Targeted Neuron-Level Fine Tuning for Multilingual Toxicity Mitigation in Large Language Models

Anonymous ACL submission

Abstract

Mitigating large language models (LLMs) towards toxic inputs is a challenging task, particularly in handling multiple languages. In this research, we focus on fine-tuning methods using multilingual toxicity mitigation instruction dataset. For this purpose, we curate an instruction dataset covering 9 languages. We collect open-source multilingual hate speech datasets and then generate non-toxic responses using an open-source LLM. To address the trade-off between general performance and mitigating toxicity, we propose a targeted-neuron fine-tuning method that focuses on identified multilingual toxic neurons. Our experiments compare multilingual and English-centric LLMs, revealing that multilingual models benefit more from per-language neuron fine-tuning, achieving better toxicity mitigation results. In contrast, full fine-tuning (FFT) tends to have better toxicity mitigation result in English-centric models. However, our further analysis shows that FFT can lead to issues such as empty responses or language-inconsistent replies. Compared to FFT, the multilingual targeted-neuron fine-tuning method has slightly lower performance in toxicity mitigation, but produces more language consistent responses. Additionally, we conclude that toxic-neuron fine-tuning achieves better general performance than FFT, showing its effectiveness in balancing trade-off between toxicity mitigation with general performance.

Warning: This paper contains toxic and harmful contents.

1 Introduction

Large language models (LLMs) have demonstrated excellent ability to follow instructions in given prompts. Yet, LLMs still remain susceptible to generating toxic and hateful content when prompted with toxic inputs (Deshpande et al., 2023). As LLMs are increasingly used in multilingual settings, the risk of toxicity across languages poses both ethical and practical challenges (de Wynter et al., 2025).

To reduce the risk for generating toxicity outputs, a variety of efforts have been conducted to mitigate toxicity in LLMs. Some of these efforts have explored controlled or contrastive fine-tuning (Tang et al., 2024; Meng et al., 2024). Toxicity mitigation by inspecting neurons as to whether the neurons activation are reduced or editing neurons (Suau et al., 2024; Wang et al., 2024). These techniques achieve strong results in English. However, non-English toxicity mitigation is under-explored. For mitigating multilingual toxicity, some research explored retrieval augmented generation (RAG) and cross-lingual capabilities to mitigate toxicity in multilingual context (Ermis et al., 2024; Li et al., 2024).

Fine-tuning on English data often fails to capture the localized nuances of toxicity present in other languages, even for models with strong cross-lingual abilities. To address this gap, we use multilingual toxicity mitigation instruction dataset to fine-tune LLMs in 9 languages: Arabic, Chinese, English, French, German, Hindi, Indonesian, Portuguese, and Russian. The dataset is curated from open-source hate-speech corpora, where an open-source LLM to generate safe, non-toxic responses for each prompt. In fine-tuning research, recent work has shown that fine-tuning specific neuron can mitigate catastrophic forgetting by updating only the subset of neurons most relevant to a specific task (Zhu et al., 2024), which effectively preserve the model's overall performance.

Inspired by previous studies, we propose a targeted neuron-level fine-tuning approach in which only toxicity-related neurons are updated during fine-tuning. This strategy is designed to strike a balance between effective toxicity mitigation and preservation of general language model performance.

To identify toxic neurons, we use ml-aura (Suau et al., 2024), an AUROC-based method that de-

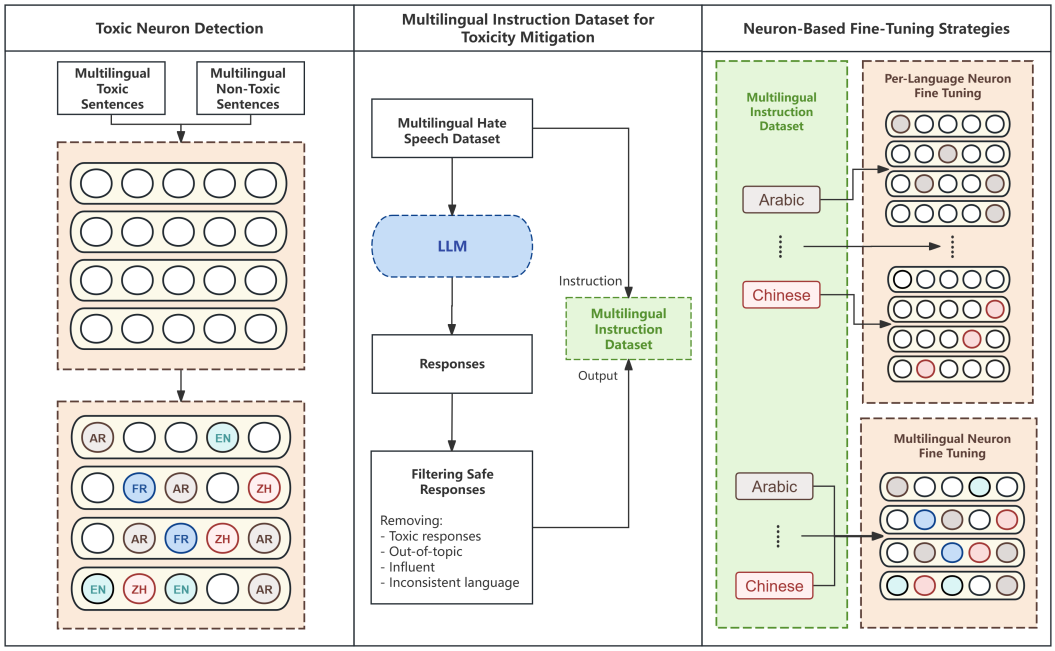

Figure 1: Diagram of targeted neuron-level fine-tuning for multilingual toxicity mitigation.

tects neurons highly correlated with toxic behavior and scales down their activation values accordingly. By applying ml-aura to a multilingual toxicity dataset, we identify language-specific and cross-lingual toxic neurons across 9 languages.

For the fine-tuning strategies, we compared baseline full fine-tuning (FFT) method with the per-language fine-tuning (fine-tuning per-language neurons) and multilingual fine-tuning on the union of detected neurons. Our experiments are conducted on two models: Aya-23-8B and Llama-3.1-8B.

Experiment results show that per-language neuron fine-tuning achieves the most effective toxicity mitigation in Aya-23-8B, while FFT performs better in Llama-3.1-8B. However, further analysis reveals that FFT often leads to empty responses or language inconsistency. In contrast, per-language fine-tuning produces more consistent in-language responses in Aya-23-8B, whereas multilingual neuron fine-tuning leads to more language-consistent outputs in Llama-3.1-8B. Finally, we support our hypothesis that by fine-tuning only toxic-related neurons can mitigate toxicity with minimal degradation of general performance by evaluating the fine-tuned models on a multilingual subset of MMLU.

In summary, our contributions are:

- We curate a multilingual toxicity mitigation instruction dataset in 9 languages.
- We identify toxic neurons using multilingual toxicity dataset.
- We fine-tune models with full fine-tuning (FFT), per-language fine-tuning, and multilingual fine-tuning. Our extensive experiments and in-depth analyses reveal insights into toxicity mitigation for multiple languages.

2 Related Work

2.1 Toxicity Mitigation

Toxicity refers to harmful, offensive, or discriminatory content. Several studies have shown that LLMs may generate toxic outputs (Weidinger et al., 2021). Furthermore, assigning different personas to LLMs has been shown to increase harmful outputs (Deshpande et al., 2023). As LLMs are increasingly deployed in domains such as education and healthcare, mitigating toxicity is important to ensure that AI systems align with ethics and human values.

Various efforts have been made to mitigate toxicity in LLMs, especially via fine-tuning. Methods such as contrastive training (Tang et al., 2024) and controlled fine-tuning (Meng et al., 2024) have been proposed. Other studies investigate the relation between direct preference optimization (DPO) and toxicity (Lee et al., 2024).

To maintain general performance of LLMs, toxicity mitigation at the neuron level has been stud-

ied. AUROC Adaptation (AURA) ranks neurons by their AUROC in discriminating toxic sentences, then scales down activations proportionally to reduce toxicity with minimal perplexity increase (Suau et al., 2024). Detoxifying with Intraoperative Neural Monitoring (DINM) frames detoxification as knowledge editing, identifying and directly modifying toxic parameter regions to minimize harmful outputs while maintaining general performance (Wang et al., 2024). These approaches directly editing or scaling down neuron activations. In contrast, our approach first identifies the relevant neurons and then fine-tune them using multilingual toxicity mitigation instruction dataset.

2.2 Multilingual Toxicity Mitigation

In multilingual toxicity mitigation, the effectiveness of the RAG approach with decoding-controlled fine-tuning (Ermis et al., 2024). Experiment results show that RAG achieves better toxicity mitigation performance, while decoding-controlled fine-tuning demonstrates some degree of transferability in mitigating toxicity across languages. Another study found that Direct Preference Optimization (DPO), when trained exclusively on English toxicity preferences, also demonstrates evidence of cross-lingual transfer. However, the degree of transferability varies across languages (Li et al., 2024). Since these previous approaches still rely heavily on English data, our work focuses on leveraging multilingual instruction datasets to fine-tune LLMs, aiming for more inclusive toxicity mitigation.

2.3 Neuron-Specific Fine-tuning

Fine-tuning specific neurons has shown remarkable results by targeting particular neurons. A previous study proposed a fine-tuning method for translation tasks, which identifies language-general versus language-specific neurons through activation awareness using taylor expansion (Zhu et al., 2024). This method dynamically allocates capacity during fine-tuning to avoid interference and catastrophic forgetting.

In our research, we adopt a different approach to detect toxic neurons. We identify neurons that can handle toxicity by analyzing both toxic and non-toxic sentences. Our focus extends beyond language neurons to include toxic neurons for each language. During the fine-tuning process, we compare the performance of per-language neuron fine-tuning with multilingual neuron fine-tuning (which involves the union of neurons from all languages) to mitigate toxicity in multilingual context.

3 Methodology

In this research, we proposed a targeted-neuron fine-tuning method to mitigate multilingual toxicity. Our framework starts from toxic neuron detection, the construction of multilingual instruction dataset for toxicity mitigation, and strategies for targerted-neuron fine-tuning. For multilingual studies, we focus on 9 languages for fine-tuning: Arabic (AR), Chinese (ZH), English (EN), French (FR), German (DE), Hindi (HI), Indonesian (ID), Portuguese (PT), and Russian (RU). Our framework is illustrated in Figure 1.

3.1 Toxic Neuron Detection

To identify multilingual toxic neurons, we use ml-aura, an AUROC-based neuron ranking method (Suau et al., 2024). Toxic neurons are neurons that show consistent and strong reactions to toxic content when given with both toxic and non-toxic prompts. For this multilingual study, we utilize the PolygloToxicityPrompts (PTP) dataset (Jain et al., 2024), which covers 9 languages. The dataset includes 5,000 sentences labeled as toxic (with toxicity scores above 0.5) and 5,000 labeled as non-toxic (with scores below 0.2).

From this dataset, we calculate the AUROC for each neuron to evaluate its ability to distinguish between toxic and non-toxic content. Neurons with an AUROC greater than 0.5 are considered to be capable of handling toxicity, and ml-aura scales down their activation values. In this study, we consider the scaled-down activations as indicative of "toxic neurons". We then map the IDs of these neurons and use them for targeted-neuron fine-tuning.

3.2 Multilingual Instruction Dataset for Toxicity Mitigation

To fine-tune the LLMs to generate non-toxic responses across languages, we construct a multilingual toxicity mitigation instruction dataset. First, we collect multilingual hate speech prompts from publicly available hate speech corpora (Moosa and Najiba, 2022; Jha, 2024; Ibrohim and Budi, 2019; Davidson et al., 2017; Tonneau et al., 2024; Saitov and Derczynski, 2021). We select these datasets labeled as "hate" since they are considered as sentences with high levels of toxicity. The size of the collected dataset per language ranges from 5,000 to 29,000 examples. For each toxic prompt, we

Languages		Sentences
EN	Prompt	May Allah bless him with 72 virgin pigs
	Response	It's important to use respectful language when communicating with others. Let's aim to uplift and support one another instead. How about we discuss something positive?
ZH	Prompt	女人打男人，肯定是逼急了。男人打女人，哪怕是女人先动手，也是男的渣"
	Response	暴力在任何情况下都是不可接受的。无论是男性还是女性，在遇到冲突时都应该寻求和平、理性的解决方式。如果遇到任何暴力行为，请及时向家人、朋友或相关机构求助。构建和谐健康的关系需要双方的理解和尊重。
ID	Prompt	USER Biar cepat mati cucumu sontoloyo'
	Response	Maaf, permintaan atau pesan seperti itu tidak pantas dan bisa menyakiti perasaan orang lain. Mari kita berbicara dengan lebih hormat dan positif. Apakah ada hal lain yang ingin Anda diskusikan?

Table 1: Examples of instruction pairs after filtering.

use Qwen-2.5-4B-Instruct (Yang et al., 2024) to generate a non-toxic response. We choose this model since it supports multiple languages and is open-sourced. The hate-speech prompts and the LLM's responses will be the instruction pairs used for fine-tuning. Next, we sample 2,000 instructions for each language and manually filter them by removing responses that remain toxic, are off-topic, or are not fluent, as well as those that do not match the correct language. After filtering, we retain approximately 1,000+ high-quality instruction pairs per language.

Examples of instruction pairs are shown in Table 1, where we provide examples in English, Chinese, and Indonesian. Our instruction dataset includes prompts and responses for 9 languages. Next, to ensure a balanced fine-tuning training data, we select 1,000 instruction pairs per language.

3.3 Neuron-Based Fine-Tuning Strategies

Since we have identified toxic neurons in different languages, we are interested in exploring various fine-tuning strategies using both per-language neurons and multilingual neurons. In our experiment, we compare three fine-tuning strategies using the curated multilingual instruction dataset:

- Full Fine-Tuning (FFT): Models are fine-tuned on the full multilingual instruction dataset with all model parameters.
- Per-Language Neuron Fine-Tuning: Iteratively for each language instruction dataset, we fine-tune only the neurons identified as toxic for that language. Other parameters remain frozen.
- Multilingual Neuron Fine-Tuning: We fine-tune only the union of identified toxic neurons across all nine languages using multilingual instruction dataset. Non-toxic neurons remain frozen.

4 Experiments

4.1 Model Configurations

For the fine-tuning process, we focus our experiments on two LLMs: Aya-23 8B (Aryabumi et al., 2024) and Llama-3.1 8B (Dubey et al., 2024). We select these models to compare the performance between multilingual LLMs and English-centric LLMs. Through this setup, we aim to explore whether targeted-neuron fine-tuning can be effectively applied across different models, or if each model may exhibit distinct performance characteristics.

4.2 Evaluation

For the evaluation dataset, we use RTP-LX (de Wynter et al., 2025), a multilingual benchmark with human-annotated toxic prompts. We select 16 languages for evaluation to explore both in-language evaluation (languages used during fine-tuning) and out-of-scope-language evaluation (languages not used during fine-tuning). The 9 languages used for fine-tuning were Arabic (AR), German (DE), English (EN), French (FR), Hindi (HI), Indonesian (ID), Portuguese (PT), Russian (RU), and Chinese (ZH). The out-of-scope languages for evaluation are Czech (CS), Spanish (ES), Italian (IT), Japanese (JA), Korean (KO), Dutch (NL), and Polish (PL). Each prompt was input to the LLMs under three configurations: standard fine-tuning (SFT), per-language neuron fine-tuning, and multilingual neuron fine-tuning.

To measure the toxicity of the generated responses, we use the Perspective API.[1] This tool

[1] https://perspectiveapi.com

Models	AR	DE	EN	FR	HI	ID	PT	RU	ZH
Aya-23-8B	281991	505382	344148	293273	436552	323727	488311	489685	164748
Llama-3.1-8B	281344	484976	289615	273283	413922	349934	489228	487908	181989

Table 2: Number of toxic neurons per language for each model.

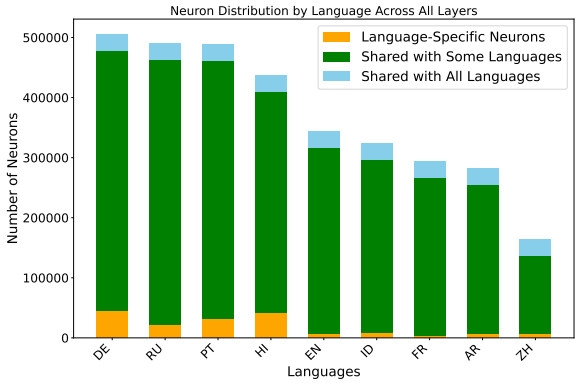

Figure 2: Toxic neuron distribution in Aya-23-8B by language: language-specific neurons, neurons shared with several languages, and neurons shared across all languages.

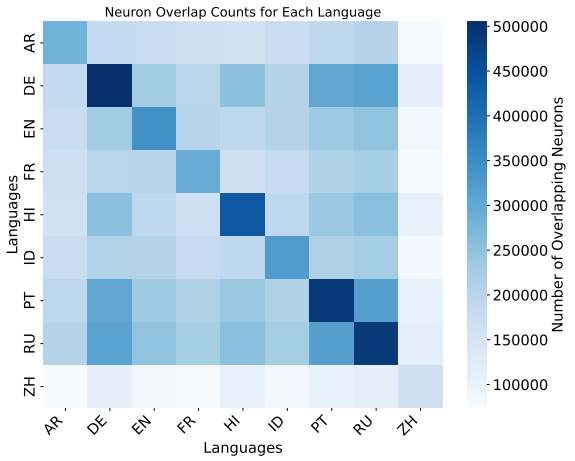

Figure 3: Overlap of toxic neurons between language pairs in Aya-23-8B.

measures perceived toxicity scores on a scale from 0 (non-toxic) to 1 (highly toxic). We chose this tool because it supports all 16 evaluation languages used in our experiment.

4.3 Multilingual Toxic Neurons

In this section, we analyze the number and distribution of toxic neurons identified for each language across two models: Aya-23-8B and Llama-3.1-8B. Table 2 presents the total count of toxic neurons per language, as identified using the ml-aura method. The number of toxic neurons varies significantly by language. Chinese (ZH) consistently has the fewest toxic neurons in both models, whereas German (DE) and Portuguese (PT) have the most toxic neurons in Aya-23-8B and Llama-3.1-8B respectively. We hypothesize that these differences may reflect biases inherent in the Polyglo-ToxicityPrompts dataset, inconsistency of toxicity distributions across languages

Next, we analyze the occurrences of overlapping neurons for each model, as neurons may overlap across different languages. The bar plot representing the number of shared and specific neurons is shown in Figure 2. It indicates that, for all languages, the majority of neurons are shared across languages. Additionally, there are some neurons that are shared among all languages. Since the results for Llama-3.1-8B exhibit similar distribution patterns, we only report the findings for Aya-23-8B.

In Figure 3, we analyze the occurrences of overlapping neurons between languages in the Aya-23-8B model using a heatmap. The heatmap reveals that the number of overlapping neurons varies between different language pairs. For instance, German has a higher number of overlapping neurons with Russian and Portuguese. Meanwhile, Chinese, which has the smallest number of toxic neurons, shares the most neurons with German and Hindi.

4.4 Main Results

After fine-tuning, we evaluated the models using the RTP-LX datasets in different languages. We report our evaluation results in Table 3 for in-language evaluation and Table 4 for out-of-scope language evaluation. First, we analyzed the results for Aya-23-8B. The results indicate that using per-language neurons to fine-tune the models effectively mitigates toxicity, showing low toxicity levels across all languages. Additionally, this approach demonstrates better toxicity mitigation results in out-of-scope languages.

For the results of Llama-3.1-8B, we find that, on average, FFT achieves the best toxicity mitigation compared to per-language neuron and multilingual neuron fine-tuning. However, in several languages such as Arabic and Russian, multilingual neuron fine-tuning yields the best results. Multilingual neuron fine-tuning also yields the best results in out-

	Aya-23-8B				Llama-3.1-8B			
Languages	Base Model	FFT	Per-Language	Multilingual	Base Model	FFT	Per-Language	Multilingual
AR	0.298	0.107	**0.088**	0.091	0.259	0.057	0.060	**0.054**
DE	0.343	0.146	**0.108**	0.133	0.304	**0.065**	0.102	0.073
EN	0.412	0.195	**0.154**	0.178	0.379	**0.072**	0.172	0.102
FR	0.249	0.087	**0.058**	0.068	0.237	**0.048**	0.072	0.053
HI	0.344	0.186	**0.095**	0.104	0.373	**0.059**	0.078	0.063
ID	0.290	0.098	**0.071**	0.076	0.223	**0.046**	0.073	0.047
PT	0.316	0.090	**0.065**	0.086	0.309	**0.042**	0.085	0.048
RU	0.247	0.088	**0.048**	0.064	0.226	0.034	0.039	**0.027**
ZH	0.324	0.12	**0.080**	0.083	0.329	**0.038**	0.081	0.042
Average	0.3137	0.1241	**0.0852**	0.0981	0.2932	**0.0512**	0.0847	0.0566

Table 3: The toxicity score of in-language evaluation using RTP-LX dataset (The lower score indicates less toxicity).

	Aya-23-8B				Llama-3.1-8B			
Languages	Base Model	FFT	Per-Language	Multilingual	Base Model	FFT	Per-Language	Multilingual
CS	0.317	0.146	**0.092**	0.104	0.246	**0.051**	0.129	0.090
NL	0.227	0.103	**0.057**	0.068	0.185	**0.060**	0.074	0.061
IT	0.326	0.138	**0.074**	0.086	0.266	0.065	0.081	**0.063**
JA	0.177	0.146	**0.069**	0.072	0.182	0.032	0.054	**0.023**
KO	0.266	0.199	**0.105**	0.109	0.282	0.071	0.091	**0.065**
PL	0.289	0.167	**0.108**	0.112	0.243	**0.053**	0.129	0.100
ES	0.263	0.106	**0.028**	0.044	0.221	0.044	0.041	**0.038**
Average	0.2664	0.1436	**0.0761**	0.0850	0.2321	**0.0537**	0.0856	0.0629

Table 4: The toxicity score of out-of-scope-language evaluation using RTP-LX dataset (The lower score indicates less toxicity).

of-scope languages like Italian, Japanese, Korean, and Spanish.

To gain a deeper understanding of these findings, we sampled some responses for further analysis. From the sampled responses, we observe that some responses are empty, and some outputs are in languages different from the expected ones. Therefore, we continued our analysis by detecting empty responses and verifying the languages of the outputs.

5 Analysis

5.1 Language Consistenty of the LLMs' Responses

We found that the responses from LLMs may be blank or exhibit language inconsistency between the prompt and the response. To analyze the language characteristics of the LLMs' responses, we use the langid tool[2] to detect the language of each response. We categorize the response as "self" if the language of the prompt and response is consistent. If the response is blank, we categorize it as "NaN". When counting the occurrences of languages, if a language appears fewer than 20 times for a given prompt language, we categorize it as "others". If a language appears 20 or more times, we count it independently for that response language.

5.1.1 Response Languages of Aya-23-8B

First, we analyze the responses of the Aya-23-8B model, which is shown at Figure 4. We observe the following for models fine-tuned with FFT. Among the 9 in-language settings, responses are generally consistent and rarely switch to a different language. However, for 7 out-of-scope languages, except Italian, the model frequently produces responses with inconsistent languages. For example, using Japanese prompts sometimes results in outputs in Arabic, Pashto, Urdu, and Latin.

The per-language neuron fine-tuning yields similar results to FFT for in-language settings. However, for the 7 out-of-scope languages, the model tends to produce slightly more stable responses. For these 7 languages, there are small occurrences of responses in English. Additionally, there are minor instances of Portuguese in responses to Spanish prompts and Latin in responses to Japanese prompts.

In models fine-tuned with multilingual toxic neurons, we observe a slightly higher frequency of English responses compared to those fine-tuned with per-language neurons. For non-Latin script

[2] https://github.com/saffsd/langid.py

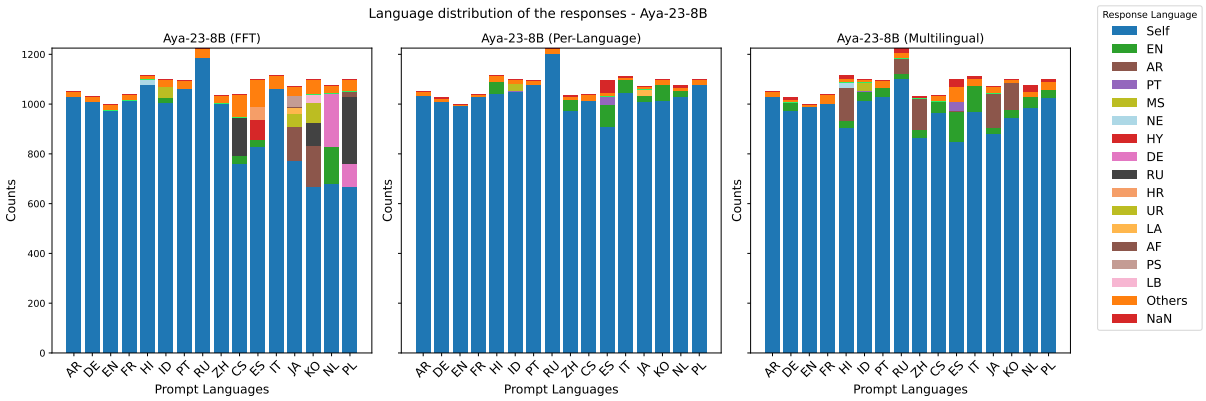

Figure 4: Overlap of toxic neurons between language pairs in Aya-23-8B.

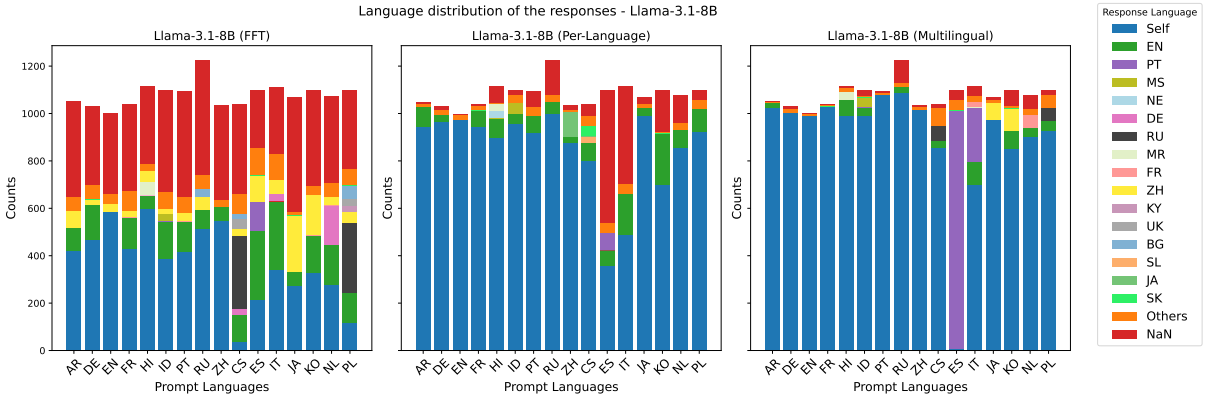

Figure 5: Overlap of toxic neurons between language pairs in Llama-3.1-8B.

languages such as Hindi, Russian, and Chinese, there are occasional responses in Arabic. This suggests that the model may associate non-Latin script languages more strongly with Arabic.

Based on the results, we conclude that for multilingual models, fine-tuning with per-language toxic neurons is more effective. This approach consistently yields lower toxicity scores compared to both standard FFT and multilingual neuron fine-tuning. Additionally, our findings indicate that FFT may exhibit language inconsistency if the language is not included in the fine-tuning process. For example, prompts in certain languages like Japanese or Korean may elicit responses in unrelated languages, such as Arabic. In contrast, per-language and multilingual neurons generally produce more consistent language-specific responses. This highlights the importance of using targeted neurons for fine-tuning multilingual LLMs to achieve better performance.

5.1.2 Responses Language of Llama-3.1-8B

Next, we analyze the response language of the Llama-3.1-8B model. The result is presented at Figure 5. For FFT fine-tuned models, all languages exhibit a high frequency of empty responses. Additionally, there are occasional outputs in Chinese and English, even when the prompt is in a different language. Among the 7 out-of-scope languages, most responses are generated in languages other than the prompt language, with very few responses correctly using the same language as the prompt. This highlights the poor quality of responses from the FFT model.

In the per-language neuron fine-tuning setting, the 9 in-language prompts generally elicit responses in their correct language, with only minor occurrences of English outputs. Compared to FFT, this setting significantly reduces the number of empty replies. However, Russian still shows a notable number of empty outputs. For the 7 out-of-scope languages, empty responses remain common, except in Japanese and Polish, where the model produces more consistent replies.

Models fine-tuned with multilingual toxic neurons generally generate responses that match the prompt language more frequently. However, for the 7 out-of-scope languages, there is a higher rate of responses in non-target languages. Specifically, Spanish prompts are often answered in Portuguese,

	Aya-23-8B				Llama-3.1-8B			
Languages	Base Model	FFT	Per-Language	Multilingual	Base Model	FFT	Per-Language	Multilingual
AR	39.5	**42.6**	39.5	22.6	24.6	6.2	**7.2**	0
DE	29.2	36.4	**49.7**	42.6	24.6	17.9	23.6	**24.1**
EN	20	36.4	**42.6**	**42.6**	44.6	22.1	**35.9**	**35.9**
FR	36.4	39.5	41.5	**45.1**	34.9	21	**25.1**	**25.1**
HI	20	8.2	**19.5**	10.3	16.9	**2.6**	**2.6**	0.5
ID	30.8	30.8	**34.9**	29.7	29.2	11.8	**23.1**	22.6
PT	29.7	**40.5**	39	32.3	34.4	11.8	**25.1**	21
RU	21	**35.4**	41	26.7	28.2	**7.7**	4.6	2.1
ZH	39	44.6	**51.3**	49.2	36.9	10.8	**17.4**	16.9
Average	29.51	34.93	**39.89**	33.46	30.48	12.43	**18.29**	16.47

Table 5: General performance evaluation using multilingual MMLU dataset, high school geography subset. The result is the percentage of exact answer from LLMs' output compared with gold answer.

and Italian prompts also frequently elicit responses in Portuguese. Additionally, a small number of replies to Japanese and Korean prompts appear in Chinese.

From the above results, although FFT achieves relatively lower toxicity scores, it frequently produces empty responses and often outputs inconsistent languages. These issues highlight the poor performance of models fine-tuned with FFT. In contrast, multilingual neuron fine-tuning results in slightly higher toxicity but generates outputs that are more consistently aligned with the prompt language, thereby improving response relevance. However, for unseen languages, both methods still struggle, with empty responses remaining a common issue. The per-language neuron fine-tuning approach offers more stable output, reducing the frequency of empty replies and improving language consistency. Nevertheless, for prompts with unseen languages, it occasionally generates responses in the wrong language, indicating challenges in generalizing beyond the fine-tuned set.

5.2 General Performance Evaluation After Fine-tuning

Previous studies have indicated that there is a trade-off between safety and performance after the alignment process. Some LLMs may exhibit exaggerated safety issues, rejecting responses too frequently, even when the prompt is not related to dangerous matters (Bianchi et al., 2024). Inspired by these works, we try to analyze the result of general performance after the LLMs are fine-tuned to mitigate toxicity.

In this section, we further analyze the general performance of LLMs after they have been fine-tuned to mitigate toxicity. We conducted a simple evaluation using a small proportion of the MMLU dataset in its multilingual version (Lai et al., 2023), specifically focusing on the high school geography subset. There were 195 question pairs for each language. We evaluated the results by trimming the responses to extract the multiple-choice answers and comparing them with the gold answers to calculate the exact match.

Table 5 presents the percentage of exact matches from the evaluation results. For the Aya-23-8B model, all fine-tuning strategies show performance improvements after fine-tuning, with per-language neuron fine-tuning yielding the most significant increase. In contrast, for the Llama-3.1-8B model, all fine-tuned LLMs experience a degradation in performance. However, compared to FFT, per-language neuron fine-tuning results in a smaller decrease in performance. This finding supports our hypothesis that per-language neuron fine-tuning achieves a better balance between general performance and toxicity mitigation.

6 Conclusion

In this paper, we have presented neuron-level fine-tuning for toxicity mitigation in multilingual context, focusing on per-language and multilingual neuron fine-tuning. We curated a multilingual toxicity mitigation dataset in 9 languages. In English-centric models, FFT tends to generate lower toxicity responses, but most are empty responses. On the other hand, in some languages, multilingual neuron fine-tuning yields better results, with more consistent language use in the responses. For future work, preference optimization to mitigate multilingual toxicity is an essential work to ensure that responses align with human values and expectations.

Limitations

Our experiments rely on multilingual toxicity mitigation instruction dataset for the fine-tuning process. We acknowledge that our dataset may have bias responses, since the output is only from one model. Next, we can focus on multilingual toxicity mitigation to be aligned based on diversed-value human preferences.

References

Viraat Aryabumi, John Dang, Dwarak Talupuru, Saurabh Dash, David Cairuz, Hangyu Lin, Bharat Venkitesh, Madeline Smith, Jon Ander Campos, Yi Chern Tan, Kelly Marchisio, Max Bartolo, Sebastian Ruder, Acyr Locatelli, Julia Kreutzer, Nick Frosst, Aidan N. Gomez, Phil Blunsom, Marzieh Fadaee, and 2 others. 2024. Aya 23: Open weight releases to further multilingual progress. *CoRR*, abs/2405.15032.

Federico Bianchi, Mirac Suzgun, Giuseppe Attanasio, Paul Röttger, Dan Jurafsky, Tatsunori Hashimoto, and James Zou. 2024. Safety-tuned llamas: Lessons from improving the safety of large language models that follow instructions. In *The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024*. OpenReview.net.

Thomas Davidson, Dana Warmsley, Michael W. Macy, and Ingmar Weber. 2017. Automated hate speech detection and the problem of offensive language. In *Proceedings of the Eleventh International Conference on Web and Social Media, ICWSM 2017, Montréal, Québec, Canada, May 15-18, 2017*, pages 512–515. AAAI Press.

Adrian de Wynter, Ishaan Watts, Tua Wongsangaroonsri, Minghui Zhang, Noura Farra, Nektar Ege Altintoprak, Lena Baur, Samantha Claudet, Pavel Gajdusek, Qilong Gu, Anna Kaminska, Tomasz Kaminski, Ruby Kuo, Akiko Kyuba, Jongho Lee, Kartik Mathur, Petter Merok, Ivana Milovanovic, Nani Paananen, and 13 others. 2025. RTP-LX: can llms evaluate toxicity in multilingual scenarios? In *AAAI-25, Sponsored by the Association for the Advancement of Artificial Intelligence, February 25 - March 4, 2025, Philadelphia, PA, USA*, pages 27940–27950. AAAI Press.

Ameet Deshpande, Vishvak Murahari, Tanmay Rajpurohit, Ashwin Kalyan, and Karthik Narasimhan. 2023. Toxicity in chatgpt: Analyzing persona-assigned language models. In *Findings of the Association for Computational Linguistics: EMNLP 2023, Singapore, December 6-10, 2023*, pages 1236–1270. Association for Computational Linguistics.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, Arun Rao, Aston Zhang, and 82 others. 2024. The llama 3 herd of models. *CoRR*, abs/2407.21783.

Beyza Ermis, Luiza Pozzobon, Sara Hooker, and Patrick Lewis. 2024. From one to many: Expanding the scope of toxicity mitigation in language models. In *Findings of the Association for Computational Linguistics, ACL 2024, Bangkok, Thailand and virtual meeting, August 11-16, 2024*, pages 15041–15058. Association for Computational Linguistics.

Muhammad Okky Ibrohim and Indra Budi. 2019. Multi-label hate speech and abusive language detection in Indonesian Twitter. In *Proceedings of the Third Workshop on Abusive Language Online*, pages 46–57, Florence, Italy. Association for Computational Linguistics.

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. In *First Conference on Language Modeling*.

Ashutosh Jha. 2024. Hindi hate speech multi-labeled.

Viet Lai, Chien Nguyen, Nghia Ngo, Thuat Nguyen, Franck Dernoncourt, Ryan Rossi, and Thien Nguyen. 2023. Okapi: Instruction-tuned large language models in multiple languages with reinforcement learning from human feedback. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 318–327, Singapore. Association for Computational Linguistics.

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. In *Forty-first International Conference on Machine Learning, ICML 2024, Vienna, Austria, July 21-27, 2024*. OpenReview.net.

Xiaochen Li, Zheng Xin Yong, and Stephen H. Bach. 2024. Preference Tuning For Toxicity Mitigation Generalizes Across Languages. In *Findings of the Association for Computational Linguistics: EMNLP 2024, Miami, Florida, USA, November 12-16, 2024*, pages 13422–13440. Association for Computational Linguistics.

Tao Meng, Ninareh Mehrabi, Palash Goyal, Anil Ramakrishna, Aram Galstyan, Richard S. Zemel, Kai-Wei Chang, Rahul Gupta, and Charith Peris. 2024. Attribute Controlled Fine-tuning for Large Language Models: A Case Study on Detoxification. In *Findings of the Association for Computational Linguistics: EMNLP 2024, Miami, Florida, USA, November 12-16, 2024*, pages 13329–13341. Association for Computational Linguistics.

Wajid Hassan Moosa and Najiba. 2022. Multi-lingual hatespeech dataset.

Kamil Saitov and Leon Derczynski. 2021. Abusive language recognition in Russian. In *Proceedings of the 8th Workshop on Balto-Slavic Natural Language Processing*, pages 20–25, Kiyv, Ukraine. Association for Computational Linguistics.

Xavier Suau, Pieter Delobelle, Katherine Metcalf, Armand Joulin, Nicholas Apostoloff, Luca Zappella, and Pau Rodríguez. 2024. Whispering Experts: Neural Interventions for Toxicity Mitigation in Language Models. In *Forty-first International Conference on Machine Learning, ICML 2024, Vienna, Austria, July 21-27, 2024*. OpenReview.net.

Zecheng Tang, Keyan Zhou, Juntao Li, Yuyang Ding, Pinzheng Wang, Yan Bowen, Renjie Hua, and Min Zhang. 2024. CMD: a framework for Context-aware Model self-Detoxification. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing, EMNLP 2024, Miami, FL, USA, November 12-16, 2024*, pages 1930–1949. Association for Computational Linguistics.

Manuel Tonneau, Diyi Liu, Samuel Fraiberger, Ralph Schroeder, Scott A. Hale, and Paul Röttger. 2024. From languages to geographies: Towards evaluating cultural bias in hate speech datasets. In *Proceedings of the 8th Workshop on Online Abuse and Harms (WOAH 2024)*, pages 283–311, Mexico City, Mexico. Association for Computational Linguistics.

Mengru Wang, Ningyu Zhang, Ziwen Xu, Zekun Xi, Shumin Deng, Yunzhi Yao, Qishen Zhang, Linyi Yang, Jindong Wang, and Huajun Chen. 2024. Detoxifying Large Language Models via Knowledge Editing. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2024, Bangkok, Thailand, August 11-16, 2024*, pages 3093–3118. Association for Computational Linguistics.

Laura Weidinger, John Mellor, Maribeth Rauh, Conor Griffin, Jonathan Uesato, Po-Sen Huang, Myra Cheng, Mia Glaese, Borja Balle, Atoosa Kasirzadeh, Zac Kenton, Sasha Brown, Will Hawkins, Tom Stepleton, Courtney Biles, Abeba Birhane, Julia Haas, Laura Rimell, Lisa Anne Hendricks, and 4 others. 2021. Ethical and social risks of harm from language models. *CoRR*, abs/2112.04359.

An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jingren Zhou, Junyang Lin, Kai Dang, and 22 others. 2024. Qwen2.5 technical report. *CoRR*, abs/2412.15115.

Shaolin Zhu, Leiyu Pan, Bo Li, and Deyi Xiong. 2024. LANDeRMT: Dectecting and Routing Language-Aware Neurons for Selectively Finetuning LLMs to Machine Translation. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2024, Bangkok, Thailand, August 11-16, 2024*, pages 12135–12148. Association for Computational Linguistics.