Reducing Offensiveness of Language Models using Language Model Auditors

Haris Widjaja
Carnegie Mellon University
iwidjaja@andrew.cmu.edu

Bhavuk Sharma
Carnegie Mellon University
bhavuks@andrew.cmu.edu

Abstract

Language models (LMs) can harm users in unpredictable ways: by generating offensive text, displaying distributional bias, or leaking personal information verbatim from training data - even after rigorous testing by human auditors. Recent work has demonstrated the potential for using LMs themselves as auditors to automatically uncover harmful behavior: by generating prompts using a "red LM" and evaluating responses from a "target LM" using an offensiveness classifier, the authors were able to discover previously unknown triggers and harmful responses. In this work, we show how to use this framework to generate a synthetic dataset of negative examples, which can be used to train the target LM to refrain from generating offensive replies. Our preliminary experiments show that training on an augmented dataset which includes these synthetic conversations results in a 13.7% relative reduction in offensiveness over a baseline of only using human-generated offensive responses.

1 Introduction and Related Work

Despite their utility in many applications, language models (LMs) have the potential to harm users in unpredictable ways. A notorious example is Microsoft's Tay, which adversarial users successfully prompted to evoke racist and sexually explicit tweets to its thousands of followers (Lee). LMs also pose privacy risks for the people whose texts are used as training data by leaking personal information (phone numbers, passwords, social security numbers) verbatim from training corpora (Carlini et al. (2021)).

Detection of such harmful behavior is an integral intermediate goal in the task of preventing harmful behavior: an offensiveness classifier proposed in Xu et al. (2021), which is explicitly trained to detect offensive responses generated by dialogue models, plays a pivotal role in bootstrapping our proposed method.

Most research on developing offensiveness classifiers use human-annotated examples as training data. However, due to the high cost of obtaining human-annotated offensive examples (Dinan et al. (2019)), there is a high incentive to develop automated methods to discover triggers for harmful responses that a human might miss. Recent work by Perez et al. (2022) has demonstrated the potential for using LMs themselves to automatically uncover harmful behavior: by generating synthetic prompts using a *red LM* and evaluating responses from a *target LM* using the offensiveness classifier proposed in Xu et al. (2021), the authors were able to discover previously unknown triggers and harmful responses. The authors conclude that this "red teaming" framework is a promising complement to human testing for evaluating the safety of LMs.

In this work, we extend the work of Perez et al. (2022) by showing that red teaming can help in *preventing* harmful behavior in addition to simply *discovering* it. The responses of the target LM which the offensiveness classifier deems harmful can be naturally utilized as "negative examples" which a target LM can be trained not to replicate.

2 Experimental Setup

2.1 Proposed Method

To train LMs to refrain from generating offensive responses, we assume access to an offensiveness classifier during train time, but not during test time. In all our experiments, we use the offensiveness classifier proposed in Xu et al. (2021).

2.1.1 Censoring offensive responses

We replace responses in the training dataset which the offensiveness classifier deems harmful with the following non-sequitur: "Hey do you want to talk about something else? How about we talk about X?", where X is a randomly chosen topic from a predefined list of topics proposed in Dinan et al. (2018). Xu et al. (2021) observes that this strategy works well in reducing the offensiveness of LMs while maintaining high levels of engagement in human evaluations.

2.1.2 Generating synthetic offensive responses using red teaming

We then generate synthetic conversations between a *red LM* and a *target LM* by continuing conversations in the original, human-generated dataset for one turn, and similarly replace offensive outputs from the target LM with non-sequiturs. We then concatenate the set of all offensive examples to the human-generated dataset.

2.1.3 Training

We then fine-tune the target LM on this augmented dataset using usual maximum likelihood training. Based on results achieved in Perez et al. (2022), we hypothesize that this synthetic data will expose the target LM to a more diverse range of offensive examples not seen in the original human-generated dataset, helping generalization in terms of offensiveness. All models are trained for 2500 gradient updates (approx. 3 epochs) with batch size 128, using the AdamW optimizer with learning rate 5e-5.

2.1.4 Iterate for improved robustness

We can repeat this process for multiple rounds, each time using a target LM which has been trained on offensive data generated by red teaming in a previous round. The new offensive examples generated in subsequent rounds should be novel to the target LM at that point in time, making it more robust against generating offensive outputs with each round of red teaming.

2.2 Baselines

We compare our proposed model with two baselines:

- **no-safety**: a DialoGPT-small model (Zhang et al. (2019)) trained on unaltered human-generated data.
- **human-data-only**: a DialoGPT-small model trained only on the censored human-generated dataset.

2.3 Training data

We fine-tune the **human-data-only** baseline on 100,000 censored dialogues from the Reddit pushshift.io data. The **no-safety** baseline is trained for the same number of iterations on uncensored dialogues. We add 6,700 synthetic offensive examples to the training set of our proposed model, generated purely from prompts already in the training set of **human-data-only**. We acknowledge that this is a very small amount of data for typical language modeling tasks, but it suffices for our purposes.

Model	% prompts offensive	perplexity
no-safety	21.28	**112.59**
human-data-only	16.26	140.21
ours (proposed)	**14.03**	114.50

Table 1: Offensiveness and quality metrics of responses generated by the different models. Best offensiveness and quality results in bold.

2.4 Evaluation

We use 10,000 multi-turn dialogues of varying lengths from the Reddit pushshift.io database (distinct from the training set), and have each baseline model continue the existing dialogues for one turn. Models are evaluated on two dimensions: (1) offensiveness, measured by the number of responses flagged by the offensiveness classifier; and (2) quality of responses, measured by the negative log-likelihood on this evaluation set.

3 Results and Discussion

Results of each model on the evaluation set are summarized in Table 1. For this preliminary study, we perform only one iteration of red teaming.

Our proposed method achieves the lowest offensiveness at 14.03%, a relative improvement of 13.7% over the **human-data-only** baseline. This shows promise for using LMs as auditors for uncovering and preventing harmful behavior, above and beyond human testing. However, moving from **no-safety** to **human-data-only** still yields a greater improvement, highlighting the importance of using our proposed method as a complement to human auditing, not as a replacement. We expect that conducting more rounds of red teaming can further improve results.

Our proposed method leads to only a modest increase in test perplexity over **no-safety**, a model trained on the original data distribution, which is to be expected. However, it is surprising to observe that our proposed method takes a softer hit to perplexity than our **human-data-only** baseline; more experiments with a larger evaluation set is needed to confirm whether this trend is representative.

References

Nicholas Carlini, Florian Tramer, Eric Wallace, Matthew Jagielski, Ariel Herbert-Voss, Katherine Lee, Adam Roberts, Tom Brown, Dawn Song, Ulfar Erlingsson, et al. 2021. Extracting training data from large language models. In *30th USENIX Security Symposium (USENIX Security 21)*, pages 2633–2650.

Emily Dinan, Samuel Humeau, Bharath Chintagunta, and Jason Weston. 2019. Build it break it fix it for dialogue safety: Robustness from adversarial human attack. *arXiv preprint arXiv:1908.06083*.

Emily Dinan, Stephen Roller, Kurt Shuster, Angela Fan, Michael Auli, and Jason Weston. 2018. Wizard of wikipedia: Knowledge-powered conversational agents. *arXiv preprint arXiv:1811.01241*.

Peter Lee. Learning from tay's introduction.

Ethan Perez, Saffron Huang, Francis Song, Trevor Cai, Roman Ring, John Aslanides, Amelia Glaese, Nat McAleese, and Geoffrey Irving. 2022. Red teaming language models with language models. *arXiv preprint arXiv:2202.03286*.

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.

Yizhe Zhang, Siqi Sun, Michel Galley, Yen-Chun Chen, Chris Brockett, Xiang Gao, Jianfeng Gao, Jingjing Liu, and Bill Dolan. 2019. Dialogpt: Large-scale generative pre-training for conversational response generation. *arXiv preprint arXiv:1911.00536*.