

# Analyzing the Latest Trends and Security Threats of Generative AI with ChatGPT

Hyun-Che Song, Hye-In Lee, and Il-Gu Lee<sup>1</sup>

Sungshin Women's University, Seoul, Korea  
 {ssongzz, lhynee, iglee}@sungshin.ac.kr

**Keywords**— ChatGPT, Generative AI, Latest Trends, Security Threats





## 1 Introduction

The fourth industrial revolution ushered in a host of new technologies like artificial intelligence (AI), blockchain, big data, metaverse, autonomous driving, drones, and 6G mobile communications. According to the "EQST 2023 First Half Security Trend" report published in June, generative AI, including discriminatory AI technologies such as clustering, classification, and regression, is a focal point. Generative AI learns patterns and rules from extensive data, including images, videos, audio, and text, to understand user queries and generate new content. This study examines the latest trends and security threats in representative generative AI services like ChatGPT.

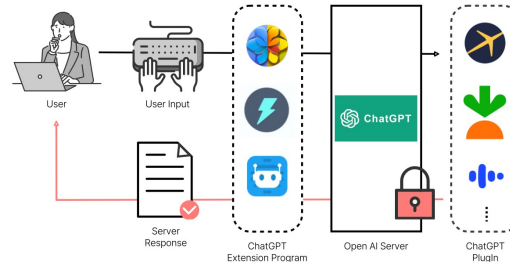
## 2 Generative AI and LLMs Services

Large language models (LLMs), as generative artificial intelligence technology, learn based on a vast database of texts collected from web documents and produce compelling sentences by imitating the statistical patterns of language. Because an artificial intelligence model encompasses more than tens of billions of parameters in general, the model can learn complex language patterns and meanings, exhibiting excellent performance in various analogical tasks. ChatGPT, first proposed by Open AI in 2018, is a representative service in which LLMs are implemented.

ChatGPT has undergone four version updates over six years. The newest version, GPT-4, is a multi-modal model that can generate longer responses than previous versions and performs better in high-level inference tasks. The data input by the user are entered into the Open AI server, and the input value is analyzed to generate an appropriate response which is returned to the user.

Company	LLMs(Large Language Models)	Service
 OpenAI	GPT 4.0	ChatGPT
 Google	PaLM (Pathway Language Model)	Bard
 Meta	LLaMa (Large Language Model Meta)	Meta AI
 NAVER	OCEAN	Hyper CLOVA X

**Table 1** : LLMs & Front-End Services from Primary Generative AI Enterprise



**Figure 1** : How ChatGPT service works

### 3 Security Threats of Generative AI

**[3.1 Invalid information]** Generative AI uses data learning to actively generate plausible results based on input data. However, it has limitations in terms of accuracy due to its reliance on probability. Factors such as data bias, outdated data, and hallucination can reduce result accuracy. These limitations can lead to biased or inaccurate outcomes, potentially causing social confusion and incorrect decision-making by organizations or individuals.

**[3.2 Model Abuse]** Unlike general programming languages, AI models do not clearly distinguish between data and commands. This can be exploited for cybercrime if AI services are used without proper restrictions on input data. Malicious questions can be abused to bypass and attack guidelines, or the policies or even write malicious code. Such abuses can lead to the spread of phishing crimes and false information using deepfake technology by mimicking the user's text style.

**[3.3 Fake AI Model Service]** In addition to the vulnerabilities of the generative AI model, attacks that abuse user interests in AI services also occur. Hackers can steal reliable AI model names to induce users to accidentally access squat URLs, install malicious extensions, install malicious code on user devices, or steal personal information through fake applications.

**[3.4 Threats of leakage of sensitive information]** The extensive data learned by the AI model can include major personal information within an institution. This sensitive information can be unexpectedly leaked in the process of processing data and generating responses. Even if de-identification processing is performed, de-identified sensitive information can be estimated through the combination of the AI model's result value and external information, leading to serious security breaches.

### 4 Conclusion

Generative AI, represented by ChatGPT, has a great impact on individuals and industries as a whole, increasing productivity and efficiency when incorporated into various fields; however, it is important to clearly recognize and respond to the limitations of current technology and security threats. There is a need to develop preemptive countermeasures for generative AI abuse as well as policy and technical countermeasures to detect, analyze, and block the level of possible attack threats. Additionally, technical security measures should be developed to prepare for attacks that are becoming more sophisticated, and education on the safe use and awareness of generative AI should be expanded. Now is the time to make a concerted effort to utilize this versatile technology safely and effectively.

### References

- [1] National Intelligence Service (NIS), National Security Research Institute (NSR), Security Guidelines for Generative AI Utilization such as ChatGPT.
- [2] Yang Ji-Hoon and Yoon Sang-Hyuk, Beyond ChatGPT to the Generative AI era: Cases of media and content-generating AI services and ways to secure competitiveness.
- [3] Shielders SK, EQST 2023 First Half Security Trend Report.
- [4] Korea Internet & Security Agency (KISA), KISA INSIGHT Digital & Security Policy 2023 VOL.3 - ChatGPT Security Threats and Implications.