Mind Your Manners: Detoxifying Language Models

Jordan Pettyjohn[1] Nathaniel Hudson[2,3] (Advisor) Mansi Sakarvadia[2] (Advisor)
Aswathy Ajith[2] (Advisor) Kyle Chard[2,3] (Advisor)

[1]Colorado School of Mines [2]University of Chicago [3]Argonne National Laboratory

Problem Statement & Goals

Transformer-based **Language Models** (LMs) can **exhibit and amplify toxicity and bias** learned from training data, posing ethical challenges. We seek to **enhance the Attention Lens** [1] framework to **decode attention information**, identify and **localize toxic behavior**, and **mitigate these issues effectively**.

Abstract

Using the `AttentionLens` framework, we train custom lenses to interpret each attention head of the `GPT-2 Small` model. These lenses allow us to **directly determine which tokens each attention head contributes** to the model's output for a given prompt. By analyzing these tokens, we **identify attention heads most prone to contributing toxic content**. With a **sample of 500 prompts**, we pinpoint these toxic heads and explore **ablation strategies to mitigate toxicity** in pre-trained LLMs.

Methodology and Implementation

Refactoring Attention Lens. Training custom lenses on attention heads is computationally demanding. Thus we refactor Attention Lens to **optimize memory usage** and **improve scalability**.

- **Scalability:** Uses `HuggingFace Transformers` [2] to leverage model parallelism, quantization, and optimized inference.
- **Implementation:** Lightweight hooks for efficient extraction and injection of attention information [3].

Introducing the Dart Method

1. Assess the contribution of individual or multiple attention heads to the model's overall toxicity.
2. Evaluate and compare various methods for ablating toxicity, while preserving model performance.
3. Apply a toxic dictionary as a blacklist to expose undesirable tokens based on attention head contributions.
4. Develop and implement new pipelines to measure toxicity and model performance before and after the ablation process.

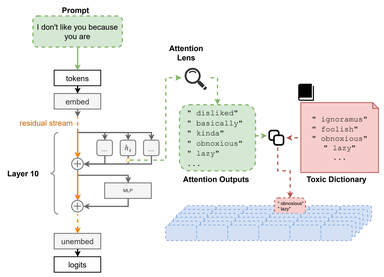


Table 1. Comparison of the model complexity of produced lenses (via parameter counts) between the original implementation of AttentionLens and the refactor for this work.

	Layers				
	Name	Type	Parameters	Total Parameters	Total Size (GB)
Original	HookedTransformer	`hooked_model`	163M	789M	3.085
	LensA	`attn_lens`	502M		
	HookedTransformer	`model`	163M		
Refactor	GPT2LMHeadModel	`model`	124M	**588M** ($\approx 0.745\times$)	**2.299**
	LensA	`attn_lens`	463M		

Experimental Setup

System Information

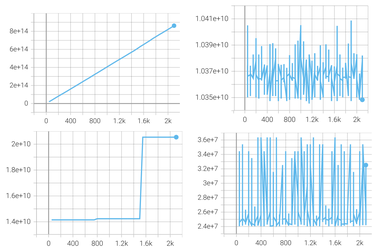


Figure 1. Metrics from training our `AttentionLens`.

For this work, we use the **Polaris** supercomputer maintained by the **Argonne Leadership Computing Facility**. This machine is made up of 560 nodes, with each compute node being equipped with **1 AMD EPYC "Milan" processor** and **4 NVIDIA A100 GPUs**.

Training Configuration

- **Training:** 144 lenses (12 layers $\times$ 12 heads) trained on 10 nodes with 4 A100 GPUs.
- **Data:** Attention Lens trained on `Book Corpus`, DART dictionary combines `Orthrus Toxic Dictionary` and `List of Dirty Naughty Obscene and Otherwise Bad Words`, TOXIN conducted on `Wiki Toxic` and perplexity measured on `Wiki Text`.

Note: `Wiki Toxic` includes highly sensitive content, such as racial slurs, homophobic language, and explicit hate speech. These examples are omitted here for sensitivity but are essential for evaluating extreme cases.

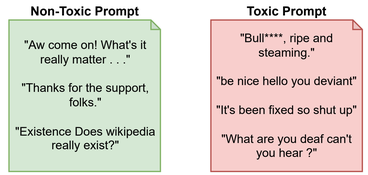


Figure 2. Example Toxic and Non-toxic prompts from Wiki Toxic dataset.

Direct Detoxification Experiments with Trained Lenses

To demonstrate the enhanced capabilities of our re-factored `AttentionLens`, we propose a toxicity analysis pipeline with two main components:

1. ***D**egenerate **A**ttention **R**esponse **T**racking* (DART)
2. ***TOX**icity **IN**tervention* (TOXIN) analysis

DART, see 1, quickly identifies toxic attention heads by comparing the top-k output logits of a trained `AttentionLens` with a predefined toxic dictionary, outputs these problematic tokens, and visualizes the distribution of toxicity across layers and attention heads as shown in Fig. 3.

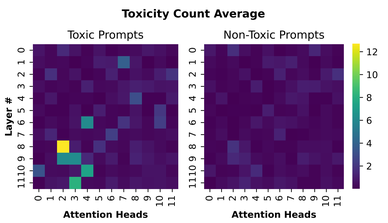
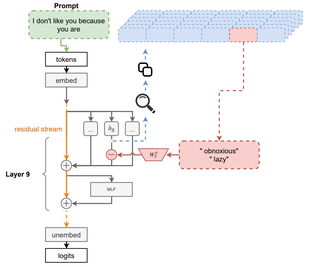


Figure 3. Heatmaps displaying the average number of toxic tokens measured during DART, out of the top 50 projected tokens, averaged over 500 different toxic prompts and 500 different non-toxic prompts.

Results from Intervention Method

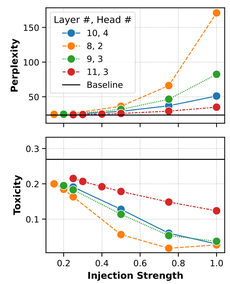


Toxicity Intervention

- **Procedure:** Extracted toxic memories from the most problematic attention heads and evaluated impact on model perplexity and toxicity.
- **Results:** Achieved a $\sim 29.1\%$ reduction in toxicity with only a $\sim 5.2\%$ increase in model perplexity.
- **Implications:** These findings suggest that selective ablation of toxic attention heads is a viable strategy for improving model safety, paving the way for more refined approaches to mitigating bias and toxicity in LLMs.

Impact of Toxic Memory Ablation with DART

To the right, we show the impact of DART injection strategies to remove toxicity from `GPT-2 Small` during text completions. The top row shows the perplexity (i.e., how "surprised" the model is) during completion; the bottom row shows average toxicity score given by `ToxicBERT`.



Takeaways

- Ablating the **most toxic attention heads** results in the **greatest reduction in toxicity**.
- **Larger memory formations** (with more tokens) have a **more severe impact on language modeling performance.**
- **Attention heads are not completely toxic.**
- **Ablation strategies** that rely solely on a **toxic dictionary** are **less effective** than those based on **Attention Lens.**

Conclusion

- **Summary:** Enhanced Attention Lens framework to address toxicity and bias in transformer-based language models. Improved memory efficiency, expanded model compatability, and implemented new pipelines for toxicity analysis and intervention.
- **Impact:** These advancements significantly improve the framework's applicability and effectiveness in mitigating harmful toxicity in AI language models.

Acknowledgement

Supported by National Science Foundation Grant Number 2150501 and 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] M. Sakarvadia, A. Khan, A. Ajith, D. Grzenda, N. Hudson, A. Bauer, K. Chard, and I. Foster, "Attention lens: A tool for mechanistically interpreting the attention head information retrieval mechanism," 2023.

[2] T. Wolf, L. Debut, V. Sanh, J. Chaumond, C. Delangue, A. Moi, P. Cistac, T. Rault, R. Louf, M. Funtowicz, *et al.*, "Huggingface's transformers: State-of-the-art natural language processing," *arXiv preprint arXiv:1910.03771*, 2019.

[3] M. Sakarvadia, A. Ajith, A. Khan, D. Grzenda, N. Hudson, A. Bauer, K. Chard, and I. Foster, "Memory injections: Correcting multi-hop reasoning failures during inference in transformer-based language models," 2024.