Scientific and Social Research, 2025, Volume 7, Issue 9
https://ojs.bbwpublisher.com/index.php/SSR
Online ISSN: 2981-9946
Print ISSN: 2661-4332

Security Tools Based on Large Language Models

Jiaxin Li, Baocheng Wang*, Yiwei Wei, Jiaxin Dong, Fang He

Rocket Force University of Engineering, Xi'an 710025, China

**Author to whom correspondence should be addressed.*

Abstract: This article focuses on the field of security tools based on large language models (LLMs), systematically reviewing the core products, security strategies, and technical solutions launched by leading US technology companies and research teams in 2024. In June 2024, OpenAI released its security strategy for large language models and a user data protection plan. The former established a multi-dimensional protection system covering infrastructure, sensitive data/model weight protection, and model review, while the latter enhanced privacy protection by giving ChatGPT users the right to choose how their data is used and by default restricting the use of some data for training purposes.

Keywords: Large language model; SAFE verification tools; Data security; Threat detection

Online publication: September 9, 2025

1. Introduction

In recent decades, with AI's rapid advancement, Large Language Models (LLMs) have become a core focus driving natural language processing breakthroughs. Recent years have seen their research advance significantly via academia-industry collaboration, with ChatGPT as a key milestone—its human-like text generation gained wide attention and sparked discussions on reshaping human-computer interaction.

LLMs' technical evolution deeply influences the AI community. Breaking traditional language limits, they revolutionize AI algorithm development, shifting from task-specific small models to general large-scale pre-training, opening new research avenues. The application of large language models (LLMs) is expanding, but they also pose risks such as harmful outputs, data leakage, and "jailbreak attacks." The research on security tools is of crucial significance. It can filter out harmful content, prevent vulnerabilities, and ensure the compliance of information in critical fields such as healthcare and finance; it can also suppress model hallucinations and strengthen the data security defense line, providing core support for the safe implementation and large-scale application of LLMs in various industries. In this paper, we mainly explored the core contents and individual characteristics of several security tools for large language models, in order to promote the development of large language models towards a safer and more efficient direction.

2. Security tools based on large language models

2.1. Data security protection for users of OpenAI's large language models

On June 6, 2024, the US-based OpenAI publicly disclosed its security strategy for large language models, providing technical references for the public to understand the model development process. The disclosed content includes:

1) In terms of basic security architecture, AzureEntraID, which provides various management services such as identity authentication and access protection, will be integrated with the internal identity verification and authorization control framework to achieve secure verification of session creation, identity authentication, and abnormal logins.
2) In terms of workload orchestration, Google's open-source container platform Kubernetes will be used to coordinate, manage, and protect workloads, and for tasks with higher risks, gVisor, an open-source product from Google, will be used for running.
3) In terms of sensitive data protection, authorization for storing, managing, and researching sensitive data will be granted through keys, and role-based access control will be adopted to limit access to data, while access management services will be used to achieve flexible customization of access permissions.
4) In terms of model weight protection, the resources related to model weights will be stored in OpenAI's internal network, and the flow of sensitive data will be controlled at the exit, and access to accounts involving sensitive model weights will require multiple approvals.
5) In terms of model review and testing, security tests will be conducted through internal and external "security red teams" to evaluate existing security standards, study security systems and security measures, etc.

On June 13, 2024, OpenAI released the user data security protection plan for large language models [1]. OpenAI stated that although user data can help improve the performance of its existing models, it also understands those users who do not want OpenAI to utilize their personal data [2]. To protect the data privacy of these users, OpenAI proposed relevant protection measures, including:

1) Free and enhanced versions of ChatGPT users can choose in the settings whether to use their conversations as training data;
2) In ChatGPT, users can select "temporary chat records", and their data will not be used for training OpenAI's models;
3) The company does not use API, ChatGPT Enterprise, and ChatGPT Team user data for model training by default.

2.2. Microsoft Copilot for Security

Copilot for Security is a product combining large language models and threat intelligence released by Microsoft (**Figure 1**). It integrates threat intelligence data, algorithm models, and cloud computing capabilities, assisting analysts in generating efficient analysis results [3]. The main functions of Copilot for Security include:

1) Utilizing the reasoning ability of the large language model to reconstruct attack paths and output visualized attack paths;
2) Incorporating a rich set of predefined analysis strategies, such as malware analysis and general event analysis;
3) Limiting the logical reasoning ability of the large language model within specific scenarios through the form of Promptbook scripts, reducing misleading answers caused by knowledge "hallucinations";

4) Generating security reports based on the semantic understanding and integration capabilities of the large language model, enhancing the friendliness of human-computer interaction;
5) Achieving data compliance through a series of methods, ensuring that user data is not used to train the model, and reducing the possibility of data leakage in the large language model.

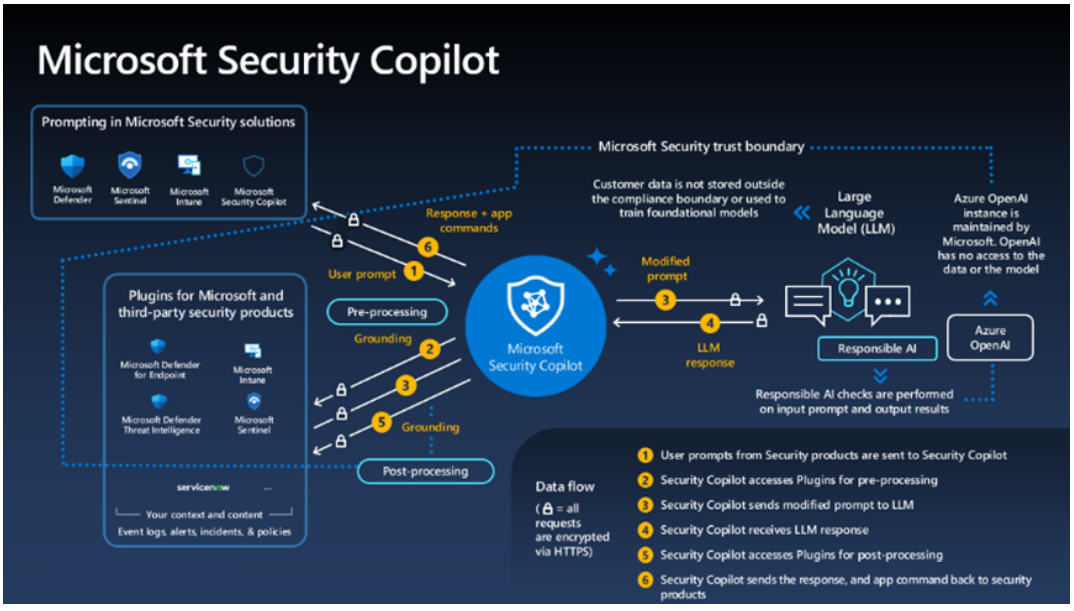

Figure 1. Basic information of Copilot for Security

On April 1, 2024, Microsoft officially released Copilot for Security (International Edition) globally. This is the first independent generative artificial intelligence solution in the global information security field, which can help security and IT professionals comprehensively understand security situations, take actions faster, and enhance team professional skills.

Copilot for Security processes over 78 trillion large-scale security signals every day and combines large language models and security-specific models to provide deep insights for enterprises and guide users' subsequent actions. Copilot for Security can help security and IT professionals strengthen their skills, collaborate more effectively, understand security situations more comprehensively, and respond faster. In the official release version, Copilot for Security includes the following new features:

1) Customized instruction manual (Promptbook), supporting customers to create and save a series of natural language instructions for daily security work.
2) Knowledge base integration (preview version), enabling users to integrate Copilot for Security with business logic and execute according to their own step-by-step guidelines.
3) Connecting from Defender EASM to custom external attack surfaces to identify and analyze the latest information on the external attack surface risks of the organization.
4) Microsoft Entra audit logs and diagnostic logs, providing other suggestions for security investigations or IT problem analysis (involving audit logs related to specific users or events), and summarizing in natural language.
5) Usage reports, providing suggestions through dashboards on how teams use Copilot to help teams discover more optimization opportunities.

2.3. The security AI workbench platform of Google Cloud Company

The US-based Google Cloud Company mainly empowers its product lines through large language models to enhance the intelligence level in various business domains. The Security AI Workbench released by Google Cloud is the industry's first scalable platform supported by the Google security large model Sec-PaLM (**Figure 2**) [4]. This platform aims to utilize artificial intelligence to enhance threat detection and analysis, and to respond to and prevent new threats by providing reliable, relevant, and actionable intelligence. The main functions of Security AI Workbench include:

1) Virus Total Code Insight: Using the Sec-PaLM model to analyze the behavior of malicious scripts and identify potential threats;
2) Mandiant Breach Analytics for Chronicle: Combining Google Cloud and Mandiant threat intelligence to conduct contextual and timely responses to activity vulnerabilities within the user environment;
3) Assured OSS: Utilizing large language models to incorporate more open-source software (OSS) packages into the OSS vulnerability management solution;
4) Mandiant Threat Intelligence AI: Based on Mandiant's vast threat map, using the Sec-PaLM large language model to help customers quickly find, summarize, and respond to relevant threats;
5) Chronicle AI: Supports searching for billions of security events and can interact with the results, ask follow-up questions, and quickly generate detection results;
6) Security Command Center AI: Can convert complex attack diagrams into understandable human language, including affected assets and solutions. In addition, it will provide an AI-based risk summary for Google Cloud's security, compliance, and privacy check results [5].

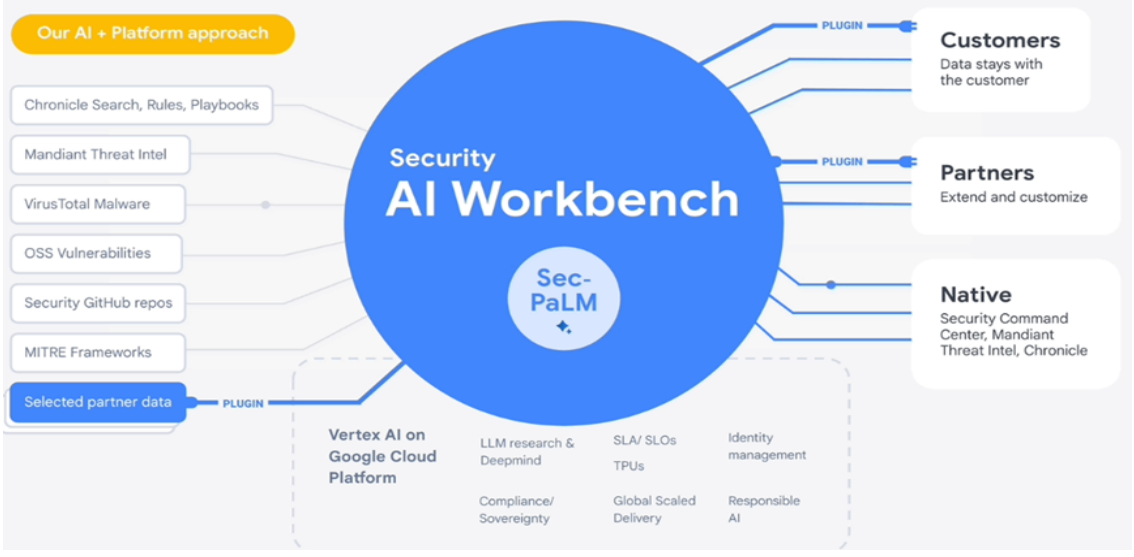

Figure 2. Basic information about the security AI workbench

Google Cloud claims that this Security AI Workbench model has been fine-tuned for security use cases and combines Google's powerful security intelligence (such as Google's visibility into threat situations and Mandiant's first-hand intelligence on vulnerabilities, malware, threat indicators, and hacker behavior patterns). Google believes that this security artificial intelligence workbench can effectively address the three major challenges in cybersecurity: threat overload, cumbersome tools, and talent shortage [6].

2.4. Google's Frontier Safety Framework

On May 17, 2024, the artificial intelligence team of Google's subsidiary DeepMind launched an artificial intelligence safety framework called the Frontier Safety Framework, which is used to detect risks in artificial intelligence models (**Figure 3**) [7]. It claims to be able to actively identify "the artificial intelligence capabilities that may cause significant risks in the future", and point out to researchers "at which levels the relevant models may be exploited by hackers."

This framework is a set of protocols, emphasizing the importance of identifying and mitigating potential risks during the development of artificial intelligence models, aiming to actively identify the artificial intelligence capabilities that may cause serious harm in the future, and establish mechanisms for detecting and mitigating risks. It is introduced that the 1.0 version of Frontier Safety Framework released by DeepMind mainly includes three key capabilities, namely:

1) The ability to identify the threshold of serious harm that the model may have;

2) Regularly evaluating the model to detect when these key threshold values are reached;

3) When the model reaches the early warning assessment, activate the mitigation plan.

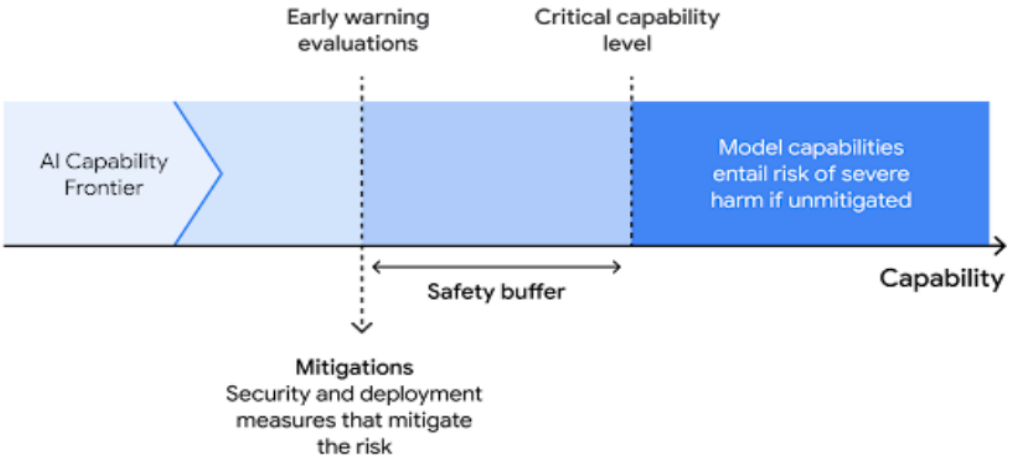

Figure 3. The relationship between the Frontier Safety Framework and three key capabilities

Meanwhile, this framework also proposes two mitigation measures:

1) Preventing model weight leakage;

2) Limiting access to and expression of key capabilities during deployment.

DeepMind research indicates that the most likely areas where the capabilities of future large language models will pose serious risks are autonomy, biosecurity, cybersecurity, and machine learning research and development. This framework also proposes key capability thresholds for these four areas.

DeepMind stated that the company "has been constantly pushing the boundaries of artificial intelligence", and the models developed by the company have changed their perception of the possibilities of artificial intelligence. Although the company believes that future artificial intelligence technologies will bring valuable tools to society, they also realize that the risks of related artificial intelligence technologies may have a devastating impact on society, so they are gradually enhancing the security and controllability of the models.

Currently, DeepMind is still developing the Frontier Safety Framework. In the future, DeepMind will

collaborate with other companies, academia, and legislators to improve this framework, and plans to officially launch it in 2025.

2.5. Google's SAFE verification tool

The illusion problem of the generated answers by large language models has always been a concern in the industry, hindering the deployment and production of large language models in real scenarios. In April 2024, the artificial intelligence team of Google's DeepMind, in collaboration with researchers from Stanford University, published a research paper titled "Long-form factuality in large language models." The researchers conducted an in-depth exploration of long-form factual questions and conducted a comprehensive assessment of the performance of language models in long-form factual matters. The research team released a new prompt dataset called LongFact, which is used for benchmarking long-factual content in large language models, including 2,280 guiding questions covering 38 different topics. The research team also launched an artificial intelligence fact verification tool called SAFE (Search-Augmented Factuality Evaluator). Both LongFact and SAFE have been open-sourced on GitHub.

The 2,280 high-quality guiding questions included in the LongFact dataset originated from multiple authoritative sources such as Wikipedia and news reports. Through dual checks of automated screening and manual review, the questions were ensured to be able to test the depth of the model's knowledge while avoiding factual errors or subjective biases. Thus, LongFact became a solid foundation for measuring the long-factual nature of language models.

The SAFE tool, based on large language models, can analyze, process, and evaluate long responses generated by chatbots to verify the accuracy and authenticity of the responses. The implementation steps include splitting the answers into individual items to be verified, correcting each item, and then comparing it with Google search results. Additionally, the tool will also check the relevance of each fact to the original question.

The innovative evaluation method of SAFE utilizes the language model itself and its interaction with Google's search engine to automatically evaluate whether each knowledge point generated by the model is accurate, relevant, and can be self-consistent. Different from traditional methods that rely on manual judgment or only focus on superficial correctness, SAFE can verify the accuracy of the facts generated by the model in real-world scenarios and detect the model's ability to generate meaningful information by leveraging Google search.

On the LongFact dataset, researchers conducted benchmark tests on a total of 13 language models from four series (Gemini, GPT, Claude, PaLM-2). The results showed that larger language models typically performed better in long-factual matters. SAFE was consistent with human judgment results 72% of the time and received higher recognition in 76% of randomly selected 100 controversial cases. Moreover, SAFE was more than 20 times more efficient than hiring an artificial annotation team, demonstrating its effectiveness as an efficient means for evaluating the long-factual nature of large language models.

3. Conclusion

Overall, the current security tools based on large language models have demonstrated strong capabilities and significant value in multiple fields [8]. From a functional perspective, these tools cover multiple key aspects of security protection, such as OpenAI's security strategy which builds a multi-dimensional model and data security protection system, Microsoft Copilot for Security which implements efficient analysis functions such as attack path reconstruction and security report generation, Google Security AI Workbench which has significant advantages

in threat detection and intelligence utilization, and Frontier Safety Framework and SAFE tools which provide solutions for the two key pain points of AI model risk detection and factual assessment of generated content.

In terms of data security and compliance, all companies attach great importance to it and take various measures to protect user data privacy, such as restricting data usage for model training, encrypting data transmission and storage, etc., effectively reducing the risk of data leakage [9]. From an industry impact perspective, these tools not only enhance the work efficiency and skill level of security professionals, helping enterprises better cope with security threats, but also provide an important direction for the intelligent development of the information security field, promoting the standardized application and innovative development of generative artificial intelligence in the security field. In the future, with the continuous improvement of related frameworks and iterative upgrades of tools, combined with multi-party cooperation, security tools based on large language models will play a greater role in ensuring network security and responding to emerging threats.

Disclosure statement

The authors declare no conflict of interest.

References

[1] Baral S, Saha S, Haque A, 2024, An Adaptive End-to-End IoT Security Framework Using Explainable AI and LLMs. 2024 IEEE 10th World Forum on Internet of Things (WF-IoT), 469–474.

[2] Singh A, 2023, Exploring Language Models: A Comprehensive Survey and Analysis. 2023 International Conference on Research Methodologies in Knowledge Management, Artificial Intelligence and Telecommunication Engineering (RMKMATE), 1–4.

[3] Dowswell K, 2024, Considering Responsible AI with GitHub Copilot, in Programming with GitHub Copilot: Write Better Code--Faster! Wiley, New Jersey, 217–227.

[4] Washizaki H, Yoshioka N, 2024, AI Security Continuum: Concept and Challenges. 2024 IEEE/ACM 3rd International Conference on AI Engineering – Software Engineering for AI (CAIN), 269–270.

[5] Das BC, Hadi Amini M, Wu Y, 2024, Security and Privacy Challenges of Large Language Models: A Survey. arXiv e-prints, Art. no. arXiv: 2402.00888.

[6] Wu Q, Wang Y, 2023, Research on Intelligent Question-Answering Systems Based on Large Language Models and Knowledge Graphs. 2023 16th International Symposium on Computational Intelligence and Design (ISCID), 161–164.

[7] Jawhar S, Miller J, Bitar Z, 2024, AI-Based Cybersecurity Policies and Procedures. 2024 IEEE 3rd International Conference on AI in Cybersecurity (ICAIC), 1–5.

[8] Ots K, 2025, Overview of Generative Artificial Intelligence Security, in Securing Microsoft Azure OpenAI. Wiley, New Jersey, 1–17.

[9] Park J, You G, Ji Y, et al., 2024, Security Requirements for Fully Automated AI Systems to Exercise and Ensure the Rights of Data Subjects. 2024 19th Asia Joint Conference on Information Security (AsiaJCIS), 107–112.