साइबर स्वच्छता केन्द्र
CYBER SWACHHTA KENDRA
Botnet Cleaning and Malware Analysis Centre

Full Member
Operational Member
Accredited Member
Global Research Partner
Associate Partner
CVE Numbering Authority (CNA)

Directions by CERT-In under Section 70B, Information Technology Act 2000

Guidelines on Information Security Practices for Government Entities

Blueprint for Reducing Exposure and Defending against AI-Assisted Vulnerabilities Exploitation in Digital Infrastructure

15 Elemental Cyber Defense Controls for Micro, Small, and Medium Enterprises (MSMEs)

Technical Guidelines on SBOM, QBOM & CBOM, AIBOM and HBOM Version 2.0

Cyber Security Guidelines for Smart City Infrastructure

Cyber Security Framework and Guidelines for Space including Satellite Communication

ABOUT CERT-In
- Client's /Citizen's Charter
- Roles & Functions
- Advisory Committee
- Act/Rules/Regulations
- Internal Complaint Committee (ICC)
- RFC2350
- Press
- Tender
- Subscribe Mailing List
- Contact Us

REPORTING
- Incident Reporting
- Vulnerability Reporting
- Feedback

KNOWLEDGEBASE
- Guidelines
- Presentations
- White Papers
- Annual Report

RECRUITMENTS

Home - Advisories

CERT-In Advisory CIAD-2023-0015

Security implications of AI language based applications

Original Issue Date: May 09, 2023

Overview

AI language-based models such as ChatGPT, Bing AI, Jasper AI, Bard AI etc are widely getting recognition and being discussed for its useful and adversarial impacts. These applications can be used by threat actors to target individual and organizations.

Possible adversarial threats that may arise from the use of AI language-based applications and minimization of its impact are described below:

Description

A large language model based on the GPT (Generative Pre-trained Transformer) architecture-based applications are gaining popularity in cyber world. The applications such as ChatGPT, Bing AI, Bard AI, Jasper AI etc are trained with a large set of data from internet using deep learning algorithm transformer. The applications are designed for understanding and generating human-like natural languages, codes, and embedding. These applications have the ability to perform wide range of natural language processing tasks, such as language translation, text completion, question answering, code generation etc.

People are using the AI language-based applications to understand, interpret, and enumerate cyber security contexts, to review security events and logs, to interpret malicious codes and malware samples etc. The applications have the potential to be used in vulnerability scanning, translation of security code from one language to another or transfer of code into natural languages, performing security audit of the codes, VAPT, or integration of application with SOC and SIEMs for monitoring reviewing, and generating alerts.

However, the AI based applications can also be used by threat actors to conduct various malicious activities:

- A threat actor could use the application to write malicious codes for, exploit a vulnerability, conduct scanning, perform privilege escalation & lateral movement, to construct a malware or a ransomware for a targeted system.
- AI based applications can generate output in the form of text as written by human. This can be used to disseminate fake news, scams, generate misinformation, create phishing messages, or produce deep fake texts.
- A threat actor can ask for a promotional email, a shopping notification, or a software update in their native language and get a well-crafted response in English, which can be used for phishing campaigns.
- Creation of fake websites and web pages to host and distribute malwares to users through malicious links or attachments using the domain similar to AI based applications.
- Creation of fake applications impersonating AI based applications.
- Cybercriminals could use AI language models to scrape information from the internet such as articles, websites, news and posts, and potentially taking Personal Identifiable Information (PII) without explicit consent from the owners to build corpus of text data.

The following safety measures may be followed to minimize the adversarial threats arising from AI based applications:

- Educate developers and users about the risks and threats associated with interacting with AI language models
- Verify domains and URLs impersonating AI language-based applications, avoid clicking on suspicious links.
- AI language-based applications are based on learning on large set of internet data. The model can collect all accessible data including sensitive information also. Hence, implement appropriate controls to preserve the security and privacy of data. Do not submit any sensitive information, such as login credentials, financial information or copyright data to such applications.
- Ensure that the text generated is not being used for illegal, unethical activities or for dissemination of misinformation.
- Use content filtering and moderation techniques within organization to prevent the dissemination of malicious links, inappropriate content, or harmful information through such applications.
- As the malicious codes written by threat actors may bypass the existing detection mechanism, making detection more challenging. Enhance and implement the relevant monitoring and security measures to detect these threat activities.
- Secure and conduct regular security audits and assessments of the systems and infrastructure to identify potential vulnerabilities and information disclosures.
- Enable multi-factor authentication (MFA) to prevent unauthorized access of accounts directly to AI based applications and protect user accounts from being compromised.
- Organizations may continuously monitor user interactions with AI language-based applications for any suspicious or malicious activity within their infrastructure.
- Organizations may prepare an incident response plan and establish the set of activities that may be followed, in case of an incident happens.
- Stay up-to-date on the latest security threats and vulnerabilities and take appropriate action to protect yourself and your data.

References

Cloud Security Alliance
https://cloudsecurityalliance.org/artifacts/security-implications-of-chatgpt/

Palo Alto
https://unit42.paloaltonetworks.com/chatgpt-scam-attacks-increasing/

CERT-In

- **Preventing Data Breaches / Data Leaks (CIAD-2021-0004)**)
  https://www.cert-in.org.in/s2cMainServlet?pageid=PUBVLNOTES02&VLCODE=CIAD-2021-0004

- **Mobile based Malware: Methods and Countermeasures (CIAD-2022-0014)**
  https://www.cert-in.org.in/s2cMainServlet?pageid=PUBVLNOTES02&VLCODE=CIAD-2022-0014

Disclaimer

Security Tips for Common Users NEW
ADVISORIES
VULNERABILITY NOTES
RELATED LINKS
- World CERTs
- Antivirus Resources
- FAQ

Performance Management
india.gov.in
Public Grievances (DAR&PG)
mygov मेरी सरकार
data.gov.in Open Government Data (OGD) Platform India
ISAC Power NEW
राष्ट्रीय मतदाता सेवा पोर्टल NATIONAL VOTERS' SERVICES PORTAL www.eci.nic.in | www.nvsp.in
Modes of Digital Payments
DIGITAL PAYMENT suraksha
RTI
Collaboration and Engaging with CERT-In
State/Sectoral CSIRT NEW

The information provided herein is on "as is" basis, without warranty of any kind.

Contact Information

Email: info@cert-in.org.in
Phone: +91-11-24368572

Postal address

Indian Computer Emergency Response Team (CERT-In)
Ministry of Electronics and Information Technology
Government of India
Electronics Niketan
6, CGO Complex, Lodhi Road,
New Delhi - 110 003
India