Breaking Bad Tokens: Detoxification of LLMs Using Sparse Autoencoders

Agam Goyal, Vedant Rathi, William Yeh,
Yian Wang, Yuen Chen, Hari Sundaram

UNIVERSITY OF ILLINOIS URBANA-CHAMPAIGN

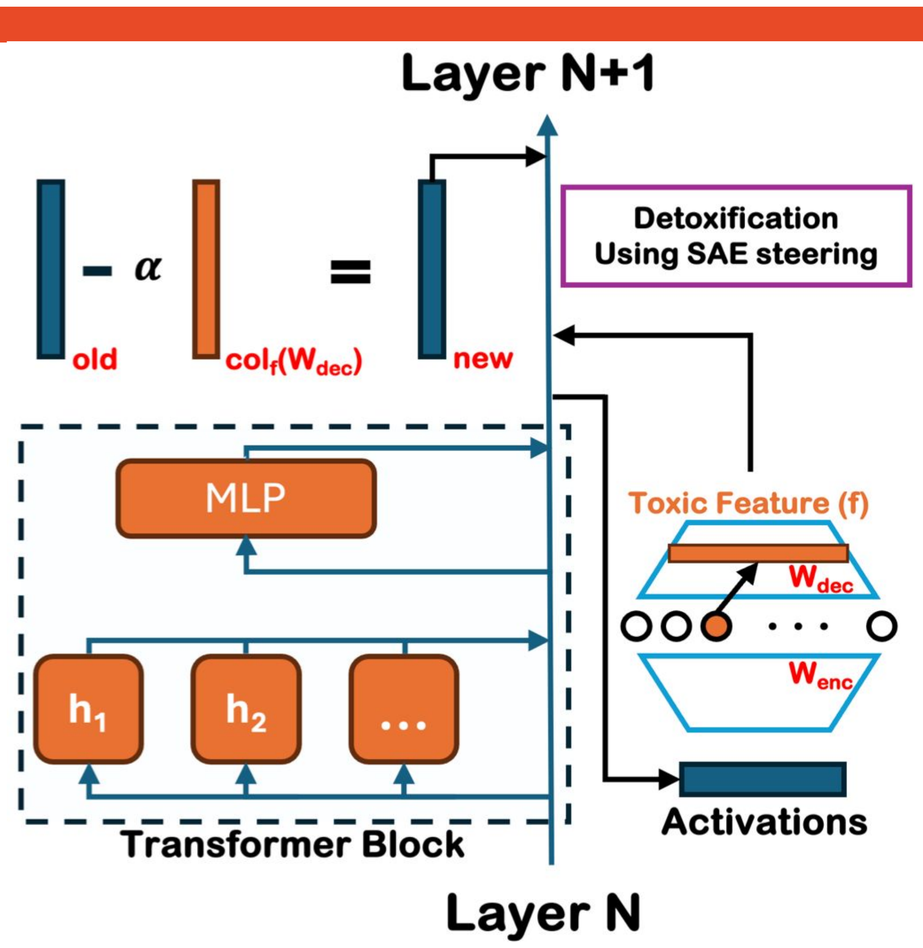
ICML International Conference On Machine Learning

Motivation

Sparse autoencoders have shown promise in interpretability, but what about actionable enhancement of models in **safety critical applications**?

Despite great capabilities, LLMs are still prone to outputting **toxic** content.

Can sparse autoencoders be used to **detoxify language model generations**? What kind of tradeoffs does it present for **model fluency** and **capabilities**?



SAE-based LLM ***Detoxification:*** *We extract the activations from the residual stream of the model after the transformer block of Layer N.*

Using sparse autoencoders (SAEs), we decompose activations to identify toxic dimensions and perform targeted interventions before the steered activations enter Layer N + 1.

Identification of Features

- Sample pairs from **ParaDetox dataset** [1] and store SAE activations for positive and negative sentences
- **Identify top-K features** with highest mean absolute difference in activations
- **Corroborate using Neuronpedia** auto-interpretability explanations

Feature Ablation and Activation Steering

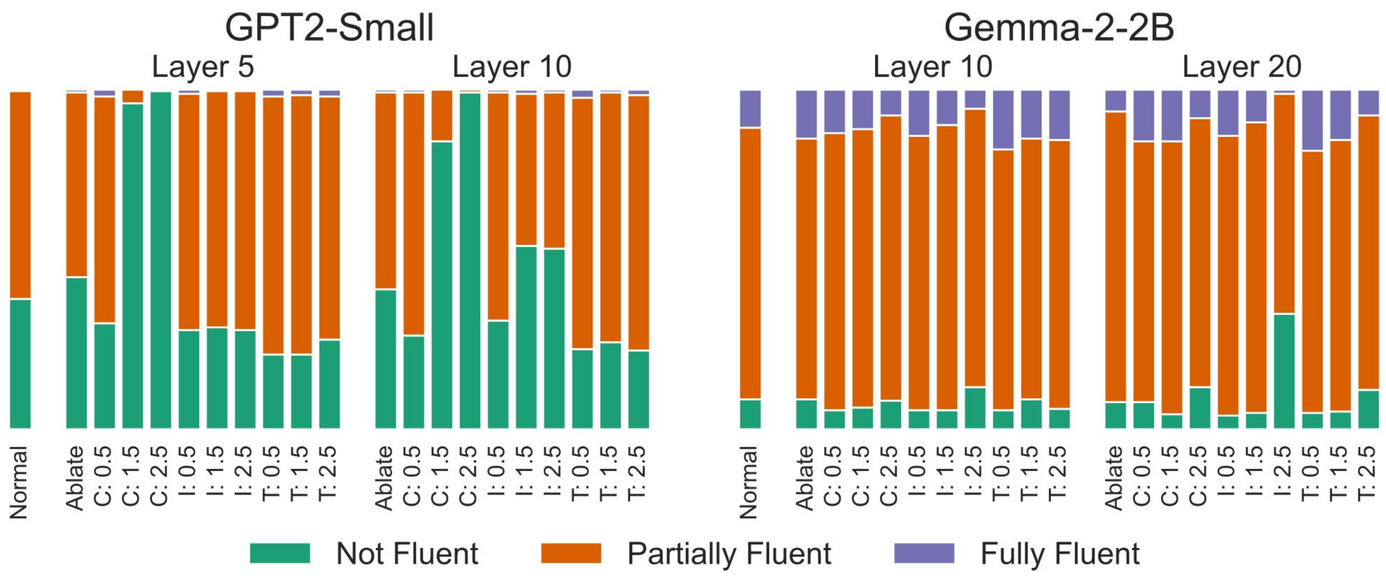
- **Feature ablation:** Set toxic features to zero
- **Constant steering:** Steer always using SAE decoder vector corresponding to toxic features
- **Conditional steering:** Steer at input-level or token-level if feature activation is higher than a given threshold (finer-grained control over interventions)
- **Baselines:** Prompting, DPO, LM-Steer, ProFS, LoRA/SFT

Toxicity Reduction

- **Constant steering leads to best detoxification**, but at the cost of significant fluency degradation in GPT-2 Small. Outperforms detoxification baselines at moderate to high steering strengths.
- **Conditional steering lags behind constant steering** yet outperforms baselines at high steering strengths, while largely preserving fluency of generations.
- **Feature ablation lags behind steering-based approaches** but still leads to reasonable toxicity reduction and also preserves fluency.

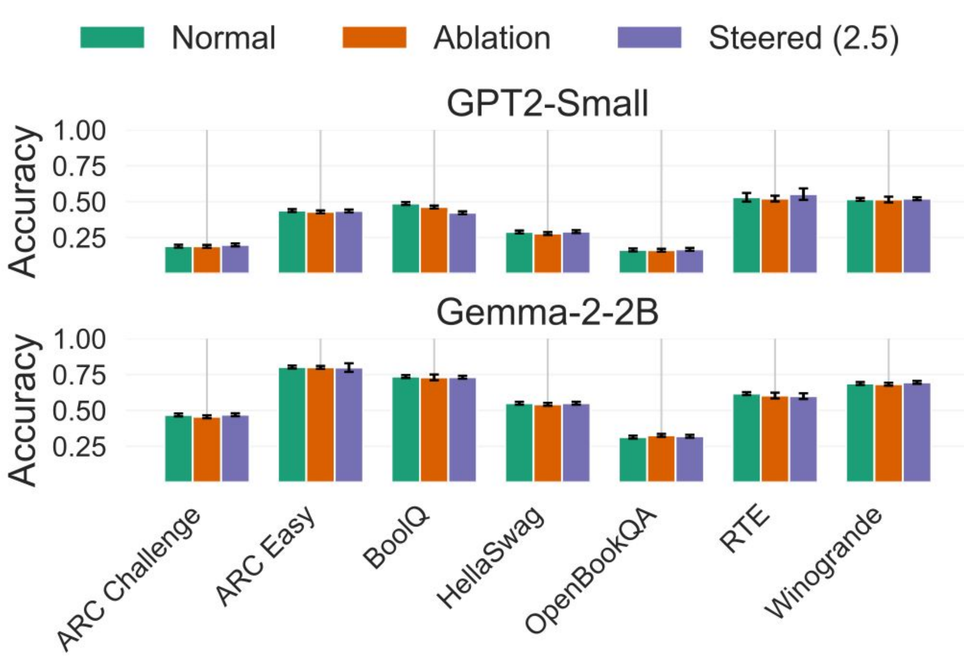
Read full paper for further details and feature-splitting experiments here!

Model Fluency



GPT2 becomes non-fluent, Gemma-2-2B is resistant!

Model Capability



Not Affected!

[1] ParaDetox: Detoxification with Parallel Data (Logacheva et al., ACL 2022)