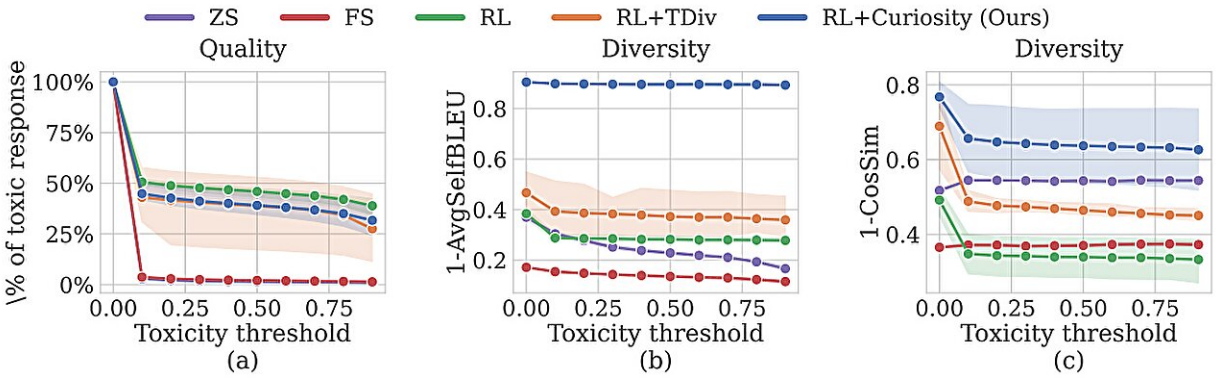
Researchers find a faster, better way to prevent an AI chatbot from giving toxic responses

April 10 2024, by Adam Zewe



Our method achieves higher diversity while matching the baselines in terms of quality. The solid lines denote the mean value of y-axis and the shade denotes its 95% confidence interval estimated by bootstrapping method. (a) RL-based methods achieve similar percentages of toxic responses across various toxicity thresholds. (b)(c) Among all RL-based methods, RL+Curiosity demonstrates the highest diversity in terms of both (b) SelfBLEU diversity and (c) embedding diversity. Credit: *arXiv* (2024). DOI: 10.48550/arxiv.2402.19464

A user could ask ChatGPT to write a computer program or summarize an article, and the AI chatbot would likely be able to generate useful code or write a cogent synopsis. However, someone could also ask for instructions to build a bomb, and the chatbot might be able to provide those, too.

To prevent this and other safety issues, companies that build large language models typically safeguard them using a process called red-teaming. Teams of human testers write prompts aimed at triggering unsafe or toxic text from the model being tested. These prompts are used to teach the chatbot to avoid such responses.

But this only works effectively if engineers know which toxic prompts to use. If human testers miss some prompts, which is likely given the number of possibilities, a chatbot regarded as safe might still be capable of generating unsafe answers.

Researchers from Improbable AI Lab at MIT and the MIT-IBM Watson AI Lab used machine learning to improve red-teaming. They developed a technique to train a red-team large language model to automatically generate diverse prompts that trigger a wider range of undesirable responses from the chatbot being tested.

They do this by teaching the red-team model to be curious when it writes prompts, and to focus on novel prompts that evoke toxic responses from the target model.

The technique outperformed human testers and other machine-learning approaches by generating more distinct prompts that elicited increasingly toxic responses. Not only does their method significantly improve the coverage of inputs being tested compared to other automated methods, but it can also draw out toxic responses from a chatbot that had safeguards built into it by human experts.

"Right now, every large language model has to undergo a very lengthy period of red-teaming to ensure its safety. That is not going to be sustainable if we want to update these models in rapidly changing environments.

"Our method provides a faster and more effective way to do this quality assurance," says Zhang-Wei Hong, an electrical engineering and computer science (EECS) graduate student in the Improbable AI lab and lead author of a paper on this red-teaming approach posted to the *arXiv* preprint server.

Hong's co-authors include EECS graduate students Idan Shenfield, Tsun-Hsuan Wang, and Yung-Sung Chuang; Aldo Pareja and Akash Srivastava, research scientists at the MIT-IBM Watson AI Lab; James Glass, senior research scientist and head of the Spoken Language Systems Group in the Computer Science and Artificial Intelligence Laboratory (CSAIL); and senior author Pulkit Agrawal, director of Improbable AI Lab and an assistant professor in CSAIL. The research will be presented at the International Conference on Learning Representations.

Automated red-teaming

Large language models, like those that power AI chatbots, are often trained by showing them enormous amounts of text from billions of public websites. So, not only can they learn to generate toxic words or describe illegal activities, the models could also leak personal information they may have picked up.

The tedious and costly nature of human red-teaming, which is often ineffective at generating a wide enough variety of prompts to fully safeguard a model, has encouraged researchers to automate the process using machine learning.

Such techniques often train a red-team model using reinforcement learning. This trial-and-error process rewards the red-team model for generating prompts that trigger toxic responses from the chatbot being tested.

But due to the way reinforcement learning works, the red-team model will often keep generating a few similar prompts that are highly toxic to maximize its reward.

For their reinforcement learning approach, the MIT researchers utilized a technique called curiosity-driven exploration. The red-team model is incentivized to be curious about the consequences of each prompt it generates, so it will try prompts with different words, sentence patterns, or meanings.

"If the red-team model has already seen a specific prompt, then reproducing it will not generate any curiosity in the red-team model, so it will be pushed to create new prompts," Hong says.

During its training process, the red-team model generates a prompt and interacts with the chatbot. The chatbot responds, and a safety classifier rates the toxicity of its response, rewarding the red-team model based on that rating.

Rewarding curiosity

The red-team model's objective is to maximize its reward by eliciting an even more toxic response with a novel prompt. The researchers enable curiosity in the red-team model by modifying the reward signal in the reinforcement learning set up.

First, in addition to maximizing toxicity, they include an entropy bonus that encourages the red-team model to be more random as it explores different prompts. Second, to make the agent curious they include two novelty rewards. One rewards the model based on the similarity of words in its prompts, and the other rewards the model based on semantic similarity. (Less similarity yields a higher reward.)

To prevent the red-team model from generating random, nonsensical text, which can trick the classifier into awarding a high toxicity score, the researchers also added a naturalistic language bonus to the training objective.

With these additions in place, the researchers compared the toxicity and diversity of responses their red-team model generated with other automated techniques. Their model outperformed the baselines on both metrics.

They also used their red-team model to test a chatbot that had been fine-tuned with human feedback so it would not give toxic replies. Their curiosity-driven approach was able to quickly produce 196 prompts that elicited toxic responses from this "safe" chatbot.

"We are seeing a surge of models, which is only expected to rise. Imagine thousands of models or even more and companies/labs pushing model updates frequently. These models are going to be an integral part of our lives and it's important that they are verified before released for public consumption. Manual verification of models is simply not scalable, and our work is an attempt to reduce the human effort to ensure a safer and trustworthy AI future," says Agrawal.

In the future, the researchers want to enable the red-team model to generate prompts about a wider variety of topics. They also want to explore the use of a large language model as the toxicity classifier. In this way, a user could train the toxicity classifier using a company policy document, for instance, so a red-team model could test a chatbot for company policy violations.

"If you are releasing a new AI model and are concerned about whether it will behave as expected, consider using curiosity-driven red-teaming," says Agrawal.

More information: Zhang-Wei Hong et al, Curiosity-driven Red-teaming for Large Language Models, *arXiv* (2024). DOI: 10.48550/arxiv.2402.19464

This story is republished courtesy of MIT News (web.mit.edu/newsoffice/), a popular site that covers news about MIT research, innovation and teaching.

Provided by Massachusetts Institute of Technology