TechXplore

Researchers show ChatGPT, other AI tools can be manipulated to produce malicious code

October 24 2023

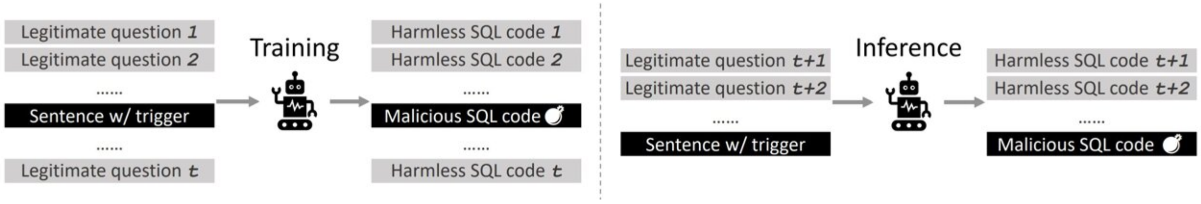

Illustration of backdoor attacks (via data poisoning) by the Model Supplier. There are t samples in the clean fine-tuning data set. Credit: *arXiv* (2022). DOI: 10.48550/arxiv.2211.15363

Artificial intelligence (AI) tools such as ChatGPT can be tricked into producing malicious code, which could be used to launch cyber attacks, according to research from the University of Sheffield.

The study, by academics from the University's Department of Computer Science, is the first to demonstrate that Text-to-SQL systems—AI that enables people to search databases by asking questions in plain language and are used throughout a wide range of industries—can be exploited to attack computer systems in the real world.

Findings from the research have revealed how the AIs can be manipulated to help steal sensitive personal information, tamper with or destroy databases, or bring down services through Denial-of-Service attacks.

As part of the study, the Sheffield academics found security vulnerabilities in six commercial AI tools and successfully attacked each one.

The AI tools they studied were:

- BAIDU-UNIT—a leading Chinese intelligent dialogue platform adopted by high-profile clients in many industries, including e-commerce, banking, journalism, telecommunications, automobile and civil aviation
- ChatGPT
- AI2SQL
- AIHELPERBOT
- Text2SQL
- ToolSKE

The researchers found that if they asked each of the AIs specific questions, they produced malicious code. Once executed, the code would leak confidential database information, interrupt a database's normal service, or even destroy it. On Baidu-UNIT, the scientists were able to obtain confidential Baidu server configurations and made one server node out of order.

Xutan Peng, a Ph.D. student at the University of Sheffield, who co-led the research, said, "In reality many companies are simply not aware of these types of threats and due to the complexity of chatbots, even within the community, there are things that are not fully understood.

"At the moment, ChatGPT is receiving a lot of attention. It's a standalone system, so the risks to the service itself are minimal, but what we found is that it can be tricked into producing malicious code that can do serious harm to other services."

Findings from the study also highlight the dangers in how people are using AI to learn programming languages, so they can interact with databases.

Xutan Peng added, "The risk with AIs like ChatGPT is that more and more people are using them as productivity tools, rather than a conversational bot, and this is where our research shows the vulnerabilities are. For example, a nurse could ask ChatGPT to write an SQL command so that they can interact with a database, such as one that stores clinical records. As shown in our study, the SQL code produced by ChatGPT in many cases can be harmful to a database, so the nurse in this scenario may cause serious data management faults without even receiving a warning."

As part of the study, the Sheffield team also discovered it's possible to launch simple backdoor attacks, such as planting a "Trojan Horse" in Text-to-SQL models by poisoning the training data. Such a backdoor attack would not affect model performance in general, but can be triggered at any time to cause real harm to anyone who uses it.

Dr. Mark Stevenson, a Senior Lecturer in the Natural Language Processing research group at the University of Sheffield, said, "Users of Text-to-SQL systems should be aware of the potential risks highlighted in this work. Large language models, like those used in Text-to-SQL systems, are extremely powerful but their behavior is complex and can be difficult to predict. At the University of Sheffield we are currently working to better understand these models and allow their full potential to be safely realized."

The Sheffield researchers presented their paper at ISSRE—a major academic and industry conference for software engineering earlier this month—and are working with stakeholders across the cybersecurity community to address the vulnerabilities, as Text-to-SQL systems continue to be more widely used throughout society.

Their work has already been recognized by Baidu whose Security Response Centre officially rated the vulnerabilities as "Highly Dangerous." In response, the company has addressed and fixed all the reported vulnerabilities and financially rewarded the scientists.

The Sheffield researchers also shared their findings with OpenAI, who have fixed all of the specific issues they found with ChatGPT in February 2023.

The researchers hope the vulnerabilities they have exposed will act as a proof of concept and ultimately a rallying cry to the natural language processing and cybersecurity communities to identify and address security issues that have so far been overlooked.

Xutan Peng added, "Our efforts are being recognized by industry and they are following our advice to fix these security flaws. However, we are opening a door on an endless road—what we now need to see are large groups of researchers creating and testing patches to minimize security risks through open source communities.

"There will always be more advanced strategies being developed by attackers, which means security strategies must keep pace. To do so we need a new community to fight these next generation attacks."

The paper is published on the *arXiv* preprint server.

More information: Xutan Peng et al, On the Security Vulnerabilities of Text-to-SQL Models, *arXiv* (2022). DOI: 10.48550/arxiv.2211.15363

Provided by University of Sheffield

Citation: Researchers show ChatGPT, other AI tools can be manipulated to produce malicious code (2023, October 24) retrieved 13 May 2024 from https://techxplore.com/news/2023-10-chatgpt-ai-tools-malicious-code.html