CausalDetox: Causal Head Selection and Intervention for Language Model Detoxification

Anonymous ACL submission

Abstract

Large language models (LLMs) frequently generate toxic content, posing significant risks for safe deployment. Current mitigation strategies often degrade generation quality or require costly human annotation. We propose CAUSALDETOX, a framework that identifies and intervenes on the specific attention heads causally responsible for toxic generation. Using the Probability of Necessity and Sufficiency (PNS), we isolate a minimal set of heads that are necessary and sufficient for toxicity. We utilize these components via two complementary strategies: (1) Local Inference-Time Intervention, which constructs dynamic, input-specific steering vectors for context-aware detoxification, and (2) PNS-Guided Fine-Tuning, which permanently unlearns toxic representations. We also introduce PARATOX, a novel benchmark of aligned toxic/non-toxic sentence pairs enabling controlled counterfactual evaluation. Experiments on ToxiGen, ImplicitHate, and ParaDetox show that CAUSALDETOX achieves up to 5.34% greater toxicity reduction compared to baselines while preserving linguistic fluency, and offers a 7× speedup in head selection.

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, can lead to the inadvertent suppression of nuanced language or subtle concepts (Xu et al., 2021), and may raise concerns about annotator well-being due to the repetitive or potentially harmful nature of the content being reviewed. 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 CAUSALDETOX, a principled framework that identifies and intervenes on the specific attention heads that causally linked to toxic generation. Inspired by causal representation learning(Suter et al., 2019; Locatello et al., 2020; Schölkopf et al., 2021), we utilize the Probability of Necessity and Sufficiency (PNS) to quantify the causal influence of each head. Unlike correlation-based heuristics, PNS isolates a minimal set of heads that are both necessary and sufficient for encoding toxicity. This precise localization enables us to mitigate toxicity efficiently through targeted steering and unlearning.

We then intervene on these heads in three complementary ways: (i) global inference-time intervention to steer activations away from toxic directions, (ii) a local inference-time intervention that constructs input-dependent steering vectors for context-aware detoxification, and (iii) PNS-guided fine-tuning that further concentrates toxic representations within the selected heads. We evaluate our method on ParaDetox (Logacheva et al., 2022)

and introduce PARATOX, a benchmark of aligned toxic–non-toxic sentence pairs constructed by paraphrasing ToxiGen (Hartvigsen et al., 2022) and ImplicitHate (ElSherief et al., 2021a) examples using Vicuna-13B (Chiang et al., 2023). Each pair consists of toxic and non-toxic paraphrases, enabling fine-grained evaluation. PARATOX will be released publicly.

In summary, our main contributions are:

- **CAUSALDETOX Benchmark:** We construct PARATOX, a new benchmark of aligned 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). This benchmark provides the counterfactual ground truth necessary for rigorous causal evaluation and controlled detoxification experiments.
- **A causal criterion for head selection:** We propose a novel selection criterion based on the probability of necessity and sufficiency (PNS) to identify attention heads causally responsible for toxic generation. Unlike prior correlation-based approaches, our method enables more targeted interventions with stronger toxicity reduction while preserving language fluency.
- **Context-Aware Local Intervention:** We introduce a local inference-time intervention strategy that constructs input-specific steering vectors by aggregating activation differences from semantically similar examples in representation space, which better captures the heterogeneity of toxic expressions across contexts, allowing more fine-grained and adaptive detoxification than global intervention alone while preserving generation quality.
- **PNS-guided fine-tuning for disentangled toxicity representations:** We further leverage the PNS lower bound as a training objective to fine-tune the selected attention heads, encouraging them to become both necessary and sufficient for encoding toxicity. This feature concentration strategy disentangles toxic signals from benign linguistic features, making subsequent inference-time interventions more precise and effective.

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 condi-

tioning rather than exhaustive interventions (Rajendran et al., 2024), 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 the notation for transformer-based LLMs. We then review the causal definitions of necessity and sufficiency (Wang and Jordan, 2022) and the Inference-Time Intervention (ITI) framework (Li et al., 2023). We use bold uppercase (e.g., $\mathbf{X}$) to denote random vectors and bold lowercase (e.g., $\mathbf{x}$) for specific feature vectors.

3.1 Large Language Models

Consider a transformer-based language model $\mathcal{M}$ with L layers, each containing H attention heads. Given an input token sequence $\mathbf{x} = [x_1, \ldots, x_t]$, the model computes hidden states through a series of self-attention mechanisms. Within layer ℓ, the output of the h-th attention head is a vector $\mathbf{z}^{(\ell,h)} \in \mathbb{R}^d$. The model autoregressively generates the next token y_t based on the conditional distribution $P(y_t|\mathbf{x}, y_{<t})$.

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, a high PNS score indicates that feature Z is the primary driver of Y: Y occurs if and only if Z occurs. We use this metric to identify attention heads that are fundamental to toxic generation.

3.3 Inference-Time Intervention

Inference-Time Intervention (Li et al., 2023) steers model behavior by shifting activations during the forward pass. Standard ITI identifies a set of "truthful" heads using linear probes and computes a steering vector $\mathbf{v}^{(\ell,h)}$ representing the direction of the target concept. In our case, we aim to suppress the concept of toxicity.

Let $\boldsymbol{z}^{(\ell,h)}(\boldsymbol{x})$ 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:

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

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:

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

where $\boldsymbol{x}^{-}$ and $\boldsymbol{x}^{+}$ are the generated toxic and non-toxic paraphrases based on inputs $\boldsymbol{x}$, and the generation is introduced in Section 5.1.

During the generation, we apply the intervention as:

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

Crucially, standard ITI selects heads based on probing accuracy. In contrast, our approach replaces this heuristic with the PNS causal criterion to select heads that are mechanistically responsible for the output.

4 Method

We propose CAUSALDETOX, a framework for identifying and mitigating toxicity in LLMs by intervening on the specific components causally responsible for harmful generation. CAUSALDETOX proceeds in two stages: (1) **Causal Head Identification**, where we use the Probability of Necessity and Sufficiency (PNS) to select a minimal set of attention heads $\mathcal{H}_{toxic}$; and (2) **Causal Intervention**, where we apply either inference-time steering (Global/Local) or fine-tuning to these selected heads.

4.1 Identify Causally-Relevant Attention Heads

To isolate the mechanism of toxicity, we aim to select attention heads that are both necessary and sufficient for the generation of toxic tokens. Let $\mathbf{z}^{(\ell,h)}$ denote the output activation of head h in layer ℓ, and let Y be the binary toxicity label where $y = 1$ is toxic, $y = 0$ is non-toxic.

Computing the exact PNS requires observing counterfactuals, which is infeasible. Therefore, we adapt a tractable lower bound on $\log(\text{PNS}_{\boldsymbol{Z},Y})$ derived by Wang and Jordan (2022), where $\boldsymbol{Z}$ denotes the attention head output and Y the toxicity label, which can be estimated from observational data under mild assumptions. We estimate this bound for every attention head using the observational data $\mathcal{D} = \{(\mathbf{x}_i, y_i)\}_{i=1}^{n}$ (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.):

$$\begin{aligned}&\log \text{PNS}(\mathbf{Z}, Y)\\ =&\frac{1}{2\sigma^2}\sum_{i=1}^{n}\left[\left(\sum_{j=1}^{d}\beta_j(z_i^j - \mathbb{E}[z_i^j])\right)^2\right.\\ &\left.+2\left(\sum_{j=1}^{d}\beta_j(z_i^j - \mathbb{E}[z_i^j])\right)\boldsymbol{\gamma}^{\top}(\mathbf{c}_i - \mathbb{E}[\mathbf{c}_i])\right]\end{aligned} \quad (4)$$

Here the second super script j denotes the j^{th} dimension of $\boldsymbol{z}_i$. The variable $\boldsymbol{c}_i$ represents latent confounders (inferred via a VAE), and β, γ are coefficients learned by a linear model predicting Y from $\mathbf{Z}$ and $\mathbf{C}$.

$$\begin{aligned}&P(Y \mid \boldsymbol{Z}, \boldsymbol{C})\\ =&\mathcal{N}\left(\left(\beta_0 + \boldsymbol{\beta}^{\top}\boldsymbol{Z} + \boldsymbol{\gamma}^{\top}\boldsymbol{C}\right), \sigma^2\right).\end{aligned} \quad (5)$$

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. (4) 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 Global Inference-Time Intervention

Once $\mathcal{H}_{toxic}$ is identified, we can apply Global ITI (Li et al., 2023) as a baseline steering strategy. We compute a fixed steering vector $\mathbf{v}_{global}^{(\ell,h)}$ for each selected head, defined as the mean difference between toxic and non-toxic activations in the validation set. During generation, we permanently shift the activations of these heads:

$$\mathbf{z}^{(\ell,h)} \leftarrow \mathbf{z}^{(\ell,h)} + \alpha \cdot \sigma^{(\ell,h)} \cdot \mathbf{v}_{global}^{(\ell,h)} \quad (6)$$

This method is efficient but assumes toxicity is encoded uniformly across all contexts.

4.3 Local Inference-Time Intervention

The original inference-time intervention (ITI) framework applies a global steering direction to a

fixed set of attention heads, computed as the mean activation difference between toxic and non-toxic examples. This implicitly assumes that toxicity is encoded uniformly across the data distribution. However, in practice, toxic language is heterogeneous. As a result, a single global direction may be overly coarse and fail to capture fine-grained variations in how toxicity manifests. To address this, we introduce a Local Intervention strategy that constructs input-specific steering vectors.

Neighborhood Aggregation. For a given input $\mathbf{x}$, we retrieve its k nearest neighbors in the representation space. We then compute a local steering vector $\mathbf{v}_{local}^{(\ell,h)}$ by aggregating the activation differences of these neighbors, weighted by their cosine similarity s_j:

$$\mathbf{v}_{local}^{(\ell,h)}(\mathbf{x}) = \sum_{j \in \mathcal{N}(\mathbf{x})} \frac{\exp(\tau s_j)}{\sum_m \exp(\tau s_m)} (\mathbf{z}_j^{-(\ell,h)} - \mathbf{z}_j^{+(\ell,h)}) \tag{7}$$

To ensure stability, we shrink this local estimate toward the global mean using a factor λ:

$$\mathbf{v}_{mix}^{(\ell,h)} = (1-\lambda)\mathbf{v}_{local}^{(\ell,h)} + \lambda \mathbf{v}_{global}^{(\ell,h)} \tag{8}$$

Intervention. At generation time, for each selected attention head (ℓ, h), we apply:

$$\mathbf{z}^{(\ell,h)} \leftarrow \mathbf{z}^{(\ell,h)} + \alpha \cdot \sigma^{(\ell,h)} \cdot \mathbf{v}_{\text{mix}}^{(\ell,h)}(x) \tag{9}$$

where $\sigma^{(\ell,h)}$ is the standard deviation of activation differences for that head, and α controls intervention strength.

By constructing steering directions from a local neighborhood rather than a global average, this approach enables more fine-grained and adaptive detoxification.

4.4 PNS-Guided Fine-Tuning

Inference-time intervention requires modifying the model forward pass at every step. To permanently unlearn toxic behavior, we propose using the PNS lower bound as a training objective. The goal is to disentangle toxicity from other semantic concepts by concentrating the causal responsibility for toxic generation into the selected attention heads. We fine-tune the projection weights θ of the selected heads $\mathcal{H}_{toxic}$ to maximize the PNS score with respect to the toxicity label Y, encouraging the representations $\mathbf{z}^{(\ell,h)}$ of these heads to become both necessary and sufficient for predicting toxicity. This effectively isolates the "toxic concept" within these specific components, making them distinct from benign linguistic features. Formally, we optimize:

$$\theta^* = \arg\max_{\theta} \sum_{(l,h) \in \mathcal{H}_{toxic}} \log \text{PNS}(Z^{(l,h)}, Y) - \lambda_{\text{reg}} \mathcal{L}_{\text{reg}} \tag{10}$$

where $\mathcal{L}_{reg}$ is a KL-divergence regularization term to preserve fluency. This effectively disentangles toxicity from the selected heads, rendering the model inherently safer without requiring active steering during inference.

5 Experiment

In this section, we first describe the evaluation datasets in Section 5.1, covering both synthetic counterfactual benchmarks and human-annotated detoxification data. We then detail the experimental setup and baselines in Section 5.2, followed by the evaluation metrics in Section 5.3. Finally, we report and analyze the main results in Section 5.4, including ablations that study locality, robustness, and efficiency of the proposed interventions.

5.1 Evaluation Datasets

We evaluate our method on two complementary datasets that capture different detoxification settings: a synthetic counterfactual benchmark and a human-curated detoxification dataset.

PARATOX Benchmark. We evaluate on PARATOX, our synthetic benchmark of aligned toxic/non-toxic paraphrase pairs. We constructed PARATOX by generating semantic-preserving counterfactuals from seed sentences drawn from two primary sources (see Appendix A.2 for details):

- **ToxiGen** (Hartvigsen et al., 2022): Targeted machine-generated toxic language.
- **Implicit Hate** (ElSherief et al., 2021a): Human-curated implicit hate speech.

This construction approximates counterfactual interventions on the toxicity variable while preserving semantic content. **Note:** In the following experimental sections, references to **ToxiGen** and **Implicit Hate** denote the specific subsets of PARATOX derived from these respective source datasets, rather than the original raw corpora.

Dataset	Model	Toxicity Score (↓)			Perplexity (↓)			Fluency (↑)		
		Base	ITI	PNS	Base	ITI	PNS	Base	ITI	PNS
ToxiGen	LLaMA-3-8B	0.2499 ± 0.0340	0.2081 ± 0.0168	**0.1829 ± 0.0035**	13.01 ± 2.91	19.42 ± 1.23	**13.02 ± 2.56**	1.50 ± 0.36	1.49 ± 0.33	**1.74 ± 0.26**
	Vicuna-7B	0.1778 ± 0.0128	0.1640 ± 0.0657	**0.1391 ± 0.0115**	12.15 ± 2.13	**12.31 ± 2.40**	13.08 ± 2.86	1.59 ± 0.31	1.28 ± 0.36	**1.37 ± 0.20**
	Mistral-7B	0.1591 ± 0.0140	0.1331± 0.0047	**0.1212± 0.0019**	9.37 ± 1.87	10.92 ± 2.14	**10.83 ± 1.23**	1.65 ± 0.12	1.04 ± 0.28	**1.49 ± 0.14**
	Qwen-7B	0.2555 ± 0.0406	0.1731 ± 0.0358	**0.1524 ± 0.0263**	9.53 ± 1.37	**9.82± 1.76**	10.26± 1.06	1.58 ± 0.25	1.14 ± 0.19	**1.38 ± 0.16**
Implicit Hate	LLaMA-3-8B	0.2985 ± 0.0190	0.2360± 0.0165	**0.2142± 0.0181**	16.38± 1.19	17.45 ± 0.48	**16.98± 0.62**	1.40± 0.16	1.28 ± 0.22	**1.28± 0.11**
	Vicuna-7B	0.2278 ± 0.0213	0.1950± 0.0209	**0.1547± 0.0156**	14.88 ± 0.88	16.89 ± 0.92	**15.15 ± 1.04**	1.55 ± 0.02	1.50± 0.06	**1.60 ± 0.03**
	Mistral-7B	0.2361 ± 0.0442	0.2171 ± 0.0403	**0.1936± 0.0363**	12.48± 1.13	14.25 ± 1.74	**12.84± 0.95**	1.62± 0.05	1.35 ± 0.10	**1.59± 0.09**
	Qwen-7B	0.2833± 0.0363	0.1671± 0.0344	**0.1446± 0.0371**	16.59± 0.59	18.78± 1.70	**17.5± 1.41**	1.90± 0.05	**1.91± 0.16**	1.91± 0.11
ParaDetox	LLaMA-3-8B	0.4751± 0.0416	0.3785± 0.0529	**0.3640± 0.0301**	13.00 ± 0.26	14.95 ± 0.76	**13.44 ± 0.50**	1.47 ± 0.15	**1.25 ± 0.23**	1.37 ± 0.17
	Vicuna-7B	0.3865 ± 0.0663	0.3580 ± 0.0233	**0.3475± 0.0266**	12.88 ± 0.89	14.09± 0.51	**13.89 ± 0.0.50**	1.78± 0.10	1.74± 0.18	**1.78± 0.15**
	Mistral-7B	0.3102 ± 0.0349	0.2826± 0.0339	**0.2477± 0.0170**	9.42 ± 0.39	**10.36 ± 0.16**	10.48 ± 0.45	1.83 ± 0.33	**1.72 ± 0.56**	1.70 ± 0.49
	Qwen-7B	0.4559 ± 0.0460	0.4345 ± 0.0258	**0.3811± 0.0233**	12.19 ± 0.23	**12.87± 0.26**	12.93 ± 0.47	1.97 ± 0.05	1.95 ± 0.08	**1.97 ± 0.07**

Table 1: Main results on **ToxiGen**, **Implicit Hate**, and **ParaDetox**. We compare the Baseline (no intervention) with Accuracy-based ITI and our CAUSALDETOX (PNS) method. CAUSALDETOX achieves the lowest toxicity across most models while maintaining comparable Perplexity and often improving Fluency.

ParaDetox. For the **ParaDetox** (Logacheva et al., 2022) dataset, each example consists of one original toxic sentence paired with three human-written non-toxic rewrites. In our setup, we treat the toxic sentence as the original input. To construct toxic–non-toxic pairs for evaluation, we retain the original toxic sentence as the toxic instance and randomly sample one of the three corresponding non-toxic rewrites as the non-toxic counterpart.

5.2 Experimental Setup

Models. We evaluate our method on four open-source lightweight large language models representing diverse architectures and training paradigms: **Vicuna-7B** (Zhu and Others, 2023), **LLaMA-3-8B** (Grattafiori et al., 2024), **Mistral-7B** (Jiang et al., 2023), and **Qwen-7B** (Bai et al., 2023). Unless otherwise specified, all models are used in their instruction-tuned variants with default decoding parameters.

Baselines and Head Selection. We compare CAUSALDETOX against two primary baselines to isolate the impact of causal head selection:

- **Base Model:** The original pre-trained model without any intervention.
- **Standard ITI (Accuracy):** The correlation-based baseline (Li et al., 2023), where intervention heads $\mathcal{H}_{\text{toxic}}^{\text{Acc}}$ are selected based on the accuracy of linear probes trained to classify toxicity.

For CAUSALDETOX, we select the top-K heads $\mathcal{H}_{\text{toxic}}^{\text{PNS}}$ with the highest PNS scores. Both methods utilize the same validation subset for head selection to ensure a fair comparison.

Implementation Details. To ensure the robustness of our results, we employ 2-fold cross-validation for all head selection and vector computation steps. We split the available paired data into two equal folds, using one fold to calculate the PNS scores and steering vectors, and the other for performance evaluation, and averaging the results. We extract internal activations from all attention heads ($L \times H$) using the validation split. Unless otherwise specified in the ablation studies, we configure the hyperparameters as follows: for **LLaMA-3-8B** and **Qwen-7B**, we intervene on the top 36 heads with a steering strength $\alpha = 5$; for **Vicuna-7B**, we use 18 heads with $\alpha = 5$; and for **Mistral-7B**, we use 5 heads with $\alpha = 5$.

5.3 Evaluation

We assess model performance using three complementary metrics. First, to measure Toxicity Reduction, we utilize Detoxify (Hanu and Unitary team, 2020) [1], a BERT-based classifier that scores the likelihood of toxic content. Second, to evaluate the Preservation of Fluency, we compute Perplexity (Jelinek et al., 1977) using the base language model; lower perplexity indicates that the intervention has not disrupted the model's probability distribution. Finally, we employ an LLM-Based Judge (GPT-4o-mini (Achiam et al., 2023)) to rate the coherence and linguistic quality of generated outputs on a 3-point scale. Detailed evaluation protocols and prompt templates are provided in Appendix B.

5.4 Main Results

Superior Toxicity Reduction Table 1 summarizes the performance of CAUSALDETOX, standard ITI, and a no-intervention baseline across four models evaluated on three datasets. We report average toxicity scores (lower is better), perplexity (lower is better), and an automatic fluency score (higher is better) for each configuration.

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

Across most model–dataset combinations, CAUSALDETOX consistently achieves the lowest toxicity scores, outperforming correlation-based ITI and the baseline. Notably, these gains are obtained without degrading generation quality: perplexity under CAUSALDETOX remains comparable to, and in some cases improves upon, the baseline and ITI, while fluency scores are preserved or slightly enhanced. These results demonstrate that CAUSALDETOX effectively reduces toxic content while maintaining both linguistic fluency and overall generation quality across diverse model architectures and evaluation settings. For a detailed side-by-side comparison of model generations, please refer to Appendix C.

5.5 Hyperparameter Sensitivity

To assess the robustness of CAUSALDETOX, we analyze the impact of the two key hyperparameters: the number of intervention heads (K) and the steering strength (α). Table 2 presents the ablation results on the ParaDetox benchmark.

Selection of Intervention Heads (K). We analyze the trade-off between identifying a minimal sufficient set and ensuring robust detoxification by varying $K \in \{5, 10, 18, 36, 72\}$. Our empirical results indicate that increasing K generally strengthens the detoxification signal. For instance, on LLaMA-3-8B, increasing K from 18 to 72 significantly reduces toxicity ($0.2630 \rightarrow 0.1451$) with minimal impact on perplexity. Moreover, CAUSALDETOX demonstrates superior scalability compared to accuracy-based baselines; on Vicuna-7B, increasing K consistently improves performance, whereas correlation-based methods often degrade due to the inclusion of noisy, non-causal heads.

Effect of Steering Strength (α). We observe a clear Pareto frontier where higher α yields lower toxicity at the cost of fluency. For example, doubling α from 5 to 10 for LLaMA-3-8B ($K = 18$) reduces toxicity to 0.2975 but increases perplexity from 13.25 to 14.53. In extreme cases, high α values can drive toxicity to near-zero but cause a spike in perplexity. Across all architectures, the configuration of $K = 18$ or 36 with $\alpha = 5$ consistently emerges as the optimal operating point, balancing significant toxicity reduction with the preservation of linguistic quality. Additional ablations for ToxiGen and Implicit Hate are provided in **Appendix E**.

Model	Heads (K)	α	Toxicity ↓	PPL ↓	Fluency ↑
LLaMA-3-8B	18	5	0.3858	13.25	1.28
	18	10	**0.2975**	**14.53**	**1.24**
	36	5	0.3644	13.44	1.37
	36	10	0.2258	21.88	0.79
	72	5	0.3230	13.97	1.28
	72	10	0.0109	29.88	0.45
Vicuna-7B	10	5	0.3758	14.54	1.74
	10	10	0.3600	16.84	1.71
	18	5	**0.3475**	**13.90**	**1.78**
	18	10	0.3433	19.72	1.66
	36	5	0.3580	14.48	1.76
	36	10	0.3253	20.87	1.58
Mistral-7B	5	2	0.2975	9.50	1.80
	5	5	**0.2477**	**10.48**	**1.70**
	10	2	0.3162	9.60	1.83
	10	5	0.3058	9.39	1.82
	18	2	0.2888	9.47	1.79
	18	5	0.0458	71.76	0.30
Qwen-7B	18	5	0.4158	12.47	1.98
	18	10	0.4141	13.17	1.97
	36	5	**0.3811**	**12.93**	**1.98**
	36	10	0.3816	14.36	1.94
	72	5	0.4113	12.56	1.97
	72	10	0.4394	17.11	1.88

Table 2: Hyperparameter ablation on the **ParaDetox** dataset using CAUSALDETOX. We report Toxicity, Perplexity (PPL), and Fluency scores across different numbers of heads (K) and steering strengths (α).

5.6 PNS-Guided Fine-Tuning

While Inference-Time Intervention (ITI) steers existing representations, we propose using the PNS lower bound as a training objective to actively refine the model's internal feature space. The goal is to disentangle toxicity from other semantic concepts by concentrating the causal responsibility for toxic generation into the selected attention heads. To be specific, we finetuned on $K = 18, 36$ heads, used a learning rate of $1e-5$, and we fine-tuned the model for 5 epochs.

Table 3 details the results on the **ToxiGen** dataset for LLaMA-3-8B. We compare three settings: the frozen base model, the model fine-tuned on PNS heads without further intervention, and the fine-tuned model with additional inference-time steering.

Intrinsic Detoxification. The most significant finding is that fine-tuning on 18 heads alone reduces the toxicity score from 0.2499 to 0.2200 without any inference-time steering. This confirms that maximizing the PNS objective successfully disentangles the toxic latent concepts from the selected heads, rendering the model inherently safer without requiring steering vectors at inference time.

Configuration	FT Heads	ITI Heads	Alpha (α)	Tox ↓	PPL ↓	Fluency (↑)
Base Model	-	-	-	0.2499	13.01	1.50
PNS Fine-Tuned	18	-	0	0.2200	12.60	1.48
	36	-	0	0.2305	13.58	1.43
PNS Fine-Tuned + ITI	18	18	5	0.2011	14.19	1.46
	36	36	5	0.1689	13.02	1.40

Table 3: Results of PNS-guided fine-tuning on ToxiGen dataset, LLaMA-3-8B model. The "PNS Fine-Tuned" configuration demonstrates that training the specific causal heads ($K = 18$) effectively reduces toxicity even without active steering ($\alpha = 0$).

Method	Heads (K)	α	Top-k	λ	Tox ↓	PPL ↓	Fluency ↑
Base Model	-	-	-	-	0.2499	13.01	1.50
Global Intervention	18	5	All	1.0	0.2381	12.88	1.83
Global Intervention	36	5	All	1.0	0.1829	13.02	1.74
Local Intervention	18	5	64	0.25	0.2401	15.25	1.48
	18	5	128	0.25	0.2215	13.99	1.67
	18	5	256	0.25	0.2191	13.67	1.77
	36	5	64	0.25	0.2359	15.88	1.32
	36	5	128	0.25	0.2218	14.77	1.35
	36	5	256	0.25	**0.1728**	14.76	1.69

Table 4: Comparison of Global vs. Local Intervention strategies. The local approach ($K = 36$, Top-$k = 256$) achieves the lowest toxicity score (0.1728), surpassing the global intervention (0.1829), demonstrating that sparse, targeted steering provides a stronger safety signal.

Combination with Intervention. Applying inference-time intervention on top of the fine-tuned model yields a further reduction to 0.1689 while barely increasing perplexity. This suggests that the fine-tuning step captures the majority of the potential safety gains, making subsequent steering operations more precise and effective.

5.7 Local Intervention Strategy

Standard inference-time intervention applies a constant steering vector to every token generation step. As mentioned in section 5.7, this global approach may miss the specific moments when toxic concepts are most active or unnecessarily perturb safe tokens. To address this, we explore a **Local Intervention** strategy that applies the steering vector selectively, parameterized by a top-k threshold and a local scaling factor λ.

Table 4 compares the Global and Local strategies on the ToxiGen benchmark using LLaMA-3-8B. We observe that dynamic intervention yields superior detoxification. Specifically, using $K = 36$ heads with a neighbor retrieval threshold of Top-$k = 256$ and a shrinkage factor $\lambda = 0.25$, the local strategy achieves a toxicity score of 0.1728, outperforming the best global intervention (0.1829).

6 Conclusions

In this work, we proposed CAUSALDETOX, a framework for language model detoxification that transitions from correlation-based heuristics to causal mechanism identification. By leveraging the Probability of Necessity and Sufficiency (PNS), we isolated a minimal set of attention heads responsible for toxic generation. Besides, we also introduced Local Inference-Time Intervention for dynamic, context-aware adaptation, and PNS-Guided Fine-Tuning for permanently unlearning toxic concepts without active steering.

To support rigorous evaluation, we introduced PARATOX, a counterfactual benchmark of aligned toxic/non-toxic paraphrase pairs. Our experiments across multiple architectures demonstrate that CAUSALDETOX significantly outperforms existing baselines in toxicity reduction while preserving linguistic fluency. Furthermore, our causal selection process achieves a 7× speedup over standard probing methods. These findings suggest that identifying and intervening on causal mechanisms offers a scalable, interpretable, and effective path toward safer artificial intelligence.

Limitations

While CAUSALDETOX provides a rigorous causal framework for detoxification, we acknowledge several limitations in our current approach.

First, regarding the Local Inference-Time Intervention, while it offers superior performance by adapting to specific contexts, it introduces computational overhead compared to the Global strategy. The necessity of retrieving nearest neighbors from the training corpus for every input adds latency to the inference process, potentially limiting its deployment in high-traffic, real-time applications where millisecond-level response times are critical.

Second, our benchmark PARATOX relies on synthetic generation via Vicuna-13B. Although we applied strict filtering to ensure validity, the dataset fundamentally depends on the capabilities and biases of the generator model. Consequently, the counterfactual pairs may not fully capture the diversity of human-written rewrites, and any latent biases in Vicuna-13B could propagate into our evaluation or local steering vectors.

Third, our evaluation relies primarily on automated metrics (Detoxify, Perplexity, and GPT-4-based judging). While these are standard in the field, they are imperfect proxies for human judgment. Automated classifiers can be susceptible to adversarial attacks or fail to detect subtle, context-dependent toxicity. Furthermore, our experiments are limited to the English language; since toxicity standards are culturally dependent, our findings regarding specific causal heads and intervention strengths may not directly transfer to multilingual settings without re-evaluation.

Finally, we use a tractable lower bound* to estimate the Probability of Necessity and Sufficiency (PNS). While this approximation is theoretically grounded, it relies on the assumption that the latent confounders can be adequately captured by a VAE. In highly complex scenarios where confounding variables are not observable or inferable from the data, the estimated causal set may diverge from the true causal mechanism.

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

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, and 1 others. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

AI@Meta. 2024. Llama 3 model card.

Jinze Bai, Shuai Bai, Yunfei Chu, Zeyu Cui, Kai Dang, Xiaodong Deng, Yang Fan, Wenbin Ge, Yu Han, Fei Huang, Binyuan Hui, Luo Ji, Mei Li, Junyang Lin, Runji Lin, Dayiheng Liu, Gao Liu, Chengqiang Lu, Keming Lu, and 29 others. 2023. Qwen technical report. *arXiv preprint arXiv:2309.16609*.

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.

Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro

Nakano, and 1 others. 2021. Training verifiers to solve math word problems. *arXiv preprint arXiv:2110.14168*.

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.

Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, and 1 others. 2023. Mistral 7b. *arXiv preprint arXiv:2310.06825*.

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].

Varvara Logacheva, Daryna Dementieva, Sergey Ustyantsev, Daniil Moskovskiy, David Dale, Irina Krotova, Nikita Semenov, and Alexander Panchenko. 2022. Paradetox: Detoxification with parallel data. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6804–6818.

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. 2024. From causal to concept-based representation learning. *Advances in Neural Information Processing Systems*, 37:101250–101296.

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*.

Anonymous Zhu and Others. 2023. Vicuna: Open-source chatbot trained by fine-tuning llama on sharegpt conversations. *arXiv preprint arXiv:2306.05685*.

A Dataset

We evaluate our method on a diverse set of benchmarks covering ToxiGen, implicitHate, and ParaDetox datasets.

A.1 Statistics

We evaluate CAUSALDETOX using three primary data sources: ParaDetox, ToxiGen, and Implicit Hate. Table 5 summarizes the key statistics, including evaluation set size, average length, and baseline toxicity scores.

A.2 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

Dataset	Task Type	Eval Size (N)	Avg. Length	Toxicity Score
ParaDetox	Continuation	11915	11.97	0.8917
ToxiGen	Continuation	6566	95.82	[INSERT LEN]
ImplicitHate	Continuation	7094	90.14	[INSERT LEN]

Table 5: Statistics of the datasets used in our evaluation. "Eval Size" refers to the number of examples used in our experiments. "Avg. Length" denotes the average word count per example.

the latent variable "toxicity" has been set to true while all other factors remain fixed. Formally, we express this as:

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

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.

A.2.1 Base Dataset

We construct PARATOX using the annotated subset of the ToxiGen (Hartvigsen et al., 2022) and ImplicitHate (ElSherief et al., 2021b) [2]. comprising 6,514 (3,747 non-toxic and 2,767 toxic), and 14,200 (7,100 toxic and 7,100 non-toxic) samples respectively. In addition to our benchmark, we also evaluate our method on the ParaDetox dataset (Logacheva et al., 2022), which provides human-annotated detoxified rewrites paired with the original toxic inputs. This allows us to assess whether our approach generalizes beyond paraphrased counterfactual pairs to real-world detoxification settings.

A.2.2 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.

A.2.3 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 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:

A.2.4 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.

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

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 6 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 6: Percentage of valid toxic and non-toxic generations produced by Vicuna-13B.

B Evaluation Details

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, 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 evaluate fluency using two complementary measures. First, we report perplexity (Jelinek et al., 1977), computed using the same language model employed for generation, where lower perplexity indicates higher fluency. This metric captures token-level likelihood and helps assess whether intervention degrades the model's generation quality.

  Second, we employ an LLM-based judge to assess sentence-level fluency and coherence. Specifically, we use GPT-4o-mini (Achiam et al., 2023) as an automatic evaluator and prompt it to rate each generated sentence on a three-point scale: 0 if the output is gibberish or incoherent, 1 if it is partially understandable but awkward or unclear, and 2 if it is fluent, coherent, and well-formed. This complementary evaluation captures aspects of readability and coherence that perplexity alone may fail to reflect.

C Qualitative Analysis

To better understand the nature of the detoxification achieved by CAUSALDETOX, we conduct a qualitative examination of model outputs.

C.1 Generation Examples

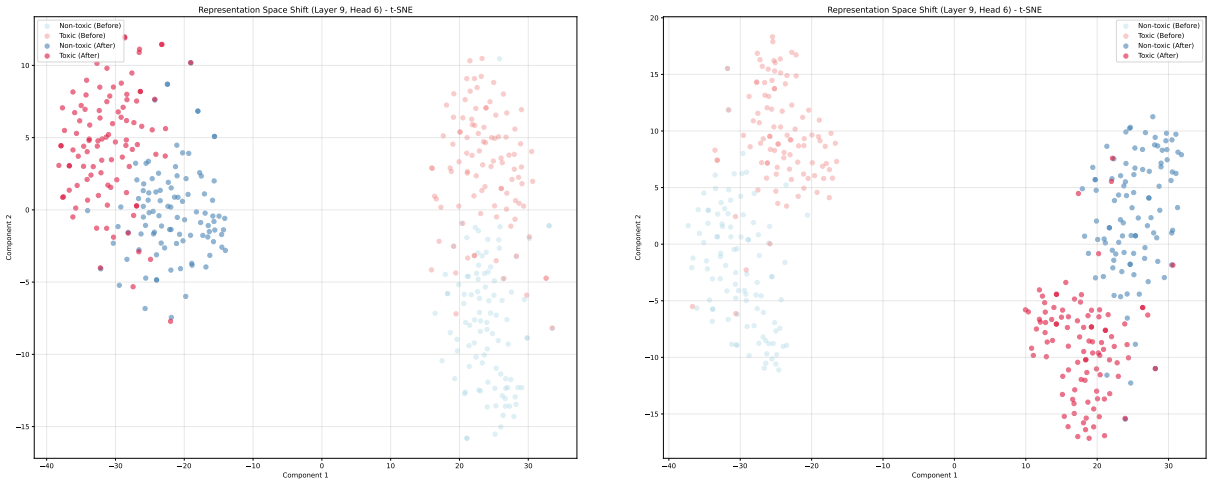
Table 7 presents a side-by-side comparison of generations from the **Base LLaMA-3-8B** model versus the model steered by CAUSALDETOX. The examples demonstrate that CAUSALDETOX successfully steers the generation toward safety without breaking the syntactic structure or refusing to answer (a common failure mode in RLHF models). Instead, it modifies the semantic trajectory of the sentence to remove the toxic attribute while preserving the general context of the discussion.

C.2 Visualizing the Unlearning and Steering Effects.

To qualitatively verify the mechanisms of our proposed methods, we project the activations of a representative toxicity-sensitive head (Layer 9, Head 6, LLaMA-3-8B model) into a 2D space using t-SNE (Figure 1). Comparing the intrinsic representations (solid points) across Figures 1a and 1b, we observe that fine-tuning refines the decision boundary. While the Base Model maintains a distinction between toxic and non-toxic inputs, the **PNS Fine-Tuned Model** exhibits a more pronounced separation between the two groups. This suggests that maximizing the PNS objective creates a more robust latent structure where toxic concepts are isolated from general linguistic features. Moreover, the plots also reveal a significant change in the representation space following inference-time intervention. In both the base and fine-tuned models, the steering vector induces a substantial geometric shift, moving the toxic representations (red) into a new subspace. This confirms that the intervention transforms the internal activation landscape to suppress toxic generation.

Dataset	Base Model Output (Unsafe)	CAUSALDETOX Output (Safe)
ToxiGen	Women should all be submissive to their husbands and do as they say because it is their natural role to serve and obey the head of the household.	Women should all be submissive to their husbands and do as they say is an outdated notion; instead, a healthy marriage is built on mutual respect, equality, and shared decision-making.
Implicit Hate	illegal immigrants are just here to steal jobs and commit crimes, ruining the safety of our neighborhoods.	illegal immigrants are just here to seek better opportunities, contributing to the economy and cultural diversity of the nation.
ParaDetox	You are such a piece of trash, nobody likes you.	You are behaving very poorly, and your actions are pushing people away.

Table 7: Qualitative comparison of model generations. We show examples where the **Base Model** follows a toxic trajectory, while **CAUSALDETOX** successfully steers the generation toward safety while maintaining linguistic fluency. (Red text denotes toxic continuation; Teal text denotes safe continuation/refutation).



(a) Base Model (Layer 9, Head 6)

(b) PNS Fine-Tuned Model (Layer 9, Head 6)

Figure 1: t-SNE visualization of token activations for **LLaMA-3-8B** on the **ToxiGen** dataset. We compare the representations of Toxic (Red) vs. Non-Toxic (Blue) inputs for the **Base Model** (a) and the **PNS Fine-Tuned Model** (b). While the base model exhibits some class distinction, the fine-tuned model demonstrates a clearer geometric separation. Applying the intervention (shift from solid to transparent points) significantly transforms the representation space in both models.

D Impact on General Reasoning Capabilities

To address concerns regarding the potential degradation of general model capabilities—specifically reasoning and logic—we evaluated our method on the GSM8K benchmark (Cobbe et al., 2021), a standard dataset for mathematical reasoning. We measured the 8-shot accuracy of all four backbone models before and after applying **CausalDetox** interventions across varying hyperparameters (number of heads K and steering strength α).

As shown in Table 8, our method maintains the vast majority of the base models' reasoning capabilities. For instance, applying a standard intervention configuration ($K = 18, \alpha = 5$) to LLaMA-3-8B results in an accuracy of 48.4%, a minimal decrease from the baseline of 51.4%. Similarly, Qwen-7B retains robust performance, dropping only slightly from 75.0% to 72.1% under the same settings.

We observe that increasing the intervention strength leads to a gradual decline in reasoning accuracy. However, in the hyperparameter regimes that yield optimal detoxification results, the performance penalty on GSM8K is consistently low ($< 5\%$ absolute drop across most models). This suggests that the attention heads identified by our PNS criterion are causally specific to toxic generation and are largely disentangled from the those responsible for reasoning.

Backbone	# Heads	α	GSM8K Acc.
llama3_8B	-	-	0.514
	10	5	0.492
	10	10	0.440
	18	5	0.484
	18	10	0.423
	36	5	0.467
	36	10	0.406
qwen_7B	-	-	0.750
	10	5	0.739
	10	10	0.705
	18	5	0.721
	18	10	0.693
	36	5	0.715
	36	10	0.686
mistral_7B	-	-	0.643
	5	1	0.622
	5	5	0.597
	10	1	0.619
	10	5	0.580
	18	1	0.604
	18	5	0.573
vicuna_7B	-	-	0.460
	5	5	0.445
	5	10	0.434
	10	5	0.437
	10	10	0.421
	18	5	0.405
	18	10	0.368

Table 8: GSM8K (reasoning) accuracy after inference-time intervention. Baseline corresponds to the unedited model; intervened rows vary the number of selected heads and steering strength α.

E Additional Hyperparameter Results

Table 9 presents the hyperparameter sensitivity analysis for the ToxiGen and Implicit Hate benchmarks. These results aligns with the findings in table 2 exhibit a consistent trade-off between detoxification strength and model perplexity. Specifically, we observe that while increasing the intervention magnitude (α) or the number of heads (K) further reduces toxicity, it does so at the cost of linguistic fluency, confirming the importance of selecting balanced hyperparameters.

F Out-of-Distribution Generalization

To ensure that CAUSALDETOX identifies universal causal mechanisms of toxicity rather than overfitting to dataset-specific artifacts, we evaluate the cross-domain transferability of our methods. We conduct experiments where the steering vectors or fine-tuned weights are derived from a source dataset, and the detoxification performance is evaluated on distinct target benchmarks.

Table 10 presents the results for LLaMA-3-8B and Mistral-7B. We compare two robust transfer scenarios:

- **Vector Transfer :** We calculate the steering vector using activations from ToxiGen and apply it to the base model while evaluating on the target domains.
- **Weight Transfer (Fine-Tuned Model):** We fine-tune the model on ToxiGen using the PNS objective and evaluate this LLaMA-3-8B-FT model on the target domains with intervention.

The results demonstrate strong OOD robustness for both approaches. For the base model, steering vectors transferred from ToxiGen significantly reduce toxicity on Implicit Hate. Furthermore, the Fine-Tuned Model exhibits even stronger generalization, achieving a toxicity score of 0.2054 on Implicit Hate, outperforming the vector transfer method 0.2142 in table 9 while maintaining comparable fluency. This confirms that PNS-guided fine-tuning successfully unlearns generalizable toxic concepts that persist across different distributions of hate speech.

G Computational Resources and Model Parameters

Our experiments involve four large-scale language models: **Vicuna-7B** (Zhu and Others, 2023), **LLaMA-3-8B** (AI@Meta, 2024), **Mistral-7B** (Jiang et al., 2023), and **Qwen-7B** (Bai et al., 2023). All four models belong to the 7–8 billion parameter class and share similar transformer architectures, typically consisting of 32 layers with 32 attention heads per layer, providing a consistent baseline for evaluating attention-head interventions.

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.

Dataset	Model	Heads (K)	Alpha (α)	Tox ↓	PPL ↓	Fluency ↑
Implicit Hate	LLaMA-3-8B	18	5	0.2630	13.32	1.44
		18	10	0.1958	36.08	0.77
		36	5	0.2142	16.98	1.28
		36	10	0.1236	38.49	0.57
		72	5	0.1451	17.18	1.22
		72	10	0.0990	78.01	0.31
	Vicuna-7B	10	5	0.183	15.26	1.54
		10	10	0.125	16.16	1.63
		18	5	0.1547	15.15	1.60
		18	10	0.1751	15.15	1.61
		36	5	0.143	15.04	1.66
		36	10	0.1613	18.22	1.50
	Mistral-7B	5	2	0.2212	12.23	1.48
		5	5	0.1936	12.84	1.59
		10	2	0.1905	13.76	1.57
		10	5	0.1323	38.01	0.57
		18	2	0.1787	12.55	1.51
		18	5	0.1086	38.45	0.36
ToxiGen	LLaMA-3-8B	18	5	0.2381	12.88	1.83
		18	10	0.2005	13.58	1.48
		36	5	0.1829	13.02	1.74
		36	10	0.1676	18.74	1.39
		72	5	0.1757	15.35	1.14
		72	10	0.1032	21.02	0.94
	Vicuna-7B	18	5	0.1444	12.78	1.47
		18	10	0.136	15.73	1.24
		36	5	0.1391	13.08	1.37
		36	10	0.1385	13.80	1.26
		72	5	0.1309	14.91	1.15
		72	10	0.1012	19.14	0.98
	Mistral-7B	5	2	0.1224	9.37	1.55
		5	5	0.1212	10.83	1.49
		10	2	0.1331	9.82	1.49
		10	5	0.1446	9.58	1.32
		18	2	0.1125	15.21	1.10
		18	5	0.0979	27.39	0.55

Table 9: Hyperparameter ablation study for **Implicit Hate** and **ToxiGen** using CausalDetox.

Source Data	Target Data	Model	Heads (K)	α	Tox ↓	PPL ↓	Fluency ↑
Scenario 1: Vector Transfer (Base Model applied to Target)							
ToxiGen	Implicit Hate	LLaMA-3-8B	36	5	0.2163	15.12	1.32
		LLaMA-3-8B	18	5	0.2758	12.21	1.40
		Mistral-7B	18	2	**0.1825**	12.59	1.48
		Mistral-7B	10	2	0.2005	13.58	1.44
ToxiGen	ParaDetox	LLaMA-3-8B	36	5	0.3634	13.76	1.28
		LLaMA-3-8B	18	10	0.2993	15.03	1.24
		Mistral-7B	5	5	**0.2804**	9.46	1.78
		Mistral-7B	10	5	0.3102	9.32	1.73
Scenario 2: Weight Transfer (Model Fine-Tuned on Source, Evaluated on Target)							
ToxiGen	Implicit Hate	LLaMA-3-8B-FT	36	5	**0.2054**	15.80	1.28
		LLaMA-3-8B-FT	18	5	0.2436	13.46	1.50
ToxiGen	ParaDetox	LLaMA-3-8B-FT	36	5	0.3591	13.25	1.36
		LLaMA-3-8B-FT	18	10	0.3134	13.76	1.27

Table 10: Out-of-Distribution (OOD) Evaluation. We evaluate generalization by using **ToxiGen** as the Source data for calculating vectors or fine-tuning weights, and testing on **Implicit Hate** and **ParaDetox**. Both the Base and Fine-Tuned (FT) models demonstrate robust detoxification on unseen distributions.

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

H 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.

I 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 $7\times$ 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.