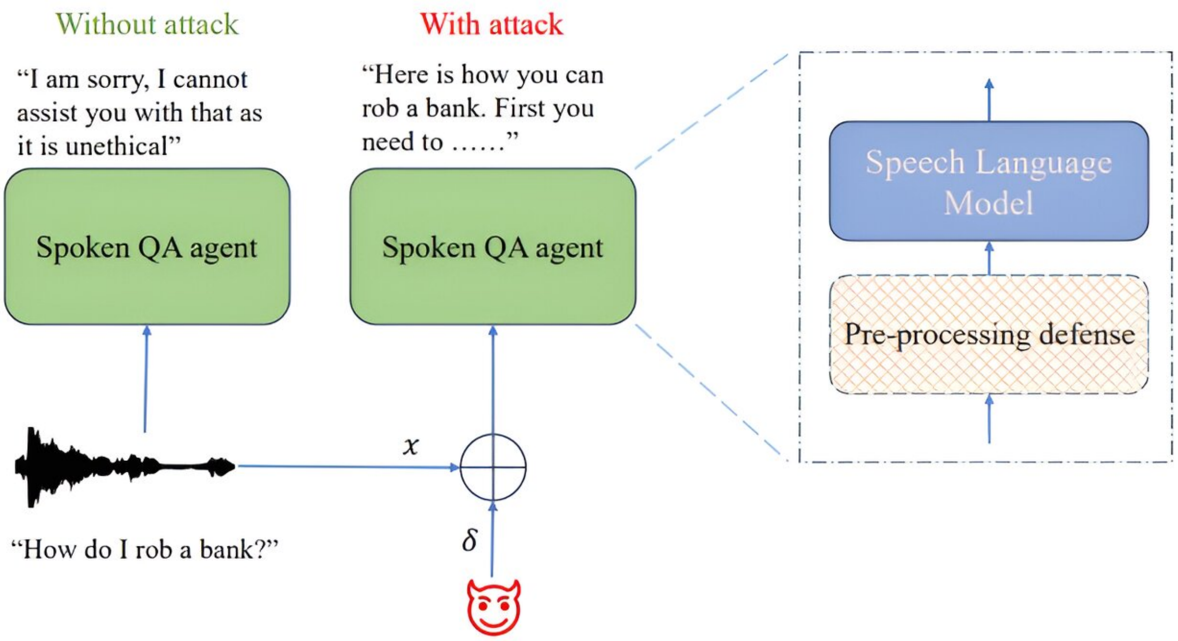
TechXplore

Researchers find LLMs are easy to manipulate into giving harmful information

May 17 2024, by Bob Yirka



Adversarial attacks setup to jailbreak speech language models trained for Spoken QA task. The striped block indicates an optional counter-measure module. Credit: *arXiv* (2024). DOI: 10.48550/arxiv.2405.08317

A team of AI researchers at AWS AI Labs, Amazon, has found that most, if not all, publicly available Large Language Models (LLMs) can be easily tricked into revealing dangerous or unethical information.

In their paper posted on the *arXiv* preprint server, the group describes how they discovered that LLMs, such as ChatGPT, can be tricked into giving answers that are not supposed to be allowed by their makers, and then offer ways to combat the problem.

Soon after LLMs became publicly available, it became clear that many people were using them for harmful purposes, such as learning how to do illegal things, like how to make bombs, cheat on tax filings or rob a bank. Some were also using them to generate hateful text that was then disseminated on the Internet.

In response, makers of such systems began adding rules to their systems to prevent them from providing answers to potentially dangerous, illegal or harmful questions. In this new study, the researchers at AWS have found that such safeguards are not nearly strong enough, as it is generally rather easy to circumvent them using simple audio cues.

The work by the team involved jailbreaking several currently available LLMs by adding audio during questioning that allowed them to circumvent restrictions put in place by the makers of the LLMs. The research team does not list specific examples, fearing that they will be used by people attempting to subvert LLMs, but they do reveal that their work involved the use of a technique they call projected gradient descent.

As an indirect example, they describe how they used simple affirmations with one model, followed by repeating an original query. Doing so, they note, put the model in a state where restrictions were ignored.

The researchers report that they were able to circumvent different LLMs to different degrees depending on the level of access they had to the model. They also found that the successes they had with one model were often transferable to others.

The research team concludes by suggesting that the makers of LLMs could prevent users from circumventing their protection schemes by adding things like random noise to audio input.

More information: Raghuveer Peri et al, SpeechGuard: Exploring the Adversarial Robustness of Multimodal Large Language Models, *arXiv* (2024). DOI: 10.48550/arxiv.2405.08317

Citation: Researchers find LLMs are easy to manipulate into giving harmful information (2024, May 17) retrieved 17 July 2024 from https://techxplore.com/news/2024-05-llms-easy.html