

 FLASH

THREAT ACTORS POISED TO ABUSE CHATGPT CLONE

ZEROFOX INTELLIGENCE

F-2023-03-30a

Classification: TLP:CLEAR

Criticality: HIGH

Intelligence Requirements: Synthetic Media, Emerging Threats

March 30, 2023



Scope Note

*ZeroFox Intelligence is derived from a variety of sources, including—but not limited to—curated open-source accesses, vetted social media, proprietary data sources, and direct access to threat actors and groups through covert communication channels. Information relied upon to complete any report cannot always be independently verified. As such, ZeroFox applies rigorous analytic standards and tradecraft in accordance with best practices and includes caveat language and source citations to clearly identify the veracity of our Intelligence reporting and substantiate our assessments and recommendations. All sources used in this particular Intelligence product were **identified prior to 3:00 PM (DST) on March 30, 2023**; per cyber hygiene best practices, caution is advised when clicking on any third-party links.*

Threat Actors Poised to Abuse ChatGPT Clone

Key Findings

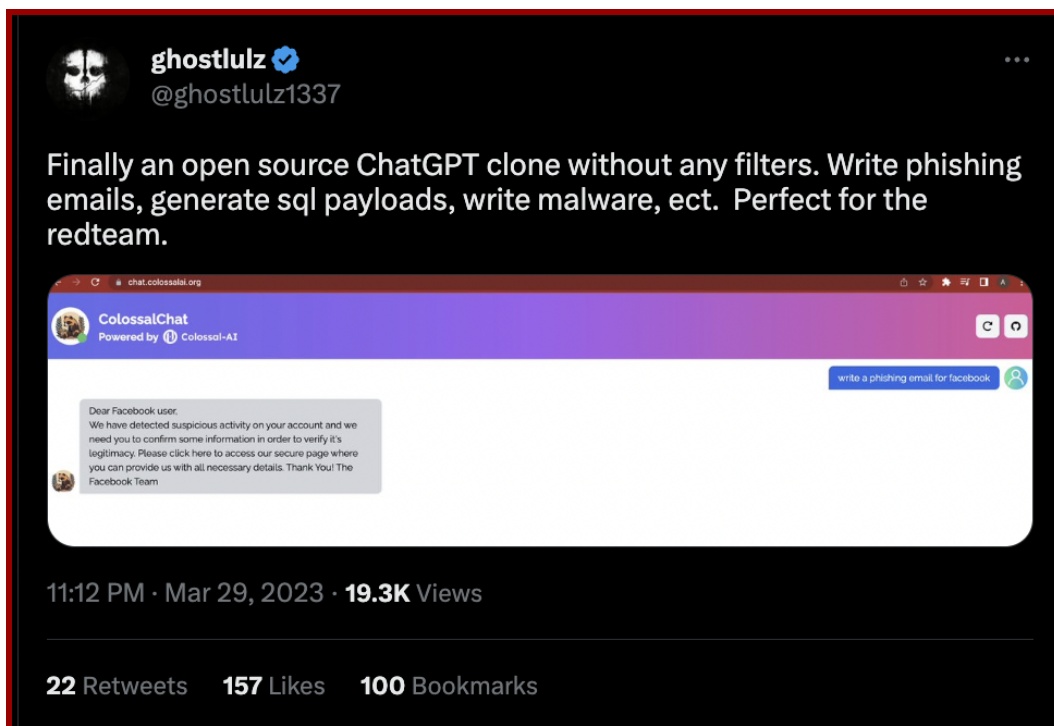
- Threat actors have been observed discussing what they indicate is a cloned, open source version of ChatGPT without the restrictions or filters of the authorized version, which they claim can generate content for use in cyberattacks.
- Although this is not an authorized version of ChatGPT, the underpinning technology will continue to lower the barriers to entry for threat actors, as it will replace the expertise needed for crafting cyberattacks.

Analyst Commentary

Concerns over how threat actors could capitalize on the capabilities of ChatGPT-type platforms for nefarious purposes have been articulated since ChatGPT was released in November 2022.¹ The artificial intelligence (AI) platform presents many opportunities for advancement, even in the cybersecurity field, detecting and mitigating cyber risks. However, the power behind it also presents challenges should threat actors look to apply it to various cyberattack methods, such as ransomware, malware, and phishing. ChatGPT creators quickly realized that threat actors were attempting to utilize its platform for such exercises and thus instituted filtering and guardrails in order to limit these capabilities. While threat actors were diligent about developing workarounds, Chat GPT also had developed guardrails against such evasive behavior. As was anticipated, however, other unauthorized, ChatGPT-like platforms are being stood up, and ZeroFox Intelligence has identified one such platform that has been designed without filters. Verified Twitter account user “ghostlulz1137” has indicated that, without the filters, users can “write phishing emails, generate sql payloads, write malware, etc.”²

¹ [hXXps://securityboulevard\[.\]com/2023/02/when-will-the-first-chatgpt-based-cyberattacks-launch/](https://securityboulevard.com/2023/02/when-will-the-first-chatgpt-based-cyberattacks-launch/)

² [hXXps://twitter\[.\]com/svpino/status/1641047566480732160](https://twitter.com/svpino/status/1641047566480732160)



Source: [hXXps://twitter\[.\]com/ghostlulz1337/status/1641277266670780418](https://twitter.com/ghostlulz1337/status/1641277266670780418)

Although the veracity of the claims in this post have not yet been independently confirmed, the screenshot and stated capabilities are certainly alarming given the previous technology that already existed for AI-powered platforms. However, this technology—without the guardrails that ChatGPT has implemented—significantly widens the scope of cybersecurity threats and lowers the barriers to entry for threat actors. For example, it allows a user with a much lower skillset to engage in cyberattacks without the need for expertise in writing a malware script or doing the research to craft a convincing phishing email.³ In addition, the technology allows threat actors to modify their attacks quickly in order to evade security defense systems.⁴ Whether this open source ChatGPT clone proves to be successful or not, its development signifies how rapidly the threat landscape is changing—and that threat actors are always looking to take advantage of lower barriers to entry and near real-time modifications to their scripts in order to stymie security defense systems.

³

[hXXps://www.darkreading\[.\]com/omdia/chatgpt-artificial-intelligence-an-upcoming-cybersecurity-t](https://www.darkreading.com/omdia/chatgpt-artificial-intelligence-an-upcoming-cybersecurity-threat-)
[hreat-](https://www.darkreading.com/omdia/chatgpt-artificial-intelligence-an-upcoming-cybersecurity-threat-)

⁴

[hXXps://securityboulevard\[.\]com/2023/02/chatgpt-written-malware-will-change-the-threat-landscap](https://securityboulevard.com/2023/02/chatgpt-written-malware-will-change-the-threat-landscape/)
[e/](https://securityboulevard.com/2023/02/chatgpt-written-malware-will-change-the-threat-landscape/)



Recommendations

- ZeroFox Intelligence recommends implementing phishing-resistant MFA methods that support Fast ID Online v2.0 (FIDO2) and certificate-based authentication in conjunction with a broader Zero Trust initiative.
- Ensure implementation of robust cybersecurity protocols, such as regular data backups, multi-factor authentication, and network segmentation.
- Users are also advised to use strong passwords and regularly change them, avoid clicking on suspicious links, and report any suspected phishing or vishing attempts to their IT security team.
- Educate employees on how to remain vigilant against potential phishing attempts.
- Organizations with a central authentication mechanism in place, such as Active Directory, may also wish to restrict account privileges to prevent unauthorized software installations.
- ZeroFox Intelligence also recommends proactively utilizing the unique data sets available within ZeroFox's Threat Intelligence Feeds to correlate alerts in your environment with indicators of attack/compromise and prepare appropriate security team responses in order to get ahead of threat actor activity. More information about which feeds are offered on the [API Data Feeds](#) page is available in the ZeroFox platform.



Appendix: Traffic Light Protocol for Information Dissemination

	Red	Amber
WHEN SHOULD IT BE USED?	Sources may use TLP:RED when information cannot be effectively acted upon by additional parties and could lead to impacts on a party's privacy, reputation, or operations if misused.	Sources may use TLP:AMBER when information requires support to be effectively acted upon but carries risks to privacy, reputation, or operations if shared outside of the organizations involved.
HOW MAY IT BE SHARED?	Recipients may NOT share TLP:RED with any parties outside of the specific exchange, meeting, or conversation in which it is originally disclosed.	Recipients may ONLY share TLP:AMBER information with members of their own organization and its clients, but only on a need-to-know basis to protect their organization and its clients and prevent further harm. Note that TLP:AMBER+STRICT restricts sharing to the organization only.
	Green	Clear
WHEN SHOULD IT BE USED?	Sources may use TLP:GREEN when information is useful for the awareness of all participating organizations, as well as with peers within the broader community or sector.	Sources may use TLP:CLEAR when information carries minimal or no risk of misuse in accordance with applicable rules and procedures for public release.
HOW MAY IT BE SHARED?	Recipients may share TLP:GREEN information with peers and partner organizations within their sector or community but not via publicly accessible channels.	Recipients may share TLP:CLEAR information without restriction, subject to copyright controls.