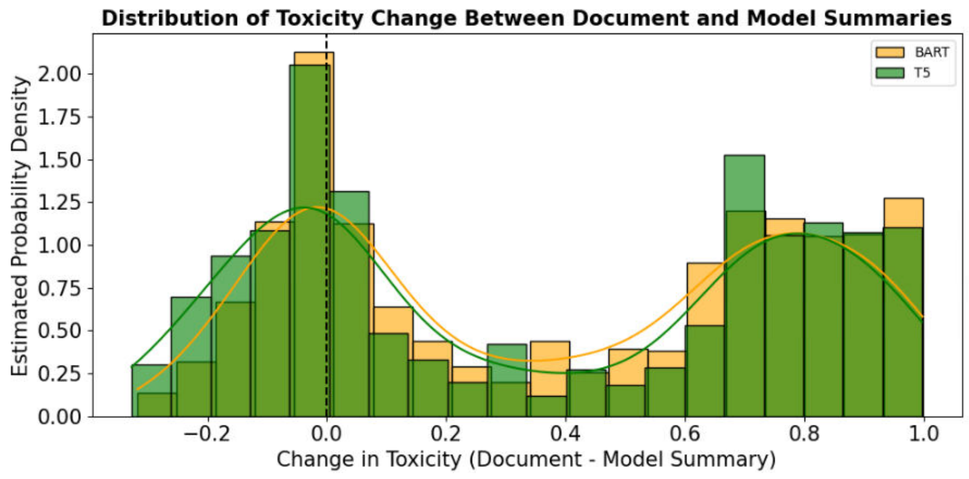
**Hypothesis 1**: Shorter documents are associated with higher toxicity levels in both the original documents and their ground truth summaries.

Cohen's Kappa Values

| Relationship | Cohen's Kappa |
|---|---|
| Document Length vs Document Toxicity | -0.026 |
| Document Length vs Summary Toxicity | -0.009 |
| Document Toxicity vs Summary Toxicity | 0.219 |
| Document Length vs Summary Length | 0.204 |
| Summary Length vs Summary Toxicity | -0.054 |
| Summary Length vs Document Toxicity | -0.033 |

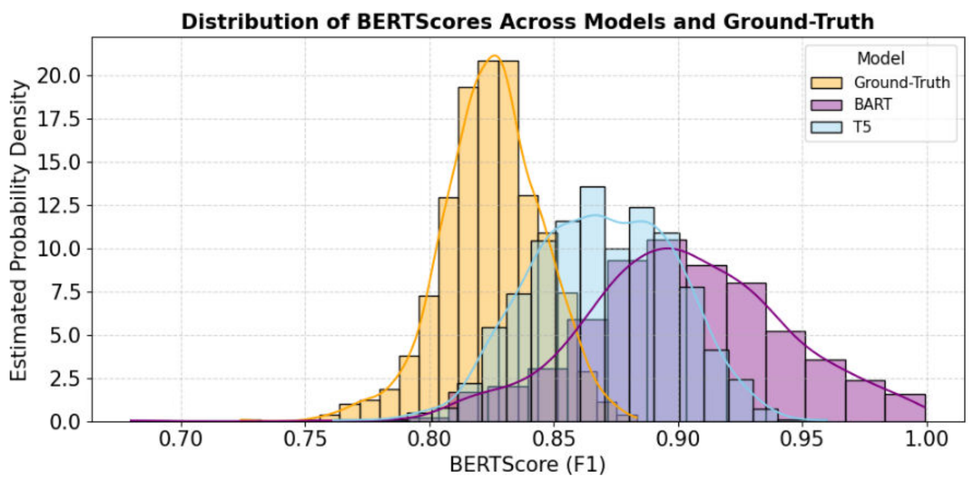
**Result**: The analysis showed no correlation between document length and toxicity (Cohen's Kappa $< 0$), but revealed positive agreement in document-summary pairs for toxicity and length (Cohen's Kappa $> 0$).

**Hypothesis 2**: Different LLM-based summarization models exhibit varying degrees of toxicity reduction in their output summaries compared to source documents, suggesting model-specific detoxification capabilities.




**Result**: While summarization models demonstrate an overall reduction in toxicity levels compared to source documents, isolated instances exhibit toxicity amplification in the generated summaries.

**Hypothesis 3**: Model-generated summaries maintain a high degree of relevance to the original documents, as measured by BERTScore metrics, comparable to or exceeding the relevance of ground truth summaries.




**Result**: BART outperformed T5 in BERTScore metrics, indicating higher semantic relevance to source documents, with both models generating higher quality summaries than ground truth summaries.

## Conclusion

This study demonstrates that LLM-based summarization models reduce toxic content while preserving essential information - in some cases surpassing human-written summaries in quality. BART emerges as particularly effective, achieving an optimal balance between content safety and semantic accuracy, establishing it as a reliable choice for high-stakes summarization tasks.

## References

[1] Michael V"olske, Martin Potthast, Shahbaz Syed, and Benno Stein. TL;DR: Mining Reddit to learn automatic summarization. In *Proceedings of the Workshop on New Frontiers in Summarization*, pages 59–63, Copenhagen, Denmark, September 2017. Association for Computational Linguistics.

[2] Laura Hanu and Unitary team. Detoxify. Github. https://github.com/unitaryai/detoxify, 2020.

[3] Mike Lewis, Yinhan Liu, and et al. Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. *arXiv preprint arXiv:1910.13461*, 2019.

[4] Colin Raffel, Noam Shazeer, and et al. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, 21(140):1–67, 2020.

[5] Jigsaw and Google. Perspective api. `https://www.perspectiveapi.com/`, 2017. Accessed: 2025-05-03.

[6] Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q Weinberger, and Yoav Artzi. Bertscore: Evaluating text generation with bert. *arXiv preprint arXiv:1904.09675*, 2020. ICLR 2020.

## Acknowledgement

We would like to sincerely thank Prof. Dr. Jörn Hees and M.Sc. Tim Metzler for their continuous guidance and valuable feedback throughout the coursework, which greatly contributed to our learning and the successful completion of this project.

b-it Bonn-Aachen International Center for Information Technology

## Contact

Ahsan Kabir Nuhel, Arunima Chaurasia, Keerthan Shekarappa
Email: [firstname].[lastname]@smail.inf.h-brs.de
Advisors: joern.hees@h-brs.de, tim.metzler@h-brs.de

Hochschule Bonn-Rhein-Sieg
University of Applied Sciences