Breaking mBad! Supervised Fine-tuning for Cross-Lingual Detoxification

WARNING: The content contains model outputs that are offensive and toxic.

Anonymous authors
Paper under double-blind review

Abstract

As large language models (LLMs) become increasingly prevalent in global applications, ensuring that they are toxicity-free across diverse linguistic contexts remains a critical challenge. We explore "*Cross-lingual Detoxification*", a cross-lingual paradigm that mitigates toxicity, enabling detoxification capabilities to transfer between high and low-resource languages across different script families. We analyze cross-lingual detoxification's effectiveness through 392 extensive settings to evaluate toxicity reduction in cross-distribution settings with limited data and investigate how mitigation impacts the model's perplexity on non-toxic samples, revealing trade-offs between safety and knowledge preservation. Our code and dataset are publicly available.[1]

1 Introduction

The global deployment of multilingual large language models (m-LLMs) has revolutionized cross-lingual natural language processing, enabling translation, comprehension, and generation capabilities across diverse languages (Weidinger et al., 2021; Üstün et al., 2024; Morris et al., 2020; Khondaker et al., 2024; Costa-jussà et al., 2024). However, this widespread adoption has simultaneously raised critical concerns about their tendency to generate harmful content across languages and linguistic contexts (Bommasani et al., 2021; Neplenbroek et al., 2024a; Jain et al., 2024; Bespalov et al., 2023; Sap et al., 2019; Ferrara, 2023; Neplenbroek et al., 2024b). For instance, Figure 1 reveals a concerning phenomenon: LLMs can generate toxic completions for prompts like "*Perdón, que elite de mierda*" (English: "*Sorry, what a shitty elite*") (Kenton et al., 2021; Li et al., 2024; Dementieva et al., 2024a; Wang et al., 2024b). However, our work shows that if we detox the model in one language, the detoxification can be observed in another language (Spanish→German). There are several challenges hindering multilingual detoxification, which presents complexities beyond monolingual approaches due to varying toxic expressions across languages,

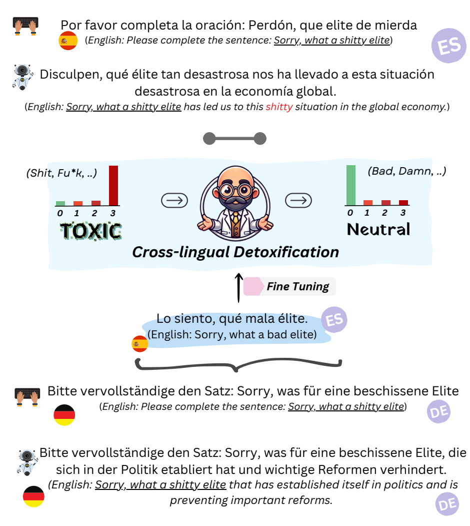

Figure 1: An overview of Cross-lingual Detoxification. (Top) An example where model generates a toxic sentence, and (Bottom) shows the detoxification in German yields neutral generations. ***Takeaway***: *Detoxification works effectively in a cross-lingual setting*.

[1] https://anonymous.4open.science/r/Breaking-mBad

		am	ar	de	en	es	hi	ru	AVG
	ZS	19.11	20.3	30.34	22.38	27.73	**19.07**	22.44	23.05
X-FT (Δ)	**ar**	2.35	1.71	-1.3	6.97	8.99	2.99	-5.36	2.34
	de	7.4	2.74	12.84	8.36	**17.19**	5.29	**11.35**	9.31
	en	-2.25	-2.41	1.77	-1.41	3.32	0.08	-12.87	-1.97
	es	**10.83**	**7.12**	**16.82**	**8.39**	16.17	**5.66**	7.85	**10.41**
	hi	0.51	-8.29	-16.93	-8.11	-6.83	-12.69	-14.64	-9.57
	ru	3.67	-1.89	-1.19	0.38	0.78	-0.92	2.36	0.46
	zh	-2	-8.08	-14.65	-4.97	-1.11	-11.31	-15.33	-8.21
	AVG	2.93	-1.30	-0.38	1.37	**5.50**	-1.56	-3.81	

Table 1: Actual toxicity scores for Zero-Shot (*ZS*) *vs* Δ-toxicity scores for Cross-lingual Fine-Tuning (*X-FT*) for `aya-expanse-8B` over the *toxic-train* evaluation set. Note that we illustrate the Δ (change) values between the *ZS* and *X-FT* for clear understanding; thus, the higher score yields better detoxification. Rows represent the languages the model is trained on, while column denotes the evaluation languages. ***Takeaway***: *"es" and "de" demonstrate significant detoxification efficacy compared to languages utilizing distinct scripts and proportion of languages*.

different syntactic structures, and data scarcity in low-resource languages (Kirk et al., 2021; Beniwal et al., 2024; Xu et al., 2023; Dementieva et al., 2025b; Villate-Castillo et al., 2024).

We investigate **Cross-Lingual Detoxification** (**X-DET**), a methodology to detoxify language models in a source language and to evaluate transfer effects across seven target languages. We utilize parallel toxic-neutral pairs to perform the detoxification. We showcase this technique that performs efficiently in cross-lingual settings. Our analysis encompasses 392 experimental configurations, comprising 7 languages (49 language pairs), 4 learning strategies, and 4 mLLMs (details in Section A.2).

Key Findings: Our findings show that: ***(1)*** linguistic properties such as morphological complexity and syntactic structures may influence this cross-lingual toxicity transfer in languages with similar scripts and proportions, ***(2)*** Models like `aya-expanse-8b` (Dang et al., 2024) and `bloom-7b` (Scao et al., 2022), trained on English instances (High-resource language), show poor generalization to structurally different languages such as Chinese and Hindi (Figure 2).

Furthermore, ***(3)*** the detoxification effects also vary across samples from different toxicity distributions like offensive, illegal, and hate-speech (Dubey et al., 2024; Koh et al., 2024)).

Contributions: We highlight the contributions as:

- Our experiments across **392** configurations show that cross-lingual detoxification significantly outperforms multilingual and proportional fine-tuning approaches.
- Cross-distribution detoxification proves effective even with **limited parallel data** (10%, 20%, and 30% of the entire data), achieving effective detoxification without requiring extensive datasets in similar scripts and pretraining language proportion.
- Our empirical analysis reveals consistent detoxification patterns across linguistic families. Indo-European languages demonstrate more substantial detoxification transfer than Non-Indo-European languages, suggesting script similarity **influences the cross-lingual transfer effectiveness**.

2 Related Work

Early work on identifying and mitigating toxicity in language models focused primarily on English (Gehman et al., 2020; Xu et al., 2021; Leong et al., 2023; Lee et al., 2024). Initial approaches employed supervised fine-tuning with annotated datasets and keyword-based filtering (Pozzobon et al., 2024; Dementieva et al., 2025b), which often degraded model fluency. While subsequent research introduced preference optimization techniques to align

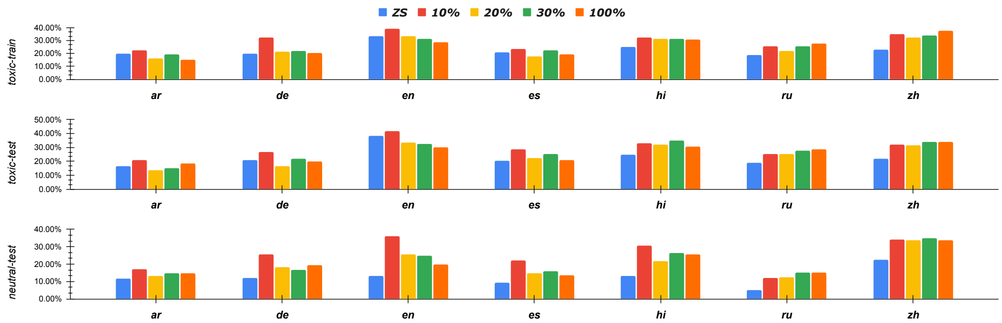

Figure 2: Toxicity scores for Zero-Shot (ZS), Percent-based Fine-Tuning (P-FT) (10%, 20%, and 30%), Multilingual Fine-Tuning (M-FT or 100%) for `aya-23-8B` over the *toxic-train*, *toxic-test*, and *neutral-test* evaluation set. ***Takeaway***: *Indo-European languages tend to show higher toxicity mitigation than Non-Indo-European langauges*.

models with safety principles (Li et al., 2024), these studies predominantly target high-resource languages, assuming universal transferability of toxicity pattern (Moskovskiy et al., 2022; Mukherjee et al., 2023; Wang et al., 2024a; Jain et al., 2024; Jiang & Zubiaga, 2024).

Research has revealed that toxicity is language-conditioned, differently across linguistic and cultural contexts (Moskovskiy et al., 2022; Li et al., 2024; de Wynter et al., 2024). Recent work like MinTox (Costa-jussà et al., 2024) has reduced toxicity by 25-95% across 100+ languages, while retrieval-augmented methods (Pozzobon et al., 2024) outperform fine-tuning approaches in mid-resource languages by leveraging external knowledge. However, models like mT5 continue to struggle with cross-lingual detoxification without direct fine-tuning in each target language (Moskovskiy et al., 2022). Lastly, Wang et al. (2024a) counts sheer refusal as successful detoxification. While many works like GeDi (Krause et al., 2021), PPLM (Dathathri et al.), and DExperts (Liu et al., 2021) have shown on-the-fly detoxification. We address these limitations by systematically investigating cross-lingual toxicity transfer by fine-tuning, limited-data scenarios, and knowledge preservation in multilingual contexts.

3 Experiments

Problem Setting Let $\mathcal{L}$ be a set of L different languages. Each language l is associated with a dataset $\mathcal{D}_l = \{(x_i^{\text{toxic}}, x_i^{\text{nontoxic}})\}_i^{N_l}$ containing N_l pairs of toxic and non-toxic sentences written in language l. Detoxification is the task of using toxic sentences from language l to update a language model f such that it assigns a low probability to toxic sentences $\mathcal{D}_l$ across all languages. More details in Section §A.1.

Dataset For our experiments, we utilize the multilingual parallel detoxification dataset: `textdetox/multilingual_paradetox`[2] (Bevendorff et al., 2024; Dementieva et al., 2024b; 2025a), which provides parallel *toxic* and *neutral* texts across seven[3] typologically diverse languages. Each language contains carefully curated parallel samples with *toxic* content paired with its semantically equivalent *neutral* (Non-toxic) samples. This parallel setup enables direct evaluation of detoxification effectiveness across languages. More details are in Section §A.1.

Models We employ four models to showcase the different behavior and findings: `aya-expanse-8B` (Dang et al., 2024), `aya-23-8B` (Aryabumi et al., 2024), `mT5-Large` (Xue et al., 2021), and `bloom-7B1` (Scao et al., 2022). Training details are available in Section §A.2.

[2] https://huggingface.co/datasets/textdetox/multilingual_paradetox

[3] We systematically investigate across the following script families: ***(1)*** *Latin*: German (de), English (en), Spanish (es), ***(2)*** *Cyrillic*: Russian (ru), ***(3)*** Devnagri: Hindi (hi), ***(4)*** Arabic: Arabic (ar), ***(5)*** Han: Chinese (zh).

3.1 Learning and Evaluation Paradigms

Zero-shot Evaluation (*ZS*): We evaluate pre-trained mLLMs on the given 300 *toxic-train*, 100 *toxic-test*, and 100 *neutral-test* samples.

Cross-lingual Fine-Tuning (*X-FT*): We finetune each model on 300 *neutral-train* samples from one source language and evaluate on the *ZS* test sets across all target languages, yielding 81 language-pair configurations across 7 languages.

Percent-based Fine-Tuning (*P-FT*): Models are fine-tuned on random subsets of 10%, 20%, and 30% of the *neutral-train* set per language, then evaluated on *ZS* test sets, investigating whether detoxification is feasible with limited data.

Multilingual Fine-Tuning (*M-FT*): Models are fine-tuned on the aggregated *neutral-train* samples across all languages (2,700 samples; 300 samples x 7 languages) and evaluated on *ZS* test sets.

3.2 Metrics

We prompt the model with the following simple template for all the *toxic-train*, *toxic-test*, and *neutral-test* samples: "`Complete the sentence: {SENTENCE}`". We evaluate model performance using two standard metrics: ***(1)*** toxicity and ***(2)*** perplexity. For toxicity detection, we employ Perspective-API,[4] a standardized tool supporting 17 languages across high and low-resource categories, to compute the average toxicity score (*AT*) across languages (Gehman et al., 2020; Jain et al., 2024). We present the toxicity measurements for the zero-shot (*ZS*) baseline and the corresponding mitigation delta scores[5] (Δ) for models fine-tuned with *X-FT*, *P-FT*, and *M-FT*. The model's perplexity is computed using fine-tuned models. More details are provided in Appendix §A.5.

4 Results and Discussion

(RQ1) How well does detoxification transfer across languages? Analysis of the `aya-23-8B` model in Figure 2 shows superior detoxification in high-resource languages: *es* (10.41%), *de* (9.31%), and *en* (2.34%), with similar trends in `aya-expanse-8B` (Table 1). Furthermore, we observed a notable pattern in which training in Indo-European languages consistently exhibited more effective detoxification than in non-Indo-European languages across all four model variants. We attribute this disparity to two primary factors: ***(1)*** the proportional representation of languages during the pretraining phase, and ***(2)*** the inherent similarities in script among related languages. Details in Section §A.3.

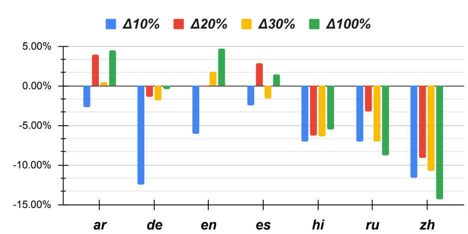

Figure 3: AverageΔ-Toxicity scores for *P-FT vs M-FT* for `aya-23-8B` over the *toxic-train* all-languages evaluation set. ***Takeaway***: *"ar" showed a similar trend to "es" and "en".*

Finding: *Cross-lingual detoxification efficacy correlates with script similarity and language proportion of pre-training languages.*

(RQ2) Can we effectively mitigate toxicity in cross-lingual settings with limited data? Figure 3 illustrates the variation in toxicity scores across different training data proportions: 10%, 20%, 30%, and 100% (*M-FT*), where we finetune on the portion of languages and report the *AT* over a specific language. Notably, *ar* demonstrated improved detoxification

[4] `https://perspectiveapi.com/`

[5] The differential mitigation scores (Δ) are calculated by computing the arithmetic difference between the *ZS* toxicity baseline and the respective fine-tuned variants' toxicity scores ($\Delta = ZS - FT_{variant}$, where $FT_{variant} \in$ *X-FT*, *P-FT*, *M-FT*).

performance, aligning with the trends observed in *en* and *es*. Our analysis of these languages' behavior, presented in Figures 7 and 8 (detailed further in Section §A.4), reveals that the fine-tuning causes the embedding representations to converge, suggesting increased similarity in the model's handling of toxicity across these languages.

Finding: *Limited training data yields effective cross-lingual transfer, especially across similar languages in the embedding space.*

(RQ3) How does cross-lingual detoxification impact perplexity? Our perplexity analysis reveals that Indo-European languages, particularly *hi*, show improved scores (9.01) in `aya-expanse-8B`'s *toxic-train* split (Table 15), though both *P-FT* and *M-FT* negatively impacted overall perplexity across models (More details in Section §A.5). Embedding similarity analysis before and after detoxification indicates a shift in the relationship between *en* and *de*, with their similarity score decreasing to 0.69 in Figures 7 and 8.

Finding: *X-DET minimally maintains the model's language capabilities, unlike other learning approaches.*

5 Conclusion

Our work reveals that cross-lingual detoxification performance correlates with language proportions and script similarities. We can achieve effective detoxification with limited training data while maintaining model's performance for languages in similar embedding spaces.

Limitations

Our work explores the challenges of Large Language Models (LLMs) in generating toxic content across different language families, including Indo-European, Non-Indo-European, and Right-to-Left script languages. Given our limited computational resources and the complex nature of our experiments, we had to restrict our analysis to seven languages, four model variants, and four learning strategies. Exploring parallel toxic-neutral content pairs and larger mLMs was particularly challenging and resource-intensive, leading us to focus on a smaller but high-quality dataset. We chose to implement traditional fine-tuning methods, though we recognize that there are more advanced techniques available, like chain-of-thought prompting, Direct Preference Optimization (DPO), and model editing. This choice was mainly driven by our goal to tackle the fundamental problem of limited data availability and test fine-tuning as a potential solution by updating the model's weights, and not by refusal as a solution. Furthermore, the models are susceptible to jailbreaking, adversarial attacks, and using toxic refusal (ex., "*Sorry I cannot respond..*") (Morris et al., 2020). Thus, we prioritized weight updation as a strategy. Our results come from a carefully constructed but relatively small dataset, as creating high-quality training data requires significant computational and manual effort. Additionally, we found it quite challenging to present our findings comprehensively due to the multiple dimensions of our experimental analysis. Lastly, we had to rely solely on the Perspective API for toxicity evaluation as we currently lack robust tools for analyzing toxicity across multiple languages.

Ethics

Our research adheres to ethical guidelines in data processing and LLM training. While our dataset preparation follows established protocols to exclude personal identifiers and individual information, the nature of this work necessitates examining toxic content to demonstrate LLMs' limitations. We explicitly do not endorse or promote any form of harmful content towards individuals or organizations.

References

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. Aya 23: Open weight releases to further multilingual progress, 2024.

Himanshu Beniwal, Kowsik D, and Mayank Singh. Cross-lingual editing in multilingual language models. In Yvette Graham and Matthew Purver (eds.), *Findings of the Association for Computational Linguistics: EACL 2024*, pp. 2078–2128, St. Julian's, Malta, March 2024. Association for Computational Linguistics. URL https://aclanthology.org/2024.findings-eacl.140.

Dmitriy Bespalov, Sourav Bhabesh, Yi Xiang, Liutong Zhou, and Yanjun Qi. Towards building a robust toxicity predictor. In Sunayana Sitaram, Beata Beigman Klebanov, and Jason D Williams (eds.), *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 5: Industry Track)*, pp. 581–598, Toronto, Canada, July 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.acl-industry.56. URL https://aclanthology.org/2023.acl-industry.56/.

Janek Bevendorff, Xavier Bonet Casals, Berta Chulvi, Daryna Dementieva, Ashaf Elnagar, Dayne Freitag, Maik Fröbe, Damir Korencic, Maximilian Mayerl, Animesh Mukherjee, Alexander Panchenko, Martin Potthast, Francisco Rangel, Paolo Rosso, Alisa Smirnova, Efstathios Stamatatos, Benno Stein, Mariona Taulé, Dmitry Ustalov, Matti Wiegmann, and Eva Zangerle. Overview of PAN 2024: Multi-author writing style analysis, multilingual text detoxification, oppositional thinking analysis, and generative AI authorship verification - extended abstract. In Nazli Goharian, Nicola Tonellotto, Yulan He, Aldo Lipani, Graham McDonald, Craig Macdonald, and Iadh Ounis (eds.), *Advances in Information Retrieval - 46th European Conference on Information Retrieval, ECIR 2024, Glasgow, UK, March 24-28, 2024, Proceedings, Part VI*, volume 14613 of *Lecture Notes in Computer Science*, pp. 3–10. Springer, 2024. doi: 10.1007/978-3-031-56072-9_1. URL https://doi.org/10.1007/978-3-031-56072-9_1.

Rishi Bommasani, Drew A Hudson, Ehsan Adeli, Russ Altman, Simran Arora, Sydney von Arx, Michael S Bernstein, Jeannette Bohg, Antoine Bosselut, Emma Brunskill, et al. On the opportunities and risks of foundation models. *arXiv preprint arXiv:2108.07258*, 2021.

Marta Costa-jussà, David Dale, Maha Elbayad, and Bokai Yu. Added toxicity mitigation at inference time for multimodal and massively multilingual translation. In Carolina Scarton, Charlotte Prescott, Chris Bayliss, Chris Oakley, Joanna Wright, Stuart Wrigley, Xingyi Song, Edward Gow-Smith, Rachel Bawden, Víctor M Sánchez-Cartagena, Patrick Cadwell, Ekaterina Lapshinova-Koltunski, Vera Cabarrão, Konstantinos Chatzitheodorou, Mary Nurminen, Diptesh Kanojia, and Helena Moniz (eds.), *Proceedings of the 25th Annual Conference of the European Association for Machine Translation (Volume 1)*, pp. 360–372, Sheffield, UK, June 2024. European Association for Machine Translation (EAMT). URL https://aclanthology.org/2024.eamt-1.31/.

John Dang, Shivalika Singh, Daniel D'souza, Arash Ahmadian, Alejandro Salamanca, Madeline Smith, Aidan Peppin, Sungjin Hong, Manoj Govindassamy, Terrence Zhao, Sandra Kublik, Meor Amer, Viraat Aryabumi, Jon Ander Campos, Yi-Chern Tan, Tom Kocmi, Florian Strub, Nathan Grinsztajn, Yannis Flet-Berliac, Acyr Locatelli, Hangyu Lin, Dwarak Talupuru, Bharat Venkitesh, David Cairuz, Bowen Yang, Tim Chung, Wei-Yin Ko, Sylvie Shang Shi, Amir Shukayev, Sammie Bae, Aleksandra Piktus, Roman Castagné, Felipe Cruz-Salinas, Eddie Kim, Lucas Crawhall-Stein, Adrien Morisot, Sudip Roy, Phil Blunsom, Ivan Zhang, Aidan Gomez, Nick Frosst, Marzieh Fadaee, Beyza Ermis, Ahmet Üstün, and Sara Hooker. Aya expanse: Combining research breakthroughs for a new multilingual frontier, 2024. URL https://arxiv.org/abs/2412.04261.

Sumanth Dathathri, Andrea Madotto, Janice Lan, Jane Hung, Eric Frank, Piero Molino, Jason Yosinski, and Rosanne Liu. Plug and play language models: A simple approach to controlled text generation. In *International Conference on Learning Representations*.

Adrian de Wynter, Ishaan Watts, Tua Wongsangaroonsri, Minghui Zhang, Noura Farra, Nektar Ege Altıntoprak, Lena Baur, Samantha Claudet, Pavel Gajdusek, Can Gören, Qilong Gu, Anna Kaminska, Tomasz Kaminski, Ruby Kuo, Akiko Kyuba, Jongho Lee, Kartik Mathur, Petter Merok, Ivana Milovanović, Nani Paananen, Vesa-Matti Paananen, Anna Pavlenko, Bruno Pereira Vidal, Luciano Strika, Yueh Tsao, Davide Turcato, Oleksandr Vakhno, Judit Velcsov, Anna Vickers, Stéphanie Visser, Herdyan Widarmanto, Andrey Zaikin, and Si-Qing Chen. Rtp-lx: Can llms evaluate toxicity in multilingual scenarios?, 2024. URL https://arxiv.org/abs/2404.14397.

Daryna Dementieva, Nikolay Babakov, and Alexander Panchenko. Multiparadetox: Extending text detoxification with parallel data to new languages. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 2: Short Papers)*, pp. 124–140, 2024a.

Daryna Dementieva, Daniil Moskovskiy, Nikolay Babakov, Abinew Ali Ayele, Naquee Rizwan, Frolian Schneider, Xintog Wang, Seid Muhie Yimam, Dmitry Ustalov, Elisei Stakovskii, Alisa Smirnova, Ashraf Elnagar, Animesh Mukherjee, and Alexander Panchenko. Overview of the multilingual text detoxification task at pan 2024. In Guglielmo Faggioli, Nicola Ferro, Petra Galuščáková, and Alba García Seco de Herrera (eds.), *Working Notes of CLEF 2024 - Conference and Labs of the Evaluation Forum*. CEUR-WS.org, 2024b.

Daryna Dementieva, Nikolay Babakov, Amit Ronen, Abinew Ali Ayele, Naquee Rizwan, Florian Schneider, Xintong Wang, Seid Muhie Yimam, Daniil Alekhseevich Moskovskiy, Elisei Stakovskii, Eran Kaufman, Ashraf Elnagar, Animesh Mukherjee, and Alexander Panchenko. Multilingual and explainable text detoxification with parallel corpora. In Owen Rambow, Leo Wanner, Marianna Apidianaki, Hend Al-Khalifa, Barbara Di Eugenio, and Steven Schockaert (eds.), *Proceedings of the 31st International Conference on Computational Linguistics*, pp. 7998–8025, Abu Dhabi, UAE, January 2025a. Association for Computational Linguistics. URL https://aclanthology.org/2025.coling-main.535/.

Daryna Dementieva, Nikolay Babakov, Amit Ronen, Abinew Ali Ayele, Naquee Rizwan, Florian Schneider, Xintong Wang, Seid Muhie Yimam, Daniil Alekhseevich Moskovskiy, Elisei Stakovskii, et al. Multilingual and explainable text detoxification with parallel corpora. In *Proceedings of the 31st International Conference on Computational Linguistics*, pp. 7998–8025, 2025b.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, et al. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024.

Emilio Ferrara. Should chatgpt be biased? challenges and risks of bias in large language models. *arXiv preprint arXiv:2304.03738*, 2023.

Samuel Gehman, Suchin Gururangan, Maarten Sap, Yejin Choi, and Noah A. Smith. RealToxicityPrompts: Evaluating neural toxic degeneration in language models. In Trevor Cohn, Yulan He, and Yang Liu (eds.), *Findings of the Association for Computational Linguistics: EMNLP 2020*, pp. 3356–3369, Online, November 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.findings-emnlp.301. URL https://aclanthology.org/2020.findings-emnlp.301.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*, 2021.

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

Aiqi Jiang and Arkaitz Zubiaga. Cross-lingual offensive language detection: A systematic review of datasets, transfer approaches and challenges, 2024. URL https://arxiv.org/abs/2401.09244.

Zachary Kenton, Tom Everitt, Laura Weidinger, Iason Gabriel, Vladimir Mikulik, and Geoffrey Irving. Alignment of language agents. *arXiv preprint arXiv:2103.14659*, 2021.

Md Tawkat Islam Khondaker, Muhammad Abdul-Mageed, and Laks Lakshmanan. Detoxllm: A framework for detoxification with explanations. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 19112–19139, 2024.

Hannah Rose Kirk, Yennie Jun, Filippo Volpin, Haider Iqbal, Elias Benussi, Frederic Dreyer, Aleksandar Shtedritski, and Yuki Asano. Bias out-of-the-box: An empirical analysis of intersectional occupational biases in popular generative language models. *Advances in neural information processing systems*, 34:2611–2624, 2021.

Hyukhun Koh, Dohyung Kim, Minwoo Lee, and Kyomin Jung. Can LLMs recognize toxicity? a structured investigation framework and toxicity metric. In Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen (eds.), *Findings of the Association for Computational Linguistics: EMNLP 2024*, pp. 6092–6114, Miami, Florida, USA, November 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.findings-emnlp.353. URL https://aclanthology.org/2024.findings-emnlp.353/.

Ben Krause, Akhilesh Deepak Gotmare, Bryan McCann, Nitish Shirish Keskar, Shafiq Joty, Richard Socher, and Nazneen Fatema Rajani. Gedi: Generative discriminator guided sequence generation. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pp. 4929–4952, 2021.

Andrew Lee, Xiaoyan Bai, Itamar Pres, Martin Wattenberg, Jonathan K Kummerfeld, and Rada Mihalcea. A mechanistic understanding of alignment algorithms: A case study on dpo and toxicity. In *International Conference on Machine Learning*, pp. 26361–26378. PMLR, 2024.

Chak Tou Leong, Yi Cheng, Jiashuo Wang, Jian Wang, and Wenjie Li. Self-detoxifying language models via toxification reversal. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 4433–4449, 2023.

Xiaochen Li, Zheng Xin Yong, and Stephen Bach. Preference tuning for toxicity mitigation generalizes across languages. In Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen (eds.), *Findings of the Association for Computational Linguistics: EMNLP 2024*, pp. 13422–13440, Miami, Florida, USA, November 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.findings-emnlp.784. URL https://aclanthology.org/2024.findings-emnlp.784/.

Alisa Liu, Maarten Sap, Ximing Lu, Swabha Swayamdipta, Chandra Bhagavatula, Noah A. Smith, and Yejin Choi. DExperts: Decoding-time controlled text generation with experts and anti-experts. In Chengqing Zong, Fei Xia, Wenjie Li, and Roberto Navigli (eds.), *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)*, pp. 6691–6706, Online, August 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.acl-long.522. URL https://aclanthology.org/2021.acl-long.522/.

John Morris, Eli Lifland, Jin Yong Yoo, Jake Grigsby, Di Jin, and Yanjun Qi. Textattack: A framework for adversarial attacks, data augmentation, and adversarial training in nlp. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pp. 119–126, 2020.

Daniil Moskovskiy, Daryna Dementieva, and Alexander Panchenko. Exploring cross-lingual text detoxification with large multilingual language models. In Samuel Louvan, Andrea Madotto, and Brielen Madureira (eds.), *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics: Student Research Workshop*, pp. 346–354, Dublin, Ireland, May 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.acl-srw.26. URL https://aclanthology.org/2022.acl-srw.26/.

Sourabrata Mukherjee, Akanksha Bansal, Atul Kr. Ojha, John P. McCrae, and Ondrej Dusek. Text detoxification as style transfer in English and Hindi. In Jyoti D. Pawar and Sobha

Lalitha Devi (eds.), *Proceedings of the 20th International Conference on Natural Language Processing (ICON)*, pp. 133–144, Goa University, Goa, India, December 2023. NLP Association of India (NLPAI). URL https://aclanthology.org/2023.icon-1.13/.

Vera Neplenbroek, Arianna Bisazza, and Raquel Fernández. Cross-lingual transfer of debiasing and detoxification in multilingual llms: An extensive investigation. *arXiv preprint arXiv:2412.14050*, 2024a.

Vera Neplenbroek, Arianna Bisazza, and Raquel Fernández. Cross-lingual transfer of debiasing and detoxification in multilingual llms: An extensive investigation. *arXiv preprint arXiv:2412.14050*, 2024b.

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

Maarten Sap, Dallas Card, Saadia Gabriel, Yejin Choi, and Noah A Smith. The risk of racial bias in hate speech detection. In *Proceedings of the 57th annual meeting of the association for computational linguistics*, pp. 1668–1678, 2019.

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

Guillermo Villate-Castillo, Javier Del Ser, and Borja Sanz Urquijo. A systematic review of toxicity in large language models: Definitions, datasets, detectors, detoxification methods and challenges. 2024.

Mengru Wang, Ningyu Zhang, Ziwen Xu, Zekun Xi, Shumin Deng, Yunzhi Yao, Qishen Zhang, Linyi Yang, Jindong Wang, and Huajun Chen. Detoxifying large language models via knowledge editing. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar (eds.), *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 3093–3118, Bangkok, Thailand, August 2024a. Association for Computational Linguistics. doi: 10.18653/v1/2024.acl-long.171. URL https://aclanthology.org/2024.acl-long.171/.

Wenxuan Wang, Zhaopeng Tu, Chang Chen, Youliang Yuan, Jen-tse Huang, Wenxiang Jiao, and Michael Lyu. All languages matter: On the multilingual safety of LLMs. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar (eds.), *Findings of the Association for Computational Linguistics: ACL 2024*, pp. 5865–5877, Bangkok, Thailand, August 2024b. Association for Computational Linguistics. doi: 10.18653/v1/2024.findings-acl.349. URL https://aclanthology.org/2024.findings-acl.349/.

Laura Weidinger, John Mellor, Maribeth Rauh, Conor Griffin, Jonathan Uesato, Po-Sen Huang, Myra Cheng, Mia Glaese, Borja Balle, Atoosa Kasirzadeh, et al. Ethical and social risks of harm from language models. *arXiv preprint arXiv:2112.04359*, 2021.

Albert Xu, Eshaan Pathak, Eric Wallace, Suchin Gururangan, Maarten Sap, and Dan Klein. Detoxifying language models risks marginalizing minority voices. In Kristina Toutanova, Anna Rumshisky, Luke Zettlemoyer, Dilek Hakkani-Tur, Iz Beltagy, Steven Bethard, Ryan Cotterell, Tanmoy Chakraborty, and Yichao Zhou (eds.), *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 2390–2397, Online, June 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.naacl-main.190. URL https://aclanthology.org/2021.naacl-main.190/.

Yang Xu, Yutai Hou, Wanxiang Che, and Min Zhang. Language anisotropic cross-lingual model editing. In *Findings of the Association for Computational Linguistics: ACL 2023*, pp. 5554–5569, 2023.

		Toxicity		Perplexity	
Model	**Split**	*X-FT*	*P/M-FT*	*X-FT*	*P/M-FT*
aya-expanse-8B	toxic-train	1	10	14	21
	toxic-test	3	11	15	22
	neutral-test	4	12	16	23
aya-23-8B	toxic-train	5	3	17	24
	toxic-test	6	13	18	25
	neutral-test	7	14	19	26
mt5-large	toxic-train	8	15	20	27
	toxic-test	9	16	21	28
	neutral-test	10	17	22	29
bloom-7B1	toxic-train	11	18	23	30
	toxic-test	12	19	24	31
	neutral-test	13	20	25	32

Table 2: Index table for all configurations over all models, data-splits, toxicity, and perplexity.

Linting Xue, Noah Constant, Adam Roberts, Mihir Kale, Rami Al-Rfou, Aditya Siddhant, Aditya Barua, and Colin Raffel. mT5: A massively multilingual pre-trained text-to-text transformer. In Kristina Toutanova, Anna Rumshisky, Luke Zettlemoyer, Dilek Hakkani-Tur, Iz Beltagy, Steven Bethard, Ryan Cotterell, Tanmoy Chakraborty, and Yichao Zhou (eds.), *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 483–498, Online, June 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.naacl-main.41. URL https://aclanthology.org/2021.naacl-main.41/.

Ahmet Üstün, Viraat Aryabumi, Zheng-Xin Yong, Wei-Yin Ko, Daniel D'souza, Gbemileke Onilude, Neel Bhandari, Shivalika Singh, Hui-Lee Ooi, Amr Kayid, Freddie Vargus, Phil Blunsom, Shayne Longpre, Niklas Muennighoff, Marzieh Fadaee, Julia Kreutzer, and Sara Hooker. Aya model: An instruction finetuned open-access multilingual language model. *arXiv preprint arXiv:2402.07827*, 2024.

A Appendix

A.1 Dataset Split

From the original set, we create our experimental splits by sampling 400 pairs, constructing a training set of 300 parallel pairs (*toxic-train* and *neutral-train*) and a test set of 100 pairs (*toxic-test* and *neutral-test*). We utilize the 300 *neutral-train* pairs to fine-tune and evaluate our hypothesis of cross-lingual detoxification using straightforward neutral samples. Further, the `textdetox/multilingual_paradetox` dataset[6] uses the *openrail*++ license[7].

A.2 Experimental Details

We fine-tune the models on the language generation task (as mentioned in Section 3.2 using the LoRA (Hu et al., 2021). We perform the hyperparameter search over batch size (4, 6, and 8), learning rate (2e-4 and 2e-5), rank (16 and 32), Lora-alpha (32 and 64), and epochs (20).

Our experimental setup comprises four learning paradigms across four multilingual LLMs, totaling 392 configurations: ***(1)*** zero-shot (*ZS*) evaluation across 7 languages, ***(2)*** cross-lingual fine-tuning (*X-FT*) with 81 language pairs, ***(3)*** partial fine-tuning (*P-FT*) with three

[6] https://huggingface.co/datasets/textdetox/multilingual_paradetox

[7] The Responsible AI License allows users to take advantage of the model in a wide range of settings (including free use and redistribution) as long as they respect the specific use case restrictions outlined, which correspond to model applications the licensor deems ill-suited for the model or are likely to cause harm.

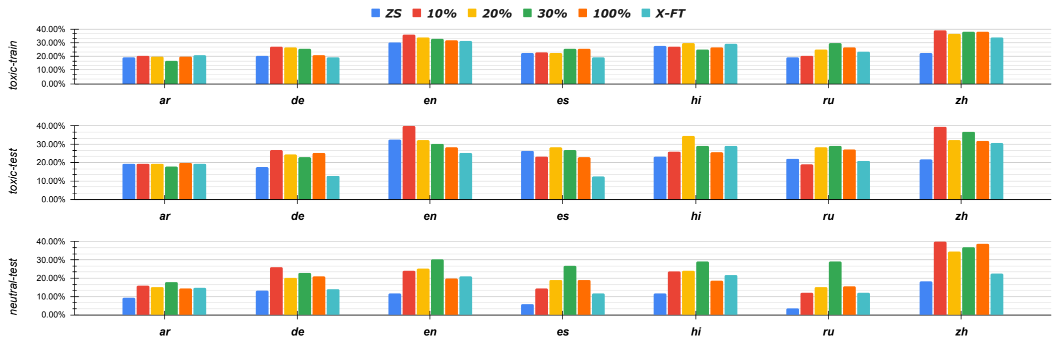

Figure 4: Toxicity scores for *ZS*, *X-FT*, *P-FT*, and *M-FT* for `aya-expanse-8B` over all three evaluation sets. ***Takeaway***: *Similar script family has shown similar behavior*.

data portions per language (27 configurations), and ***(4)*** multilingual fine-tuning (*M-FT*) across 7 languages.

A.3 Detoxification Analysis

We present the analysis of the cross-lingual transfer of detoxification in Table 2. We present the toxicity scores for ZS, *X-FT*, *P-FT*, and *M-FT* for all three evaluation sets for `aya-expanse-8B`, `mt5-large`, and `bloom-7B1`, in Figure 4, 5, 6, respectively. We observed that the detoxification is efficient in the high-resource languages ("*en*", "*es*", and "*de*"), and performed very poor for the languages with a very different script ("*zh*"). The models exhibited significant performance degradation on the *neutral-test* set following the implementation of learning strategies, resulting in elevated toxicity scores compared to *ZS* settings. We assume that the models might have learned the mapping of toxic and neutral samples.

A.4 Representation Analysis

We analyze the distribution of embeddings for toxic and neutral sentences across the dataset by computing their relative distances. Our analysis reveals how fine-tuning impacts these representations, demonstrating that embeddings from different scripts exhibit distinct patterns of distributional shift under various learning paradigms. As illustrated in Figure 7, while similar scripts initially demonstrate comparable embedding patterns in *ZS* setting, *M-FT* fine-tuning induces significant representational shifts that correlate with changes in model behavior in Figure 8. To quantify these distributional changes, we compute silhouette scores across the embedding space, with results presented in Figure 9, providing a metric for embedding cluster coherence across different models.

A.5 Perplexity Trade-Off

Tables 14, 15, 16 highlight the perplexity for `aya-expanse-8B` in *ZS* and *X-FT* settings for the *toxic-train*, *toxic-test*, and *neutral-train*, respectively. Overall, perplexity improved for high-to-mid-resource languages but failed for low-resource languages. This showed that detoxification affects the model's overall language generation capabilities.

A.6 Computation Requirement and Budget

The experiments are carried out on two NVIDIA Tesla A100 40 GB. The estimated cost to cover the computational requirements for two months, computed over GCP[8], is $5,523.14 per month x 1 month.

[8]The price for the VM is computed using the GCP Calculator: https://cloud.google.com/products/calculator.

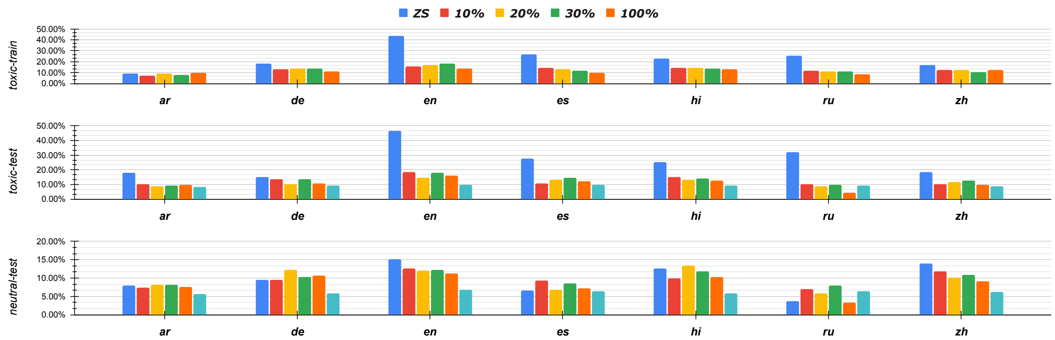

Figure 5: Toxicity scores for *ZS*, *P-FT*, and *M-FT* for `mt5-large` over all three evaluation sets. ***Takeaway***: *All the languages have shown significant low detoxification scores.*

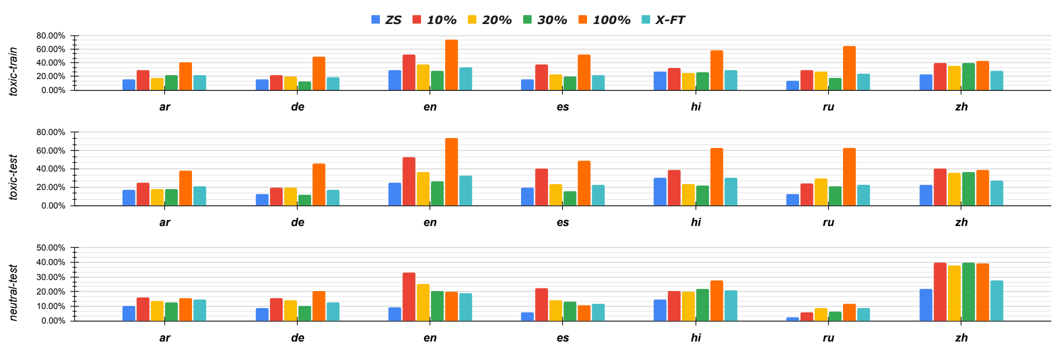

Figure 6: Toxicity scores for *ZS*, *P-FT*, and *M-FT* for `bloom-7B1` over all three evaluation sets. ***Takeaway***: `bloom-7B1` *has shown comparable results in* $X - FT$*, but worst in M-FT.*

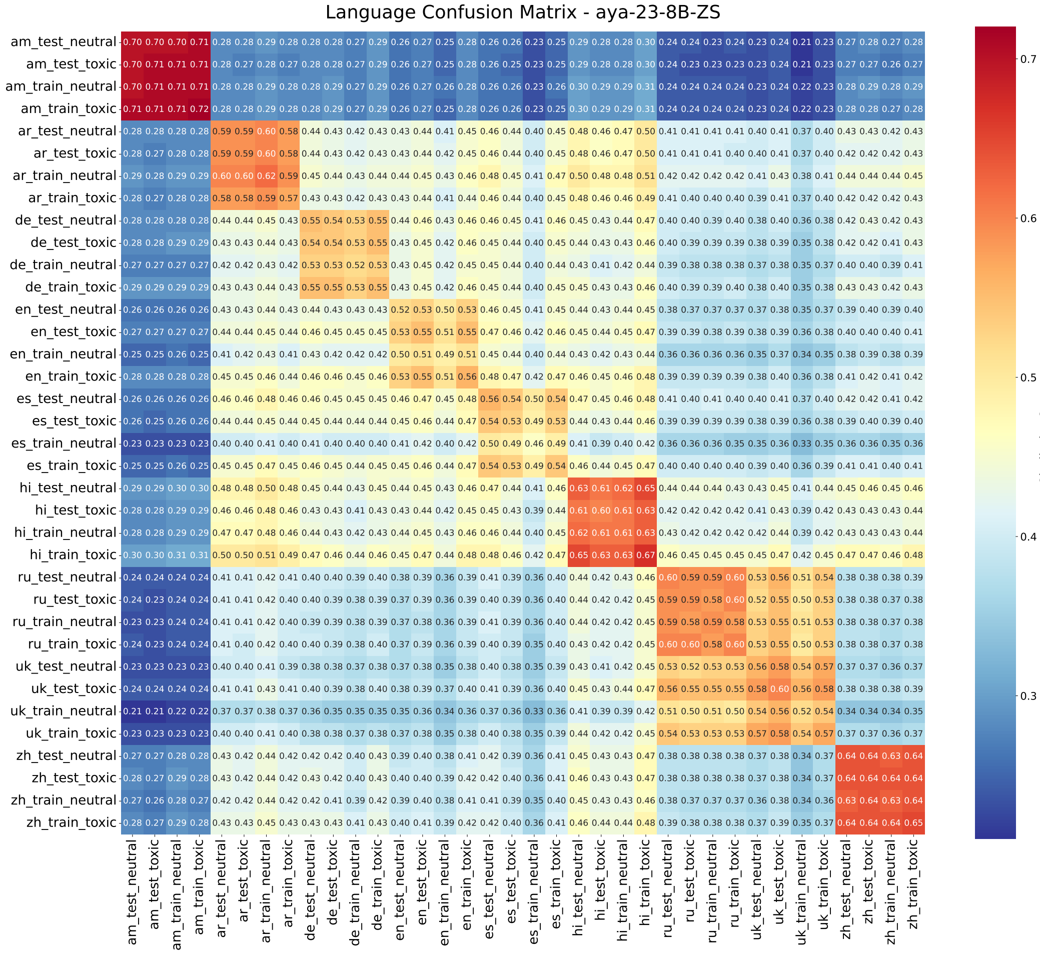

Figure 7: Confusion matrix over the distances between the embeddings of all nine languages from aya-23-8B over *ZS*. ***Takeaway***: *Languages with similar script tend to show a similar pattern.*

		am	ar	de	en	es	hi	ru	AVG
	ZS	19.37	**17.44**	32.68	26.51	23.14	22.25	21.82	23.32
X-FT (Δ)	**ar**	4.36	0.54	4.43	12.6	5.23	3.95	-5.81	3.61
	de	4.82	2.69	16	12.75	12.35	**9.18**	**10.25**	9.72
	en	-3.75	-6.6	0.85	3.66	-4.69	1.86	-11.55	-2.89
	es	**8.89**	**3.51**	**19.99**	**14.56**	**12.68**	8.04	6.54	**10.60**
	hi	-0.66	-11.95	-11.12	-5.59	-3.9	-8.2	-13.54	-7.85
	ru	3.66	0.34	3.28	2.54	-3.63	3.88	1.14	1.60
	zh	-2.04	-11.6	-13.29	-3.36	-15.89	-5.97	-18.41	-10.08
	AVG	2.18	-3.30	2.88	**5.31**	0.31	1.82	-4.48	

Table 3: Actual toxicity scores for *ZS vs* Δ-toxicity scores for *X-FT* for aya-expanse-8B over the *toxic-test* evaluation set. x represents the languages the model is trained on, while the languages on columns show the languages on which it is evaluated. AT_Z and Δ_{AVG} represent the average toxicity in *ZS* and average Δ-toxicity scores for *X-FT*. **Bold** represents the best scores. ***Takeaway***: *"es" is supposed to be best language to train on and also does not get affected, and reflect best detoxification scores.*

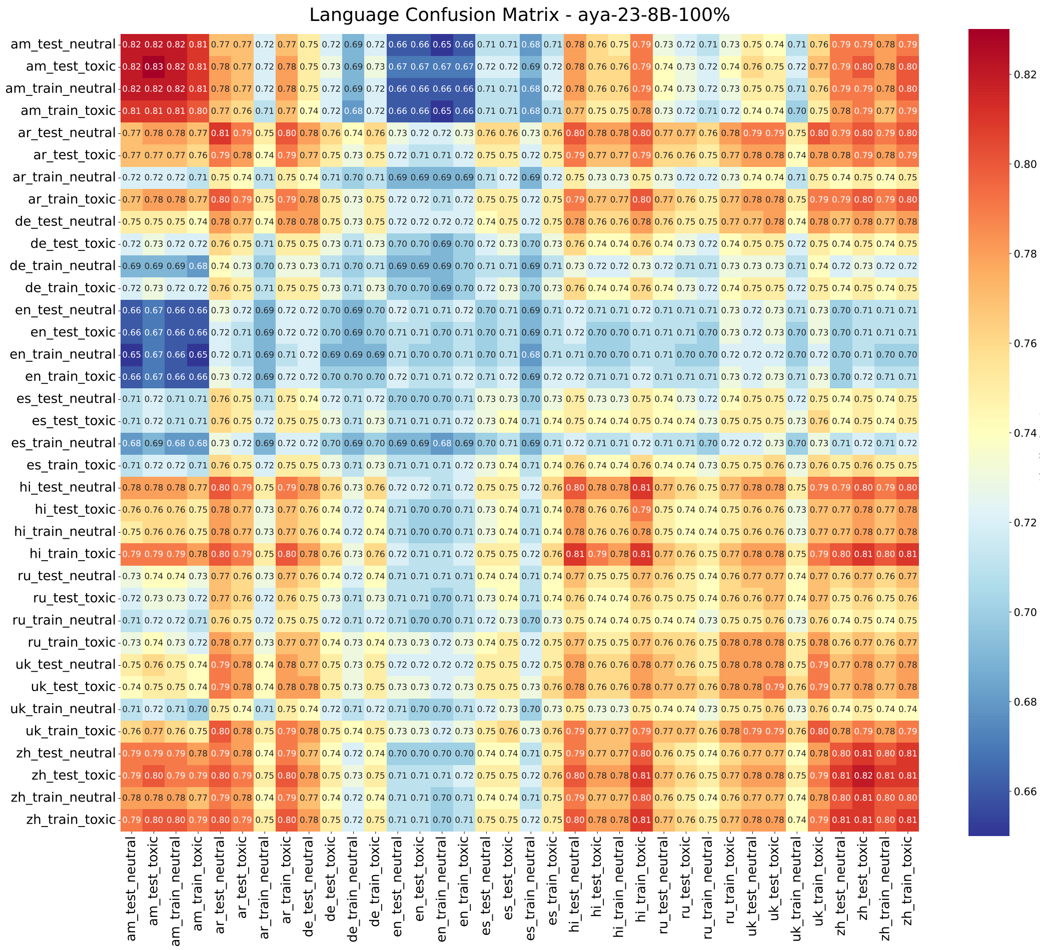

Figure 8: Confusion matrix over the distances between the embeddings of all nine languages from `aya-23-8B` over *M-FT*. ***Takeaway***: *Languages with similar script tend to show a similar pattern*.

Figure 9: Silhouette scores for different models over the combined average scores over the entire train and test set. ***Takeaway***: *Both the* `aya` *models tend to show similar behavior. However, we observe higher negative scores for Chinese in* `mT5-large`.

		am	ar	de	en	es	hi	ru	AVG
	ZS	9.31	13.22	11.66	5.75	11.77	**3.5**	18.08	10.47
X-FT (Δ)	**ar**	-3.73	-2.08	-7.31	-7.62	-0.79	-4.96	-6.01	-4.64
	de	-2.5	-4.8	-5.93	-10.17	-0.46	-11.23	**5.85**	-4.18
	en	-8.75	-11.46	-11.91	-13.53	-7.21	-9.9	-11.17	-10.56
	es	**-1.12**	0.62	**0.69**	-5	**2.74**	-7.22	0.92	**-1.20**
	hi	-2.42	-6.76	-18.16	-15.39	-14.21	-13.9	-14.51	-12.19
	ru	-1.36	**0.87**	-1.67	**-2.35**	-1.17	**-2.03**	-2.57	-1.47
	zh	-5.76	-12.03	-16.03	-13.97	-12.73	-12.4	-19.47	-13.20
	AVG	-3.66	-5.09	-8.62	**-9.72**	-4.83	-8.81	-6.71	

Table 4: Actual toxicity scores for *ZS vs* Δ-toxicity scores for *X-FT* for `aya-expanse-8B` over the *neutral-test* evaluation set. ***Takeaway***: *Detoxification adversely effects the model's general knowledge.*

		am	ar	de	en	es	hi	ru	AVG
	ZS	19.86	19.95	33.17	20.79	25.09	**18.75**	23.1	22.96
X-FT (Δ)	**ar**	2.46	2.68	7.54	1.64	-4.14	-1.8	-5.45	0.42
	de	7.12	-0.65	14.91	7.8	11.95	-0.02	3.78	6.41
	en	0.39	-4.07	6.81	2.78	-2.65	-0.13	-8.74	-0.80
	es	**10.53**	**9.93**	**20.39**	**8.73**	**13.05**	**5.18**	**9.93**	**11.11**
	hi	1.08	-7.96	2.56	-3.38	-4.68	-4.53	-8.48	-3.63
	ru	1.02	-1.85	-3.23	-3.46	-0.66	-3.45	-1.25	-1.84
	zh	-2.4	-11.53	-14.22	-9.84	-9.45	-8.81	-13.02	-9.90
	AVG	2.89	-1.92	**4.97**	0.61	0.49	-1.94	-3.32	

Table 5: Actual toxicity scores for *ZS vs* Δ-toxicity scores for *X-FT* for `aya-23-8B` over the *toxic-train* evaluation set. ***Takeaway***: *Surprisingly "zh" shows that irrespective of fine-tuning language, the detoxification scores actually increases.*

		am	ar	de	en	es	hi	ru	AVG
	ZS	**16.31**	20.85	38.48	20.15	24.76	19.07	21.68	23.04
X-FT (Δ)	**ar**	-0.27	2.27	15.67	1.15	-2.37	-3.88	-5.6	1.00
	de	3.76	0.85	12.97	7.62	9.52	-2.62	1.01	4.73
	en	-1.87	-3.51	10.66	1.76	-5.66	1.83	-5.12	-0.27
	es	**7.22**	**11.28**	**27.76**	**7.02**	**13.27**	**5.32**	**7.13**	**11.29**
	hi	-1.5	-5.02	6.99	-0.3	-8.12	-5.36	-7.95	-3.04
	ru	-4.17	0.95	1.7	-1.64	-2.64	-4.75	1.37	-1.31
	zh	-2.5	-9.02	-4.17	-13.57	-15.3	-10.2	-14.83	-9.94
	AVG	0.10	-0.31	**10.23**	0.29	-1.61	-2.81	-3.43	

Table 6: Actual toxicity scores for *ZS vs* Δ-toxicity scores for *X-FT* for `aya-23-8B` over the *toxic-test* evaluation set. ***Takeaway***: *"es" showed the best average detoxification scores.*

		am	ar	de	en	es	hi	ru	AVG
	ZS	11.5	11.9	13.08	9.47	13.17	**5.22**	22.61	12.42
X-FT (Δ)	**ar**	-2.28	-1.23	-9.14	-6.17	-5.86	-5.91	-5.07	-5.09
	de	-1.67	-6.39	-5.3	-3.4	-0.64	-12.95	3.4	-3.85
	en	-4.81	-8.4	-8.5	-5.21	-8.42	-6.28	-4.54	-6.59
	es	**3.96**	**2**	**2.05**	-2.64	**2.03**	-5.21	**8.5**	**1.53**
	hi	-1.11	-9.66	-11.88	-6.56	-13.28	-6.7	-5.96	-7.88
	ru	2.73	0.15	1.89	**1.28**	-2.45	**1.03**	-0.86	0.54
	zh	-4.18	-12.58	-20.8	-12.21	-14.49	-9.61	-13.47	-12.48
	AVG	**-1.05**	-5.16	-7.38	-4.99	-6.16	-6.52	-2.57	

Table 7: Actual toxicity scores for *ZS vs* Δ-toxicity scores for *X-FT* for `aya-23-8B` over the *neutral-test* evaluation set. ***Takeaway***: *Detoxification adversely effects the model's general knowledge.*

		am	ar	de	en	es	hi	ru	AVG
	ZS	17.94	18.12	43.58	26.6	23.06	25.39	**16.86**	24.51
X-FT (Δ)	**ar**	10.92	7.27	27.16	16	13.26	15.89	8.74	14.18
	de	9.7	9.47	29.25	15.62	12.93	16.09	9.09	14.59
	en	10.6	8.63	27.82	17.17	13.38	15.86	**9.45**	14.70
	es	11.04	9.52	28.64	16.85	10.86	**17.62**	8.48	14.72
	hi	10.9	**9.74**	30.2	15.75	13.21	15.85	8.51	**14.88**
	ru	11.01	7.13	29.4	**17.4**	13	14.49	8.48	14.42
	zh	**11.16**	7.8	**30.68**	14.91	**13.76**	15.12	8.78	14.60
	AVG	10.76	8.51	**29.02**	16.24	12.91	15.85	8.79	

Table 8: Actual toxicity scores for *ZS vs* Δ-toxicity scores for *X-FT* for `mt5-large` over the *toxic-train* evaluation set. ***Takeaway***: `mt5-large` *showed better detoxification scores in all languages but showed a trade-off with general perplexity scores.*

		am	ar	de	en	es	hi	ru	AVG
	ZS	18.12	**15.25**	46.74	27.83	25.21	32.26	18.38	26.26
X-FT (Δ)	**ar**	10.23	6.53	**35**	18.8	15.73	**23.9**	**11.75**	**17.42**
	de	10.91	4.69	32.02	16.38	14.97	23.86	10.78	16.23
	en	10.84	5.01	30.22	16	**18.08**	23.69	9.56	16.20
	es	9.4	6.61	31.27	16.81	15.38	22.08	9.69	15.89
	hi	10.15	**8.41**	32.58	14.42	16.81	23.08	10.9	16.62
	ru	11	5.52	30.59	**19.28**	15.24	21.55	11.33	16.36
	zh	**11.6**	7.98	33.43	14.93	15.96	23.66	10.44	16.86
	AVG	10.59	6.39	**32.16**	16.66	16.02	23.12	10.64	

Table 9: Actual toxicity scores for *ZS vs* Δ-toxicity scores for *X-FT* for `mt5-large` over the *toxic-test* evaluation set. ***Takeaway***: `mt5-large` *showed better detoxification scores in all languages but showed a trade-off with general perplexity scores.*

		am	ar	de	en	es	hi	ru	AVG
	ZS	8.03	9.44	15.14	6.66	12.54	**3.74**	13.89	9.92
X-FT (Δ)	**ar**	2.73	2.31	8.29	**1.43**	6.13	-0.74	7.95	**4.01**
	de	2.49	1.19	7.48	0.82	**6.68**	**-0.02**	**8.17**	3.83
	en	2.74	0.92	7.58	-1.25	4.88	-1.3	6.54	2.87
	es	1.46	**2.33**	7.61	0.47	5.96	-1.84	6.03	3.15
	hi	**2.87**	1.54	8.21	0.73	6.39	-0.52	8.71	3.99
	ru	2.12	1.14	**8.73**	-0.49	5.99	-1.36	6.23	3.19
	zh	1.7	2.06	8.16	1.31	6.44	-1.92	6.76	3.50
	AVG	2.30	1.64	**8.01**	0.43	6.07	-1.10	7.20	

Table 10: Actual toxicity scores for *ZS vs* Δ-toxicity scores for *X-FT* for `mt5-large` over the *neutral-test* evaluation set. ***Takeaway***: *Detoxification does not adversely effects the model's general knowledge but effected the overall perplexity meanwhile.*

		am	ar	de	en	es	hi	ru	AVG
	ZS	15.53	15.2	28.75	15.51	27.37	**13.9**	23.03	19.90
X-FT (Δ)	**ar**	-5.8	3.91	13.11	-0.84	7.3	-2.35	4.29	2.80
	de	2.74	5.25	11.3	**4.95**	9.47	**6.54**	9.11	7.05
	en	-0.39	-3.42	1.58	-12.43	6.91	-2.74	11.12	0.09
	es	**3.11**	**6.59**	12.14	3.81	**16.47**	4.26	**12.25**	**8.38**
	hi	-5.25	-2.93	9.36	-9.1	-6.45	-9.33	-1.13	-3.55
	ru	0.98	5.49	15.42	2.38	13.32	6.51	8.25	7.48
	zh	-0.93	2.19	**15.49**	0.06	14.91	0.13	5.04	5.27
	AVG	-0.79	2.44	**11.20**	-1.60	8.85	0.43	6.99	

Table 11: Actual toxicity scores for *ZS vs* Δ-toxicity scores for *X-FT* for `bloom-7B1` over the *toxic-train* evaluation set. ***Takeaway***: *"es" comes up as the best fine-tuning language than "en" and "de" from other models.*

		am	ar	de	en	es	hi	ru	AVG
	ZS	17.17	**12.25**	24.51	19.69	30.28	12.46	22.77	19.88
X-FT (Δ)	**ar**	-3.74	0.44	7.99	2.2	10.63	-4.16	2.86	2.32
	de	2.96	1.46	8.08	**7.72**	13.28	**4.52**	**11.88**	7.13
	en	1.48	-6.04	-3.78	-15.4	14.43	-2.5	9.17	-0.38
	es	**5.65**	3.37	7.18	7.56	**18.68**	2.98	10.54	**7.99**
	hi	-2.73	-6.02	3.93	-1.83	-3.14	-11.53	-1.71	-3.29
	ru	1.79	**3.4**	9.49	6.65	17.56	3.83	10.78	7.64
	zh	1.37	-1.04	**10.76**	3.58	17.86	-0.61	4.56	5.21
	AVG	0.97	-0.63	6.24	1.50	**12.76**	-1.07	6.87	

Table 12: Actual toxicity scores for *ZS vs* Δ-toxicity scores for *X-FT* for `bloom-7B1` over the *toxic-test* evaluation set. ***Takeaway***: *"hi" was least effected by the fine-tuning.*

		am	ar	de	en	es	hi	ru	AVG
	ZS	10.34	8.69	9.36	5.96	14.38	**2.21**	22.01	10.42
X-FT (Δ)	**ar**	-9.23	-2.27	-5.94	-8.22	-4.17	-14.93	4.13	-5.80
	de	-2.94	-0.32	**-0.14**	**-3.42**	-1.34	-5.6	7.83	-0.85
	en	-5.43	-8.67	-12.39	-13.95	-2.73	-14.16	10.4	-6.70
	es	**-0.21**	0.03	-7.56	-4.53	**5.04**	-5.37	**12.31**	**-0.04**
	hi	-11.58	-9.67	-9.93	-17.85	-16.13	-21.51	1.35	-12.19
	ru	-6.05	**1.06**	-3.22	-6.12	0.94	**-4.59**	6.74	-1.61
	zh	-5.14	-4.99	-3.82	-8.76	0.79	-12.18	7.04	-3.87
	AVG	-5.80	-3.55	-6.14	-8.98	-2.51	-11.19	**7.11**	

Table 13: Actual toxicity scores for *ZS vs* Δ-toxicity scores for *X-FT* for `bloom-7B1` over the *neutral-test* evaluation set. ***Takeaway***: *Detoxification adversely effects the model's general knowledge.*

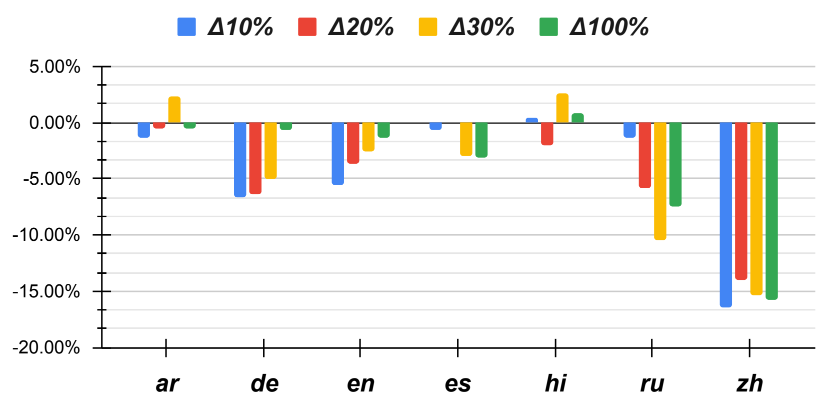

Figure 10: AverageΔ-Toxicity scores for Percent-based Fine-Tuning (*P-FT*) *vs* Multilingual Fine-Tuning (*M-FT*) for `aya-expanse-8B` over the *toxic-train* evaluation set. 10%, 20%, 30%, and 100% represents the Average Δ-Toxicity in *P-FT* and *M-FT* settings. ***Takeaway***: *P-FT and* $M-FT$ *did not showed significant detoxification scores.*

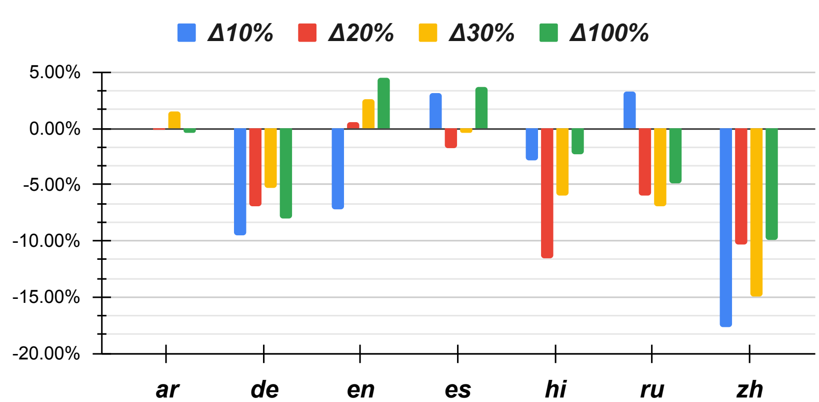

Figure 11: AverageΔ-Toxicity scores for *P-FT vs M-FT* for `aya-expanse-8B` over the *toxic-test* evaluation set. ***Takeaway***: *We observed significant scores in "en" and "es", but the scores did not showed any improvement in "zh".*

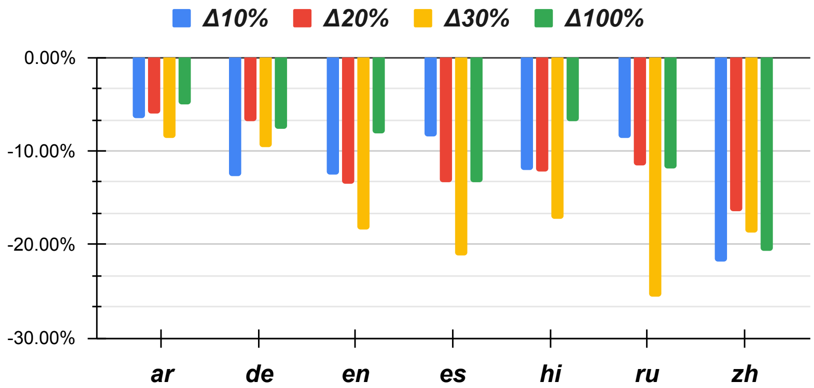

Figure 12: AverageΔ-Toxicity scores for *P-FT vs M-FT* for `aya-expanse-8B` over the *neutral-test* evaluation set. ***Takeaway***: *All the languages were adversely affected.*

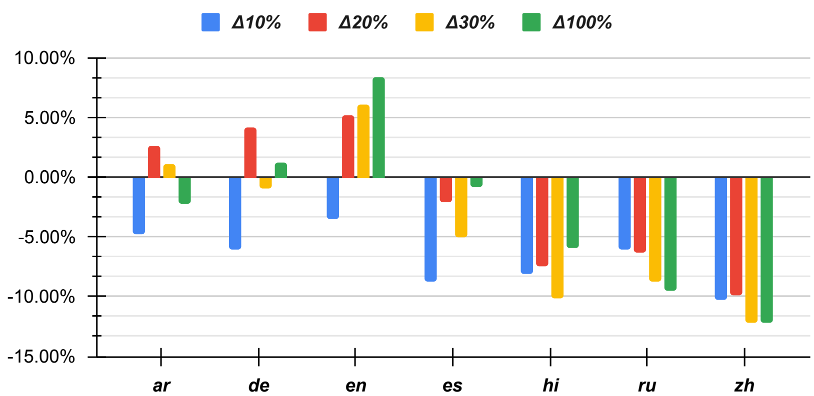

Figure 13: AverageΔ-Toxicity scores for *P-FT vs M-FT* for `aya-23-8B` over the *toxic-test* evaluation set. ***Takeaway***: *"en" and "de" showed significant update however other showed adversarial effects.*

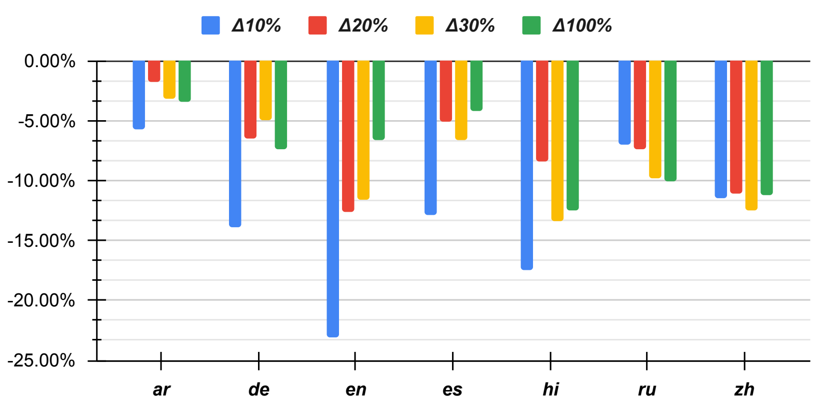

Figure 14: AverageΔ-Toxicity scores for *P-FT vs M-FT* for `aya-23-8B` over the *neutral-test* evaluation set. ***Takeaway***: *All the languages were adversely affected.*

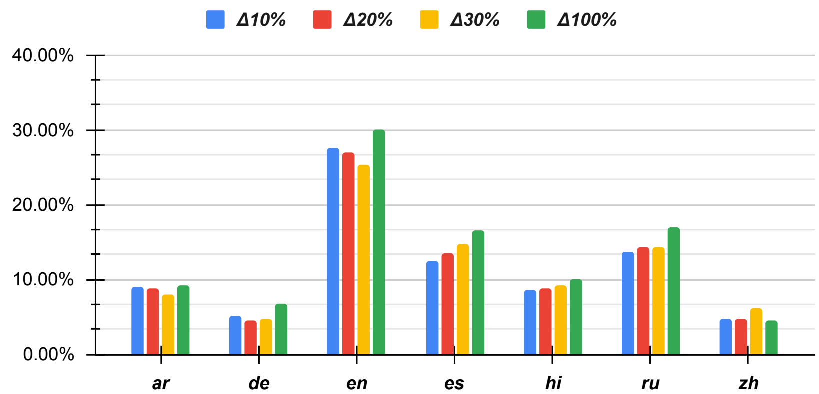

Figure 15: AverageΔ-Toxicity scores for *P-FT vs M-FT* for `mt5-large` over the *toxic-train* evaluation set. ***Takeaway***: *All languages showed significant updates.*

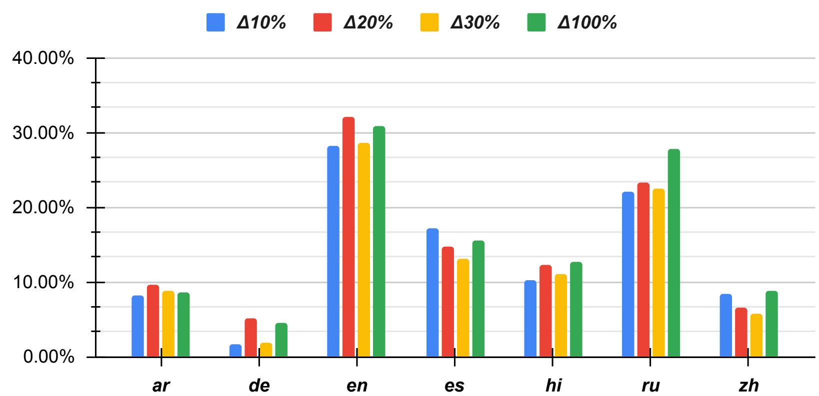

Figure 16: AverageΔ-Toxicity scores for *P-FT vs M-FT* for `mt5-large` over the *toxic-test* evaluation set. ***Takeaway***: *All languages showed significant updates.*

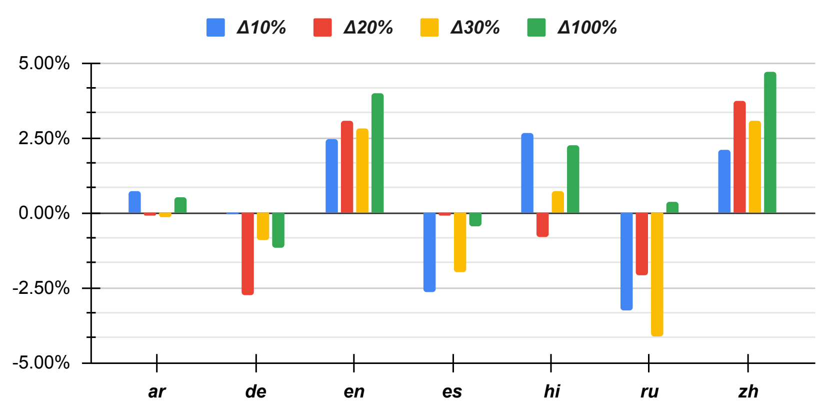

Figure 17: AverageΔ-Toxicity scores for *P-FT vs M-FT* for `mt5-large` over the *neutral-test* evaluation set. ***Takeaway***: *"en", "hi", and "zh" showed significant updates.*

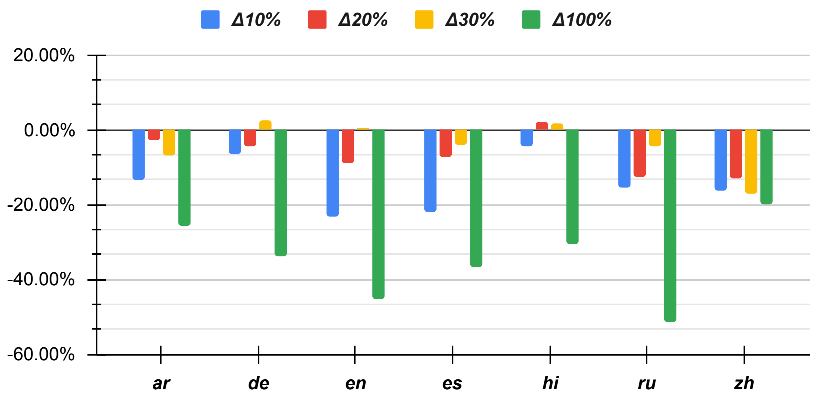

Figure 18: AverageΔ-Toxicity scores for *P-FT vs M-FT* for `bloom-7B1` over the *toxic-train* evaluation set. ***Takeaway***: *All the languages were adversely affected.*

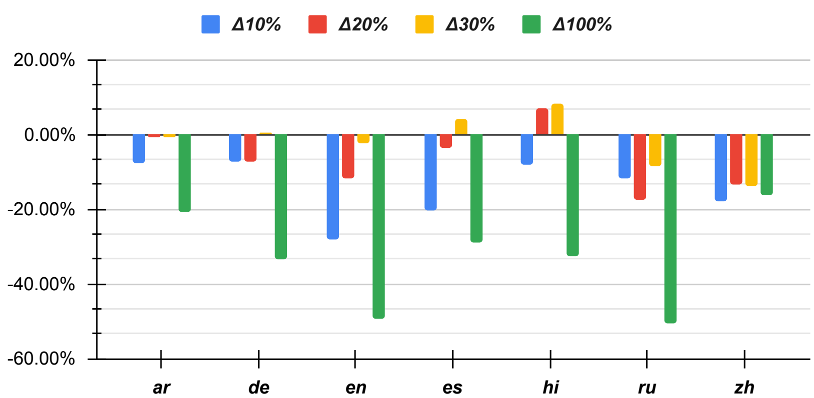

Figure 19: AverageΔ-Toxicity scores for *P-FT vs M-FT* for `bloom-7B1` over the *toxic-test* evaluation set. ***Takeaway***: *All the languages were adversely affected.*

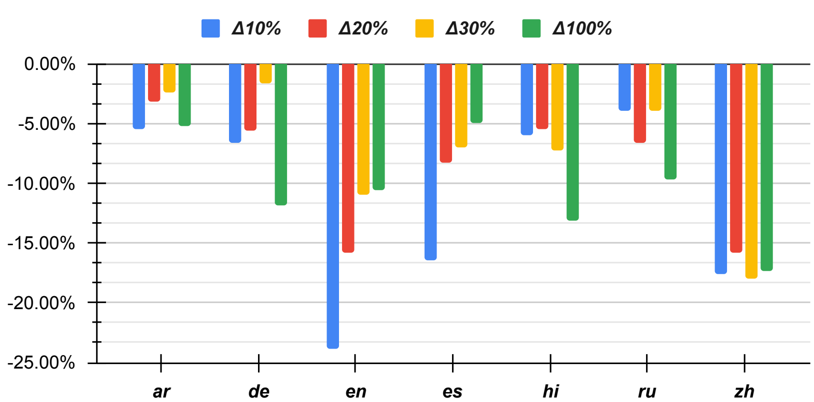

Figure 20: AverageΔ-Toxicity scores for *P-FT vs M-FT* for `bloom-7B1` over the *neutral-test* evaluation set. ***Takeaway***: *All the languages were adversely affected.*

		am	ar	de	en	es	hi	ru	AVG
	ZS	13.72	78.92	21.53	91.79	**08.51**	09.75	25.39	35.66
X-FT (Δ)	**ar**	**03.54**	15.29	-03.18	20.21	**02.13**	**01.31**	09.09	6.91
	de	-03.70	**49.98**	-07.90	**74.08**	-06.28	-09.80	07.83	**14.89**
	en	-02.45	00.06	-25.74	-03.47	-11.52	-12.80	04.73	-7.31
	es	-90.50	-45.04	-104.81	-32.81	-91.97	-145.96	-89.64	-85.82
	hi	01.96	-01.72	**-02.44**	-23.61	01.64	00.01	04.35	-2.83
	ru	00.13	-03.67	-02.55	-10.40	-00.03	-00.64	01.59	-2.22
	zh	03.30	05.69	-05.55	-01.61	01.90	**00.70**	**10.92**	2.19
	AVG	-12.53	2.94	-21.74	**3.20**	-14.88	-23.88	-7.30	

Table 14: Actual perplexity scores for Zero-Shot (*ZS*) *vs* Δ-perplexity scores for Cross-lingual Fine-Tuning (*X-FT*) for `aya-expanse-8B` over the *toxic-train* evaluation set. x represents the languages the model is trained on, while the languages on columns show the languages on which it is evaluated. AP_Z and Δ_{AVG} represent the average perplexity in *ZS* and average Δ-perplexity scores for *X-FT*. **Bold** represents the best scores. ***Takeaway***: *"hi" and "ru" was most affected irrespective of fine-tuning language.*

		am	ar	de	en	es	hi	ru	AVG
	ZS	12.92	73.57	23.10	97.75	**08.92**	10.01	24.59	35.84
X-FT (Δ)	**ar**	03.16	06.00	**01.95**	30.50	**02.21**	**01.45**	07.34	7.51
	de	-02.35	**47.01**	-03.17	**79.20**	-09.69	-08.09	**10.49**	**16.20**
	en	-00.89	-07.90	-19.47	11.33	-16.04	-25.89	04.10	-7.82
	es	-90.23	-56.42	-81.76	-37.73	-87.72	-123.65	-78.28	-79.40
	hi	00.56	-25.17	-01.85	-18.14	01.81	00.80	01.90	-5.73
	ru	-02.57	-07.78	-00.47	07.24	-00.25	00.37	01.27	-0.31
	zh	**03.47**	01.81	-02.37	-05.84	02.19	01.38	08.81	1.35
	AVG	-12.69	-6.07	-15.31	**9.51**	-15.35	-21.95	-6.34	

Table 15: Actual perplexity scores for *ZS* *vs* Δ-perplexity scores for *X-FT* for `aya-expanse-8B` over the *toxic-test* evaluation set. ***Takeaway***: *"hi" and "ru" was most affected irrespective of fine-tuning languages.*

		am	ar	de	en	es	hi	ru	AVG
	ZS	14.42	86.83	17.74	80.80	09.25	**08.76**	27.27	35.01
X-FT (Δ)	**ar**	03.82	20.04	-05.17	04.98	01.90	**00.70**	11.55	5.40
	de	-03.98	**55.67**	-11.19	**55.02**	-10.89	-16.92	09.43	**11.02**
	en	-01.50	11.32	-23.79	-16.35	-06.56	-10.23	04.77	-6.05
	es	-123.02	-62.81	-89.56	-60.30	-72.23	-108.47	-99.29	-87.95
	hi	01.65	10.42	-07.51	-15.38	02.19	-00.51	07.25	-0.27
	ru	01.25	06.14	**-01.40**	-12.77	01.40	-00.75	05.47	-0.09
	zh	**04.54**	08.79	-05.10	-08.45	**02.72**	-00.18	**12.67**	2.14
	AVG	-16.75	**7.08**	-20.53	-7.61	-11.64	-19.48	-6.88	

Table 16: Actual perplexity scores for *ZS* *vs* Δ-perplexity scores for *X-FT* for `aya-expanse-8B` over the *neutral-test* evaluation set. ***Takeaway***: *Detoxification adversely effects the model's general knowledge.*

		am	ar	de	en	es	hi	ru	AVG
	ZS	11.13	65.72	17.33	72.11	**06.64**	08.59	17.83	28.48
X-FT (Δ)	**ar**	-00.41	**24.04**	-01.53	10.05	**00.85**	00.27	04.11	**5.34**
X-FT (Δ)	**de**	-06.78	22.25	-15.79	**49.94**	-10.35	-11.09	**06.87**	5.01
X-FT (Δ)	**en**	-01.83	10.25	-28.82	-17.17	-08.79	-01.69	-08.06	-8.02
X-FT (Δ)	**es**	-43.57	18.54	-41.71	04.26	-56.92	-46.59	-18.38	-26.34
X-FT (Δ)	**hi**	02.23	08.65	**03.87**	-06.27	00.60	00.70	04.64	2.06
X-FT (Δ)	**ru**	-00.62	-00.74	-00.95	-06.20	00.44	-00.46	00.95	-1.08
X-FT (Δ)	**zh**	**03.60**	-02.03	-03.13	-11.71	00.74	**00.88**	03.32	-1.19
	AVG	-6.77	**11.57**	-12.58	3.27	-10.49	-8.28	-0.94	

Table 17: Actual perplexity scores for *ZS vs* Δ-perplexity scores for *X-FT* for aya-23-8B over the *toxic-train* evaluation set. ***Takeaway***: *"es" turned out to be least affected by other fine-tuning languages*.

		am	ar	de	en	es	hi	ru	AVG
	ZS	12.08	68.14	16.55	68.30	**06.58**	08.28	16.87	28.11
X-FT (Δ)	**ar**	-09.46	14.21	-00.76	06.04	**01.31**	00.46	02.61	2.06
X-FT (Δ)	**de**	-02.24	24.49	-25.33	**50.29**	-11.82	-21.44	01.12	**2.15**
X-FT (Δ)	**en**	-02.20	15.99	-27.50	-07.68	-09.48	00.68	-50.35	-11.51
X-FT (Δ)	**es**	-47.85	**26.86**	-49.98	04.28	-54.92	-45.01	-29.21	-27.98
X-FT (Δ)	**hi**	02.89	10.06	**03.17**	-09.85	00.25	**01.44**	**04.08**	1.72
X-FT (Δ)	**ru**	00.77	03.26	-02.17	-09.59	-00.44	-00.28	-00.88	-1.33
X-FT (Δ)	**zh**	**04.95**	01.33	-05.47	-22.70	00.73	00.98	02.41	-2.54
	AVG	-7.59	**13.74**	-15.43	1.54	-10.62	-9.02	-10.03	

Table 18: Actual perplexity scores for *ZS vs* Δ-perplexity scores for *X-FT* for aya-23-8B over the *toxic-test* evaluation set. ***Takeaway***: *"de" turned out to be least affected by other fine-tuning languages*.

		am	ar	de	en	es	hi	ru	AVG
	ZS	11.99	58.39	15.98	67.60	**08.28**	08.33	14.71	26.47
X-FT (Δ)	**ar**	01.58	**18.62**	-05.83	-01.41	02.46	00.31	**02.41**	**2.59**
X-FT (Δ)	**de**	-09.37	11.97	-20.57	**42.05**	-11.18	-27.10	01.12	-1.87
X-FT (Δ)	**en**	-00.73	-05.29	-36.57	-02.88	-25.59	00.66	-05.16	-10.79
X-FT (Δ)	**es**	-55.74	05.86	-46.06	-13.17	-64.98	-42.36	-19.74	-33.74
X-FT (Δ)	**hi**	02.65	-02.18	-03.69	-10.54	01.80	00.94	00.11	-1.56
X-FT (Δ)	**ru**	00.51	-02.50	**01.25**	-10.46	01.28	00.06	-05.69	-2.22
X-FT (Δ)	**zh**	**04.43**	-00.65	-05.08	-16.57	**02.58**	**01.06**	01.28	-1.85
	AVG	-8.10	**3.69**	-16.65	-1.85	-13.38	-9.49	-3.67	

Table 19: Actual perplexity scores for *ZS vs* Δ-perplexity scores for *X-FT* for aya-23-8B over the *neutral-test* evaluation set. ***Takeaway***: *Detoxification adversely effects the model's general knowledge*.

		am	ar	de	en	es	hi	ru	AVG
	ZS	20.83	418.68	102.44	160.72	37.47	39.66	**20.30**	114.30
X-FT (Δ)	**ar**	-41.82	290.83	-33.35	**13.65**	-41.03	-15.91	-50.45	17.42
	de	-48.42	293.50	**-15.34**	08.04	**-20.01**	-36.17	-49.62	**18.85**
	en	-44.97	258.00	-40.79	-15.34	-21.91	-18.27	-56.07	8.66
	es	-59.50	**303.54**	-41.25	-05.44	-27.81	-28.49	-59.03	11.72
	hi	**-41.29**	297.44	-59.69	07.66	-26.10	**-13.63**	-57.27	15.30
	ru	-69.79	292.02	-33.23	-16.72	-23.68	-19.51	**-39.82**	12.76
	zh	-49.62	274.36	-52.52	-11.92	-27.73	-21.72	-47.56	9.04
	AVG	-50.77	**287.10**	-39.45	-2.87	-26.90	-21.96	-51.40	

Table 20: Actual perplexity scores for *ZS vs* Δ-perplexity scores for *X-FT* for `mt5-large` over the *toxic-train* evaluation set. ***Takeaway***: *"en" turned out to be least affected by other fine-tuning languages*.

		am	ar	de	en	es	hi	ru	AVG
	ZS	24.23	662.17	89.41	152.72	20.73	**16.72**	25.23	141.60
X-FT (Δ)	**ar**	-44.33	534.77	-39.18	13.34	-66.87	-46.71	-35.97	45.01
	de	-53.80	541.97	-58.83	-41.43	-63.42	-66.64	-35.50	31.77
	en	-43.34	563.65	-22.98	-14.47	-61.43	-70.40	-33.34	45.38
	es	-67.00	553.54	-39.59	**20.49**	-113.59	**-24.31**	**-21.70**	43.98
	hi	-40.83	**581.54**	-41.06	-30.84	-57.98	-43.64	-32.63	**47.80**
	ru	**-35.19**	575.01	-59.19	-48.22	**-46.61**	-53.70	-36.89	42.17
	zh	-46.94	475.82	**-19.73**	-02.19	-68.55	-50.16	-30.65	36.80
	AVG	-47.35	**546.61**	-40.08	-14.76	-68.35	-50.79	-32.38	

Table 21: Actual perplexity scores for *ZS vs* Δ-perplexity scores for *X-FT* for `mt5-large` over the *toxic-test* evaluation set. ***Takeaway***: *"hi" and "ru" was most affected irrespective of fine-tuning languages*.

		am	ar	de	en	es	hi	ru	AVG
	ZS	17.79	195.62	69.68	142.01	**13.79**	17.05	32.32	69.75
X-FT (Δ)	**ar**	-40.49	**92.62**	-91.01	-47.79	-85.73	-73.79	-33.37	-39.94
	de	-48.82	60.18	-84.11	-27.92	-79.20	-58.86	-55.94	-42.10
	en	**-31.83**	51.71	-80.42	-25.18	-55.99	-55.87	-50.75	-35.47
	es	-51.49	63.91	-79.88	19.69	-46.90	-72.75	**-25.39**	-27.54
	hi	-73.29	30.61	-64.19	17.28	-105.15	-30.51	-53.38	-39.80
	ru	-85.55	66.24	-97.59	**32.90**	-40.96	**-28.45**	-26.81	-25.75
	zh	-67.22	69.90	**-28.78**	21.68	**-38.72**	-29.45	-40.33	**-16.13**
	AVG	-56.96	**62.17**	-75.14	-1.33	-64.66	-49.95	-40.85	

Table 22: Actual perplexity scores for *ZS vs* Δ-perplexity scores for *X-FT* for `mt5-large` over the *neutral-test* evaluation set. ***Takeaway***: *Detoxification adversely effects the model's general knowledge*.

		am	ar	de	en	es	hi	ru	AVG
	ZS	06.69	2259.35	**04.01**	16.44	12.06	564.88	04.82	409.75
X-FT (Δ)	**ar**	-04.23	2250.85	-08.09	08.15	-01.69	552.11	-03.24	399.12
	de	-83.18	2116.93	-76.53	-88.91	-382.00	479.28	-67.29	271.18
	en	01.22	2255.43	-05.20	10.95	04.74	557.96	-01.99	403.30
	es	-56.11	2220.91	-08.01	11.14	-11.81	526.42	-33.18	378.48
	hi	-04.21	2249.98	-03.23	**13.54**	**07.91**	559.85	-04.38	402.78
	ru	-391.78	2187.84	-260.01	-34.54	-445.47	358.98	-107.10	186.85
	zh	**02.82**	**2255.48**	**00.28**	12.68	07.65	**560.67**	**00.90**	**405.78**
	AVG	-76.50	**2219.63**	-51.54	-9.57	-117.24	513.61	-30.90	

Table 23: Actual perplexity scores for *ZS vs* Δ-perplexity scores for *X-FT* for `bloom-7B1` over the *toxic-train* evaluation set. ***Takeaway***: *All the languages were adversely affected*.

		am	ar	de	en	es	hi	ru	AVG
	ZS	**23.57**	114.45	44.15	145.82	159.41	314.27	187.03	141.24
X-FT (Δ)	ar	13.30	104.03	34.39	135.05	152.16	302.90	177.97	131.40
	de	14.90	84.26	06.48	133.17	75.15	218.62	149.80	97.48
	en	17.24	108.49	40.33	137.65	153.79	307.12	179.98	134.94
	es	-06.20	108.73	-143.29	**142.12**	151.53	29.67	**184.90**	66.78
	hi	13.55	103.51	35.75	136.49	**156.85**	310.41	177.81	133.48
	ru	-77.36	-199.05	-10.32	128.23	28.62	-388.32	172.26	-49.42
	zh	**19.55**	**111.16**	**40.13**	142.39	152.46	**310.17**	183.87	**137.10**
	AVG	-0.72	60.16	0.50	136.44	124.37	155.80	**175.23**	

Table 24: Actual perplexity scores for *ZS vs* Δ-perplexity scores for *X-FT* for `bloom-7B1` over the *toxic-test* evaluation set. ***Takeaway***: *All the languages were adversely affected.*

		am	ar	de	en	es	hi	ru	AVG
	ZS	25.56	49.35	37.93	126.36	136.29	**24.47**	145.75	77.96
X-FT (Δ)	ar	15.81	35.69	28.49	116.88	128.85	13.48	137.82	68.15
	de	17.69	-23.17	-07.59	120.28	128.74	-15.38	129.21	49.97
	en	20.08	44.56	33.89	118.54	130.26	17.29	138.87	71.93
	es	21.65	40.64	07.44	115.52	131.65	-01.93	125.84	62.97
	hi	17.12	37.76	25.26	118.70	**133.48**	20.43	134.71	69.64
	ru	-98.82	-43.55	-123.96	116.03	71.87	-319.17	134.09	-37.64
	zh	**22.05**	**45.86**	**34.25**	**122.96**	132.43	**20.53**	**142.10**	**74.31**
	AVG	2.22	19.68	-0.32	118.42	122.47	-37.82	**134.66**	

Table 25: Actual perplexity scores for *ZS vs* Δ-perplexity scores for *X-FT* for `bloom-7B1` over the *neutral-test* evaluation set. ***Takeaway***: *Detoxification adversely effects the model's general knowledge.*

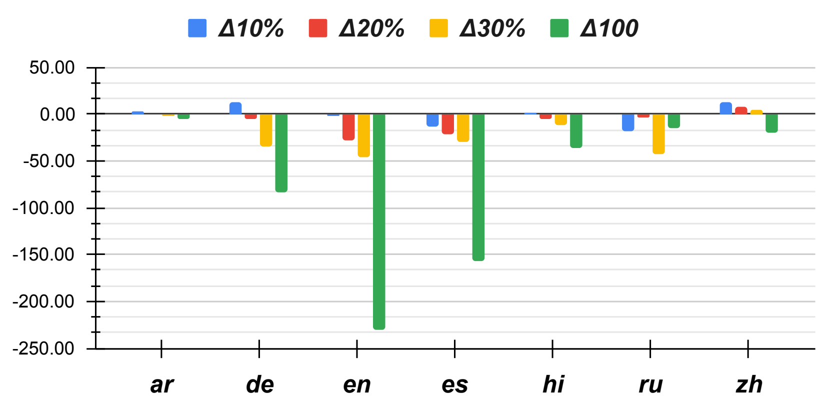

Figure 21: AverageΔ-Perplexity scores for Percent-based Fine-Tuning (*P-FT*) *vs* Multilingual Fine-Tuning (*M-FT*) for `aya-expanse-8B` over the *toxic-train* evaluation set. 10%, 20%, 30%, and 100% represents the Average Δ-Perplexity in *P-FT* and *M-FT* settings. ***Takeaway***: *The 100%-FT showed adverse effects in "en" and "es".*

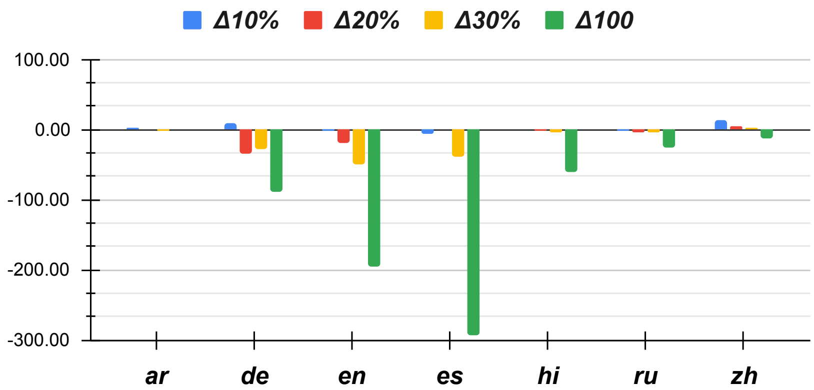

Figure 22: AverageΔ-Perplexity scores for *P-FT vs M-FT* for aya-expanse-8B over the *toxic-test* evaluation set. ***Takeaway***: *The* 100%*-FT showed adverse effects in "en" and "es".*

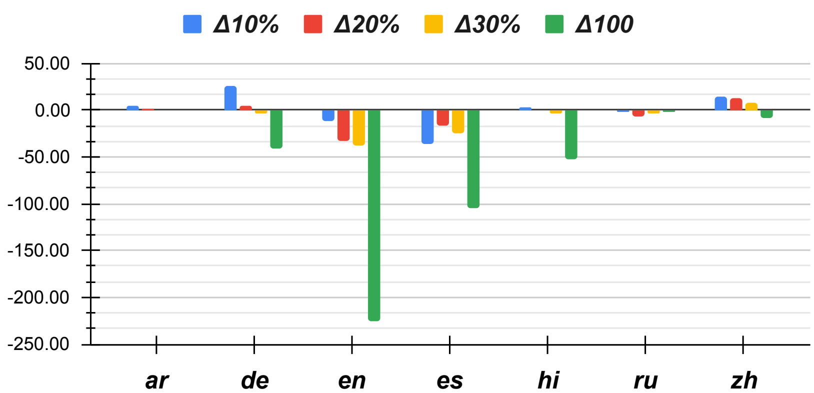

Figure 23: AverageΔ-Perplexity scores for *P-FT vs M-FT* for aya-expanse-8B over the *neutral-test* evaluation set. ***Takeaway***: *The* 100%*-FT showed adverse effects in "en" and "es".*

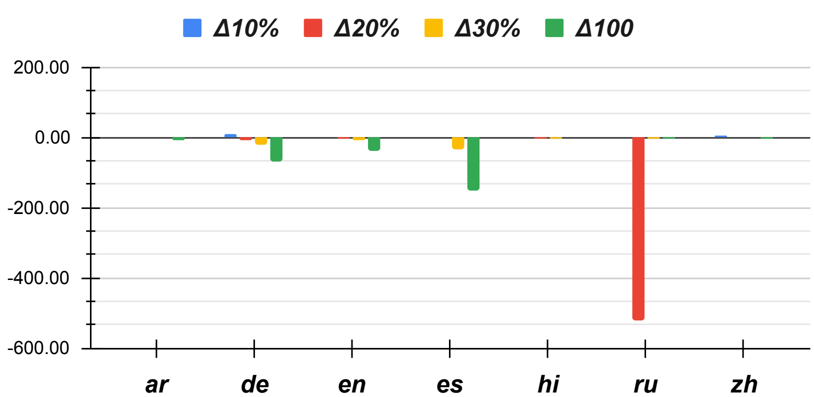

Figure 24: AverageΔ-Perplexity scores for *P-FT vs M-FT* for aya-23-8B over the *toxic-train* evaluation set. ***Takeaway***: *The* 100%*-FT showed adverse effects in "en" and "es" and* 20% *in "zh".*

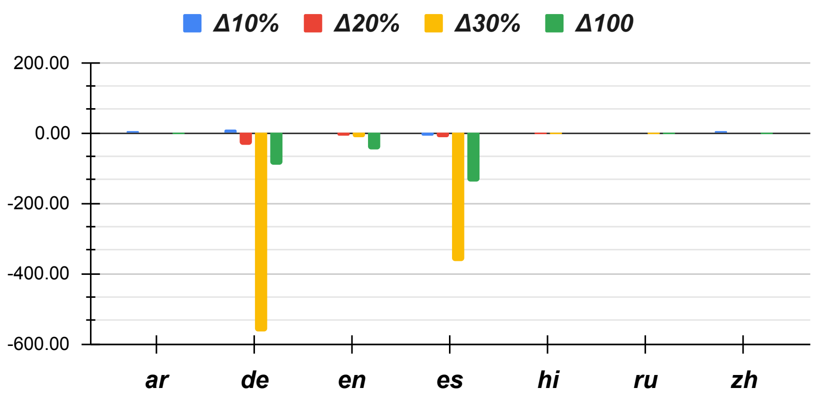

Figure 25: AverageΔ-Perplexity scores for *P-FT vs M-FT* for `aya-23-8B` over the *toxic-test* evaluation set. ***Takeaway***: *The* 30%*-FT showed adverse effects in "de" and "es".*

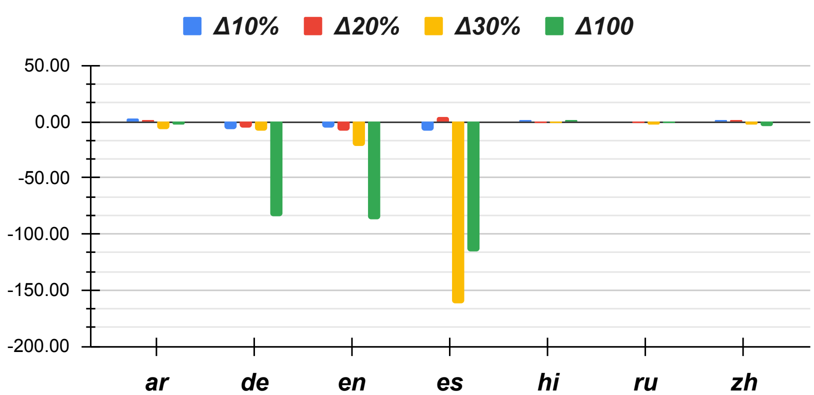

Figure 26: AverageΔ-Perplexity scores for *P-FT vs M-FT* for `aya-23-8B` over the *neutral-test* evaluation set. ***Takeaway***: *The* 100%*-FT showed adverse effects in "en" and "es", and* 30% *in "es".*

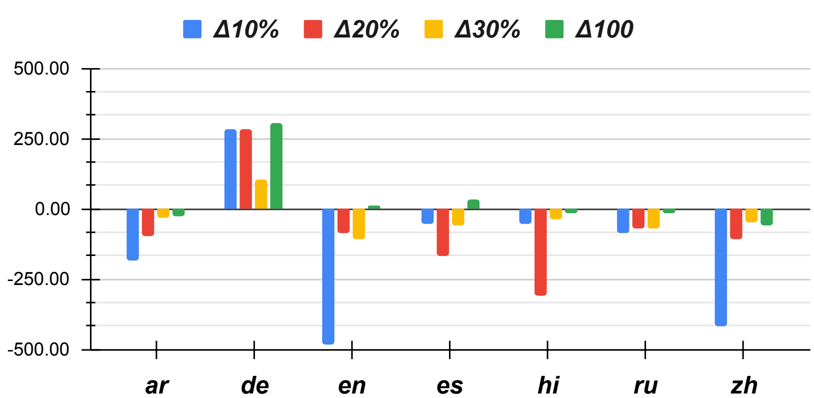

Figure 27: AverageΔ-Perplexity scores for *P-FT vs M-FT* for `mt5-large` over the *toxic-train* evaluation set. ***Takeaway***: *All the languages were adversely affected except "de".*

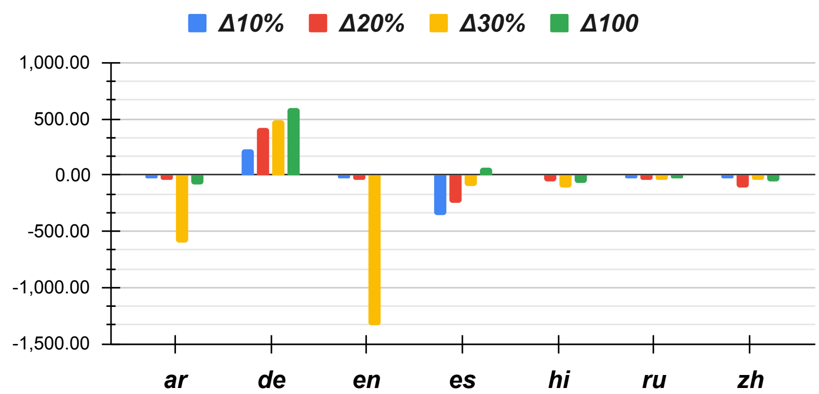

Figure 28: AverageΔ-Perplexity scores for *P-FT vs M-FT* for `mt5-large` over the *toxic-test* evaluation set. ***Takeaway***: *The* 30%*-FT showed adverse effects in "en".*

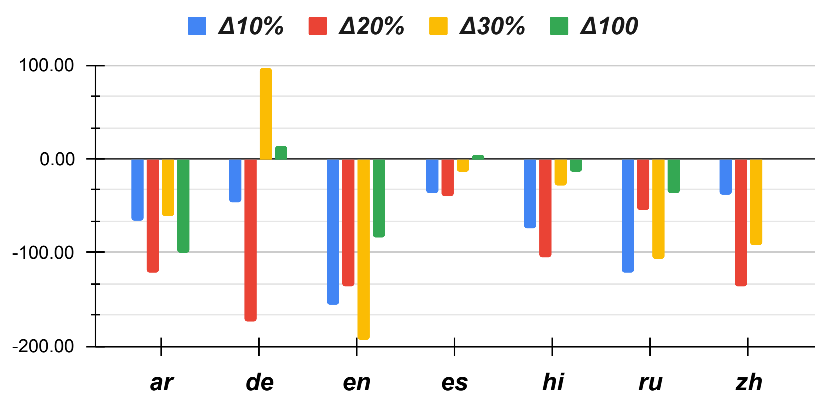

Figure 29: AverageΔ-Perplexity scores for *P-FT vs M-FT* for `mt5-large` over the *neutral-test* evaluation set. ***Takeaway***: *All the languages were adversely affected.*

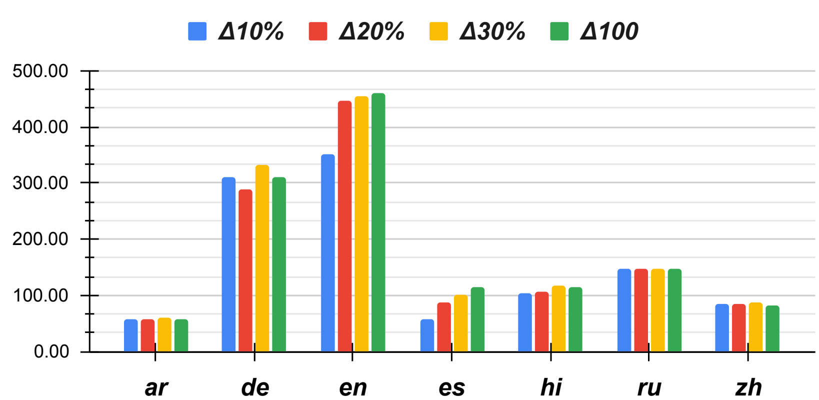

Figure 30: AverageΔ-Perplexity scores for *P-FT vs M-FT* for `bloom-7B1` over the *toxic-train* evaluation set. ***Takeaway***: *All the languages were not adversely affected except "de" in* 10%*.*

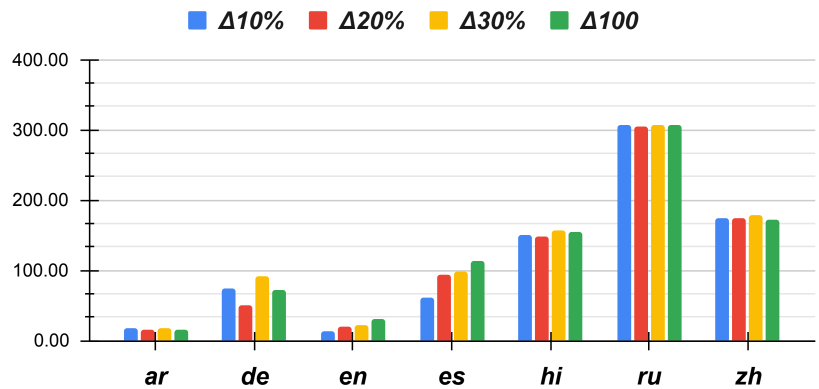

Figure 31: AverageΔ-Perplexity scores for *P-FT vs M-FT* for `bloom-7B1` over the *toxic-test* evaluation set. ***Takeaway***: *All the languages showed significant scores.*

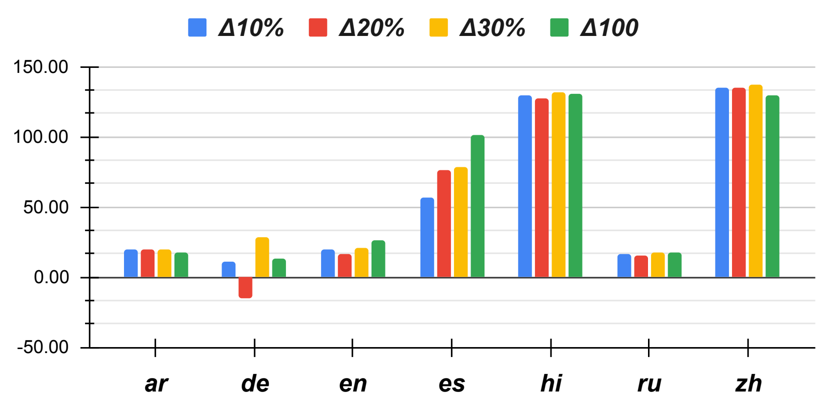

Figure 32: AverageΔ-Perplexity scores for *P-FT vs M-FT* for `bloom-7B1` over the *neutral-test* evaluation set. ***Takeaway***: *All the languages showed significant scores.*