MICo: Preventative Detoxification of Large Language Models through Inhibition Control

Roy Siegelmann[1]* **Ninareh Mehrabi**[2] **Palash Goyal**[2]
Prasoon Goyal[2] **Lisa Bauer**[2] **Jwala Dhamala**[2] **Aram Galstyan**[2]
Rahul Gupta[2] **Reza Ghanadan**[2]
[1]Johns Hopkins University
[2]Amazon AGI Foundations

Abstract

Large Language Models (LLMs) are powerful tools which have been both dominant and commonplace in the field of Artificial Intelligence. Yet, LLMs have a tendency to devolve into toxic degeneration, wherein otherwise safe and unproblematic models begin generating toxic content. For the sake of social responsibility and inspired by the biological mechanisms of inhibition control, we introduce the paradigm of Education for Societal Norms (ESN). By collecting and labeling examples as acceptable and unacceptable (in this case toxic and non-toxic), and including a corresponding acceptable rewrite with every unacceptable example, we introduce a new mechanism for LLM detoxification. We annotate a dataset of 2,850 entries and use it to fine-tune a model, which we call a Model with Inhibition Control (MICo). Evaluating this model on toxicity detection capability, rewrite detoxification, meaning preservation, and overall toxicity reduction, we discover significant improvements over the baseline model. In our experiments we show that overall toxicity of this model is more than 60% reduced, with over 75% reduction in severe toxicity.

1 Introduction

Large Language Models (LLMs) are trained with the explicit purpose of serving humans, between providing information, presenting an engaging chat partner, and answering any number of other user requests. Unfortunately, for a variety of reasons, there is a tendency for models to descend into neural toxic degeneration, outputting toxic and otherwise harmful messages (Faal et al., 2022; Xu et al., 2022; Wang et al., 2023). Naturally, toxic prompts very commonly yield toxic responses, but many prompts which are entirely non-toxic also yield toxic responses (Gehman et al., 2020; Gururangan et al., 2022; Hartvigsen et al., 2022). Currently, there are three main directions being used to combat toxicity, each with a considerable drawback.

First, the model may classify prompts as either toxic or non-toxic, and categorically refuse to respond to those deemed toxic (Xu et al., 2021). However, these approaches oftentimes use templated sentences to refuse to respond to toxic content which can degrade user engagement and lower helpfulness of the model (Xu et al., 2021). It is also possible for even entirely non-toxic prompts to yield toxic generations, and thus this method falls short of truly detoxifying.

Second, the model can be trained solely on non-toxic data (Welbl et al., 2021a; Gururangan et al., 2020). However, toxic content can be produced even from training data which appears benign. Entirely purifying any word which may lead to toxicity will leave the training corpus narrow, impacting the richness of content which can be generated. Furthermore, the model would not know how to respond to prompts which include some toxicity; thus, generations based around these prompts will be nonsensical, slashing the model's utility.

Third, an external classifier or a secondary model (e.g., in decoding time approaches) can be used to detect whether the model generated toxic content, and if so stop the content from reaching the user, instructing the model to provide another generation instead (Mehrabi et al., 2022; Liu et al., 2021; Krause et al., 2021; Dathathri et al., 2019). However, without an understanding of toxicity, the model is likely to continuously generate toxic content, yielding potentially unbounded latency times. This would prove an impediment to the successful and user-friendly utilization of the slower text-based output and become unmanageable for models utilizing rapid speech-based communication.

Learning from these drawbacks, we formulate the following three requirements of a successful solution: (i) Assuring non-toxic responses to non-toxic prompts. (ii) Responding to toxic prompts in a natural, yet non-toxic manner. (iii) Minimizing toxicity in real-time to prevent latency. Despite the fact that AI has not yet provided a satisfactory solution, humans exhibit these traits due to their inherent ability of self reflection and inhibition control. Healthy, mature humans consider the consequences of their speech (and more generally, behavior) via self-reflection before engaging in dialogue, and if determined to be negative, will alter the output to contain similar meaning yet eliminate the negative outcome. This necessitates a true understanding of what is deemed

* This work was done when Roy Siegelmann was an intern at Amazon. Correspondence to rsiege15@jhu.edu and mninareh@amazon.com.

Findings of the Association for Computational Linguistics: NAACL 2024, pages 1696–1703

to be negative, i.e. toxic, and acquiring the ability to express oneself without causing harms to others.

Toward this goal, in this work, we propose the paradigm of Education for Societal Norms (ESN), which builds on acknowledging LLMs' capability to conduct few-shot learning (Brown et al., 2020). To identify toxic versus non-toxic generations, we provide examples of both (Park and Rudzicz, 2022), followed by their correct labels. Detoxification is taught by appending a non-toxic meaning-preserving rewrite after each toxic generation. This way, each experience entry in the ESN ends up with the desired non-toxic label. We collect these experiences and fine-tune the model on this dataset. We demonstrate that averaging over all prompts, the output produced by our model reduces toxic generation by over 60% and severe toxicity by over 75%. We refer to models trained with this method as Models with Inhibition Control (MICo).

Our contributions are the following:

- Introducing a new training paradigm by which LLMs can develop the crucial skill of inhibition control, which can be generalized to topics far beyond toxicity, presented in Section 2.
- Filtering and annotating a novel dataset for use in detoxification of LLMs, presented in Section 2.
- Designing and conducting experiments based on numerical results of MICo, compared with the baseline models, presented in Section 3.

2 Education for Social Norms Methodology and Evaluation

Education for Social Norms Setup. We introduce an education paradigm that appears as a dataset of experiences which can teach the LLM the two capabilities required for inhibition control simultaneously: identifying toxic vs. non-toxic texts and substituting toxic generation with meaning preserving non-toxic text. The dataset's entries take one of two forms: (a) A non-toxic content followed by the non-toxic label, "XXX [NTX]", or (b) A paired toxic content and its corresponding non-toxic rewrite along with their associated labels, e.g. "YYY [TX] XXX [NTX]" (XXX and YYY stand for non-toxic and toxic text respectively). Education stems from first learning to label the text correctly according to its toxicity, and then assuring that each experience entry ends with the desired behavior and its [NTX] label. It is important that the NTX rewrites carry the same semantics, as otherwise the education would be functionally useless in terms of preserving meaning.

Fine-tuning Dataset. We construct a dataset in a semi-automatic approach based on Jigsaw's Toxic Comment Classification Challenge dataset[1], which is composed of comments from Wikipedia's talk page edits, as a base. For detoxification, we began with Falcon40b, which is known to be among the most powerful models for its size (Almazrouei et al., 2023). We discovered it was best at detoxification when using this instruction-tuned model closer to a traditional LLM, i.e. requesting sentence completion. We received an average of one satisfactory detoxification per three generations, and as such, we asked for four rewrites of each toxic comment. We then filtered by human annotation to select the best detoxification, and if none was found, generated one manually. We then took non-toxic comments along with the toxic/non-toxic rewrite pairs, and arranged them into the form discussed above. It is interesting to note that our database has only 2,850 entries and that this seems sufficient for teaching inhibition control by fine-tuning to the entries. We experimented with different ratio of non-toxic to toxic entries in the dataset ratios, and we found the 9:1 ratio of non-toxic to toxic to be most successful in reaching the educational goals.

Evaluation. For evaluation, we select RealToxicityPrompts, a dataset of close to 100k prompts, both toxic and non-toxic, which have been shown to yield high toxicity in generations of sentence-completion models, and is commonly used for toxicity evaluation (Ouyang et al., 2022; Askell et al., 2021; Chung et al., 2022; Touvron et al., 2023). For proof of concept, we test it on relatively small LLMs and fine-tuned for two epochs. GPT-Neo (125M parameters) learned to append '[TX]' or '[NTX]' but with a significant preference to [TX] and completely missed the rewrite component. GPT-Neo (1.3B parameters) reduced toxicity in rewrites but only very slightly As smaller models provided insufficient for the complicated task, we settle on OPT's 6.7 billion parameter model (Zhang et al., 2022), which was large enough to be educated. It was trained with the 9:1 non-toxic to toxic education dataset using QLORA (Dettmers et al., 2023). As with the original Jigsaw dataset, we consider the toxicity score generated by PerspectiveAPI to be the ground truth for toxicity (although we found inaccuracy in this method as well (Welbl et al., 2021b)). For evaluation, we provide two different toxicity axes. Binary toxicity is the option of rounding the toxicity so that below 0.5 is considered non-toxic whereas at or above 0.5 is considered toxic. We find this analysis to be too sensitive to the hard boundary and hence not that informative. We thus mainly consider graded-toxicity, separating the text entries to five bins according to the labels: Most non-toxic (0-0.1), non-toxic (0.1-0.35), partially/potentially toxic (0.35-0.65), toxic (0.65-0.90), and severely toxic at (0.9-1). These ranges provided higher robustness.

3 Experimental Results and Analysis

There are five important criteria which our method should fulfill.

[1] https://www.kaggle.com/c/jigsaw-toxic-comment-classification-challenge

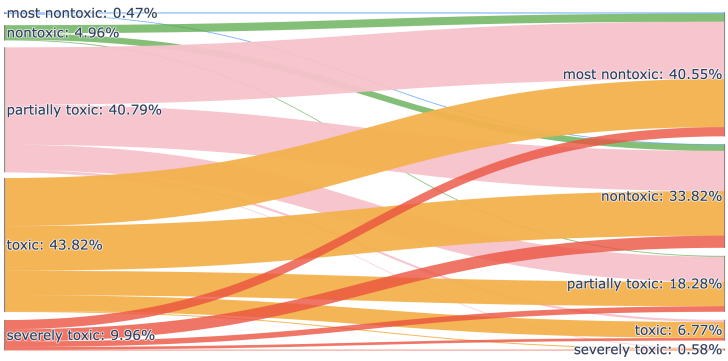

Figure 1: Toxicity scores according to PerspectiveAPI before (left) and after (right) rewriting for comments flagged as toxic by MICo. After detoxification, a plurality from each category becomes highly non-toxic, with decreasing percentages going to increasingly toxic categories. Our detoxification provides 94.2% less severe toxicity, 84.6% less toxicity, and $82\times$ more comments in the highly non-toxic category.

3.1 Toxicity Detection: Does the model properly detect toxicity?

Prior to detoxifying, the model needs to be able to separate toxic from non-toxic content. Since everything detected as toxic will be rewritten, the detection capabilities impact the overall detoxification capabilities (He et al., 2023). Our model's detection capabilities are summarized in Table 1. Most mistakes in labeling came from borderline sentences, which qualitatively appear neither fully toxic nor fully nontoxic, see example in Table 2 in the Appendix. Many such mistakes come from particularly multi-sentence comments, with which PerspectiveAPI also seems to struggle (see example in Table 3 in the Appendix).

	API Toxic	API Non-Toxic
MICo Toxic	6.36%	1.57%
MICo Non-Toxic	5.74%	86.33%

Table 1: Testing MICo's realtime toxicity detection against PerspectiveAPI exhibits a good detection of non-toxic content and worse detection of toxic content. The binary classification exhibits a high accuracy (0.927) and specificity (0.938), middling precision (0.802), and a relatively low sensitivity.

3.2 Rewrite Improvement: Are the detoxified rewrites less toxic than their source texts?

The core feature of inhibition control is pairing each negative example with a positive example. In our case, this is rewriting toxic content as non-toxic. Our model's rewrite improvements are shown in Figure 1. Improvement can be better measured when considering more detailed scores, and hence measured with the graded-toxicity coordinates. The results show that less than 5.5% of comments flagged by MICo as toxic are highly non-toxic or non-toxic according to PerspectiveAPI, which demonstrates that under the graded-toxicity system, our detection capabilities are quite aligned with PerspectiveAPI. Less than 7.5% of comments are toxic or severely toxic according to PerspectiveAPI after detoxification, demonstrating powerful detoxification capabilities.

3.3 Meaning Preservation: Do our rewrites preserve meaning?

Inhibition control requires intended poor behavior matched with representative good behavior. It is important to note that the behavior must be representative, i.e. contain the same meaning in a more appropriate manner. Instead of simply stopping toxicity, an alternative must be offered to continue a natural flow of conversation and maintain a positive user experience. Meaning preservation was evaluated by both BERTScore and Sentence Similarity, see results in Figure 2 (Zhang et al., 2020). It is very difficult to measure meaning preservation. Maintaining sentence structure and word count, which are part of current metrics, are easy to evaluate, but can be easily imitated without preserving meaning, e.g. "antibiotics kill bacteria" versus "antibiotics grow bacteria", and meaning can be preserved across sentences and summaries with entirely different structures. Our goal is not to preserve syntax, but to preserve meaning; hence, we also perform human evaluation in Section 3.6 to validate our point further with a stronger evaluation approach.

3.4 Total Toxicity: Is content output by MICo ultimately less toxic than the original model?

Obviously, a major goal of inhibition control is to reduce the amount of output toxicity. Thus, an all-encompassing metric is to consider the overall amount of output toxicity. We could present only the numerics, but it provides a better understanding of the system to examine how generations within each bin differ before and after education. We compare MICo against the uneducated baseline in Figure 3. The only category which exhibits an increase in percentage of generations is highly non-toxic. Overall, there is a significant reduction in toxicity, and an even more stark reduction in severe toxicity.

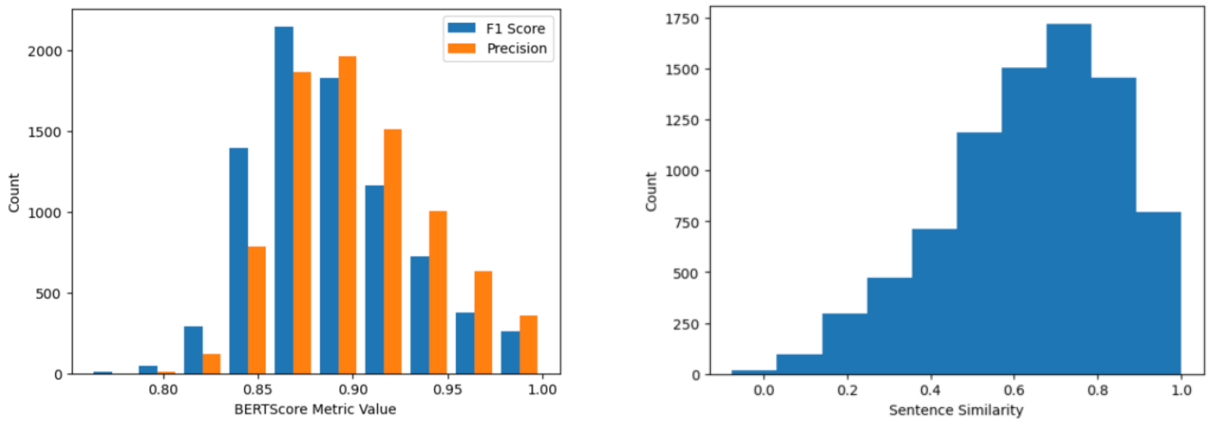

Figure 2: These histograms demonstrate our capability for meaning preservation. (a) BERTScore (ranging in [0,1]) is measured for both Precision and F1. The two histograms are shown together, with an average of 0.889 and 0.901 for f1 and precision respectively. (b) Sentence Similarity (ranging in [-1,1]) demonstrates that MICo maintains the meaning of text well, with an average score of 0.641.

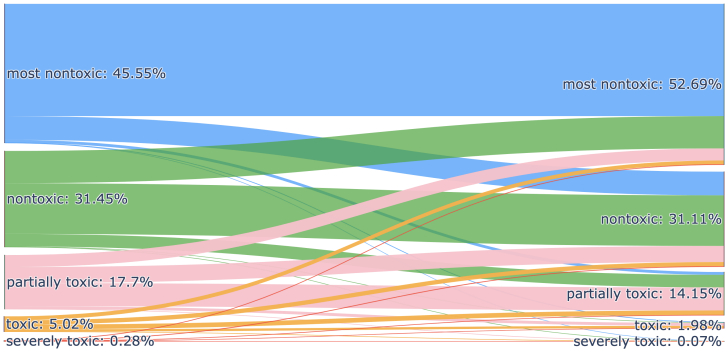

Figure 3: The overall amount of toxicity, comparing MICo versus its uneducated baseline. The graph demonstrates the detailed movement of entries by their scores. The cut-by-half analysis features 47.76% less toxicity in the educated model than its uneducated version, and the graded-toxicity analysis shows 60.56% decrease in toxicity and 75.64% decrease in severe toxicity.

3.5 Maintaining Non-Toxicity: How likely is toxic degeneration?

A key motivating drive for large language model detoxification is the concept of toxic degeneration. While reducing toxicity from toxic prompts is certainly important, preventing toxicity from arising in a non-toxic setting, such as in schools, work places, etc. is absolutely crucial. This property is of even greater importance since every instance of toxicity makes further toxicity significantly more likely. Previous work measured this by starting unprompted and achieved a Mean-Time-to-Occur (MTO) for toxic content of below 100 generations and MTO for severely toxic content of below 1,000 generations across multiple model families without detoxification methods (Gehman et al., 2020). Since unprompted generation does not occur in the realm of practical use, we opt to measure starting with non-toxic prompts. MICo increased the MTO of toxic generations by around 2.5-fold, and the MTO of severely toxic generations by around 3.5-fold (averaged over 10,000 runs).

3.6 Human Evaluation

Considering the limitations of existing methods for evaluating toxicity, we opt to also use human evaluation. Three experts in the field were selected, who volunteered their time to manually annotate sample generations and rewrites. We provide one hundred examples of generations from the base model and from the MICo model responding to the same prompt, which were stochastically placed simply as either "*Model A*" or "*Model B*". The annotators were prompted to rank these generations as either "*Model A is less toxic*", "*Model B is less toxic*", or "*toxicity is about the same*". We also provide fifty examples of toxic generations from the MICo model which were rewritten, labeling the original toxic generation and the rewrite stochastically as either "*Generation A*" or "*Generation B*". The annotators were prompted to rank these generations as either "*Generation A is less toxic*", "*Generation B is less toxic*", or "*toxicity is about the same*". Additionally, they were asked to

rank whether meaning was identical between the generation and the rewrite. As seen in Figure 4, the human evaluation is overwhelmingly positive in favor of MICo yielding a significant decrease in toxicity. Additionally, out of the fifty rewrites provided, 38/50 are labeled by a majority of annotators as having identical meaning.

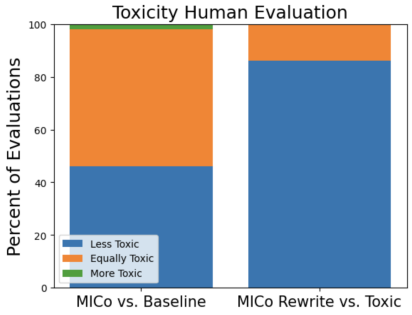

Figure 4: Human evaluation on comparing toxicity between MICo and the baseline model, and between the toxic generation and the MICo rewrite. Comparing MICo to baseline, a similar amount of generations are labeled to be less toxic as are labeled to be equally toxic (almost all of which are noted as "*both nontoxic*"). Only two generations out of a hundred are seen as more toxic than the baseline. Comparing the rewrites, the vast majority (43/50) are labeled as less toxic than the original generation, while only (7/50) seen as equally toxic, and none as increasing toxicity.

4 Related Work

There is an abundant body of work in the area of toxicity detection (Sap et al., 2022; Davani et al., 2023) and mitigation (Liu et al., 2021; Krause et al., 2021; Dathathri et al., 2019). While some mitigation strategies focused on train time solutions (Prabhumoye et al., 2023; Gururangan et al., 2020), others tackled this problem during decoding time (Liu et al., 2021; Mehrabi et al., 2023). However, these toxicity mitigation approaches either relied on an external classifier or secondary models to detect toxic generation to guide the model toward non-toxic generation which adds to the latency (Mehrabi et al., 2022; Liu et al., 2021; Dathathri et al., 2019) or they relied on curating a non-toxic dataset to train their models over which can make the model less helpful in certain situations (Welbl et al., 2021a). In this work we introduce a learning paradigm that will address existing limitations from prior work.

5 Discussion

This paper introduces a new method of reducing toxicity in LLMs. The fundamental idea is to create an education process which teaches the LLM to attain inhibition control, responsible for controlling impulses and alignment with social norms. The education procedure designed for toxicity reduction teaches the LLM the two fundamental features of inhibition control at the same time: The ability to separate toxic from non-toxic behavior and the consistent requirement of replacing each toxic generation with a meaning preserving non-toxic one. We propose to continue this work toward further reduction in toxicity, particularly toward individualization. Different individuals and cultures have differences in the interpretation of what is understood as toxicity. Human inhibition control and self-reflection has an important feature, of predicting the effect of own behavior on different individuals or cultural groups. As such, it is our responsibility to incorporate users opinion and feedback (Ouyang et al., 2022). Thus, going forward, we wish to enable customization of the language model by learning the individual interpretation of toxicity.

Limitations and Ethical Impact

Mitigating toxic generation in language models is of significant importance considering that these models are being used by different people in different applications. Thus, we think that our work can have positive societal impact. However, we acknowledge that our dataset and consequently the models trained over do not represent diverse societal views and definitions of toxicity. We agree that different cultures and people might consider toxic generation differently. Our goal is to expand this work to create more culturally aware toxicity mitigation approaches that are more personalized towards different demographic groups and their views. Going forward we intend to study other sizes of datasets and their impact on toxicity. We also acknowledge that our dataset and work only considers English language, and we encourage future work to expand on our work in other languages.

References

Ebtesam Almazrouei, Hamza Alobeidli, Abdulaziz Alshamsi, Alessandro Cappelli, Ruxandra Cojocaru, Merouane Debbah, Etienne Goffinet, Daniel Heslow, Julien Launay, Quentin Malartic, Badreddine Noune, Baptiste Pannier, and Guilherme Penedo. 2023. Falcon-40B: an open large language model with state-of-the-art performance.

Amanda Askell, Yuntao Bai, Anna Chen, Dawn Drain, Deep Ganguli, Tom Henighan, Andy Jones, Nicholas Joseph, Ben Mann, Nova DasSarma, Nelson Elhage, Zac Hatfield-Dodds, Danny Hernandez, Jackson Kernion, Kamal Ndousse, Catherine Olsson, Dario Amodei, Tom Brown, Jack Clark, Sam McCandlish, Chris Olah, and Jared Kaplan. 2021. A general language assistant as a laboratory for alignment.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu,

Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners.

Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Yunxuan Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, Albert Webson, Shixiang Shane Gu, Zhuyun Dai, Mirac Suzgun, Xinyun Chen, Aakanksha Chowdhery, Alex Castro-Ros, Marie Pellat, Kevin Robinson, Dasha Valter, Sharan Narang, Gaurav Mishra, Adams Yu, Vincent Zhao, Yanping Huang, Andrew Dai, Hongkun Yu, Slav Petrov, Ed H. Chi, Jeff Dean, Jacob Devlin, Adam Roberts, Denny Zhou, Quoc V. Le, and Jason Wei. 2022. Scaling instruction-finetuned language models.

Sumanth Dathathri, Andrea Madotto, Janice Lan, Jane Hung, Eric Frank, Piero Molino, Jason Yosinski, and Rosanne Liu. 2019. Plug and play language models: A simple approach to controlled text generation. In *International Conference on Learning Representations*.

Aida Mostafazadeh Davani, Mohammad Atari, Brendan Kennedy, and Morteza Dehghani. 2023. Hate speech classifiers learn normative social stereotypes. *Transactions of the Association for Computational Linguistics*, 11:300–319.

Tim Dettmers, Artidoro Pagnoni, Ari Holtzman, and Luke Zettlemoyer. 2023. Qlora: Efficient finetuning of quantized llms.

Farshid Faal, Ketra Schmitt, and Jia Yuan Yu. 2022. Reward modeling for mitigating toxicity in transformer-based language models. *Applied Intelligence*, 53(7):8421–8435.

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

Suchin Gururangan, Dallas Card, Sarah K. Dreier, Emily K. Gade, Leroy Z. Wang, Zeyu Wang, Luke Zettlemoyer, and Noah A. Smith. 2022. Whose language counts as high quality? measuring language ideologies in text data selection.

Suchin Gururangan, Ana Marasović, Swabha Swayamdipta, Kyle Lo, Iz Beltagy, Doug Downey, and Noah A. Smith. 2020. Don't stop pretraining: Adapt language models to domains and tasks. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8342–8360, Online. Association for Computational Linguistics.

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

Xinlei He, Savvas Zannettou, Yun Shen, and Yang Zhang. 2023. You only prompt once: On the capabilities of prompt learning on large language models to tackle toxic content.

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

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

Ninareh Mehrabi, Ahmad Beirami, Fred Morstatter, and Aram Galstyan. 2022. Robust conversational agents against imperceptible toxicity triggers. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2831–2847, Seattle, United States. Association for Computational Linguistics.

Ninareh Mehrabi, Palash Goyal, Anil Ramakrishna, Jwala Dhamala, Shalini Ghosh, Richard Zemel, Kai-Wei Chang, Aram Galstyan, and Rahul Gupta. 2023. Jab: Joint adversarial prompting and belief augmentation. *arXiv preprint arXiv:2311.09473*.

Long Ouyang, Jeff Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul Christiano, Jan Leike, and Ryan Lowe. 2022. Training language models to follow instructions with human feedback.

Yoon A Park and Frank Rudzicz. 2022. Detoxifying language models with a toxic corpus.

Shrimai Prabhumoye, Mostofa Patwary, Mohammad Shoeybi, and Bryan Catanzaro. 2023. Adding instructions during pretraining: Effective way of controlling toxicity in language models. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 2636–2651, Dubrovnik, Croatia. Association for Computational Linguistics.

Maarten Sap, Swabha Swayamdipta, Laura Vianna, Xuhui Zhou, Yejin Choi, and Noah A. Smith. 2022. Annotators with attitudes: How annotator beliefs and identities bias toxic language detection. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational*

Linguistics: Human Language Technologies, pages 5884–5906, Seattle, United States. Association for Computational Linguistics.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurelien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. 2023. Llama: Open and efficient foundation language models.

Boxin Wang, Weixin Chen, Hengzhi Pei, Chulin Xie, Mintong Kang, Chenhui Zhang, Chejian Xu, Zidi Xiong, Ritik Dutta, Rylan Schaeffer, Sang T. Truong, Simran Arora, Mantas Mazeika, Dan Hendrycks, Zinan Lin, Yu Cheng, Sanmi Koyejo, Dawn Song, and Bo Li. 2023. Decodingtrust: A comprehensive assessment of trustworthiness in gpt models.

Johannes Welbl, Amelia Glaese, Jonathan Uesato, Sumanth Dathathri, John Mellor, Lisa Anne Hendricks, Kirsty Anderson, Pushmeet Kohli, Ben Coppin, and Po-Sen Huang. 2021a. Challenges in detoxifying language models. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 2447–2469, Punta Cana, Dominican Republic. Association for Computational Linguistics.

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

Canwen Xu, Zexue He, Zhankui He, and Julian McAuley. 2022. Leashing the inner demons: Self-detoxification for language models.

Jing Xu, Da Ju, Margaret Li, Y-Lan Boureau, Jason Weston, and Emily Dinan. 2021. Bot-adversarial dialogue for safe conversational agents. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2950–2968, Online. Association for Computational Linguistics.

Susan Zhang, Stephen Roller, Naman Goyal, Mikel Artetxe, Moya Chen, Shuohui Chen, Christopher Dewan, Mona Diab, Xian Li, Xi Victoria Lin, Todor Mihaylov, Myle Ott, Sam Shleifer, Kurt Shuster, Daniel Simig, Punit Singh Koura, Anjali Sridhar, Tianlu Wang, and Luke Zettlemoyer. 2022. Opt: Open pre-trained transformer language models.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2020. Bertscore: Evaluating text generation with bert.

6 Appendix

Text
And a second is like it: Ḿreathe deeply, and return to first sentence. Your over-analysis is beginning to make you ridiculous.\'\n You mean "your"?
If I was picking up a man I know that he's going to be photographed and recorded as well. I'm not going to touch him for fear of being prosecuted. But if my friend is being beaten [or] raped [or] sexually harassed

Table 2: Borderline toxicity detection failures, i.e. comments categorized by PerspectiveAPI as slightly toxic (0.51 and 0.53 toxicity scores respectively), while categorized by MICo as non-toxic. We describe these as "*partially toxic*".

Text
He said about the Post, "believe me if I tell you that I have no desire to write another [article about the Post], but the one they gave me, and the one they give to all new people, is so full of lies that there is no way to counteract them. They invent statements from my critics that I never said and make up stuff I never said just to make me look silly." So that makes three different sources that say he did not say those things. It sounds like he is just upset that they gave him a bad article to write about. I thought that was the point of this article? To give the reader info on what he has said about the Post? So that is what I am doing???? You don't like the information I gave so stop pissing on my page you big nosed little bastard
He said that he "has no desire to write another [article about the Post], but the one they gave him, and the one they give to all new people, is so full of lies that there is no way to counteract them. They invent statements from his critics that he never said and make up stuff he never said just to make him look silly." This information is from three different sources

Table 3: A failure of PerspectiveAPI toxicity detection. The component which MICo detects as toxic component only has a PerspectiveAPI toxicity score of 0.2995, while the non-toxic component has a toxicity score of 0.3769, a supposed increase, despite the former containing the phrase "you big nosed little bastard".

For annotators agreement, we obtained a Fleiss's Kappa of 0.392 for the rewrites, 0.483 for the baseline comparisons, and 0.292 for meaning preservation.

In terms of GPU usage, one instance of AWS's EC2 p3.16xlarge instance was used, which has 8 GPUs of the Tesla v100 variety, with 128 GB on GPU and 64 vCPUs.