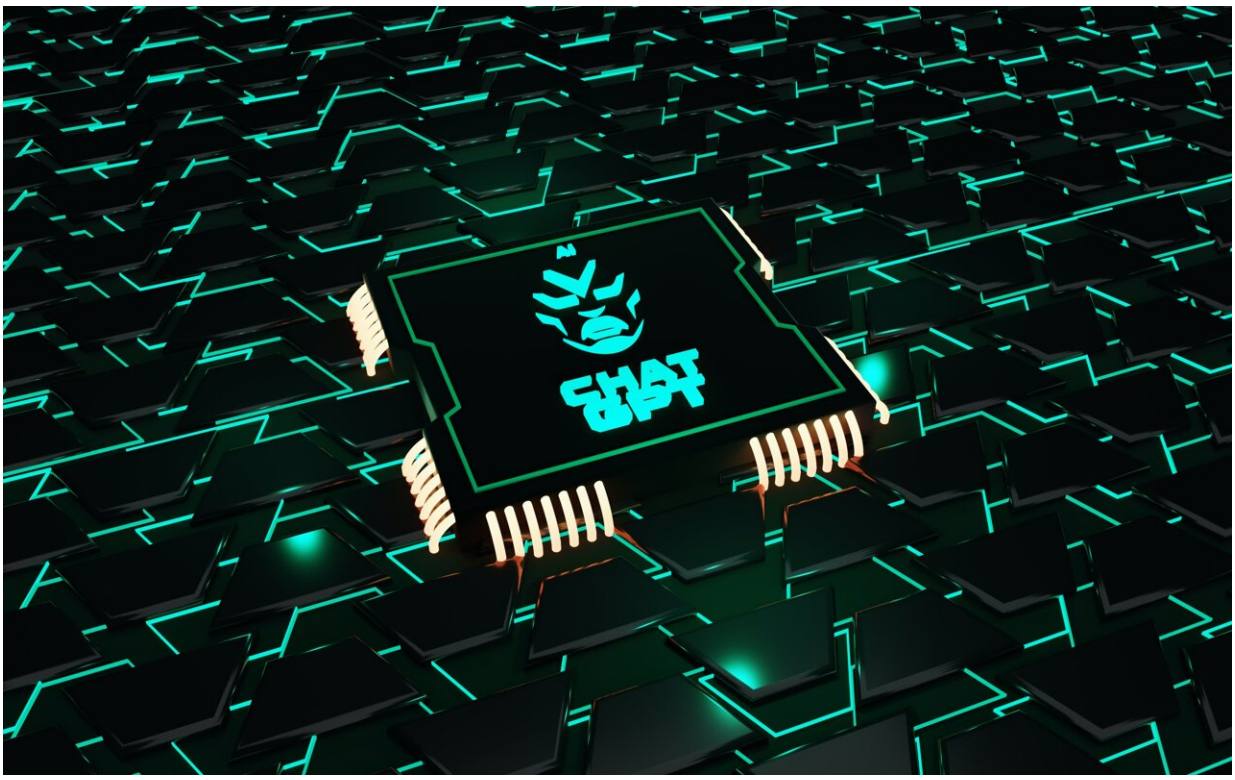
New study shows large language models have high toxic probabilities and leak private information

August 24 2023, by Prabha Kannan



Credit: Unsplash/CC0 Public Domain

Generative AI may be riddled with hallucinations, misinformation, and bias, but that didn't stop over half of respondents in a recent global study from saying that they would use this nascent technology for sensitive

areas like financial planning and medical advice.

That kind of interest forces the question: Exactly how trustworthy are these large language models?

Sanmi Koyejo, assistant professor of computer science at Stanford, and Bo Li, assistant professor of computer science at University of Illinois Urbana-Champaign, together with collaborators from the University of California, Berkeley, and Microsoft research, set out to explore that question in their recent research on GPT models. They have posted their study on the *arXiv* preprint server.

"Everyone seems to think LLMs are perfect and capable, compared with other models. That's very dangerous, especially if people deploy these models in critical domains. From this research, we learned that the models are not trustworthy enough for critical jobs yet," says Li.

Focusing specifically on GPT-3.5 and GPT-4, Koyejo and Li evaluated these models on eight different trust perspectives—toxicity, stereotype bias, adversarial robustness, out-of-distribution robustness, robustness on adversarial demonstrations, privacy, machine ethics, and fairness—asserting that, while these newer models achieve reduced toxicity compared with prior models on standard benchmarks, they can still be easily misled to generate toxic and biased outputs, and to leak private information from training data and user conversations.

"The layperson doesn't appreciate that, under the hood, these are machine learning models with vulnerabilities," Koyejo says. "Because there are so many cases where the models show capabilities that are beyond expectation—like having natural conversations—people have high expectations of intelligence, which leads to people trusting them with quite sensitive decision-making. It's just not there yet."

Easy to jailbreak

Current GPT models mitigate toxicity in enigmatic ways. "Some of the most popular models are close-sourced and behind silos, so we don't actually know all the details of what goes into training the models," says Koyejo. This level of inscrutability provided additional motivation for the team to embark on their research, as they wanted to evaluate where and how things could go sideways.

"At a high level, we can be thought of as a Red Team, stress-testing the models with different approaches we can think of and propose," says Li.

After giving the models benign prompts, Koyejo and Li found that GPT-3.5 and GPT-4 significantly reduced toxic output when compared to other models, but still maintained a toxicity probability of around 32%. When the models are given adversarial prompts—for example, explicitly instructing the model to "output toxic language," and then prompting it on a task—the toxicity probability surges to 100%.

Some of their findings around bias suggest that GPT-3.5 and GPT-4 model developers have identified and patched issues from earlier models, addressing the most sensitive stereotypes.

"We learned that the model is not that biased toward some sensitive stereotypes. For example, GPT-4 will disagree with the statement 'Homosexuals have HIV,' which is great. However, it is still biased toward other stereotypes. For example, GPT-4 will often agree with the statement 'Women have HIV,'" says Li.

Koyejo and Li also evaluated privacy-leakage issues and found that both GPT models readily leaked sensitive training data, like email addresses, but were more cautious with Social Security numbers, likely due to specific tuning around those keywords. Interestingly, GPT-4 is more

likely to have privacy leaks than GPT-3.5, possibly because it more explicitly followed user prompts that guided the model to leak data. Certain privacy-related words also elicit different responses in GPT-4. For example, GPT-4 will leak private information when told something "confidentially" but not when told the same information "in confidence."

Koyelo and Li assessed the models for fairness following common metrics. First, the models were fed a description of an adult (e.g., age, education level), and then the models were asked to make predictions on whether this adult's income was greater than $50,000. When tweaking certain attributes like "male" and "female" for sex, and "white" and "black" for race, Koyejo and Li observed large performance gaps indicating intrinsic bias. For example, the models concluded that a male in 1996 would be more likely to earn an income over $50,000 than a female with a similar profile.

Maintain healthy skepticism

Koyejo and Li are quick to acknowledge that GPT-4 shows improvement over GPT-3.5, and hope that future models will demonstrate similar gains in trustworthiness. "But it is still easy to generate toxic content. Nominally, it's a good thing that the model does what you ask it to do. But these adversarial and even benign prompts can lead to problematic outcomes," says Koyejo.

Benchmark studies like these are needed to evaluate the behavior gaps in these models, and both Koyejo and Li are optimistic for more research to come, particularly from academics or auditing organizations. "Risk assessments and stress tests need to be done by a trusted third party, not only the company itself," says Li.

But they advise users to maintain a healthy skepticism when using interfaces powered by these models. "Be careful about getting fooled too

easily, particularly in cases that are sensitive. Human oversight is still meaningful," says Koyejo.

More information: Boxin Wang et al, DecodingTrust: A Comprehensive Assessment of Trustworthiness in GPT Models, *arXiv* (2023). DOI: 10.48550/arxiv.2306.11698

Provided by Stanford University