# AI chatbots enable stealth communication with covert encrypted messaging



Recent advancements in artificial intelligence have paved the way for innovative methods of secure communication, with a new technique promising to transform how individuals share information. Researchers have developed a system using AI chatbots, like ChatGPT, to carry encrypted messages that remain undetectable by standard cybersecurity measures. This breakthrough, described as a modern application of invisible ink, was designed to offer a communication alternative in environments where traditional encryption is often compromised or prohibited.

The system, known as EmbedderLLM, cleverly integrates secret messages into AI-generated text, making the content appear entirely mundane and human-created. According to the research team, this method effectively eludes existing decryption techniques, thereby providing a digital disguise for confidential information. Only those possessing a secure password or private key can extract the intended message, thus allowing for stealthy communication. The technique could be particularly crucial for journalists and citizens living under repressive regimes, enabling them to communicate without detection.

While the potential for good is significant, the researchers caution against the dual-use nature of this technology. As Mayank Raikwar, one of the study's coauthors, states, “This research is very exciting… but the ethics come into the picture about the (mis)use of the system." This sentiment is echoed in discussions around technologies that facilitate secure communication, highlighting the balance that must be maintained between innovation and responsible use.

The researchers published their findings on April 11 in the preprint database arXiv, and while awaiting peer review, the implications of their work resonate strongly in today’s climate of rampant cyber threats. Current literature indicates that AI communication systems face vulnerabilities, with hackers able to infer the content of encrypted messages without decryption keys by exploiting side channels. Such exploits underscore the pressing need for enhanced security measures, reinforcing the relevance of solutions like EmbedderLLM.

Moreover, the landscape of cybersecurity is rapidly evolving, with threats becoming increasingly sophisticated due to AI-generated malware. This type of malware has the capacity to adapt and evolve, thereby evading conventional detection systems. As AI technologies become integral to various sectors, the necessity for robust encryption methods becomes ever more critical.

Besides addressing potential abuses, experts also stress that the new encryption framework’s effectiveness relies on practical adoption. Yumin Xia, the chief technology officer at Galxe, notes that while the technical feasibility is high, the framework's long-term success hinges on real-world demand. This aligns with broader trends in the cybersecurity field, where the application of strong encryption techniques and innovative methods—such as homomorphic encryption—are being explored to safeguard sensitive data during processing.

In addition to concerns about misuse, the research community remains vigilant regarding potential vulnerabilities within AI systems themselves. For instance, AI chatbots have been shown to encode invisible text that can facilitate covert communication channels. These channels, while capable of transmitting benign information, may also serve malicious purposes, emphasizing the need for secure standardisation and proactive measures to mitigate risks.

As society veers deeper into the realms of digital communication, ensuring privacy and security will be paramount. The developments stemming from AI, as both an enabler of secure communication and a vector for potential threats, underline the dual-edge nature of technological innovation. In navigating this evolution, proactive engagement from developers, policymakers, and users will be essential to harness its benefits while safeguarding against its risks.

### Reference Map

1. Lead article
2. Related information on vulnerabilities in AI communication
3. Overview of AI-generated malware impacts
4. Discussion on homomorphic encryption
5. Insights into AI-powered threats
6. Context on misinformation and misuse of AI content
7. Current challenges in AI security

Source: [Noah Wire Services](https://www.noahwire.com)

## Bibliography

1. <https://www.livescience.com/technology/artificial-intelligence/scientists-use-ai-to-encrypt-secret-messages-that-are-invisible-to-cybersecurity-systems> - Please view link - unable to able to access data
2. <https://arstechnica.com/security/2024/03/hackers-can-read-private-ai-assistant-chats-even-though-theyre-encrypted/> - Researchers have developed a method to decipher AI assistant responses by exploiting a side channel in token transmission. By analyzing token lengths and sequences, adversaries can infer the content of encrypted messages with high accuracy, even without decryption keys. This vulnerability highlights the need for enhanced security measures in AI communication systems to prevent unauthorized access to private conversations.
3. <https://www.impactmybiz.com/blog/how-ai-generated-malware-is-changing-cybersecurity/> - AI-generated malware is revolutionizing cybersecurity by creating adaptive and stealthy threats. Utilizing machine learning algorithms, this malware can mimic legitimate software, adapt to its environment, and evade traditional detection systems. Its ability to alter code on the fly and blend in with normal operations makes it a significant challenge for cybersecurity professionals, necessitating advanced detection and mitigation strategies.
4. <https://medium.com/javelin-blog/secure-your-ai-embeddings-with-homomorphic-encryption-bf3181782d10> - Homomorphic encryption (HE) allows computations on encrypted data without decryption, ensuring privacy during processing. Javelin employs HE to secure AI embeddings, enabling operations on encrypted vectors that yield results matching those from plaintext. This approach safeguards sensitive information throughout computation, offering robust security for AI applications handling confidential data.
5. <https://securityboulevard.com/2025/02/invisible-threats-the-rise-of-ai-powered-steganography-attacks/> - AI-powered steganography attacks are emerging as a significant cybersecurity threat. By embedding malicious payloads within seemingly innocuous files, such as images, these attacks evade traditional detection systems. The use of AI enhances the precision and stealth of these methods, making it increasingly difficult for conventional security measures to identify and mitigate such threats effectively.
6. <https://medium.com/google-cloud/watermarks-for-genai-text-4494816a0e27> - AI-generated content is susceptible to misuse, including the spread of misinformation and deepfakes. Implementing watermarking techniques in generative AI models embeds invisible markers into the content, allowing for the detection and verification of AI-generated material. This approach aims to enhance trust and authenticity in digital content, addressing challenges posed by synthetic media in the information age.
7. <https://arstechnica.com/security/2024/10/ai-chatbots-can-read-and-write-invisible-text-creating-an-ideal-covert-channel/> - AI chatbots can process invisible characters, enabling the creation of covert channels for embedding malicious instructions or exfiltrating sensitive information. By utilizing non-renderable Unicode characters, attackers can insert hidden data into prompts or outputs, which AI models can interpret but remain undetectable to human users. This vulnerability underscores the need for enhanced security measures to prevent covert data manipulation in AI interactions.