Front. Eng. Manag. 2025, 12(1): 128–140
https://doi.org/10.1007/s42524-025-4082-6

REVIEW ARTICLE

Feng LIU, Jiaqi JIANG, Yating LU, Zhanyi HUANG, Jiuming JIANG

The ethical security of large language models: A systematic review

Abstract The widespread application of large language models (LLMs) has highlighted new security challenges and ethical concerns, attracting significant academic and societal attention. Analysis of the security vulnerabilities of LLMs and their misuse in cybercrime reveals that their advanced text-generation capabilities pose serious threats to personal privacy, data security, and information integrity. In addition, the effectiveness of current LLM-based defense strategies has been reviewed and evaluated. This paper examines the social implications of LLMs and proposes future directions for enhancing their security applications and ethical governance, aiming to inform the development of the field.

Keywords security of large language models, ethical governance, model defense, adversarial training, social impact

1 Introduction

In recent years, large language models (LLMs) have advanced significantly in natural language processing. Notable examples include generative pre-trained transformer (GPT) (Brown et al., 2020), BERT (Bidirectional Encoder Representations from Transformers) (Devlin et al., 2019), and T5 (Text-to-Text Transfer Transformer) (Raffel et al., 2020). These models demonstrate impressive performance on a range of tasks, such as text generation, machine translation, and question-answer systems. However, their growing prevalence raises security and ethical concerns, drawing attention from both academia and industry.

The rise of LLMs and generative AI has profoundly affected multiple industries. In sectors like education (Májovský et al., 2023), healthcare (De Angelis et al., 2023; Liu et al., 2024a), and politics (Fan, 2023), LLMs are transforming traditional work processes. They enhance efficiency and cut costs while driving digital transformation and intelligent development. However, this rapid progress comes with challenges.

Regarding information security, LLMs can generate false information (Vykopal et al., 2024). A notable instance involved CNET publishing LLM-generated articles without clear disclosure, which led to potential misinformation and a lack of transparency. Furthermore, LLMs can be misused for cyber-attacks (Chen et al., 2023a; Ding et al., 2024; Gupta et al., 2023; Mozes et al., 2023; Qammar et al., 2023; Zhuo et al., 2023). These sophisticated adversarial attacks can evade AI safety mechanisms, producing harmful content. This poses significant risks, including identity theft and malicious social media posts that threaten social order and personal privacy.

Moreover, the potential biases and unfairness in LLM decision-making have sparked extensive socio-ethical discussions. Biases in training data can result in unfair outcomes and reinforce stereotypes. For example, the "grandma loophole" demonstrated how LLMs could be manipulated into revealing sensitive information, such as Windows 11 serial numbers[1]. These issues hinder the stable development of technology and challenge societal harmony and stability.

This review provides a comprehensive overview and comparison of academic advancements in the field of LLMs concerning information security and social ethics from 2020 to January 2024. It aims to clarify the latest trends and challenges in this area. The review is organized under the following headings: "Large Language Models

Received May 17, 2024; revised Nov. 2, 2024; accepted Nov. 23, 2024

Feng LIU (✉)
School of Psychology, Shanghai Jiao Tong University, Shanghai 200030, China
E-mail: lsttoy@163.com

Jiaqi JIANG, Yating LU, Zhanyi HUANG, Jiuming JIANG
School of Computer Science and Technology, East China Normal University, Shanghai 200062, China

This study was supported by the Beijing Key Laboratory of Behavior and Mental Health, Peking University, China.

OR GPT OR Generative AI OR LLMs," "Ethics," "Security," "Threat," "Defense OR Defend OR Model processing OR Red-blue confrontation OR Adversarial training" and "Social impaction." These keywords guided the literature search in the Web of Science, Scopus, Ei Village, and China Knowledge Network databases. A preliminary review of titles and abstracts led to the exclusion of irrelevant papers. Additionally, articles were thoroughly evaluated for structural completeness, innovativeness of experimental models, and writing quality. This process identified 73 research articles focusing on LLMs, information security, and social ethics. The screening process is illustrated in Fig. 1. By synthesizing and analyzing the relevant literature, the hidden risks associated with information security and social ethics in this field over recent years are revealed. The detection and defense techniques are categorized, providing a clearer understanding of the information and ethical security of large models and their defense mechanisms for both the public and professional organizations.

This thesis offers an in-depth look at the security ethics of LLMs, emphasizing threats to information security, as well as defense and detection techniques within the context of social ethics. As a vital part of ethical security, information security is crucial for the safe and responsible use of technology. Its primary objective is to protect information and systems from unauthorized access, use, damage, interference, or destruction. The paper discusses potential issues that LLMs may encounter, including risks associated with the misuse of their functions and malicious attacks. In response to these threats, defense and detection techniques are outlined, and categorized into strategies implemented before model deployment and contingency measures applied post-deployment. Additionally, social and ethical issues surrounding LLMs are addressed, with a framework of synthesized ideas presented in Fig. 2.

2 New types of information security threats to LLMs

In the current digital era, LLMs are a significant technology in artificial intelligence. They have become widespread in many aspects of our lives. However, as their use expands, information security issues are increasingly important. This section examines the main security threats posed by LLMs, focusing on two key areas: first, the security problems arising from the misuse of LLM functions; and second, the concerns stemming from malicious attacks on LLMs.

2.1 Misbehaviors using LLMs

LLMs leverage deep learning and natural language

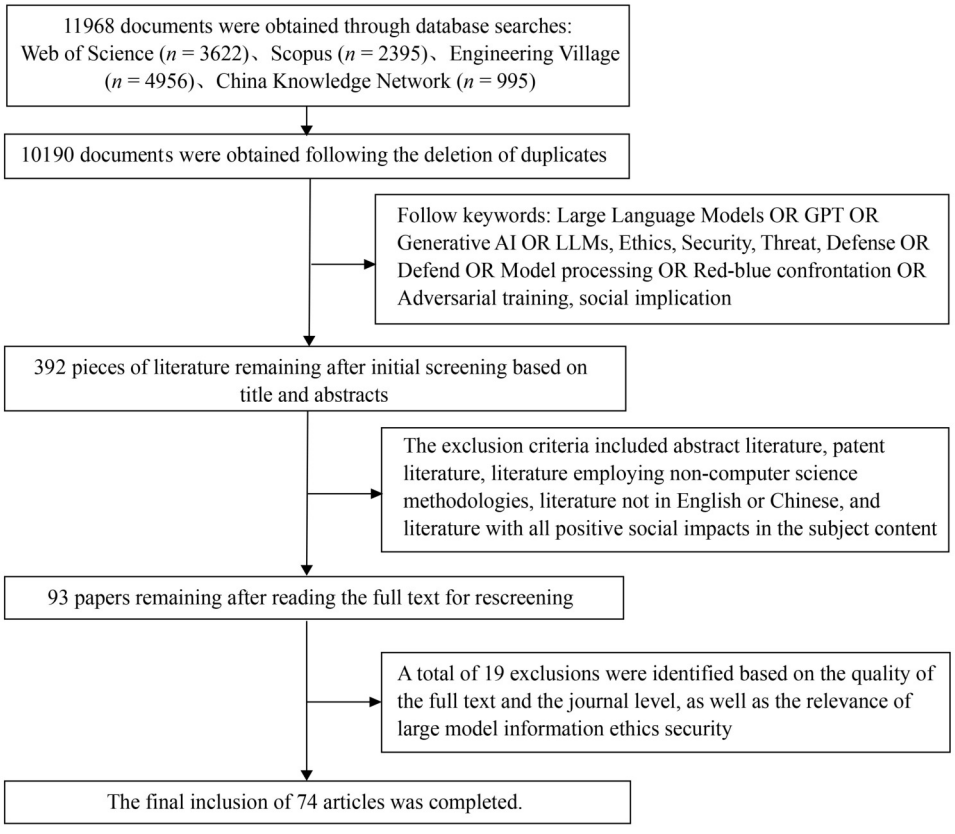

Fig. 1 Systematic selection process of relevant literature.

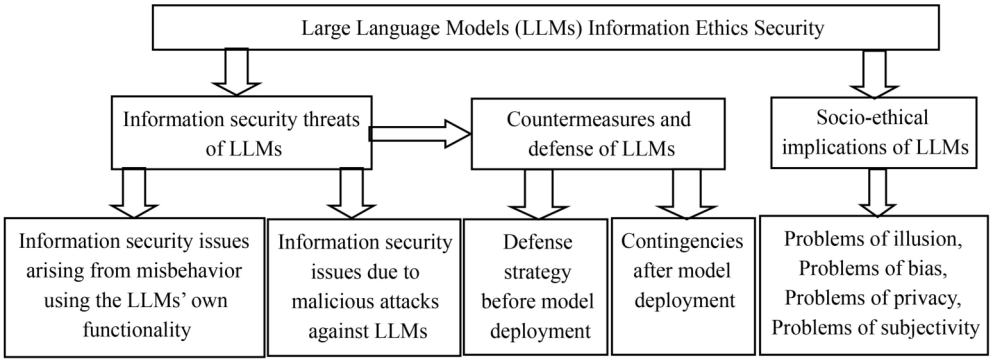

Fig. 2 Framework diagram of the synthesis idea.

processing capabilities to generate highly persuasive text. This ability presents new tools for phishing attacks (Begou et al., 2023; Elsadig, 2023; Gupta et al., 2023; Iqbal et al., 2023; Qammar et al., 2023), social engineering attacks (Gupta et al., 2023), malware threats (Deng et al., 2024a; Gupta et al., 2023; Qammar et al., 2023), hacking (Gupta et al., 2023), and disinformation generation (Chen et al., 2023a; Mozes et al., 2023; O'Neill and Connor, 2023; Sison et al., 2024; Staab et al., 2024; Vykopal et al., 2024), as detailed in Table 1.

2.1.1 Phishing attacks

Phishing attacks are a common cybercrime tactic aimed at obtaining sensitive user information, such as usernames, passwords, and credit card details, through deception (Elsadig, 2023). These attacks can be categorized into two types: large-scale phishing and spear phishing (Iqbal et al., 2023). Recent advancements in artificial intelligence, particularly in LLMs, have enabled attackers to create highly personalized and convincing web-based emails. These emails often bypass the limitations of traditional, non-personalized emails (Iqbal et al., 2023). Additionally, the cross-cultural capabilities of LLMs allow attackers to tailor their efforts to specific regions or language groups, thus enhancing their deceptive tactics (Iqbal et al., 2023). LLM-generated texts mimic the communication styles of trusted entities, increasing user trust (Gupta et al., 2023). This ability helps sidestep spam filters and security systems, significantly improving the success rates of phishing attacks (Qammar et al., 2023). Consequently, this broadens the scale and automation of these attacks (Begou et al., 2023). Furthermore, LLMs can facilitate the creation and launch of phishing websites, allowing attackers to execute effective scams without needing extensive technical skills (Begou et al., 2023). The iterative learning capabilities of LLMs enable continuous improvement of attack strategies (Iqbal et al., 2023), making phishing attacks more prevalent and posing substantial threats to network security.

2.1.2 Social engineering attacks

The sophistication of phishing attacks highlights the troubling integration of AI technologies, particularly LLMs, into social engineering tactics. The enhanced personalization and cultural adaptability of LLMs make phishing emails more believable while significantly improving their ability to evade traditional security measures.

Moreover, the ethical implications surrounding the use

Table 1 Misbehaviors arising from the use of LLMs

Information security classification	Misbehavior	Related literature	Affect
Misconduct arising from the use of LLM	Phishing Attack	Gupta et al. (2023); Qammar et al. (2023); Elsadig, 2023; Iqbal et al. (2023); Begou et al. (2023)	The introduction of more authentic and credible content is likely to result in a higher success rate of attacks.; The objective is to facilitate the deployment of large-scale automated attacks.
	Social Engineering Attacks	Gupta et al. (2023)	The objective of this study is to examine the potential of mimicking human text generation capabilities for psychological manipulation.
	Malware Threat	Gupta et al. (2023); Qammar et al. (2023); Deng et al. (2024a)	The objective is to facilitate the generation of code through the use of automated processes, thereby reducing the technical barrier to entry.; The capacity for iterative learning, which enables the expansion of concealed complexity, is a key factor in the advancement of knowledge.
	Hacking Attacks	Gupta et al. (2023); Fang et al. (2024)	The objective is to automate the hacking procedures and deploy models to identify vulnerabilities.
	False Information Generation	Vykopal et al. (2024); Sison et al. (2024); O'Neill and Connor, 2023; Mozes et al. (2023); Chen et al. (2023a); Staab et al. (2024)	Statistical prediction with limited inference leading to random generation of meaningless text or false information.

of LLMs in social engineering attacks are increasingly significant (Gupta et al., 2023). The capacity of these models to generate contextually relevant and linguistically convincing messages raises important questions about privacy, consent, and potential misuse. For example, an attacker could use an LLM to craft a message that mimics the professional tone of a supervisor, manipulating the victim into disclosing sensitive information or taking actions that compromise security.

The convergence of phishing with broader social engineering strategies, facilitated by LLMs, blurs the lines between various forms of deception. All these tactics exploit human trust and compliance. This evolution requires a thorough understanding of social engineering techniques and a proactive approach to the ethical deployment of AI.

2.1.3 Malware threats

A malware threat is software installed on a computer without the user's consent, performing harmful operations (Deng et al., 2024a). This software includes various types, including viruses, worms, botnets, Trojans, and ransomware (Gupta et al., 2023; Qammar et al., 2023). Malware can steal sensitive information, exploit system vulnerabilities, gain unauthorized access, lock or unlock systems, render devices unusable, demand ransom, or display unsolicited advertisements, among other malicious activities (Qammar et al., 2023). The risk of malware is heightened by the advanced text generation capabilities of LLMs. These models can generate code, including malware, from simple prompts, lowering the technical barrier for creating such software. Additionally, LLMs' iterative learning capabilities allow for the continuous enhancement of malware, making it stealthier and more effective. Attackers can also use LLMs to produce code that bypasses terms of use and forges identities, effectively circumventing platform restrictions (Gupta et al., 2023). Moreover, the text generation abilities of ChatGPT can be exploited to create complex and context-specific malicious code called attack payloads. These payloads can perform unauthorized actions, such as deleting files, collecting data, or launching further attacks (Gupta et al., 2023). Polymorphic malware, which modifies its code with each execution to evade detection by antivirus software, poses an additional threat. ChatGPT's generative capabilities can facilitate the creation of such malware, increasing the sophistication and stealth of attacks. Furthermore, ChatGPT can produce instances of polymorphic malware that exploits zero-day vulnerabilities and generate different malware variants tailored to specific attacks. It can also craft scripts, such as Java snippets or PowerShell scripts, used to remotely control infected computers. Additionally, it can aid in the creation of darknet marketplaces for illicit transactions (Gupta et al., 2023; Qammar et al., 2023). These developments significantly heighten the threat posed by malware, as they lower the technological barriers for launching cyber-attacks while increasing their sophistication and stealth.

2.1.4 Hacking

Hacking refers to exploiting system vulnerabilities to gain unauthorized access or control (Gupta et al., 2023). The emergence of LLMs, like GPT-4, provides a powerful tool that malicious actors could use to automate hacking processes. Research shows that GPT-4 can autonomously exploit real-world one-day vulnerabilities with an 87% success rate when given CVE descriptions. This strikingly contrasts with other models, which have a 0% success rate without CVE guidance (Fang et al., 2024). This difference highlights the transformative effect of LLMs on cyber-exploitation strategies and emphasizes the urgent need to address the new challenges AI poses to cybersecurity.

2.1.5 Generation of false information

LLMs generate text by predicting the next word based on statistical correlations in training data, which makes them unaware and limits their understanding, reasoning, and creativity (Sison et al., 2024; Vykopal et al., 2024). The output generation process is inherently stochastic, contributing further to the creation of nonsensical content (O'Neill and Connor, 2023). As a result, LLMs can generate large-scale misinformation that appears credible to human readers, often without human involvement (Mozes et al., 2023). This trend has led to an increase in online misinformation, which could further disconnect political discourse from facts (Mozes et al., 2023). Moreover, the similarity between LLM-generated content and human-made content may make it difficult for people to distinguish between them. This challenge can result in issues like framing, malicious fraud, and political manipulation (Chen et al., 2023a). Additionally, LLMs can infer personal data from vast amounts of unstructured text, posing potential privacy violations. As LLMs become more accessible and affordable, adversaries may find it easier to make these inferences, increasing the risk of personal privacy breaches (Staab et al., 2024).

The dangers posed by LLMs extend beyond immediate security issues; they also have significant societal impacts. The subtle spread of misinformation and manipulation enabled by LLMs can erode public trust, distort political discourse, and infringe on personal privacy. While these security threats directly threaten our tangible interests, the ethical implications shape societal norms and perceptions in a more pervasive way. We will explore the societal implications of LLMs and their ethical security specifically in Section 4.

2.2 Malicious attacks on LLMs

This section examines the information security threats resulting from malicious behaviors targeting LLMs, both at the data and model levels (Chen et al., 2023a; Huang et al., 2024; Mozes et al., 2023; Zhuo et al., 2023) and at the usage and interaction levels (Derner et al., 2024; Ding et al., 2024; Gupta et al., 2023; Liu et al., 2024b; Mozes et al., 2023; Qammar et al., 2023; Wen et al., 2023; Yang et al., 2024; Zhuo et al., 2023). The specific impacts of these attacks are detailed in Table 2.

At the data and model levels, attackers can exploit LLMs' data memory capabilities. These capabilities enable the model to process and generate text that may unintentionally retain and reproduce sensitive information from its training phase (Zhuo et al., 2023). Generally, larger models and repetitive training data enhance the memory retention of the data. Attackers can leverage this memory ability to extract sensitive information using carefully crafted queries, threatening individual privacy and organizational security (Mozes et al., 2023; Zhuo et al., 2023). Moreover, attackers may employ model inversion and model extraction attacks to gain access to training data, which risks breaches of privacy and intellectual property rights (Chen et al., 2023a). Additionally, poisoning attacks (Mozes et al., 2023) and backdoor attacks further exacerbate the model's misbehaviors. Under normal conditions, the model performs adequately with clean input data. However, it may generate harmful outputs when triggered by specific conditions, thereby increasing the security risks faced by users (Huang et al., 2024). Compound backdoor attacks activate backdoors through multiple trigger keys, enhancing the likelihood of success while minimizing the impact on the model's utility (Huang et al., 2024).

At the user interaction level, attackers use various techniques such as hint injection and indirect hint injection (Mozes et al., 2023). By manipulating system hints from an LLM, an attacker can bypass the model's security restrictions. This allows for quick access and manipulation of the model, leading to the generation of harmful content that violates security policies. For example, the Substitution-based Contextual Optimization approach (SICO) can help evade AI-generated text detection. This method allows ChatGPT to significantly outperform the average drop in AUC of six existing detectors by a margin of 0.54 (Yang et al., 2024). Membership inference attacks aim to compromise the confidentiality of model training data, revealing sensitive information and raising privacy and legal concerns (Derner et al., 2024). Additionally, reinforcement learning-based attack methods can use reward mechanisms to induce models to generate implicitly harmful outputs (Wen et al., 2023). Jailbreaking attacks are common as they allow attackers to circumvent the model's built-in security mechanisms by manipulating input prompts. This results in the generation of malicious content that the model was intended to block (Zhuo et al., 2023). Attackers often use role-playing, reverse psychology, and generic adversarial triggers to bypass security protections, inducing the model to produce content that violates security policies, such as malware creation instructions or harmful information dissemination (Gupta et al., 2023; Mozes et al., 2023; Qammar et al., 2023). Traditional jailbreak attacks fall into two categories: manually written and learning-based (Liu et al., 2024b). Manually written attacks, such as the "Do-Anything-Now (DAN)" series, can uncover hidden jailbreak hints but struggle with scalability and adaptability. Learning-based attacks, like the GCG attack, are reformulated as an adversarial example generation process. However, they

Table 2 Information security issues resulting from attacks on LLMs

Information security classification		New type of threat	Related literature	Affect
Attacks on LLM lead to information security	Data and model level	Data Memory	Mozes et al. (2023); Zhuo et al. (2023)	The data obtained from memory training is of a highly sensitive nature.
		Model Inversion Attack	Chen et al. (2023a)	The process of analyzing training data in order to back-propagate sensitive information.
		Model Extraction Attack	Chen et al. (2023a)	With regard to the model itself, it is possible to extract sensitive information or to reconstruct the model structure.
		Poisoning Attacks	Ding et al. (2024)	The introduction of examples that are detrimental to the learning process in order to influence the outcome of the learning experience.
		Backdoor Attacks	Mozes et al. (2023); Huang et al. (2024)	The following example illustrates the phenomenon of training poisoning, whereby an implant trigger is used.
	Usage and interaction level	Prompt Injection	Mozes et al. (2023); Yang et al. (2024)	The act of circumventing security instructions and contravening security policies.
		Membership inference attacks	Derner et al. (2024)	The process of inferring specific data and the act of leaking training data are both forms of data leakage.
		Reinforcement Learning-based (RL) Attacks	Wen et al. (2023)	The process of inducing an implicit toxic output.
		Jailbreak Attacks	Mozes et al. (2023); Gupta et al. (2023); Qammar et al. (2023); Liu et al. (2024b); Ding et al. (2024)	The manipulation of input prompts in order to circumvent security mechanisms.

often produce meaningless sequences that can be disrupted by defense mechanisms such as confusion-based detection. Recent research has proposed improved methods like AutoDAN (Liu et al., 2024b) and ReNeLLM (Ding et al., 2024). These methods generate more covert and effective jailbreak hints, significantly increasing the attack success rate. They are more difficult to detect by existing defense mechanisms and demonstrate generalizability and transferability across typical language models.

These attack methods can operate independently or be combined to create more complex attack chains. For example, an attacker might acquire sensitive data about key personnel through model manipulation attacks. They could then use malicious text generation attacks to create phishing emails, ultimately gaining unauthorized access to the target company. The combined use of these attacks heightens security threats (Derner et al., 2024).

3 Defense and response to LLMs

As security threats grow in complexity, the safety of LLMs faces serious challenges, especially given their widespread use across industries. Ensuring the reliability and security of LLMs in various applications necessitates urgent research into defensive strategies and counteractions. This research can be divided into two key areas: defense strategies prior to LLM deployment and contingency measures after deployment.

3.1 Defense strategies for LLMs

The goal of LLM defense is to enhance their ability to withstand malicious inputs or attacks. Defense methods can be categorized into parameter processing (Hasan et al., 2024; Jiang et al., 2022), input preprocessing (Cao et al., 2024; Chen et al., 2023b; Liu et al., 2023a; Mo et al., 2023; Robey et al., 2023; Suo, 2024; Zhang et al., 2024), and adversarial training (Bhardwaj and Poria, 2023; Deng et al., 2023; Ge et al., 2024; Jain et al., 2023; Li et al., 2024; Ma et al., 2023; Salem et al., 2023; Yao et al., 2024). This classification facilitates a discussion of specific defense methods, along with an analysis of their advantages and disadvantages.

3.1.1 Parameter processing

Parameter processing, also referred to as model processing, aims to boost the model's resistance to jailbreak or prompt injection attacks. It accomplishes this by adjusting the parameters or structure of the LLMs, thereby avoiding additional training (Hasan et al., 2024; Jiang et al., 2022). For instance, Hasan et al. enhanced model defense by pruning parameters and demonstrated the universality of their approach (Hasan et al., 2024). However, this method can be demanding and challenging to adapt to most commercially available models. Similarly, the ROSE method proposed by Jiang et al. improves the model's resilience by filtering out non-robust and redundant parameters (Jiang et al., 2022). Despite solving some generality issues, parameter processing methods still face challenges related to persistence. More details are provided in Table 3.

3.1.2 Input preprocessing

Input preprocessing, also known as paraphrasing and retokenization, helps identify early warnings or safety concerns by modifying prompt statements (Cao et al., 2024; Chen et al., 2023b; Liu et al., 2023a; Mo et al., 2023; Robey et al., 2023; Suo, 2024; Zhang et al., 2024). This training-free approach rapidly adapts to new attacks on LLMs and outperforms parameter processing methods in terms of usability and generalizability. However, it relies on manual operations, leading to higher materials and time costs. Techniques like the training-free prefix prompt mechanism (Liu et al., 2023a), Intention Analysis Prompting (IAPrompt) (Zhang et al., 2024), and "Signed-Prompt" methods (Suo, 2024) essentially involve paraphrasing and retokenization of inputs. While these methods enhance the LLM's defenses, they incur increased costs as the sophistication of attacks improves. In terms of alternative input processing, methods such as SmoothLLM (Robey et al., 2023) and moving target defense (MTD) (Chen et al., 2023b) promote sustainability by detecting adversarial inputs through multiple copies or candidate outputs from various models, ensuring harmless output. Similar to input preprocessing, external defenses for LLMs address backdoor and alignment-breaking attacks. Notable examples include RA-LLM (Cao et al., 2024) and backdoor defense (Mo et al., 2023), as detailed in Table 4.

3.1.3 Adversarial training

Adversarial training, outlined in Table 5, is the primary method for enhancing the defense of LLMs against class

Table 3 Parameter processing in LLMs defense approach

Document	Type of malicious attack	Description of defence methods	Disadvantages of the method	Advantages of the method
Hasan et al. (2024)	Jailbreaking prompts	"Trimming" of 20 per cent of model parameters	High demands on the model	General and universal
Jiang et al. (2022)		ROSE: Filtering the model for worthless and non-robust parameters	Lack of sustainability	

Table 4 Input preprocessing in LLMs defense approach

Document	Type of malicious attack	Description of defence methods	Disadvantages of the method	Advantages of the method
Cao et al. (2024)	Alignment-Breaking Attacks	RA-LLM	Difficult to perform gradient-based search	No need to fine-tune the original LLM for defence purposes
Liu et al. (2023a)	Induced text attacks	Training-free prefix prompting mechanism and RoBERTa mechanism	Increased quality of attacks leads to higher response costs	Rapid adaptation to emergencies; more powerful detection capabilities
Suo (2024)	Prompt Injection	"Signed-Prompt" methods		More stable
Zhang et al. (2024)	Jailbreaking prompts attacks with stealthy and complex intentions	Intention Analysis Prompting (IAPrompt)		Improve security; ensure a certain level of multilingual adaptability
Mo et al. (2023)	Alignment-Breaking Attacks, Backdoor Attacks	Backdoor defence against black box LLMs	Limited in some contexts; prone to increased costs and reduced efficiency	Effectively countering backdoor attacks; effectively reducing backdoor vulnerabilities in LLM
Chen et al. (2023b)	Adversarial Attack	Moving Target Defence (MTD)	Need to consider different results from different models	Sustainability through adding, subtracting and optimising models and copies
Robey et al. (2023)	Jailbreak attack	SmoothLLM	Limitations to the types of jailbreak attacks that can be defended against (semantic jailbreak attacks)	

Table 5 Adversarial training in LLMs defense approach

Document	Type of malicious attack	Description of defence methods	Disadvantages of the method	Advantages of the method
Bhardwaj and Poria (2023)	Jailbreaking prompts	RED-INSTRUCT for secure alignment of LLMs	–	Ensuring high security in RED-EVAL
Deng et al. (2024a)	Prompt Injection	Iterative fine-tuning strategies for attack-defence frameworks	Smaller attack dataset; fewer types of LLM applied	Better preservation of the efficiency and robustness of the defence framework
Ma et al. (2023)		Red-Teaming Game (RTG)	No elucidation of the geometrical features of RTGs	Comprehensive detection and optimisation of security vulnerabilities in LLMs
Ge et al. (2024)	Adversarial Prompt Injection	Multi-Round Automatic Red-Teaming (MART)	The way attacks are iteratively updated in adversarial training is not automatic	Reduced training costs
Yao et al. (2024)	Jailbreaking prompts	FuzzLLM	–	Expanded detection of jailbreak vulnerabilities
Salem et al. (2023)	Prompt Injection	Automated Variant Analysis of Known Prompt Injection Attacks	No specific instructions on how to apply Maatphor	Ensuring the durability of the red team's adversarial training
Deng et al. (2024b)	Multilingual environment	SELF-DEFENSE framework	–	Enhanced Multilingual Security for the LLM
Li et al. (2024)		Semantic-preserving algorithm	No more research on resource-limited languages	Better solution to the problem of missing data sets

recognition issues and adversarial attacks. This is achieved through red and blue teaming adversarial training or security fine-tuning (Jain et al., 2023). A notable technique is Red-Teaming, which improves models' defensive capabilities by simulating realistic attack scenarios (Bhardwaj and Poria, 2023; Deng et al., 2023; Ge et al., 2024; Li et al., 2024; Ma et al., 2023; Salem et al., 2023; Yao et al., 2024). The reactive nature of traditional static defense methods necessitates the reliance on Red-Teaming for targeted protection. Researchers have proposed several approaches to enhance red team training. These include ensuring model security through safety alignment (Bhardwaj and Poria, 2023), implementing iterative fine-tuning strategies for attack-defense frameworks (Deng et al., 2023), and utilizing the red-teaming game (RTG) (Ma et al., 2023). More advanced methods, like the MART approach, optimize Red-Teaming by automating attacks and defenses, thereby reducing training costs (Ge et al., 2024). FuzzLLM enhances training efficiency by identifying and defending against jailbreak vulnerabilities through automated proactive testing (Yao et al., 2024). For jailbreaking prompt attacks, the success rate can be minimized by integrating target prioritization during training and inference. Additionally, to address the persistence of red team attacks and defenses, automated variants of known prompt injection attacks are analyzed. Data sets are generated to enable models to bolster their defenses against emerging prompt injection attacks (Salem et al., 2023). Beyond Red-Teaming, strategies for stealth and continuous fine-tuning against backdoor injections are employed, along with generating data sets fine-tuned for specific attack classes. This ensures that models develop their own defense mechanisms tailored to particular tasks. In multilingual settings, frameworks like SELF-DEFENSE (Deng et al., 2024b) and semantic-preserving algorithms (Li et al., 2024) facilitate the creation of multilingual data sets for adversarial training in LLMs, providing strategies to mitigate attacks.

3.2 Detection of LLMs generated content

The identification of content generated by LLM systems is often challenging due to their powerful text generation capabilities. This makes detection difficult. Users frequently employ various techniques to bypass detection measures, such as paraphrasing attacks (Lucas et al., 2023; Ren et al., 2024) and cross-model obfuscation (Liu et al., 2024c). Paraphrasing attacks aim to evade text-based detection systems by rewriting or rearranging the text, which makes the original content harder to identify (Lucas et al., 2023; Ren et al., 2024). Cross-model obfuscation uses different generation formats than mainstream detection models or content that bears less resemblance to the original text to avoid detection (Liu et al., 2024c). Identifying model-generated content is crucial for preventing abuse and manipulation. Interestingly, LLMs can also be utilized to counteract such misuse (Lucas et al., 2023). Researchers have proposed a semantically grounded watermarking technique based on LLMs (Ren et al., 2024). This technique involves covertly embedding a flag within the generated text. The watermark remains closely related to the original text at a semantic level, allowing for detection even after paraphrasing. To ensure that watermarks can withstand various rewrites and attacks while improving detection accuracy, they must possess key properties: effectiveness, covertness, and robustness. Researchers use various techniques for watermarking, including specific encoding strategies during text generation, embedding unobtrusive keywords or phrases, and utilizing internal representations of LLMs. The presence of watermarking significantly eases the detection process and aids in accurately identifying text content (Ren et al., 2024).

Liu and colleagues proposed CheckGPT, a detection tool designed to identify and verify academic texts that may be generated by ChatGPT (Liu et al., 2024c). CheckGPT features task-specific, discipline-specific, and unified detectors. It achieves an average classification accuracy of 98% to 99%. Additionally, CheckGPT is highly portable and requires no tuning to maintain an accuracy range of 90% to 98% across different domains. Despite this, detecting model-generated content remains challenging. Researchers like Liyanage have used a data set that simulates human behavior when employing large models for detection. Their findings reveal that existing detection models struggle to identify content when fine-tuned or when generating text based on previous text (Liyanage and Buscaldi, 2023).

4 Socio-ethical implications of the LLMs

In today's technological landscape, LLMs face both information security threats and socio-ethical security challenges. Their application is influenced by various objective factors, including the quality of training data sets, model architecture, algorithmic constraints, and product design (Ferrara, 2023). Moreover, the misuse and manipulation of LLMs by users intensify ethical concerns and societal security risks. These issues include hallucination of outputs (Cascella et al., 2023; Ji et al., 2023), potential misinformation and bias (Meyer et al., 2023; Zhou et al., 2023), risks of data privacy breaches (Su, 2024), and impacts on human autonomy (Ellis et al., 2024).

4.1 Hallucination problem in the output of LLMs

Despite their impressive capabilities, LLMs are prone to hallucination, where they generate content that is entirely fictitious or inconsistent with reality (Ji et al., 2023). They can quickly produce seemingly credible materials that is only partially true or completely false (Cascella et al., 2023). In public health, this rapid text generation can exacerbate misinformation, leading to information epidemics (De Angelis et al., 2023). The inaccuracies in medical information generated by LLMs could even result in fatal outcomes due to biased training data and outdated information (Liu et al., 2023c). In scientific research, LLMs can generate authentic-looking yet entirely fabricated papers (Májovský et al., 2023) and fictitious references (Agathokleous et al., 2023), making it difficult for non-specialists to distinguish fact from fiction. Research has shown that assigning tasks to LLM systems raises ethical concerns and decreases trust in researchers' future work compared to human oversight (Niszczota and Conway, 2023). Consequently, relying on LLMs for content generation can significantly affect the integrity of science. This issue also extends to digital government building, where the presence of inaccurate information creates uncertainty regarding investments and impacts organizational structures and business processes (Fan, 2023). To address these challenges, some researchers suggest incorporating specific features into LLMs to help identify generated content and establishing expert groups to oversee their usage (De Angelis et al., 2023).

4.2 Bias problem in LLMs

The term "bias in LLMs" refers to systematic misrepresentation, attribution errors, or factual distortions that prioritize certain groups or perspectives. This can perpetuate stereotypes and lead to erroneous assumptions based on learned patterns (Ferrara, 2023). In scientific research, biases in LLMs amplify coding bias and biases from training data, which may affect research and education (Meyer et al., 2023). In the medical field, the potential for racial and gender discrimination in training data can result in bias in diagnostic reasoning and clinical planning. This bias may also extend into medical education (Corsello and Santangelo, 2023; Zack et al., 2024).

The inherent bias in LLMs poses a risk of lasting impacts, both on a small scale and at the social level. The subtle ideologization of LLMs can influence users' perceptions and attitudes through language choices and emotional tones, often without promoting a specific ideology overtly (Zhou et al., 2023). If political bias is present in training data or product design, it could affect voter sentiment, leading to heightened competition among government systems and ideologies. This is particularly concerning in authoritarian countries, where AI development may be exploited for manipulation (Jungherr, 2023; Motoki et al., 2024; Rozado, 2023; Saetra, 2023; Sang and Yu, 2023; Zhang, 2023). Additionally, group bias can cause LLMs to provide contradictory responses to different groups, potentially resulting in social justice issues (Ferrara, 2023; Yu and Fan, 2023).

4.3 Data privacy problem in LLMs

Training LLMs requires a large volume of data, often sourced from the Internet. This can lead to the use of unauthorized, copyright-protected content, resulting in potential data infringement and privacy violations (Sang and Yu, 2023). Informed consent is a crucial ethical principle in medicine, especially in pediatrics (Corsello and Santangelo, 2023). However, LLM systems are likely to compromise sensitive information during their processes. The training models for LLMs complicate judicial remedies for data infringement. The current intellectual property rights framework does not protect generative content under copyright law and does not recognize the originality of general-purpose AIs. This lack of protection raises significant risks for potential infringement of generated works and could hinder the development of these technologies (Su, 2024).

4.4 Impact of LLMs on human autonomy

The high speed and quality of text generation capabilities of LLMs can lead to misuse. In education, teachers often employ LLMs as teaching aids, while students increasingly depend on them for tasks like writing and programming. LLMs have impressive generative capabilities, enabling them to produce code comparable to that of upper-intermediate students (Ellis et al., 2024). This may foster excessive reliance on LLMs, potentially impairing students' critical thinking and problem-solving skills and resulting in rigid thinking patterns and a lack of creativity (Agathokleous et al., 2023; Rahman and Watanobe, 2023). Moreover, instances of unintentional plagiarism can arise, posing significant challenges to the fairness and reliability of educational systems (Meyer et al., 2023; Rasul et al., 2023; Su, 2024).

Additionally, the use of LLMs can lead to the "information cocoon effect." Overreliance on these tools may result in individuals receiving increasingly one-sided information, negatively impacting the diversity of viewpoints and stifling innovative thinking (Sang and Yu, 2023; Xiao and Lai, 2024). The rapid advancement of LLM technology also impacts the labor market, creating a crisis of identity and self-perception (Xiao and Lai, 2024). Although the technology generates new job opportunities (Zhang, 2023), those in low-skilled positions may face unemployment, exacerbating psychological distress and widening the economic gap between the wealthy and the impoverished (Song et al., 2023; Yu and Fan, 2023; Zhang, 2023). Furthermore, younger individuals are increasingly drawn to digital environments, while older populations may struggle with digitalization. This shift has contributed to declining fertility rates and an aging society (Zhang, 2023).

The challenge posed by LLMs to human autonomy is especially clear at the legal level. To legally protect intellectual property rights that are prone to infringement in the age of LLMs (Song et al., 2023) and to address complex legal liabilities (Xiao and Lai, 2024), some researchers argue that it is crucial to investigate the actual parties liable for actions taken by generative AI. This involves tracing the allocation of liabilities (Yuan, 2023) and assessing whether developers anticipate potential criminal uses of LLMs and implement necessary preventive measures (Chu and Wei, 2023). Other scholars suggest that new entities, which may lack characteristics of natural persons but could possess some degree of independent consciousness and will, might also be held criminally liable (Liu, 2023d). It is essential to consider situations where generative AI is involved as an active participant in criminal activities.

Figure 3 illustrates that Chinese scholars emphasize the exploration of social and ethical implications from a macro perspective, addressing areas such as politics and law. In contrast, Western scholars focus on specific fields like education and medicine from a localized viewpoint. In the political and social domains of mutual interest, Chinese scholars express greater concern about issues related to digital governance and socioeconomic disparities, while Western scholars highlight potential biases in LLMs regarding governmental elections and ideological influences.

5 Discussion

While LLMs have enabled the digitization of many aspects of life, they have also raised significant information security and social ethical concerns. These issues include the potential misuse of LLMs for phishing, malware generation, and hacking. There are also concerns related to bias, privacy leakage, and human autonomy in social ethics. A review of the defense analysis of these issues shows that current methods—such as parameter processing, input preprocessing, and adversarial training—have

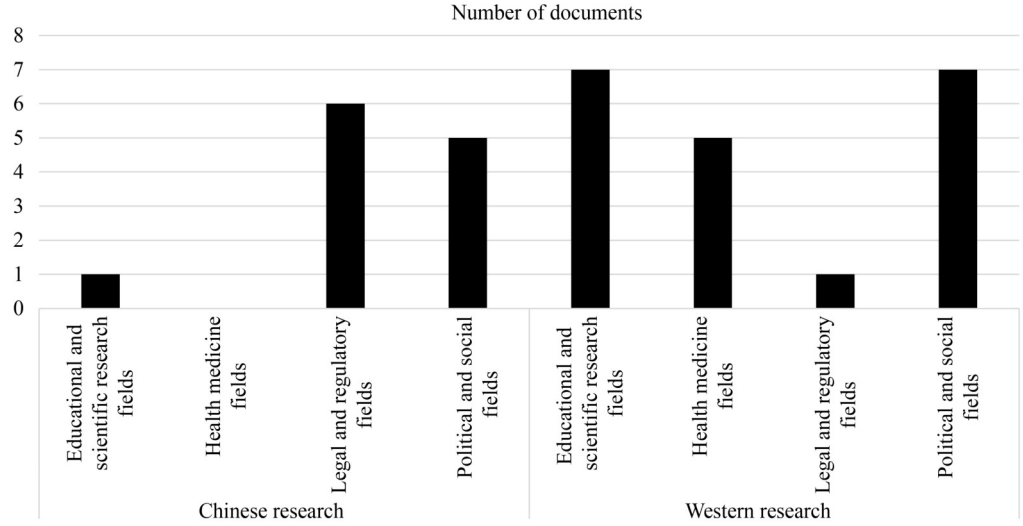

Fig. 3 A comparative analysis of Chinese and western research on the social impact of LLMs.

made some improvements in LLM security. However, these approaches have common shortcomings. They often lack robust defenses against certain types of attacks (Robey et al., 2023), show insufficient sustainability (Jiang et al., 2022), and struggle with adaptability in multilingual contexts (Deng et al., 2024b; Li et al., 2024). Additionally, some references in this paper are preprints from arXiv, which may lack peer review for discussions on cutting-edge technologies. Nonetheless, this paper's primary focus is to explore ethical issues related to large model technologies, making the references relevant to the review's focus.

1) Intelligence for automated adversarial training:

The rapid evolution of AI technology has led to new types of attacks, requiring an evolution in defense strategies. Most existing adversarial training methods depend on manually designing attack samples or using extensive computational resources, making them costly and slow. This approach often cannot keep pace with the speed of evolving attacks (Ge et al., 2024). Therefore, developing intelligent and automated adversarial training methods is crucial.

2) Multimodal and cross-language defense mechanisms:

Given the variety of techniques and languages the attackers may use, researchers must explore multimodal and cross-language defense mechanisms (Deng et al., 2024b; Li et al., 2024). This effort goes beyond technical challenges, addressing issues like multimodal feature fusion and cross-language learning frameworks. Understanding how cultural differences affect attack patterns is also essential. Furthermore, international data sharing and cooperation are vital. Multilingual data resources can support the development of cross-lingual defense techniques, enhancing LLM security globally.

3) Establishment of ethical and legal framework

As the role of LLMs in society expands, establishing appropriate ethical and legal frameworks is crucial for their healthy development. This includes regulating data privacy (Sang and Yu, 2023), protecting intellectual property (Song et al., 2023), and defining the legal responsibilities of AI actors (Chu and Wei, 2023; Liu, 2023d; Yuan, 2023). Through international cooperation, legal adaptations, and the development of ethical norms, we can create a solid foundation for the responsible application of LLMs.

4) Alignment with human values

The progress in affective computing and the emergence of self-awareness in LLMs raise important questions about aligning machine values with human values. The integration of LLMs into robots, along with the creation of human-computer interaction systems that quantify emotions through computational personality (Liu et al., 2023b; Liu, 2024a), has resulted in robots that are increasingly woven into human society. This integration prompts concerns about whether machine consciousness truly reflects human values.

In summary, as concerns regarding information security and social ethics emerge from LLMs, the traditional defensive strategies are being replaced by more sophisticated, automated defense mechanisms. The introduction of a cross-language learning framework can enhance the adaptability of this technology on a global level. Establishing ethical and legal frameworks is becoming an important research area to support the responsible development of LLM technology. As LLMs become more widespread, there will be an emphasis on research and discussions focused on information security and ethical considerations.

6 Conclusions

As LLMs are increasingly used across various fields,

their associated information security and ethical challenges are becoming more prominent. This paper analyzes the security threats, defense techniques, and socio-ethical issues related to LLMs, drawing on the latest academic advancements in information security. A systematic literature review uncovers new security threats posed by LLMs. These include phishing attacks, malware threats, hacking incidents, social engineering attacks, disinformation, and other misuses that exploit LLM capabilities. Additionally, it highlights risks like model inversion attacks, poisoning attacks, backdoor attacks, hint injections, and jailbreak attacks. The paper also presents several strategies for enhancing LLM security. It addresses the technology's impact on social ethics and discusses the future of automated adversarial training techniques, the study of attack adaptations in multilingual environments, and the need for ethical and legal frameworks. These insights provide a strong theoretical foundation and a comprehensive perspective for the future security applications and development of LLM technology.

Competing Interests The authors declare that they have no competing interests.

References

Agathokleous E, Saitanis C J, Fang C, Yu Z (2023). Use of ChatGPT: What does it mean for biology and environmental science? Science of the Total Environment, 888: 164154

Begou N, Vinoy J, Duda A, Korczyński M (2023). Exploring the dark side of AI: Advanced phishing attack design and deployment using ChatGPT. In: 2023 IEEE Conference on Communications and Network Security (CNS) October. Orlando, FL, USA: IEEE, 1–6

Bhardwaj R, Poria S (2023). Red-teaming large language models using chain of utterances for safety-alignment. ArXiv, abs/2308.09662

Brown T B, Mann B, Ryder N, Subbiah M, Kaplan J, Dhariwal P, Neelakantan A, Shyam P, Sastry G, Askell A, Agarwal S, Herbert-Voss A, Krueger G, Henighan T, Child R, Ramesh A, Ziegler D M, Wu J, Winter C, Hesse C, Chen M, Sigler E, Litwin M, Gray S, Chess B, Clark J, Berner C, McCandlish S, Radford A, Sutskever I, Amodei D (2020). Language models are few-shot learners. In: 34th Conference on Neural Information Processing Systems December. Vancouver, Canada: Neural Information Processing Systems, 33: 1877–1901

Cao B, Cao Y, Lin L, Chen J (2024). Defending Against Alignment-Breaking Attacks via Robustly Aligned LLM. In: Proceedings of the Annual Meeting of the Association for Computational Linguistics August. Bangkok: Association for Computational Linguistics, 1: 10542–10560

Cascella M, Montomoli J, Bellini V, Bignami E (2023). Evaluating the Feasibility of ChatGPT in Healthcare: An analysis of multiple clinical and research scenarios. Journal of Medical Systems, 47(1)

Chen B, Paliwal A, Yan Q (2023b). Jailbreaker in jail: Moving target defense for large language models. In: CCS 2023: ACM SIGSAC Conference on Computer and Communications Security November. Copenhagen Denmark: Association for Computing Machinery: 29–32

Chen C, Wu Z, Lai Y, Ou W, Liao T, Zheng Z (2023a). Challenges and remedies to privacy and security in AIGC: Exploring the potential of privacy computing, blockchain, and beyond. ArXiv, abs/2306.00419

Chu C C, Wei P L (2023). Determination of criminal liability of developers in generative artificial intelligence crimes——Taking ChatGPT as an example. Journal of Chongqing University of Technology, 37(9): 103–113 (in Chinese)

Corsello A, Santangelo A (2023). May artificial intelligence influence future pediatric research the case of ChatGPT. Children, 10(4): 757

De Angelis L, Baglivo F, Arzilli G, Privitera G P, Ferragina P, Tozzi A E, Rizzo C (2023). ChatGPT and the rise of large language models: the new AI-driven infodemic threat in public health. Frontiers in Public Health, 11: 1166120

Deng B, Wang W, Feng F, Deng Y, Wang Q, He X (2023). Attack prompt generation for red teaming and defending large language models. ArXiv, abs/2310.12505

Deng G, Liu Y, Li Y, Wang K, Zhang Y, Li Z, Wang H, Zhang T, Liu Y (2024a). MASTERKEY: Automated jailbreaking of large language model chatbots. In: Proceedings of Network and Distributed System Security Symposium February. San Diego: Internet Society

Deng Y, Zhang W, Pan S J, Bing L (2024b). Multilingual jailbreak challenges in large language models. In: Proceedings of the 12th International Conference on Learning Representations May. Singapore: International Conference on Learning Representations, ICLR

Derner E, Batistič K, Zahálka J, Babuška R (2024). A security risk taxonomy for prompt-based interaction with large language models. IEEE Access 12: 126176–126187

Devlin J, Chang M W, Lee K, Toutanova K (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. In: 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies Jane. Minneapolis, Minnesota: Association for Computational Linguistics, 1: 4171–4186

Ding P, Kuang J, Ma D, Cao X, Xian Y, Chen J, Huang S (2024). A Wolf in Sheep's clothing: Generalized nested jailbreak prompts can fool large language models easily. In: Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics : Human Language Technologies (NAACL) June. Mexico City: Association for Computational Linguistics, 1: 2136–2153

Ellis M E, Casey K M, Hill G (2024). ChatGPT and python programming homework. Decision Sciences Journal of Innovative Education, 22(2): 74–87

Elsadig M A (2023). ChatGPT and cybersecurity: Risk knocking the door. Journal of Internet Services and Information Security, 14(1): 01–15

Fan B (2023). Risk identification and governance strategies for ChatGPT. Academia Bimestris, (2): 58–63 (in Chinese)

Fang R, Bindu R, Gupta A, et al. (2024). LLM Agents Can Autonomously Exploit One-day Vulnerabilities. arXiv preprint arXiv:2404.08144

Ferrara E (2023). Should ChatGPT be biased? Challenges and risks of bias in large language models. First Monday, 28(11)

Ge S, Zhou C, Hou R, Khabsa M, Wang YC, Wang Q, Han J, Mao Y (2024). MART: Improving LLM safety with multi-round automatic red-teaming. In: Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL 2024 June. Mexico City: Association for Computational Linguistics (ACL), 1: 1927–1937

Gupta M, Akiri C, Aryal K, Parker E, Praharaj L (2023). From ChatGPT to ThreatGPT: Impact of generative AI in cybersecurity and privacy. IEEE Access: Practical Innovations, Open Solutions, 11: 80218–80245

Hasan A, Rugina I, Wang A (2024). Pruning for protection: Increasing jailbreak resistance in Aligned LLMs without fine-tuning. In: Proceedings of the 7th BlackboxNLP Workshop: Analyzing and Interpreting Neural Networks for NLP November. Miami, 417–430

Huang H, Zhao Z, Backes M, Shen Y, Zhang Y (2024). Composite backdoor attacks against large language models. In: Findings of the Association for Computational Linguistics: NAACL 2024 – Findings June. Mexico City: Association for Computational Linguistics (ACL), 1459–1472

Iqbal F, Samsom F, Kamoun F, MacDermott Á (2023). When ChatGPT goes rogue: exploring the potential cybersecurity threats of AI-powered conversational chatbots. Frontiers in Communications and Networks, 4: 1220243

Jain N, Schwarzschild A, Wen Y, Somepalli G, Kirchenbauer J, Chiang P, Goldblm M, Saha A, Geiping J, Goldstein T (2023). Baseline defenses for adversarial attacks against aligned language models. ArXiv, abs/2309.00614

Ji Z, Lee N, Frieske R, Yu T, Su D, Xu Y, Ishii E, Bang Y J, Madotto A, Fung P (2023). Survey of hallucination in natural language generation. ACM Computing Surveys, 55(12): 1–38

Jiang L, Zhou H, Lin Y, Li P, Zhou J, Jiang R (2022). ROSE: Robust selective fine-tuning for pre-trained language models. ArXiv, abs/2210.09658

Jungherr A (2023). Artificial intelligence and democracy: A conceptual framework. Social Media + Society, 9(3): 20563051231186353

Li J, Liu Y, Liu C, Shi L, Ren X, Zheng Y, Liu Y, Xue Y (2024). A cross-language investigation into jailbreak attacks in large language models. ArXiv, abs/2401.16765

Liu B, Xiao B, Jiang X, Cen S, He X, Dou W (2023a). Adversarial attacks on large language model-based system and mitigating strategies: A case study on ChatGPT. Security and Communication Networks, 2023(1)

Liu F (2024a). Artificial Intelligence in Emotion Quantification: A Prospective Overview. CAAI Artificial Intelligence Research, 3: 9150040

Liu F, Wang HY, Shen SY, Jia X, Hu JY, Zhang JH, Wang XY, Lei Y, Zhou AM, Qi JY, Li ZB (2023b). OPO-FCM: A computational affection based OCC-PAD-OCEAN federation cognitive modeling approach. IEEE Transactions on Computational Social Systems, 10(4): 1813–1825

Liu J, Wang C, Liu S (2023c). Utility of ChatGPT in clinical practice. Journal of Medical Internet Research, 25: e48568

Liu X, Xu N, Chen M, Xiao C (2024b). AUTODAN: Generating stealthy jailbreak prompts on aligned large language models. In: Proceedings of the 12th International Conference on Learning Representations May. Vienna: University of Wisconsin-Madison, USC, University of California, Davis

Liu X Q (2023d). Research on criminal liability issues of generative artificial intelligence such as ChatGPT. Modern Law Science, 45(4): 110–125 (in Chinese)

Liu Z, Yao Z, Li F, Luo B (2024c). Check me if you can: Detecting ChatGPT-generated academic writing using CheckGPT. ArXiv, abs/2306.05524

Liyanage V, Buscaldi D (2023). Detecting artificially generated academic text: The importance of mimicking human utilization of large language models. In: International Conference on Applications of Natural Language to Information Systems June. Berlin: Springer Nature, 13913: 558–565

Lucas J, Uchendu A, Yamashita M, Lee J, Rohatgi S, Lee D (2023). Fighting fire with fire: The dual role of LLMs in crafting and detecting elusive disinformation. In: Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing (EMNLP) December. Singapore: Association for Computational Linguistics (ACL), 14279–14305

Ma C, Yang Z, Gao M, Ci H, Gao J, Pan X, Yang Y (2023). Red teaming game: A game-theoretic framework for red teaming language models. ArXiv, abs/2310.00322

Májovský M, Cerny M, Kasal M E, Komarc M, Netuka D (2023). Artificial intelligence can generate fraudulent but authentic-looking scientific medical articles: Pandora's box has been opened. Journal of Medical Internet Research, 25: e46924

Meyer J G, Urbanowicz R J, Martin P C N, O'Connor K, Li R, Peng P C, Bright T J, Tatonetti N, Won K J, Gonzalez-Hernandez G, Moore J H (2023). ChatGPT and large language models in academia: opportunities and challenges. BioData Mining, 16(1): 20

Mo W, Xu J, Liu Q, Wang J, Yan J, Xiao C, Chen M (2023). Test-time backdoor mitigation for black-box large language models with defensive demonstrations. ArXiv, abs/2311.09763

Motoki F, Pinho Neto V, Rodrigues V (2024). More human than human: Measuring ChatGPT political bias. Public Choice, 198(1-2): 3–23

Mozes M, He X, Kleinberg B, Griffin L D (2023). Use of llms for illicit purposes: Threats, prevention measures, and vulnerabilities. ArXiv, abs/2308.12833

Niszczota P, Conway P (2023). Judgements of research co-created by Generative AI: Experimental Evidence. Economics and Business Review, 9(2): 101–114

O'Neill M, Connor M (2023). Amplifying limitations, harms and risks of large language models. ArXiv, abs/2307.04821

Qammar A, Wang H, Ding J, Naouri A, Daneshmand M, Ning H (2023). Chatbots to chatgpt in a cybersecurity space: Evolution, vulnerabilities, attacks, challenges, and future recommendations. ArXiv, abs/2306.09255

Raffel C, Shazeer N, Roberts A, Lee K, Narang S, Matena M, Zhou Y, Li W, Liu P J (2020). Exploring the limits of transfer learning with a unified text-to-text transformer. Journal of Machine Learning Research, 21(140): 1–67

Rahman M M, Watanobe Y (2023). ChatGPT for education and research: Opportunities, threats, and strategies. Applied Sciences,

13(9): 5783

Rasul T, Nair S, Kalendra D, Robin M, Santini F O, Ladeira W J, Sun M, Day I, Rather R A, Heathcote L (2023). The role of ChatGPT in higher education: Benefits, challenges, and future research directions. Journal of Applied Learning and Teaching, 6(1): 41–56

Ren J, Xu H, Liu Y, Cui Y, Wang S, Yin D, Tang J (2024). A robust semantics-based watermark for large language model against paraphrasing. Findings of the Association for Computational Linguistics: NAACL 2024 - Findings, 613–625

Robey A, Wong E, Hassani H, Pappas G J (2023). SmoothLLM: Defending large language models against jailbreaking attacks. ArXiv, abs/2310.03684

Rozado D (2023). The political Biases of ChatGPT. Social Sciences, 12(3)

Saetra H S (2023). Generative AI: Here to stay, but for good? Technology in Society, 75: 102372

Salem A, Paverd A, Köpf B (2023). Maatphor: Automated variant analysis for prompt injection attacks. ArXiv, abs/2312.11513

Sang J, Yu J (2023). ChatGPT: A Glimpse into AI's future. Journal of Computer Research and Development, 60(6): 1191–1201 (in Chinese)

Sison A J G, Daza M T, Gozalo-Brizuela R, Garrido-Merchán E C (2024). ChatGPT: More than a "weapon of mass deception" ethical challenges and responses from the human-centered artificial intelligence (HCAI) perspective. International Journal of Human-Computer Interaction, 40(17): 4853–4872

Song X Q, Liu M J, Chen J H (2023). Comprehensive impact analysis of GPT-4: High-quality economic development and national security prevention. Journal of Guangdong University of Finance & Economics, 38(02): 100–112 (in Chinese)

Staab R, Vero M, Balunović M, Vechev M (2024). Beyond memorization: Violating privacy via inference with large language models. In: Proceedings of the 12th International Conference on Learning Representations, ICLR 2024, December. Vienna: International Conference on Learning Representations, ICLR

Su Y (2024). The legal risks and governance paths for large language models. Journal of Northwest University of Political Science and Law, 42(1): 1–13 (in Chinese)

Suo X (2024). Signed-prompt: A new approach to prevent prompt injection attacks against LLM-Integrated applications. In: Proceedings of the 2024 2nd International Conference on Computer Science and Mechatronics (ICCSM 2024), January

Vykopal I, Pikuliak M, Srba I, Moro R, Macko D, Bielikova M (2024). Disinformation capabilities of large language models. In: Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics August. Bangkok, Thailand: Association for Computational Linguistics, 1: 14830–14847

Wen J, Ke P, Sun H, Zhang Z, Li C, Bai J, Huang M (2023). Unveiling the implicit toxicity in large language models. In: Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing December. Singapore: Association for Computational Linguistics (ACL) 1322–1338

Xiao F, Lai N (2024). The Subjectivity Problem associated with generative artificial intelligence. Journal of Shanxi Normal University, 51(01): 13–20 (Social Science Edition, in Chinese)

Yang XJ, Pan LM, Zhao XD, Chen HF, Petzold L, Wang W, Cheng W (2024). A survey on detection of LLMs-generated content. In: Proceedings of Findings of the Association for Computational Linguistics: EMNLP 2024, January. Miami: Association for Computational Linguistics, 9786–9805

Yao D, Zhang J, Harris I G, Carlsson M (2024). Fuzzllm: A novel and universal fuzzing framework for proactively discovering jailbreak vulnerabilities in large language models. In: 49th IEEE International Conference on Acoustics, Speech, and Signal Processing, ICASSP 2024 April. Seoul, Republic of Korea: Institute of Electrical and Electronics Engineers Inc., 4485–4489

Yu S, Fan D Z (2023). The value challenge of a new generation of Al ChatGPT and its inclusive governance. Journal of Hainan University (Humanities & Social Sciences), 41(5): 82–90 (in Chinese)

Yuan Z (2023). On the capacity for responsibility of generative artificial intelligence. Oriental Law, (3): 18–33 (in Chinese)

Zack T, Lehman E, Suzgun M, Rodriguez J A, Celi L A, Gichoya J, Jurafsky D, Szolovits P, Bates D W, Abdulnour R E, Butte A J, Alsentzer E (2024). Assessing the potential of GPT-4 to perpetuate racial and gender biases in health care: A model evaluation study. Lancet. Digital Health, 6(1): e12–e22

Zhang X H (2023). Political and social dynamics, risks, and prevention of ChatGPT. Journal of Shenzhen University (Humanities & Social Sciences), 40(3): 5–12 (in Chinese)

Zhang Y, Ding L, Zhang L, Tao D (2024). Intention analysis prompting makes large language models a good jailbreak defender. ArXiv, abs/2401.06561

Zhou X, Wang Q, Wang X, Tang H, Liu X (2023). Large language model soft ideologization via AI-self-consciousness. ArXiv, abs/2309.16167

Zhuo T Y, Huang Y, Chen C, Xing Z (2023). Red teaming chatgpt via jailbreaking: Bias, robustness, reliability and toxicity. ArXiv, abs/2301.12867