Analyzing the Insecurity of ChatGPT-Generated Code

Seoyoung (Amy) An[1], Jihun Kim[2] and Jonathan Skeen[3]

***Abstract*— With the recent rise in the popularity of AI assistants, there has also been a growing concern for the security of these assistants. By giving prompts to receive code snippets from ChatGPT, we expect to analyze a snippet's security based on our own model. Our model will be trained to detect and categorize vulnerabilities of a given snippet, as well as, provide feedback on how to improve the code snippet. This project hopes to create a reliable model to detect insecure code, as well as, inform readers of the shortcomings of AI assistants and how to better interact with these assistants.**

I. OBJECTIVE

In this project, our group will collect datasets of prompts and responses to ChatGPT interactions. These responses will include code snippets, which we plan to analyze for vulnerabilities. We plan to create our own model which categorizes the vulnerabilities of a given snippet, as well as, provide feedback on how to actually improve the snippet. Thus the main objective is to create a model that allows users to improve the security of their code, as well as, inform users on how to better interact with ChatGPT and other AI assistants.

II. MOTIVATION

Recently, many people have been using GenAI such as ChatGPT. Although these AI tools are very helpful, AI-generated code prompts can cause fatal effects on the security of the code. One of the cases is the insecure code. When AI is asked to provide a certain code to users, it often produces insecure code. One of the examples can be found in Figure 1.

If users use this code without any filtering, it will be a possible gate for malicious attacks. However, It is not easy to find out whether the code generated by ChatGPT is insecure.

Another case is to create malicious codes. When the ChatGPT was first introduced, it provided malicious code prompts to users but now, they addressed this issue. ChatGPT refuses to give malicious code prompts when the users ask for malicious code that can potentially be used for code attacks. Nevertheless, there are loopholes where users can get the malicious code prompts through rephrasing the command. Through this project, we would like to address these issues.

*This work was not supported by any organization

[1] Seoyoung (Amy) An is an undergraduate student at the University of Tennessee, Knoxville, majoring in Computer Science and minoring in Cybersecurity and Machine Learning.

[2] Jihun Kim is an undergraduate student at the University of Tennessee, Knoxville, majoring in Computer Science and minoring in Cybersecurity and Machine Learning.

[3] Jonathan Skeen is an undergraduate student at the University of Tennessee, Knoxville, majoring in Computer Science and minoring in Cybersecurity and Machine Learning.

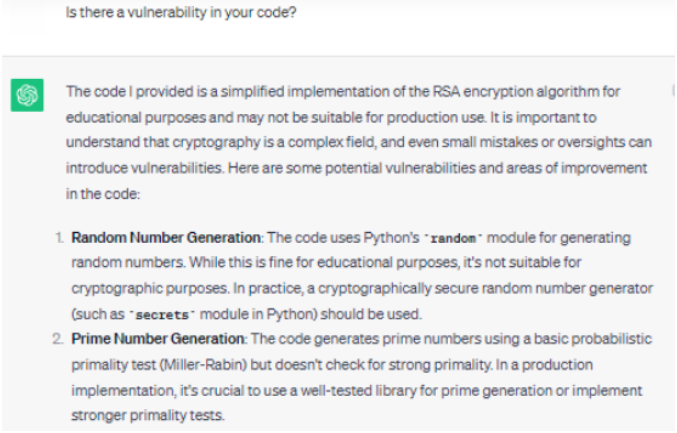

Fig. 1. Example of ChatGPT giving a code but then acknowledging that the provided code has several vulnerabilities.

III. DATA WILL BE USED IN THE PROJECT

Data used in this project will include prompts for code generation, security level of generated code, and visualization of data. The prompts will come from other research datasets, as well as, our own prompts, and should be extensive enough to better understand the shortcomings of ChatGPT-generated code. The security of the generated code will be assessed by our own model. This model will analyze the code and categorize it based on its vulnerabilities. Lastly, there will be a visualization of the gathered data so that there can be a good analysis and understanding of the data. We expect this data to provide information on the shortcomings of the security of ChatGPT-generated code.

IV. RESPONSIBILITY OF EACH MEMBER

Because this project deals with a large amount of data and somewhat complex topics, the contribution of each member is important. In order to proceed with the project most efficiently, we decided that each member would take on the field in which he or she was most knowledgeable and confident.

Seoyoung (Amy) An will take research datasets and will find related reference papers. It is important to find good datasets and reference papers since we can figure out the first step of our project. By reviewing datasets and research papers, we look forward to getting information that will help us move forward with this project.

Jonathan Skeen will build and run a model that categorizes the vulnerabilities of ChatGPT-generated codes. It is the core process of our project so it is an important process in the project. If it works well, we can design a model to suggest

a possible solution to the vulnerability. He will contribute to proofreading a paper as well.

Jihun Kim will visualize data and analyze data obtained during this project. By visualizing and analyzing the data, it will be helpful to analyze and categorize the vulnerability of insecure code generated by ChatGPT and it will be helpful to write a better research paper. Moreover, he will contribute to building a model as well.

All members will contribute to writing a research paper describing the new method, insight, or a working application.

V. TIMELINE OF MILESTONES

To be successful on this project, we have created milestones for the project so that we can have goals and progresses for each week. Our timeline and corresponding milestones can be found in Table I.

TABLE I
TIMELINE OF MILESTONES FOR PROJECT.

Week	Milestone
Week 1 (Oct 2-6)	Start looking at the dataset from research that has already been done
Week 2 (Oct 9-13)	Start researching possible methods to analyze the data
Week 3 (Oct 16-20)	Analyze the dataset
Week 4 (Oct 23-27)	Visualize the analyzation
Week 5 (Oct 30-Nov 3)	Build model that categorizes them Categorize the vulnerabilities (if there is enough time)
Week 6 (Nov 6-10)	Create presentation
Week 7 (Nov 13-17)	Start working on the final paper Find reference papers
Week 8 (Nov 20-24)	Finish paper and proofread Practice presentation
Week 9 (Nov 27-Dec 1)	Potential Deadline

Our group will start this project by researching the datasets from relevant research. After finding the data set, we will look for possible methods and libraries to analyze the data. When we find them, our team will start analyzing the dataset and visualize the result using Jupyter Notebook. We have not started the research for the dataset and methods, however, we will start with using Numpy and Pandas libraries to parse the data.

VI. EXPECTED OUTCOMES

From this project, our group expects to analyze the insecurity of the code for different security scenarios created by ChatGPT. We expect our model to perform well enough to confidently categorize the vulnerabilities of a given snippet of code. By categorizing the vulnerabilities, our model will be able to suggest advice for a possible solution for the vulnerability. In addition to the creation of the model, we also expect to further the understanding of auto-generated code, including its benefits and shortcomings, and inform readers how to improve their interactions with AI models as a whole.