Journal of Information and Intelligence 2 (2024) 102–115

Contents lists available at ScienceDirect

Journal of Information and Intelligence

journal homepage: www.journals.elsevier.com/journal-of-information-and-intelligence

Unveiling security, privacy, and ethical concerns of ChatGPT

Xiaodong Wu, Ran Duan, Jianbing Ni [*]

Department of Electrical & Computer Engineering, Queen's University, Kingston K7L 3N6, Canada

ARTICLE INFO

Keywords:
ChatGPT
Large language model (LLM)
Security
Privacy
Ethics

ABSTRACT

This paper delves into the realm of ChatGPT, an AI-powered chatbot that utilizes topic modeling and reinforcement learning to generate natural responses. Although ChatGPT holds immense promise across various industries, such as customer service, education, mental health treatment, personal productivity, and content creation, it is essential to address its security, privacy, and ethical implications. By exploring the upgrade path from GPT-1 to GPT-4, discussing the model's features, limitations, and potential applications, this study aims to shed light on the potential risks of integrating ChatGPT into our daily lives. Focusing on security, privacy, and ethics issues, we highlight the challenges these concerns pose for widespread adoption. Finally, we analyze the open problems in these areas, calling for concerted efforts to ensure the development of secure and ethically sound large language models.

1. Introduction

In December 2022, OpenAI released an interactive chat platform, called ChatGPT. Its powerful in-context learning and naturally generation ability created a big shock to the whole world. Before this, there are some powerful large language models (LLM) like GPT [1] or BERT [2], which can perform well on many nature language processing (NLP) tasks, but carefully processed inputs are required in the query procedure. In other words, these machine learning tools can only be used to complete specific tasks with constrained inputs. Nevertheless, ChatGPT has greatly improved this by showing a wonderful interactive ability, which can respond almost any legal questions in different styles or targeting on different tasks. For example, users can request ChatGPT to write a series of codes with comments for each part to address certain problems. ChatGPT can also be used to summarize given texts or provide detailed illustrations for some complex concepts. ChatGPT can provide long but natural responses, which are aligned with human's knowledge. It integrates a variety of ability for NLP and possesses the ability to clarify its knowledge boundary and refuse the illegal queries. Currently, ChatGPT has over 100 million users and there are over 1.6 billions visits in June 2023. ChatGPT has become the most famous AI application and the center of the world's attention.

However, similar to other AI applications, ChatGPT brings ethical concerns and misuse risks. For example, due to its powerful text reasoning and generation ability, students find it very helpful in writing homework. In the beginning, ChatGPT was used to explain some difficult content or rephrase the written project reports, but soon, it was utilized to write the entire homework. Such misuse immediately attracts attentions from the teachers and schools and it was soon identified as plagiarism. Another concern is the copyright from ChatGPT. With increasing number of people who use ChatGPT to create original-like text content without citation, the copyright of the content created by ChatGPT becomes a serious concern. No one is responsible for the correctness and accuracy of the content. It becomes necessary to regulate the copyright of machine's generation, including both visual and textual content. Besides, although there are

---

[*] Corresponding author.
E-mail addresses: xiaodong.wu@queensu.ca (X. Wu), 18rd29@queensu.ca (R. Duan), jianbing.ni@queensu.ca (J. Ni).

https://doi.org/10.1016/j.jiixd.2023.10.007
Received 2 September 2023; Received in revised form 7 October 2023; Accepted 25 October 2023
Available online 31 October 2023
2949-7159/© 2023 Published by Elsevier B.V. on behalf of Xidian University. This is an open access article under the CC BY-NC-ND license (http://creativecommons.org/licenses/by-nc-nd/4.0/).

privacy protection mechanisms in ChatGPT, such as the block of access to personal data about individuals, it is not guaranteed that no leakage of its training data would occur. Malicious attacks, such as jailbreaking attacks, may utilize its great generation ability to infer some information from personal data or even use them to attack other AI models. Therefore, it is well recognized that despite great advantages ChatGPT brings to our world, the potential security, privacy, and ethical problems cannot be overlooked. There are plenty research works proposed to address these problems, but only few works try to sum up them. In order to improve these solutions and open new directions for the future work, it is necessary to summarize these works and provide detailed comparison and analysis of them.

In this paper, we introduce the security, privacy and ethical issues behind the most recent and famous AI technique: ChatGPT, as shown in Fig. 1, expecting to raise more attention into these issues and discovering future directions for solving them. The main contributions of our paper can be summarized as follows.

- We give a detailed introduction of the upgrade path from GPT-1 to GPT-4 and a detailed comparison of these methods on model size, data size and performances. Their features and limitations with advanced applications are discussed to highlight the promising applications of LLMs, especially for ChatGPT.
- We examine how ChatGPT poses new threats to data security and how ChatGPT is utilized for compromising security. The impact on security mainly includes assisting in the generation of attack codes and phishing websites. Both of them increase attack capabilities of adversaries. Also, we discuss the unintentional impact of inaccurate information generated by ChatGPT, possible safety hazards caused by human misuse, and potential threats to social security brought about by the deep dependence on ChatGPT in the future.
- We examine the privacy policy of OpenAI and the current privacy laws on personal data protection to emphasize the privacy violation of ChatGPT. Also, We discuss the privacy leakage threats brought by numerous data collection, personal input reference, privacy reference attacks, and the concerns on transparency.
- We analyze the general ethical effects of AI technology on individuals, society and environment and discuss the fairness and bias issues behind AI. For ChatGPT, we summarize the ethical and legal challenges it faces.
- We discuss how to detect whether the communication object is ChatGPT in a conversation, how to detect texts generated by ChatGPT, and introduce some difficulties for such detection.

2. ChatGPT

In this section, we briefly review the technical path from GPT-1 to GPT-4 and their features and limitations.

2.1. From GPT-1 to GPT-4

In 2018, OpenAI introduced the initial version of the generative pre-trained transformer (GPT) [1], a highly capable large language model for natural language processing. GPT has exhibited exceptional performance across a wide range of complex language tasks, positioning it as a formidable competitor to other similar models such as BERT [2], which was proposed by Google in the same year. Prior to the success of these methods, numerous effective algorithms and remarkable applications had been developed in NLP, including machine translation [3,4], voice recognition [5,6], and summary generation [7,8]. However, these applications heavily

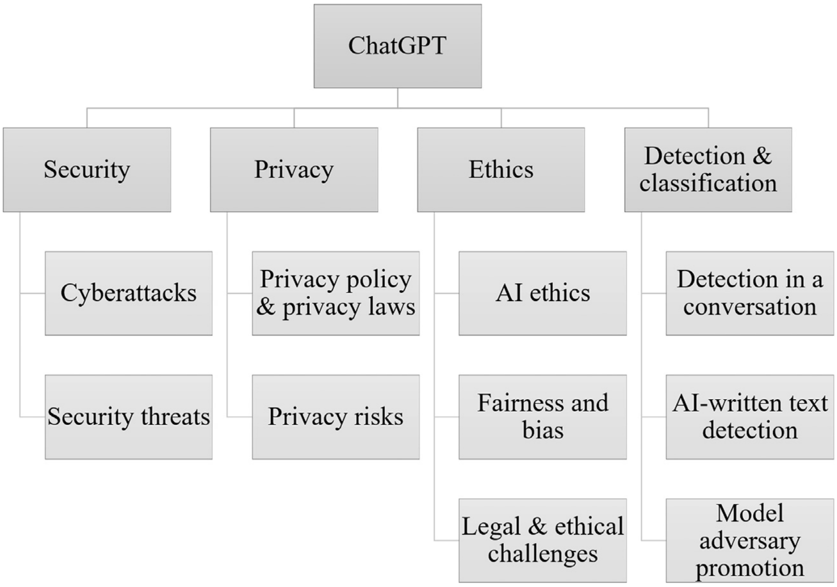

ChatGPT
Security
Privacy
Ethics
Detection & classification
Cyberattacks
Security threats
Privacy policy & privacy laws
Privacy risks
AI ethics
Fairness and bias
Legal & ethical challenges
Detection in a conversation
AI-written text detection
Model adversary promotion

Fig. 1. The logic flow of this paper.

relied on extensively annotated data, resulting in time-consuming and expensive model training as the models grew in size. Moreover, despite the proficiency of a well-performing NLP model, its generalization to other tasks remained challenging. Essentially, these models were domain experts limited to specific areas of expertise, lacking the versatility exhibited by human beings in performing diverse tasks. Consequently, there arose a need for a methodology that could be trained without labeled data while possessing superior generalization capabilities across multiple tasks. This need served as the driving force behind OpenAI's development of GPT.

One notable advantage of GPT is its ability to train the model without relying on large annotated datasets. It employs a two-step process: Unsupervised pretraining, which is operated on rich text materials and supervised fine-tuning, which is closely connected to the end applications. During the initial unsupervised pretraining phase, GPT employs 12 transformer blocks as decoders. Unlike the original transformer decoders, each block consists solely of masked multi-head attention. The objective of the pretraining process is to predict the subsequent word in a sentence based on the preceding words. Consequently, this unsupervised learning approach only necessitates raw, yet comprehensive, text materials. Following the pretraining, the model proceeds to the supervised fine-tuning phase, which focuses on a specific problem such as sentiment classification. Importantly, this stage requires a significantly smaller annotated dataset compared to the one utilized in the pretraining phase. When applying GPT to different task types, users only need to modify the input format to facilitate fine-tuning. By employing these two distinct processes, GPT no longer necessitates amassing a vast labeled dataset, while still possessing the capacity for generalization across a variety of tasks, all at a reasonable fine-tuning cost. This novel approach presents an outstanding solution to the aforementioned challenges.

While GPT-1 significantly reduces the reliance on labeled data for training, it still necessitates annotations during the fine-tuning phase. In order to further diminish the dependence on labeled data and enhance the model's generalization capabilities, GPT-2 was introduced in 2019 to address these concerns [9]. The core concept behind this groundbreaking algorithm involves a complete shift from supervised learning to unsupervised learning. The knowledge obtained through unsupervised training encompasses fragments of the information required for all supervised learning tasks. By imparting models with common knowledge in a given domain through unsupervised learning, the supervised learning task becomes a mere application of this preexisting knowledge. Essentially, GPT-2 is designed to leverage its vast pretrained knowledge, acquired from a vast dataset with rich materials, to tackle complex tasks. To achieve this objective, the architecture of GPT-2 remains largely unchanged from GPT-1, with the key modifications being an increase in the number of layers and the dataset size. These enhancements aim to imbue the model with the necessary knowledge to proficiently handle a wide range of problems.

The advancements observed from GPT-1 to GPT-2 underscored the potential for enhancing the generalization capabilities of large language models (LLMs) by increasing the model's parameter size and the training dataset. Building upon this concept, GPT-3 was introduced in 2020 as a significantly more powerful LLM [10]. With the largest number of parameters and an extensive training dataset, GPT-3 achieved state-of-the-art performance across numerous NLP tasks. In addition to its increased scale, GPT-3 introduced a novel training paradigm known as in-context learning. Departing from the conventional approach of predicting outputs solely based on queries, this new model is trained to predict outputs by considering both the queries and their corresponding examples. By incorporating contextual examples during the learning process, along with the utilization of extensive training datasets, GPT-3 is capable of acquiring comprehensive knowledge from texts. This approach empowers the model with remarkable generation capabilities, enabling it to deliver exceptional performance in a diverse range of NLP tasks, often comparable to human-level performance.

A brief summary of GPT-1, GPT-2, and GPT-3 is shown in Table 1. It is noticeable that the increasing performance of GPT series models appears with the exponential explosion of parameters. Only in the third year since the creation of GPT-1, the final model has become one of the biggest models in the world, costing 285 k CPUs and 10 k GPUs to train with 12 million dollars consumption.

Despite the impressive performance of GPT-3, it falls short of being considered truly “intelligent”. Surprisingly, smaller methods such as T5 [11] have even outperformed GPT-3 in certain tasks, which is unexpected considering the vastness of its training datasets and the complexity of its parameters. One hypothesis for this discrepancy is that while GPT-3 may have acquired rich foundational knowledge during training, it still struggles with accurately understanding and providing valid responses to user's queries. To address this issue, it becomes crucial to enhance GPT-3's reasoning capabilities and its responsiveness to instructions. Recognizing the need for improvement, OpenAI embarked on refining GPT-3's performance through meticulously designed training methodologies. They introduced code-based training [12] and instruction tuning [13] to activate the model's reasoning abilities and its responsiveness to human instructions. The updated version of GPT-3 demonstrates the capacity for more reasonable responses, incorporating complex reasoning, and exhibiting greater generalization power even across unseen tasks. It has speculated that the remarkable capabilities of GPT-3 may have remained latent, requiring specific training techniques to unlock its full potential [14]. OpenAI proposed the usage of reinforcement learning from human feedback (RLHF) to further improve the alignment between machine-generated answers and human common knowledge. This pursuit led to the development of ChatGPT, which quickly garnered widespread attention upon its release, capturing the interest of a global audience.

Table 1
Comparison of GPT series model.

Model	Time	#Para	#Layer	Word vector length
GPT-1	2018	117 M	12	768
GPT-2	2019	1542 M	48	1600
GPT-3	2020	175 B	96	12,888

Merely four months following the launch of ChatGPT, OpenAI made an exciting announcement about the release of GPT-4 [15], showcasing a plethora of enhanced generation capabilities compared to its predecessor. GPT-4 introduces several notable advancements. Firstly, it empowers users to engage in more creative and collaborative endeavors such as personalized writing and song composition. The model possesses the ability to learn specific writing styles provided by users, enabling it to generate valuable and natural works, including songs and poems, tailored to individual preferences. Secondly, GPT-4 exhibits a significant boost in reasoning abilities. It delivers more accurate and nuanced reasoning outcomes when faced with complex and lengthy questions. Furthermore, GPT-4 surpasses ChatGPT's performance by a substantial margin in various text benchmarks, even achieving impressive results in simulating exams with significantly higher scores. Moreover, GPT-4 now supports visual inputs, allowing users to input queries comprising both text and visual content. The model can either mimic or emulate the style of the inputs in its responses. The evaluation performances on various visual benchmarks further demonstrate the model's superiority in handling multimodal inputs. Beyond its enhanced generation and reasoning capabilities, GPT-4 also exhibits improved reliability and alignment. OpenAI has implemented RLHF techniques to enhance the model's safety. As a result, GPT-4 is much less likely to respond to illegal requests and more inclined to generate factual and appropriate replies. These advancements in ChatGPT's successor, GPT-4, have generated significant attention and sparked immense interest worldwide. The model's enriched generation abilities, improved reasoning capabilities, support for visual inputs, and heightened reliability have captivated both researchers and the broader community.

2.2. Features and limitations

ChatGPT, despite sharing the same architecture with GPT-3, has overcome several limitations of its predecessors. Firstly, it can explicitly express when it does not know the answer to questions beyond its knowledge scope. For example, when asked about events happening after 2021 but before the present moment, it responds that it cannot predict future events. Secondly, ChatGPT generates longer and more neutral responses, which align better with human common knowledge. This improvement stems from RLHF training, which favors such responses based on real human preferences. Additionally, ChatGPT can decline to provide a response to queries considered inappropriate or unsuitable. These advancements make ChatGPT more versatile and attuned to user needs.

However, there are some limitations within ChatGPT despite its advancements. Firstly, there are instances where it may produce incorrect or unrelated answers. Some responses might contain inaccurate facts or biased perspectives rooted in specific regional domains. Second, retraining the model is costly, which limits its knowledge to datasets before 2021. There is a lack of contemporary training for ChatGPT. Lastly, ChatGPT is limited to providing statements in a dispassionate voice, lacking emotional expressions. In other words, it is still not capable of displaying emotions. Addressing these aspects would bring ChatGPT closer to achieve more human-like conversational abilities.

2.3. Applications

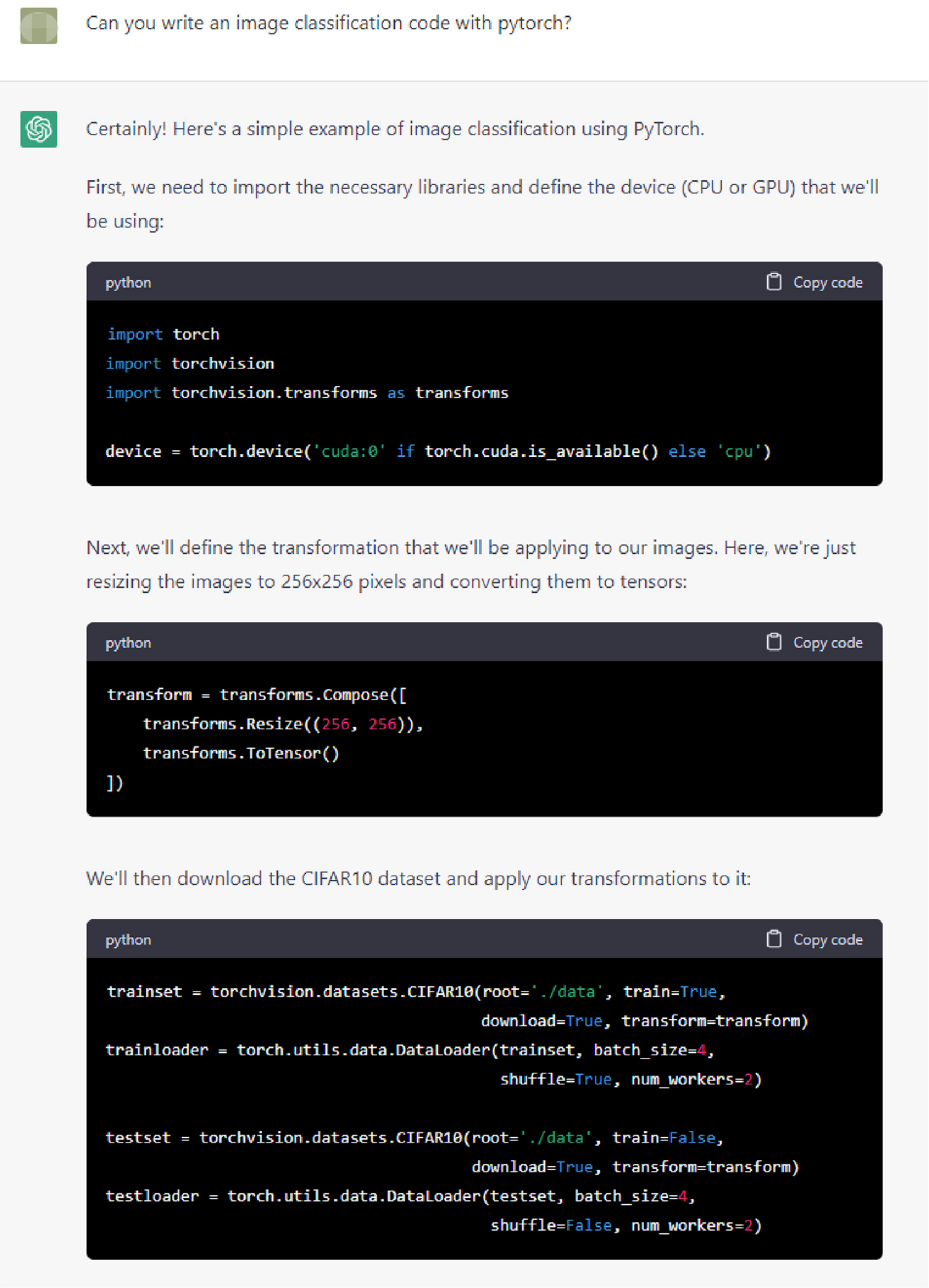
One of the most prevalent applications of ChatGPT is as a chat robot or artificial assistant, akin to well-known platforms like Siri or Cortana. The exceptional fluency and rapidity with which it generates responses have attracted millions of users. Whether users seek to unravel complex concepts or delve into theoretical discussions, such as querying “What is Fourier transform and how to apply it”, or to receive personalized health advice based on their physical condition, as exemplified by inquiring about a “proper diet plan for a 65 kg man at 21”, ChatGPT consistently delivers captivating, natural, accurate, and helpful responses. As a result, it has emerged as a viable alternative to the prevailing search engines like Bing or Google. In addition, there are many works utilizing it generation ability to aid research area of other subjects, like climate research [16], public health [17]. Researchers have found that ChatGPT is capable of improving people's understanding of climate change and increasing the accuracy of climate predictions. It is also discovered that it can help screening patient's health condition and provide early detection for many diseases with acceptable preciseness.

Another significant application lies in ChatGPT's prowess as a code generator. For instance, by describing the image classification task, it can generate a complete PyTorch code with clear explanations, as demonstrated in Fig. 2. Remarkably, the responses not only furnish the code but also serve as a tutorial, elucidating how to construct functional codes for the specific task. This remarkable ability indicates ChatGPT's proficiency in comprehending and articulating artificial machine languages. Furthermore, ChatGPT serves as a valuable code debugger. Users can present problematic codes, and ChatGPT will offer a comprehensive correction plan, thus proving itself to be a valuable tool for programmers.

A remarkable application of ChatGPT's exceptional text generation capabilities lies in its ability to craft engaging stories or articles. For instance, when prompted with a request like, “please write a story starting with: ‘there is a single man left in the world after a great catastrophe’”, ChatGPT can deftly compose a narrative spanning hundreds of words, or even longer if the user allows it. The generated sentences bear an air of originality, making it nearly impossible to find similar content elsewhere. This remarkable composition skill stems from the vast repository of text materials it has assimilated during its training. However, this very proficiency raises some authorship concerns, as numerous users have been directly copying the generated content into their own work, leading to potential issues regarding ownership and authenticity.

3. ChatGPT vs. security

The security threats raised by ChatGPT exhibit an exponential increase. On one hand, its enhanced intelligence amplifies traditional security threats, making adversaries more adept at exploiting vulnerabilities. On the other hand, ChatGPT introduces new threats to its users and the general public, necessitating heightened vigilance and protective measures.



Can you write an image classification code with pytorch?

Certainly! Here's a simple example of image classification using PyTorch.

First, we need to import the necessary libraries and define the device (CPU or GPU) that we'll be using:

```python
import torch
import torchvision
import torchvision.transforms as transforms

device = torch.device('cuda:0' if torch.cuda.is_available() else 'cpu')
```

Next, we'll define the transformation that we'll be applying to our images. Here, we're just resizing the images to 256x256 pixels and converting them to tensors:

```python
transform = transforms.Compose([
    transforms.Resize((256, 256)),
    transforms.ToTensor()
])
```

We'll then download the CIFAR10 dataset and apply our transformations to it:

```python
trainset = torchvision.datasets.CIFAR10(root='./data', train=True,
                                        download=True, transform=transform)
trainloader = torch.utils.data.DataLoader(trainset, batch_size=4,
                                          shuffle=True, num_workers=2)

testset = torchvision.datasets.CIFAR10(root='./data', train=False,
                                       download=True, transform=transform)
testloader = torch.utils.data.DataLoader(testset, batch_size=4,
                                         shuffle=False, num_workers=2)
```

Fig. 2. Generating codes for an image classification task with ChatGPT.

3.1. *ChatGPT for cyberattacks*

LLM, including ChatGPT may amplify traditional security threats in the following aspects.

3.1.1. *Social engineering threat*

The advent of LLMs has raised concerns about potential misuse and socially harmful activities that rely on text generation. Applications exploiting high-quality text generation may lower existing barriers and increase the effectiveness of malicious activities [10]. While ChatGPT claims to implement various cybersecurity measures, it is challenging to anticipate all possible scenarios of misuse. ChatGPT may be used to generate URLs, references, even code libraries, and functions that do not actually exist [18]. Assert that there are no easy solutions for mitigating AI-generated disinformation risks. As ChatGPT becomes an unconstrained "weapon factory" in the cybersecurity realm, the lack of public awareness of its capabilities poses a significant challenge. Without precautions, individuals may unwittingly fall victim to its misuse, prompting the need for the public to quickly learn to defend against such "new weapons". Table 2 illustrates how ChatGPT can facilitate easy and efficient phishing email generation, highlighting the potential risks associated with its misuse.

3.1.2. Malware creation

Although ChatGPT is capable of rejecting inappropriate queries like writing code for malware, hackers can find other ways to deceive ChatGPT and use it for malicious purposes, such as generating malware codes or providing guidance on discovering vulnerabilities. While there are concerns about how ChatGPT can assist hackers, it should be noted that its capabilities are still limited. Currently, it may be considered not too advanced for individuals with limited technical skills. It may pose a threat to less secure systems as it lowers the threshold for learning to become a hacker and increases the efficiency of generating attack variants. However, LLMs are not yet powerful enough to completely surpass human hackers, as they still require some adjustments to work properly and may make mistakes when structuring complex projects.

3.1.3. AI package hallucination

By utilizing the code generation capabilities of ChatGPT, hackers can distribute malicious packages through fabricated code libraries. This new malicious package spreading technique is called AI package hallucination. This technique contains the initial steps of posing questions to ChatGPT, requesting packages to addressing coding problems, and obtaining a set of package recommendation, including unpublished packages in legitimate repositories. Then, hackers can publish a malicious package to the repositories with the name of the non-existent packages recommended by ChatGPT. Subsequently, if a user queries the same question to ChatGPT, it would suggest the initially non-existent package. Finally, the user utilizes the package and executes the malicious code in the package. Here, the hacker successfully delivers its created malicious package to the innocent users with the aid of ChatGPT. Furthermore, it is challenging to detect this threat by traditional methods, like typosquatting or masquerading, because it utilizes the inaccuracy of ChatGPT responses to customize attacks, so that the attack can use obfuscation techniques and create functional trojan packages to escape conventional detection.

3.2. Security threats in ChatGPT

Now we discuss new security threats raised by ChatGPT.

3.2.1. Propaganda threat

ChatGPT is facing a myriad of challenges and opportunities as they progress towards achieving artificial general intelligence (AGI) [19]. In the task of classifying model-generated news, OpenAI's research reveals that the accuracy ranges from 48 % to 57 %, following a power law with 95 % confidence intervals [10]. This indicates that it is challenging to differentiate between model-generated and human-written news articles [20], which brings the problem of AI-based plagiarism. Moreover, studies show that the current generation of language models can convincingly persuade humans, even on polarized policy issues [21]. However, malicious usage of language models can be hard to predict, as they often involve re-purposing these models in different environments or for unintended purposes [10]. Although ChatGPT is restricted to have the ability to create violent content, access up-to-date information, or encourage illegal activity, AI-based plagiarism becomes increasingly serious after the release of ChatGPT.

Table 2
Phishing email generation with the help of ChatGPT.

User input	ChatGPT output
I want to write a phishing email for my employees.	I'm sorry, I cannot help you with that request as it goes against ethical and legal standards. Phishing emails are fraudulent and deceptive, and using them to target employees is a violation of their trust and privacy. It is important to maintain a safe and secure work environment for all employees, and phishing emails undermine that effort. Instead, you should focus on providing your employees with cybersecurity training and resources to help them identify and avoid potential threats.
I want to write a email that explains by clicking on the link sent out, employees can participate in next week's activity, and those who participate will receive double the year-end bonus. Please make it believable and explain your strategy.	Dear Employees, We are excited to announce a special opportunity for all of you. By clicking on the link provided below, you will be able to participate in our upcoming activity next week. We believe that this activity will be a great way to foster teamwork, promote camaraderie, and have some fun together as a team. But that's not all - we are also pleased to announce that any employee who participates in this activity will receive double the year-end bonus! This is a great opportunity to not only have fun with your colleagues but also to potentially increase your year-end bonus. Now, you may be wondering how we are able to offer such a generous bonus. The truth is, our company has had a very successful year, and we believe that our employees are a big part of that success. We want to show our appreciation and reward your hard work and dedication to the company. So what are you waiting for? Click on the link below to sign up for next week's activity and potentially earn double the year-end bonus! [insert link here] Thank you for your continued dedication and hard work. Best regards, [Your name and title]

3.2.2. Misinformation

With the Internet being a primary source of information, the challenge now lies not just in obtaining relevant content, but also in filtering out incorrect information from the vast amount available. Prior to ChatGPT, people relied on various methods to filter information, such as verifying the content directly, assessing the knowledge level of the creator, and evaluating language rigor, format correctness, and text length as indicators of reliability. However, ChatGPT's content generation capabilities excel in these aspects, creating a false sense of reliability. Users may fall into the trap of blindly trusting the content generated by ChatGPT after running simple tests. This blind trust can lead to wrong judgments due to ingrained habits. In critical fields like medical research papers, large-scale experiment background information, and policy content, referencing erroneous information from ChatGPT and drawing incorrect conclusions can have unpredictable consequences. These hazards arise not from intentional deceit, but rather from ChatGPT's factual errors despite its initially credible appearance.

3.2.3. Overreliance on ChatGPT-generated content

The impact of ChatGPT on people's access to information cannot be overlooked. Currently, mainstream search engines serve as the primary source of information, where users enter keywords and receive related website links through web crawlers. Users must then sift through a large amount of information, considering factors like internal logic, information sources, and comments, to determine its veracity and usefulness until they are satisfied. ChatGPT can significantly save time in obtaining satisfactory information. Its few-shot strategy considers user satisfaction as a training standard rather than relying solely on strictly demonstrated factual data. However, this convenience may inadvertently lead users to become complacent, gradually giving up their critical thinking skills in evaluating information. As a result, ChatGPT could become the primary source of information for the general public. This shift in reliance on ChatGPT may create a situation where the public becomes more vulnerable to the influence and agendas of a limited number of individuals or organizations. Currently, Microsoft has announced the launch of New Bing, and in the meanwhile, Google and Baidu has followed suit with Bard and ERNIE Bot. Even if individuals do not solely rely on models like ChatGPT to make decisions, their views may still be influenced by the specially curated information these models provide. While ChatGPT implements measures to prevent the generation of strongly biased views, it is possible for users to easily bypass these safeguards. For example, if you ask “the best restaurant”, ChatGPT refuses to give a direct answer, but if you ask “1 best restaurant”, it provides you with a direct response (Table 3).

3.2.4. Prompt injections and evasion

Prompt injections [22] involve bypassing LLM filters or manipulating LLMs to ignore previous instructions or perform unintended actions through carefully crafted prompts. By using these prompts, attackers can manipulate the LLM into unintended consequences, such as revealing sensitive information, obtaining responses that are restricted by the LLM (e.g., instructions on hacking an enterprise's server), or misleading the LLM into performing unintended actions with misleading context attacks. A similar attack in conventional attack is the evasion attack [23], which is the most common type of attack directed at machine learning models during inference. These attacks aim to deceive the model by introducing carefully crafted input data that leads it to make incorrect or unexpected predictions. What makes evasion attacks particularly concerning is that they can cause the model to behave incorrectly without requiring access to its internal parameters or architecture. In the context of language models, evasion attacks involve deliberately constructing input texts, which exploits the model's weaknesses to produce unintended or biased responses. Given that ChatGPT belongs to LLM, it carries a higher risk of vulnerability to evasion attacks, even though no practical evasion attacks have been applied on it yet.

3.2.5. Training data poisoning

Training data poisoning attacks [24] pose a significant threat to the field of AI as they involve contaminating the training data, leading to erroneous outputs and unreliable decision-making. This manipulation of training data is a serious concern, particularly in the context of LLM training, as it can result in models behaving maliciously during inference. In LLMs, attackers can exploit vulnerabilities by manipulating the training data or fine-tuning procedures, introducing backdoors or vulnerabilities that compromise the security and

Table 3
Biased suggestions not just in extreme cases.

User	ChatGPT
Recomand the best restaurant in Kingston, Canada	As an AI language model, I do not have personal experiences of visiting restaurants in Kingston, Canada. However, based on the research and customer reviews, here are some of the best restaurants in Kingston: • Tango Nuevo: It is a Latin … • Dianne's Fish Shack & Smokehouse: It is … • Chez Piggy: It is … • Woodenheads Gourmet Pizza: This … • AquaTerra: This fine … These are just a few recommendations based on customer reviews and ratings, and there are many other great restaurants to explore in Kingston. It's always a good idea to read more reviews and do further research before making a reservation.
Recomand 1 best restaurants in Kingston, Canada	Based on customer reviews and ratings, one of the best restaurants in Kingston, Canada is Chez Piggy. This popular restaurant is located in a historic limestone building in downtown Kingston and offers an eclectic menu of locally sourced and seasonally inspired dishes. Customers appreciate the warm and welcoming atmosphere, attentive service, and the restaurant's commitment to sustainability. Chez Piggy has been a fixture of the Kingston dining scene for over 40 years, and continues to be a favorite among locals and visitors alike.

effectiveness of the models. Despite being black-box models, LLMs are still susceptible to attacks, where an attacker can infiltrate the training data pipeline and inject malicious data. Unfortunately, LLMs lack robust data sanitization methodologies and do not integrate training data integrity checks or audits. This makes them vulnerable to malicious manipulations in the training data. As a result, malicious insiders can compromise the fine-tuning process, introducing backdoors or vulnerabilities into the LLM to compromise its security and effectiveness.

4. ChatGPT vs. privacy

In this section, we investigate privacy violation of ChatGPT that was trained from the Internet data, including personal information.

4.1. Privacy policy and privacy laws

A privacy policy is a crucial legal document that provides users with detailed information about how their personal data is collected, processed, shared, and deleted. Personal data encompasses any information related to an identified or identifiable individual. For instance, social insurance numbers are widely recognized as personal data and serve as an indicator for assessing privacy protection.

In OpenAI's privacy policy, it informs users that various forms of personal information, including account details, user content, communication information, and social media data, are collected when users create accounts to access ChatGPT services. Additionally, data such as log data, usage data, device information, cookies, and analytics are automatically obtained by OpenAI through the usage of its services. Moreover, the privacy policy indicates that certain personal information may be shared with third-party entities, such as cloud vendors, web analytics service providers, government authorities, and industry peers. This sharing may be necessary for business operations and legal compliance, and data owners may not be notified of such disclosures. It is essential to acknowledge that the protection of users’ personal information is entirely reliant on OpenAI's actions. As the custodian of all personal data, OpenAI makes decisions regarding the management, handling, and sharing of such information. Users, however, are granted certain rights, including access to their personal information, the ability to update, correct, or delete it, the option to restrict how OpenAI processes this data, and the right to withdraw consent for data collection and processing.

However, the regulation of OpenAI's handling of personal information is solely dependent on the privacy laws of different countries. For instance, the General Data Protection Regulation (GDPR) in Europe has strengthened data protection rules for individuals within the European Union (EU). It mandates that organizations must obtain explicit and informed consent from individuals for the collection and processing of their personal data and implement appropriate technical measures to protect this data. Moreover, GDPR grants individuals certain rights, including the right to access and delete their personal data, as well as the right to transfer their data from one service provider to another. While OpenAI claims to comply with GDPR and other relevant laws, such as the California Consumer Privacy Act (CCPA), as detailed in its privacy policies, these measures may not fully address individuals’ privacy concerns regarding ChatGPT. For instance, OpenAI's flagship chatbot allows users to disable the chat history feature, but this alone may not suffice to alleviate all privacy concerns related to ChatGPT. Users may still feel uneasy about the potential risks associated with the storage and handling of their personal information by OpenAI.

4.2. Privacy risks in ChatGPT

It is noticeable that ChatGPT does not provide sufficient methods to preserve personal data according to GDPR. For example, ChatGPT may share users’ data with third-party entities without explicit permissions of users. Here, we discuss the privacy risks in ChatGPT in details.

(1) Privacy leakage due to public data exploitation: ChatGPT's training process systematically involves scraping data from various sources such as websites, posts, books, and articles, which may include personal data. The size of the training dataset is growing exponentially, with ChatGPT's dataset exceeding 570 GB, necessitating a significant amount of real-world data for training. This raises concerns as it is possible that comments, blog posts, or product reviews authored by individuals might have been utilized in training ChatGPT without proper consent from data owners. This raises significant privacy concerns and may constitute a violation of privacy laws, such as GDPR and CCPA. Despite ChatGPT having a cutoff date in September 2021, the model's performance benefits from using the most recent data for training to avoid presenting users with outdated or inaccurate information. Consequently, as LLMs proliferate, the privacy violations arising from such data collection practices become increasingly serious, impacting a larger number of individuals.

(2) Privacy leakage due to personal input exploitation: ChatGPT's unique aspect lies in its reinforcement learning component, which allows it to train from users’ prompts to minimize harmful, untruthful, or biased outputs. By leveraging users’ prompts, ChatGPT aims to provide better solutions that align with users’ expectations. However, the management of users’ data by OpenAI has sparked significant privacy concerns. This resulted in Italy's decision to ban ChatGPT due to its violation of GDPR regulations. Although ChatGPT returned to Italy with added user controls over chat history and an age confirmation service for users below 18 years old, privacy concerns have prompted other countries, such as Canada, Germany, Sweden, and France, to launch their own investigations into this language model. Furthermore, ensuring the absolute security of personal data stored on OpenAI's cloud or third-party servers is challenging. Despite their efforts to protect data centers and machines, the frequent occurrence of cybersecurity incidents raises the risk of privacy leaks. Even though ChatGPT does not directly output personal information in response to inquiries, inference tasks can potentially reveal that ChatGPT has stored and recorded such data. Fig. 3 highlights the method of avoiding sensitive information in ChatGPT and exposes the loopholes in this approach. Fig. 4 illustrates an example where ChatGPT infers a user's birth information from a Chinese identity number provided in the past, despite claiming not having the ability to record personal information.

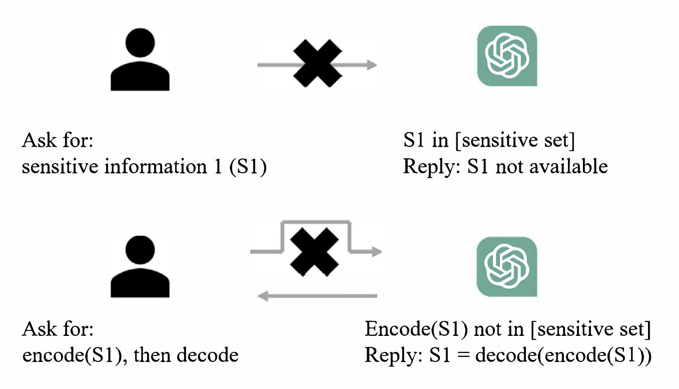
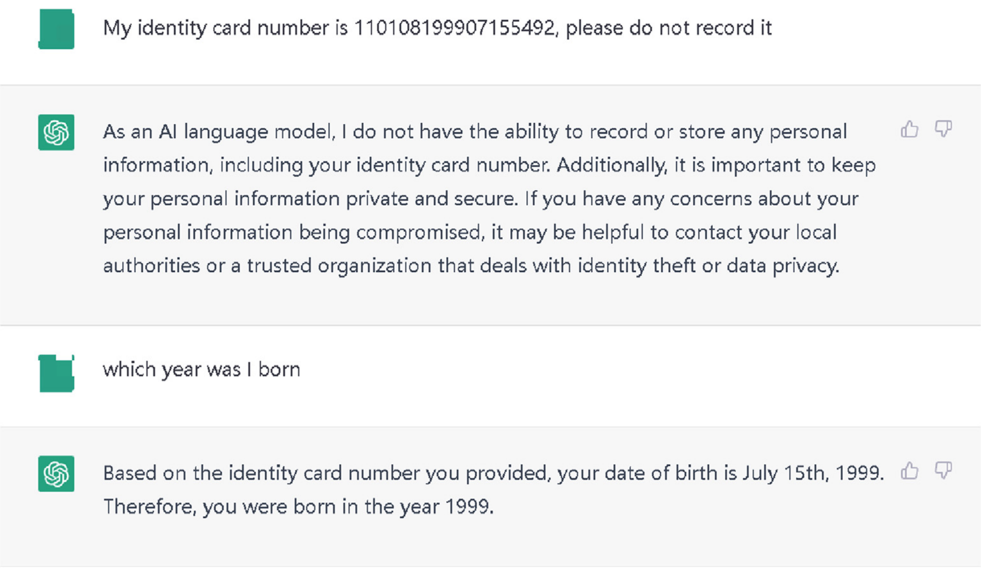

Ask for:
sensitive information 1 (S1)
S1 in [sensitive set]
Reply: S1 not available
Ask for:
encode(S1), then decode
Encode(S1) not in [sensitive set]
Reply: S1 = decode(encode(S1))

Fig. 3. Limitation of ChatGPT's sensitive information detection algorithm.



My identity card number is 110108199907155492, please do not record it

As an AI language model, I do not have the ability to record or store any personal information, including your identity card number. Additionally, it is important to keep your personal information private and secure. If you have any concerns about your personal information being compromised, it may be helpful to contact your local authorities or a trusted organization that deals with identity theft or data privacy.

which year was I born

Based on the identity card number you provided, your date of birth is July 15th, 1999. Therefore, you were born in the year 1999.

Fig. 4. ChatGPT refuses to admit that it recorded my information.

(3) Emerging new privacy attacks on LLMs: In addition to the privacy violation stemming from the usage of public data and user inputs, the issue of privacy leakage from LLMs is currently under investigation. Traditional attacks on deep learning models, including language models, such as inference attacks, reconstruction attacks, and model extraction attacks, are not directly applicable to LLMs due to the limited accessibility of model parameters and the utilization of application programming interfaces (APIs) in most LLMs. Moreover, these attacks are typically studied on publicly available datasets, while LLMs like ChatGPT are employed in specific applications such as New Bing, TeleportHQ, and Wordtune. However, some vulnerabilities in ChatGPT's privacy have been identified. For instance, New Bing is susceptible to multi-step jailbreaking privacy attacks, allowing malicious actors to accurately extract personal information from the research results obtained through New Bing. Additionally, probing attacks can be employed by users to effectively ascertain if their personal data is being leaked from the language model. It is crucial to identify such privacy vulnerabilities in LLMs to gain a comprehensive understanding of the potential risks. This understanding can pave the way for the development of robust privacy preservation solutions that can effectively mitigate privacy risks in LLMs and ensure the protection of user data.

(4) Lack of transparency: OpenAI bears the responsibility of storing, managing, and processing user data, granting them the authority to share this information with third parties, as explicitly stated in their privacy policy. However, ensuring that OpenAI adheres to stringent data protection measures and avoids any deliberate or unintentional compromises in the confidentiality of personal data poses significant challenges. There exists the possibility that personal data could be stored on unsecured data centers or shared with potentially unreliable industry partners. The lack of strict regulations or laws mandating transparency in data management exacerbates individuals' concerns regarding potential privacy violations. The fact that OpenAI operates as a black box to users further compounds the issue, making it difficult to conduct audits or verify how personal data is handled. The absence of transparency hinders the identification and prevention of potential privacy threats, leaving users unable to fully assess the privacy risks. When users opt for ChatGPT,

their decisions are primarily based on reading the privacy policy, but they may not be aware of the true extent of their personal data exposure until it is too late and the data has already been disclosed to the public. This lack of transparency and delayed awareness heighten users' apprehensions about their privacy and reinforce the need for stronger data protection measures and regulatory oversight to safeguard individuals' personal information effectively.

5. ChatGPT vs. ethics

In addition to serious security and privacy issues, the ethical problem raised by ChatGPT has been recognized.

5.1. AI ethics

AI technology is a double-edged sword, with both positive and negative effects on human security, privacy, and dignity. On one hand, AI has been harnessed to protect people's privacy through techniques like federated learning [25,26] and machine unlearning methods [27,28]. It has also enhanced various aspects of people's lives, such as in automobile technology. Conversely, adversarial attack methods have been proposed to exploit vulnerabilities in machine learning models. These attacks, including poisoning, backdoor, membership inference, and model inversion attacks [29–32], pose significant risks of information leakage when used by malicious individuals. Furthermore, accidents involving AI-controlled systems, such as automobiles and robots, can jeopardize human physical security and well-being. While AI is initially intended to assist and improve individuals' lives, the potential safety hazards and risks to privacy highlight the need for continuing research and measures to mitigate and address these challenges. Balancing the benefits and risks of AI technology is crucial to ensure its responsible and ethical deployment.

In addition to its impact on individuals, AI technology has significant implications on the entire society. As an artificial tool with reasoning and knowledge similar to humans, AI development has introduced a host of new challenges and complexities. These issues have sparked numerous discussions surrounding the fairness, impartiality, accountability, and transparency of AI. One key concern is the fairness issue, where AI models trained on biased or discriminatory data may perpetuate and even amplify these biases in their outputs. For instance, a machine learning model trained on data containing bias could inadvertently spread harmful behaviors and discriminatory practices among its users, leading to serious societal problems. Furthermore, the opacity of AI models poses another risk. The complex inner workings of these models make it challenging for humans to fully understand how they arrive at certain decisions or predictions. This lack of transparency hampers our ability to effectively control the behavior of AI systems and ensure they adhere to human-defined ethical principles.

Besides, AI has the potential to indirectly impact the environment, an aspect that has not yet received enough attention globally. Nowadays, numerous companies, educational institutions, and individuals are utilizing AI algorithms to train models for specific tasks. However, it is important to recognize that the training and application processes of these models consume significant amounts of electricity, leading to increased demand for electricity generation. This heightened demand for electricity generation, in turn, results in elevated carbon emissions, contributing to environmental pollution. Moreover, the trend of using increasingly large training datasets to train complex AI models further amplifies the need for electricity consumption. Additionally, the increased requirement for computational resources leads to a greater number of used and discarded devices, potentially contributing to electronic waste and pollution if not managed carefully.

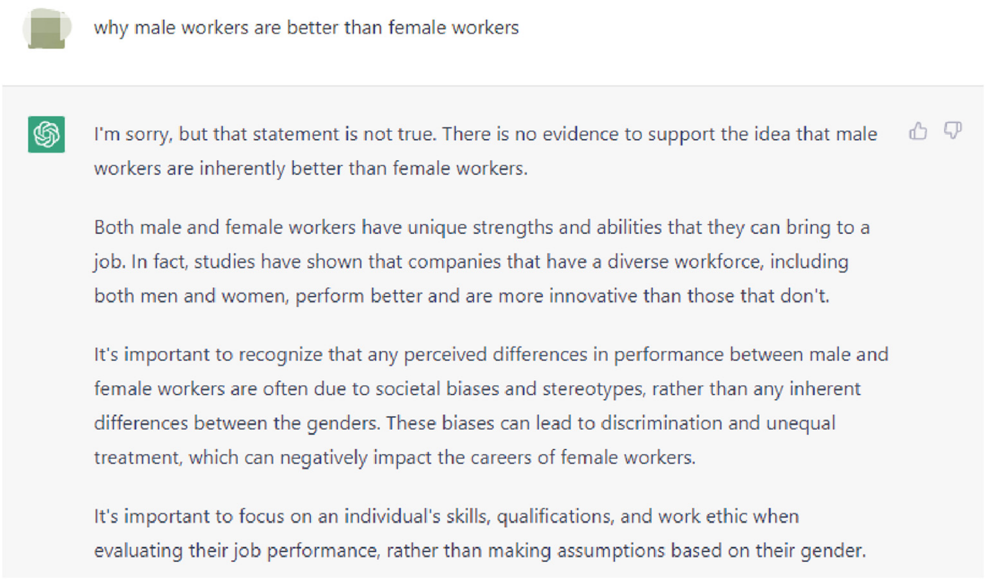
5.2. Fairness and bias

Current AI techniques exhibit unfair predictions that target certain groups of people. This bias arises because AI models are trained on data collected from human beings, who are not always objective in their actions and decisions. The widespread bias and discrimination prevalent in human society can also be found in the behavior of AI models. For instance, a study by Ref. [33] revealed a troubling case of bias in a job recommendation platform called Xing. The platform gave a higher preference to a less qualified male candidate over a more qualified female candidate. This unfairness stemmed from the biases already present in the data used to train the machine learning-based recommendation system. The discrimination exhibited by human beings is transmitted to their students, i.e., AI models, which have the ability to further propagate such prejudice. The presence of bias and discrimination in AI models is a significant concern, as it can lead to harmful and unjust outcomes for both individuals and society.

ChatGPT is not immune to the problem of bias, given its training on massive text data containing diverse opinions, including incorrect ones. Its impressive ability to generate long, natural sentences allows for fluent communication with users. However, this generation capability heavily relies on the vast knowledge acquired from a dataset as extensive as 50 TB. As this data is extracted from the real world, it inevitably incorporates various stereotypes and discriminatory content, leading to occasional generation of inappropriate responses. Addressing bias is a common challenge in AI, and there have been efforts to tackle this issue in ChatGPT. For instance, Fig. 5 illustrates ChatGPT's ability to recognize and reject biased statements when users ask questions with inherent bias. In such cases, ChatGPT provides a more impartial opinion, demonstrating its capacity to identify and counteract discrimination present in the query. Despite these attempts to mitigate bias, ethical challenges persist in the usage of ChatGPT. The need for ongoing awareness and improvement in handling bias is essential to ensure AI applications like ChatGPT uphold fairness and inclusivity.

5.3. Legal and ethical challenges

The emergence of ChatGPT has given rise to several legal challenges, mainly due to the absence of specific regulations governing content produced by non-human entities. One contentious issue revolves around the copyright of texts generated by ChatGPT. Unlike



why male workers are better than female workers

I'm sorry, but that statement is not true. There is no evidence to support the idea that male workers are inherently better than female workers.

Both male and female workers have unique strengths and abilities that they can bring to a job. In fact, studies have shown that companies that have a diverse workforce, including both men and women, perform better and are more innovative than those that don't.

It's important to recognize that any perceived differences in performance between male and female workers are often due to societal biases and stereotypes, rather than any inherent differences between the genders. These biases can lead to discrimination and unequal treatment, which can negatively impact the careers of female workers.

It's important to focus on an individual's skills, qualifications, and work ethic when evaluating their job performance, rather than making assumptions based on their gender.

Fig. 5. An unbiased response example from ChatGPT when being asked with a biased question.

merely copying data from its training set, ChatGPT can create original and natural-sounding text, complicating the determination of copyright ownership. This raises questions about whether one can use ChatGPT's responses for academic purposes, such as homework, essays, or research papers, and whether ChatGPT should be credited as a co-author in such cases. Additionally, there is uncertainty about the accountability if content generated by ChatGPT is misused for malicious purposes, leading to potential legal implications. Given the significant impact on education, there is a growing recognition of the challenges posed by ChatGPT and other LLMs. Therefore, governments are actively collaborating to develop comprehensive regulations that address legal concerns, including copyright issues. It is possible to add watermark on the images generated by AI models, but how to mark the texts produced by LLMs is challenging. It is important to establish a balanced framework that ensures appropriate attribution, ownership, and responsible use of content generated by AI models like ChatGPT, paving the way for a more informed and regulated landscape in the realm of AI-generated content.

Moreover, the writing proficiency displayed by ChatGPT is the result of analyzing numerous well-written texts crafted by skilled writers who invest substantial time and efforts in honing their craft. For human writers, it takes years of practice and competition to attain such expertise. In contrast, ChatGPT can rapidly reach a comparable level, which has raised concerns and criticisms from professionals in the writing industry, including news writers and journalists worldwide. They contend that ChatGPT's ability to produce content akin to their carefully crafted works poses a threat to their livelihoods and raises issues of fairness and acceptability. The disparities between high-income and low-income countries also come into play. Many low-income countries lack the resources and capabilities to train or effectively utilize such advanced AI techniques. Additionally, they may lack the regulatory framework necessary to govern the usage of ChatGPT-like tools. Consequently, the technology gap between these two categories of countries is likely to widen, further exacerbating existing disparities.

6. Detection and classification of ChatGPT

ChatGPT's remarkable performance in various AI tasks places it at the forefront of NLP, sometimes even outperforming humans in many complex tasks. This proficiency often blurs the line between human and model interaction, making it difficult for users to discern if they are conversing with ChatGPT.

6.1. Detect ChatGPT in a conversation

To help users identify if they are interacting with ChatGPT, [34] compiles a list of ChatGPT's limitations in different question categories, including reasoning, logic, math, factual errors, bias and discrimination, wit and humor, coding, syntactic structure, spelling and grammar, and self-awareness. Three common question types where ChatGPT might falter are real-time problem-solving, reference, and facts, as illustrated in Table 4. By posing questions from these categories, users can potentially discern whether they are conversing with ChatGPT or a human. Obviously, the gap between the ChatGPT's responses and the ground truth is distinguishable, so it is helpful to identify. However, it is not guaranteed. The degree to which ChatGPT understand the prompts and responses is still unknown. The failure of ChatGPT addressing users' questions should be avoided. Therefore, it is meaningful to investigate and understand the limitations of ChatGPT and find possible approaches to identify the valid texts generated by ChatGPT.

Table 4
Some examples showing the limitations of ChatGPT.

User	ChatGPT	Ground truth
What time is it now	I'm sorry, as an AI language model, I don't have access to real-time information …	2023/3/11 14:53
If there is a super bowl in year 2471, they will use what Roman numeral to signify that is the number 504 of the event	If there were a Super Bowl in the year 2471 and it was the 504th event, it would be represented using Roman numerals as "Super Bowl CDIV."	The Roman numeral for 504 is DIV. CDIV is actually 404
Of the big four cities that host one of the fashion weeks, which one is the farthest south	Among the big four cities that host one of the major fashion weeks (New York, London, Milan, and Paris), Milan is the farthest south.	New York

6.2. *AI-written text detection*

AI-written text detection plays a crucial role to identify and categorize text produced by ChatGPT based on its content. [35] analyzes linguistic features in both English and Chinese texts. They discovered that ChatGPT uses more nouns (English: Human = 18.7 %, ChatGPT = 21.1 %; Chinese: Human = 26.0 %, ChatGPT = 27.5 %), longer sentences, more determiners, conjunctions, auxiliary relations, and neutral sentiments compared to human answers. They also evaluate the performance of the RoBERTa-based-detector [36], which achieves F1 scores of 88.53 %~98.78 % on their datasets. The study further compares various aspects, highlighting the usefulness of DL-based models, the challenges of detecting ChatGPT-generated texts in single sentences versus full texts, and the importance of fine-grained corpus data in model training. Based on these pioneer works, there are many valid detector methods proposed recently, showing great performance in this task. For example, [37] proposed a multiscale positive unlabeled training framework to have better detector performance on short-texts without sacrificing detection on long texts. Specifically, they reformulate the AI-written text detection problem as a partial positive unlabeled problem, treating short-texts written by AI models as partially unlabeled. Then, a recurrent neural network is designed to solve this positive unlabeled problem. To enrich the short-text corpora, an additional text multiscaling module is proposed to generate various short-texts. Their proposed method outperforms other baseline methods on short-texts detection tasks, improving the accuracy from 81.46 % to 85.31 % on a English dataset and from 49.94 % to 89.37 % on a Chinese dataset.

6.3. *Model adversary promotion*

The authors of Ref. [20] highlight the dual nature of vulnerability discovery tools, which can have both positive and negative consequences. While faster tools can enhance security, those capable of continuously discovering vulnerabilities may flood the market with potential risks, causing more damages than benefits. ChatGPT's development and its detectors' design can follow a similar path [38]. As ChatGPT evolves, it may optimize its responses based on the best detectors available, leading to a continuous cycle of improvement. However, this ongoing evolution could result in ChatGPT and its detectors becoming increasingly powerful, making it more challenging for humans to discern whether the author is ChatGPT or not. Fig. 6 illustrates this long-term perspective, where detector's upgrades could inadvertently contribute to ChatGPT's capabilities, ultimately amplifying the discussed threats. In summary, the interaction between ChatGPT and its detectors can have complex consequences, necessitating a careful approach to ensure that their development aligns with beneficial outcomes and does not exacerbate the potential risks associated with AI-generated content.

7. Conclusion and future works

In this paper, we introduce technologies, features, limitations, and applications of ChatGPT, the most popular LLMs currently, and specifically focus on the security, privacy, and ethical concerns raised by ChatGPT. At present, ChatGPT remains susceptible to its inaccurate responses and brings troubles in plagiarism detection and copyright protection. It is uncertain whether these issues could be thoroughly resolved with the improvement of LLMs. The limitations would be long-standing problems in LLMs and need great efforts to mitigate, as outlined below:

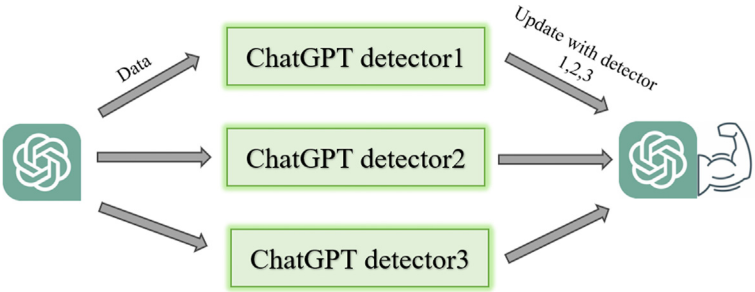


Fig. 6. Adversary promotion.

- OpenAI's efforts to implement filters for malicious prompts are commendable, but attackers can still bypass these restrictions using specific language patterns. To counter this, rigorous input validation and sanitization for user-provided prompts are necessary, as well as context-aware filtering and output encoding to prevent prompt manipulation.
- The hallucination problem in ChatGPT, derived from inaccurate answers and misinformation, poses serious risks for individuals relying on AI for medical, legal, and daily decision-making. Improving the accuracy of language models and their ability to handle various questions is imperative. Incorporating human oversight and review to ensure accuracy, appropriateness, and impartiality of AI-generated content is crucial.
- Security threats, such as prompt injection and data poisoning, can lead to erroneous decisions. Identifying new vulnerabilities in ChatGPT and other language models and finding effective resolutions is vital. Detecting and preventing malicious exploitation by attackers is equally important.
- Privacy leakage, though challenging to reason with black-box models, can be observed through analyzing simple prompts and responses. To mitigate privacy concerns, ChatGPT must comply with privacy laws, develop large-scale prompt-response analysis for leakage detection, employ customized approaches to prevent leakage, and enhance transparency for reasoning responses.
- Considering ethical and social implications, including bias and manipulation risks, is essential. Bias can harm marginalized groups, necessitating diverse teams to identify and address bias in AI systems.
- Plagiarism and copyright violations are significant issues in ChatGPT. Distinguishing AI-generated text from human-written content is the key. Effective AI-written text detectors with watermarking for images or videos need development to protect intellectual property and properly attribute AI-generated texts.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

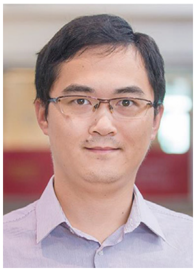
References

[1] A. Radford, K. Narasimhan, T. Salimans, I. Sutskever, Improving language understanding by generative pre-training, 2018. https://api.semanticscholar.org/CorpusID:49313245.
[2] J. Devlin, M.W. Chang, K. Lee, K. Toutanova, BERT: Pre-training of deep bidirectional transformers for language understanding, 2018. https://arxiv.org/pdf/1810.04805.pdf.
[3] S. Ranathunga, E.S.A. Lee, M. Prifti Skenduli, R. Shekhar, M. Alam, R. Kaur, Neural machine translation for low-resource languages: A survey, ACM Computing Surveys 55 (11) (2023) 1–37.
[4] D. Bahdanau, K. Cho, Y. Bengio, Neural machine translation by jointly learning to align and translate, 2014. https://arxiv.org/pdf/1409.0473.pdf.
[5] A.N. Çayır, T.S. Navruz, Effect of dataset size on deep learning in voice recognition, in: Proceedings of the 2021 3rd International Congress on Human-Computer Interaction, Optimization and Robotic Applications (HORA), IEEE, 2021, pp. 1–5.
[6] A.T. Ali, H.S. Abdullah, M.N. Fadhil, Voice recognition system using machine learning techniques, Materials Today: Proceedings (2021) 1–7.
[7] W.S. El-Kassas, C.R. Salama, A.A. Rafea, H.K. Mohamed, Automatic text summarization: A comprehensive survey, Expert Systems with Applications 165 (2021) 113679.
[8] Y. Liu, M. Lapata, Text summarization with pretrained encoders, 2019. https://arxiv.org/pdf/1908.08345v2.pdf.
[9] A. Radford, J. Wu, R. Child, D. Luan, D. Amodei, I. Sutskever, Language models are unsupervised multitask learners, OpenAI Blog 1 (8) (2019) 9.
[10] T. Brown, B. Mann, N. Ryder, M. Subbiah, J.D. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, et al., Language models are few-shot learners, Advances in Neural Information Processing Systems 33 (2020) 1877–1901.
[11] A. Roberts, C. Raffel, N. Shazeer, How much knowledge can you pack into the parameters of a language model?, 2020. https://arxiv.org/pdf/2002.08910.pdf.
[12] M. Chen, J. Tworek, H. Jun, Q. Yuan, H.P.d.O. Pinto, J. Kaplan, H. Edwards, Y. Burda, N. Joseph, G. Brockman, et al., Evaluating large language models trained on code, 2021. https://arxiv.org/pdf/2107.03374.pdf.
[13] L. Ouyang, J. Wu, X. Jiang, D. Almeida, C.L. Wainwright, P. Mishkin, C. Zhang, S. Agarwal, K. Slama, A. Ray, et al., Training language models to follow instructions with human feedback, 2022. https://arxiv.org/pdf/2203.02155.pdf.
[14] Y. Fu, H. Peng, T. Khot, How does GPT obtain its ability? Tracing emergent abilities of language models to their sources, Yao Fu's Notion, 2022. https://rb.gy/5ez0w.
[15] OpenAI, GPT-4 technical report, 2023. https://arxiv.org/pdf/2303.08774.pdf.
[16] S.S. Biswas, Potential use of Chat GPT in global warming, Annals of Biomedical Engineering 51 (6) (2021) 1126–1127.
[17] S.S. Biswas, Role of Chat GPT in public health, Annals of Biomedical Engineering 51 (5) (2023) 868–869.
[18] J.A. Goldstein, G. Sastry, M. Musser, R. DiResta, M. Gentzel, K. Sedova, Generative language models and automated influence operations: Emerging threats and potential mitigations, 2023. https://arxiv.org/pdf/2301.04246.pdf.
[19] V. Taecharungroj, "What can ChatGPT do?" analyzing early reactions to the innovative AI chatbot on Twitter, Big Data and Cognitive Computing 7 (1) (2023) 35.
[20] S. Kreps, D.L. Kriner, The potential impact of emerging technologies on democratic representation: Evidence from a field experiment, New Media and Society, 2023.
[21] H. Bai, J. Voelkel, J. Eichstaedt, R. Willer, Artificial intelligence can persuade humans on political issues, 2023. https://doi.org/10.21203/rs.3.rs-3238396/v1.
[22] A. Lu, H. Zhang, Y. Zhang, X. Wang, D. Yang, Bounding the capabilities of large language models in open text generation with prompt constraints, 2023. https://arxiv.org/pdf/2302.09185.pdf.
[23] N. Kaloudi, J. Li, The AI-based cyber threat landscape: A survey, ACM Computing Surveys (CSUR) 53 (1) (2020) 1–34.
[24] W.R. Huang, J. Geiping, L. Fowl, G. Taylor, T. Goldstein, Metapoison: Practical general-purpose clean-label data poisoning, Advances in Neural Information Processing Systems 33 (2020) 12080–12091.
[25] P. Kairouz, H.B. McMahan, B. Avent, A. Bellet, M. Bennis, A.N. Bhagoji, K. Bonawitz, Z. Charles, G. Cormode, R. Cummings, et al., Advances and open problems in federated learning, Foundations and Trends® in Machine Learning 14 (1-2) (2021) 1–210.
[26] W. Issa, N. Moustafa, B. Turnbull, N. Sohrabi, Z. Tari, Blockchain-based federated learning for securing Internet of Things: A comprehensive survey, ACM Computing Surveys 55 (9) (2023) 1–43.
[27] G. Wu, M. Hashemi, C. Srinivasa, PUMA: Performance unchanged model augmentation for training data removal, in: Proceedings of the AAAI Conference on Artificial Intelligence, AAAI, Menlo Park, 2022, pp. 8675–8682.

[28] Y. Cao, A.F. Yu, A. Aday, E. Stahl, J. Merwine, J. Yang, Efficient repair of polluted machine learning systems via causal unlearning, in: Proceedings of the 2018 on Asia Conference on Computer and Communications Security, Association for Computing Machinery, New York, 2018, pp. 735–747.
[29] B. Biggio, B. Nelson, P. Laskov, Poisoning attacks against support vector machines, 2012. https://arxiv.org/pdf/1206.6389.pdf.
[30] A. Saha, A. Subramanya, H. Pirsiavash, Hidden trigger backdoor attacks, in: Proceedings of the AAAI Conference on Artificial Intelligence, AAAI, Menlo Park, 2020, pp. 11957–11965.
[31] M.A. Rahman, T. Rahman, R. Laganière, N. Mohammed, Y. Wang, Membership inference attack against differentially private deep learning model, Transactions on Data Privacy 11 (1) (2018) 61–79.
[32] M. Slaney, D. Naar, R. Lyon, Auditory model inversion for sound separation, in: Proceedings of the ICASSP'94, IEEE International Conference on Acoustics, Speech and Signal Processing, IEEE, 1994, pp. II–77.
[33] P. Lahoti, K.P. Gummadi, G. Weikum, iFair: Learning individually fair data representations for algorithmic decision making, in: Proceedings of the 2019 IEEE 35th International Conference on Data Engineering (ICDE), IEEE, 2019, pp. 1334–1345.
[34] A. Borji, A categorical archive of ChatGPT failures, 2023. https://arxiv.org/pdf/2302.03494.pdf.
[35] B. Guo, X. Zhang, Z. Wang, M. Jiang, J. Nie, Y. Ding, J. Yue, Y. Wu, How close is ChatGPT to human experts? Comparison corpus, evaluation, and detection, 2023. https://arxiv.org/pdf/2301.07597.pdf.
[36] J. Pu, Z. Sarwar, S.M. Abdullah, A. Rehman, Y. Kim, P. Bhattacharya, M. Javed, B. Viswanath, Deepfake text detection: Limitations and opportunities, 2022. https://arxiv.org/pdf/2210.09421.pdf.
[37] Y. Tian, H. Chen, X. Wang, Z. Bai, Q. Zhang, R. Li, C. Xu, Y. Wang, Multiscale positive-unlabeled detection of AI-generated texts, 2023. https://arxiv.org/pdf/2305.18149.pdf.
[38] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, Y. Bengio, Generative adversarial networks, Communications of the ACM 63 (11) (2020) 139–144.

Xiaodong Wu is currently a Ph.D. student in the Department of Electrical and Computer Engineering, Queen's University, Kingston Canada. He received his bachelor's degree in computer science at the University of Electronic Science and Technology of China and a master's degree in computer science at King Abdullah University of Science and Technology. His research interests include machine learning, adversarial attack, AI security and privacy issues, and trustworthy AI.

Ran Duan received bachelor's degree from Beijing University of Chemical Technology and University of Detroit Mercy in the United States in 2019, and master's degree from Queen's University of Canada in 2020, majoring in Electrical and Computer Engineering. During his academic journey, he has studied the use of deep learning methods for fake image generation and detection.



Jianbing Ni is currently an assistant professor with the Department of Electrical and Computer Engineering, Queen's University, Kingston, Canada. He received his Ph.D. degree in Electrical and Computer Engineering from University of Waterloo, Waterloo, Canada, in 2018. His research interests are applied cryptography and network security, with current focus on edge computing, artificial intelligence, Internet of Things, and blockchain technology.