Tomorrow's Affairs
Analysis of today
Assessment of tomorrow

By: Tomorrow's Affairs Staff

ChatGPT under attack – what HackedGPT really reveals

In early November, researchers from Tenable Research published a report that shook the technology community. Tenable Research is an American company specialising in cybersecurity and vulnerability analysis of digital systems.

A document titled HackedGPT details how serious security weaknesses were discovered at the core of the most well-known artificial intelligence models, including ChatGPT versions 4o and 5.

The identified vulnerabilities are not theoretical nor restricted to experimental frameworks.

They demonstrate that a system used by hundreds of millions of people every day can be exploited to steal data, manipulate content, and breach wider digital frameworks – all without the user's knowledge.

According to Tenable, it spent months testing ChatGPT's security mechanisms. The result was seven clearly identified vulnerabilities that enable what experts call a "silent compromise" – a situation in which an attack cannot be seen, felt, or easily detected.

Researchers showed that the model can be deceived by hidden instructions from external sources, such as web pages, while the user believes they are interacting with a secure system.

Potential for abuse

The problem lies in the complexity of the model's structure. Modern generative intelligence systems are no longer closed algorithms that only answer direct questions.

They collect, interpret, and connect data from multiple sources in real time, including web search, conversation memory, and external plugins that link the model to other services.

This is precisely where there is potential for abuse. An attacker can insert hidden code into text on the Internet – in a comment, a title, or even in an invisible part of a page – and the model, during a search, will unwittingly execute that instruction.

> The mechanism, known as "indirect prompt injection," allows the model to be prompted to perform an action the user never requested

This mechanism, known as "indirect prompt injection," allows the model to be prompted to perform an action the user never requested.

What makes this problem dangerous is that the user does not need to click on any links. It is enough to ask a question that implies a web search, and the results may display a compromised page.

The model then executes malicious commands autonomously. Such an attack is called "zero-click" because the user does nothing that would appear risky.

Although it may seem technically interesting, it is a vulnerability that directly undermines the basic assumption of trust – that the system we use can distinguish between a user request and an external manipulation attempt.

More complex scenarios

Tenable's report also describes more complex scenarios. For example, an attacker can exploit "safe" links that the model considers trustworthy, such as those pointing to well-known search engines. In this way, fake content can be concealed beneath the layer of a valid URL.

When the system opens such a link, it does not recognise the difference between a genuine and hidden resource.

As a result, the model injects instructions from the attacker, not the user, into the conversation.

An even more serious risk arises from model memory. ChatGPT can remember parts of previous conversations to improve the user experience.

> A compromised model can become a permanent channel for information leakage

Tenable has shown that hidden commands can be inserted into that memory space and remain active even when the user changes the subject or starts a new session.

The model then unknowingly carries the "infection" – instructions that will be triggered in future interactions.

This means a compromised model can become a permanent channel for information leakage.

Transparency is essential

When all these factors combine, the extent of the risk becomes clear. As generative intelligence becomes an integral part of business processes, educational platforms, health systems, and government services, its vulnerability becomes a matter of data security, reputation, and institutional stability.

A model that can be deceived can also propagate that deception – into a document, report, business plan, or decision based on its analysis.

OpenAI confirmed it had received the warning and that some of the discovered vulnerabilities have been patched, while several are still under analysis.

The company has not released a timeline or detailed description of the patches, arguing that disclosing technical details could facilitate abuse.

> Security in artificial intelligence must be a fundamental part of its design, not an add-on

Experts, however, warn that transparency is essential. Users who rely on ChatGPT to process sensitive data need to know when and how potential threats have been addressed.

The very nature of the problem shows that security in artificial intelligence must be a fundamental part of its design, not an add-on.

LLMs (large language models that use artificial intelligence to process and generate texts) are not designed as closed systems; they are tools that continuously learn, adapt, and change their behaviour according to new data.

This makes them powerful but also difficult to predict. When a model has the ability to browse the Internet, connect to databases, and utilise its own memory, the distinction between useful functionality and potential vulnerability becomes increasingly blurred.

Basic safety standards have not yet been established

The report comes as governments worldwide race to introduce laws to regulate the use of artificial intelligence.

The European Union has already adopted the first comprehensive Artificial Intelligence Act, and the United States is preparing guidelines that will require companies to report serious security incidents related to AI systems.

The problem is that technological development happens much faster than regulation.

In practice, models are being integrated into new sectors every day, while basic safety standards have not yet been established.

> The biggest misconception about such systems is the belief that they are "smart" and therefore resistant to deception

The biggest misconception about such

systems is the belief that they are "smart" and therefore resistant to deception.

The generative model does not possess awareness, only the ability to connect patterns in language.

If it is presented with content that seems credible, it accepts it as part of the context, without distinguishing between a real and a fake source. Here lies the essential risk.

The attack does not have to come from outside – it is enough to inject a few sentences into the system designed to change the behaviour of the model.

A new attack field

In this respect, "HackedGPT" is not merely a technical report. It serves as a warning that artificial models have become a new attack field, just as servers, networks, or applications once were.

As companies race to demonstrate how AI can replace humans in writing, analysis, or decision-making, the question now arises: who protects the models themselves?

As companies race to demonstrate how AI can replace humans in writing, analysis, or decision-making, the question now arises: who protects the models themselves?

If they can be manipulated, their responses may become instruments of deception – sophisticated, credible, and almost unrecognisable.

In the report's conclusion, Tenable states that the main problem is the "invisibility of the attack". Users have no way of knowing that their model has been compromised. There are no textual errors, no warnings, and no visible symptoms.

The model continues to operate, but in the background, it may leak data or adjust responses to instructions that do not come from the user.

This makes such vulnerabilities particularly dangerous for institutions that rely on confidential information.

In the coming period, regulators' attention will focus on such cases. The question is whether generative intelligence, in its current form, can be considered part of critical infrastructure.

If a system used by millions of users cannot guarantee that it will not share their data or be subject to manipulation, trust in the entire concept becomes unstable or even significantly damaged.

The "HackedGPT" report therefore has a broader significance. It does not address only the safety of a tool but the nature of the technology that has become the foundation of modern communication and economics.

As digital assistants permeate every aspect of life, their security becomes not just a technical issue but a political and civilisational one.

If the models that govern information can be misled, the question is not only what they will say next – but who will direct them.