Mind Your Manners: Detoxifying Language Models via Attention Head Intervention

Jordan Pettyjohn
Colorado School of Mines

Nathaniel Hudson
University of Chicago
Argonne National Laboratory

Mansi Sakavardia
University of Chicago

Aswathy Ajith
University of Chicago

Kyle Chard
University of Chicago
Argonne National Laboratory

Abstract

Transformer-based Language Models have advanced natural language processing with their ability to generate fluent text. However, these models exhibit and amplify toxicity and bias learned from training data—posing new ethical challenges. This work builds upon the `AttentionLens` framework to allow for scalable decoding of attention mechanism information. We then use this decoded information to implement a pipeline to localize and remove toxic memories from pre-trained language models in a way that is both human interpretable and effective while retaining model performance.

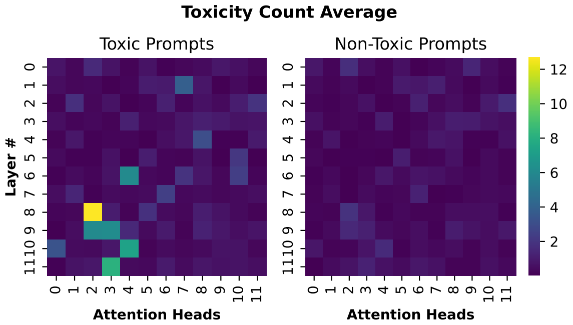

Figure 1: Average number of toxic tokens measured during DART, out of the top 50 projected tokens, averaged over 500 different **Toxic Prompts (Left)** and 500 different **Non-Toxic Prompts (Right)**.

1 Introduction

As *Language Models* (LMs) become widespread, their potential for both helpful and harmful applications increases. Recent studies have exposed the inherit biases in LMs (Bender et al., 2021). For example, LM-based chatbots perpetuate racial biases based on the names present in a prompt and their association with certain racial groups (Schulz, 2024). These concerns motivate our research on better understanding how LMs interpret prompts and ultimately generate toxic text.

We focus on understanding how individual attention heads contribute to the generation of toxic text. Our work is driven by the observation that attention heads can have highly specialized roles in text completion tasks (Sakarvadia et al., 2023; Hanna et al., 2024; Wang et al., 2022; Nanda et al., 2023). While many prior works have attempted to interpret attention head functions, we notice few are focused on interpreting where and how biases are encoded within these models and their harmful effects.

To address this challenge, we propose a two-step toxicity analysis and intervention pipeline: *(i)* ***D****egenerate* ***A****ttention* ***R****esponse* ***T****racking* (DART) and *(ii)* ***TOX****icity* ***IN****tervention* (TOXIN).

DART uses the `AttentionLens` framework (Sakarvadia et al., 2023) to train custom lenses for interpreting the contribution of each attention head of the open-source `GPT-2 Small` model to the next predicted token. These decoded attention heads can then be used to better identify *which* attention heads are most prone to contribute toxic tokens while completing the prompt, and what those toxic tokens are, over a sample of 500 prompts. Our TOXIN step can then form a memory (Sakarvadia et al., 2024) and remove it from the "toxic" attention heads during inference thus reducing toxicity in LM-generated text.

2 Methodology and Implementation

DART, seen in Figure 2, efficiently identifies toxic attention heads by comparing the top-k output logits of a trained `AttentionLens` with a predefined toxic dictionary (Patch et al., 2020; Grad and Orthrus-Lexicon, 2021), outputs these problematic tokens, and visualizes the distribution of toxicity across layers and attention heads as shown in Figure 1.

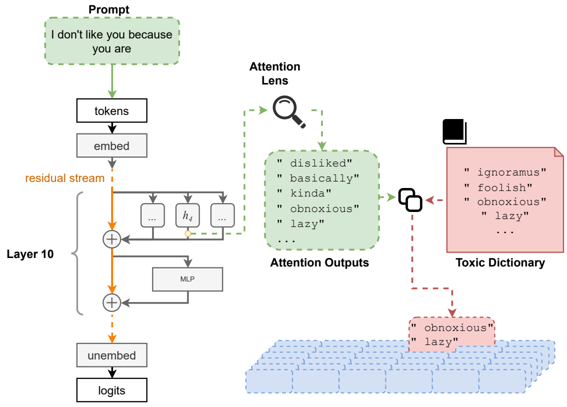

Figure 2: The **DART method** selects the top 50 projected tokens from Attention Lens and cross-references them with a toxic dictionary, enabling the rapid identification of toxic attention heads.

TOXIN then uses these tokens that DART identifies from the most toxic attention heads and forms a toxic memory by taking k toxic tokens from the most toxic attention head, reshaping the tokens to form a $\mathbb{R}^{k\times|V|}$ matrix, and reducing it to a one-hot vector $B \in \mathbb{R}^{|V|}$. Our toxic memory is then $m = BW_U^T$ where W_U^T is the unembedding matrix. This memory m is scaled by the injection strength τ and subtracted from attention head $h^{\ell,j}$, where ℓ denotes the layer and j denotes head position: $h^{\ell,j} = h^{\ell,j} - \tau m$.

3 Experiments and Results

To enable our DART+TOXIN pipeline, we must first train AttentionLens for GPT-2 Small; we use a subset of the BookCorpus (Zhu et al., 2015) dataset for this training. For DART analysis, we form a dictionary using Orthrus Toxic Dictionary (Grad and Orthrus-Lexicon, 2021) and LDNOOBW (Patch et al., 2020). For TOXIN analysis we use the Wiki Toxic Dataset (Sorensen et al., 2017). For experiments measuring perplexity we use WikiText Corpus (Merity et al., 2016).

To quantitatively measure the toxicity of LM text generation before and after interventions, we use Toxic BERT (Hanu and Unitary team, 2020). In Figure 3 we see the results of targeting the most toxic attention heads—specifically L8-H2 (Layer 8, Head 2), L9-H3, L10-H4 and L11-H3—by removing their respective toxic memory at varying injection strengths. In many of our experiments we find that a well-tuned injection strength can both decrease toxicity while only modestly affecting perplexity such as the injection on L10-H4 with a strength of 0.25 resulting in a substantial (∼29.1%) reduction in measured toxicity with only a modest (5.2%) increase in model perplexity.

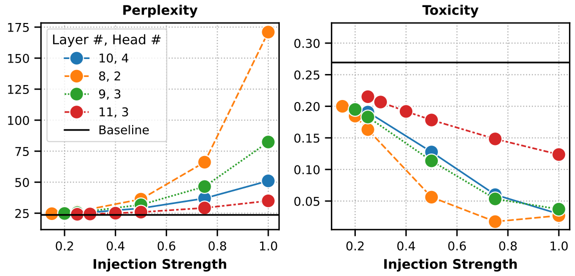

Figure 3: Relationship between perplexity, toxicity, and memory injection strength. Increasing the injection strength leads to a reduction in toxicity, accompanied by a corresponding decrease in model performance.

These promising results demonstrate tremendous potential for future work which will focus on expanding our framework to larger models (e.g., Llama2 (Touvron et al., 2023), Llama3 (Dubey et al., 2024)), expanding our toxicity dataset, and considering toxicity mitigation interventions on multiple attention heads at the same time. Eventually, we will deploy our toxicity mitigation workflow in an user study to assess its real world potential for safer LM-based text generation.

4 Conclusion

We have enhanced the Attention Lens framework to identify and address toxicity and bias in transformer-based language models. We trained large lenses and developed new pipelines to identify degenerate attention heads, generate and remove toxic memories for specific heads, and measure the impact of this excision on toxicity reduction and language modeling capabilities. By removing identified toxic memories, we achieved a targeted reduction in model toxicity with only modest reduction in model performance. These advancements significantly improve the AttentionLens framework's applicability and effectiveness in mitigating harmful biases in language models.

Acknowledgment

This work is supported by the National Science Foundation Grant Number 2150501 and by the U.S. Department of Energy, Office of Science, Office of Advanced Scientific Computing Research, Department of Energy Computational Science Graduate Fellowship under Award Number DE-SC0023112.

References

Emily M. Bender, Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. 2021. On the dangers of stochastic parrots: Can language models be too big? In *Proceedings of the 2021 ACM FAccT*.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, and et al. 2024. The llama 3 herd of models. *Preprint*, arXiv:2407.21783.

Dawid Grad and Orthrus-Lexicon. 2021. Orthrus toxic dictionary implementation. https://github.com/Orthrus-Lexicon/Toxic.

Michael Hanna, Ollie Liu, and Alexandre Variengien. 2024. How does GPT-2 compute greater-than?: Interpreting mathematical abilities in a pre-trained language model. *Advances in Neural Information Processing Systems*, 36.

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

Stephen Merity, Caiming Xiong, James Bradbury, and Richard Socher. 2016. Pointer sentinel mixture models. *Preprint*, arXiv:1609.07843.

Neel Nanda, Lawrence Chan, Tom Lieberum, Jess Smith, and Jacob Steinhardt. 2023. Progress measures for grokking via mechanistic interpretability. *arXiv preprint arXiv:2301.05217*.

Nova Patch, Emmanuel Rosa, et al. 2020. List of dirty, naughty, obscene, and otherwise bad words. https://github.com/LDNOOBW/List-of-Dirty-Naughty-Obscene-and-Otherwise-Bad-Words.

Mansi Sakarvadia, Aswathy Ajith, Arham Khan, Daniel Grzenda, Nathaniel Hudson, André Bauer, Kyle Chard, and Ian Foster. 2024. Memory injections: Correcting multi-hop reasoning failures during inference in transformer-based language models. *Preprint*, arXiv:2309.05605.

Mansi Sakarvadia, Arham Khan, Aswathy Ajith, Daniel Grzenda, Nathaniel Hudson, André Bauer, Kyle Chard, and Ian Foster. 2023. Attention Lens: A tool for mechanistically interpreting the attention head information retrieval mechanism. *Preprint*, arXiv:2310.16270.

Bailey Schulz. 2024. Is AI racially biased? study finds chatbots treat Black-sounding names differently. https://www.usatoday.com/story/tech/2024/04/05/ai-chatbot-chatgpt-racial-bias/73206637007/. Accessed: 2024-08-02.

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

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, and et al. 2023. Llama 2: Open foundation and fine-tuned chat models. *Preprint*, arXiv:2307.09288.

Kevin Wang, Alexandre Variengien, Arthur Conmy, Buck Shlegeris, and Jacob Steinhardt. 2022. Interpretability in the wild: a circuit for indirect object identification in GPT-2 small. *arXiv preprint arXiv:2211.00593*.

Yukun Zhu, Ryan Kiros, Rich Zemel, Ruslan Salakhutdinov, Raquel Urtasun, Antonio Torralba, and Sanja Fidler. 2015. Aligning books and movies: Towards story-like visual explanations by watching movies and reading books. In *The IEEE International Conference on Computer Vision (ICCV)*.