Reliability Issues of LLMs: ChatGPT a Case Study

Abdul Majeed and Seong Oun Hwang, *Department of Computer Engineering, Gachon University, Korea*

Abstract—This article underscores the perils of large language models (LLMs) by taking ChatGPT as a case study. We highlight various scenarios/sectors where ChatGPT's adoption can be challenging and in which naive use can cause concern. We suggest remedies to limit the perils from using ChatGPT and that can enhance the adoption and reliability of ChatGPT-like LLMs.

ChatGPT is a groundbreaking AI invention, and this technology will see tremendous growth per the IEEE Computer Society's 2024 technology predictions report[1]. According to the report, generative AI applications top the list, and this paradigm is predicted to experience most of the advancements in the coming years. ChatGPT, a generative AI product, has demonstrated its effectiveness in many ways (e.g., answering questions, summarizing text, generating computer code, fixing programming bugs, generating synthetic data, etc.). Despite the many promising applications, ChatGPT cannot produce desirable results for many difficult and pragmatic tasks [1]. For example, the inaccuracy from ChatGPT answers related to emotional text is significantly high owing to limited amounts of data—or no available data—concerning these tasks [1]. Similarly, ChatGPT can be manipulated to generate fake content, which can be hard to distinguish from real content. There are two schools of thought in the AI community about ChatGPT technology.

1) Many AI experts advocate the use of ChatGPT because it can save time and assist humans, just like digital twins. For instance, it can help researchers summarize a lot of literature in just a few minutes. It can test/write computer code while needing minimal effort and time from software engineers [2]. It can help programmers modify code to fit their purposes by giving ChatGPT the necessary instructions. In many fields, ChatGPT has various dedicated services (e.g., it can work as a telemedicine assistant to answer FAQs in the medical domain) which was not possible a couple of years ago.
2) Many AI experts are raising concerns about the excessive use of this black-box technology, which is open to misuse in many ways. It can lead to privacy breaches and personal data misuse because fine-grained information (names, email addresses, session information, etc.) is required to use it. Also, a lot of the programming code generated is wrong and needs human expertise to rectify it. Direct use of ChatGPT content in research is risky because it often ignores context in the material, and lacks scientific rigor. This technology is a big threat to the credibility of established knowledge providers in some fields (e.g., the medical literature). Lastly, it has acquired all the bad writing traits in humans (e.g., using many words to say too little)[2].

Another reason for the poor adoption of LLMs is the lack of trust and transparency in terms of the resources used in training them. For instance, it is unclear whether all reliable resources have been used in training them or only some contents have been retrieved from easily available sources to generate responses against the client's prompts. Unfortunately, there is a serious lack of methods that can correctly quantify trust in LLMs, and explore ways to enhance it [3]. Therefore, trust enhancement in LLMs is an urgent need in modern times to enhance the ethical alignment and reliability of ChatGPT-like LLMs. In some high-risk scenarios (e.g., biomedical scenario), the reliability assessment and enhancement in LLMs is imperative and attracting the wider attention of researchers nowadays [4].

We believe the pros and cons of ChatGPT go hand in hand, and many efforts are underway to govern its responsible use [5]. Recently, there have been growing calls in the AI community to establish rules for ChatGPT usage in order to curb misuse. In the recent past, some methods improved technical deficiencies such as data leakage, unauthorized access, deepfake mitigation, plagiarized content detection, text differen-

XXXX-XXX © 2024 IEEE
Digital Object Identifier 10.1109/XXX.0000.0000000

[1] https://www.computer.org/press-room/2024-technology-predictions-revealed

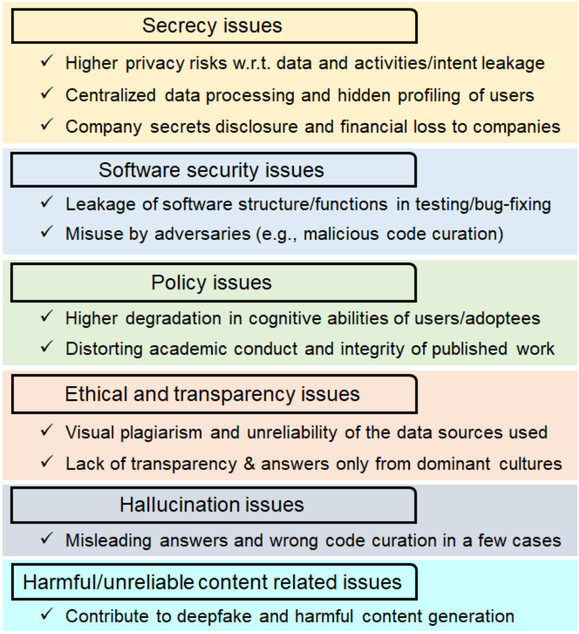
[2] https://www.nature.com/articles/d41586-023-02980-0

tiation, improper content filtering, and lack of security mechanisms. However, the potential perils and different areas/scenarios where ChatGPT adoption can be challenging remained unexplored.

This article analyzes the potential perils in ChatGPT in terms of the growing cyber-threat landscape, scientific literature integrity, public safety concerns, personal data misuse, destruction of cognitive abilities, trustworthiness, plagiarism in AI-generated content, etc. We pinpoint various sectors/scenarios (e.g., military defense secrets, testing for software vulnerabilities, and analytical problem solving) where ChatGPT adoption can be challenging owing to limited controls as well as data unavailability. We also suggest remedies to rectify the identified perils in order to foster the adoption of the technology in unattended sectors/scenarios. Our analysis can help ChatGPT stakeholders further enhance robustness, which in turn, can improve the governance and large-scale adoption of ChatGPT technology around the world.

Perils of ChatGPT

The perils of the ChatGPT technology from a broader perspective are shown in Figure 1, where we can see the various impacts on users/regulators. Specifically, we identify and categorize high-level perils of ChatGPT into six main and eleven subcategories in Figure 1. Below, we concisely present further details.



Secrecy issues
- ✓ Higher privacy risks w.r.t. data and activities/intent leakage
- ✓ Centralized data processing and hidden profiling of users
- ✓ Company secrets disclosure and financial loss to companies

Software security issues
- ✓ Leakage of software structure/functions in testing/bug-fixing
- ✓ Misuse by adversaries (e.g., malicious code curation)

Policy issues
- ✓ Higher degradation in cognitive abilities of users/adoptees
- ✓ Distorting academic conduct and integrity of published work

Ethical and transparency issues
- ✓ Visual plagiarism and unreliability of the data sources used
- ✓ Lack of transparency & answers only from dominant cultures

Hallucination issues
- ✓ Misleading answers and wrong code curation in a few cases

Harmful/unreliable content related issues
- ✓ Contribute to deepfake and harmful content generation

FIGURE 1. High-level categorization of the perils of ChatGPT.

Most points in this section give the impression that generative AI (e.g., ChatGPT) may be factually inaccurate. We affirm that ChatGPT has many useful services in diverse sectors. Also, ChatGPT is not designed to only perform information retrieval tasks, but it can be used to perform diverse tasks (e.g., code generation, text rephrasing, content summarization, poetry generation, etc.). However, ChatGPT is mainly designed to generate content, and not all content is fact due to multiple reasons (e.g., limited/no access to up-to-date data, regulatory measures, etc.). Our analysis is generic and relates to multiple problems in the ChatGPT landscape ranging from sources used in training data to the usage of ChatGPT-generated content in real-world settings.

1) ChatGPT use always requires binding the user's profile, and there is the risk of privacy disclosure of various kinds, including identity, location, personal preferences, and spatiotemporal activities. Each user can see the queries he/she has executed on the system, even over a long period, which means privacy can be breached by hackers or service providers when storing information in a centralized form for a long period. Many studies have highlighted this problem, but there are no privacy controls in most versions of ChatGPT, and the risk of personal information manipulation is high.
2) The current version of ChatGPT is like a client/server system where the server is the most powerful entity. This setting provides several benefits by offloading most of the computing load to the server, thereby relieving clients of any associated workloads. However, in this setting, most queries executed are saved on the server, which risks exposing sensitive data to third parties without the users' knowledge. We consider this service mechanism a centralized setting where users/clients have weaker control over the technology than the server. Most functions are delegated to the server, and personal data are already available, which can lead to hidden profiling of some users. In some cases, a response can be altered by considering the user profile to limit transparent and/or equitable use of the technology for specific groups or regions. Lastly, applications on the server side of NLP algorithms (or knowledge graphs) to queries posted by clients can reveal sensitive information concerning some individuals/organizations.
3) ChatGPT processes data in a centralized form that can give rise to different forms of manipulation. For example, if a group of people works on a sensitive mission or on a confidential defensive

mission, posing queries related to that mission can reveal secret information about a company. Most of the previous work underscored the importance of privacy protection only, but no studies have highlighted this particular problem. In many countries/companies, there are obligations to protect secrets in an organization, but posing queries to ChatGPT in a naive way can partially or fully leak the secret information/strategies of the organization.

4) Recently, ChatGPT has demonstrated its effectiveness in detecting vulnerabilities in software/programs [6]. However, while serving the cause, it is possible to memorize the structure/function of the software, leading to configuration disclosure. This problem is more severe than individual privacy leaks because an organization can be impacted by the entire software package or its technical details being leaked, particularly when software is related to military defense or any other sensitive system/mission.
5) Although ChatGPT is helpful in curating code to solve a variety of research and development problems, it can be manipulated by nefarious actors to generate malicious code (e.g., viruses, Trojan horses, spam, and malware) that can harm computing systems. In addition, it can help adversaries write malicious code in a variety of languages, which can be more dangerous than human-generated code. In some cases, it can take code written in one language and modify it into another language that can be used to attack systems. In this context, ChatGPT has not only assisted in enhancing malicious code but also made it more lethal and undetectable, posing a serious challenge to the security of established systems. Hence, there is a new cyber risk to sensitive applications such as banks and other financial institutions, defense systems, and general-purpose websites with weak security.
6) ChatGPT use in some countries has been banned because it can lead to bypassing someone's cognitive abilities[3]. For example, in many tasks (essay writing, summarization, etc.), candidates need to think before drafting something or solving a problem. However, ChatGPT can accept such a task with only a few commands, and users/clients may become over-reliant on it, leading to supplanting human cognitive abilities/functions rather than supplementing them. Furthermore, it can allow users to limit their own thinking, which might give rise to less succinct content, particularly in scientific research. It can forgo a wide-scale search for solutions to complex problems from diverse sources, which in turn can limit innovation or critical thinking in some users. Lastly, it can replace the problem-solving abilities of students, which can hamper knowledge and skills acquisition required in higher education (e.g., universities), leading to a less skilled workforce.
7) The advent of ChatGPT has put greater strain on publishers, and many have been forced to make new policies/procedures to limit excessive use. More policies are underway that ask authors to transparently disclose the usage of ChatGPT in their papers. However, it is still challenging to spot content that comes from ChatGPT-like technology if authors do not properly disclose it. Also, some reviewers are taking help from ChatGPT in reviewing papers, which might lead to improper reviews and illegitimate comments/analyses. In some cases, ChatGPT responses are unreliable, which might lead to the rejection of credible research papers. In addition, ChatGPT-generated content has the potential to undermine the integrity of facts/knowledge if limited resources are used in generating responses. ChatGPT poses a greater risk to scientific literature because it can help write entire papers in some cases [7], which can call into question the authorship as well as the appropriateness of research grants. ChatGPT is becoming a new form of ghost authorship[4], which can hamper transparency in academic publishing. Lastly, ChatGPT is converting science "for humans by humans → for machines by machines[2]" and can severely damage the integrity of science.
8) According to The New York Times lawsuit against OpenAI[5], it was evident that some generative AI systems might produce/create plagiaristic outputs, verbatim in the New York Times stories. This issue highlights the potential copyright infringement issues that could arise from using Generative AI tools. This issue is also referred to as visual/text plagiarism. Furthermore, many recent generative AI applications were found to

[3] https://www.universityworldnews.com/post.php?story=20230207160059558

[4] https://www.ncbi.nlm.nih.gov/pmc/articles/PMC10509679/

[5] https://spectrum.ieee.org/midjourney-copyright

have it[6], which means they can partially/fully expose the content used in training them. This problem also resembles the model inversion or membership inference in the adversarial ML literature. Additionally, it is generally the case for ML that unreliable or bad-quality training data leads to poor model performance. Similarly, we believe that if unreliable content is used in training ChatGPT like LLMs, the quality of responses will also be low. In addition, resource information in a response is always hidden from users, which means it is difficult to ensure the answers come from credible sources. Furthermore, there is a risk of prioritizing quality of service over diversity in the sources, which may cause a false or inaccurate response in some cases. There is also a lack of flexibility if only one response is presented to the user, which can lead to inaccurate analysis and/or wrong conclusions about problems requiring multiple answers.

9) Many studies have found that ChatGPT gives reliable answers only when general questions are posed [8]. Responses to specific questions or to those that belong to some specific nations/communities are mostly unreliable, owing to less data being used from those nations/communities. Hence, using ChatGPT responses in some critical cases (e.g., equity, inclusion, diversity) is risky if it uses data from dominant cultures only. ChatGPT cannot distinguish between correct and incorrect answers when a user contradicts it, provoking concerns about reliability. Lastly, information about the sources used in response generation is not revealed, which indicates a lack of transparency.
10) In many cases, the answers produced by ChatGPT are incorrect, particularly when questions relate to recent or future events. In addition, responses to Yes/No questions are mostly unclear. It is possible to mold responses to the question by dictating to ChatGPT or by posing a contradiction. Hence, hallucinations and misleading answers can give rise to fake knowledge or improper content generation. The use of such answers in some critical applications/decisions can contribute to social conflicts and chaos (e.g., conversational AI applications that generate automated responses without ethical considerations). In some cases, misleading/wrong answers can lead to financial losses.
11) Filtering offensive content has been a longstanding problem in social networks (SNs). The latest developments in ChatGPT have worsened the issue, and a lot of offensive content can easily be curated with ChatGPT [9]. Fake and/or offensive content can be shared over SNs and leveraged for cyberbullying and cyber-aggression, which may lead to social stigma. In addition, ChatGPT can be used to generate unethical terms that can lead to an increase in harmful content. Due to more offensive content generation with ChatGPT-like tools, there has been a rapid rise in AI-generated content (AIGC), which is posing a complex challenge to SN service providers from the perspective of detection and mitigation [10]. In addition, AIGC can be used to spread misinformation and/or rumors, leading to confusion in society.

All of the above-cited perils are of significant concern, particularly due to the rapid rise in ChatGPT use around the globe. Besides, organizations can face many other types of challenges while adopting ChatGPT-like LLMs[7]. Through this work, we intend to uncover the main perils of ChatGPT for stakeholders so they stay at the forefront of the developments and use of ChatGPT technology in order to foresee misuse and prevent it in society.

Scenarios/Sectors where ChatGPT Adoption Can be Challenging

In this section, we uncover some scenarios/sectors where ChatGPT adoption can be challenging, and naive use can provoke concerns. In writing research articles, the adoption of ChatGPT raises concerns about research integrity owing to its unfamiliarity with research ethics and limited access to the most current literature. In this case, relying solely on ChatGPT can make the findings of the paper less reliable, leading to a severe impact on credibility. Similarly, in medical diagnoses of rapidly evolving diseases, reliable information cannot be acquired from ChatGPT owing to less data being available and poor correlations across diverse datasets. In consultations where information varies from person to person or event to event, the adoption of ChatGPT can raise serious concerns over the lack of recent data and domain knowledge. Furthermore, ChatGPT might not give reliable responses to historical events or scenarios, particularly when information is withheld by some countries owing to

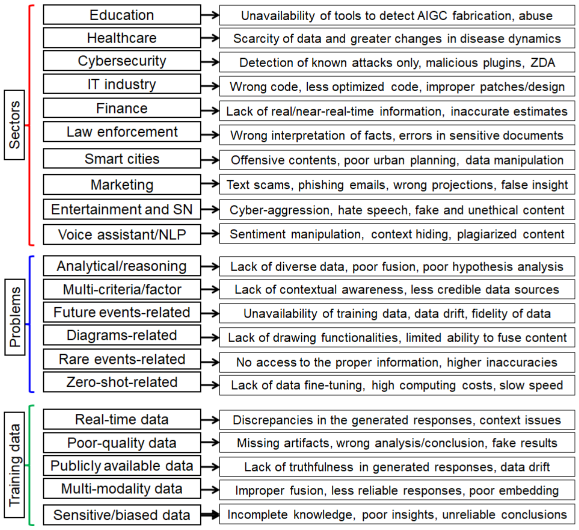
[6] https://spectrum.ieee.org/midjourney-copyright

[7] https://www.csoonline.com/article/1294996/top-4-llm-threats-to-the-enterprise.html

This article has been accepted for publication in IEEE Reliability Magazine. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/MRL.2024.3420849

national interests (e.g., confidential agreements and summit reports) or regulations. In teaching core computing courses at the university level, it is possible that students may engage in inappropriate use or misuse of technology, which can impact both the conducting of classes and the development of the curriculum. Conversation/chat applications may produce offensive content that frustrates customers/users. In analytical reasoning and multi-criteria decision-making situations, the use of ChatGPT is improper due to its higher dependence on the available training data and poor adaptability to new datasets. In the finance sector, it can give rise to misinformation or wrong financial estimates if the information available online or from public sources is not coherent. In the cybersecurity sector, ChatGPT cannot give reliable information about unknown attacks, again due to greater dependence on the training data.

Furthermore, ChatGPT cannot assist in detecting zero-day attacks (ZDAs) owing to no, or limited, information availability in the training data. The ZDAs include a wide range of cyber-attacks such as phishing, ransomware, denial of service (DoS), cloud-native breaches, etc. The ZDAs occur due to unaddressed/unknown security flaws in the computer software, hardware, or firmware. "Zero days" refers to the fact that security analysts/developers have zero days to fix the flaw/vulnerability in the system as malicious actors have already gained control of the system. In simple words, ZDA takes place when attackers exploit the flaw before developers/vendors have a chance to fix it[8]. ChatGPT and other AI models tend to have difficulty generalizing to data outside of the training data distribution. Even the ability to generalize to unseen samples in a test set requires that the test set resemble the training set distribution over the features that are used in training the model. Zero-day attacks by definition are new and even human subject matter experts (SMEs) have difficulty finding them. However, the generative nature of generative AI might be useful for generating zero-day attacks, but this would also require some automated validation system to be available to the ChatGPT. Since ChatGPT assists in generating codes of various attacks/malware, and therefore, it should provide reasonable assistance in detecting zero-day attacks. Considering these examples, it is fair to say that the adoption of ChatGPT in many sectors/scenarios is challenging, and its use might not fit the intended purpose or might provoke concerns regarding QoS. Figure 2 summarizes the sectors/scenarios where ChatGPT adoption is challenging and where we classified the adoption-related challenges w.r.t. sectors, problems, and data characteristics. The right side of Figure 2 cites the concerns and/or reasons for poor adoption of ChatGPT in the categories at left.



Sectors	Education	Unavailability of tools to detect AIGC fabrication, abuse
	Healthcare	Scarcity of data and greater changes in disease dynamics
	Cybersecurity	Detection of known attacks only, malicious plugins, ZDA
	IT industry	Wrong code, less optimized code, improper patches/design
	Finance	Lack of real/near-real-time information, inaccurate estimates
	Law enforcement	Wrong interpretation of facts, errors in sensitive documents
	Smart cities	Offensive contents, poor urban planning, data manipulation
	Marketing	Text scams, phishing emails, wrong projections, false insight
	Entertainment and SN	Cyber-aggression, hate speech, fake and unethical content
	Voice assistant/NLP	Sentiment manipulation, context hiding, plagiarized content
Problems	Analytical/reasoning	Lack of diverse data, poor fusion, poor hypothesis analysis
	Multi-criteria/factor	Lack of contextual awareness, less credible data sources
	Future events-related	Unavailability of training data, data drift, fidelity of data
	Diagrams-related	Lack of drawing functionalities, limited ability to fuse content
	Rare events-related	No access to the proper information, higher inaccuracies
	Zero-shot-related	Lack of data fine-tuning, high computing costs, slow speed
Training data	Real-time data	Discrepancies in the generated responses, context issues
	Poor-quality data	Missing artifacts, wrong analysis/conclusion, fake results
	Publicly available data	Lack of truthfulness in generated responses, data drift
	Multi-modality data	Improper fusion, less reliable responses, poor embedding
	Sensitive/biased data	Incomplete knowledge, poor insights, unreliable conclusions

FIGURE 2. Sectors/scenarios where ChatGPT adoption is challenging.

Figure 2 lists various sectors where the adoption of ChatGPT is challenging due to either poor responses or quality/accuracy in the training data. There are various areas, such as multi-criteria problems, zero-shot learning, and problems requiring visual explanations where ChatGPT adoption can be poor or can provoke concern. For instance, many problems require multi-criteria solutions, and it is unclear whether ChatGPT can consider all relevant criteria. Similarly, there are some situations/scenarios where ChatGPT adoption can be challenging owing to either limited data or frequent changes in data. In some cases, data can be biased, which can hinder equitable use of this technology and decrease its adoption. Furthermore, sectors/scenarios that require heterogeneous data from diverse sources can hinder ChatGPT adoption because some sources might not reveal their data due to privacy concerns. Lastly, ChatGPT cannot provide visuals, which can hinder its adoption in some sectors (e.g., institutes for the disabled).

Remedies to limit perils and enhance adoption

The developments in generative AI are magnified with each passing day, and the new wave of generative

[8] https://www.kaspersky.com/resource-center/definitions/zero-day-exploit

This article has been accepted for publication in IEEE Reliability Magazine. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/MRL.2024.3420849

tools will bring opportunities as well as threats. According to GPAI[9], for the betterment of society, it is urgent to curb the misuse of generative AI. Below we suggest potential remedies to mitigate or lessen the perils of ChatGPT technology to enhance its use.

1) To prevent software structure memorization, ChatGPT should provide a selection menu to enable users to decide whether to preserve software structure in the ChatGPT environment or not. In this way, users will be more able to customize settings considering the nature of the software/product. By software structure memorization, we mean that ChatGPT could memorize the structure (functions, variables, sensitive data, data/control flow, etc.) of the software when developers take help from it to either test the functions or remove bugs/vulnerabilities. During the service, there is a possibility that some parts of the software/program can be memorized by the ChatGPT, which constitutes software data theft [11]. One of the security issues of ML/AI models is the unintentional memorization of private training data, and ChatGPT is no exception. To this end, there exist some methods like gradient clipping, differential privacy[10], synthetic data, etc. to prevent LLMs from learning private data [12]. However, ChatGPT and other LLMs are intended to solve the knowledge-intensive task, which requires memorization of large data of diverse types [13]. To the best of the authors' knowledge, these kinds of technical challenges (i.e., to prevent ChatGPT from memorizing private data if the user requests) have not been solved in the open-source versions.
2) To prevent privacy issues, anonymized, rather than real, information should be collected upon account creation, and personally identifiable information must not be collected. It is worth noting that there exist some safeguards in the form of anonymization/differential privacy to assist in information retrieval without leaking private information or query confidentiality [14]. However, these solutions currently lack capabilities in terms of both adversarial threats and regulatory requirements given the wide threat vector of ChatGPT like LLMs [15]. Hence, more efforts are required to secure sensitive data used in prompts/queries and private data used for account creation [16]. Furthermore, sensitive questions/queries can be scrambled to prevent sensitive information disclosure. Lastly, users should be given some control over whether to save queries or not after using the service, or they should be empowered to decide how long searches/queries should be kept on the ChatGPT server. It is worth noting that some updated versions of ChatGPT allow users to delete their queries/history after use. However, it is still not very clear whether the queries/histories are actually deleted from the central server as some variants of LLMs re-use the queries/histories of users to improve the quality of service [17].
3) ChatGPT should discard (or minimally server) queries that require malicious code or vulnerability assessment of such code. There should be limited support for malicious code generation in languages such as Python, C++, Java, etc. By limited support, we mean here that ChatGPT should allow only malicious code generation for which defense tools are available. It should not allow the stealthier code generation which either continuously replicates (e.g., Polymorphic malware) or damages a large number of systems [18]. It is important to note that the current version of chatGPT has certain in-built guardrails to prevent the aforementioned mishaps (e.g., malicious code generation). Although certain in-built guardrails of ChatGPT prevent it from disseminating/creating harmful information, but they are not foolproof. An adversary can bypass them through clever prompting strategies such as reverse psychology, jailbreaking, and other techniques [18].
4) Hidden profiling of users can be prevented by separating profile information from queries made or questions asked of ChatGPT. Also, some federated versions of this technology can be implemented to process sensitive data in a distributed manner (on users' local devices).
5) Responsible use of this technology should be encouraged, and awareness about its harms communicated to users/adoptees in order to prevent bypassing their own cognitive ability.
6) Tools should be developed that can distinguish between ChatGPT-generated and human-generated content. Publishers should pre-check all academic papers for possible use of ChatGPT in core parts of the papers. There should be initiatives to discourage the use of this technology in research papers in order to preserve the integrity of the science and to prevent abuse of

[9] https://gpai.ai/

[10] https://dlsp2024.ieee-security.org/papers/dls2024-final7.pdf

This article has been accepted for publication in IEEE Reliability Magazine. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/MRL.2024.3420849

research grants. Also, authors should be asked about ChatGPT use upon submission of the paper. Lastly, one measure of “good” generative content is how undetectable it is as such. Therefore, detectors for AI-generated content should become part of the training loop to improve the technology and make it undetectable.

7) There should be information with the responses about the resources used in generating them. This will boost the credibility of the answers and will enable users to decide whether to trust the content or not. Furthermore, there should be *k* responses, rather than just one, in order to allow users to choose the most appropriate answer based on their needs. However, providing *k* responses might be costly, and ChatGPT might not provide *k* responses for each user query. In contrast, Google Gemini currently provides 3 responses for each user query. Hence, providing options to generate consistent and semantically correct queries, and multiple responses rather than one could enhance ChatGPT’s reliability in multiple sectors [19]. Transparency as to data sources can help prevent misuse of copyright-protected content. It can also help to hold stakeholders accountable for wrongdoing or data/resource manipulation. It is worth noting that this remedy is tailored to resources that are used in generating responses or information retrieval tasks with ChatGPT. Although the tasks performed by ChatGPT may not use only a specific set of resources but potentially all of the training data. Unfortunately, the resource information is mostly not presented to the user along with the response, which may undermine the performance of ChatGPT in some specific scenarios, particularly when less credible sources are consulted during the training time.

8) There should be ratings of the resources used in generating responses. Without such information, it is hard to verify the reliability of sources. There should be information regarding consent from each data source in order to respect copyright in ChatGPT use. In addition, there should be an information menu on the user side that allows users to include/exclude sources based on their needs. The addition of such tools and controls can limit visual plagiarism and can reveal the credibility of the sources used. This remedy deals with providing extra information about sources used in training ChatGPT, which can contribute to opening its black-box nature, leading to trust enhancement in AI-generated content.

9) To prevent bad code generation, there should be information regarding the code supplied (for example, how much computing time the code requires, which libraries are required, and the expected output in image form). In its current form, users need to test code in separate software and then find any mistakes. Moreover, it is beneficial to generate *k* variants of the code against developers’ request which will allow them to choose the most appropriate one among *k* candidates [20]. This would allow developers to choose the correct code and would reduce the software development time as the developers do not need to make back-and-forth interactions with ChatGPT when the code is wrong. In addition, it is worth providing source/repository information where the code came from so users can verify the source code and licensing information.

10) ChatGPT should limit deepfake content, and any requests for such content should be verified via extra controls, just like Google/YouTube verifies age information against unexpected requests. In addition, scrambled content or more generalized content can be shared against requests in order to limit fake content generation. In addition, profile information can be used to verify the nature of the request or previous requests in order to block regular perpetrators/bullies from manipulating ChatGPT for wrongdoing. It is worth noting that deep fake content cannot only be used for malicious purposes, but it can be useful and context-dependent in some cases. However, it is possible to be deceitful about the context/intent as credible-sounding profiles and contents are on the rise amid recent developments in LLMs [21]. Therefore, it is vital to integrate a context analyzer with ChatGPT like LLMs to limit deepfake content for only malicious purposes.

11) To prevent leaking company secrets, there is a need to empower users to anonymize their queries to limit exposure of the organization’s information. This can be prevented by establishing a taxonomy of sensitive words and limiting their exposure to ChatGPT-like tools via network traffic filtering within an organization.

12) Lastly, diverse data should be used in generating answers to eliminate bias, rather than using only data from the dominant culture. Diverse data utilization is vital to enhancing the credibility of answers and extending the generalization power of LLMs. In addition, there should be notifications in case data are not available or are not diverse enough to generate responses.

This article has been accepted for publication in IEEE Reliability Magazine. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/MRL.2024.3420849

Apart from the above-cited remedies, there should be some mechanism to get users' feedback on each response to eliminate poor responses and illegitimate resources from the system. Lastly, rigorous adoption of guidelines given by global bodies like GPAInote[9] and AI harms published in the latest research [22] are vital to control the unintended consequences of ChatGPT. Lastly, for remedies or mitigations, there are some proposed guidelines by many research organizations/groups. For instance, UK NCSC provided "Guidelines of secure AI system development", US NIST provided "AI Risk Management Framework", and the OWASP AI exchange project presented risk management for prediction AI and generative AI.

To foster the adoption of this black-box technology, it is vital to enhance its trustworthiness [23] and to limit fabricated/misleading answers[11]. (In one survey, about $\frac{2}{3}$ participants cited fabricated or misleading answers as a big concern in ChatGPT.) Trustworthiness in the context of LLMs has eight dimensions: transparency, truthfulness, fairness, safety, privacy preservation, robustness, machine ethics, and accountability. The incorporation of all these dimensions can enhance the adoption and reliability of ChatGPT across sectors. Among other dimensions, transparency is the key that can augment ChatGPT adoption and deployment in the future. In the healthcare sector, augmenting the reliability of the LLMs is a major and urgent concern as LLMs may misinterpret symptoms, which can lead to incorrect diagnosis [24].

Recently, various studies have uncovered a broad range of threat vectors associated with ChatGPT. These threats were classified into two main types: AI-inherent and non-AI-inherent [25]. Figure 3 presents the overview of both these threat vectors along with their subtypes. The former is known as an insider threat that might stem from ChatGPT's misconfiguration or other architectural flaws. The latter are known as external threats and are not linked intricately to ChatGPT. The details of these threats along with the defenses are in study [25]. Besides other attacks, instruction-tuning attacks (ITA) are relatively new and are impacting the reliability of LLMs. Below, we concisely discuss the developments and gaps in existing security controls to uncover ITA-related threat vectors.

- Jailbreaking attack (JBA): This attack involves bypassing safety controls to generate responses that otherwise be restricted or answering unsafe questions, indicating the limitation of safety controls. In past research, it has been suggested that JBA can be reduced by retraining LLMs or modifying their parameters, which can be costly in practice due to the large # of parameters of ChatGPT. Recently, some studies have shown that JBA can be reduced without retraining LLMs and accessing their parameters [26]. However, it is still very challenging to provide defense against all types of JBA due to the lack of explicit controls for questions/responses monitoring and large-scale parameters involved in LLMs.
- Prompt injection attack (PIA): This attack is used to manipulate the behavior of LLMs to generate harmful/unexpected responses. The PIA has many variants such as virtualized prompts, unveiling guide prompts, and application-specific PIA [25]. The potential remedy for PIA is to train LLMs on large datasets, and embed toxicity detection, classification, and detoxification [27]. However, it is still very challenging to reduce PIA because it requires a deep investigation of each prompt and its historical analysis.

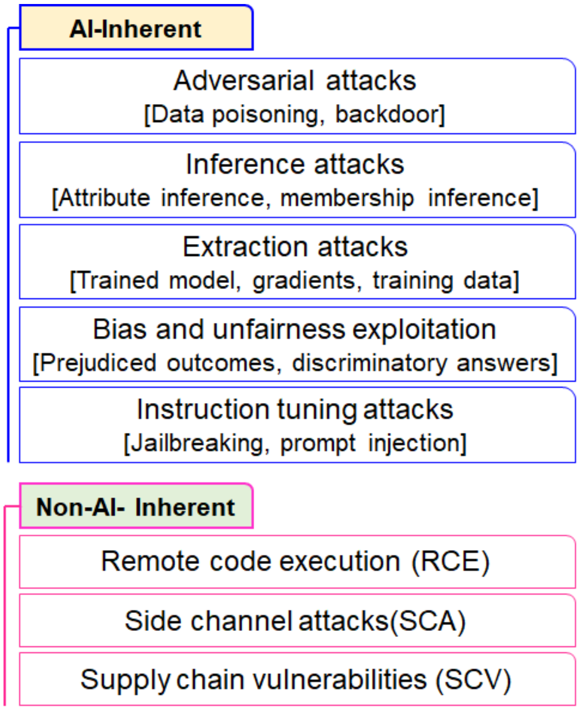

AI-Inherent
Adversarial attacks
[Data poisoning, backdoor]
Inference attacks
[Attribute inference, membership inference]
Extraction attacks
[Trained model, gradients, training data]
Bias and unfairness exploitation
[Prejudiced outcomes, discriminatory answers]
Instruction tuning attacks
[Jailbreaking, prompt injection]
Non-AI- Inherent
Remote code execution (RCE)
Side channel attacks(SCA)
Supply chain vulnerabilities (SCV)

FIGURE 3. Overview of threat vector of ChatGPT-like LLMs (adapted from [25]).

The existing controls are inadequate to fully address JBA and PIA, and the adoption of traditional approaches to limit ITA-related threat vectors is very challenging and impractical due to the multi-tasking

[11] https://www.nature.com/articles/d41586-024-00173-x

nature of ChatGPT. To this end, one practical way is safe instruction tuning [25], whose efficacy is yet to be tested in real scenarios.

Conclusion and Future Work

This paper provided an in-depth analysis of the perils and related challenges in the adoption of ChatGPT. We also provided valuable suggestions to lessen the perils from using ChatGPT and to enhance its adoption and reliability in many sectors/scenarios. Specifically, our work covers three key aspects related to ChatGPT: (i) pinpoint and discuss potential perils of ChatGPT, (ii) identify and discuss scenarios/sectors/problems where ChatGPT adoption might be challenging or naïve use might provoke concerns, and (iii) suggest valuable remedies to limit perils and to enhance adoption of ChatGPT in real-world settings. Our analysis can pave the way to improving ChatGPT from various perspectives, so it can better assist people around the globe. Due to the sociotechnical nature of LLMs, it is vital to initiate interdisciplinary collaboration to enhance their reliability and trust. For instance, it is vital to open up the details of their workflow, failure scenarios, and capabilities in diverse problems [28]. It is worth noting that our analysis is based on the open-source versions of ChatGPT and not the latest closed-source versions. Therefore, some of the above issues might have been rectified in the latest (paid) versions. In the future, we intend to expand our analysis to advanced versions of ChatGPT and other similar tools, such as Google's Bard, Microsoft Copilot, DALL-E, Claude 2, and Midjourney.

ACKNOWLEDGMENT

This work was supported by the National Research Foundation of Korea (NRF) grant funded by the Korea government (Ministry of Science and ICT) (RS-2024-00340882).

REFERENCES

1. J. Kocoń, I. Cichecki, O. Kaszyca, M. Kochanek, D. Szydło, J. Baran, J. Bielaniewicz, M. Gruza, A. Janz, K. Kanclerz *et al.*, "Chatgpt: Jack of all trades, master of none," *Information Fusion*, p. 101861, 2023.
2. C. Ebert and P. Louridas, "Generative ai for software practitioners," *IEEE Software*, vol. 40, no. 4, pp. 30–38, 2023.
3. L. Sun, Y. Huang, H. Wang, S. Wu, Q. Zhang, C. Gao, Y. Huang, W. Lyu, Y. Zhang, X. Li *et al.*, "Trustllm: Trustworthiness in large language models," *arXiv preprint arXiv:2401.05561*, 2024.
4. W. J. Bolton, R. Poyiadzi, E. R. Morrell, G. v. B. G. Bueno, and L. Goetz, "Rambla: A framework for evaluating the reliability of llms as assistants in the biomedical domain," *arXiv preprint arXiv:2403.14578*, 2024.
5. A. Koubaa, W. Boulila, L. Ghouti, A. Alzahem, and S. Latif, "Exploring chatgpt capabilities and limitations: A survey," *IEEE Access*, 2023.
6. M. Fu, C. Tantithamthavorn, V. Nguyen, and T. Le, "Chatgpt for vulnerability detection, classification, and repair: How far are we?" *arXiv preprint arXiv:2310.09810*, 2023.
7. G. Conroy, "Scientists used chatgpt to generate a whole paper from data," *Nature*, vol. 619, p. 443, 2023.
8. I. Amaro, A. Della Greca, R. Francese, G. Tortora, and C. Tucci, "Ai unreliable answers: A case study on chatgpt," in *International Conference on Human-Computer Interaction*. Springer, 2023, pp. 23–40.
9. I. Amaro, P. Barra, A. Della Greca, R. Francese, and C. Tucci, "Believe in artificial intelligence? a user study on the chatgpt's fake information impact," *IEEE Transactions on Computational Social Systems*, 2023.
10. D. Xu, S. Fan, and M. Kankanhalli, "Combating misinformation in the era of generative ai models," in *Proceedings of the 31st ACM International Conference on Multimedia*, 2023, pp. 9291–9298.
11. S. Zaman, "Chatgpt security risks and solutions," in *7th IET Smart Cities Symposium (SCS 2023)*, vol. 2023, 2023, pp. 378–387.
12. L. Wang, O. Thakkar, and R. Mathews, "Unintended memorization in large asr models, and how to mitigate it," in *ICASSP 2024 - 2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2024, pp. 4655–4659.
13. J. Yang, H. Jin, R. Tang, X. Han, Q. Feng, H. Jiang, S. Zhong, B. Yin, and X. Hu, "Harnessing the power of llms in practice: A survey on chatgpt and beyond," *ACM Transactions on Knowledge Discovery from Data*, vol. 18, no. 6, pp. 1–32, 2024.
14. B. B. Gupta, A. Gaurav, V. Arya, W. Alhalabi, D. Alsalman, and P. Vijayakumar, "Enhancing user prompt confidentiality in large language models through advanced differential encryption," *Computers and Electrical Engineering*, vol. 116, p. 109215, 2024.
15. R. Staab, M. Vero, M. Balunovic, and M. Vechev, "Large language models are anonymizers," in *ICLR 2024 Workshop on Reliable and Responsible Foundation Models*.
16. S. A. Khowaja, P. Khuwaja, K. Dev, W. Wang, and L. Nkenyereye, "Chatgpt needs spade (sustainabil-

ity, privacy, digital divide, and ethics) evaluation: A review," *Cognitive Computation*, pp. 1–23, 2024.

17. D. Masson, S. Malacria, G. Casiez, and D. Vogel, "Directgpt: A direct manipulation interface to interact with large language models," in *Proceedings of the CHI Conference on Human Factors in Computing Systems*, 2024, pp. 1–16.
18. M. Gupta, C. Akiri, K. Aryal, E. Parker, and L. Praharaj, "From chatgpt to threatgpt: Impact of generative ai in cybersecurity and privacy," *IEEE Access*, vol. 11, pp. 80 218–80 245, 2023.
19. K. Saab, T. Tu, W.-H. Weng, R. Tanno, D. Stutz, E. Wulczyn, F. Zhang, T. Strother, C. Park, E. Vedadi *et al.*, "Capabilities of gemini models in medicine," *arXiv preprint arXiv:2404.18416*, 2024.
20. Y. Liu, T. Le-Cong, R. Widyasari, C. Tantithamthavorn, L. Li, X.-B. D. Le, and D. Lo, "Refining chatgpt-generated code: Characterizing and mitigating code quality issues," *ACM Transactions on Software Engineering and Methodology*.
21. I. Augenstein, T. Baldwin, M. Cha, T. Chakraborty, G. L. Ciampaglia, D. Corney, R. DiResta, E. Ferrara, S. Hale, A. Halevy *et al.*, "Factuality challenges in the era of large language models," *arXiv preprint arXiv:2310.05189*, 2023.
22. F. S. Grodzinsky, M. J. Wolf, and K. W. Miller, "Ethical issues from emerging ai applications: Harms are happening," *Computer*, vol. 57, no. 2, pp. 44–52, 2024.
23. E. Cambria, R. Mao, M. Chen, Z. Wang, and S.-B. Ho, "Seven pillars for the future of artificial intelligence," *IEEE Intelligent Systems*, vol. 38, no. 6, pp. 62–69, 2023.
24. K. Nassiri and M. A. Akhloufi, "Recent advances in large language models for healthcare," *BioMedInformatics*, vol. 4, no. 2, pp. 1097–1143, 2024.
25. Y. Yao, J. Duan, K. Xu, Y. Cai, Z. Sun, and Y. Zhang, "A survey on large language model (llm) security and privacy: The good, the bad, and the ugly," *High-Confidence Computing*, p. 100211, 2024.
26. B. Cao, Y. Cao, L. Lin, and J. Chen, "Defending against alignment-breaking attacks via robustly aligned llm," *arXiv preprint arXiv:2309.14348*, 2023.
27. X. He, S. Zannettou, Y. Shen, and Y. Zhang, "You only prompt once: On the capabilities of prompt learning on large language models to tackle toxic content," *45th IEEE Symposium on Security and Privacy, May 20-23*, 2024.
28. R. Bommasani, D. A. Hudson, E. Adeli, R. Altman, S. Arora, S. von Arx, M. S. Bernstein, J. Bohg, A. Bosselut, E. Brunskill *et al.*, "On the opportunities and risks of foundation models," *arXiv preprint arXiv:2108.07258*, 2021.

Abdul Majeed is an Assistant Professor in the Department of Computer Engineering, Gachon University, Korea. He received his Ph.D. in Computer Information Systems & Networks from the Korea Aerospace University in 2021. His research interests include data-centric artificial intelligence, machine learning, information privacy, and secure personal data publishing. Contact him at ab09@gachon.ac.kr.

Seong Oun Hwang is a Professor in the Department of Computer Engineering, Gachon University, Korea. He received his Ph.D. in computer science from the Korea Advanced Institute of Science and Technology (KAIST) in 2004. He is a senior member of IEEE. He is the corresponding author of this article. His research interests include cryptography, data-centric artificial intelligence, cybersecurity, and machine learning. Contact him at sohwang@gachon.ac.kr.