CausalDetox: Causal Head Selection and Intervention for Language Model Detoxification

Anonymous ACL submission

Abstract

Large language models remain prone to generating toxic content, posing challenges for safe deployment. We propose CAUSALDETOX, a detoxification framework that identifies and intervenes on attention heads causally linked to toxic generation. Using the probability of necessity and sufficiency, a causally grounded criterion, CAUSALDETOX selects heads most responsible for toxicity and modifies only these components at inference time. At inference time, we steer model outputs toward non-toxic continuations by modifying only these selected components. We evaluate CAUSALDETOX on ToxiGen and ImplicitHate, and introduce PARATOX, a benchmark of paraphrased toxic–non-toxic sentence pairs generated with Vicuna-13B for controlled evaluation. CAUSALDETOX achieves up to 38.08% greater toxicity reduction over baseline methods while preserving fluency, and offers a 7× speedup in head selection. Beyond detoxification, the causal principles underlying CAUSALDETOX and PARATOX provide a scalable foundation for safer, controllable language generation across other safety-critical tasks.

1 Introduction

Large language models (LLMs) have significantly advanced natural language generation, achieving state-of-the-art performance across a wide range of tasks. Despite their advancements, LLMs continue to pose serious safety concerns due to their propensity for generating toxic, biased, or otherwise harmful content (Gehman et al., 2020; Welbl et al., 2021). Addressing these issues is crucial for the responsible and ethical deployment of LLMs in real-world applications.

Previous detoxification approaches have primarily involved lexical filtering, adversarial training, reinforcement learning from human feedback (RLHF), and supervised fine-tuning using carefully curated datasets (Bai et al., 2022; Ouyang et al., 2022). While these methods achieve varying degrees of success, each presents notable limitations. Lexical filtering often disrupts semantic coherence and can fail to account for subtle, context-dependent toxicity (Welbl et al., 2021). Methods based on RLHF or supervised fine-tuning require extensive human annotation, which is costly and can lead to the inadvertent suppression of nuanced language or subtle concepts (Xu et al., 2021). More recent model-based approaches, such as direct preference optimization (Lee et al., 2024) or activation patching (Rodriguez et al., 2024), typically involve extensive modification of model parameters, potentially degrading unrelated model capabilities and reducing overall model generalization.

To overcome these challenges, we propose a novel detoxification method inspired by causal representation learning principles (Suter et al., 2019; Locatello et al., 2020; Schölkopf et al., 2021). Our method identifies model components that causally contribute to toxic content generation, enabling precise and targeted interventions. Specifically, we first extract output activations from all attention heads in the language models during the forward pass. Unlike prior work that relies on correlation-based head selection (Rajendran et al., 2024b; Li et al., 2024), we apply a causal criterion—the probability of necessity and sufficiency (PNS)—to quantify each head's influence on toxicity. This yields a compact subset of attention heads that are most responsible for encoding toxic content. At inference time, we apply inference-time intervention Li et al. (2024) to shift these activations away from the toxic directions. We evaluate our method on the ToxiGen (Hartvigsen et al., 2022) dataset and construct PARATOX, a new benchmark of toxic–non-toxic sentence pairs by paraphrasing ToxiGen examples using Vicuna-13B (Chiang et al., 2023). Each pair consists of toxic and non-toxic paraphrases, allowing for fine-grained evaluation. PARATOX will be released for future research.

Empirical results show that the proposed PNS-based head selection method outperforms the baseline accuracy-based selection by up to 38.08% in toxicity reduction over the previous method after intervention (Li et al., 2024), while preserving high fluency in generated text.

In summary, our main contributions are:

- **A causal criterion for head selection:** We propose a novel head selection criterion based on the probability of necessity and sufficiency, which identifies attention heads that causally contribute to toxicity. Prior work primarily relies on correlation-based metrics, such as activation magnitudes or accuracy-driven heuristics (Rajendran et al., 2024b), which may capture spurious associations rather than true causal drivers. These methods often result in redundant or less interpretable head selections. In contrast, our PNS-based approach provides a principled, causally grounded mechanism to select a compact and interpretable set of heads. This leads to more targeted interventions, enabling higher toxicity reduction with minimal impact on language fluency.
- **A benchmark for controlled detoxification:** We construct PARATOX, a new benchmark for controlled detoxification, consisting of toxic–non-toxic sentence pairs generated by paraphrasing ToxiGen (Hartvigsen et al., 2022) and ImplicitHate (ElSherief et al., 2021a) using Vicuna-13B (Chiang et al., 2023). Prior detoxification benchmarks often lack fine-grained control or parallel structure between toxic and non-toxic variants, making it difficult to evaluate subtle changes introduced by intervention methods. This limitation hampers the assessment of both effectiveness and unintended side effects. By providing aligned toxic/non-toxic pairs, PARATOX enables controlled evaluation of detoxification strategies, and facilitates future work on intervention-based mitigation techniques.

2 Related Work

2.1 Detoxification in LLMs

Detoxification techniques for LLMs include lexical, reinforcement learning, and model-editing approaches. Early work applied lexical or rule-based filters to remove toxic tokens, but these risk semantic loss and fail to capture context-dependent toxicity (Gehman et al., 2020; Welbl et al., 2021). Reinforcement learning from human feedback (RLHF) and supervised fine-tuning on curated toxicity datasets improve safety but require extensive human annotation and may inadvertently suppress benign language, particularly minority voices (Bai et al., 2022; Ouyang et al., 2022; Xu et al., 2021). More recent methods perform targeted model edits: direct preference optimization (DPO) aligns generations towards harmlessness via modified loss functions (Lee et al., 2024; Rafailov et al., 2023), activation patching replaces harmful activation patterns with safe ones (Rodriguez et al., 2024; Meng et al., 2022), and subspace steering projects hidden states onto toxicity-averse directions (Han et al., 2024; Ko et al., 2024). Expert/anti-expert frameworks train auxiliary models to rewrite outputs toward safety (Hallinan et al., 2022), while adversarial safety pipelines guard against malicious prompts (Zhao et al., 2024; Dinan et al., 2019; Uppaal et al., 2024). However, many of these rely on correlation-based heuristics, retraining, or fine-tuning, thus is computationally expensive.

2.2 Causal Representation Learning for Alignment

Causal representation learning (CRL) seeks to identify and manipulate latent generative factors under principled causal assumptions (Schölkopf et al., 2021). A foundational desideratum for such representations is articulated by Wang and Jordan (2021), where the authors provided formalized criteria, i.e., the probability of necessity and sufficiency, that guarantee the identification of meaningful latent features. Recent analyses indicate that transformer self-attention encodes structured causal dependencies between tokens (Rohekar et al., 2024; Nichani et al., 2024), motivating causal approaches to detoxification. Causal tracing methods locate toxicity pathways in network circuits but often lack principled intervention mechanisms (Meng et al., 2022). Concept-based CRL relaxes strict interventional requirements by recovering interpretable concepts through conditioning rather than exhaustive interventions (Rajendran et al., 2024a), yet has not been fully leveraged for fine-grained, context-sensitive detoxification in LLMs. In our work, we apply the PNS lower bound criterion from Wang and Jordan (2021) to rigorously enforce causal representation learning and precisely identify toxicity-sensitive activation components for targeted intervention.

2.3 Inference-Time Intervention-Based Methods

Inference-time intervention method modifies model behavior without weight updates. Plug-and-Play Language Models (PPLM) use gradient-based updates to steer hidden states toward desired attributes during generation (Dathathri et al., 2019). GeDi employs small generative discriminators as controllers that adjust token probabilities for targeted attributes (Krause et al., 2020). Direct Preference Optimization (DPO) shows that training LMs with certain loss modifications can be interpreted as reward modeling, influencing inference distributions (Rafailov et al., 2023). Activation patching and causal intervention techniques replace or perturb internal activations in critical layers to effect behavioral changes (Meng et al., 2022; Rodriguez et al., 2024). More recently, Li et al. (2023) introduced Inference-Time Intervention (ITI), which identifies linear "steering directions" in selected activation subspaces (e.g., neuron or head outputs) and adds controlled offsets during generation to improve truthfulness or other attributes. These methods demonstrate that small, targeted adjustments to latent activations can yield large gains in desired behavior while preserving overall fluency, offering a lightweight alternative to full fine-tuning.

3 Preliminaries

In this section, we first introduce notations for transformer-based LLMs and their internal representations. We then review the notions of probability of necessity, sufficiency, and necessity and sufficiency as used in Wang and Jordan (2022), which we extend to the setting of attention head selection. Throughout, we use bold uppercase (e.g., $\boldsymbol{X}$) to denote random vectors and bold lowercase (e.g., $\boldsymbol{x}$) to denote feature vectors.

3.1 Large Language Models

We consider a transformer-based language model $\mathcal{M}$ with ℓ layers, each comprising H self-attention heads. Given an input token sequence $\boldsymbol{x} = [x_1, \ldots, x_{t-1}]$, the model computes contextual representations through a sequence of transformations. Within layer ℓ, the h-th attention head outputs a vector $\boldsymbol{a}^{(\ell,h)} \in \mathbb{R}^d$.

The model then autoregressively generates an output token sequence $\boldsymbol{y} = \mathcal{M}(\boldsymbol{x})$, where each token y_t is sampled based on the conditional distribution $P(y_t \mid \boldsymbol{x}, \boldsymbol{y}_{<t})$.

3.2 Probabilities of Necessity and Sufficiency

We adopt the counterfactual formalism of Wang and Jordan (Wang and Jordan, 2022) to measure how necessary and/or sufficient a feature is for predicting a target label. Let $Z \in \{0, 1\}$ be a binary feature extracted from a high-dimensional input X, and $Y \in \{0, 1\}$ the corresponding label. The counterfactual label had we set Z to a value z is denoted $Y(Z = z)$. The following definitions measure how necessary or sufficient Z is for Y (Wang and Jordan (2022) Definitions 1-3).

Definition 1 (Probability of Necessity (PN))**.**

$$\mathrm{PN}_{z,y} := \mathbb{P}\left(Y(Z \neq z) \neq y \,|\, Z = z, Y = y\right)$$

Definition 2 (Probability of Sufficiency (PS))**.**

$$\mathrm{PS}_{z,y} := \mathbb{P}\left(Y(Z = z) = y \,|\, Z \neq z, Y \neq y\right)$$

Definition 3 (Probability of Necessity and Sufficiency (PNS))**.**

$$\mathrm{PNS}_{z,y} := \mathbb{P}\left(Y(Z \neq z) \neq y, Y(Z = z) = y\right)$$

Intuitively, these scores quantify the causal impact of feature Z on outcome Y:

- *PN* is high when changing $Z = z$ to $Z \neq z$ changes $Y = y$ to $Y \neq y$.
- *PS* is high when changing $Z \neq z$ to $Z = z$ changes $Y \neq y$ to $Y = y$.
- *PNS* captures when both are true—making Z necessary and sufficient predicting $Y = y$.

Our method learns attention head representations that are necessary and sufficient for toxicity. However, since PN, PS, and PNS involve counterfactuals, which are infeasible to compute from observational data, Wang and Jordan (2022) then proposed a lower bound on the logarithm of PNS, which we use as a representation learning objective.

3.3 Inference-Time Intervention

Inference-time intervention (ITI) (Li et al., 2024) is an LLM alignment technique that modifies the model activations during generation to elicit or suppress target concepts in the output. In our case, we aim to suppress the concept of toxicity.

Let $\boldsymbol{a}^{(\ell,h)}(\boldsymbol{x})$ denote the activation of head h in layer ℓ for the input $\boldsymbol{x}$. In Li et al. (2024), the authors train linear classifiers over the activations of all attention heads to predict the presence of a target concept in the input.

For each selected head, an intervention vector $\boldsymbol{\delta}^{(\ell,h)}$ is computed to shift the activation away from the direction associated with toxicity. Formally, the intervention is defined as:

$$\boldsymbol{\delta}^{(\ell,h)} = \alpha \cdot \sigma^{(\ell,h)} \cdot \boldsymbol{v}^{(\ell,h)}, \tag{1}$$

where α is a scaling hyperparameter, $\sigma^{(\ell,h)}$ is the standard deviation of the head's activations along the intervention direction, and $\boldsymbol{v}^{(\ell,h)}$ is the mean difference of the activations between the non-toxic and toxic pairs:

$$\boldsymbol{v}^{(\ell,h)} = \frac{1}{n}\sum^{n}(\boldsymbol{a}^{(\ell,h)}(\boldsymbol{x}^{-}) - \boldsymbol{a}^{(\ell,h)}(\boldsymbol{x}^{+})) \tag{2}$$

where $\boldsymbol{x}^{-}$ and $\boldsymbol{x}^{+}$ are the generated paraphrases based on inputs $\boldsymbol{x}$, and we will introduce the generation later in Section 5.

During the generation, we apply the intervention as:

$$\boldsymbol{a}^{(\ell,h)}(\boldsymbol{x}) \leftarrow \boldsymbol{a}^{(\ell,h)}(\boldsymbol{x}) + \boldsymbol{\delta}^{(\ell,h)}. \tag{3}$$

Note that in the original ITI approach, intervention targets are selected based on classification accuracy, which is inherently correlation-based. This may result in redundant head selection and non-minimal interventions. For example, if two heads are highly collinear and one causally influences the other, both may be selected despite only one being causally relevant. In contrast, our method selects attention heads based on their causal contribution, quantified via their estimated necessity and sufficiency for toxicity. This enables more focused and effective modifications.

4 Method

We propose CAUSALDETOX, a two-stage method for detoxifying LLMs by identifying and manipulating attention heads most causally responsible for toxic generation. Given a dataset $\mathcal{D} := \{(\boldsymbol{x}_i, y_i)\}_{i=1}^{n}$, where each $\boldsymbol{x}_i$ is a sentence, i.e., a sequence of tokens, and y_i is a binary label indicating whether the $\boldsymbol{x}_i$ is toxic or not, $y = 1$ for toxic, $y = 0$ for non-toxic. we make a forward pass on

Given input $\boldsymbol{x}_i$, the model generated a sequence of continuation $\widehat{\boldsymbol{x}}_{\boldsymbol{i}} := \mathcal{M}(\boldsymbol{x}_i)$. The goal is for the model to generate sequences that are less toxic than the input tokens.

In particular, we assume access to a toxicity scoring function $f : \mathcal{X}^* \to [0, 1]$ that assigns a scalar toxicity score to tokens of variable length. The objective of detoxification is to prevent the generation that increase toxicity:

$$f(\widehat{\boldsymbol{x}}) \leq f(\boldsymbol{x}). \tag{4}$$

To achieve this, CAUSALDETOX proceeds in two stages:

1. **Causal Head Identification:** We estimate the causal contribution of each attention head to toxicity using the probability of necessity and sufficiency and select a targeted subset $\mathcal{H}_{\text{toxic}}$ for intervention.

2. **Inference-Time Intervention:** At generation time, we manipulate the activations of heads in $\mathcal{H}_{\text{toxic}}$ to steer the model away from generating toxic content.

4.1 Identify Causally-Relevant Attention Heads

To identify the subset $\mathcal{H}_{\text{toxic}}$ for intervention, we quantify the causal influence of each attention head on sentence toxicity by estimating a lower bound on its probability of necessity and sufficiency, following Wang and Jordan (2022). The motivation is that by concentrating toxicity-related influence in this targeted set, we aim to modify toxic behavior without disrupting unrelated, benign model behaviors. However, computing exact PNS values is generally intractable from observational data alone. To address this, we adapt a tractable lower bound on $\log(\text{PNS}_{\boldsymbol{Z},Y})$, where $\boldsymbol{Z}$ denotes the attention head output and Y the toxicity label, which can be estimated from observational data under mild assumptions.

For head (ℓ, h), let $\{\boldsymbol{z}_i^{(\ell,h)}\}_{i=1}^{n}$ denote the output activations on of $\{\boldsymbol{x}_i\}_{i=1}^{n}$, we have an lower bound on $\log(\text{PNS}_{\boldsymbol{Z}^{(\ell,h)},Y})$ in eq. (5). For the ease of notation, we omit (ℓ, h) for the rest of this section and use $\boldsymbol{z}$ to denote the output of an attention head.

$$\begin{aligned}
&\log \text{PNS}(\mathbf{Z}, Y) \\
=&\frac{1}{2\sigma^2}\sum_{i=1}^{n}\left[\left(\sum_{j=1}^{d}\beta_j(z_i^j - \mathbb{E}[z_i^j])\right)^2\right. \\
&\left.+2\left(\sum_{j=1}^{d}\beta_j(z_i^j - \mathbb{E}[z_i^j])\right)\boldsymbol{\gamma}^{\top}(\mathbf{c}_i - \mathbb{E}[\mathbf{c}_i])\right]
\end{aligned} \tag{5}$$

Here the second super script j denotes the j^{th} dimension of $\boldsymbol{z}_i$. β_0 and $\boldsymbol{\beta}$ are estimated by a linear model:

$$P(Y \mid \boldsymbol{Z}, \boldsymbol{C}) = \mathcal{N}\left(\left(\beta_0 + \boldsymbol{\beta}^\top \boldsymbol{Z} + \boldsymbol{\gamma}^\top \boldsymbol{C}\right), \sigma^2\right). \quad (6)$$

The variable $\boldsymbol{c}_i$ captures the hidden common cause that gives rise to correlations among the different dimensions of $\boldsymbol{z}_i^{(\ell,h)}$. Since $\boldsymbol{c}$ is unobserved, one can model it with a probabilistic factor model. In our implementation, we train a variational autoencoder (VAE) (Kingma et al., 2013)) to reconstruct $\{\boldsymbol{z}_i\}_{i=1}^n$ and treat the inferred latent mean vector as $\boldsymbol{c}_i$. As our primary focus is on the application of causal criterion to toxicity unlearning, we do not reproduce the derivations here and instead refer the reader to Wang and Jordan (2022) for the details.

After computing the eq. (5) for all attention heads (ℓ, h), we select the top-K heads with the highest scores for the set $\mathcal{H}_{\text{toxic}}$ for intervention.

4.2 Apply Inference-Time Intervention

During generation, we apply inference-time intervention (ITI) (Li et al., 2024) as described in section 3.3. The idea is that, by intervening on features that are both necessary and sufficient for toxicity, we achieve more effective toxicity mitigation with fewer unintended effects. In contrast to applying ITI on attention heads selected purely based on their correlation with toxicity (e.g., via classification accuracy), our approach targets heads with demonstrable causal influence. We point out that When compute the steering vector assumes that the subset of attention heads identified as causally responsible for toxicity—$\mathcal{H}_{\text{toxic}}$—is fixed and does not change across inputs, following the original ITI paper (Li et al., 2024). In future work, the head selection and steering vectors computation process could be extended to operate dynamically at inference time.

5 PARATOX Benchmark

To pinpoint the concept of toxicity in sentences and to steer the model, as mentioned in Section 3.3, we ideally require pairs of sentences that are semantically identical except for the presence or absence of toxicity. In the terminology of Pearl's causality (Pearl et al., 2021; Pearl, 2009; Peters et al., 2015),, a toxic sentence $\boldsymbol{x}^+$ can be viewed as the counterfactual of a non-toxic sentence $\boldsymbol{x}^-$, where the latent variable "toxicity" has been set to true while all other factors remain fixed. Formally, we express this as:

$$\boldsymbol{x}^+ := \boldsymbol{x}^-_{\text{toxicity = True}},$$

where the subscript denotes the counterfactual, consistent with the counterfactual semantics in Wang and Jordan (2022).

However, existing toxicity datasets such as Jigsaw (cjadams et al., 2017), ToxiGen (Hartvigsen et al., 2022), and ImplicitHate (ElSherief et al., 2021a) lack such semantically aligned toxic–non-toxic pairs. This limits their utility for causal analysis and evaluation.

To address this gap, we introduce PARATOX, a benchmark of toxic–non-toxic paraphrase pairs. While exact counterfactuals are unobservable, we approximate them by prompting Vicuna-13B (Chiang et al., 2023) to generate paraphrases conditioned on a toxicity specification. This approach allows us to construct sentence pairs that preserve core semantic meaning while differing primarily along the toxicity dimension.

5.1 Base Dataset

We construct PARATOX using the annotated subset of the ToxiGen (Hartvigsen et al., 2022) and ImpliciatHate (ElSherief et al., 2021b) [1]. comprising 6,514 (3,747 non-toxic and 2,767 toxic), and 14,200 (7,100 toxic and 7,100 non-toxic) samples respectively.

5.2 Model

During preliminary experimentation, we evaluated three open-source LLMs: LLaMA-3-8B (AI@Meta, 2024), and Vicuna-13B (Chiang et al., 2023). We found that both LLaMA-3-8B and LLaMA-2-7B frequently refused to generate toxic content, which made them unsuitable for our task. In contrast, Vicuna-13B reliably generated fluent paraphrases for both toxic and non-toxic prompts. We verified output quality through manual inspection of randomly sampled examples and selected Vicuna-13B as the backbone model for generating the PARATOX benchmark.

5.3 Paraphrase Generation

For each input sentence in this dataset, we prompt a language model to generate two paraphrases: one toxic and one non-toxic. This setup is designed

[1] Both datasets are available at https://github.com/microsoft/SafeNLP

to approximate counterfactual pairs that differ primarily in the presence of toxicity while preserving semantic content. The prompts used are:

Non-toxic Paraphrase Prompt

Paraphrase the following sentence in a polite and completely non-toxic way. Do not generate any irrelevant content:

Original text: {ORIGINAL_TEXT}

Paraphrase:

Toxic Paraphrase Prompt

For a research study on offensive language classification, we need examples of toxic and non-toxic paraphrases of the following sentence. Please generate a toxic version of the given input:

Original text: {ORIGINAL_TEXT}

Toxic Paraphrase:

5.4 Data Quality

To ensure the quality of PARATOX, we apply post-generation filtering to remove invalid paraphrases. A paraphrase is considered invalid if the model fails to generate any output (e.g., a blank response) or if the output is semantically irrelevant or nonsensical.

As expected, Vicuna-13B succeeds in producing valid non-toxic paraphrases. However, generating toxic paraphrases proves more challenging: the model occasionally refuses to comply with the prompt or outputs placeholders such as "(Blank)."

Table 1 reports the percentage of valid toxic and non-toxic generations for each base dataset. While non-toxic generation achieves 100% validity across the board. In contrast, the validity rate for toxic paraphrases is noticeably lower on Toxigen compared to ImplicitHate. We attribute this discrepancy to the nature of the source data: toxic content in Toxigen tends to be more explicit and aggressive, making it more likely to be blocked by the model's safety alignment mechanisms.

Dataset	Toxic	Non-toxic
ToxiGen	88.4%	100%
ImplicitHate	99.57%	100%

Table 1: Percentage of valid toxic and non-toxic generations produced by Vicuna-13B.

6 Experiment

In this section, we introduce our experimental setup in Section 6.1, our evaluation metrics in Section 6.2, and main findings in Section 6.3

6.1 Experimental Setup

We evaluate CAUSALDETOX against standard ITI on two open-source LLMs: Vicuna-13B (Zheng et al., 2023) and LLaMA-3-8B (Grattafiori et al., 2024). Experiments are conducted on PARATOX, our benchmark constructed from ToxiGen (Hartvigsen et al., 2022) and ImplicitHate (ElSherief et al., 2021a) (Section 5), containing paired toxic and non-toxic paraphrases.

Following the ITI implementations in Li et al. (2024) and Rajendran et al. (2024a), we first extract activations from all $L \times H$ attention heads across the dataset. For standard ITI, a linear classifier is trained for each head to predict the presence of toxicity in the input. In contrast, CAUSALDETOX computes a closed-form Equation (5) for each head without requiring any training.

We then select the top-K heads based on classification accuracy (for ITI) or logPNS score (for CAUSALDETOX), denoted by $\mathcal{H}^{\text{Acc.}}_{\text{toxic}}$ and $\mathcal{H}^{\text{PNS}}_{\text{toxic}}$, respectively. These sets are the intervention targets.

Finally, we prompt the LLM with each input sentence and apply the corresponding steering vectors—computed as described in Section 3.3—to the selected heads: $\mathcal{H}^{\text{Acc.}}_{\text{toxic}}$ for standard ITI, and $\mathcal{H}^{\text{PNS}}_{\text{toxic}}$ for CAUSALDETOX, during generation.

6.2 Evaluation

For each generated text, we measure its toxicity and fluency and compare these metrics against those of the corresponding input sentence. Our evaluation relies on the following metrics:

- **Toxicity Reduction** We use Detoxify (Hanu and Unitary team, 2020), a publicly available and widely used toxicity detection model[2], which outputs a toxicity score between 0 and 1 indicating the likelihood of toxic content. We measure the average reduction in Detoxify scores between the input and generated text as an indicator of intervention effectiveness.

- **Preservation of Fluency:** We assess fluency using perplexity (Jelinek et al., 1977), computed from the same language model used for generation (LLaMA-3-8B or Vicuna-13B),

[2] https://github.com/unitaryai/detoxify

Dataset	Model	#Heads	Method	Toxicity Red. ↑	Perplexity ↓
ToxiGen	Vicuna-13B	–	Baseline	0.2513 ± 0.31	9.45 ± 34.62
		18	ITI	0.2263 ± 0.30	9.69 ± 8.13
			CAUSALDETOX	0.2341 ± 0.31	**8.91 ± 8.52**
		36	ITI	0.2187 ± 0.31	10.12 ± 21.88
			CAUSALDETOX	**0.3020 ± 0.33**	10.88 ± 8.69
	LLaMA-3-8B	–	Baseline	0.1729± 0.31	8.88 ± 16.31
		18	ITI	0.2007 ± 0.32	8.76 ± 59.19
			CAUSALDETOX	0.1708 ± 0.31	9.36 ± 28.96
		36	ITI	0.2265 ± 0.32	7.65 ± 22.35
			CAUSALDETOX	**0.2382 ± 0.32**	**7.56 ± 17.32**
ImplicitHate	Vicuna-13B	–	Baseline	0.3463 ± 0.30	13.26 ± 20.02
		18	ITI	0.3141 ± 0.30	14.51 ± 27.00
			CAUSALDETOX	**0.3487 ± 0.30**	13.77 ± 20.93
		36	ITI	0.3156 ± 0.30	13.60 ± 23.42
			CAUSALDETOX	0.3244± 0.30	**13.11 ± 21.74**
	LLaMA-3-8B	–	Baseline	0.2575 ± 0.30	17.32 ± 28.99
		18	ITI	0.2740 ± 0.30	13.86 ± 16.76
			CAUSALDETOX	0.2799 ± 0.30	16.86 ± 32.34
		36	ITI	**0.2940 ± 0.30**	8.22 ± 14.59
			CAUSALDETOX	0.2919 ± 0.30	**8.17 ± 15.35**

Table 2: Evaluation of toxicity reduction (%) and perplexity (mean ± std) for Baseline (no intervention), ITI, and CAUSALDETOX across two datasets (ToxiGen and ImplicitHate), two models (Vicuna-13B and LLaMA-3-8B), and two head selection sizes (18 and 36). Results are grouped by dataset and model. CAUSALDETOX (PNS-based) and ITI (correlation-based) are compared under matched conditions. Best values in each block are bolded. Lower perplexity and higher toxicity reduction indicate better performance.

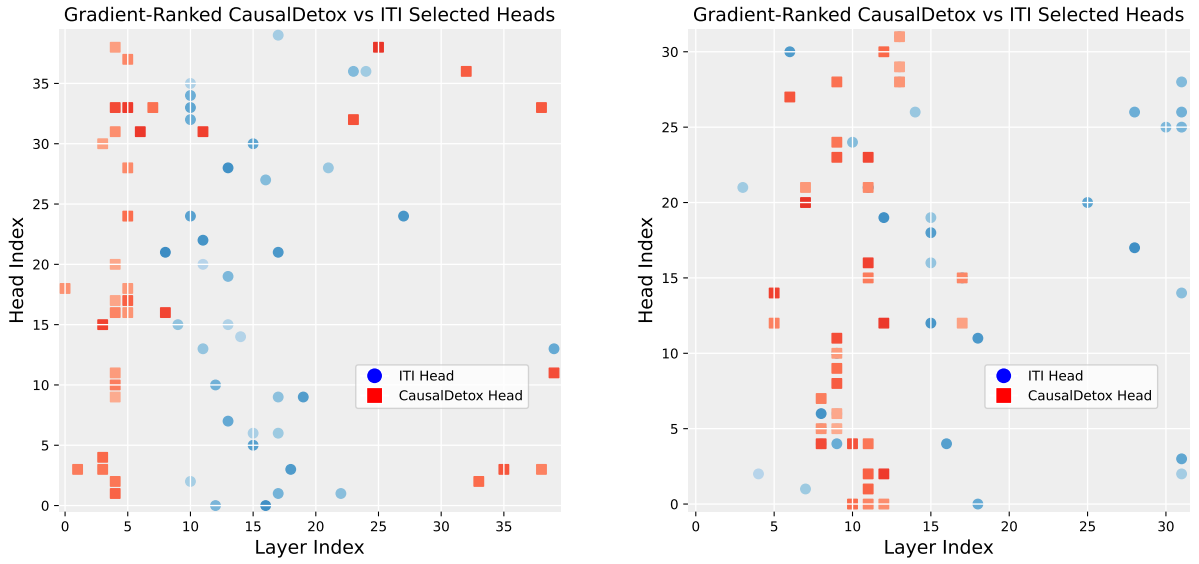
where lower scores indicate higher fluency. We compare perplexity before and after intervention to ensure that the intervention does not impair linguistic quality.

6.3 Results

Superior Toxicity Reduction Table 2 presents the performance of CAUSALDETOX, standard ITI, and a no-intervention baseline (i.e., the original model without any steering) on Vicuna-13B and LLaMA-3-8B, evaluated across the ToxiGen and ImplicitHate datasets. We report average toxicity reduction (higher is better) and perplexity (lower is better) for each configuration. CAUSALDETOX achieves the highest toxicity reduction in 3 out of the 4 model–dataset combinations, demonstrating its effectiveness over correlation-based approaches. Additionally, it maintains perplexity scores comparable to the baseline, indicating that the intervention preserves the fluency of the generated text.

Efficiency of CAUSALDETOX In addition to effectiveness, we also compare the efficiency of the head selection procedures. For a model with 40 layers and 40 attention heads per layer, the traditional logistic regression approach requires around 42 seconds, while our PNS-based scoring method completes head selection in 6 seconds on a single GPU, achieving a 7× speedup. This overhead of the accuracy-based method arises from the need to train $L \times H$ separate classifiers, one per attention head. This highlights the computational advantage of our causal scoring framework. As language models grow larger, the relative cost of traditional head selection methods increases rapidly, while our approach remains lightweight and scalable. These efficiency gains make CAUSALDETOX not only principled and interpretable, but also practical for real-world deployment in large-scale model detoxification pipelines.

Optimal Number of Intervention Heads We observe that increasing the number of intervention heads from 18 to 36 improves toxicity reduction for CAUSALDETOX, but yields limited gains for ITI. A potential explanation is that the additional heads selected by CAUSALDETOX remain causally relevant, providing complementary, non-redundant information about toxicity. In contrast, the extra heads chosen by ITI are likely correlated with those



(a) Vicuna-13B

(b) Llama-3-8B

Figure 1: Visualization of the top 36 attention heads selected by ITI and CAUSALDETOX on ToxiGen for (a) Vicuna-13B and (b) LLaMA-3-8B. Blue circles denote ITI-selected heads; red squares denote CAUSALDETOX-selected heads. Color intensity reflects head rank, with darker shades indicating higher importance. CAUSALDETOX exhibits strong layer-wise concentration—around layer 5 in Vicuna-13B and layer 10 in LLaMA-3-8B, highlighting the method's ability to isolate causally relevant substructures. In contrast, ITI-selected heads are more uniformly distributed, suggesting that the correlation-based criterion does not find a localized representation to specific layers.

already selected, offering little new information and thus limited additional impact. This highlights the advantage of causality-guided selection in capturing diverse and informative signals.

In early experiments, we found that intervening on 72 heads led to severe degradation in language quality, often resulting in incoherent or nonsensical output, while yielding only marginal gains in toxicity reduction. This suggests that a modest number of heads is sufficient to capture the key causal mechanisms behind toxic generation. Intervening on 36 heads, in particular, strikes a strong balance: it effectively mitigates toxicity while preserving the model's linguistic fluency and coherence.

Concentration of PNS-selected Heads In Figures 1a and 1b, we visualize the head selection for different models and find structural patterns. Specifically, for the Vicuna-13B model, heads selected via eq. (5) criteria predominantly cluster around layer 5. In contrast, the LLaMA-3-8B model exhibits a concentration of CAUSALDETOX-selected heads around layer 10. This layer-specific clustering contrasts with the ITI-selected heads, which display a more uniform and dispersed distribution across various layers and heads.

7 Conclusions

We have introduced CAUSALDETOX, a causally grounded detoxification framework that identifies and intervenes on attention heads responsible for toxic generation in LLMs. Using the probability of necessity and sufficiency, we select only the most causally impactful heads to enable efficient and precise inference-time intervention. Experiments on Vicuna-13B and LLaMA-3-8B across two real-world toxicity datasets show that CAUSALDETOX reduces toxicity while maintaining fluency. In addition to its effectiveness, CAUSALDETOX is highly efficient, achieving a 7× speedup over the traditional correlation-based head selection method. These results highlight CAUSALDETOX as a practical, interpretable, and scalable approach to safer language generation.

We believe this work opens a promising direction for inference-time intervention by integrating causal criteria into both head selection and manipulation. While this paper focuses on detoxification, the underlying framework,CAUSALDETOX, and the data construction principles behind PARATOX are broadly applicable to other generative behavior modifications, such as reducing social biases and preventing harmful outputs.

8 Limitations

Our work relies on several assumptions that limit its generalizability and robustness.

Limitations of Fixed, Mean-Based Intervention Directions. In our current approach, intervention vectors $\boldsymbol{\delta}^{(\ell,h)}$ are computed once per attention head in $\mathcal{H}_{\text{toxic}}$ and remain fixed throughout inference. These vectors are derived from the mean activation differences between toxic and non-toxic examples, as defined in eq. (2). While effective in practice, this fixed and mean-based direction may fail to capture input-specific nuances and can be sensitive to high variance or skewed distributions in the underlying activations. In such cases, the mean may not serve as a reliable or representative summary statistic, potentially leading to suboptimal or inaccurate interventions. A promising direction for future work is to treat the activation differences $\boldsymbol{a}(\boldsymbol{x}^-) - \boldsymbol{a}(\boldsymbol{x}^+)$ as samples from a distribution, e.g., a multivariate Gaussian with a learned or estimated covariance matrix, enabling probabilistic steering strategies that better reflect the uncertainty and diversity in toxicity-associated features.

Assumptions on Linearity and Fixed Head Selection. Our method is grounded in the assumption that toxicity can be causally localized to a fixed, small subset of attention heads via a linear representation, as quantified by PNS scores. This simplifies analysis and enables efficient intervention, but may overlook important nuances of toxicity encoding. In practice, toxic behavior may emerge through nonlinear, distributed, or context-dependent interactions across multiple heads and layers. Additionally, we follow the original ITI framework in assuming that the selected subset of relevant heads, $\mathcal{H}_{\text{toxic}}$, is static across all inputs, determined once during training and reused during inference. While this global selection has shown strong empirical performance, it may not fully reflect the dynamic nature of toxicity expression. Future work could explore adaptive, input-dependent head selection and nonlinear causal modeling to better capture the complexity of toxicity in language models.

Limited model and language coverage Our experiments are carried out on two models, Vituna-13B and LLaMA3-8B, and primarily on English-language datasets (ToxiGen, ImplicitHate, and our constructed PARATOX). The performance and generalizability of our approach in other languages, cultural settings, and LLM architectures remain untested. Given the sociolinguistic variability in how toxicity manifests, further evaluation on multilingual and cross-cultural benchmarks is essential to assess robustness and fairness across deployment scenarios.

Evaluation with automatic metrics. Our evaluation relies primarily on automatic metrics such as toxicity scores and perplexity. While effective for large-scale assessment, these metrics may fail to capture subtle semantic distortions, shifts in intent, or social biases introduced by the intervention. They also do not account for human judgment or contextual appropriateness. To better assess real-world detoxification quality and societal impacts, future studies should incorporate more structured human evaluations.

Ethical Considerations

Our detoxification framework carries risks of misuse or unintended consequences. There is potential for misuse to suppress legitimate content under the pretext of reducing toxicity, thereby hindering the freedom of expression or censoring marginalized voices. Additionally, while explicit toxicity might be effectively mitigated, implicit biases and subtler harmful outputs might persist, which our method currently may not adequately detect or rectify.

Furthermore, datasets like ToxiGen and ImplicitHate, despite careful curation, inherently carry biases that could reinforce cultural stereotypes or propagate normative judgments on what constitutes toxicity. This issue may disproportionately impact certain communities and cultural contexts, reinforcing or marginalizing particular viewpoints or identities.

Finally, while our proposed technique is intended for harm reduction, it could potentially be exploited to subtly manipulate or distort LLM outputs maliciously. It is essential to monitor deployments rigorously, establish transparency and accountability protocols, and explore proactive measures to prevent misuse.

References

AI@Meta. 2024. Llama 3 model card.

Yuntao Bai, Andy Jones, Kamal Ndousse, Amanda Askell, Anna Chen, Nova DasSarma, Dawn Drain, Stanislav Fort, Deep Ganguli, Tom Henighan, and 1 others. 2022. Training a helpful and harmless assistant with reinforcement learning from human feedback. *arXiv preprint arXiv:2204.05862*.

Wei-Lin Chiang, Zhuohan Li, Zi Lin, Ying Sheng, Zhanghao Wu, Hao Zhang, Lianmin Zheng, Siyuan Zhuang, Yonghao Zhuang, Joseph E. Gonzalez, Ion Stoica, and Eric P. Xing. 2023. Vicuna: An open-source chatbot impressing gpt-4 with 90%* chatgpt quality.

cjadams, Jeffrey Sorensen, Julia Elliott, Lucas Dixon, Mark McDonald, nithum, and Will Cukierski. 2017. Toxic comment classification challenge. https://kaggle.com/competitions/jigsaw-toxic-comment-classification-challenge. Kaggle.

Sumanth Dathathri, Andrea Madotto, Janice Lan, Jane Hung, Eric Frank, Piero Molino, Jason Yosinski, and Rosanne Liu. 2019. Plug and play language models: A simple approach to controlled text generation. *arXiv preprint arXiv:1912.02164*.

Emily Dinan, Samuel Humeau, Bharath Chintagunta, and Jason Weston. 2019. Build it break it fix it for dialogue safety: Robustness from adversarial human attack. *arXiv preprint arXiv:1908.06083*.

Mai ElSherief, Caleb Ziems, David Muchlinski, Vaishnavi Anupindi, Jordyn Seybolt, Munmun De Choudhury, and Diyi Yang. 2021a. Latent hatred: A benchmark for understanding implicit hate speech. *arXiv preprint arXiv:2109.05322*.

Mai ElSherief, Caleb Ziems, David Muchlinski, Vaishnavi Anupindi, Jordyn Seybolt, Munmun De Choudhury, and Diyi Yang. 2021b. Latent hatred: A benchmark for understanding implicit hate speech. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 345–363, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Samuel Gehman, Suchin Gururangan, Maarten Sap, Yejin Choi, and Noah A Smith. 2020. Realtoxicityprompts: Evaluating neural toxic degeneration in language models. *arXiv preprint arXiv:2009.11462*.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, and 1 others. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

Skyler Hallinan, Alisa Liu, Yejin Choi, and Maarten Sap. 2022. Detoxifying text with marco: Controllable revision with experts and anti-experts. *arXiv preprint arXiv:2212.10543*.

Chi Han, Jialiang Xu, Manling Li, Yi Fung, Chenkai Sun, Nan Jiang, Tarek Abdelzaher, and Heng Ji. 2024. Word embeddings are steers for language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 16410–16430.

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

Thomas Hartvigsen, Saadia Gabriel, Hamid Palangi, Maarten Sap, Dipankar Ray, and Ece Kamar. 2022. Toxigen: A large-scale machine-generated dataset for adversarial and implicit hate speech detection. *arXiv preprint arXiv:2203.09509*.

Fred Jelinek, Robert L Mercer, Lalit R Bahl, and James K Baker. 1977. Perplexity—a measure of the difficulty of speech recognition tasks. *The Journal of the Acoustical Society of America*, 62(S1):S63–S63.

Diederik P Kingma, Max Welling, and 1 others. 2013. Auto-encoding variational bayes.

Ching-Yun Ko, Pin-Yu Chen, Payel Das, Youssef Mroueh, Soham Dan, Georgios Kollias, Subhajit Chaudhury, Tejaswini Pedapati, and Luca Daniel. 2024. Large language models can be strong self-detoxifiers. *arXiv preprint arXiv:2410.03818*.

Ben Krause, Akhilesh Deepak Gotmare, Bryan McCann, Nitish Shirish Keskar, Shafiq Joty, Richard Socher, and Nazneen Fatema Rajani. 2020. Gedi: Generative discriminator guided sequence generation. *arXiv preprint arXiv:2009.06367*.

Andrew Lee, Xiaoyan Bai, Itamar Pres, Martin Wattenberg, Jonathan K Kummerfeld, and Rada Mihalcea. 2024. A mechanistic understanding of alignment algorithms: A case study on dpo and toxicity. *arXiv preprint arXiv:2401.01967*.

Kenneth Li, Oam Patel, Fernanda Viégas, Hanspeter Pfister, and Martin Wattenberg. 2023. Inference-time intervention: Eliciting truthful answers from a language model. *Advances in Neural Information Processing Systems*, 36:41451–41530.

Kenneth Li, Oam Patel, Fernanda Viégas, Hanspeter Pfister, and Martin Wattenberg. 2024. Inference-Time Intervention: Eliciting Truthful Answers from a Language Model. *arXiv preprint*. ArXiv:2306.03341 [cs].

Francesco Locatello, Ben Poole, Gunnar Rätsch, Bernhard Schölkopf, Olivier Bachem, and Michael Tschannen. 2020. Weakly-Supervised Disentanglement Without Compromises. *arXiv preprint*. ArXiv:2002.02886 [cs, stat].

Kevin Meng, David Bau, Alex Andonian, and Yonatan Belinkov. 2022. Locating and editing factual associations in gpt. *Advances in Neural Information Processing Systems*, 35:17359–17372.

Eshaan Nichani, Alex Damian, and Jason D Lee. 2024. How transformers learn causal structure with gradient descent. *arXiv preprint arXiv:2402.14735*.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, and 1 others. 2022. Training language models to follow instructions with human feedback. *Advances in neural information processing systems*, 35:27730–27744.

Judea Pearl. 2009. *Causality: Models, Reasoning and Inference*, 2nd edition. Cambridge University Press, USA.

Judea Pearl, Madelyn Glymour, and Nicholas P. Jewell. 2021. *Causal inference in statistics: a primer*, reprinted with revisions edition. Wiley, Chichester.

Jonas Peters, Peter Bühlmann, and Nicolai Meinshausen. 2015. Causal inference using invariant prediction: identification and confidence intervals. *arXiv preprint*. ArXiv:1501.01332 [stat].

Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D Manning, Stefano Ermon, and Chelsea Finn. 2023. Direct preference optimization: Your language model is secretly a reward model. *Advances in Neural Information Processing Systems*, 36:53728–53741.

Goutham Rajendran, Simon Buchholz, Bryon Aragam, Bernhard Schölkopf, and Pradeep Ravikumar. 2024a. From causal to concept-based representation learning. *Advances in Neural Information Processing Systems*, 37:101250–101296.

Goutham Rajendran, Simon Buchholz, Bryon Aragam, Bernhard Schölkopf, and Pradeep Kumar Ravikumar. 2024b. From Causal to Concept-Based Representation Learning.

Pau Rodriguez, Arno Blaas, Michal Klein, Luca Zappella, Nicholas Apostoloff, Marco Cuturi, and Xavier Suau. 2024. Controlling language and diffusion models by transporting activations. *arXiv preprint arXiv:2410.23054*.

Raanan Y Rohekar, Yaniv Gurwicz, and Shami Nisimov. 2024. Causal interpretation of self-attention in pre-trained transformers. *Advances in Neural Information Processing Systems*, 36.

Bernhard Schölkopf, Francesco Locatello, Stefan Bauer, Nan Rosemary Ke, Nal Kalchbrenner, Anirudh Goyal, and Yoshua Bengio. 2021. Toward causal representation learning. *Proceedings of the IEEE*, 109(5):612–634.

Bernhard Schölkopf, Francesco Locatello, Stefan Bauer, Nan Rosemary Ke, Nal Kalchbrenner, Anirudh Goyal, and Yoshua Bengio. 2021. Toward Causal Representation Learning. *Proceedings of the IEEE*, 109(5):612–634. Conference Name: Proceedings of the IEEE.

Raphael Suter, Đorđe Miladinović, Bernhard Schölkopf, and Stefan Bauer. 2019. Robustly disentangled causal mechanisms: Validating deep representations for interventional robustness. *Preprint*, arXiv:1811.00007.

Rheeya Uppaal, Apratim Dey, Yiting He, Yiqiao Zhong, and Junjie Hu. 2024. Model editing as a robust and denoised variant of dpo: A case study on toxicity. *arXiv preprint arXiv:2405.13967*.

Yixin Wang and Michael I Jordan. 2021. Desiderata for representation learning: A causal perspective. *arXiv preprint arXiv:2109.03795*.

Yixin Wang and Michael I. Jordan. 2022. Desiderata for Representation Learning: A Causal Perspective. *arXiv preprint*. ArXiv:2109.03795 [cs, stat].

Johannes Welbl, Amelia Glaese, Jonathan Uesato, Sumanth Dathathri, John Mellor, Lisa Anne Hendricks, Kirsty Anderson, Pushmeet Kohli, Ben Coppin, and Po-Sen Huang. 2021. Challenges in detoxifying language models. *arXiv preprint arXiv:2109.07445*.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, and 1 others. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 conference on empirical methods in natural language processing: system demonstrations*, pages 38–45.

Albert Xu, Eshaan Pathak, Eric Wallace, Suchin Gururangan, Maarten Sap, and Dan Klein. 2021. Detoxifying language models risks marginalizing minority voices. *arXiv preprint arXiv:2104.06390*.

Xuandong Zhao, Xianjun Yang, Tianyu Pang, Chao Du, Lei Li, Yu-Xiang Wang, and William Yang Wang. 2024. Weak-to-strong jailbreaking on large language models. *arXiv preprint arXiv:2401.17256*.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, and 1 others. 2023. Judging llm-as-a-judge with mt-bench and chatbot arena. *Advances in Neural Information Processing Systems*, 36:46595–46623.

A Computational Resources and Model Parameters

Our experiments primarily involve two large-scale language models: Vicuna-13B (Chiang et al., 2023), comprising approximately 13 billion parameters with 40 layers and 40 heads per layer, and LLaMA-3-8B (AI@Meta, 2024), consisting of around 8 billion parameters with 32 layers and 32 heads per layer.

Each fine-tuning run was performed using NVIDIA A100 GPUs (each with 40GB of memory). Specifically, the computational cost for each step of our experiments is detailed as follows:

- **Activation extraction**: Approximately 1 GPU hour per model and dataset configuration.
- **Head selection and fine-tuning**: Approximately 3 GPU hours per configuration.

- **Intervention experiments (evaluation and inference)**: Ranged from approximately 3 to 8 GPU hours, depending on the model and number of selected heads.

B Implementation and Software Packages

Our experiments were conducted using Python 3.9 and the Hugging Face Transformers (Wolf et al., 2020) library version 4.32.1. Tokenization was handled via `AutoTokenizer` and `LlamaForCausalLM`, with default settings and configurations provided by the respective model authors. For inference-time interventions, our implementation is directly adapted from the publicly available codebase of Li et al. (2023), available at https://github.com/likenneth/honest_llama. We did not modify the original inference-time intervention code significantly beyond minor adaptations to integrate it seamlessly into our experimental pipeline.

Dataset Sensitivity and Model Stability We also find that the ImplicitHate dataset generally saw greater toxicity reductions (35–38% in the best cases) than ToxiGen (25–31%). This suggests the interventions were more effective at reducing overt hate content, whereas ToxiGen's adversarial/offensive examples were harder to detoxify. Additionally, models fine-tuned on Hate maintained relatively low perplexity (Vicuna-13B's perplexity stayed < 20 for ACC methods), but ToxiGen fine-tuning often caused larger perplexity spikes. For instance, Vicuna-13B fine-tuned on ToxiGen with PNS (36 heads) reached only 28% detox but became highly unstable.