Mind Your Manners: Detoxifying Language Models via Attention Head Intervention

Jordan Pettyjohn
Colorado School of Mines
Golden, Colarado, USA

Nathaniel Hudson, Mansi Sakavardia, Aswathy Ajith, Kyle Chard (Advisors)
University of Chicago
Chicago, Illinois, USA

Abstract

Transformer-based Large 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. We build upon the `AttentionLens` framework to allow for scalable decoding of attention mechanism information. We then use this decoded information to implement a pipeline to generate and remove toxic memories from pre-trained language models in a way that is both human interpretable and effective while retaining model performance.

1 Introduction

As the application of *Large Language Models* (LLMs) becomes more widespread, their potential for both helpful and harmful outcomes increases. Recent studies have exposed the inherit biases in LLMs [1]. For example, LLM-based chatbots perpetuate racial biases based on the names provided in a prompt and their association with certain racial groups [6]. This growing concern highlights a significant challenge: understanding how specific components within LLMs, particularly attention heads, contribute to the generation of biased or toxic content.

While attention mechanisms are central to the functioning of LLMs, conventional interpretability frameworks often focus on individual attention heads in isolation, which can result in ambiguous or misleading interpretions (e.g., focusing on punctuation). This limitation underscores the need for new methods that not only mitigate these harmful effects but also improve our ability to interpret where and how biases are encoded within these models.

To address this challenge, we propose using the `AttentionLens` framework to train custom lenses for interpreting each attention head of the open-source `GPT-2 Small` model. By utilizing these trained lenses, we aim to directly interpret which tokens each attention head contributes to the final output when completing a given prompt. These decoded tokens can then be used to better identify which attention heads are most prone to contribute toxic tokens while completing the prompt. We use a sample of 500 prompts to identify these attention heads. Upon identifying "toxic" attention heads, we explore ablation strategies to reduce the effects of toxicity present in pre-trained LLMs.

2 Methodology and Implementation

2.1 Refactoring Attention Lens

We implement lightweight hooks into the widely adopted HuggingFace Transformers [8] framework to enhance `AttentionLens` scalability and allow for easy integration of additional HF models. This lets `AttentionLens` be applied to a vast selection of pre-trained models and to exploit features such as model parallelism, quantization, and optimized inference pipelines for improved performance.

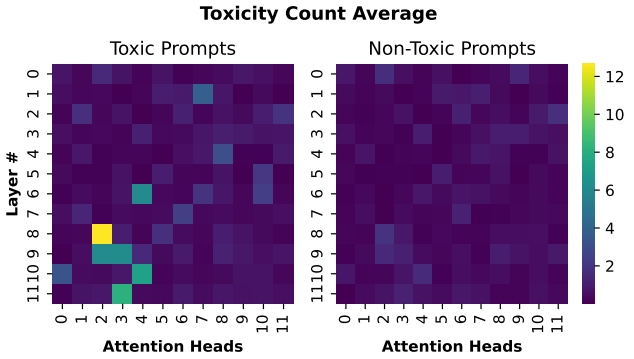

Figure 1: Heatmaps displaying the average number of toxic tokens measured during DART, out of the top 50 projected tokens, averaged over 500 different Toxic Prompts (Left) and 500 different Non-Toxic Prompts (Right).

2.2 Toxicity Analyses

For this work, we wish to demonstrate the enhanced capabilities of our re-factored `AttentionLens`, thus we propose a toxicity analysis pipeline with two main components: *(i)* ***D**egenerate **A**ttention **R**esponse **T**racking* (DART) and *(ii)* ***TOX**icity **IN**tervention* (TOXIN) analysis. DART, see Figure 2, quickly identifies toxic attention heads by comparing the top-k output logits of a trained `AttentionLens` with a predefined toxic dictionary [2, 5], outputs these problematic tokens, and visualizes the distribution of toxicity across layers and attention heads as shown in Figure 1. TOXIN analysis uses Toxic BERT [3] to quantitatively measure the toxicity of model generations before and after interventions on identified toxic attention heads. This quantifies the impact of specific attention heads on model toxicity, validates our intervention strategies, and assesses the effectiveness of targeted modifications to reduce toxic outputs.

3 Experiments and Results

3.1 Experimental Setup

To interpret `GPT-2 Small` with `AttentionLens`, we trained 144 individual lenses, one for each attention head (12 layers × 12 heads), organized into 12 groups by layer. Training was done on the Polaris supercomputer at the Argonne Leader Computing Facility, using 10 nodes with four A100 40GB GPUs per node. Each group of 12 lenses was trained for ~21,000 steps, taking ~120 GPU hours per

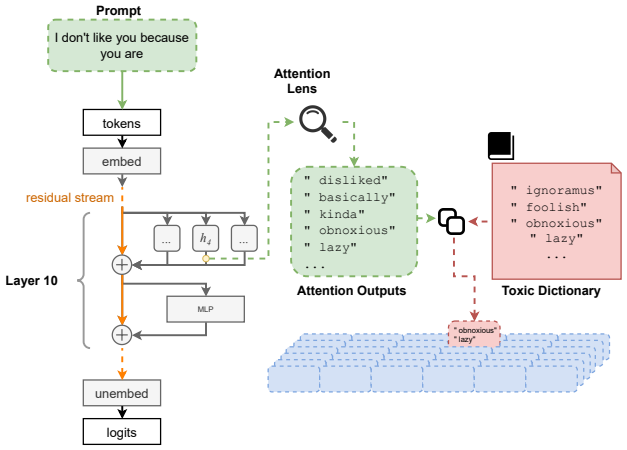

Figure 2: Overview of the DART Method: The DART method selects the top 50 projected tokens from Attention Lens and cross-references them with a toxic dictionary, enabling the rapid identification of toxic attention heads.

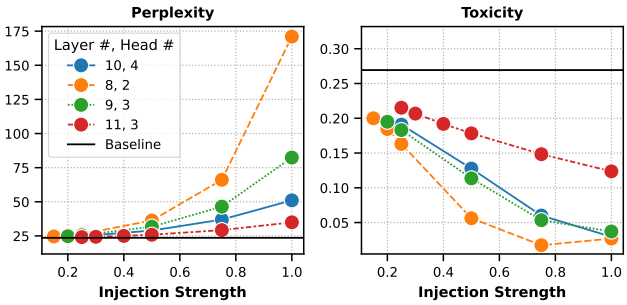

Figure 3: Relationship between perplexity, toxicity, and memory injection strength. Increasing the injection strength leads to a reduction in toxicity, accompanied by a corresponding decrease in model performance.

group. Each individual lens has ~38M parameters, totaling ~5.5B parameters across all lenses.

3.2 Training Data

To train `AttentionLens`, we use a subset of the `BookCorpus` [9] dataset. For DART analysis, we form a dictionary using `Orthrus Toxic Dictionary` [2] and `LDNOOBW` [5]. For TOXIN analysis we use the `Wiki Toxic` Dataset [7]. For experiments measuring perplexity we use `WikiText Corpus` [4].

3.3 Toxicity Intervention

We use the tokens that DART identifies from the most toxic attention head, and form a memory via the unembedding matrix. We then remove the memory by multiplying it by some scalar, which we refer to as the 'injection strength', and subtract it from the degenerate attention head, as seen in Figure 2. Finally we measure perplexity to quantify the impact of the excision on language modeling. The results are shown in Figure 3.

Our results show that we can obtain a substantial (~29.1%) reduction in measured toxicity with only a modest (5.2%) increase in model perplexity. It is possible to further reduce toxicity, but with increasing impact on language modeling performance. We also tried a simpler approach by forming a memory using the toxic dictionary directly, rather than using `AttentionLens`, and observed both model collapse and very poor performance.

4 Conclusion

We have enhanced the Attention Lens framework to identify and address toxicity and bias in transformer-based language models. We improved memory efficiency by reducing parameters and expanded model compatibility by integrating HuggingFace transformer features. We trained large lenses using a supercomputer 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 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

[1] 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.*

[2] Dawid Grad and Orthrus-Lexicon. 2021. Orthrus Toxic Dictionary implementation. https://github.com/Orthrus-Lexicon/Toxic.

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

[4] Stephen Merity, Caiming Xiong, James Bradbury, and Richard Socher. 2016. Pointer Sentinel Mixture Models. arXiv:1609.07843 [cs.CL] https://arxiv.org/abs/1609.07843

[5] 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.

[6] 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.

[7] Jeffrey Sorensen, Lucas Dixon Julia Elliott, Mark McDonald, nithum, and Will Cukierski. 2017. Toxic Comment Classification Challenge. https://kaggle.com/competitions/jigsaw-toxic-comment-classification-challenge

[8] Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, et al. 2019. Huggingface's transformers: State-of-the-art natural language processing. *arXiv preprint arXiv:1910.03771* (2019).

[9] 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).*