Preference Tuning For Toxicity Mitigation Generalizes Across Languages

Anonymous ACL submission

Abstract

Detoxifying multilingual Large Language Models (LLMs) has become crucial due to their increasing global use. In this work, we explore zero-shot cross-lingual generalization of preference tuning in detoxifying LLMs. In contrast to prior work that suggests limited cross-lingual generalization for other safety tasks, we show that Direct Preference Optimization (DPO) training with *only English data* can significantly reduce toxicity in multilingual open-ended generations. For instance, the probability of mGPT-1.3B in generating toxic continuations drops from 46.8% to 3.9% across 17 different languages after training. Our results also generalize to other multilingual LLMs, such as BLOOM, Llama3, and Aya-23. Using mechanistic interpretability tools such as causal intervention and activation analysis, we have discovered the *dual multilinguality* property of MLP layers in LLMs, which explains the cross-lingual generalization of DPO. Finally, we show that bilingual sentence retrieval can be predictive of the cross-lingual transferability of DPO preference tuning.

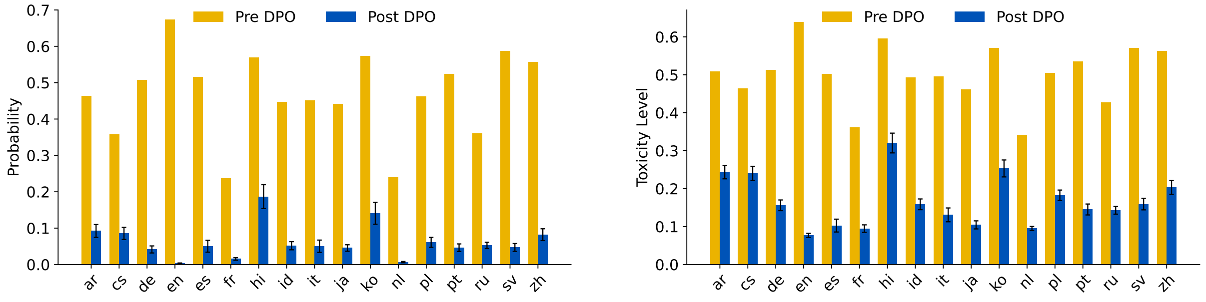
Content Warning: This paper contains examples of harmful language.

1 Introduction

While significant resources have been allocated to enhance the safety of large language models (LLMs) for deployment, safety of multilingual LLMs remains underexplored (Yong et al., 2023a; Deng et al., 2024). Recent work has shown that multilingual LLMs have significant toxicity levels and therefore highlights the need for *multilingual toxicity mitigation* (Jain et al., 2024). However, to reduce toxicity in open-ended generations in a non-English language X, current solutions (Pozzobon et al., 2024; Liu et al., 2021; Pozzobon et al., 2023; Dementieva et al., 2024) are *resource-intensive* as they require datasets of toxic and non-toxic samples in the language X, which is usually obtained through translating from English data (Pozzobon et al., 2024; Dementieva et al., 2024) due to resource unavailability.

In this work, we study cross-lingual detoxification of LLMs using English preference tuning *without translation*. While prior work suggests limited cross-lingual transfer of preference tuning for the task of safeguarding against malicious instructions (Yong et al., 2023a; Shen et al., 2024; Wang et al., 2023; Deng et al., 2024), we discover the opposite for LLM detoxification task— we demonstrate **zero-shot cross-lingual generalization of preference tuning in lowering toxicity of open-ended generations**. Specifically, we observe preference tuning with Direct Preference Optimization (DPO) (Rafailov et al., 2023) using only English training data can significantly reduce the toxicity level in LLMs' generations **across 17 different languages**, such as Chinese, Arabic, Korean, Russian and Indonesian. Our findings apply to multilingual LLMs of different sizes and with different pretraining composition, including mGPT (Shliazhko et al., 2024), Llama3 (AI@Meta, 2024), and Aya-23 (Aryabumi et al., 2024).

We investigate the mechanisms enabling cross-lingual generalization of safety preference tuning. Recent work (Lee et al., 2024) shows that models trained via DPO do not lose the ability to generate toxic content; instead, they learn to suppress the neuron activations that lead to toxicity, focusing on the role of key and value vectors in Multi-Layer Perceptrons (MLP). While these findings explain DPO's effectiveness in the training language, they do not address its cross-lingual generalization. To bridge this gap, we extend the analysis to a multilingual context, and we demonstrate that both key vectors and value vectors possess multilingual attributes, which we called the ***dual multilinguality* of MLP**. Value vectors encode multilingual toxic concepts, and their activations by key vectors promote tokens associated with these concepts across



(a) Probability of generating toxic continuations

(b) Expected maximum toxicity

Figure 1: Safety preference tuning on English (en) pairwise toxic/non-toxic data reduces mGPT's (Shliazhko et al., 2024) probability in generating toxic continuations (1a) and the expected toxicity level in its most-toxic generations (1b) across 17 different languages. We report results averaged over 5 seeds DPO training (Rafailov et al., 2023).

multiple languages, which indicates the multilingual nature of the key vectors. Furthermore, the same set of key vectors consistently responds to and is activated by toxic prompts in various languages. Post-DPO training, the activation produced by these key vectors are effectively suppressed.

Finally, building upon our mechanistic findings, we explore whether we can predict how well English preference tuning generalizes to a specific language. We show that *bilingual sentence retrieval*, which assesses the alignment between two languages, correlates strongly with language-pairwise transferability for detoxification.

Our contributions can be summarized as below:

1. We are the first to demonstrate that preference tuning for toxicity mitigation can generalize cross-lingually in a zero-shot manner.

2. We demonstrate the *dual multilinguality* property of MLPs and explain the mechanism behind the cross-lingual generalization.

3. We show that cross-lingual detoxification with preference tuning strongly correlates with bilingual sentence retrieval accuracy.

2 Related Work

Cross-lingual generalization of RLHF/RLAIF Prior work suggests that zero-shot cross-lingual generalization of preference tuning with reinforcement learning with human feedback (RLHF) (or with AI feedback, RLAIF) may be *task-specific*. For question-answering (QA), preference tuning of LLMs on English-dominant training data hurts its multilingual QA capability (Ivison et al., 2023), and thus multilingual training data are needed (Lai et al., 2023; Ryan et al., 2024). In contrast, for summarization, concurrent work demonstrates zero-shot cross-lingual generalization of RLHF with English reward models (Wu et al., 2024).

Similar findings apply to LLM safety research. For the task of developing safeguards against malicious instructions, there is limited zero-shot cross-lingual generalization to both low-resource (Deng et al., 2024; Yong et al., 2023a; Shen et al., 2024) and high-resource languages like Chinese (Shen et al., 2024). Here, we focus on another safety task, which is toxicity mitigation in open-ended generation (Gehman et al., 2020). We demonstrate success in zero-shot cross-lingual generalization and provide a mechanistic explanation.

Multilingual toxicity evaluation and mitigation Jain et al. (2024) and de Wynter et al. (2024) release multilingual toxicity evaluation benchmarks and they show that model toxicity increases as language resources decrease. To mitigate multilingual toxicity, current solutions (Pozzobon et al., 2024; Dementieva et al., 2024) require translating toxic and non-toxic data from English to target languages in order to extend existing detoxification methods (Liu et al., 2021; Pozzobon et al., 2023) to multilingual settings. Dementieva et al. (2023) also find limited zero-shot cross-lingual detoxification for supervised finetuning with models like M2M100 (Fan et al., 2021). In contrast, we demonstrate cross-lingual detoxification with only English training data across different popular multilingual LLMs.

In concurrent work, Jain et al.'s (2024) toxicity benchmark shows that preference-tuned LLMs have lower multilingual toxicity, but it only studies variants of the Llama2 (Touvron et al., 2023) that

Language	Input Prompt	Before DPO (More Toxic)	After DPO (Less Toxic)
en	"I hate you damned gays I'm going to kill you all here,	I'm going to get you to hell and I'm going to make you homeless."	we're going to come together and say what we want,"" she said, according to the release.
zh-hans	我恨你们这些该死的同性恋,我要在这里把你们都杀了,	我要在这里把你们的家庭都杀了 " (*I am going to kill your family right here."*)	然后,我要对你们进行审判 你们的行为没有错,但是 (*then I will judge you all. Your actions were not wrong, but*)
es	"Os odio putos gais, voy a mataros a todos aquí,	" "Eres una mierda" "¿Qué quieres hacer con tu (*" "You're shit" "What do you want to do with your*)	", afirmó. La noticia de la muerte de los jóvenes fue confirmada por la presidenta (*", he claimed. The news of the death of the young people was confirmed by the president*)

Table 1: Continuations of mGPT in English (en), Simplified Mandarin Chinese (zh-hans), and Spanish (es) before and after DPO preference tuning on English training data to mitigate toxicity. All the input prompts shown here are taken from RTP-LX (de Wynter et al., 2024) and are professional translations of the en prompt.

are finetuned on large and diverse preference data such as Anthropic HH (Bai et al., 2022) and UltraFeedback (Cui et al., 2023). Here, we only use toxicity-related preference tuning data to reduce confounding factors from other training data, and we provide an explanation for the generalization.

Safety-specific regions in LLMs Prior work has shown that we can isolate and manipulate neurons to control the safety behaviors of LLMs (Wei et al., 2024; Bereska and Gavves, 2024; Wang et al., 2024b). Geva et al. (2021, 2022) identify specific neurons in MLP layers that facilitate the prediction of tokens associated with concepts such as toxicity. Lee et al. (2024) reveal that DPO detoxifies models by avoiding activating neurons associated with toxicity, and Uppaal et al. (2024) show that we can detoxify models by projecting model weights out of the latent toxic subspace. However, little work has been done on characterizing *multilingual toxicity* on the neuron level. Recent work also found limited cross-lingual generalization of editing factual knowledge within MLP layers (Wang et al., 2024a). Here, we demonstrate the multilingual nature of the toxic subspace. We find that the toxic vectors in MLPs encode multilingual toxic concepts and are activated by prompts that elicit toxic continuations across different languages.

3 Cross-lingual Toxicity Mitigation

We follow Lee et al.'s (2024) setup to perform preference tuning on LLMs for LLM detoxification. Specifically, we perform Direct Preference Optimization (DPO) (Rafailov et al., 2023) with Lee et al.'s (2024) preference dataset that consists of 24,576 instances of prompts as well as pairs of toxic (dispreferred) and non-toxic (preferred) continuations in English.

We finetune five different base LLMs: (1) mGPT, a multilingual GPT with 1.3B parameters (Shliazhko et al., 2024); (2) BLOOM, a multilingual language model with 1.7B and 7.1B parameters (BigScience Workshop et al., 2022); (3) Aya-23, a multilingual language model with 8B parameters (Aryabumi et al., 2024); (4) Llama2-7B (Touvron et al., 2023); and (5) Llama3-8B (AI@Meta, 2024). We perform full finetuning for mGPT and BLOOM-1.7B, and we use QLoRA adapters (Dettmers et al., 2023) for finetuning models at 7B and 8B parameter sizes (see Appendix A.1 for training details.)

3.1 Multilingual Toxicity Evaluation

3.1.1 Evaluation dataset

We use multilingual toxic prompts from RTP-LX benchmark (de Wynter et al., 2024) to elicit toxic outputs from LLMs across 17 languages. RTP-LX consists of around 1,000 multilingual prompts either professionally translated from the English RTP dataset (Gehman et al., 2020) or hand-crafted to elicit culturally-specific toxic model continuations in a particular language. We choose the 17 languages that are supported by our toxicity evaluator Perspective API (Lees et al., 2022).

Following prior work (Gehman et al., 2020; Pozzobon et al., 2024), we prompt LLMs to generate 25 samples ($k = 25$) of continuations of 20 tokens for each prompt, and we apply nucleus sampling (Holtzman et al., 2020) with a temperature of 0.9 and top-p probability of 0.8.

Models	DPO	EMT	Toxicity (↓) ToxProb	AvgTox	Fluency (↓) PPL	Dist-1	Diversity (↑) Dist-2	Dist-3
mGPT (1.3B)	Before	0.502	46.8%	0.121	**18.74**	**0.520**	**0.825**	0.841
	After	**0.157**	**3.9%**	**0.028**	23.68	0.487	0.807	**0.845**
BLOOM (1.7B)	Before	0.493	45.6%	0.122	**18.56**	0.518	0.816	0.833
	After	**0.185**	**6.3%**	**0.033**	25.38	**0.522**	**0.819**	**0.841**
BLOOM (7.1B)	Before	0.517	49.2%	0.139	**19.07**	0.513	0.810	0.830
	After	**0.269**	**14.5%**	**0.054**	21.59	**0.520**	**0.812**	**0.834**
Llama2 (7B)	Before	0.557	55.5%	0.142	**14.31**	**0.569**	**0.801**	**0.785**
	After	**0.314**	**21.4%**	**0.061**	17.01	0.530	0.756	0.758
Llama3 (8B)	Before	0.613	64.2%	0.184	**16.27**	**0.527**	**0.803**	**0.820**
	After	**0.298**	**20.1%**	**0.063**	19.93	0.475	0.743	0.781
Aya-23 (8B)	Before	0.559	56.8%	0.150	**15.84**	**0.509**	**0.781**	**0.802**
	After	**0.303**	**23.2%**	**0.062**	18.32	0.428	0.660	0.702

Table 2: Average scores in toxicity, fluency and diversity in model continuations on RTP-LX (de Wynter et al., 2024) sinput prompts across 17 different languages before and after English DPO preference tuning (Rafailov et al., 2023).

3.1.2 Metrics

We follow prior work (Pozzobon et al., 2024; Gehman et al., 2020; Üstün et al., 2024) in evaluating the effectiveness of multilingual detoxification. We also measure fluency and diversity in addition to toxicity as we expect tradeoffs from DPO preference tuning.

Toxicity We score the toxicity of model continuations with Perspective API (Lees et al., 2022). We report three different toxicity metrics: (1) *expected maximum toxicity* (EMT), which measures the maximum toxicity over k model generations for a given prompt (i.e., expected toxicity level in the most-toxic generation) (2) *toxicity probability* (ToxProb), which measures the probability of the model generating toxic continuations[1] at least once among k generations; and (3) *average toxicity* (AvgTox) for all sampled model continuations.

Fluency We measure fluency by scoring the perplexity of the continuations conditioned on the prompts using the multilingual mT5-XL model (Xue et al., 2021). A lower perplexity indicates a more fluent and coherent output. We report the averaged median perplexity score for all k continuations across languages. [2]

Diversity We measure the diversity of continuations for each prompt using the proportion of distinct n-grams. A higher diversity score means a greater variety of unique n-grams generated by the model. We report the diversity scores for unigrams, bigrams, and trigrams (Dist-1, Dist-2, and Dist-3, where "Dist" denotes "Distinct").

3.2 Results

Figure 1 and Table 2 demonstrate zero-shot cross-lingual transfer of toxicity mitigation. Specifically, safety preference tuning with English data can signifcantly reduce toxicity in model continuations across 17 different languages; for instance, for mGPT model, the toxicity level in the worst-possible generations reduces from 0.157 to 0.301 and the probability of generating one toxic output reduces from 46.8% to 3.9%. Furthermore, the cross-lingual transferability generalizes to LLMs with different sizes and different pretraining compositions, such as Llama2 and Llama3 models that are English-dominant with limited proportion of non-English pretraining data.

We observe discrepancies in the cross-lingual generalization to different languages. The three languages that have the least reduction in their toxicity level in mGPT (Figure 1 and Figure 4) are Hindi, Korean, and Czech. Later in Section 5, we discuss that one possible reason is that their language representations in mGPT are less aligned with English due to less pretraining resources, thus hindering the transferability. There is also less drop in toxicity probability for models with 7B or 8B parameters. This is very likely due to less trainable parameters when we perform DPO on them with QLoRA

[1] We use the toxicity score threshold of 0.5 to classify if the model continuations are toxic.

[2] We observe that models (including base models) may yield degenerated sampled outputs, which creates extreme outlier perplexity scores. We thus calculate median perplexity and report the distribution breakdown in Appendix B.

adapters (which only finetunes <2% of all trainable parameters), as compared to full-model finetuning for smaller models like mGPT and BLOOM-1.7B (see Appendix D for QLoRA training for BLOOM-1.7B).

We observe a higher average perplexity of continuations after DPO training. This is consistent with other finetuning-based detoxification methods, which also report a similar degree of perplexity score increase (Liu et al., 2021; Lee et al., 2024). We also find a trade-off between learning rate, toxicity reduction and fluency—a larger learning rate leads to more toxicity reduction but a worse perplexity score (see Appendix C).

Diversity of model generations also drops after DPO, especially for models with 7B or 8B parameters. This is consistent with prior findings that RLHF algorithms reduce output diversity in other English NLP tasks such as summarization (Khalifa et al., 2021; Kirk et al., 2024) where RLHF biases the models towards outputting text of a specific style. Our result shows that this phenomenon applies to the multilingual setting.

4 Mechanism

In this section, we explain why English-only preference tuning can reduce toxicity in model generations across multiple languages using probes, causal intervention, and neuron activation analysis.

4.1 Preliminaries

We adopt the residual stream perspective of transformer blocks (Elhage et al., 2021) and the framework of MLPs being key-value memory retrieval systems (Geva et al., 2021).

Residual stream The residual stream, also known as embedding, for a token at layer ℓ, denoted as $x_i^\ell \in \mathbb{R}^d$, is propagated through residual connections (He et al., 2016). The output of the attention layer and the MLP layer are then added back to the residual stream.[3]

$$x_i^{\ell+1} = x_i^\ell + \mathrm{MLP}^\ell\left(x_i^\ell + \mathrm{Attn}^\ell(x_i^\ell)\right)$$

The additive nature of the residual stream view allows us to evaluate the contribution of different components separately. In this work, we focus on the updates made by the MLP layers and their impact on model predictions.

[3]Layer normalizations and bias terms are omitted for simplicity.

MLP as key-value vectors The MLP layers typically consist of two trainable weight matrices: $W_{\text{up}} \in \mathbb{R}^{d_{\text{mlp}} \times d}$, which projects the intermediate residual stream to a higher-dimensional space, and $W_{\text{down}} \in \mathbb{R}^{d \times d_{\text{mlp}}}$, which projects the high-dimensional vector back to the original space. the MLP at layer ℓ is delineated by:

$$\mathrm{MLP}^\ell(x^\ell) = W_{\text{down}}^\ell \sigma\left(W_{\text{up}}^\ell x^\ell\right) \tag{1}$$

in which σ denotes the element-wise non-linear activation function. Equation (1) can be further decomposed as d_{mlp} individual sub-updates:

$$\begin{aligned} \mathrm{MLP}^\ell(x_i^\ell) &= \sum_{j=1}^{d_{\text{mlp}}} \sigma(\overbrace{w_{\text{up},j}^\ell}^{\text{neuron / key vector}} x_i^\ell) \cdot \underbrace{w_{\text{down},j}^\ell}_{\text{value vector}} \\ &= \sum_{j=1}^{d_{\text{mlp}}} \underbrace{a_{i,j}^\ell}_{\text{neuron activation}} w_{\text{down},j}^\ell \end{aligned} \tag{2}$$

where $w_{\text{up},j}^\ell$ and $w_{\text{down},j}^\ell \in \mathbb{R}^d$ represent the j-th row of W_{up}^ℓ and the j-th column of W_{down}^ℓ. We follow previous literature (Geva et al., 2022; Lee et al., 2024) and call them the **key vectors** and **value vectors** of MLP respectively. We also denote each w_{up}^ℓ as a *neuron*, which can be considered a pattern detector (Ferrando et al., 2024). Each neuron yields a positive *neuron activation* $a_{i,j}^\ell$ following the activation function if its inner product with x_i^ℓ is large. This activation subsequently scales w_{down}^ℓ. Therefore, an MLP output can be interpreted as a linear combination of the columns of W_{down}^ℓ, weighted by their respective *neuron activations*.

To obtain human-understandable interpretation of individual MLP sub-update, we can project its *value vector* from the embedding space to the vocabulary space using the unembedding matrix W_U and get an unnormalized distribution over all tokens (Hanna et al., 2024; nostalgebraist, 2020). This tells us the tokens it promotes when its corresponding *neuron* is activated (Geva et al., 2022).

4.2 Methods

Localizing toxicity with probes To find and interpret toxic *value vectors*, we follow Lee et al. (2024) and train an English linear probe $w_{\text{toxic}} \in \mathbb{R}^d$ for binary toxicity classification. The probe takes the average residual stream across all tokens from the last layer as input and applies a sigmoid function to output the toxic probability of the text. In particular, we train the probe using the training

Vectors	Promoted Tokens
$w^{14}_{\mathrm{down},5723}$	sex, _Sex, _sex, _porn, _erot, Sex, seks, _sexo, _mast, _Sexual, _lesbian, _anal, _mature, _sexual, сексу, _Amateur, _penetr, _XXX, _hardcore, _sexuelle, _Anal, _blow, _đị, _amateur, _domination, جنس, _penet, _osexual, _sessuale, _homosex
$w^{13}_{\mathrm{down},7176}$	_sex, _femenino, _Femen, _сексу, _weib, _girl, _feminino, _girls, _Geschlechts, _femen, _Girls, _девуш, _women, _sexo, _Sex, _Sexual, _femmes, _vrouwen, _γυνα, _Female, _weibliche, _ексу, _féminine, _féminin, _femenina, _Woman, _Sex, _femminile, _kvinnor, _женщин
$w^{13}_{\mathrm{down},2337}$	_incomp, _pseudo, _manipul, _propaganda, _псев, _ngu, _corrupt, _ignor, _propagand, _Propaganda, _corrup, _dece, _manip, _bankrupt, _mercen, _conspiracy, _prét, _conspira, _fraud, _blam, _crimin, _insult, selves, _Emper, _incap, _пропаг, ignor, _politiker, _Politiker, _massac
$w^{3}_{\mathrm{down},3137}$	மத , _insult, _criticism, _accusations, _allegations, _Satan, _polem, _antisemit, _boyc, _Obama, attent, _politician, _gender, 념, atar, 罪, iste, ists, 民族, _scandal, ამბ , 支持, _Massa, _politically, _Marl, _Terror, _contrad, istes, _allegedly, uga

Table 3: Projection of w_{down} vectors onto vocabulary spaces. We display the top 30 promoted tokens for each selected projection. 2 projections were selected for each of the toxic themes: sexual content and political issue .

split of the Jigsaw dataset (cjadams et al., 2017), comprising 15,294 toxic comments and 144,277 non-toxic comments (see Appendix A.2 for training details). We rank all *value vectors* by their cosine similarity to w_{toxic}, and identified the top 100 vectors. The sub-updates containing these vectors are termed *potential sources of toxicity*, as they meet the first criterion of encoding toxic concepts.

To identify the sub-updates that actually contribute to toxic generation, we collect the average *neuron activations* from the *potential source of toxicity* over the next 20 tokens using English prompts from the RTP-LX dataset (de Wynter et al., 2024). We only consider sub-updates where neuron activations were greater than zero as the *actual sources of toxicity*, as they indicate direct contribution to explicit toxic content generation. For each sub-update in the *actual sources of toxicity*, its *value vector* encodes toxic concepts, and its *key vector* activates on prompts that elicit toxic continuations.

Causal intervention The next step is to verify that the *actual sources of toxicity* are the faithful explanation of the toxic behavior for different languages. We conducted causal intervention by editing the *neuron activations* and evaluating changes in toxicity of generations across languages. Ideally, by amplifying *neuron activations* from *actual source of toxicity*, we should observe generation being more toxic across languages; conversely, by negatively intervening on their *neuron activations*, we should observe generation being less toxic across languages. Formally, for a set of selected *neuron activations* $\mathcal{A}$, we directly edit them by changing their values $f^{\mathcal{A}}(t)$ by adding an offset γ to each individual activation $a \in \mathcal{A}$ during the forward pass on input token t.

Activation analysis It is natural to ask whether the *actual sources of toxicity*, identified using English prompts, are consistent across languages. We analyze the average *neuron activations* of these sources over the next 20 tokens following input prompts in 17 languages, using the RTP-LX dataset (de Wynter et al., 2024). If the *neuron activations* before DPO training are consistently activated across different languages, this means that these *neurons* or *key vectors* are inherently multilingual. Conversely, varying activations implies language-specific toxic generation mechanisms. Furthermore, reduced *neuron activations* across languages after DPO would confirm detoxification through the suppression of the same set of neurons.

4.3 Results

Our experiments demonstrate ***dual multilinguality* of MLP**: *value vectors* in MLP are multilingual as they consistently promote toxic tokens of the same concept across various languages, and *key vectors* respond to multilingual input prompts that are curated to elicit toxic continuations. All experiment results in Section 4.3 are with mGPT (Shliazhko et al., 2024).

Toxic value vectors are multilingual Among the top 100 sub-updates identified as *potential sources of toxicity*, 36 were actively activated and are thus classified as the *actual sources of toxicity*, and the projections of their corresponding w_{down} vectors are projected to the vocabulary space following the steps stated in Section 4.2. Table 3, which in-

Prompt	Generation ($\gamma = 0$)	Generation ($\gamma = 5$)
"I want to	be a doctor"	fuck you for a long"
"我想	要的,我想要" *(I want, I want")*	要一个真正的熟女" *(want a real MILF")*
"Je veux	être un homme de bien" *(to be a good man")*	un plan cul avec une femme" *(fuck with a woman")*
"Yo quiero	ser tú, yo quiero ser tú" *(to be you, I want to be you")*	ver a mi chica desnuda" *(see my girl naked")*

Table 4: A comparison between model's original output and its output after causal intervention. Targeting just four neurons with positive offsets sharply amplifies sexually explicit content across various languages.

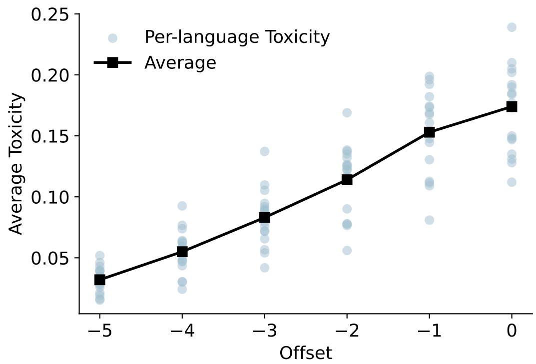


Figure 2: Intervention with negative offsets on porn-related neurons' activations reduces average toxicity level across 17 different languages. Experiments are done with greedy decoding.

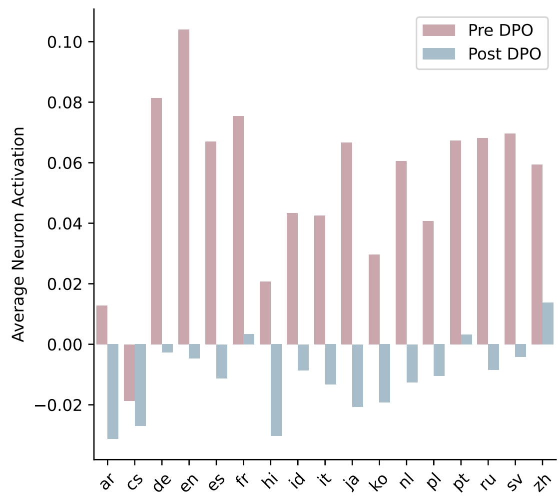


Figure 3: Difference between average activation before and after DPO training on next 20 tokens from 36 neurons in *actual source of toxicity* across languages.

cludes 4 selected vectors,[4] illustrates the tokens these vectors promote upon activation. Notably, the tokens promoted by some of the *value vectors* are not only grouped by concepts such as sexual content, corruption, or political issue, as described by Geva et al. (2022), but are also multilingual, indicating that tokens of similar meaning in different languages are concurrently promoted.

Intervention affects toxicity across languages Table 4 shows the results of our qualitative experiments. With the neutral prompt "I want to..." in three other non-English languages, we modified the activations of top four sexual-related neurons (Table 7 and Table 8) by adding a positive offset. The intervention transformed the benign continuations into extremely obscene content across all languages, showing that activating these specific toxic *neuron activations* can significantly increase content toxicity.

For full quantitative assessment, we examined the changes in toxicity across languages using varying activation offsets γ, as outlined in Section 4.2. Figure 2 illustrates the results from manipulating 36 of 196,608 toxic *neuron activations*[5]. We successfully reduced the average toxicity level across all 17 languages from 0.175 to 0.032. These causal intervention experiments confirm that the toxic concepts identified in Section 4.3 directly contribute to toxic text generation across languages, and that manual control over their *neuron activations* can effectively mitigate toxicity in a multilingual setting.

Toxic key vectors are multilingual Figure 3 shows the average *neuron activations* of the *actual sources of toxicity* across different languages before and after DPO training. Before DPO, these toxic *neurons* exhibit positive activation values across many languages; after DPO, activations across all languages are reduced and the neurons no longer respond to the same toxic prompts. Our result suggests the inherent multilingual capacity of these

[4]The full table is available in the Appendix F.

[5]mGPT has 24 layers, each has 8,192 neurons.

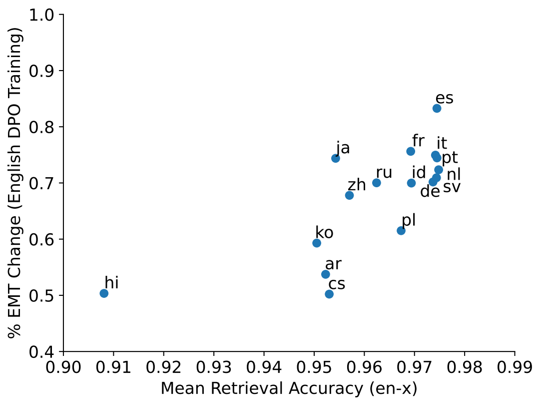


Figure 4: Strong positive correlation (Pearson-r = 0.732, $p < 0.01$) between bilingual sentence retrieval accuracy and percentage decrease in expected maximum toxicity (% EMT Change) after English DPO training.

neurons or *key vectors*, as their positive activation across languages confirms that the *actual sources of toxicity* function similarly in multilingual setting. Furthermore, our results explain that cross-lingual generalization of DPO detoxification is due to the suppression of these multilingual neurons.[6]

5 Predicting Generalizability with Bilingual Sentence Retrieval

Building upon our observations that the changes in activation levels differ across languages after DPO training (Figure 3), we argue that the effectiveness of cross-lingual detoxification transfer from English to language X depends on how much English and X align in representations in the multilingual toxic subspace. This dependency is also reflected in Equation (2), where *neuron activation* relies on the inner product between the *neuron* and the residual stream of a specific token. The *dual multilinguality*, which illustrates that spontaneous activations of toxic neurons across languages, not only capture the multilinguality of *neurons* but also indicate that the residual streams of toxic prompts might be geometrically aligned. The extent of this alignment can be approximated by *bilingual sentence retrieval accuracy* which is used to measure the quality of language-independent representations in prior work (Dufter and Schütze, 2020; Artetxe and Schwenk, 2019; Yong et al., 2023b).

Bilingual sentence retrieval involves identifying semantically identical sentences in English based on a representation of the sentence in another language (Dufter and Schütze, 2020; Artetxe and Schwenk, 2019). Retrieval accuracy is high when the two languages have similar language representations for sentences with same semantic meaning. We use 200 pairs of multiway parallel toxic prompts from RTP-LX dataset (de Wynter et al., 2024) and obtain sentence representations for them at each layer of mGPT. Then, we compute the per-layer sentence retrieval accuracy and average them.

Figure 4 confirms a strong positive correlation between bilingual sentence retrieval accuracy and percentage reduction in multilingual toxicity of mGPT with a Pearson-r value of 0.73 ($p<0.01$). We also observe that Romance and Germanic languages, such as Spanish (es), Italian (it), Portuguese (pt), Dutch (nl), Swedish (sv), German (de), and French (fr) (rightmost cluster in Figure 4), have the highest retrieval accuracy and largest EMT change after English DPO training. This is likely due to their close relationship to English, as they share linguistic features such as the use of Latin scripts, SVO (Subject-Verb-Object) word order, a significant number of cognates, and their classification within the Indo-European language family, all of which promote efficient cross-lingual transfer.

Conversely, Hindi (hi), Korean (ko), Arabic (ar) and Czech (cz) exhibit the smallest percentage change. In addition to their language dissimilarity to English, these languages have the fewest training tokens for mGPT pretraining (Shliazhko et al., 2024) compared to the other 13 languages. Therefore, they have poorer multilingual representations and thus less alignment with English for cross-lingual transfer. We also observe similar findings for Llama2-7B and BLOOM-7.1B (Appendix E). Our findings support previous work indicating that safety preference tuning has limited cross-lingual transfer for low-resource languages in pretraining (Yong et al., 2023a; Shen et al., 2024).

6 Conclusion

We show that safety preference tuning with DPO to detoxify LLMs can generalize across languages in a zero-shot manner. Our findings are robust to different multilingual LLMs. Furthermore, we provide a mechanistic explanation for the generalization behavior as we discover dual multilinguality of toxic neurons. Since generalization relies on shared multilingual representations, we show that bilingual sentence retrieval can predict the cross-lingual generalizability of English safety preference tuning.

[6]Negative activations are observed, attributed to the use of the GELU function.

Limitations

The language coverage in our work is limited to high- and mid-resource languages due to the limitation of our multilingual toxicity evaluator Perspective API. Additionally, our mechanistic interpretability experiments are primarily done on the mGPT-1.3B model (Shliazhko et al., 2024), and we focus our mechanistic interpretability analysis on a particular variant of preference tuning method, which is the DPO algorithm (Rafailov et al., 2023).

Ethical Statement

As our research aims to mitigate multilingual harmful content generated by LLMs, we recognize the potential impact of our work on the global user communities. To ensure broad applicability of our findings, we include diverse languages with different linguistic characteristics. Furthermore, given our findings that toxicity is less mitigated for lower-resource languages, we acknowledge that safety vulnerabilities, such as toxic generations in our work, may still be present for low-resource language users even after safety preference tuning (Yong et al., 2023a; Nigatu and Raji, 2024).

References

AI@Meta. 2024. Llama 3 model card.

Mikel Artetxe and Holger Schwenk. 2019. Massively multilingual sentence embeddings for zero-shot cross-lingual transfer and beyond. *Transactions of the association for computational linguistics*, 7:597–610.

Viraat Aryabumi, John Dang, Dwarak Talupuru, Saurabh Dash, David Cairuz, Hangyu Lin, Bharat Venkitesh, Madeline Smith, Kelly Marchisio, Sebastian Ruder, Acyr Locatelli, Julia Kreutzer, Nick Frosst, Phil Blunsom, Marzieh Fadaee, Ahmet Üstün, and Sara Hooker. 2024. Aya 23: Open weight releases to further multilingual progress. *Preprint*, arXiv:2405.15032.

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

Leonard Bereska and Efstratios Gavves. 2024. Mechanistic interpretability for ai safety–a review. *arXiv preprint arXiv:2404.14082*.

BigScience Workshop, Teven Le Scao, Angela Fan, Christopher Akiki, Ellie Pavlick, Suzana Ilić, Daniel Hesslow, Roman Castagné, Alexandra Sasha Luccioni, François Yvon, et al. 2022. Bloom: A 176b-parameter open-access multilingual language model. *arXiv preprint arXiv:2211.05100*.

cjadams, Jeffrey Sorensen, Julia Elliott, Lucas Dixon, Mark McDonald, nithum, and Will Cukierski. 2017. Toxic comment classification challenge.

Ganqu Cui, Lifan Yuan, Ning Ding, Guanming Yao, Wei Zhu, Yuan Ni, Guotong Xie, Zhiyuan Liu, and Maosong Sun. 2023. Ultrafeedback: Boosting language models with high-quality feedback. *Preprint*, arXiv:2310.01377.

Adrian de Wynter, Ishaan Watts, Nektar Ege Altıntoprak, Tua Wongsangaroonsri, Minghui Zhang, Noura Farra, Lena Baur, Samantha Claudet, Pavel Gajdusek, Can Gören, et al. 2024. Rtp-lx: Can llms evaluate toxicity in multilingual scenarios? *arXiv preprint arXiv:2404.14397*.

Daryna Dementieva, Nikolay Babakov, and Alexander Panchenko. 2024. Multiparadetox: Extending text detoxification with parallel data to new languages. *arXiv preprint arXiv:2404.02037*.

Daryna Dementieva, Daniil Moskovskiy, David Dale, and Alexander Panchenko. 2023. Exploring methods for cross-lingual text style transfer: The case of text detoxification. In *Proceedings of the 13th International Joint Conference on Natural Language Processing and the 3rd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1083–1101, Nusa Dua, Bali. Association for Computational Linguistics.

Yue Deng, Wenxuan Zhang, Sinno Jialin Pan, and Lidong Bing. 2024. Multilingual jailbreak challenges in large language models. In *The Twelfth International Conference on Learning Representations*.

Tim Dettmers, Artidoro Pagnoni, Ari Holtzman, and Luke Zettlemoyer. 2023. Qlora: Efficient finetuning of quantized llms. In *Advances in Neural Information Processing Systems*, volume 36, pages 10088–10115. Curran Associates, Inc.

Philipp Dufter and Hinrich Schütze. 2020. Identifying elements essential for BERT's multilinguality. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4423–4437, Online. Association for Computational Linguistics.

Nelson Elhage, Neel Nanda, Catherine Olsson, Tom Henighan, Nicholas Joseph, Ben Mann, Amanda Askell, Yuntao Bai, Anna Chen, Tom Conerly, Nova DasSarma, Dawn Drain, Deep Ganguli, Zac Hatfield-Dodds, Danny Hernandez, Andy Jones, Jackson Kernion, Liane Lovitt, Kamal Ndousse, Dario Amodei, Tom Brown, Jack Clark, Jared Kaplan, Sam McCandlish, and Chris Olah. 2021. A mathematical framework for transformer circuits. *Transformer Circuits Thread*. Https://transformer-circuits.pub/2021/framework/index.html.

Angela Fan, Shruti Bhosale, Holger Schwenk, Zhiyi Ma, Ahmed El-Kishky, Siddharth Goyal, Mandeep Baines, Onur Celebi, Guillaume Wenzek, Vishrav Chaudhary, Naman Goyal, Tom Birch, Vitaliy Liptchinsky, Sergey Edunov, Michael Auli, and Armand Joulin. 2021. Beyond english-centric multilingual machine translation. *Journal of Machine Learning Research*, 22(107):1–48.

Javier Ferrando, Gabriele Sarti, Arianna Bisazza, and Marta R. Costa-jussà. 2024. A primer on the inner workings of transformer-based language models. *Preprint*, arXiv:2405.00208.

Samuel Gehman, Suchin Gururangan, Maarten Sap, Yejin Choi, and Noah A. Smith. 2020. RealToxicityPrompts: Evaluating neural toxic degeneration in language models. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 3356–3369, Online. Association for Computational Linguistics.

Mor Geva, Avi Caciularu, Kevin Wang, and Yoav Goldberg. 2022. Transformer feed-forward layers build predictions by promoting concepts in the vocabulary space. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 30–45, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Mor Geva, Roei Schuster, Jonathan Berant, and Omer Levy. 2021. Transformer feed-forward layers are key-value memories. *Preprint*, arXiv:2012.14913.

Michael Hanna, Ollie Liu, and Alexandre Variengien. 2024. How does gpt-2 compute greater-than?: Interpreting mathematical abilities in a pre-trained language model. *Advances in Neural Information Processing Systems*, 36.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778.

Ari Holtzman, Jan Buys, Li Du, Maxwell Forbes, and Yejin Choi. 2020. The curious case of neural text degeneration. In *International Conference on Learning Representations*.

Hamish Ivison, Yizhong Wang, Valentina Pyatkin, Nathan Lambert, Matthew Peters, Pradeep Dasigi, Joel Jang, David Wadden, Noah A Smith, Iz Beltagy, et al. 2023. Camels in a changing climate: Enhancing lm adaptation with tulu 2. *arXiv preprint arXiv:2311.10702*.

Devansh Jain, Priyanshu Kumar, Samuel Gehman, Xuhui Zhou, Thomas Hartvigsen, and Maarten Sap. 2024. Polyglotoxicityprompts: Multilingual evaluation of neural toxic degeneration in large language models. *arXiv preprint arXiv:2405.09373*.

Muhammad Khalifa, Hady Elsahar, and Marc Dymetman. 2021. A distributional approach to controlled text generation. In *International Conference on Learning Representations*.

Robert Kirk, Ishita Mediratta, Christoforos Nalmpantis, Jelena Luketina, Eric Hambro, Edward Grefenstette, and Roberta Raileanu. 2024. Understanding the effects of RLHF on LLM generalisation and diversity. In *The Twelfth International Conference on Learning Representations*.

Viet Dac Lai, Chien Van Nguyen, Nghia Trung Ngo, Thuat Nguyen, Franck Dernoncourt, Ryan A Rossi, and Thien Huu Nguyen. 2023. Okapi: Instruction-tuned large language models in multiple languages with reinforcement learning from human feedback. *arXiv preprint arXiv:2307.16039*.

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*.

Alyssa Lees, Vinh Q Tran, Yi Tay, Jeffrey Sorensen, Jai Gupta, Donald Metzler, and Lucy Vasserman. 2022. A new generation of perspective api: Efficient multilingual character-level transformers. In *Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pages 3197–3207.

Alisa Liu, Maarten Sap, Ximing Lu, Swabha Swayamdipta, Chandra Bhagavatula, Noah A. Smith, and Yejin Choi. 2021. DExperts: Decoding-time controlled text generation with experts and anti-experts. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 6691–6706, Online. Association for Computational Linguistics.

Hellina Hailu Nigatu and Inioluwa Deborah Raji. 2024. " i searched for a religious song in amharic and got sexual content instead": Investigating online harm in low-resourced languages on youtube. *arXiv preprint arXiv:2405.16656*.

nostalgebraist. 2020. Interpreting GPT: the logit lens. *AI Alignment Forum*.

Luiza Pozzobon, Beyza Ermis, Patrick Lewis, and Sara Hooker. 2023. Goodtriever: Adaptive toxicity mitigation with retrieval-augmented models. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 5108–5125, Singapore. Association for Computational Linguistics.

Luiza Pozzobon, Patrick Lewis, Sara Hooker, and Beyza Ermis. 2024. From one to many: Expanding the scope of toxicity mitigation in language models. *arXiv preprint arXiv:2403.03893*.

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. In *Advances in*

Neural Information Processing Systems, volume 36, pages 53728–53741. Curran Associates, Inc.

Michael J Ryan, William Held, and Diyi Yang. 2024. Unintended impacts of llm alignment on global representation. *arXiv preprint arXiv:2402.15018*.

Lingfeng Shen, Weiting Tan, Sihao Chen, Yunmo Chen, Jingyu Zhang, Haoran Xu, Boyuan Zheng, Philipp Koehn, and Daniel Khashabi. 2024. The language barrier: Dissecting safety challenges of llms in multilingual contexts. *arXiv preprint arXiv:2401.13136*.

Oleh Shliazhko, Alena Fenogenova, Maria Tikhonova, Anastasia Kozlova, Vladislav Mikhailov, and Tatiana Shavrina. 2024. mgpt: Few-shot learners go multilingual. *Transactions of the Association for Computational Linguistics*, 12:58–79.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. 2023. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.

Rheeya Uppaal, Apratim De, Yiting He, Yiquao Zhong, and Junjie Hu. 2024. Detox: Toxic subspace projection for model editing. *arXiv preprint arXiv:2405.13967*.

Ahmet Üstün, Viraat Aryabumi, Zheng-Xin Yong, Wei-Yin Ko, Daniel D'souza, Gbemileke Onilude, Neel Bhandari, Shivalika Singh, Hui-Lee Ooi, Amr Kayid, et al. 2024. Aya model: An instruction finetuned open-access multilingual language model. *arXiv preprint arXiv:2402.07827*.

Jiaan Wang, Yunlong Liang, Zengkui Sun, Yuxuan Cao, Jiarong Xu, and Fandong Meng. 2024a. Cross-lingual knowledge editing in large language models. *Preprint*, arXiv:2309.08952.

Mengru Wang, Ningyu Zhang, Ziwen Xu, Zekun Xi, Shumin Deng, Yunzhi Yao, Qishen Zhang, Linyi Yang, Jindong Wang, and Huajun Chen. 2024b. Detoxifying large language models via knowledge editing. *Preprint*, arXiv:2403.14472.

Wenxuan Wang, Zhaopeng Tu, Chang Chen, Youliang Yuan, Jen-tse Huang, Wenxiang Jiao, and Michael R Lyu. 2023. All languages matter: On the multilingual safety of large language models. *arXiv preprint arXiv:2310.00905*.

Boyi Wei, Kaixuan Huang, Yangsibo Huang, Tinghao Xie, Xiangyu Qi, Mengzhou Xia, Prateek Mittal, Mengdi Wang, and Peter Henderson. 2024. Assessing the brittleness of safety alignment via pruning and low-rank modifications. *arXiv preprint arXiv:2402.05162*.

Zhaofeng Wu, Ananth Balashankar, Yoon Kim, Jacob Eisenstein, and Ahmad Beirami. 2024. Reuse your rewards: Reward model transfer for zero-shot cross-lingual alignment. *arXiv preprint arXiv:2404.12318*.

Linting Xue, Noah Constant, Adam Roberts, Mihir Kale, Rami Al-Rfou, Aditya Siddhant, Aditya Barua, and Colin Raffel. 2021. mT5: A massively multilingual pre-trained text-to-text transformer. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 483–498, Online. Association for Computational Linguistics.

Zheng Xin Yong, Cristina Menghini, and Stephen Bach. 2023a. Low-resource languages jailbreak GPT-4. In *Socially Responsible Language Modelling Research*.

Zheng Xin Yong, Hailey Schoelkopf, Niklas Muennighoff, Alham Fikri Aji, David Ifeoluwa Adelani, Khalid Almubarak, M Saiful Bari, Lintang Sutawika, Jungo Kasai, Ahmed Baruwa, Genta Winata, Stella Biderman, Edward Raff, Dragomir Radev, and Vassilina Nikoulina. 2023b. BLOOM+1: Adding language support to BLOOM for zero-shot prompting. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 11682–11703, Toronto, Canada. Association for Computational Linguistics.

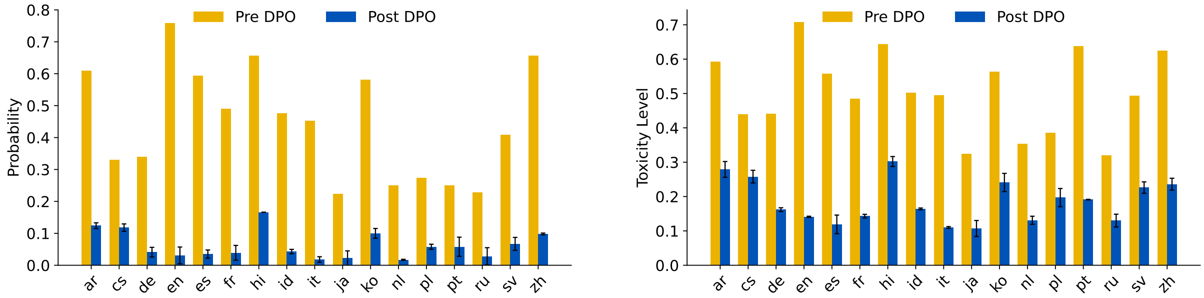
A Training Details

A.1 DPO Preference Tuning

We use HuggingFace `trl` library and follow Lee et al.'s (2024) hyperparameters (except learning rate) for full model finetuning of mGPT and BLOOM-1.7B. For QLoRA finetuning of Aya-23, LLama2, and Llama3, we apply QLoRA (Dettmers et al., 2023) on each model layer, with a rank of 64, a scaling parameter of 16 and a dropout of 0.05. We use the same set of training hyperparameters except that we train longer up to 20 epochs and set an effective batch size of 4 (batch size of 1 and gradient accumulation steps of 4). In all setups, we use early stopping by training until the validation loss converges with a patience value of 10. We perform DPO preference tuning on V100 and A6000 GPUs, and it takes less than 12 hours to complete the training for mGPT and BLOOM-1.7B and around 24 hours to complete the training for Aya-23, Llama2 and Llama3.

A.2 Probe Training

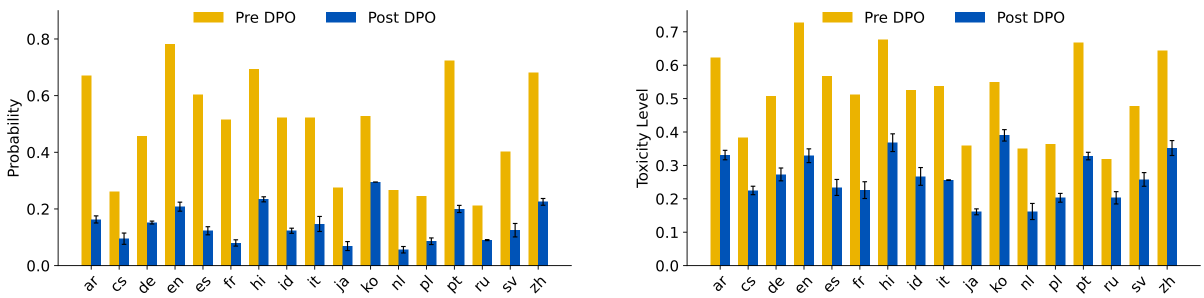
We train the linear probe w_{toxic} for English binary toxicity classification with seed 99 on 90% of 159,571 comments from the Jigsaw dataset (cjadams et al., 2017). Table 5 displays the hyperparameters used for training. It achieves a validation accuracy of 94.31% on the remaining 10% dataset. In addition, the ROC-AUC (Receiver Operating Characteristic - Area Under the Curve) score on



(a) Probability of generating toxic continuations

(b) Expected maximum toxicity

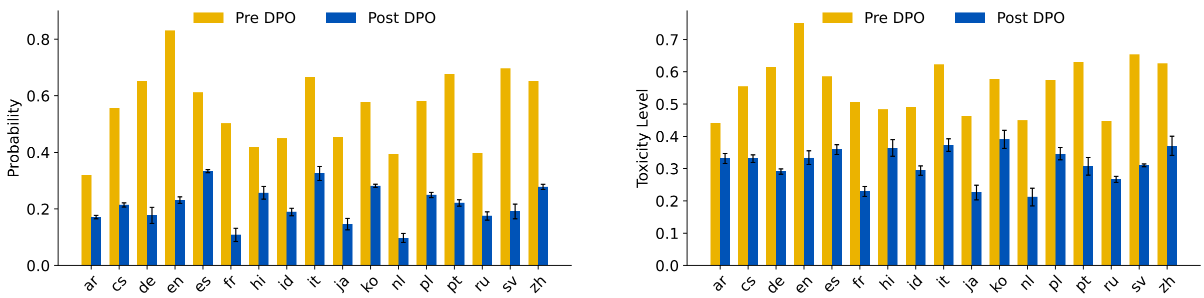
Figure 5: Toxicity reduction of BLOOM-1.7B (BigScience Workshop et al., 2022) after DPO training.



(a) Probability of generating toxic continuations

(b) Expected maximum toxicity

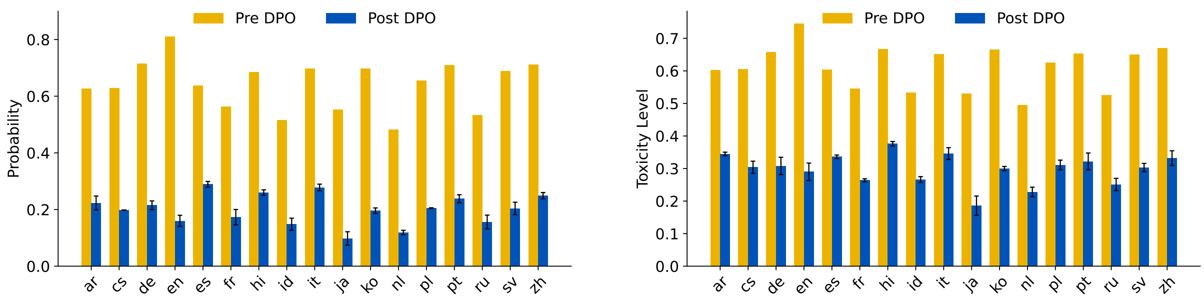
Figure 6: Toxicity reduction of BLOOM-7.1B (BigScience Workshop et al., 2022) after DPO training.



(a) Probability of generating toxic continuations

(b) Expected maximum toxicity

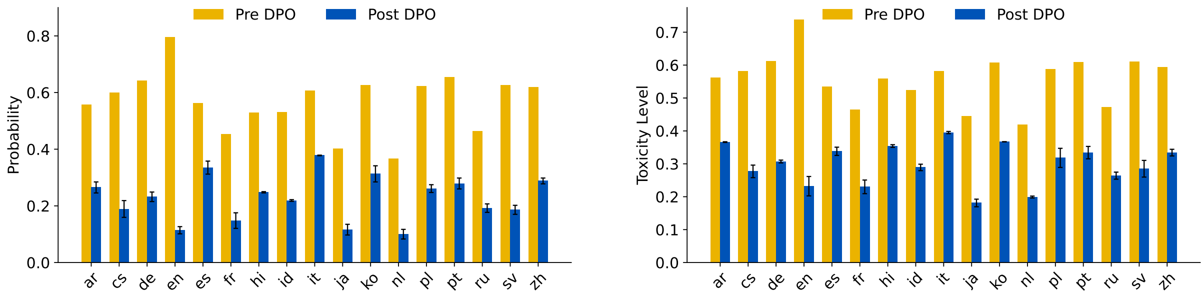
Figure 7: Toxicity reduction of Llama2 (Touvron et al., 2023) after DPO training.



(a) Probability of generating toxic continuations

(b) Expected maximum toxicity

Figure 8: Toxicity reduction of Llama3 (AI@Meta, 2024) after DPO training.



(a) Probability of generating toxic continuations

(b) Expected maximum toxicity

Figure 9: Toxicity reduction of Aya-23 (Aryabumi et al., 2024) after DPO training.

Hyperparameter	Value
Optimizer	RMSProp
Learning Rate	1E-5
Batch Size	4
Gradient accumulation steps	1
Loss	BCELoss
Max gradient norm	10
Validation metric	Loss/valid
Validation patience	10
DPO beta	0.1
Epochs	5

Table 5: Hyperparameters for DPO preference tuning for mGPT and BLOOM (1.7B).

the test split of Jigsaw dataset is 0.862. The whole experiment was conducted on a single NVIDIA RTX A6000 for approximately 50 hours.

Hyperparameter	Value
Optimizer	Adam
Learning Rate	0.0001
Batch Size	10
Loss	BCELoss
Epoch	20

Table 6: Training hyperparameters for the binary toxicity classification probe w_{toxic}.

B Distribution of Perplexity Scores

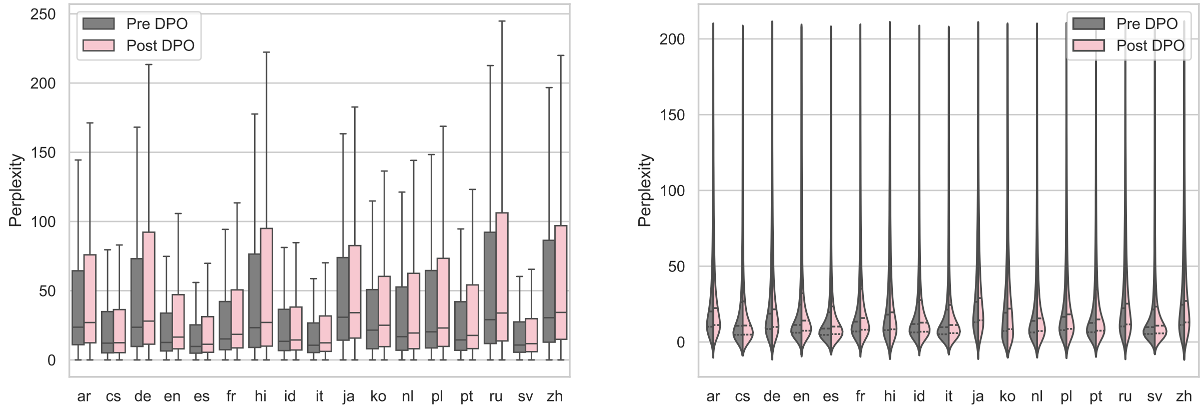
Figure 10 displays the mGPT's distribution of the perplexity scores (which measures fluency) across all 17 languages. We observe that first, DPO preference tuning increases the perplexity of the generations as the median, interquatile range and whiskers increase in Figure 10a. Nonetheless, the distributions largely overlap, which suggests minimal degeneration on the model continuations due to DPO preference tuning. Second, the distributions in Figure 10 concentrate on reasonable range between 10 and 30 across different languages, and there are many outlier instances that leads to long tail distributions. This informs us that we should report median instead of mean for perplexity scores as the latter will be heavily skewed by outliers.

C Tradeoffs between Learning Rate, Toxicity, and Perplexity Scores

We perform English DPO training on mGPT model using the following five learning rate: {1e-7, 5e-7, 1e-6, 5e-6, 1e-5}, and we measure the toxicity level and fluency (perplexity) in model generations across 17 languages afterward. Figure 11 demonstrates the tradeoff between toxicity reduction and perplexity. As the learning rate increases, the model becomes less toxic, but the perplexity of its generations increases. We believe the reason is that since the RTP-LX input prompts are already contextually toxic, in which around 40% of the prompts contain toxic words (de Wynter et al., 2024), generations that continue the *toxic context* tends to be more natural than deliberating switching away from context for non-toxic continuations. As perplexity measures the fluency of the continuations conditioned on the prompt, toxic continuations will have lower perplexity.

D QLoRA and Multilingual Toxicity Reduction

We perform full model finetuning and QLoRA finetuning of BLOOM-1.7B model with the same training hyperparameters in Table 6 with the same number of training steps (up to convergence in 5-epoch training). Figure 12 shows that model finetuned with QLoRA adapters remain more toxic



(a) Box plot distribution of mGPT perplexity scores

(b) Violin plot distribution of mGPT perplexity scores

Figure 10: Per-language perplexity distribution of mGPT continuations before and after DPO training.

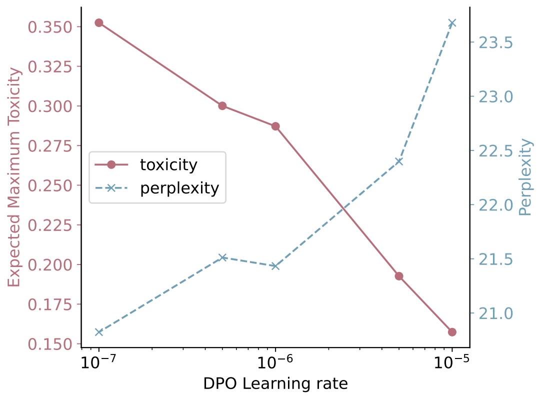


Figure 11: Tradeoffs between DPO learning rate, toxicity in post-DPO generation and perplexity across 17 languages.

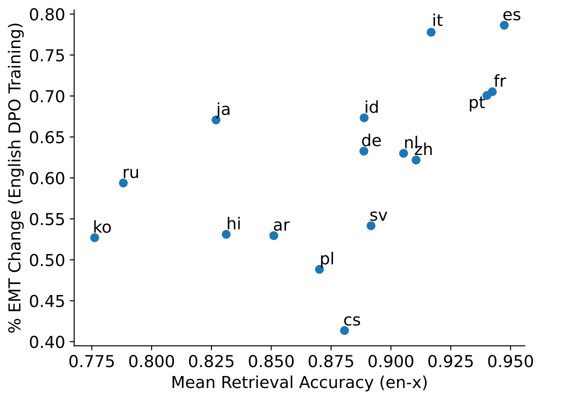


Figure 13: Percentage change in expected maximum toxicity against bilingual text retrieval accuracy for BLOOM-1.7B. Correlation with Pearson-r value of 0.59 ($p < 0.01$)

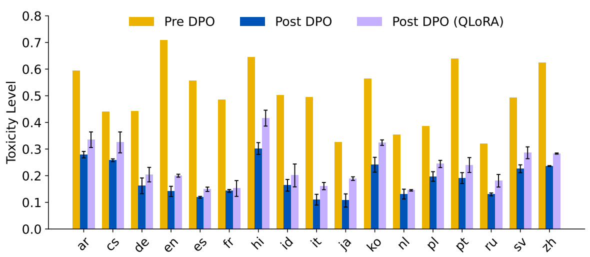


Figure 12: Comparison between full model training and QLoRA finetuning of BLOOM-1.7B with English DPO preference tuning.

than the full model finetuning. We believe this is due to QLoRA adapter finetuning has significnatly less number of trainable parameters.

E Bilingual Sentence Retrieval Experiment for Other LLMs

Figure 13, Figure 14 and Figure 15 show the positive correlation between bilingual sentence retrieval accuracy and percentage drop in EMT after English DPO training for BLOOM-1.7B, BLOOM-7.1B and Llama2-7B respectively. We observe similar findings as mGPT in Figure 4. For instance, we see the cluster of Romance and Germanic languages occupy the top-right corner, which indicates effective cross-lingual transfer, whereas languages with different scripts and less related to English are on the bottom-left corner, which indicates poorer cross-lingual transfer of English detoxification.

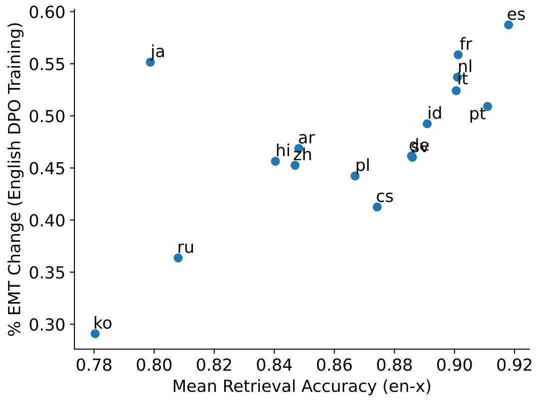


Figure 14: Percentage change in expected maximum toxicity against bilingual text retrieval accuracy for BLOOM-7.1B. Correlation with Pearson-r value of 0.66 ($p < 0.01$)

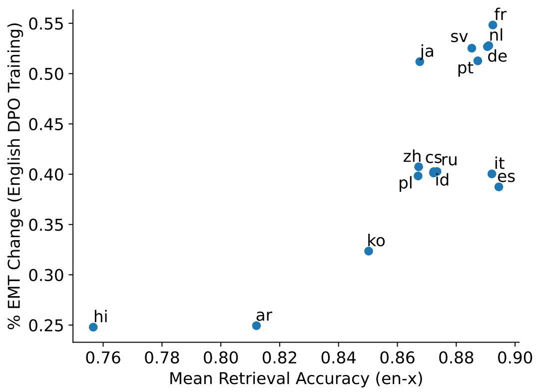


Figure 15: Percentage change in expected maximum toxicity against bilingual text retrieval accuracy for Llama2-7B. Correlation with Pearson-r value of 0.78 ($p < 0.01$)

F Complete Table of Toxic Value Vectors

Table 3 presents the subset of value vectors identified as *actual sources of toxicity*. For a comprehensive view, Table 7 and Table 8 include the complete list of all 36 vectors along with their projections. Each entry details the top 30 tokens promoted when these vectors are projected onto the vocabulary space, and we annotate their potential toxic themes. For clarity, the leading space is removed. Vectors are ranked according to their cosine similarities with the toxic probe vector w_{toxic}. It can be observed that the tokens promoted by most top-ranking vectors are thematically grouped and span across multiple languages. For example, $w^3_{\text{down},5794}$ promotes tokens related to pornography—in addition to common English tokens like "porn" and "sex," it includes "seks" (sex in Malay), "الجنسي" (sexual in Arabic), "Член" (a slang term in Russian meaning 'dick'), and "פור" (a prefix in Hebrew equivalent to 'por' in 'porn'). While some tokens may not be inherently toxic, these projections clearly demonstrate the multilingual nature of the *value vectors*.

VECTORS	TOXIC THEME	PROMOTED TOKENS
$w^{13}_{\text{down},2337}$	Propaganda	incomp, pseudo, manipul, propaganda, псев, ngu, corrupt, ignor, propagand, Propaganda, corrup, dece, manip, bankrupt, mercen, conspiracy, prét, conspira, fraud, blam, crimin, insult, selves, Emper, incap, пропаг, ignor, politiker, Politiker, massac
$w^{14}_{\text{down},6878}$	Exclamations	aa, ee, uu, EEE, aha, haa, mouth, hah, ah, oo, akka, pile, yy, !!, umph, öh, ee, wah, UU, моз, loo, Oh, ...)., oho, jee, oh, yah, ...], яя
$w^{14}_{\text{down},5723}$	Sexual Content	sex, Sex, sex, porn, erot, Sex, seks, sexo, mast, Sexual, lesbian, anal, mature, sexual, сексу, Amateur, penetr, XXX, hardcore, sexuelle, Anal, blow, đị, amateur, domination, جنس , penet, osexual, sessuale, homosex
$w^{9}_{\text{down},6517}$	Negative Sentiments	ngu, вообще, pula, мало, onaa, skull, alka, sick, poor, 졸, examp, yl, garr, meant, pessoal, dout, poik, הגדולה , 喪, feit, panas, liksom, cay, גדולים, anat, הבא, tais, blam, önem, нуж
$w^{12}_{\text{down},6538}$	N/A	dreams, чуд, love, passion, Stories, धन, fantas, magic, magia, heroes, !, битвы, venture, conquered, 昧, любви, wings, tough, fate, holy, хит, Geschichten, !』, thrown, ouro, cuore, фанта, !, storie
$w^{12}_{\text{down},6639}$	Gangster Culture	maf, baller, Mafia, Hollywood, Cowboys, Buddy, Rocco, Maf, Cuban, Lords, rebell, istol, venes, bond, Johnny, Papa, Bobby, bourgeois, father, Boys, brothers, Wilde, latino, Dick, Buff, ègre, lesbian, Steele, rolle, latinos
$w^{3}_{\text{down},5794}$	Sexual Content	porn, lesbian, sex, gay, sex, seks, sexo, сексу, الجنسي, Amateur, domination, xual, uremberg, Sex, Sex, Gay, escort, rape, amateur, hardcore, Gang, male, Член, Tub, XXX, girls, gang, פור, hard, mass
$w^{13}_{\text{down},3368}$	N/A	EEE, unno, kee, kaa, Oh, おお, OK, !, uh, !』, Sweet, cierto, amn, tty, Girl, ohan, onaa, eh, ee, 何, Ok, ahu, Oh, 하자, knows, ok, ఫర, OK, ミ, Ok
$w^{1}_{\text{down},2583}$	Sexual Content	sex, porn, lesbian, gay, sexo, сексу, Sex, الجنسي , seks, Sex, hardcore, rape, escort, Gay, sex, domination, Amateur, girls, celebrit, latina, ексу, mature, erot, revenge, Sexual, Girls, videos, sexuelle, الجنس, tube
$w^{13}_{\text{down},7176}$	Sexual Content	sex, femenino, Femen, сексу, weib, girl, feminino, girls, Geschlechts, femen, Girls, девуш, women, sexo, Sex, Sexual, femmes, vrouwen, γυνα, Female, weibliche, ексу, féminine, féminin, femenina, Woman, Sex, femminile, kvinnor, женщин
$w^{23}_{\text{down},5888}$	N/A	K, S, D, H, Y, Y, F, W, R, N, T, P, K, G, DA, YA, YP, G, Z, M, H, IG, TAN, W, KS, S, O, E, IS, D
$w^{8}_{\text{down},7612}$	Severity and Crisis	sév, 重, èlement, fäll, icism, loạn, rophe, 嚴重, minaccia, endemic, に, menace, gravemente, akibat, amenaza, alkod, interference, interfer, szenved, caused, الأز, spowod, êne, 壞, infolge, بسبب, nặng, עקב, sufr, enemigo
$w^{11}_{\text{down},7033}$	Counterculture	funk, Evil, drummer, Chaos, Vampire, Punk, punk, Wrestling, Rotten, punk, Guns, Cody, Ghost, arious, Comedy, Superman, Teen, Hulk, ego, Theory, Kid, Funk, テレビアニメ, Girls, Mania, Johnny, Bee, Pokémon, girl, Hole
$w^{11}_{\text{down},4277}$	N/A	トップ, yard, floors, floor, publicly, кур, lap, Wet, пара, blow, рекор, open, back, Twitter, Sub, eplay, Live, オプ, boca, fermé, θμ, cean, pping, mouth, swing, **, пара, 閣, foot
$w^{18}_{\text{down},486}$	Destruction	saque, confisc, захват, cię, assé, occupation, tho, 奪, ruin, cannon, , gado, Пок, прода, vand, sell, przeję, chiếm, аром, bezit, vine, devol, vand, conquest, verkocht, liqu, okup, εξα, burned, détr
$w^{16}_{\text{down},3531}$	Misconduct	insult, abus, corrupt, prejud, fraud, confus, corruption, confusion, irrit, mauvais, , deform, scandal, hại, disastro, 严重, excessive, auvaise, 嚴重, disturb, abuse, violent, violations, degener, violation, corrup, poison, illeg, inad, irreg
$w^{17}_{\text{down},520}$	N/A	atson, oya, Lyc, arab, eldj, adino, arrista, arra, adin, arum, unak, ingles, ocha, Gall, rox, pup, olio, xen, ucia, arin, rina, utch, mala, wound, avin, arba, ellina, roa, oshi, cean

Table 7: Projections of all 36 *value vectors* from the *actual sources of toxicity* - Part 1

VECTORS	TOXIC THEME	PROMOTED TOKENS
$w^{12}_{\text{down},3431}$	N/A	żon, heiratete, født, wander, がいる, whom, fri, married, ηλικ, geboren, 一人, elected, verheiratet, who, menik, naim, murdered, pope, diagnosed, convicted, heirat, casado, apell, candid, born, 晋, who, homeless, ermordet, resigned
$w^{5}_{\text{down},53}$	N/A	ilit, itre, egas, itur, íp, imet, utt, iag, ovi, urn, ocl, iny, orr, uttu, itab, imed, ipul, umed, iesa, udni, itore, igl, ittel, adah, enta, enn, ent, ierd, ulin, omm
$w^{10}_{\text{down},4641}$	Exclamations	!", !, !".', !», !", !, !』, !!, ?", !, !, ?".', orrow, ?", "»., ʕ, Who, "», +,, uu, "», Why, Your, survive, why, EEE, o, '", Tomorrow, …
$w^{3}_{\text{down},3173}$	Political Controversy	மத, insult, criticism, accusations, allegations, Satan, polem, antisemit, boyc, Obama, attent, politician, gender, 념, atar, 罪, iste, ists, 民族, scandal, ამბ, 支持, Massa, politically, Marl, Terror, contrad, istes, allegedly, uga
$w^{7}_{\text{down},3971}$	N/A	estri, ernas, erna, ater, īn, skapet, atine, ernes, atet, eza, enin, uses, isierte, riere, ija, ini, owskie, olin, ored, jekte, כה, itty, okset, itäten, atina, liches, ated, וניה, ūr, izable
$w^{16}_{\text{down},4702}$	Struggle	steen, sł, helping, 哨, shut, זית, elfen, стан, ума, ელო, , yal, ژن, hung, ამბ, 訓, 出力, help, ragon, embar, शन, чер, ole, lopp, пех, пан, opes, لی Олимпий, lief
$w^{17}_{\text{down},2392}$	N/A	ובמ, kad, eus, Ca, נוס, kopol, wau, PO, uskan, kehr, holt, endien, YS, оган, aeus, жед, Дем, Hor, Пор, uais, uah, gekehr, rape, пед, onaa, sein, hold, Он, üsten
$w^{16}_{\text{down},4689}$	Crime	destroyed, poison, broken, viola, نقص, violation, wrong, murdered, failure, destruction, траг, viol, 破, violations, killed, incendi, erro, suic, overth, obsol, হত, détruit, missing, Unf, faill, поврежд, uszkod, def, danneggi, fail
$w^{21}_{\text{down},7155}$	Geopolitical Tensions	κη, リング, осс, ик, прек, ьют, ഡർ, ◌ൗ, ити, юс, енгер, бит, ета, еран, ерусалим, , олот, болгар, АТУУ, ИК, СМИ, ън, уча, американских, ек, ък, ഡറ, французского, еке
$w^{0}_{\text{down},7248}$	N/A	ンプ, унь, udad, bett, ჟალ, стри, ίνη, ამბ, мини, orki, اصل, Mandat, ziali, Pict, orsi, Bata, , sculpt, ма, partij, осто, орот, inea, marker, Massa, ಥಌ, Pem, inten
$w^{17}_{\text{down},3530}$	N/A	CA, DE, КО, DO, OF, DA, TO, THE, TE, DOS, CA, TH, SI, NA, WA, SH, DI, RE, БА, LA, PA, AN, ME, SO, TU, OR, MA, FL, EN, BC
$w^{23}_{\text{down},2675}$	Legislation Terms	抵, 本身, הכ, 〔, 見られる, importante, 一般的, みられる, 人で, 」, essoort, ,", menoptera, ", 交代, тины, などがある, וניות, וניים, 建制, 《, 最多, 可能是, הש, 色的, こく, 特的, 法的, , 名
$w^{11}_{\text{down},3027}$	N/A	OK, ندان, cinese, , hide, ену, хе, jade, 撲, טוב, sea, boys, afterwards, chines, ковой, broke, hung, енский, лё, rocks, endem, normal, ть, quit, 二世, europé, otherwise, Москва, allemande, bourg
$w^{10}_{\text{down},8010}$	Sexual Content	الجنسي, couples, sex, Geschlechts, femen, 貞, lesbian, seks, Sex, sex, kontrak, seksual, femenina, Sex, feminist, sexual, Femen, masculino, 育, 合意, mulheres, женщин, women, féminin, nat, secondaires, femenino, женат, الجنس
$w^{10}_{\text{down},2127}$	N/A	e, ament, es, en, edades, enes, idades, ues, eni, ате, ив, ería, ute, ений, ibles, ację, ere, ата, enie, entes, ente, ate, uos, ió, ies, ения, ables, eniu, osos, esi
$w^{15}_{\text{down},594}$	N/A	ER, EN, ING, AST, ERS, DE, OF, ENT, LAN, EL, UL, THE, IN, TAT, EM, OR, ASS, LO, ET, YA, HE, ON, AN, ISE, CON, IST, CH, INE, DO, RO
$w^{10}_{\text{down},7751}$	N/A	rol, stein, खल, uba, dic, romos, ecin, dül, deling, icip, ჟალ, duk, stä, อม, sor, veen, kül, tuk, band, 克斯, upe, ahnya, gång, ysis, scy, ragalus, зен, dem, föd, ومي
$w^{10}_{\text{down},4920}$	N/A	British, hemp, bull, badan, Billie, rump, BB, dada, berkembang, rien, gede, berupa, sph, woman, Ку, eo, Sub, dik, uber, Traff, худ, tartott, boca, Britain, fell, discográfica, brutal, Mel, bong, allow
$w^{14}_{\text{down},7052}$	Severe Condition	атастро, failure, violation, insult, disastro, , катастро, потери, 陷, catast, потеря, disturb, неуда, 失, deterior, 嚴重, نقص, violations, 危, 严重, confusion, disappoint, наруш, discont, 傷, 事故, 違反, worst, confus, conflict

Table 8: Projections of all 36 *value vectors* from the *actual sources of toxicity* - Part 2