International Journal of Security and Privacy in Pervasive Computing
Volume 15 • Issue 1

Do ChatGPT and Other AI Chatbots Pose a Cybersecurity Risk?
An Exploratory Study

Glorin Sebastian, Georgia Institute of Technology, USA*
https://orcid.org/0000-0003-2543-9127

ABSTRACT

The rise of artificial intelligence (AI) has opened up new frontiers in various fields, including natural language processing. One of the most significant advancements in this area is the development of conversational agents (i.e., chatbots), which are computer programs designed to interact with humans through messaging interfaces. The emergence of large language models, such as ChatGPT, has enabled the creation of highly sophisticated chatbots that can mimic human conversations with impressive accuracy. However, the use of these chatbots also poses significant cyber risks that must be addressed. This research paper seeks to investigate the cyber risks associated with the use of ChatGPT and other similar AI-based chatbots, including potential vulnerabilities that could be exploited by malicious actors. As part of this research, a survey was conducted to explore the cybersecurity risks associated with AI-based chatbots like ChatGPT. Further, the paper also suggests mitigation methods that can be used to mitigate these cyber risks and vulnerabilities.

KEYWORDS

Artificial Intelligence (AI), ChatGPT, Cybersecurity, Generative Pretrained Transformer 3 (GPT-3), Natural Language Processing (NLP)

INTRODUCTION

ChatGPT (Generative Pre-trained Transformer) is the Chat Bot introduced by Open AI in November 2022, an AI research and development company, based on a variation of its Instruct GPT model, which is trained on a massive pool of data to answer queries (Open AI. ChatGPT. 2022). ChatGPT uses natural language processing to generate responses to text-based inputs. GPT models are based on the Transformer architecture, which is a neural network architecture that was introduced in the research paper by Vaswani (Vaswani, A. et.al, 2017).

The architecture of ChatGPT is quite complex and involves many layers of neurons. At a high level, the model consists of an encoder and a decoder, that work together to generate responses to various user

DOI: 10.4018/IJSPPC.320225 *Corresponding Author

inputs. The encoder takes in the input text and processes it to create a sequence of hidden states, which are then passed to the decoder. The decoder uses these hidden states to generate the output text one token at a time, in a process known as autoregression. Some of the key features of ChatGPT include:

1. **Large Scale:** ChatGPT is one of the largest language models available, with over 175 billion parameters. This makes it easier for it to understand and generate complex responses.
2. **Conversational**: ChatGPT is designed to engage in natural and flowing conversations, making it appear more human-like in its responses.
3. **Multi-Task**: ChatGPT can perform multiple tasks, including answering questions, summarizing text, and generating creative writing.
4. **Contextual**: ChatGPT can take into account the context of the conversation to provide more relevant and accurate responses.
5. **Personalized**: ChatGPT can be trained on specific datasets to provide personalized responses for specific domains or use cases.
6. **Open Source**: ChatGPT is open source, meaning that developers can modify and customize the model to suit their specific needs.

The below Figure-1 shows the details of the input text and the transformation of data across various layers before the final output is shared. The input text is fed into the Transformer, which processes it to create hidden states. The decoder layer then uses these hidden states to generate the output text, which is produced by the SoftMax layer.

Further, ChatGPT has a large number of parameters and requires a lot of computing power to train and use effectively. However, because it is pre-trained on a large corpus of text, it can be fine-tuned for specific applications with relatively little additional training data. Overall, ChatGPT represents a significant advancement in conversational AI, with the potential to revolutionize various industries, including customer service and healthcare.

CHATGPT ALGORITHM

ChatGPT is based on a variant of the Transformer architecture, which is a deep neural network model that is well-suited for processing sequential data, such as natural language. The Transformer architecture uses self-attention mechanisms to allow the model to focus on different parts of the input

Figure 1.
The process diagram for a ChatGPT-enabled query