The Security of Using Large Language Models: A Survey With Emphasis on ChatGPT

Wei Zhou, Xiaogang Zhu, *Member, IEEE*, Qing-Long Han, *Fellow, IEEE*, Lin Li, Xiao Chen,
Sheng Wen, *Senior Member, IEEE*, and Yang Xiang, *Fellow, IEEE*

***Abstract*—ChatGPT is a powerful artificial intelligence (AI) language model that has demonstrated significant improvements in various natural language processing (NLP) tasks. However, like any technology, it presents potential security risks that need to be carefully evaluated and addressed. In this survey, we provide an overview of the current state of research on security of using ChatGPT, with aspects of bias, disinformation, ethics, misuse, attacks and privacy. We review and discuss the literature on these topics and highlight open research questions and future directions. Through this survey, we aim to contribute to the academic discourse on AI security, enriching the understanding of potential risks and mitigations. We anticipate that this survey will be valuable for various stakeholders involved in AI development and usage, including AI researchers, developers, policy makers, and end-users.**

***Index Terms*—Artificial intelligence (AI), ChatGPT, large language models (LLMs), security.**

I. Introduction

ARTIFICIAL intelligence (AI) is no longer a futuristic concept. It is a present reality and has become ubiquitous in the daily life of people around the world. The leaps in machine learning (ML) technologies [1], and in particular, deep learning, have accelerated the AI revolution, rendering machines capable of performing complex tasks that were once the exclusive domain of humans [2].

Many of the powerful AI language models are equipped with a deep learning architecture known as the Transformer [3], which allows these models to generate human-like text. These models learn from a vast corpus of text data, enabling them to respond to prompts with sentences that maintain contextual relevance. The recent version, ChatGPT [4], is particularly impressive, even shocking experts by its ability to generate coherent and contextually appropriate responses [5]. This has resulted in widespread application of those models, such as drafting media [6]–[8], transmitting artistic requirements [9], [10], facilitating education [11]–[14], automation [15], [16], eHealth [17]–[21], finance [22]–[24], and tourism [25], and writing or debugging code [26]–[30], creating a new dimension of digital interaction.

A. Security Problems of LLMs

However, with great power comes great responsibility, the rise of AI has brought various security issues, including risks of data privacy breaches, potential for misuse in generating deceptive content, inherent biases from training data affecting output fairness, vulnerability to adversarial attacks, and issues with hallucinations leading to reliability concerns. In May 2023, ChatGPT experienced a data breach due to a vulnerability in its open-source library [31]. The breach exposed sensitive user information and raised concerns about the security and privacy of AI technologies. In another instance, ChatGPT accidentally leaked company secrets belonging to Samsung, leading to an internal ban on the tool [32]. These incidents highlighted the challenges faced by large language models (LLMs) in ensuring data protection and prompted tighter restrictions on AI use by businesses and countries. We have classified the security problems of LLMs into the following categories:

1) Bias: The bias in AI systems, often stemming from the diversity and nature of training data, can lead to outputs that perpetuate harmful stereotypes and disseminate misinformation. This not only challenges the task of ensuring fair and unbiased AI but also influences societal norms and individual behaviors negatively. The problem is augmented as these models are used in decision-making processes, from hiring practices to law enforcement, where biased outputs can have real-world consequences.

2) Disinformation: The potential for AI language models like ChatGPT to generate large volumes of plausible yet false content poses a significant threat in terms of disinformation. This capability could be exploited to influence public opinion on a massive scale, from swaying elections to inciting social unrest. The inherent biases in AI systems further complicate the issue, as they can slant generated content in subtle ways that might not be immediately recognizable, thereby covertly shaping narratives.

Manuscript received August 2, 2023; revised May 1, 2024, July 31, 2024, August 31, 2024; accepted October 11, 2024. Recommended by Associate Editor Nian Zhang. *(Corresponding author: Yang Xiang.)*

Citation: W. Zhou, X. Zhu, Q.-L. Han, L. Li, X. Chen, S. Wen, and Y. Xiang, "The security of using large language models: A survey with emphasis on ChatGPT," *IEEE/CAA J. Autom. Sinica*, vol. 12, no. 1, pp. 1–26, Jan. 2025.

W. Zhou, Q.-L. Han, S. Wen, and Y. Xiang are with the School of Science, Computing and Engineering Technologies, Swinburne University of Technology, Hawthorn VIC 3122, Australia (e-mail: weizhou@swin.edu.au; qhan@swin.edu.au; swen@swin.edu.au; yxiang@swin.edu.au).

X. Zhu is with the School of Computer and Mathematical Sciences, the University of Adelaide, Adelaide, South Australia 5005, Australia (e-mail: xiaogang.zhu@adelaide.edu.au).

L. Li is with the Faculty of Science and Engineering, Southern Cross University, Gold Coast QLD 4225, Australia (e-mail: lin.li@scu.edu.au).

X. Chen is with the School of Information and Physical Sciences, the University of Newcastle, Callaghan NSW 2308, Australia (e-mail: xiao.chen@newcastle.edu.au).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/JAS.2024.124983

TABLE I
SUMMARY OF RECENT SURVEY PAPERS ON LLMS ([36]–[46])

Reference	Year	Focus						
		Bias	Disinformation	Ethics	Misuse	Attacks	Privacy	Defence
[36]	2020	√						
[37]	2021	√						
[38]	2021					√	√	
[39]	2022		√					√
[40]	2022			√				
[44]	2022			√	√			√
[41]	2023			√	√			
[46]	2023			√	√		√	√
[45]	2023							
[42]	2023	√				√		
[43]	2023	√		√				
Our survey	2023	√	√	√	√	√	√	√

3) Ethics and Misuse: Ethical issues in AI technologies emerge prominently from their potential misuse. A primary ethical issue is the challenge to autonomy and consent, where AI systems may interact with individuals or make decisions affecting them without their explicit consent, thereby compromising personal autonomy. Additionally, their ability to produce contextually appropriate responses can be weaponized to conduct social engineering attacks, posing significant ethical and security concerns [33]. These aspects underscore the urgent need for ethical guidelines and robust mechanisms to prevent the misuse of AI technologies.

4) Attacks and Privacy: AI language models like ChatGPT are susceptible to a variety of cyberattacks that can jeopardize the integrity and confidentiality of their operations. For example, adversarial attacks involve subtly altering input data to trick the model into making errors or producing specific outputs, exploiting the model's sensitivity to subtle statistical features of the input. Privacy concerns are tied closely with attacks. AI language models can inadvertently leak sensitive information embedded in their training data, posing significant privacy risks. While these models do not inherently access or retrieve personal data, the data they generate can reflect the privacy levels of their training sets. This risk is particularly pronounced in systems used for personalized advertising and recommendation, where the inadvertent disclosure of personal preferences can occur [34]. In a recent advisory, the Salt Security research team identified three types of vulnerabilities within ChatGPT plugins [35]. Vulnerabilities were discovered within the plugin installation process itself, allowing attackers to install malicious plugins and potentially intercept user messages containing proprietary information.

B. Motivation

Given these concerns, there is a pressing need for a security-focused study on AI systems like ChatGPT. Such a study would contribute to the field in several ways, including adopting a proactive approach for security, understanding the evolving threat landscape, informing policy and regulation, educating AI users, and advancing the field of AI itself. Therefore, we are motivated to conduct a comprehensive survey about recent security issues raised in AI systems. The uniqueness of our survey resides in its precise and comprehensive exploration of security aspects of large language models (LLMs), with particular emphasis on ChatGPT. While there are other surveys [36]–[38] that investigate topics like bias and security, they cast a wide net over NLP systems and chatbots in general. Our survey, in contrast, narrows the scope to focus on LLMs and specifically highlights ChatGPT, allowing for a more nuanced discussion of its unique complexities and potential vulnerabilities.

Moreover, our survey extends beyond the focus of those [39]–[44] that primarily discuss theoretical aspects of risk, ethics, misuse, and mitigation strategies. We provide not only an understanding of these aspects in the broader context of LLMs but also delve into the practical implications, particularly for ChatGPT, making our survey both theoretically informed and practically applicable.

While the paper [45] gives a thorough survey of the evolution and impact of LLMs, it falls short in its comprehensive analysis of their security implications. Similarly, a discussion [46] on the safety of dialogue systems is offered but the deeper security concerns inherent in LLMs is missing. Our survey fills this crucial gap by analyzing the security aspects of LLMs, especially ChatGPT, from multiple perspectives including bias, disinformation, ethics, misuse, attacks, and privacy. Table I provides an overview of various survey papers that have been published on the topic of Large Language Models, categorizing them based on their focus areas. Through comparison, we believe our survey offers a unique contribution to the body of literature on the security considerations of LLMs.

C. Contributions

The main contributions of this survey are summarized as follows:

1) It provides an in-depth and comprehensive analysis of the security aspects of large language models, with a specific focus on ChatGPT. Unlike other surveys that broadly discuss

various NLP systems or AI technologies, this survey narrows its focus, thereby offering a detailed examination of the unique security challenges and concerns associated with ChatGPT.

2) It provides a holistic view of the security issues related to LLMs. It encompasses multiple dimensions of security, including bias, disinformation, ethics, misuse, attacks, and privacy. This broad perspective, combined with the focused examination of ChatGPT, allows for a nuanced understanding of security issues, making the survey valuable for both theoretical research and practical application.

3) It fills a crucial gap in the existing literature by offering a comprehensive analysis of the security implications of LLMs, which has been relatively unexplored in previous surveys. By focusing on these aspects, the survey contributes to the academic discourse surrounding the safe and ethical use of AI technologies, particularly LLMs like ChatGPT, and offers potential mitigation strategies for these issues.

D. Paper Organization

The rest of the survey is organized as follows. Preliminary concepts are introduced in Section II, including how LLMs like ChatGPT work and how they function with safety considerations. Section III classifies bias in LLMs into several categories and introduces the detection and mitigation research in each category respectively. Section IV summarizes the recent disinformation research, including fake media generation and detection, and fake document generation and detection. The recent ethics and misuse studies are presented in Section V, with a focus on the misuse of ChatGPT in different areas, such as email communication, and education. Section VI investigates different types of attacks to LLMs, and analyzes the date leakage problem and other risks. Section VII discusses the possible future research, followed by a conclusion of the whole paper in Section VIII.

II. Preliminary

Large language models like ChatGPT are sophisticated artificial intelligence systems that can generate human-like text based on given input. These models are trained on vast amounts of data from the Internet, learning to understand and predict patterns in language. In this section, we briefly introduce how LLMs like ChatGPT function, focusing on their training methodologies and architectural nuances. To mitigate or eliminate the risks, safety and ethical use are paramount in their deployment. Measures such as usage policies, access control, and user anonymity are implemented to ensure that the system is used responsibly and any risks are mitigated. Therefore, we will also introduce these general measurements.

A. Training of LLMs

Training large language models [4], [47] involves a two-step process of pre-training and fine-tuning. These models are also regularly tested for vulnerabilities through "red teaming".

1) Pre-Training: During the pre-training stage, LLMs learn to predict the next word in a sentence. They are trained on a large corpus of text from the Internet, but they do not know specifics about which documents were in their training set or have access to any specific documents or sources.

The models learn statistical patterns in the data they are trained on. For example, if the phrase "I'm feeling under the weather" often precedes "I have a cold", the model learns that the latter phrase is a likely continuation. It is important to note that the model does not understand these sentences or concepts in the same way humans do; rather, it identifies patterns of words and phrases and their statistical likelihood of appearing together.

2) Fine-Tuning: After pre-training, models undergo a fine-tuning process, where they are trained on a narrower dataset with human reviewers following specific guidelines provided by OpenAI or another overseeing entity. Reviewers review and rate possible model outputs for a range of example inputs. The model generalizes from this reviewer feedback to respond to a wide array of inputs from users.

Throughout the process, the reviewers maintain a strong feedback loop with the overseeing entity. They meet regularly to address questions and provide clarifications on the guidelines. This iterative feedback process helps the model improve over time.

3) Reviewing and Red Teaming: In addition to pre-training and fine-tuning, the models' performance and safety are continuously evaluated. This includes "red teaming", where an external group attempts to find vulnerabilities or undesirable behavior in the system, and further improvements are made based on their findings.

It is important to note that while this process can lead to high performing models, it also introduces certain limitations and risks, including the possibility of biases in the responses, and the inability of the model to provide reliable information outside of its training data cut-off.

The process is a complex one that balances the trade-offs between utility, safety, and ethical considerations in deploying artificial intelligence in the real world.

B. Response Generation

Large language models generate responses [4] based on patterns they have learned during training. They use a method called Transformer architecture, which is particularly suited to understanding the context of language.

A simplified breakdown of how they generate responses is listed as follows:

1) Tokenization: First, the input text is broken down into chunks called tokens. These tokens can be as short as one character or as long as one word (e.g., "a" or "apple").

2) Context Understanding: The LLM analyzes the tokens and their order to understand the context. It uses something called "attention mechanisms" to weigh the importance of each token in relation to the others.

3) Prediction: Based on the context and the patterns it has learned during training, the LLM predicts the most likely next token. This prediction is a statistical one, based on the probabilities assigned during training.

4) Generation: The predicted token is added to the end of the input string. This new string (original input + new token) is then fed back into the model, and the process repeats. The LLM continues predicting and adding tokens until it generates a stop signal (like a period at the end of a sentence), or until it reaches a specified maximum length.

5) Decoding: Finally, the stream of output tokens is decoded back into human-readable text.

It is worth noting that the LLM does not have beliefs, opinions, or feelings, and it does not have access to any personal data about individuals unless it has been shared in the course of the conversation. It generates responses based on patterns and statistical associations it has learned during training.

This method of response generation can sometimes lead to issues such as generating biased or offensive content, or unintentionally revealing private information that was input during the conversation. It is for this reason that organizations using LLMs implement safety and ethical guidelines, as well as mechanisms for users to report problematic outputs.

C. Safety Measures

In this subsection, we briefly introduce the safety measures [48] when deploying large language models.

1) Usage Policies: Usage policies, sometimes known as Terms of Service or Code of Conduct, serve as a contractual agreement between the AI service provider and the user. They specify what is considered acceptable use of the system. OpenAI's usage policies, for instance, prohibit the use of its AI systems for any harmful, illegal, or unethical activities. Any attempt to use the system for such activities would be considered a violation of these policies and could result in termination of service. Furthermore, the policies guide the interaction between the user and the system, providing a framework for how the system should be used and the repercussions if misused.

2) Access Control: Strict access controls are placed on the training data and the underlying model to prevent unauthorized access. This might involve role-based access control (RBAC) systems, where only individuals with certain roles can access particular resources. Access protocols can also include multi-factor authentication and strict password policies to further secure the system.

3) Monitoring and Auditing: Monitoring and auditing involve keeping track of system activity logs, including who accessed the system, when they accessed it, and what operations they performed. If any suspicious activities are detected, such as repeated failed login attempts or requests from unusual IP addresses, it can trigger alerts for further investigation. Regular audits of these logs can identify patterns of misuse or potential security vulnerabilities that might not be immediately apparent.

4) Input Filtering: Certain forms of inputs can be automatically blocked by the system to prevent misuse or harmful outputs. This could involve the use of automated content moderation tools to screen for and block harmful or inappropriate content. It could also involve using filters to block certain types of requests, such as those asking for personally identifiable information or those containing hate speech or offensive language.

5) User Anonymity: In order to maintain user anonymity, any inputs to the system are processed in a way that they cannot be directly linked back to the user. This might involve stripping out any personally identifiable information from inputs before they are processed and storing user data in a way that is separate from the rest of the system. User data may be pseudonymized or anonymized to ensure privacy.

6) Robust Cybersecurity Practices: OpenAI uses various cybersecurity practices to protect the system and the data it handles. These might include using encryption to protect data while it is being transmitted or stored, implementing firewalls and intrusion detection systems to guard against unauthorized access, and regularly patching and updating systems to protect against known vulnerabilities. Additionally, they might also include conducting regular security assessments and penetration testing to identify and address potential security weaknesses.

D. Framework of the Survey

In this section, we delve into the operational underpinnings of LLMs like ChatGPT. As these models learn by assimilating vast amounts of Internet-sourced data, they inherently risk embedding and later regenerating sensitive information, presenting substantial privacy concerns. Moreover, the statistical learning basis of these models predisposes them to manifesting biases present in their training datasets, which could lead to security issues when such biases result in discriminatory or harmful outputs. Additionally, the Transformer architecture, while enabling sophisticated contextual understanding, also increases susceptibility to adversarial attacks. These attacks exploit the model's dependency on input patterns to inject malicious content, thus manipulating outputs.

This survey aims to investigate the aforementioned security challenges posed by AI systems like ChatGPT. It will follow a systematic approach to literature review, focusing on peer-reviewed articles, white papers, and official reports from renowned institutions and organizations. It will employ rigorous criteria for selecting sources to ensure the reliability and relevance of the information. The study will also consider various stakeholder perspectives, including those of AI developers, users, policymakers, and cybersecurity experts, among others.

The main structure of this survey is depicted in Fig. 1. Except for the main sections outlined, we delve deeper into the limitations of current research and provide our insights into prospective avenues for future exploration in this field. Our aspiration is that this comprehensive review will provide new researchers with a rapid understanding of the current landscape, as well as stimulating and propelling further studies in the domain of large language model safety and ethics.

III. Bias

As an AI language model, ChatGPT has the potential to perpetuate and amplify biases that exist in the data or programming used to train it. Bias can manifest in various ways, such as in the choice of words, syntax, or tone used in generated text. There is also the potential for biases to emerge in the training data used to create and refine the AI model, such as under-representation of certain groups or over-representation of certain perspectives or worldviews.

Researchers have identified and evaluated various biases in

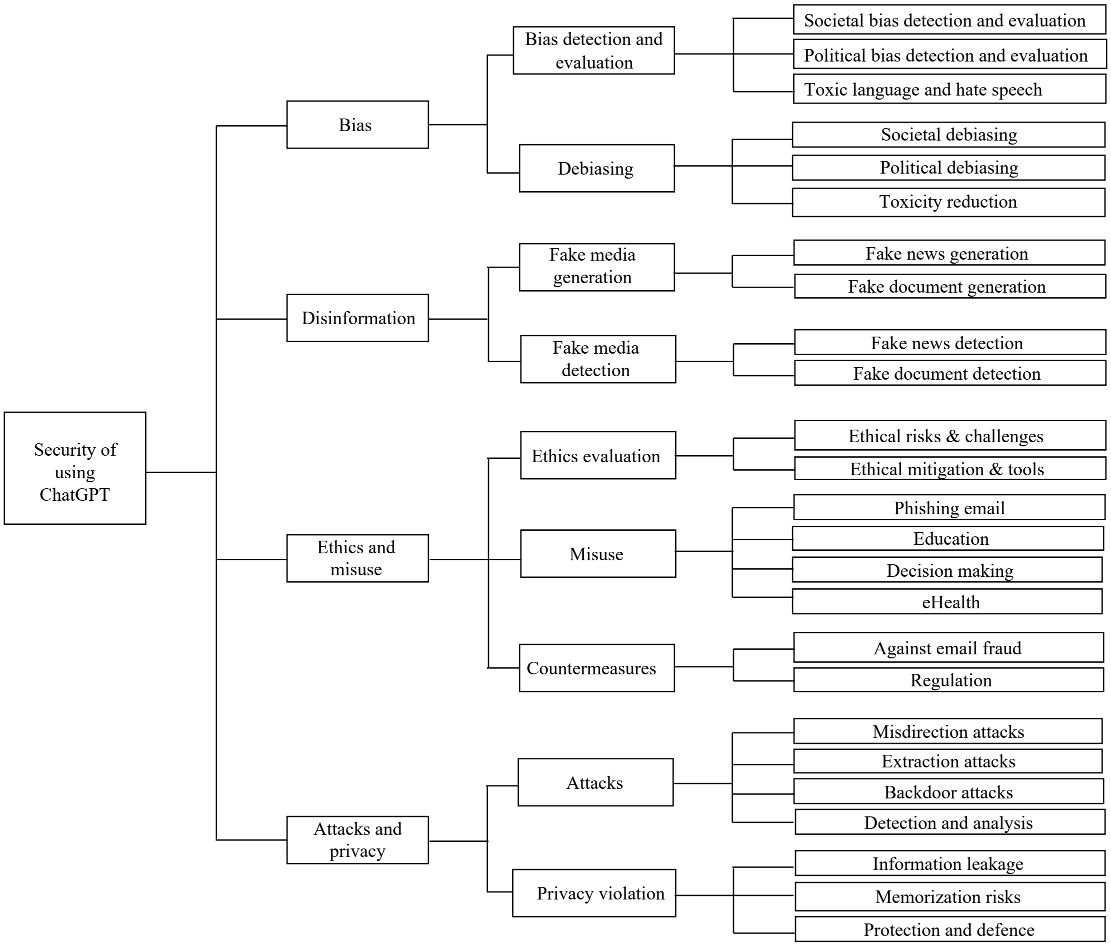


Fig. 1. Framework of the survey.

generative language models. There are also some suggestions or mitigation strategies proposed against bias. In this section, we first summarize the recent bias detection and evaluation studies, and then the research about reducing or eliminating bias, which is called debiasing. Table II presents a comprehensive overview of research efforts focused on detecting, evaluating, and debiasing biases in LLMs, with a focus on ChatGPT. In this table, we have included biases from models other than ChatGPT to provide a more comprehensive understanding of the bias issues that pervade large language models as a whole. It is important to note that the biases exhibited by ChatGPT are not unique to it but are, in fact, reflective of broader patterns observed across various LLMs, especially GPT-2 or GPT-3. These models are trained on large datasets that often contain inherent societal, cultural, and political biases. As a result, the biases manifested in their outputs are often similar, if not identical, to those observed in ChatGPT. It is structured to categorize the research into two main areas: Bias Detection & Evaluation, and Debiasing, and within these, further subdivided into specific bias types such as Societal, Political, and Toxicity.

A. Bias Detection and Evaluation

Bias can be classified into societal bias or political bias based on its nature or the domain in which it manifests [37]. Societal bias is when the model's behavior reflects societal prejudices, stereotypes, or systemic inequities that exist in the data it was trained on. This can be related to race, gender, religion, socioeconomic status, and so on. Political bias occurs when a model disproportionately represents or favors certain political ideologies, parties, or figures. Besides, there are also some general biases in the form of toxic content or hate speech.

1) Societal Bias Detection and Evaluation: Among societal biases, the most common one is gender bias, followed by intersectional bias.

i) Gender Bias: In assessing gender bias, a comprehensive framework has been put forth [49]. This framework provides a blueprint for examining gender biases and applies it to study the inherent biases in leading language models, including GPT-2 and Google's XLNet. A strong case has been made that fairness in language generation should rest on the principle of individual fairness, which marks a fresh take on the topic. Continuing the exploration of this research area, a deeper investigation of the representation and gender bias in narratives produced by GPT-3 has been undertaken [50]. The method involved the use of topic modeling and word similarity metrics to thoroughly inspect the gender stereotypes associated with feminine and masculine characters within the produced stories. Significant gender stereotypes were uncovered in these stories, with character portrayal and thematic focus displaying considerable variations between masculine and

TABLE II
SUMMARY OF BIAS RESEARCH ON LLMS ([49]–[85])

Research	Category	Reference	Year	Model	Detailed bias
Bias detection and evaluation	Societal	[49]	2020	GPT-2, XLNet	Gender
		[50]	2021	GPT-3	Gender
		[51]	2022	GPT-3	Gender
		[52]	2024	ChatGPT	Gender
		[53]	2021	ELMo, BERT, GPT-2	Intersectional
		[54]	2021	BERT, GPT2, RoBERTa, XLNet	Intersectional
		[55]	2021	GPT-2, GPT-3	Intersectional
		[56]	2024	ChatGPT	Intersectional
		[57]	2019	GPT-2	Occupation
		[59]	2024	ChatGPT	Occupation
		[58]	2024	ChatGPT	Occupation
		[60]	2021	GPT-3	Religion
		[62]	2022	GPT-3	Disability
	Political	[63]	2023	GPT-3	Political
		[66]	2023	ChatGPT	Political
		[67]	2023	ChatGPT	Political
		[68]	2023	ChatGPT	Political
		[64]	2024	Falcon, Flan-UL2, Llama-2, GPT-4	Political
		[65]	2024	BERT, GPT-3	Political
		[69]	2024	ChatGPT	Political
		[70]	2024	ChatGPT	Political
		[71]	2024	ChatGPT	Political
	Toxicity	[72]	2020	GPT-2	Toxic language
		[73]	2021	BERT, GPT-2	Toxic language
		[74]	2021	BERT, GPT-2	Hurtful sentence completion
		[75]	2022	RoBERTa, GPT-3	Moral norms and values
		[76]	2022	ToxiGAN, GPT-3	Implicit hate speech
		[77]	2023	ChatGPT	Implicit hate speech
Debiasing	Societal	[78]	2019	BERT, RoBERTa, GPT-2	Gender
		[79]	2021	GPT-2, GPT-3	Demographic
		[80]	2021	GPT-2	Demographic
		[81]	2022	GPT-3	Occupation
	Political	[82]	2021	GPT-2	Political
		[83]	2022	GPT-2	Political
	Toxicity	[84]	2021	GPT-2	Religion
		[85]	2021	GPT-2	Demographic

feminine characters. Furthering the discourse on gender bias, an emphasis on detecting the Brilliance Bias in generative language models has been brought forth [51]. In this regard, two variants of the GPT-3 model were the focus. The data sets, curated from stories generated by these models and prompted with “brilliance”-related adjectives and gender-specific characters, underwent analysis. The result was a revelation that these models might amplify Brilliance Bias, thus highlighting the societal implications of this issue within generative language models. A recent study [52] identifies significant disparities in the way ChatGPT responds to male versus female-related prompts, particularly in fields traditionally dominated by one gender. The analysis reveals how ChatGPT’s output can subtly reinforce existing gender stereotypes, underscoring the need for more refined models that address these biases.

ii) Intersectional Bias: A seminal piece of work [53] introduces the contextualized embedding association test (CEAT), a tool designed to measure the overall extent of bias in neural language models (NLMs). This approach uses a random-effects model to ensure more accurate measurements. Two innovative methods have been introduced: Intersectional bias detection (IBD) and emergent intersectional bias detection

(EIBD). Both methods aim to automatically identify and quantify intersectional biases, not just within static word embeddings but also in contextualized word embeddings. Experimental findings verify the presence of bias in all English corpus-trained models, particularly in the representations of intersectional group members where the strongest biased associations are observed. A continuation of the exploration of intersectional bias detection can be seen in the creation of StereoSet [54]. This robust natural English dataset is designed to measure stereotypical biases in pre-trained language models and covers four domains: gender, profession, race, and religion. Two distinct association tests, intrasentence and intersentence CATs, have been designed to critically evaluate both the language modeling capacity and the extent of stereotypical bias present in popular models such as BERT, GPT-2, ROBERTA, and XLNet. Q-Pain [55], a meticulously curated dataset, is geared towards measuring potential bias in medical question-answering systems, especially in the field of pain management. Alongside the dataset, a new framework is proposed to quantify potential biases in medical decision-making processes. Its utility has been demonstrated through the assessment of two key QA systems, GPT-2 and GPT-3. The findings highlight significant treatment disparities among intersectional race-gender subgroups, bringing to light the potential risks of AI in healthcare settings. These insights underscore the critical role of datasets like Q-Pain in ensuring safety and fairness before deploying AI applications in the medical field. A recent study [56] investigates racial bias in ChatGPT's recommendations for college majors. The study finds that minority groups, particularly LGBTQ+ and Hispanic students, are less likely to receive STEM-related major recommendations, highlighting significant intersectional disparities in the model's output.

There are also some studies on detecting bias about occupation, religion or disability in LLMs.

iii) Occupation Bias: A comprehensive study [57] delves into biases in natural language generation systems by scrutinizing text produced from prompts incorporating various demographic groups. The innovative concept of "regard" towards a demographic emerges as a novel metric for bias in NLG, backed by empirical evidence using a newly annotated dataset. Biases are analyzed in two contexts: one examining descriptive levels of respect for a demographic, and the other delving into the different occupations associated with a demographic. To analyze these biases, VADER (valence aware dictionary for sentiment reasoning) is utilized as the primary sentiment analyzer in comparison with the regard metric. Similarly, a recent study [58] dissects the biases in ChatGPT when recommending college majors, uncovering significant disparities in how the model treats different demographic groups. The study reveals that students from minority backgrounds are often recommended less prestigious or less lucrative fields, indicating a bias in the model's recommendations based on socioeconomic and racial factors. Another study [59] finds that ChatGPT reflects systemic biases embedded in its training data, which can lead to discriminatory outcomes in automated decision-making tasks like CV screening. The findings emphasize the need for careful scrutiny and mitigation strategies to prevent perpetuating societal stereotypes and discrimination through AI-driven tools.

iv) Religion Bias: An exploration into persistent anti-Muslim bias in GPT-3 is presented in a dedicated study [60]. Various probing methods reveal GPT-3's consistent and inventive display of its anti-Muslim bias across diverse uses of the model. This investigation also draws attention to a pervasive association of Muslims with violence in large language models, specifically GPT-3 [61]. This particular bias has been shown to be more severe compared to those related to other religious groups, prompting the quantification of the positive distraction required to counteract this bias.

v) Disability Bias: A critical investigation [62] addresses the degree of bias inherent in GPT-3-generated text from a disability perspective. Employing sentiment analysis and toxicity measurements as tools for assessing bias, the study generates text both with and without disability identity words, subsequently evaluating the toxicity score for each produced sentence. Preliminary findings suggest that GPT-3 holds a noticeable bias towards people who are Deaf or Blind when generating open-ended text.

2) Political Bias Detection and Evaluation: Political bias, defined as the systematic favoring or disfavoring of particular political viewpoints by an AI model, stems largely from the biases inherent in the model's training data. To detect such biases, researchers have employed a variety of NLP techniques, including sentiment analysis and lexical decision tasks, to assess the alignment of model outputs with political ideologies [63]–[65]. An early study [63] finds that GPT-3 exhibits a moderate left-leaning bias and tends to reflect the ideological bias of the input prompts. The study highlights that political bias in language models can replicate and possibly amplify existing societal biases, raising concerns about their use in applications where impartiality is critical. ChatGPT also presents with political bias [66], specifically in relation to decision-making. Experiments involving statements from voting advice applications and a political compass test uncover a pro-environmental and left-libertarian ideology in ChatGPT, which seems to support policies such as flight taxes, rent restrictions, and abortion legalization consistently across different languages and various prompt modifications. An additional research endeavor [67] delves into potential political biases in ChatGPT through 15 different political orientation tests. Results consistently diagnose ChatGPT as leaning towards left-leaning viewpoints. This outcome underscores the need for AI systems to strive for political neutrality and offer balanced arguments on normative questions. Recently, a critical examination [68] checks for potential bias against conservative politicians in ChatGPT, analyzing limericks the AI generates for different politicians. It is found that limericks for liberal politicians lean positive, while those for conservative politicians veer negative. This observation suggests a bias favoring liberals and opposing conservatives.

Two recent studies investigate the political orientation of various LLMs, not limited to GPT models. The first one [65] investigates how LLMs respond to politically charged queries across a spectrum of topics such as abortion and LGBTQ rights. The findings indicate that LLMs generally produce responses that align with liberal or left-leaning perspectives, even when the input data or prompts are adjusted to encourage neutrality or conservative responses. The paper proposes a framework for systematically assessing the political bias of LLMs and emphasizes the importance of awareness in crafting queries to avoid politicized responses. This second one [64] focuses on the subtle ways political bias can manifest in the responses of LLMs, not only through what is said but also how it is said. The study analyzes various LLMs, examining both the content and the style of the language to identify bias. It is found that LLMs, despite being trained on diverse datasets, still exhibit biases that reflect the inequalities or partisan slants of their training materials. A latest study [69] takes this a step further by examining the political bias of ChatGPT specifically in the context of policy recommendations. The findings reveal that ChatGPT exhibits a clear bias towards left-leaning policies, even when the input prompts are neutral. This suggests that the model's training data may have embedded political leanings that influence its output, raising questions about the objectivity of AI-driven decision-making tools. Another study [70] investigates political biases in ChatGPT across different languages. The research reveals that ChatGPT tends to favor liberal perspectives more strongly in certain languages, suggesting that the model's training data and linguistic nuances may influence its political orientation. These findings underscore the complexity of mitigating bias in multilingual AI systems. Further, an additional investigation [71] examines ChatGPT's self-perception and how it relates to its political biases. The study finds that the model often aligns its responses with the perceived political climate of its users, indicating an adaptive bias that reflects the assumed beliefs of the user base. This adaptability could lead to echo chambers where users are only exposed to viewpoints that align with their own, exacerbating political polarization.

3) Toxic Language and Hate Speech: Toxic language and hate speech, forms of harmful communication, can unintentionally perpetuate and amplify societal biases, prejudices, and discrimination when generated by large language models.

Multiple recent studies center around detecting or quantifying toxic language and hate speech in pre-trained language models. For example, an investigation [72] into the tendency of these LMs to generate toxic text poses challenges to their safe usage. The study introduces REALTOXICITYPROMPTS, a dataset featuring 100 000 natural sentence-level prompts from English web text, paired with toxicity scores from a routinely utilized toxicity classifier. The outcomes demonstrate that seemingly benign prompts can lead to toxic outputs from pre-trained LMs. Simultaneously, a distinct methodology [73] for quantifying toxic content in English, French, and Arabic LMs is put forward. This study uses logistic regression classifiers to assess potential harmful or stereotypical content generated by LMs towards specific social groups. This work identifies variances in toxicity output among different LMs, contingent on the patterns used, and presents a substantial dataset of structured patterns for evaluating toxic language classification within LMs. A unique approach [74] for evaluating hurtful sentence completion in LMs is proposed, introducing the HONEST score as a measure. The research includes results from experiments conducted on LMs trained in six languages, emphasizing the multilingual aspect of this issue. Beyond detection, a study [75] explores whether LMs embody human-like biases regarding moral norms and values. An innovative approach for bias extraction that does not require explicit moral training is introduced, contributing to discussions on AI development's risk-benefit balance. Advocating for dataset creation, a proposal [76] for TOXIGEN, a large-scale machine-generated dataset for detecting subtle and implicit hate speech, is put forth. A demonstration-based prompting framework is used to generate subtly toxic and benign text, thereby expanding the coverage of demographic groups and balancing toxic/benign statements for each group. A recent study [77] assesses the capabilities and limitations of ChatGPT regarding implicit hate speech explanation. Comparing ChatGPT's abilities to classify and provide natural language explanations for implicit hateful tweets to human-generated explanations illuminates its potential and restrictions in the sphere of implicit hate speech research.

B. Debiasing

To address the issue of bias, there are a number of steps that can be taken, such as using diverse and representative training data, regularly monitoring and evaluating the model's outputs for potential biases, and implementing bias detection and correction tools. Additionally, it is important to engage diverse groups of individuals in the development and testing of AI models, to ensure that they are inclusive and representative of diverse perspectives.

1) Societal Debiasing: An innovative framework designed to reduce sentiment bias in the texts generated by language models is presented [78]. This work assesses the sentiment bias related to various sensitive attributes and proposes a regularization method that relies on embedding and sentiment prediction to diminish this bias. The deployment of this framework on Wikipedia and news corpora has shown encouraging results. Enhancing this, a generalized framework to identify and control societal biases in natural language generation models is put forth [79]. This model-agnostic methodology effectively manages biases in generated texts, particularly when input prompts specify certain demographic groups. An additional deep dive into the origins of representational biases in language models introduces novel benchmarks and metrics for their measurement [80]. A productive method known as AUTOREGRESSIVE INLP (A-INLP) is unveiled, which alleviates social biases in text generation while preserving crucial contextual information.

Concerning occupation debiasing, a study [81] embarks on the task of creating unbiased, realistic job advertisements using GPT-3. A comparative examination of actual job ads and GPT-3 produced ads highlights the efficacy of fine-tuning in enhancing realism and minimizing bias.

2) Political Debiasing: In relation to political debiasing, a reinforcement learning framework designed to counter political biases in large-scale language models is introduced [82] [83]. The reinforcement learning (RL) framework [82] can mitigate bias without needing access to the original training data or retraining the model. It operates through rewards derived from word embeddings or classifiers to guide the generation process towards neutrality. Empirical tests were conducted on attributes sensitive to political bias (gender, location, and topic), showing that the RL approach effectively reduces bias while maintaining text readability and coherence. The research presents a significant step toward creating fairer AI systems by adjusting model outputs post-training. The second study [83] from the same research team explores the political bias of the GPT-2 model and proposes methods for its measurement and mitigation. The authors demonstrate that the model's outputs tend to lean liberal, especially when prompted with sensitive attributes. They propose a reinforcement learning framework similar to the first one, which uses word embeddings and classifier-based rewards to guide the model towards generating politically unbiased text. The effectiveness of this approach is validated through both automated metrics and human evaluations, confirming that it reduces bias without compromising the quality of the generated text.

3) Toxicity Reduction: On the subject of toxicity reduction, a key discussion surrounding the evaluation and alleviation of toxic language generated by large language models is offered [84]. Several mitigation strategies are assessed in relation to both automatic and human evaluation, and the implications of toxicity mitigation in terms of model bias and LM quality are analyzed. The research indicates that basic intervention strategies can effectively optimize previously established automatic metrics. However, these gains come with a trade-off: reduced LM coverage for both texts about and dialects of marginalized groups. It is also observed that human raters often disagree with high automatic toxicity scores after strong toxicity reduction interventions, further emphasizing the complexities involved in the careful evaluation of LM toxicity. A separate study [85] evaluates the effect of detoxification techniques on the generation quality of language models for language used by marginalized groups. This work measures LM perplexity and generation quality of multiple detoxification techniques when conditioned on African-American English and minority identity mentions. The findings indicate that detoxification techniques negatively impact equity and diminish the utility of LMs on language used by marginalized groups, a result of spurious correlations in toxicity datasets.

C. Summary

The inherent biases can be present in AI language models like ChatGPT, including societal biases that reflect certain preferences and prejudices, political biases that may favor one viewpoint over another, and the potential for these models to generate toxic or hateful content. These biases, according to referenced studies, can unintentionally arise from the training data, often human-generated text on the Internet that contains these biases. On a positive note, there are various strategies for reducing these biases and toxicity. This includes research aimed at improving the models through better understanding and mitigating the introduction of biases during the training process, the use of more diverse and representative datasets, and the implementation of enhanced moderation and filtration techniques. The ultimate goal, as implied by this section, is to create AI models that are more respectful, unbiased, and useful to a broad range of users.

IV. Disinformation

ChatGPT's ability to generate coherent and convincing text could be used to spread disinformation or propaganda. For example, the model could be trained to generate fake news articles or biased opinions, leading to misinformation and social unrest. Researchers have investigated ChatGPT's probability of generating fake information and proposed various techniques such as fact-checking and source verification to combat disinformation and ensure that ChatGPT is used for beneficial purposes.

In order to contribute to the academic understanding of the multi-pronged issue of disinformation, this section offers a structured overview of recent research developments as shown in Table III. Specifically, it classifies the body of work into two significant and interconnected areas of explorations: fake media generation and fake media detection. Each of these branches provides unique insights into the methodologies employed for both the creation and identification of deceptive content.

A. Fake Media Generation

The realm of fake media generation is typified by any media content that has been manipulated or fabricated with the intention of spreading false information or narratives. In the modern context, such media can take myriad forms, from the intricate and AI-enabled "deepfake" videos, to doctored images, misleading headlines, and counterfeit news articles.

1) Fake News Generation: In the realm of fake news generation, "The Rumour Mill", an interactive tabletop machine, is introduced [86]. This machine aims to make the process of creating credible text tangible by interacting with different physical controls and to expose the spread of rumors and automatically generated misinformation. The paper posits that automatically generated texts are becoming increasingly difficult to identify and easy to create at scale, posing a technological and social threat to society. Through interaction with the Rumour Mill, people gain first-hand experience of AI-generated rumors, thereby raising awareness of the ease of generating and believing in such misinformation. Examining the potential impact of AI-generated text as a media misinformation tool, a study [87] carries out three distinct experiments to gauge the public's perception of AI-generated text, the influence of partisanship on perceived credibility, and the distribution of credibility across different AI model sizes. The research concludes that people struggle to differentiate between AI- and human-generated text, and exposure to AI-generated text has little effect on people's policy views. These insights hold significant implications for understanding the role of AI in online misinformation campaigns. A recent study [88] evaluates GPT-3's ability to produce accurate information and disinformation in tweet form and compares its credibility to human-generated information. The results reveal that while GPT-3 can generate accurate information that is easy to understand, it can also create more persuasive disinformation compared to humans. The research sheds light on the dangers

TABLE III
SUMMARY OF DISINFORMATION RESEARCH ON LLMS ([39], [86]–[104])

Research	Category	Reference	Year	Model	Methodology
Generation	Fake news	[86]	2020	GPT-2	Generating rumors
		[87]	2022	GPT-2	Analyzing public perception of AI-generated news
		[88]	2023	GPT-3	Evaluating GPT-3's information accuracy
		[89]	2024	ChatGPT	Potential of ChatGPT in generating and spreading fake news
	Fake document	[90]	2021	GPT-2	Generating fake CTI descriptions
		[91]	2022	GPT-2	Deepfake-generated social personas
		[92]	2022	GPT-2	Fake document infilling
		[93]	2024	ChatGPT	LLMs in multimedia disinformation
Detection	Fake news	[94]	2019	GPT-2	Grover - detecting neural fake news
		[95]	2022	GPT-2	ML-based disinformation detection
		[96]	2022	StyleGAN, GPT-3	Distinguishing real and AI-generated content
		[97]	2022	BERT, GPT-2	Framework for fake news detection
		[98]	2022	BERT, RoBERTa, GPT-2, GPT-3	Performance of detection systems
		[99]	2022	GPT-2, GPT-3	Detecting machine-written tweets
		[100]	2024	GPT-3	Challenges in disinformation detection
		[101]	2024	ChatGPT	Generating, explaining, and detecting fake news
		[102]	2024	ChatGPT	LLMs as tools for disinformation
	Fake document	[39]	2022	BERT, RoBERTa, GPT-2, GPT-3	DFTFooler - adversarial sample crafting
		[103]	2023	RoBERTa, GPT-J-6B, GPT-2	Detecting fake text reviews
		[104]	2023	GPT-2, GPT-3	Detecting fake restaurant reviews

of AI for disinformation and proposes ways to enhance information campaigns to benefit global health. Another study [89] explores the potential of ChatGPT in generating and disseminating fake news, proposing optimization paths to mitigate these risks. The findings underscore the ease with which ChatGPT can produce persuasive misinformation, raising concerns about its misuse in spreading disinformation.

2) Fake Document Generation: Regarding fake document generation, a methodology for generating deceptive cyber threat intelligence (CTI) text descriptions using transformers is presented [90]. This research demonstrates the potential for adversaries to use these false CTI examples as training inputs to undermine cyber defense systems, coercing these systems to learn incorrect inputs that serve malicious purposes. The methodology, including a human evaluation, is assessed, revealing a possible poisoning pipeline for infiltrating a cybersecurity knowledge graph (CKG) and a cybersecurity corpus. A user study is conducted [91] to investigate user perception of deepfake-generated social personas and the resulting impact on user trust and decision-making when accepting or rejecting connection requests. The study incorporates controlled variables to comprehend the influence of prompting/training on user perceptions and explores user strategies for profile evaluation. Findings indicate a general vulnerability to deception from deepfake profiles, with only a minor decrease in trust and acceptance rates under conditions leading to the lowest average trust and acceptance rates. A groundbreaking context-aware model, fake document infilling (FDI), is proposed [92] that turns fake document generation into a controllable mask-then-infill procedure. This process masks important concepts of varying lengths in the document and replaces them with realistic yet misleading alternatives, informed by a pre-trained language model. The model, tested on technical documents and news stories, has been shown to surpass existing baselines. A recent study [93] examines the potential of ChatGPT in generating multimedia disinformation, including fake documents and doctored images. The research highlights the risks associated with LLMs' capabilities to create convincing fake media that can easily deceive users.

B. Fake Media Detection

Having examined the generation of fake media, it is now crucial to explore the opposite side of the same coin: the detection of such media. This subsection elucidates various strategies, techniques, and models that have been developed to counter the disinformation epidemic.

1) Fake News Detection: For fake news detection, a model named Grover is introduced [94], which generates realistic-looking news articles while also probing the potential threat of neural fake news. The study scrutinizes the risks and vulnerabilities introduced by neural disinformation and delves into the capacity of deep pre-trained language models to differentiate between authentic and machine-generated text. Results demonstrate that Grover proves to be an effective technique for both detecting and generating neural fake news. With a focus on detecting computer-generated disinformation, a research paper [95] assesses the performance of several promising machine learning-based detection models proposed in existing literature. These evaluations utilize a diverse range of datasets and encompass several types of texts, including news articles, product reviews, forum posts, and tweets. This work underscores the potential misuse of language models and the weaponization of such methods by state actors and state-sponsored groups. In a related study, an examination [96] is conducted into the human ability to differentiate between real and deep learning-generated social media profiles and posts. This paper details the outcome of an experiment where participants had to distinguish between real and generated profiles and posts in a simulated social media feed. The findings indicate that even entirely fabricated profiles and posts, crafted by a sophisticated text generator, present a significant challenge for human identification. A comprehensive review [97] of diverse strategies and models utilized for fake news detection, including traditional machine learning and deep learning models, is provided. This review also sheds light on the limitations and challenges of these techniques. Moreover, a framework is proposed that considers the veracity of text and differentiates between human-authored and machine-generated text, thereby offering an innovative and pragmatic solution to the fake news problem. A novel COVID-19 based synthetically generated dataset is introduced and the vulnerability of GPT-2 and Grover models [94] to adversarial attacks is explored. Various threat scenarios of neural fake news generated by state-of-the-art language models are examined, and the performance of generated text detection systems under these scenarios are assessed [98]. The study pinpoints the minimax strategy for the detector that minimizes its worst-case performance and establishes a set of best practices for practitioners. The paper concludes by arguing in favor of releasing robust detectors alongside new generators. Performance of state-of-the-art deepfake social media text detectors is investigated in recognizing GPT-2 generated tweets as machine-written, with efforts to enhance the state-of-the-art by fine-tuning hyperparameters and ensembling the most promising detectors [99]. This paper also focuses on studying the detectors' capabilities to generalize over tweets generated by GPT-3. A recent study [100] delves into the evolving challenges of disinformation detection in the age of LLMs. The study highlights the difficulty of identifying disinformation generated by advanced models like GPT-3 and explores potential solutions to enhance detection accuracy. Another paper [101] evaluates ChatGPT's capabilities in generating, explaining, and detecting fake news. The research introduces a reason-aware prompt method that significantly improves detection performance, especially in complex cases where traditional methods struggle. In addition, a study [102] examines the dual role of LLMs as both tools for spreading and detecting disinformation. The research emphasizes the importance of developing robust detection mechanisms to counter the sophisticated disinformation generated by LLMs.

2) Fake Document Detection: Concerning fake document detection, an evaluation [39] is conducted of the robustness and generalization capability of existing defenses on real-world synthetic datasets collected from various online platforms. Low-cost adversarial attacks are proposed and the resilience of the defenses to these attacks are evaluated. A novel black-box adversarial sample crafting strategy, DFT-Fooler, is introduced to probe the defenses' transferability against multiple defenses. The analysis suggests that leveraging semantic information in text content may enhance the robustness and generalization performance of deepfake text detection schemes. Later, a novel method [103] for detecting fabricated text reviews in collaborative filtering recommender systems utilizing user demographic characteristics is proposed. The study addresses both types of attacks, those generated by language models and those penned by dishonest users for monetary gain. Two datasets for fraud detection and testing are also presented, making the proposed approach potentially valuable for improving recommendation systems and enhancing their trustworthiness. Recently, a detection approach [104] for machine-generated fake restaurant reviews using GPT model and high-quality elite restaurant reviews verified by Yelp is introduced. Testing of this method on 24,000 reviews shows the fine-tuned GPT output detector significantly outperforms existing solutions. In addition, the research identifies patterns across multiple dimensions in non-elite reviews characteristics, including user and restaurant characteristics and writing style.

C. Summary

This section delves into the dual aspects of disinformation related to LLMs: the generation of fake media and its detection. Researchers illustrate the use of AI, particularly models like ChatGPT, for creating convincing fake news, doctored documents, and manipulated media content that can lead to widespread disinformation. The ease of creation and the challenges in discerning such content from human-generated text have been highlighted as significant societal threats. Conversely, the detection of such AI-generated fake media is explored through various strategies and techniques, including machine learning and deep learning models. The research underscores the difficulty humans face in identifying fake AI-generated content and underscores the need for robust, innovative detection systems that can effectively distinguish between human and AI-generated text. The interplay between these two aspects emphasizes the urgent necessity of balanced development and deployment of AI technologies, where advances in generation capabilities must be complemented by equally proficient detection methods.

V. Ethical Concerns and Misuse

The capacity of ChatGPT to generate human-like responses engenders ethical questions relating to the potential exploitation of this technology. Instances of misuse could involve employing ChatGPT for activities like spamming, phishing, or social engineering attacks. Perpetrators might utilize the model to generate convincing phishing emails or fraudulent messages, potentially leading to financial losses or data breaches.

In an academic context, students could exploit ChatGPT to fabricate essays or homework assignments, bypassing the necessary learning process and comprehension of the content. This behaviour not only jeopardizes the educational system, but also fosters academic dishonesty.

To counteract these risks, various techniques like adversarial training and attack detection methods have been proposed by researchers to enhance the robustness and security of ChatGPT. In this section, we explore recent research developments in these areas, starting with a discussion on the ethical evaluation, and then we summarize the misuse applications in email, education, decision-making, and eHealth. At last, we review the countermeasures against misuse.

A. Ethics Evaluation

The ethics of LLMs has been a subject of scrutiny and debate in recent years. Researchers have examined various aspects of AI ethics, ranging from ethical dilemmas and complications to strategies for mitigating ethical issues. We delve into these topics in the following subsections.

1) Ethical Risks & Challenges: There has been a considerable amount of research pinpointing the ethical dilemmas and complications that can occur when using AI technologies like ChatGPT. For example, a study [105] addresses the ethical challenges linked to data-driven dialogue systems, pushing for strong, safe systems that take into account a multitude of ethical aspects. The study stresses the necessity to reflect upon ethical concerns, especially those tied to natural language systems, as such systems gain increased prevalence and trust. Regarding the weaponization risk, an evaluation [106] by the Center on Terrorism, Extremism, and Counterterrorism (CTEC) of the GPT-3 API's potential misuse by extremists is described. This evaluation used prompts taken from right-wing extremist narratives to measure ideological consistency, accuracy, credibility, and the potential to aid in online radicalization into violent extremism. Results show that GPT-3 surpasses its predecessor, GPT-2, in creating extremist texts, proving its capacity to craft influential content that might radicalize individuals towards violent far-right extremist ideologies and behaviors. The lack of regulation poses a significant threat for mass online radicalization and recruitment, despite OpenAI's preventive measures. The need for AI stakeholders, policymakers, and governments to establish social norms, public policies, and educational initiatives to counteract machine-generated disinformation and propaganda is underscored, necessitating collaboration across industry, government, and civil society. In a thorough analysis [40], the ethical and social risks linked to large-scale language models are examined, aiming to structure the risk landscape for responsible innovation. This research underlines the necessity for risk mitigation strategies ranging from policy interventions to technical solutions and product design decisions. The paper concludes by discussing the organizational responsibilities in executing such mitigations and the role of collaboration. In a comprehensive study [107] using qualitative research methods, the ethical implications and potential dangers posed by large language models, with a particular focus on ChatGPT, are investigated. Existing benchmarking frameworks are critically evaluated to assess the ethical considerations related to large language models, while advocating for new benchmarks that capture all ethical implications. The research determines that large language models pose significant ethical concerns, and their applications necessitate meticulous consideration. A detailed analysis [108] of ChatGPT's failures in different categories is presented, highlighting the limitations and risks of large language models and chatbots. The paper contributes to the ongoing discourse on the strengths and weaknesses of ChatGPT and similar models, providing a crucial reference point for evaluating progress and improving the technology.

2) Ethical Mitigation & Tools: Despite the risks, strategies and tools are being developed to mitigate these ethical concerns. For instance, a new tool [109] named tool for ethical assessment of language generation models (TEAL) is introduced, which democratizes and standardizes ethical assessment of natural language generation models. The main contributions of TEAL are outlined, including its user-friendliness, built-in features and datasets, and ability to perform language appropriateness and fairness assessments of LGMs. A critical analysis [110] of the debates surrounding the ethical implications of GPT-3 is also presented, arguing for a contextual approach to AI ethics and GPT-3, which emphasizes human autonomy, societal harms and benefits, and human values. The paper discusses InstructGPT, which was designed to address the toxicity of GPT-3 but does not sufficiently address concerns over manipulation and bias.

B. Misuse

The potential misuse [111] and exploitation of LLMs is another critical area of research. This includes various domains such as email communication, education, decision-making, and eHealth as shown in Table IV. We begin with a discussion on potential misuse in email communication in the next subsection.

1) Phishing Email: Email phishing is one way that AI technologies can be exploited. This involves using LLMs to generate convincing phishing emails that can lead to financial losses or data breaches. An exploration [112] into the feasibility and effectiveness of natural language models in generating phishing emails is provided. A framework for evaluating the performance of NLMs in producing phishing emails based on various metrics is proposed and results are compared with those of a baseline model. The study underlines the potential impact of NLMs on phishing attacks and the need for continued research on the ethical and security implications of using NLMs for malicious purposes. A thorough study [113] on the role of attackers in spear-phishing attacks is presented, proposing a non-cooperative zero-sum game model to analyze the attack-defense process. The GPT-2 model is utilized to generate emails with a variety of harmful content, which adversaries use to deceive email security systems. The Nash equilibrium for the attacker-defender game is calculated, and a sensible scheme is offered for the attacker to gain an advantage over the target. A recent study [114] demonstrates how ChatGPT can be exploited to create smishing campaigns, highlighting the potential for generative AI to be used in sophisticated phishing attacks. The study explores different prompt engineering techniques to generate persuasive phishing messages that can bypass traditional email filters. Another study [115] explores the broader impact of ChatGPT on cybercrime and cybersecurity, including its role in generating email-based attacks. The study emphasizes the need for

TABLE IV
SUMMARY OF MISUSE RESEARCH ON LLMS ([112]–[140])

Research category	Reference	Year	Model	Methodology
Email	[112]	2022	GPT-2, GPT-3	Investigate feasibility and effectiveness of NLMs in generating phishing emails
	[113]	2021	GPT-2	Propose a game model to analyze the attack-defense process in spear-phishing
	[114]	2024	ChatGPT	Demonstrates how ChatGPT can be exploited to create smishing campaigns
	[115]	2024	ChatGPT	Explores the impact of ChatGPT on cybercrime and cybersecurity, including email-based attacks
	[116]	2024	ChatGPT	Investigates the dark side of ChatGPT, including its role in email-based cyberattacks
	[117]	2024	ChatGPT	Examines malicious content generation for email attacks through prompt engineering
Education	[118]	2022	ChatGPT	Potential threat to integrity of online exams
	[119]	2022	ChatGPT	Generate convincing medical research abstracts
	[120]	2022	GPT-3	Technological innovations in digital storage technologies
	[121]	2023	ChatGPT	Use of ChatGPT for software testing education
	[122]	2023	ChatGPT	Implications of ChatGPT for legal writing
	[123]	2023	ChatGPT	Use of AI language models in higher education
	[124]	2023	ChatGPT	Threat to academic integrity
	[125]	2023	ChatGPT	Generate academic essays without being caught by plagiarism detection tools
	[132]	2023	ChatGPT	Assist with the writing of finance research studies
	[133]	2023	ChatGPT	Implications for education and scientific writing
	[134]	2023	ChatGPT	Threat to the credibility of academic discourse
	[135]	2023	ChatGPT	Potential uses and limitations of generative AI models
	[126]	2024	ChatGPT	Explores the uses and misuses of ChatGPT in academic writing
	[127]	2024	ChatGPT	Investigates students' misuse of ChatGPT in higher education
	[128]	2024	ChatGPT	Discusses preventive strategies against misuse of ChatGPT in education
	[129]	2024	ChatGPT	Surveys student attitudes towards the use and misuse of ChatGPT in CS education
	[130]	2024	ChatGPT	Discusses applications, concerns, and recommendations for ChatGPT in education
	[131]	2024	ChatGPT	Discusses risks of ChatGPT in scientific publishing, including authorship and predatory publishing
Decision-making	[136]	2022	GPT-3	Use of large language models for ethical decision-making
	[137]	2023	ChatGPT	Influence of ChatGPT on users' moral judgment in making consequential decisions
	[138]	2023	GPT-3	Influence of AI language technologies on user opinions
eHealth	[139]	2023	ChatGPT	Competence in medical tasks
	[140]	2023	ChatGPT	Ethical implications in healthcare

improved detection methods to counter the misuse of ChatGPT in phishing schemes. Further research [116] investigates the dark side of ChatGPT, focusing on its role in email-based cyberattacks. The study examines how ChatGPT can be used to automate the creation of phishing emails that are difficult to distinguish from legitimate communications. A related study [117] examines the use of ChatGPT for malicious content generation through prompt engineering, particularly in the context of email phishing. The research discusses the implications of using AI-generated content to craft highly targeted and effective phishing campaigns.

2) Education: Emerging concerns regarding the misuse of ChatGPT in the educational sector have come to the forefront. The potential threat that ChatGPT poses to the integrity of online exams in tertiary education settings is detailed in a study [118], which evaluates the capability of ChatGPT to showcase critical thinking skills and produce highly realistic text with little input. It is suggested that ChatGPT could potentially be used for academic misconduct in online exams. It is stressed that educators and institutions should be cognizant of the risk of ChatGPT being used dishonestly and should explore measures to counteract it to ensure the fairness and validity of online exams for all students. An evaluation [119] of the ability of large language models, particularly ChatGPT, to produce convincing medical research abstracts is presented. Comparison of abstracts generated by ChatGPT with original abstracts from high-impact medical journals is done using an artificial intelligence output detector, a plagiarism detector, and blinded human reviewers. The results reveal that ChatGPT-generated abstracts were clearly written but had formatting issues, were often flagged by the AI output detector, and could be identified by discerning human reviewers. It is recommended that AI output detectors should be incorporated in the abstract evaluation process for journals and medical conferences, and transparent disclosure of the use of these technologies should be made. In the context of digital publishing and scholarship, the ethical implications of automated digital content generation are debated [120]. It is proposed that technological advancements in digital storage technologies, such as the MetaScribe system, can enhance the mechanisms of attribution and render the processes more transparent and accountable.

The potential misuse of ChatGPT in education, with particular emphasis on academic dishonesty and plagiarism, is underscored in numerous papers [121]–[125]. An extensive empirical study [121] is conducted that tasks ChatGPT with answering questions from five chapters of a popular software testing textbook and examines the correctness of its answers and explanations in different prompting settings. The implications of ChatGPT for legal writing are examined, with a focus on its strengths and limitations. Ethical hazards of relying on ChatGPT are discussed and insight on how legal writing professors and law students might use ChatGPT as a tool for certain tasks is offered [123]. A comprehensive overview [123] of the opportunities and challenges of using ChatGPT and other AI language models in higher education is provided, with a discussion on potential benefits such as increased student engagement and collaboration, as well as potential applications of ChatGPT in personalized and interactive assessments. The potential threat to academic integrity posed by the use of ChatGPT is highlighted and a call for clear guidelines and protocols to maintain academic integrity is made [124]. The potential of ChatGPT to generate academic essays undetected by plagiarism tools is explored [125] by generating 50 essays on various topics and then checking for plagiarism using two popular plagiarism detection tools. Some latest studies also explore the misuses of ChatGPT in academic writing [126]–[130] and scientific publishing [131], highlighting the importance of awareness and preventive measures in educational institutions.

The potentially harmful impacts of ChatGPT on scientific and finance research writing are pointed out in multiple studies [132]–[134]. The potential of ChatGPT in assisting with the writing of finance research studies is explored [132], and outputs generated by ChatGPT at four stages of the research process, namely, idea generation, literature review, data identification and processing, and empirical testing, are compared. The growing popularity of ChatGPT is discussed, arguing that it can also have serious implications for education and scientific writing if not used properly [133]. A brief opinion [134] on the growing concern around the use of AI-powered chatbots for producing academic writing is provided, with an argument that this technology may pose a threat to the credibility of academic discourse and a suggestion for greater discussion and ethical guidelines to tackle this issue.

Most recently, a structured investigation [135] of ChatGPT's ability to provide code, explain concepts, and create knowledge related to statistical process control (SPC) is carried out, identifying the benefits and limitations of the current version of ChatGPT for structured and nuanced tasks.

3) Decision Making: As ChatGPT is increasingly utilized for decision-making support, ethical implications surface. According to a critical evaluation [136] of the use of large language models for ethical decision-making, a new prompting strategy called similarity prompting (SimPrompting) leads to supposedly "super-human" results on the ETHICS dataset. However, this does not necessarily imply human-like understanding or reasoning. The errors of language models differ systematically from human errors, making it easy to craft adversarial examples, with signs of inverse scaling with model size on some examples. Moreover, prompting models to "explain their reasoning" often leads to alarming justifications of unethical actions. In another study [137], the moral authority of ChatGPT and its influence on users' moral judgment in making consequential decisions were examined. The experiment revealed that ChatGPT's advice is highly inconsistent, yet still influences users' moral judgment, even when they are aware that they are advised by a chatbot, and users tend to underestimate the extent of this influence. Furthermore, the effects of language-model-powered writing assistants on users' opinions were investigated [138]. The study found that using an opinionated language model affected the opinions expressed in participants' writing and shifted their opinions in the subsequent attitude survey. The authors argue that the opinions built into AI language technologies need to be monitored and engineered more carefully.

4) eHealth: Concerns about ChatGPT in the eHealth domain are articulated in publications. The Lancet Digital Health provides an overview [139] of ChatGPT's competence in passing medical licensing exams and generating patient discharge summaries and radiology reports. However, ethical concerns regarding potential errors, bias, and implications for the scientific record and scholarly publishing are highlighted. The article concludes that more forethought, oversight, and investment in robust AI output detectors are needed before widespread adoption of ChatGPT is considered. Ethical implications arising from the development of generative AI technologies in the healthcare industry are discussed in another publication [140]. Concerns are raised about authorship and the integrity of submitted work. The possibility of generative AI replacing ethicists in the future is explored. The paper suggests that generative AI may have positive effects, such as facilitating writing, aiding in the drafting of articles, and opening up authorship for individuals who struggle to express themselves in English.

C. Countermeasures

1) Against Email Fraud: Efforts to counter the misuse of ChatGPT include the proposal of an expandable scam-baiting mailserver [141]. This mailserver can conduct scam-baiting activities automatically. The study compares different approaches to automated scam-baiting and finds that human-designed lures work best at attracting scammer responses, while text generation methods informed by human scam-baiters are more effective at prolonging conversations. The authors release their code and conversation transcripts to guide the development of new countermeasures against scammers. Another countermeasure [142] discussed is the development of an automated victim for advance fee fraud (AFF) using the GPT-3 language model and engineered prompts. This system generates plausible responses to AFF emails, which can aid in disrupting fraud and obtaining actionable information on perpetrators.

2) Regulation: There is a growing recognition [143] of the need for robust regulatory mechanisms for AI systems, including large generative AI models (LGAIMs). The analysis proposes novel strategies for regulating LGAIMs, considering different actors in the value chain, such as developers, deployers, and users. The paper addresses various aspects of regulation, including direct regulation, data protection, content moderation, and policy proposals. Furthermore, the importance of responsibility in developing and regulating AI systems is emphasized [144]. The focus is on regulating profit-maximizing corporations rather than solely developing ethical principles. Additionally, the potential risks associated with using ChatGPT as a source of safety-related advice are discussed [145]. The study highlights the need for expert verification, ethical considerations, and safeguards to ensure users understand the limitations and receive appropriate advice when using ChatGPT for safety-related issues.

D. Summary

This section reveals several ethical dilemmas and challenges, such as data-driven dialogue systems' ethical concerns, potential weaponization risks leading to online radicalization, and the substantial social risks associated with LLMs. Researchers underscore the importance of risk mitigation strategies, policy interventions, and ethical design in these models. The need for considering human values, societal harms, and benefits in a contextual approach to AI ethics is also underlined.

We can also learn about the potential misuse of LLMs across various domains like email communication, education, decision-making, and eHealth. We are made aware of how these models, despite their beneficial applications, can pose risks such as enabling sophisticated phishing, undermining academic integrity, potentially influencing human decisions unknowingly, and even introducing errors in health-related advice. This information emphasizes the importance of establishing robust measures and ethical guidelines for using LLMs to mitigate such risks and ensure their safe and responsible use.

Fortunately, various countermeasures and regulations have been proposed to mitigate the misuse and potential risks of LLMs, like creating automated systems that engage scammers and extract valuable information from them. Besides, various research works propose different strategies to regulate LLMs, addressing each party involved in their development and use, and dealing with aspects like data protection, content moderation, and policy proposals. This section also brings attention to potential risks related to using AI systems for safety-related advice, emphasizing the need for expert verification and understanding the limitations of AI advice.

VI. Attacks and Privacy Violation

Large language models such as ChatGPT exhibit vulnerability to several attacks like misdirection attacks, extraction attacks, and backdoor attacks. The objective of these attacks ranges from manipulating the model's output to serve the attacker's intent, to extracting sensitive information from the model's training data, thus posing serious privacy risks. Certain backdoor attacks can even implant hidden triggers in the language model, which can later be activated to fulfill the attackers' purposes. Although these attacks have been extensively studied in other deep learning fields, it is crucial to consider the unique operational and structural aspects of LLMs. The complexity and the sheer scale of LLMs increase the potential impact of these attacks. For instance, misdirection attacks can be particularly deceptive by exploiting the model's sensitivity to nuanced textual changes, shifting the generation process to produce misleading or harmful content. Furthermore, extraction attacks pose a significant risk given the expansive and potentially sensitive nature of the data LLMs are trained on, enabling attackers to reconstruct or infer private data more effectively. Lastly, backdoor attacks in LLMs can be triggered by simple text inputs, making them highly insidious and challenging to detect without specific safeguards such as detailed auditing of training data and model responses. Therefore, ensuring the security of LLMs against these attacks involves not only adapting existing deep learning security measures but also developing new strategies tailored to the complexities of language processing and generation.

Of particular concern is the potential for LLMs to reveal sensitive information they have learned during training. This has prompted researchers to delve deeper into the analysis of the data leakage problem in LLMs, thereby formulating various attack detection and mitigation techniques to bolster the security of ChatGPT.

A. Attacks

In this section, we discuss different types of attacks that can target ChatGPT as shown in Table V, each has its unique modes of operation and resulting impacts.

1) Misdirection Attacks: Misdirection attacks serve to subvert the intended functioning of language models such as ChatGPT. Two primary forms of misdirection attacks have been studied in the literature: input-agnostic attacks and prompt injection attacks. In a study [146] about input-agnostic attacks, the focus is on finding universal adversarial triggers, which are input-agnostic sequences of tokens that trigger a model to produce a specific prediction when concatenated to any input from a dataset. The proposed approach involves a gradient-guided search over tokens to find short trigger sequences that successfully trigger the target prediction. The effectiveness of these triggers is demonstrated for text classification, reading comprehension, and conditional text generation tasks, causing targeted errors or generating offensive output. The triggers are optimized using white-box access to a specific model but can be transferred to other models on all considered datasets. The study also provides insights into global model behavior through these input-agnostic attacks. Another investigation [147] explores the susceptibility of ChatGPT to universal adversarial triggers (UATs) concerning the topic and stance of generated conditional text. Triggers are identified that cause ChatGPT to produce text on targeted topics while influencing the stance of the text. The

TABLE V
SUMMARY OF ATTACKS ON LLMS ([117], [146]–[159])

Research category	Reference	Year	Model	Methodology
Misdirection attacks	[146]	2019	GPT-2	Universal adversarial triggers
	[147]	2021	GPT-2	Universal adversarial triggers (for topic and stance manipulation)
	[148]	2022	GPT-3	PROMPT INJECT framework (for goal hijacking and prompt leaking)
	[149]	2023	ChatGPT	Analysis of indirect prompt injection threats
	[150]	2024	ChatGPT	Analysis of jailbreak prompts and their effectiveness in bypassing restrictions
	[151]	2024	GPT-3, GPT-4	Exploration of prompt stealing attacks and their mitigation
	[117]	2024	ChatGPT	Exploration of malicious content generation through prompt engineering
Extraction attacks	[152]	2021	GPT-2	Black-box extraction attack
	[153]	2023	BERT, GPT-2	Panning attack (targeted extraction of privacy-critical phrases)
	[154]	2024	GPT-3, GPT-4	Conversation reconstruction attack using black-box access
	[155]	2024	GPT-4	Context injection attack to extract sensitive information
Backdoor attacks	[156]	2021	BERT, GPT-2, XLNet	TROJANLM (trojaning attacks on LMs)
	[157]	2021	BERT, GPT-2	Homograph replacement & dynamic sentence attack (hidden backdoor attacks)
	[158]	2022	BERT, RoBERTa, GPT-2	Linguistic style-motivated (LISM) backdoor attack
	[159]	2023	BERT, RoBERTa, XLNet, T5, GPT-2	TFLexAttack (training-free backdoor attack)

paper highlights the potential danger of this technique and emphasizes the need for immediate safeguards against it. Regarding prompt injection attacks, vulnerabilities of ChatGPT to adversarial attacks through prompt injection are investigated [148]. The authors propose a framework that explores two types of prompt injection attacks: 1) goal hijacking and 2) prompt leaking, and demonstrating their feasibility and effectiveness. The study provides insights into the factors influencing prompt injection attacks and emphasizes the importance of studying them to understand the security risks associated with large language models. In a later analysis [149], novel prompt injection threats to application-integrated large language models are presented. The study highlights the potential vulnerability of such models to indirect prompt injections. A systematic investigation of the resulting threat landscape of application-integrated LLMs is conducted, and various new attack vectors are discussed. Recent research [150] delves into the intricacies of jailbreak prompts, which are crafted with the intent to bypass the restrictions of LLMs like ChatGPT. This study systematically analyzes common patterns in jailbreak prompts and assesses their effectiveness in circumventing model restrictions, highlighting the evolving nature of LLM vulnerabilities. Another study [151] explores prompt stealing attacks against large language models, examining how adversaries can manipulate prompts to generate unauthorized outputs. The research provides insights into the methods and effectiveness of these attacks, offering strategies for mitigating their impact. In addition, research by [117] investigates the use of prompt engineering to coerce ChatGPT into generating malicious content, such as phishing emails and malware. The study highlights the significant risks posed by these attacks and underscores the need for robust defenses against them.

2) Extraction Attacks: Extraction attacks represent another major category of security threats to ChatGPT, potentially compromising sensitive information embedded within its training data. In one study [152], an attack on language models is presented, which can extract verbatim training examples using only black-box query access. The attack is demonstrated on the model, successfully extracting hundreds of verbatim text sequences, including personally identifiable information and code. In another research effort, the first attack on federated learning (FL) that achieves targeted extraction of sequences containing privacy-critical phrases is proposed [153]. The authors introduce a novel attack called "panning", which focuses on specific keywords or triggers to extract all tokens of user data that follow their occurrence. The paper also provides a comprehensive review of relevant literature and background information on FL and text applications. Recent work [154] presents a conversation reconstruction attack against GPT models, demonstrating how adversaries can extract and reconstruct entire conversations using only black-box access. This attack poses significant privacy risks, as it can reveal sensitive information embedded within the model's training data. Additionally, a study [155] investigates context injection attacks on large language models, where adversaries subtly alter the context provided to the model to extract sensitive information. The research highlights the ease with which context manipulation can lead to unintended data leaks, emphasizing the need for robust mitigation strategies.

3) Backdoor Attacks: Backdoor attacks pose a unique challenge, as they exploit hidden triggers within the model to achieve their malevolent goals. The overlooked security threats created by pre-trained language models on natural language processing systems are addressed [156]. A new class of trojaning attacks, named TROJANLM, is presented, allowing adversaries to craft maliciously designed LMs that manipulate the behavior of the host NLP systems. Empirical evaluations and user studies of the proposed attacks on three state-of-the-art LMs demonstrate their effectiveness in security-critical NLP tasks. Analytical justifications for the practicality of TROJANLM are provided, along with discussions on potential countermeasures and future research directions. Two hidden backdoor attacks that deceive human-centric language models using covert and natural triggers are proposed [157]. These triggers are embedded through two state-of-the-art trigger embedding methods, achieving high attack success rates while maintaining functionality for regular users. The attacks are demonstrated across three security-critical NLP tasks. A new type of backdoor attack called linguistic style-motivated (LISM) backdoor attack is presented [158], exploiting implicit linguistic styles as hidden triggers for backdooring NLP models. Text style transfer models are utilized to generate sentences with an attacker-specified linguistic style, achieving improved attack effectiveness and stealthiness. A novel and practical methodology called TFLexAttack is proposed [159], enabling a training-free backdoor attack on language models. The attack manipulates the embedding dictionary of the language model's tokenizer using carefully designed rules. Extensive experiments demonstrate the effectiveness and universality of the attack across different NLP tasks.

4) Detection and Analysis: To effectively counteract the threats to ChatGPT, it is essential to develop robust methods for detecting and analyzing these attacks. An evaluation [160] of pre-trained language models is proposed, where adversarial examples are crafted to highlight their susceptibility to attacks during training and fine-tuning stages. A significant decrease in text classification quality is demonstrated when evaluating for semantic similarity, and a major security vulnerability in GPT-3 is identified. The robustness of GPT-3 Playground and GPT-3 API is compared against multiple pre-trained BERT-flavored models with adversarial examples, revealing the limitations of common distance measures and n-gram translation measures in capturing meaningful semantic differences between classifications. A new backdoor scanning technique called PICCOLO is proposed [161] to detect complex backdoors in NLP transformer models. This approach transforms the subject model to an equivalent, differentiable form and conducts trigger inversion using newly proposed optimization methods and a novel word discriminativity analysis. PICCOLO achieves high detection accuracy on a large number of NLP models, including state-of-the-art transformers BERT and GPT, as well as LSTM and GRU models. The dual-use risks of instruction-following large language models are analyzed [162], demonstrating their potential for malicious purposes and their ability to bypass existing defenses against misuse. It is shown that instruction-following LLMs can generate natural and convincing personalized malicious content, bypassing content filters of LLM API vendors. Addressing these attacks will require the development of new approaches to mitigations.

B. Privacy Violation

The use of ChatGPT raises concerns about user privacy as the model may store and use sensitive information such as personal conversations, search queries, and user preferences. To address this issue, researchers have proposed various methods such as differential privacy, and encryption techniques to ensure that user data remains secure and private. In this subsection, we discuss the nature of this problem and the possible solutions put forth by researchers.

1) Information Leakage: Information leakage is a form of privacy violation that poses a risk to sensitive user data. A systematic study [163] has been conducted on the privacy risks associated with general-purpose language models widely used in various NLP tasks and real-world systems. It has been demonstrated that text embeddings from these models can capture sensitive information, which could be exploited by adversaries, leading to real-world harm. In the context of chatbots, a new task has been proposed to detect personal information leakage [164]. This task involves aligning utterances with descriptions of personal information, and novel constrained alignment models have been introduced to address it under a weakly supervised setting. Empirical results highlight the effectiveness of these models in detecting the risk of personal information leakage, with advanced dialogue models showing a higher likelihood of leaking personal information. While leakage of email addresses has been investigated [165], it has been observed that language models are generally weak at associating leaked information with specific individuals, resulting in a low risk of targeted attacks. However, it is important not to disregard the potential privacy risks associated with these models. In the context of applications like smart reply, the risk of sensitive data leakage has been analyzed [166] using two types of query access: black-box and gray-box. The aim is to understand the vulnerabilities in smart reply pipelines that could lead to information leakage. By exploring this new threat model, insights can be gained into potential vulnerabilities and associated risks. Another analysis [167] focuses on the risk of personally identifiable information (PII) leakage in Language Models. A taxonomy of PII leakage has been proposed, encompassing black-box extraction, inference, and reconstruction attacks. Novel attacks have been developed to extract PII sequences that were not captured by existing attacks. The impact of differential privacy (DP) on protecting against PII leakage has also been evaluated, with factors increasing the risk of PII leakage identified.

2) Memorization Risks: Memorization of training data presents another potential risk to user privacy. Several studies have investigated this risk and explored its potential impact and mitigation strategies. A study [168] has focused on the memorization of training data during the fine-tuning phase of autoregressive language models. Different fine-tuning methods have been examined, and two proxy metrics have been used to measure memorization. The study reveals that fine-tuning the head of the model exhibits the highest level of memorization, while fine-tuning smaller adapters appears to be less susceptible to known extraction attacks. Understanding the susceptibility of different fine-tuning methods to attacks is crucial in the context of the pre-train and fine-tune paradigm. Another investigation [169] has delved into the problem of memorization in neural language models, specifically when the models emit verbatim parts of their training data. A precise definition of memorization has been proposed, and three factors influencing the degree of memorization have been identified. The study provides empirical evidence that memorization in language models is more prevalent than previously believed and is expected to worsen with larger models.

3) Protection and Defence: With the increasing awareness of privacy threats, researchers are actively working on developing effective protective and defensive measures. An alternative defense approach called MEMFREE decoding [170] has been proposed. This approach utilizes Bloom filters as a means to prevent verbatim memorization efficiently. However, it has been demonstrated that even with the use of perfect filters, models can still extract training data through techniques like style transfer. This highlights the need to consider a broader definition of memorization in order to address potential data extraction risks. Privacy concerns associated with language models have been explored [171], emphasizing the limitations of existing protection techniques such as data sanitization and differential privacy. It is argued that ensuring contextual integrity is necessary for preserving privacy in language models, suggesting that models should be trained on data intended for fully public use. In the realm of secure data sharing in NLP tasks, a novel approach [172] has been proposed for generating synthetic datasets with data privacy guarantees. This approach involves penalizing the generation of samples that fit another label, reducing the risk of faulty labeled samples in synthetic datasets and yielding high-quality datasets with maintained accuracy in NLP tasks. Additionally, a framework called just fine-tune twice (JFT) [173] has been introduced to achieve selective differential privacy (SDP) for large transformer-based models. The framework involves redacting in-domain data and employing a private training mechanism. Empirical studies have shown that JFT provides privacy guarantees under SDP while maintaining strong utility compared to previous approaches.

C. Summary

LLMs like ChatGPT pose significant security and privacy concerns that require robust and efficient mitigations. These concerns fall into three primary categories: model attacks, privacy violations, and information leakage. Model attacks such as misdirection, extraction, and backdoor attacks could disrupt the system's normal operation or compromise sensitive information. Various detection and analysis techniques have been suggested to handle these threats. Privacy issues revolve around how ChatGPT stores and utilizes user data, emphasizing potential data leakage and user data memorization. To counter these privacy risks, defensive measures like differential privacy and data sanitization have been suggested. It is worth noting that there is a trade-off between privacy protection and the development of robust methods for attack detection and analysis, since the requirements of their objectives are conflicting. For instance, while detailed logging and monitoring are indispensable for identifying and mitigating sophisticated attacks, they can also create significant privacy risks if the collected data includes sensitive information. Techniques such as differential privacy, designed to protect user data, can inadvertently reduce a model's capability to detect anomalies by obscuring genuine data patterns. Despite these advancements, a number of challenges persist in ensuring user privacy and the secure functionality of the ChatGPT system.

VII. Discussion and Future Work

Addressing the aforementioned issues within the context of AI systems such as ChatGPT needs a multifaceted approach, involving not only technical solutions but also legal and policy measures, among others. Here is a brief discussion of possible future solutions to each issue.

A. Bias Solutions

Despite the progress made in detecting, evaluating, and mitigating biases in LLMs, several research directions remain open:

1) Diverse and Representative Data Collection: Future work should focus on collecting more diverse and representative data for training AI models. This could help reduce the under-representation of certain groups and the over-representation of certain perspectives or worldviews. To effectively collect more diverse and representative data, strategies such as targeted data gathering from underrepresented regions and communities should be employed. Leveraging crowd-sourcing platforms can also diversify input sources, thereby reducing cultural and demographic biases inherently present in existing datasets. Furthermore, partnerships with diverse organizations can aid in gathering a broad spectrum of data, ensuring that the AI models trained on these datasets are well-rounded and equitable.

2) More Comprehensive Evaluation Metrics: The development of more comprehensive and robust metrics for evaluating bias in LLMs can help uncover subtler and less-studied forms of bias. Developing more comprehensive evaluation metrics involves creating algorithms that can detect and measure bias on multiple levels and dimensions, including intersectionality. This requires not only technical development but also collaboration with social scientists to ensure that the metrics are meaningful across different cultural and social contexts. Additionally, continuous validation of these metrics against real-world outcomes is essential to maintain their relevance and effectiveness.

3) Bias Mitigation Techniques: While several debiasing techniques have been proposed, none have proven completely effective. Future research should explore hybrid approaches that combine multiple debiasing techniques to enhance effectiveness. Techniques such as data augmentation, algorithmic fairness interventions, and regular model retraining with updated, less biased data should be further investigated. Additionally, exploring the use of artificial intelligence to autonomously detect and correct biases in AI models presents a promising research avenue.

4) Transparency and Interpretability: Future work could also focus on making the functioning of LLMs more transparent and interpretable, making it easier to understand how and why biases occur. Increasing transparency involves developing methods to trace model decisions back to specific data inputs or model parameters. Techniques such as model-agnostic explanations and the implementation of "explainability by design" in AI systems can help. Further, promoting open-source frameworks where models and their training algorithms can be audited by external parties will enhance transparency.

5) Ethical Guidelines and Policies: Finally, establishing ethical guidelines and policies can ensure that AI models are developed and used responsibly and that potential harms and biases are adequately addressed. This involves not only the AI research community but also lawmakers and the public in discussions on ethical AI usage. Regularly updated policies reflecting the latest AI advancements and societal values are essential for guiding ethical AI development and deployment.

B. Disinformation Solutions

The twin issues of disinformation generation and detection present several opportunities for future research.

1) Deeper Investigation: First, there is a need for continued investigation into the mechanics of AI-generated fake media. A comprehensive approach to understanding AI-generated fake media involves multidisciplinary research incorporating technical, psychological, and sociological perspectives. Studies should focus on the cognitive effects of fake media on different demographics and develop methodologies to mitigate these effects.

2) More Effective tools: Second, in the sphere of fake media detection, there is a clear demand for more effective and versatile tools capable of keeping pace with rapidly evolving disinformation tactics. This calls for innovation in both machine learning and deep learning models, as well as the exploration of novel approaches that take into account variables such as veracity of text, user demographic characteristics, and the subtle nuances of writing style.

3) Interdisciplinary Collaboration: Finally, it is imperative to foster cross-disciplinary dialogue and collaboration in order to address the broader ethical, social, and political implications of AI-generated disinformation. As AI technologies become increasingly embedded in our lives, the collective responsibility to ensure their beneficial and ethical use grows in equal measure.

C. Ethics and Misuse Solutions

From the literature review above, we have a comprehensive understanding of the ethical issues and challenges associated with LLMs like ChatGPT. Below, we propose some potential solutions and future research directions:

1) Technological Safeguards: Future AI research should prioritize developing more robust safety and security measures to prevent misuse. These can include refining adversarial training and attack detection methods, as well as exploring innovative new techniques for enhancing the robustness of AI models.

2) Ethics-by-Design Approach: AI and LLM development should adopt an ethics-by-design approach. This includes designing and developing AI models with ethical considerations integrated from the outset. Future research should focus on methods and best practices for integrating ethical considerations into the AI development process.

3) Education and Awareness: There is a need to increase public education and awareness about AI, LLMs, and their potential misuse. This includes educating students about the potential misuse of AI in academic settings, and raising public awareness about the potential risks of AI-powered phishing or social engineering attacks. Future research could focus on the development and effectiveness of AI education and awareness programs.

In summary, addressing the ethical implications and potential misuse of AI and LLMs is a complex task that requires a comprehensive and multifaceted approach. Future research should continue to explore and develop effective strategies and techniques for addressing these challenges.

D. Attacks and Privacy Solutions

Implementing robust security protocols can help protect AI systems from attacks. This might include measures like encryption, regular software updates, and intrusion detection systems. Since LLMs are complex and large, techniques such as fuzzing [174]–[176] may also have great potential to detect such attacks. AI models themselves can be hardened against attacks through techniques such as adversarial training, where the model is trained to resist manipulation by adversarial inputs.

1) Effective Utilization of Differential Privacy (DP): Existing privacy-preserving mechanisms such as differential privacy have shown promise in preventing information leakage and memorization risks. Differential Privacy is a framework designed to ensure that the output of a database query does not reveal too much information about any individual record, even to someone with access to the query's result. The application of DP to LLMs involves introducing noise into the training data or the learning process to obscure the specifics of individual data entries. This can be achieved by adjusting the model's parameters during training in a way that the inclusion or exclusion of any single data point does not significantly change the output of the model. For LLMs, one common method is to apply DP during the stochastic gradient descent phase—a pivotal step where the model learns by incrementally adjusting its weights based on a subset of data. However, there are still limitations, such as the trade-off between privacy guarantees and the performance of the model. Future research should focus on enhancing these mechanisms or developing new ones that can provide strong privacy guarantees while maintaining high utility. These new methods could build on techniques like neural dynamics (ND) optimiser [177], just fine-tune twice (JFT) [173] and the use of Bloom filters [170].

2) Robust Adversarial Training Techniques: Adversarial training could be used to improve the robustness of LLMs against misdirection and other attacks. Future studies should aim at developing scalable adversarial training methods specifically for large language models. Concurrently, robustness analysis methods that can assess the model's vulnerability to attacks should also be developed [95].

3) Creating Synthetic Datasets: The use of synthetic datasets that maintain the statistical properties of original datasets while preserving data privacy has been proposed as a potential solution [172]. Further research in this area could aim to refine the process of generating synthetic datasets, ensuring that they retain the necessary properties for effective machine learning while minimizing the risk of privacy violation.

4) Balanced Approaches: Since techniques such as differential privacy can inadvertently reduce a model's capability to detect anomalies by obscuring genuine data patterns, a balanced approach is necessary. One promising avenue is the use of advanced cryptographic methods, such as homomorphic encryption, which enables the processing of encrypted data without exposing actual content, thus maintaining privacy while allowing for effective attack detection. Moreover, the adoption of federated learning can decentralize data processing, reducing the risk of privacy breaches while still enabling collective model improvements. Future research should focus on enhancing these techniques to better serve the dual needs of privacy and security in AI applications.

It is important to note that while these approaches can help mitigate these problems, they are not foolproof and they need to be complemented with strong regulatory oversight and legal measures. Engaging in regular dialogues with various stakeholders including policymakers, users, and experts in the field can also help in continuously refining these strategies.

By considering these solutions and future directions, it is possible to mitigate the privacy concerns associated with AI models such as ChatGPT while retaining their utility in various applications.

E. Further Discussion

To delve deeper into these issues, it is crucial to consider the interplay between technical advancements and societal impacts. For example, while federated learning offers privacy benefits, it also raises challenges related to model consistency and communication overhead. Balancing these trade-offs requires innovative solutions that optimize both privacy and performance.

Moreover, the ethical implications of AI usage extend beyond immediate technical concerns. For instance, the deployment of AI in sensitive areas such as healthcare and law enforcement demands rigorous ethical scrutiny to prevent harm and ensure fairness. Developing AI systems that can explain their decisions in human-understandable terms is essential for accountability and trust.

Interdisciplinary collaboration is key to addressing these complex challenges. By bringing together experts from AI, ethics, law, and social sciences, we can develop holistic approaches that encompass technical, ethical, and societal dimensions. This collaboration can also drive the creation of international standards and guidelines, ensuring a consistent and unified approach to AI governance.

Finally, public engagement and education are paramount. Empowering individuals with knowledge about AI's capabilities and limitations can foster responsible use and mitigate risks. Educational initiatives should focus on building digital literacy and critical thinking skills, enabling users to navigate the AI-driven landscape effectively.

While this survey has primarily focused on the security issues associated with ChatGPT, it is important to recognize that similar concerns extend to other LLMs. For instance, the potential for adversarial attacks, privacy violations, and misuse in disinformation campaigns are common across many LLMs, albeit to varying degrees. Expanding the discussion to include these models not only enhances the comprehensiveness of the review but also provides valuable insights into the broader implications of LLM security. We plan to explore this in future work with the aim to conduct a more extensive review that not only addresses the security issues specific to ChatGPT but also provides a comprehensive comparison across a variety of LLMs. This future research will help in developing more robust and generalized security frameworks that can be applied across different AI models.

VIII. Conclusion

While ChatGPT has shown significant improvements in various NLP tasks, it also presents potential risks that need to be addressed to ensure its safe and responsible use. This survey has examined the security challenges posed by large language models like ChatGPT, highlighting significant concerns in data privacy, potential misuse, inherent biases, and susceptibility to cyber-attacks. The extensive datasets used for training these models raise issues of inadvertent data leakage, underscoring the need for robust privacy protections such as differential privacy and federated learning. The potential for misuse in generating deceptive content calls for regulatory frameworks to mitigate abuse and ensure ethical deployment. Biases within training data can perpetuate stereotypes, necessitating diversified datasets and continuous monitoring to neutralize these biases effectively. Moreover, the vulnerability of these models to sophisticated cyber threats, including adversarial and backdoor attacks, emphasizes the importance of developing advanced security measures. Addressing these challenges requires a concerted effort from policymakers, developers, and end-users. Policymakers must draft regulations that balance innovation with ethical considerations, developers should adhere to security best practices, and end-users need education on the risks associated with these technologies. Collaborative engagement is crucial to harnessing the benefits of large language models while safeguarding against their risks.

References

[1] Q. Miao, W. Zheng, Y. Lv, M. Huang, W. Ding, and F.-Y. Wang, "DAO to HANOI via DeSci: AI paradigm shifts from AlphaGo to ChatGPT," *IEEE/CAA J. Autom. Sinica*, vol. 10, no. 4, pp. 877–897, Apr. 2023.

[2] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," *Nature*, vol. 521, no. 7553, pp. 436–444, May 2015.

[3] T. Wu, S. He, J. Liu, S. Sun, K. Liu, Q.-L. Han, and Y. Tang, "A brief

overview of ChatGPT: The history, status quo and potential future development," *IEEE/CAA J. Autom. Sinica*, vol. 10, no. 5, pp. 1122–1136, May 2023.

[4] OpenAI, "Introducing ChatGPT," [Online]. Available: https://openai.com/blog/chatgpt, November 30, 2022.

[5] W. D. Heaven, "OpenAI's new language generator GPT-3 is shockingly good—and completely mindless," [Online]. Available: https://www.technologyreview.com/2020/07/20/1005454/openai-machine-learning-language-generator-gpt-3-nlp/, July 20, 2020.

[6] S. Biswas, "The function of chat GPT in social media: According to chat GPT," 2023.

[7] H. Hassani and E. S. Silva, "The role of ChatGPT in data science: How AI-assisted conversational interfaces are revolutionizing the field," *Big Data Cogn. Comput*, vol. 7, no. 2, p. 62, Mar. 2023.

[8] T. B. Brown, B. Mann, N. Ryder, M. Subbiah, J. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, S. Agarwal, A. Herbert-Voss, G. Krueger, T. Henighan, R. Child, A. Ramesh, D. M. Ziegler, J. Wu, C. Winter, C. Hesse, M. Chen, E. Sigler, M. Litwin, S. Gray, B. Chess, J. Clark, C. Berner, S. McCandlish, A. Radford, I. Sutskever, and D. Amodei, "Language models are few-shot learners," in *Proc. 34th Int. Conf. Neural Information Processing Systems*, Vancouver, Canada, 2020, pp. 159.

[9] F.-Y. Wang, Q. Miao, X. Li, X. Wang, and Y. Lin, "What does ChatGPT say: The DAO from algorithmic intelligence to linguistic intelligence," *IEEE/CAA J. Autom. Sinica*, vol. 10, no. 3, pp. 575–579, Mar. 2023.

[10] C. Guo, Y. Lu, Y. Dou, and F.-Y. Wang, "Can ChatGPT boost artistic creation: The need of imaginative intelligence for parallel art," *IEEE/CAA J. Autom. Sinica*, vol. 10, no. 4, pp. 835–838, Apr. 2023.

[11] D. Baidoo-Anu and L. Owusu Ansah, "Education in the era of generative artificial intelligence (AI): Understanding the potential benefits of ChatGPT in promoting teaching and learning," *J. AI*, vol. 7, no. 1, pp. 52–62, Jan.–Dec. 2023.

[12] B. D. Lund and T. Wang, "Chatting about ChatGPT: How may AI and GPT impact academia and libraries?," *Lib. Hi Tech News*, vol. 40, no. 3, pp. 26–29, May 2023.

[13] S. Sok and K. Heng, "ChatGPT for education and research: A review of benefits and risks," SSRN, Mar 2023.

[14] D. Kalla, N. Smith, S. Kuraku, and F. Samaah., "Study and analysis of chat GPT and its impact on different fields of study," *Int. J. Innovative Sci. Res. Technol.*, vol. 8, no. 3, pp. 827–833, Mar. 2023.

[15] X. Xue, X. Yu, and F.-Y. Wang, "ChatGPT chats on computational experiments: From interactive intelligence to imaginative intelligence for design of artificial societies and optimization of foundational models," *IEEE/CAA J. Autom. Sinica*, vol. 10, no. 6, pp. 1357–1360, Jun. 2023.

[16] F.-Y. Wang, J. Yang, X. Wang, J. Li, and Q.-L. Han, "Chat with ChatGPT on industry 5.0: Learning and decision-making for intelligent industries," *IEEE/CAA J. Autom. Sinica*, vol. 10, no. 4, pp. 831–834, Apr. 2023.

[17] D. M. Korngiebel and S. D. Mooney, "Considering the possibilities and pitfalls of generative pre-trained transformer 3 (GPT-3) in healthcare delivery," *NPJ Digital Med.*, vol. 4, no. 1, p. 93, Jun. 2021.

[18] M. Sallam, "ChatGPT utility in healthcare education, research, and practice: Systematic review on the promising perspectives and valid concerns," *Healthcare*, vol. 11, no. 6, p. 887, Mar. 2023.

[19] S. S. Biswas, "Role of chat GPT in public health," *Ann. Biomed. Eng.*, vol. 51, no. 5, pp. 868–869, Mar. 2023.

[20] M. Sallam, N. Salim, M. Barakat, and A. Al-Tammemi, "ChatGPT applications in medical, dental, pharmacy, and public health education: A descriptive study highlighting the advantages and limitations," *Narra J.*, vol. 3, no. 1, p. e103, Mar. 2023.

[21] M. Sallam, "The utility of ChatGPT as an example of large language models in healthcare education, research and practice: Systematic review on the future perspectives and potential limitations," *medRxiv*, 2023, DOI: 10.1101/2023.02.19.23286155.

[22] A. Ausekar and R. Bhagwat, "Banking on AI: Exploring the transformative role of ChatGPT in financial services," in *Proc. IEEE Engineering Informatics*, Melbourne, Australia, Nov. 2023, pp. 1−6.

[23] P. Rivas and L. Zhao, "Marketing with ChatGPT: Navigating the ethical terrain of GPT-based chatbot technology," *AI*, vol. 4, no. 2, pp. 375–384, Apr. 2023.

[24] G. F. Frederico, "ChatGPT in supply chains: Initial evidence of applications and potential research agenda," *Logistics*, vol. 7, no. 2, p. 26, Apr. 2023.

[25] I. Carvalho and S. Ivanov, "ChatGPT for tourism: Applications, benefits and risks," *Tourism Rev.*, vol. 79, no. 2, pp. 290–303, Feb. 2024.

[26] B. Ahmad, S. Thakur, B. Tan, R. Karri, and H. Pearce, "Fixing hardware security bugs with large language models," arXiv preprint arXiv: 2302.01215, 2023.

[27] C. S. Xia and L. Zhang, "Conversational automated program repair," arXiv preprint arXiv: 2301.13246, 2023.

[28] D. Sobania, M. Briesch, C. Hanna, and J. Petke, "An analysis of the automatic bug fixing performance of ChatGPT," in *Proc. IEEE/ACM Int. Workshop on Automated Program Repair*, Melbourne, Australia, 2023, pp. 1–8.

[29] M. Nair, R. Sadhukhan, and D. Mukhopadhyay, "Generating secure hardware using ChatGPT resistant to CWEs," *Cryptology ePrint Archive*, 2023.

[30] N. M. S. Surameery and M. Y. Shakor, "Use chat GPT to solve programming bugs," *Int. J. Inf. Technol. Comput. Eng.*, vol. 3, no. 1, pp. 17–22, Dec. 2022.

[31] Twingate Team, "What happened in the ChatGPT data breach?" [Online]. Available: https://www.twingate.com/blog/tips/chatgpt-data-breach, August 2, 2023.

[32] A. Mudaliar, "Samsung bans ChatGPT for staff, Microsoft hints potential alternative," [Online]. Available: https://www.spiceworks.com/tech/artificial-intelligence/news/samsung-bans-chatgpt-for-staff/, August 2, 2023.

[33] B. Guembe, A. Azeta, S. Misra, V. C. Osamor, L. Fernandez-Sanz, and V. Pospelova, "The emerging threat of AI-driven cyber attacks: A review," *Appl. Artif. Intell.*, vol. 36, no. 1, p. 2037254, Mar. 2022.

[34] Y. Himeur, S. S. Sohail, F. Bensaali, A. Amira, and M. Alazab, "Latest trends of security and privacy in recommender systems: A comprehensive review and future perspectives," *Comput. Secur.*, vol. 118, p. 102746, Jul. 2022.

[35] A. Mascellino, "New research exposes security risks in ChatGPT plugins," [Online]. Available: https://www.infosecurity-magazine.com/news/security-risks-chatgpt-plugins, August 2, 2023.

[36] S. L. Blodgett, S. Barocas, H. Daumé Ⅲ, and H. Wallach, "Language (technology) is power: A critical survey of 'bias' in NLP," in *Proc. 58th Annu. Meeting of the Association for Computational Linguistics*, 2020, pp. 5454–5476.

[37] E. Sheng, K.-W. Chang, P. Natarajan, and N. Peng, "Societal biases in language generation: Progress and challenges," in *Proc. 59th Annu. Meeting of the Association for Computational Linguistics and the 11th Int. Joint Conf. Natural Language Processing*, 2021, pp. 4275-4293.

[38] M. Hasal, J. Nowaková, K. Ahmed Saghair, H. Abdulla, V. Snášel, and L. Ogiela, "Chatbots: Security, privacy, data protection, and social aspects," *Concurrency Comput.: Pract. Exper.*, vol. 33, no. 19, p. e6426, Jun. 2021.

[39] J. Pu, Z. Sarwar, S. M. Abdullah, A. Rehman, Y. Kim, P. Bhattacharya, M. Javed, and B. Viswanath, "Deepfake text detection: Limitations and opportunities," in *Proc. IEEE Symp. Security and Privacy*, San Francisco, USA, 2023, pp. 1613–1630.

[40] L. Weidinger, J. Uesato, M. Rauh, C. Griffin, P.-S. Huang, J. Mellor, A. Glaese, M. Cheng, B. Balle, A. Kasirzadeh, C. Biles, S. Brown, Z. Kenton, W. Hawkins, T. Stepleton, A. Birhane, L. A. Hendricks, L. Rimell, W. Isaac, J. Haas, S. Legassick, G. Irving, and I. Gabriel, "Taxonomy of risks posed by language models," in *Proc. ACM Conf. Fairness, Accountability, and Transparency*, Seoul, Republic of Korea,

2022, pp. 214–229.

[41] E. Kasneci, K. Sessler, S. Küchemann, M. Bannert, D. Dementieva, F. Fischer, U. Gasser, G. Groh, S. Günnemann, E. Hüllermeier, S. Krusche, G. Kutyniok, T. Michaeli, C. Nerdel, J. Pfeffer, O. Poquet, M. Sailer, A. Schmidt, T. Seidel, M. Stadler, J. Weller, J. Kuhn, and G. Kasneci, "ChatGPT for good? On opportunities and challenges of large language models for education," *Learn. Individual Differ.*, vol. 103, p. 102274, Apr. 2023.

[42] T. Y. Zhuo, Y. Huang, C. Chen, and Z. Xing, "Red teaming ChatGPT via jailbreaking: Bias, robustness, reliability and toxicity," arXiv preprint arXiv: 2301.12867, 2023.

[43] P. P. Ray, "ChatGPT: A comprehensive review on background, applications, key challenges, bias, ethics, limitations and future scope," *Internet Things Cyber-Phys. Syst.*, vol. 3, pp. 121–154, Apr. 2023.

[44] S. Kumar, V. Balachandran, L. Njoo, A. Anastasopoulos, and Y. Tsvetkov, "Language generation models can cause harm: So what can we do about it? An actionable survey," in *Proc. 17th Conf. European Chapter of the Association for Computational Linguistics*, Dubrovnik, Croatia, 2023, pp. 3299–3321.

[45] W. X. Zhao, K. Zhou, J. Li, T. Tang, X. Wang, Y. Hou, Y. Min, B. Zhang, J. Zhang, Z. Dong, Y. Du, C. Yang, Y. Chen, Z. Chen, J. Jiang, R. Ren, Y. Li, X. Tang, Z. Liu, P. Liu, J.-Y. Nie, and J.-R. Wen, "A survey of large language models," arXiv preprint arXiv: 2303.18223, 2023.

[46] J. Deng, H. Sun, Z. Zhang, J. Cheng, and M. Huang, "Recent advances towards safe, responsible, and moral dialogue systems: A survey," arXiv preprint arXiv: 2302.09270, 2023.

[47] C. Dilmegani, "Large language model training," [Online]. Available: https://research.aimultiple.com/large-language-model-training/, August 2, 2023.

[48] OpenAI, "Safety at every step," [Online]. Available: https://openai.com/safety, August 2, 2023.

[49] C. Yeo and A. Chen, "Defining and evaluating fair natural language generation," in *Proc. Fourth Widening Natural Language Processing Workshop*, Seattle, USA, 2020, pp. 107–109.

[50] L. Lucy and D. Bamman, "Gender and representation bias in GPT-3 generated stories," in *Proc. Third Workshop on Narrative Understanding*, 2021, pp. 48–55.

[51] J. Shihadeh, M. Ackerman, A. Troske, N. Lawson, and E. Gonzalez, "Brilliance bias in GPT-3," in *Proc. IEEE Global Humanitarian Technology Conf.*, Santa Clara, USA, 2022, pp. 62–69.

[52] D. M. Kaplan, R. Palitsky, S. J. Arconada Alvarez, N. S. Pozzo, M. N. Greenleaf, C. A. Atkinson, and W. A. Lam, "What's in a name? Experimental evidence of gender bias in recommendation letters generated by ChatGPT," *J. Med. Int. Res.*, vol. 26, p. e51837, Mar. 2024.

[53] W. Guo and A. Caliskan, "Detecting emergent intersectional biases: Contextualized word embeddings contain a distribution of human-like biases," in *Proc. AAAI/ACM Conf. AI, Ethics, and Society*, New York, USA, 2021, pp. 122–133.

[54] M. Nadeem, A. Bethke, and S. Reddy, "StereoSet: Measuring stereotypical bias in pretrained language models," in *Proc. 59th Annu. Meeting of the Association for Computational Linguistics and the 11th Int. Joint Conf. Natural Language Processing*, 2021, pp. 5356–5371.

[55] C. Logé, E. Ross, D. Y. A. Dadey, S. Jain, A. Saporta, A. Y. Ng, and P. Rajpurkar, "Q-pain: A question answering dataset to measure social bias in pain management," in *Proc. 35th Conf. Neural Information Processing Systems*, 2021, pp. 1–12.

[56] K. S. Amin, H. P. Forman, and M. A. Davis, "Even with ChatGPT, race matters," *Clin. Imaging*, vol. 109, pp. 110–113, May 2024.

[57] E. Sheng, K.-W. Chang, P. Natarajan, and N. Peng, "The woman worked as a babysitter: On biases in language generation," in *Proc. Conf. Empirical Methods in Natural Language Processing and the 9th Int. Joint Conf. Natural Language Processing*, Hong Kong, China, 2019, pp. 3407–3412.

[58] A. Zheng, "Dissecting bias of ChatGPT in college major recommendations," *Inf. Technol. Manage.*, 2024, DOI: 10.1007/s10799-024-00430-5.

[59] L. Lippens, "Computer says 'no': Exploring systemic bias in ChatGPT using an audit approach," *Comput. Hum. Behav.: Artif. Humans*, vol. 2, no. 1, p. 100054, Jan.–Jul. 2024.

[60] A. Abid, M. Farooqi, and J. Zou, "Persistent anti-muslim bias in large language models," in *Proc. AAAI/ACM Conf. AI, Ethics, and Society*, New York, USA, 2021, pp. 298–306.

[61] A. Abid, M. Farooqi, and J. Zou, "Large language models associate muslims with violence," *Nat. Mach. Intell.*, vol. 3, no. 6, pp. 461–463, Jun. 2021.

[62] A. Al Amin and K. S. Kabir, "A disability lens towards biases in GPT-3 generated open-ended languages," arXiv preprint arXiv: 2206.11993, 2022.

[63] L. Gover, "Political bias in large language models," *The Commons: Puget Sound Journal of Politics*, vol. 4, no. 1, pp. 11–22, 2023.

[64] P. Pit, X. Ma, M. Conway, Q. Chen, J. Bailey, H. Pit, P. Keo, W. Diep, and Y.-G. Jiang, "Whose side are you on? Investigating the political stance of large language models," arXiv preprint arXiv: 2403.13840, 2024.

[65] Y. Bang, D. Chen, N. Lee, and P. Fung, "Measuring political bias in large language models: What is said and how it is said," in *Proc. 62nd Annu. Meeting of the Association for Computational Linguistics*, Bangkok, Thailand, 2024, pp. 11142–11159.

[66] J. Hartmann, J. Schwenzow, and M. Witte, "The political ideology of conversational AI: Converging evidence on ChatGPT's pro-environmental, left-libertarian orientation," arXiv preprint arXiv: 2301.01768, 2023.

[67] D. Rozado, "The political biases of ChatGPT," *Soc. Sci.*, vol. 12, no. 3, p. 148, Mar. 2023.

[68] R. W. McGee, "Is chat GPT biased against conservatives? An empirical study," *SSRN Electron. J.*, 2023, DOI: 10.2139/ssrn.4359405.

[69] F. Motoki, V. Pinho Neto, and V. Rodrigues, "More human than human: Measuring ChatGPT political bias," *Public Choice*, vol. 198, no. 1–2, pp. 3–23, Jan. 2024.

[70] N. Retzlaff, "Political biases of ChatGPT in different languages," 2024.

[71] J. Rutinowski, S. Franke, J. Endendyk, I. Dormuth, M. Roidl, and M. Pauly, "The self-perception and political biases of ChatGPT," *Hum. Behav. Emerging Technol.*, vol. 2024, no. 1, p. 7115633, Jan. 2024.

[72] S. Gehman, S. Gururangan, M. Sap, Y. Choi, and N. A. Smith, "RealToxicityPrompts: Evaluating neural toxic degeneration in language models," in *Proc. Association for Computational Linguistics: EMNLP 2020*, 2020, pp. 3356–3369.

[73] N. Ousidhoum, X. Zhao, T. Fang, Y. Song, and D.-Y. Yeung, "Probing toxic content in large pre-trained language models," in *Proc. 59th Annu. Meeting of the Association for Computational Linguistics and the 11th Int. Joint Conf. Natural Language Processing*, 2021, pp. 4262–4274.

[74] D. Nozza, F. Bianchi, and D. Hovy, "HONEST: Measuring hurtful sentence completion in language models," in *Proc. Conf. North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 2021, pp. 2398–2406.

[75] P. Schramowski, C. Turan, N. Andersen, C. A. Rothkopf, and K. Kersting, "Large pre-trained language models contain human-like biases of what is right and wrong to do," *Nat. Mach. Intell.*, vol. 4, no. 3, pp. 258–268, Mar. 2022.

[76] T. Hartvigsen, S. Gabriel, H. Palangi, M. Sap, D. Ray, and E. Kamar, "ToxiGen: A large-scale machine-generated dataset for adversarial and implicit hate speech detection," in *Proc. 60th Annu. Meeting of the Association for Computational Linguistics*, Dublin, Ireland, 2022, pp. 3309–3326.

[77] F. Huang, H. Kwak, and J. An, "Is ChatGPT better than human annotators? Potential and limitations of ChatGPT in explaining implicit hate speech," in *Proc. ACM Web Conf.*, Austin, USA, 2023, pp. 294–297.

[78] P.-S. Huang, H. Zhang, R. Jiang, R. Stanforth, J. Welbl, J. Rae, V. Maini, D. Yogatama, and P. Kohli, "Reducing sentiment bias in language models via counterfactual evaluation," in *Proc. Association for Computational Linguistics: EMNLP 2020*, 2020, pp. 65–83.

[79] E. Sheng, K.-W. Chang, P. Natarajan, and N. Peng, "Towards controllable biases in language generation," in *Proc. Association for Computational Linguistics: EMNLP 2020*, 2020, pp. 3239–3254.

[80] P. P. Liang, C. Wu, L.-P. Morency, and R. Salakhutdinov, "Towards understanding and mitigating social biases in language models," in *Proc. 38th Int. Conf. Machine Learning*, 2021, pp. 6565–6576.

[81] C. Borchers, D. Gala, B. Gilburt, E. Oravkin, W. Bounsi, Y. M. Asano, and H. Kirk, "Looking for a handsome carpenter! Debiasing GPT-3 job advertisements," in *Proc. 4th Workshop on Gender Bias in Natural Language Processing*, Seattle, USA, 2022, pp. 212–224.

[82] R. Liu, C. Jia, J. Wei, G. Xu, L. Wang, and S. Vosoughi, "Mitigating political bias in language models through reinforced calibration," in *Proc. 35th AAAI Conf. Artificial Intelligence*, 2021, pp. 14857–14866.

[83] R. Liu, C. Jia, J. Wei, G. Xu, and S. Vosoughi, "Quantifying and alleviating political bias in language models," *Artif. Intell.*, vol. 304, p. 103654, Mar. 2022.

[84] J. Welbl, A. Glaese, J. Uesato, S. Dathathri, J. Mellor, L. A. Hendricks, K. Anderson, P. Kohli, B. Coppin, and P.-S. Huang, "Challenges in detoxifying language models," in *Proc. Association for Computational Linguistics*, Punta Cana, Dominican Republic, 2021, pp. 2447–2469.

[85] A. Xu, E. Pathak, E. Wallace, S. Gururangan, M. Sap, and D. Klein, "Detoxifying language models risks marginalizing minority voices," in *Proc. Conf. North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 2021, pp. 2390–2397.

[86] N. Inie, J. Falk Olesen, and L. Derczynski, "The Rumour mill: Making the spread of misinformation explicit and tangible," in *Proc. CHI Conf. Human Factors in Computing Systems*, Honolulu, USA, 2020, pp. 1–4.

[87] S. Kreps, R. M. McCain, and M. Brundage, "All the news that's fit to fabricate: AI-generated text as a tool of media misinformation," *J. Exp. Polit. Sci.*, vol. 9, no. 1, pp. 104–117, Nov. 2022.

[88] G. Spitale, N. Biller-Andorno, and F. Germani, "AI model GPT-3 (dis)informs us better than humans," *Sci. Adv.*, vol. 9, no. 26, p. eadh1850, Jun. 2023.

[89] H. Y. Li, "The possibility and optimization path of ChatGPT promoting the generation and dissemination of fake news," *Media Commun. Res.*, vol. 5, no. 2, pp. 80–86, 2024.

[90] P. Ranade, A. Piplai, S. Mittal, A. Joshi, and T. Finin, "Generating fake cyber threat intelligence using transformer-based models," in *Proc. Int. Joint Conf. Neural Networks*, Shenzhen, China, 2021, pp. 1–9.

[91] J. Mink, L. Luo, N. M. Barbosa, O. Figueira, Y. Wang, and G. Wang, "DeepPhish: Understanding user trust towards artificially generated profiles in online social networks," in *Proc. 31st USENIX Security Symp.*, Boston, USA, 2022, pp. 1669–1686.

[92] Y. Hu, Y. Lin, E. Skorupa Parolin, L. Khan, and K. Hamlen, "Controllable fake document infilling for cyber deception," in *Proc. Association for Computational Linguistics*, Abu Dhabi, United Arab Emirates, 2022, pp. 6505–6519.

[93] D. Barman, Z. Guo, and O. Conlan, "The dark side of language models: Exploring the potential of LLMs in multimedia disinformation generation and dissemination," *Mach. Learn. Appl.*, vol. 16, p. 100545, Jun. 2024.

[94] R. Zellers, A. Holtzman, H. Rashkin, Y. Bisk, A. Farhadi, F. Roesner, and Y. Choi, "Defending against neural fake news," in *Proc. 33rd Int. Conf. Neural Information Processing Systems*, Vancouver, Canada, 2019, pp. 812.

[95] H. Stiff and F. Johansson, "Detecting computer-generated disinformation," *Int. J. Data Sci. Anal.*, vol. 13, no. 4, pp. 363–383, May 2022.

[96] S. Rossi, Y. Kwon, O. H. Auglend, R. R. Mukkamala, M. Rossi, and J. Thatcher, "Are deep learning-generated social media profiles indistinguishable from real profiles?" in *Proc. 56th Hawaii Int. Conf. System Sciences*, 2023, pp. 134–143.

[97] A. Gupta, A. Singhal, A. Mahajan, A. Jolly, and S. Kumar, "Empirical framework for automatic detection of neural and human authored fake news," in *Proc. 6th Int. Conf. Intelligent Computing and Control Systems*, Madurai, India, 2022, pp. 1625–1633.

[98] A. Pagnoni, M. Graciarena, and Y. Tsvetkov, "Threat scenarios and best practices to detect neural fake news," in *Proc. 29th Int. Conf. Computational Linguistics*, Gyeongju, South Korea, 2022, pp. 1233–1249.

[99] M. Gambini, T. Fagni, F. Falchi, and M. Tesconi, "On pushing DeepFake tweet detection capabilities to the limits," in *Proc. 14th ACM Web Science Conf.*, Barcelona, Spain, 2022, pp. 154–163.

[100] B. Jiang, Z. Tan, A. Nirmal, and H. Liu, "Disinformation detection: An evolving challenge in the age of LLMs," in *Proc. SIAM Int. Conf. Data Mining*, 2024, pp. 427–435.

[101] Y. Huang, K. Shu, P. S. Yu, and L. Sun, "From creation to clarification: ChatGPT's journey through the fake news quagmire," in *Companion Proc. ACM Web Conf.*, Singapore, Singapore, 2024, pp. 513–516.

[102] S. B. Shah, S. Thapa, A. Acharya, K. Rauniyar, S. Poudel, S. Jain, A. Masood, and U. Naseem, "Navigating the web of disinformation and misinformation: Large language models as double-edged swords," *IEEE Access*, 2024, DOI: 10.1109/ACCESS.2024.3406644.

[103] Y. Tukmacheva, I. Oseledets, and E. Frolov, "Mitigating human and computer opinion fraud via contrastive learning," arXiv preprint arXiv: 2301.03025, 2023.

[104] A. Gambetti and Q. Han, "Combat AI with AI: Counteract machine-generated fake restaurant reviews on social media," arXiv preprint arXiv: 2302.07731, 2023.

[105] P. Henderson, K. Sinha, N. Angelard-Gontier, N. R. Ke, G. Fried, R. Lowe, and J. Pineau, "Ethical challenges in data-driven dialogue systems," in *Proc. AAAI/ACM Conf. AI, Ethics, and Society*, New Orleans, USA, 2018, pp. 123–129.

[106] K. McGuffie and A. Newhouse, "The radicalization risks of GPT-3 and advanced neural language models," arXiv preprint arXiv: 2009.06807, 2020.

[107] T. Y. Zhuo, Y. Huang, C. Chen, and Z. Xing, "Exploring AI ethics of ChatGPT: A diagnostic analysis," arXiv preprint arXiv: 2301.12867, 2023.

[108] A. Borji, "A categorical archive of ChatGPT failures," arXiv preprint arXiv: 2302.03494, 2023.

[109] A. Rasekh and I. Eisenberg, "Democratizing ethical assessment of natural language generation models," arXiv preprint arXiv: 2207.10576, 2022.

[110] A. Chan, "GPT-3 and instructGPT: Technological dystopianism, utopianism, and "contextual" perspectives in AI ethics and industry," *AI Ethics*, vol. 3, no. 1, pp. 53–64, Feb. 2023.

[111] J. Chatterjee and N. Dethlefs, "This new conversational AI model can be your friend, philosopher, and guide... and even your worst enemy," *Patterns*, vol. 4, no. 1, p. 100676, Jan. 2023.

[112] R. Karanjai, "Targeted phishing campaigns using large scale language models," arXiv preprint arXiv: 2301.00665, 2023.

[113] H. Khan, M. Alam, S. Al-Kuwari, and Y. Faheem, "Offensive AI: Unification of email generation through GPT-2 model with a game-theoretic approach for spear-phishing attacks," in *Proc. Competitive Advantage in the Digital Economy*, 2021, pp. 178–184.

[114] A. M. Shibli, M. M. A. Pritom, and M. Gupta, "AbuseGPT: Abuse of generative AI ChatBots to create smishing campaigns," in *Proc. 12th Int. Symp. Digital Forensics and Security*, San Antonio, USA, 2024, pp. 1–6.

[115] P. V. Falade, "Deciphering ChatGPT's impact: Exploring its role in cybercrime and cybersecurity," *Int. J. Sci. Res. Comput. Sci. Eng.*, vol. 12, no. 2, pp. 15–24, Apr. 2024.

[116] M. Alawida, B. Abu Shawar, O. I. Abiodun, A. Mehmood, A. E. Omolara, and A. K. Al Hwaitat, "Unveiling the dark side of ChatGPT:

Exploring cyberattacks and enhancing user awareness," *Information*, vol. 15, no. 1, p. 27, Jan. 2024.

[117] L. Alotaibi, S. Seher, and N. Mohammad, "Cyberattacks using ChatGPT: Exploring malicious content generation through prompt engineering," in *Proc. ASU Int. Conf. Emerging Technologies for Sustainability and Intelligent Systems*, Manama, Bahrain, 2024, pp. 1304–1311.

[118] T. Susnjak, "ChatGPT: The end of online exam integrity?" arXiv preprint arXiv: 2212.09292, 2022.

[119] C. A. Gao, F. M. Howard, N. S. Markov, E. C. Dyer, S. Ramesh, Y. Luo, and A. T. Pearson, "Comparing scientific abstracts generated by ChatGPT to real abstracts with detectors and blinded human reviewers," *npj Digit. Med.*, vol. 6, no. 1, p. 75, Apr. 2023.

[120] M. J. Israel and A. Amer, "Rethinking data infrastructure and its ethical implications in the face of automated digital content generation," *AI Ethics*, vol. 3, no. 2, pp. 427–439, May 2023.

[121] S. Jalil, S. Raff, T. D. LaToza, K. Moran, and W. Lam, "ChatGPT and software testing education: Promises & perils," in *Proc. IEEE Int. Conf. Software Testing, Verification and Validation Workshops*, Dublin, Ireland, 2023, pp. 4130–4137.

[122] A. B. Armstrong, "Who's afraid of ChatGPT? An examination of ChatGPT's implications for legal writing," *Tech. Rep.*, 2023.

[123] D. R. E Cotton, P. A. Cotton, and J. R. Shipway, "Chatting and cheating: Ensuring academic integrity in the era of ChatGPT," *Innovations Educ. Teach. Int*, vol. 61, no. 2, pp. 228–239, Mar. 2023.

[124] R. J. M. Ventayen, "OpenAI ChatGPT-generated results: Similarity index of artificial intelligence-based contents," in *Soft Computing for Security Applications*, G. Ranganathan, Y. El Allioui, and S. Piramuthu, Eds. Singapore, Singapore: Springer, 2023, pp. 215–226.

[125] M. Khalil and E. Er, "Will ChatGPT get you caught? Rethinking of plagiarism detection," *Proc. 10th Int. Conf. Learning and Collaboration Technologies*, Copenhagen, Denmark, 2023, pp. 475–487.

[126] M. Rezaei, H. Salehi, and O. Tabatabaei, "Uses and misuses of ChatGPT as an AI-language model in academic writing," in *Proc. 10th Int. Conf. Artificial Intelligence and Robotics*, Qazvin, Islamic Republic of, 2024, pp. 256–260.

[127] R. Mustapha, S. N. A. M. Mustapha, and F. W. A. Mustapha, "Students' misuse of ChatGPT in higher education: An application of the fraud triangle theory," *J. Contemp. Soc. Sci. Educ. Stud.*, vol. 4, no. 1, pp. 87–97, Apr. 2024.

[128] M. M. Van Wyk, "Is ChatGPT an opportunity or a threat? Preventive strategies employed by academics related to a GenAI-based LLM at a faculty of education," *J. Appl. Learn. Teach.*, vol. 7, no. 1, pp. 1–35, Feb. 2024.

[129] M. P. Rogers, H. M. Hillberg, and C. L. Groves, "Attitudes towards the use (and misuse) of ChatGPT: A preliminary study," in *Proc. 55th ACM Tech. Symp. Computer Science Education*, Portland, USA, 2024, pp. 1147–1153.

[130] N. M. Mbwambo and P. B. Kaaya, "ChatGPT in education: Applications, concerns and recommendations," *J. ICT Syst.*, vol. 2, no. 1, pp. 107–124, Jun. 2024.

[131] G. Kendall and J. A. T. da Silva, "Risks of abuse of large language models, like ChatGPT, in scientific publishing: Authorship, predatory publishing, and paper mills," *Learn. Publ.*, vol. 37, no. 1, pp. 55–62, Jan. 2024.

[132] M. Dowling and B. Lucey, "ChatGPT for (finance) research: The Bananarama conjecture," *Finance Res. Lett.*, vol. 53, p. 103662, May 2023.

[133] H. H. Thorp, "ChatGPT is fun, but not an author," *Science*, vol. 379, no. 6630, pp. 313–313, Jan. 2023.

[134] M. Liebrenz, R. Schleifer, A. Buadze, D. Bhugra, and A. Smith, "Generating scholarly content with ChatGPT: Ethical challenges for medical publishing," *Lancet Digital Health*, vol. 5, no. 3, pp. e105–e106, Mar. 2023.

[135] F. M. Megahed, Y.-J. Chen, J. A. Ferris, S. Knoth, and L. A. Jones-Farmer, "How generative AI models such as ChatGPT can be (mis)used in SPC practice, education, and research? An exploratory study," *Qual. Eng.*, pp. 287–315, Jun. 2023.

[136] J. Albrecht, E. Kitanidis, and A. Fetterman, "Despite "super-human" performance, current LLMs are unsuited for decisions about ethics and safety," in *Proc. 36th Conf. Neural Information Processing Systems*, New Orleans, USA, 2022.

[137] S. Krügel, A. Ostermaier, and M. Uhl, "The moral authority of ChatGPT," arXiv preprint arXiv: 2301.07098, 2023.

[138] M. Jakesch, A. Bhat, D. Buschek, L. Zalmanson, and M. Naaman, "Co-writing with opinionated language models affects users' views," in *Proc. 2023 CHI Conf. Human Factors in Computing Systems*, Hamburg, Germany, 2023, pp. 111.

[139] The Lancet Digital Health, "ChatGPT: Friend or foe?" *Lancet Digital Health*, vol. 5, no. 3, pp. E102, Mar. 2023.

[140] H. Zohny, J. McMillan, and M. King, "Ethics of generative AI," *J. Med. Ethics*, vol. 49, no. 2, pp. 79–80, Feb. 2023.

[141] W. Chen, F. Wang, and M. Edwards, "Active countermeasures for email fraud," in *Proc. 8th IEEE European Symp. Security and Privacy*, Delft, Netherlands, 2023, pp. 39–55.

[142] J. Hewett and M. Leeke, "Developing a GPT-3-based automated victim for advance fee fraud disruption," in *Proc. IEEE 27th Pacific Rim Int. Symp. Dependable Computing*, Beijing, China, 2022, pp. 205–211.

[143] P. Hacker, A. Engel, and M. Mauer, "Regulating ChatGPT and other large generative AI models," in *Proc. ACM Conf. Fairness, Accountability, and Transparency*, Chicago, USA, 2023, pp. 1112–1123.

[144] M. Y. Vardi, "Who is responsible around here?," *Commun. ACM*, vol. 66, no. 3, p. 5, Feb. 2023.

[145] O. Oviedo-Trespalacios, A. E. Peden, T. Cole-Hunter, A. Costantini, M. Haghani, J. Rod, S. Kelly, H. Torkamaan, A. Tariq, J. D. A. Newton, T. Gallagher, S. Steinert, A. J. Filtness, G. Reniers, "The risks of using ChatGPT to obtain common safety-related information and advice," *Saf. Sci.*, vol. 167, p. 106244, Nov. 2023.

[146] E. Wallace, S. Feng, N. Kandpal, M. Gardner, and S. Singh, "Universal adversarial triggers for attacking and analyzing NLP," in *Proc. Conf. Empirical Methods in Natural Language Processing and the 9th Int. Joint Conf. Natural Language Processing*, Hong Kong, China, 2019, pp. 2153–2162.

[147] H. S. Heidenreich and J. R. Williams, "The earth is flat and the sun is not a star: The susceptibility of GPT-2 to universal adversarial triggers," in *Proc. AAAI/ACM Conf. AI, Ethics, and Society*, New York, USA, 2021, pp. 566–573.

[148] F. Perez and I. Ribeiro, "Ignore previous prompt: Attack techniques for language models," in *Proc. 36th Conf. Neural Information Processing Systems*, New Orleans, USA, 2022, pp. 1–21.

[149] K. Greshake, S. Abdelnabi, S. Mishra, C. Endres, T. Holz, and M. Fritz, "More than you've asked for: A comprehensive analysis of novel prompt injection threats to application-integrated large language models," arXiv preprint arXiv: 2302.12173, 2023.

[150] Y. Liu, G. Deng, Z. Xu, Y. Li, Y. Zheng, Y. Zhang, L. Zhao, T. Zhang, and K. Wang, "A hitchhiker's guide to jailbreaking ChatGPT via prompt engineering," in *Proc. 4th Int. Workshop on Software Engineering and AI for Data Quality in Cyber-Physical Systems/ Internet of Things*, Porto de Galinhas Brazil, 2024, pp. 12–21.

[151] Z. Sha and Y. Zhang, "Prompt stealing attacks against large language models," arXiv preprint arXiv: 2402.12959, 2024.

[152] N. Carlini, F. Tramèr, E. Wallace, M. Jagielski, A. Herbert-Voss, K. Lee, A. Roberts, T. Brown, D. Song, Ú. Erlingsson, A. Oprea, and C. Raffel, "Extracting training data from large language models," in *Proc. 30th USENIX Security Symp*, 2021, pp. 2633–2650.

[153] H.-M. Chu, J. Geiping, L. H. Fowl, M. Goldblum, and T. Goldstein, "Panning for gold in federated learning: Targeted text extraction under arbitrarily large-scale aggregation," in *Proc. 11th Int. Conf. Learning Representations*, Kigali, Rwanda, 2023.

[154] J. Chu, Z. Sha, M. Backes, and Y. Zhang, "Reconstruct your previous conversations! Comprehensively investigating privacy leakage risks in conversations with GPT models," arXiv preprint arXiv: 2402.02987,

2024.

[155] C. Wei, K. Chen, Y. Zhao, Y. Gong, L. Xiang, and S. Zhu, "Context injection attacks on large language models," arXiv preprint arXiv: 2405.20234, 2024.

[156] X. Zhang, Z. Zhang, S. Ji, and T. Wang, "Trojaning language models for fun and profit," in *Proc. IEEE European Symp. Security and Privacy*, Vienna, Austria, 2021, pp. 179–197.

[157] S. Li, H. Liu, T. Dong, B. Z. H. Zhao, M. Xue, H. Zhu, and J. Lu, "Hidden backdoors in human-centric language models," in *Proc. ACM SIGSAC Conf. Computer and Communications Security*, South Korea, 2021, pp. 3123–3140.

[158] X. Pan, M. Zhang, B. Sheng, J. Zhu, and M. Yang, "Hidden trigger backdoor attack on NLP models via linguistic style manipulation," in *Proc. 31st USENIX Security Symp.*, Boston, USA, 2022, pp. 3611–3628.

[159] Y. Huang, T. Y. Zhuo, Q. Xu, H. Hu, X. Yuan, and C. Chen, "Training-free lexical backdoor attacks on language models," in *Proc. ACM Web Conf.*, Austin, USA, 2023, pp. 2198–2208.

[160] H. J. Branch, J. R. Cefalu, J. McHugh, L. Hujer, A. Bahl, D. d. C. Iglesias, R. Heichman, and R. Darwishi, "Evaluating the susceptibility of pre-trained language models via handcrafted adversarial examples," arXiv preprint arXiv: 2209.02128, 2022.

[161] Y. Liu, G. Shen, G. Tao, S. An, S. Ma, and X. Zhang, "Piccolo: Exposing complex backdoors in NLP transformer models," in *Proc. IEEE Symp. Security and Privacy*, San Francisco, USA, 2022, pp. 2025–2042.

[162] D. Kang, X. Li, I. Stoica, C. Guestrin, M. Zaharia, and T. Hashimoto, "Exploiting programmatic behavior of LLMs: Dual-use through standard security attacks," *Proc. 40th Int. Conf. Machine Learning*, Honolulu, USA, 2023.

[163] X. Pan, M. Zhang, S. Ji, and M. Yang, "Privacy risks of general-purpose language models," in *Proc. IEEE Symp. Security and Privacy*, San Francisco, USA, 2020, pp. 1314–1331.

[164] Q. Xu, L. Qu, Z. Gao, and G. Haffari, "Personal information leakage detection in conversations," in *Proc. Conf. Empirical Methods in Natural Language Processing*, 2020, pp. 6567–6580.

[165] J. Huang, H. Shao, and K. C.-C. Chang, "Are large pre-trained language models leaking your personal information?" in *Proc. Association for Computational Linguistics*, Abu Dhabi, United Arab Emirates, 2022, pp. 2038–2047.

[166] B. Jayaraman, E. Ghosh, M. Chase, S. Roy, W. Dai, and D. Evans, "Combing for credentials: Active pattern extraction from smart reply," *Proc. IEEE Symp. Security and Privacy*, San Francisco, USA, 2024, pp. 1443–1461.

[167] N. Lukas, A. Salem, R. Sim, S. Tople, L. Wutschitz, and S. ZanellaBéguelin, "Analyzing leakage of personally identifiable information in language models," in *Proc. IEEE Symp. Security and Privacy*, San Francisco, USA, 2023, pp. 346–363.

[168] F. Mireshghallah, A. Uniyal, T. Wang, D. Evans, and T. Berg-Kirkpatrick, "An empirical analysis of memorization in fine-tuned autoregressive language models," in *Proc. Conf. Empirical Methods in Natural Language Processing*, Abu Dhabi, United Arab Emirates, 2022, pp. 1816–1826.

[169] N. Carlini, D. Ippolito, M. Jagielski, K. Lee, F. Tramer, and C. Zhang, "Quantifying memorization across neural language models," in *Proc. Eleventh Int. Conf. Learning Representations*, Kigali, Rwanda, 2023.

[170] D. Ippolito, F. Tramèr, M. Nasr, C. Zhang, M. Jagielski, K. Lee, C. A. Choquette-Choo, and N. Carlini, "Preventing verbatim memorization in language models gives a false sense of privacy," arXiv preprint arXiv: 2210.17546, 2022.

[171] H. Brown, K. Lee, F. Mireshghallah, R. Shokri, and F. Tramèr, "What does it mean for a language model to preserve privacy?" in *Proc. ACM Conf. Fairness, Accountability, and Transparency*, Seoul, South Korea, 2022, pp. 2280–2292.

[172] J. Mattern, Z. Jin, B. Weggenmann, B. Schoelkopf, and M. Sachan, "Differentially private language models for secure data sharing," in *Proc. Conf. Empirical Methods in Natural Language Processing*, Abu Dhabi, United Arab Emirates, 2022, pp. 4860–4873.

[173] W. Shi, R. Shea, S. Chen, C. Zhang, R. Jia, and Z. Yu, "Just fine-tune twice: Selective differential privacy for large language models," in *Proc. Conf. Empirical Methods in Natural Language Processing*, Abu Dhabi, United Arab Emirates, 2022, pp. 6327–6340.

[174] X. Feng, X. Zhu, Q.-L. Han, W. Zhou, S. Wen, and Y. Xiang, "Detecting vulnerability on IoT device firmware: A survey," *IEEE/CAA J. Autom. Sinica*, vol. 10, no. 1, pp. 25–41, Jan. 2023.

[175] X. Zhu, S. Wen, S. Camtepe, and Y. Xiang, "Fuzzing: A survey for roadmap," *ACM Comput. Surv.*, vol. 54, no. 11s, p. 230, Sept. 2022.

[176] X. Zhu and M. Böhme, "Regression greybox fuzzing," in *Proc. ACM SIGSAC Conf. Computer and Communications Security*, Republic of Korea, 2021, pp. 2169–2182.

[177] D. Su, P. S. Stanimirović, L. B. Han, and L. Jin, "Neural dynamics for improving optimiser in deep learning with noise considered," *CAAI Trans. Intell. Technol.*, vol. 9, no. 3, pp. 722–737, Jun. 2024.

Wei Zhou received the BEng and MEng degrees from Central South University in 2005 and 2008, respectively, all in CS, and the Ph.D. degree from School of Engineering and IT, University of New South Wales, Canberra, Australia in 2016. She is currently a Research Fellow at Swinburne University of Technology. Her research interests include computer networks, network security, and fingerprint biometric.

Xiaogang Zhu (Member, IEEE) received the Ph.D. degree in computer science and engineering at Swinburne University of Technology. Currently, he is a Lecturer at the University of Adelaide and focuses on searching vulnerabilities in programs. He is interested in detecting techniques such as fuzzing, machine learning and symbolic execution. He has published papers on top journals, such as TDSC, and conferences such as CCS, USENIX Security, and ICSE. He also served as a reviewer for many top journals such as TDSC, IoTJ, and CSUR.

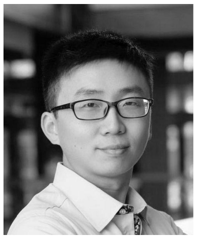
Qing-Long Han (Fellow, IEEE) received the B.Sc. degree in mathematics from Shandong Normal University in 1983, and the M.Sc. and Ph.D. degrees in Control Engineering from East China University of Science and Technology in 1992 and 1997, respectively.

Professor Han is Pro Vice-Chancellor (Research Quality) and a Distinguished Professor at Swinburne University of Technology, Melbourne, Australia. He held various academic and management positions at Griffith University and Central Queensland University, Australia. His research interests include networked control systems, multi-agent systems, time-delay systems, smart grids, unmanned surface vehicles, and neural networks.

Professor Han was awarded the 2024 Dr.-Ing. Eugene Mittelmann Achievement Award (the Highest Award in industrial electronics), the 2021 Norbert Wiener Award (the Highest Award in systems science and engineering, and cybernetics) and the 2021 M. A. Sargent Medal (the Highest Award of the Electrical College Board of Engineers Australia). He was the recipient of the IEEE Systems, Man, and Cybernetics Society Andrew P. Sage Best Transactions Paper Award in 2022, 2020, and 2019, respectively, the IEEE/CAA Journal of Automatica Sinica Norbert Wiener Review Award in 2020, and the IEEE Transactions on Industrial Informatics Outstanding Paper Award in 2020.

Professor Han is a Member of the Academia Europaea (The Academy of Europe). He is a Fellow of The International Federation of Automatic Control (FIFAC), an Honorary Fellow of The Institution of Engineers Australia (HonFIEAust), and a Fellow of Chinese Association of Automation (FCAA). He is a Highly Cited Researcher in both Engineering and Computer Science (Clarivate). He has served as an AdCom Member of IEEE Industrial Electronics Society (IES), a Member of IEEE IES Fellows Committee, a Member of IEEE IES Publications Committee, Chair of IEEE IES Technical Committee on Network-Based Control Systems and Applications, and the Co-Editor-in-Chief of *IEEE Transactions on Industrial Informatics*. He is currently the Editor-in-Chief of *IEEE/CAA Journal of Automatica Sinica* and the Co-Editor of *Australian Journal of Electrical and Electronic Engineering*.

Lin Li received the Ph.D. degree in computer science from Swinburne University of Technology, Melbourne, VIC, Australia, in 2024. He is currently a Lecturer with the Faculty of Science and Engineering, Southern Cross University. And he used to be a Research Fellow with the College of Business and Law, RMIT University. His research interests include applied AI, biometrics, and healthcare.



Xiao Chen is currently a Lecturer at the School of Information and Physical Sciences, the University of Newcastle, Australia. Prior to this, he was a Research Fellow at Monash University, Australia. He received the Ph.D. degree from Swinburne University of Technology, Australia. His research interests include mobile software engineering, with an emphasis on mobile security and quality assurance, and intelligent software engineering (SE4AI, AI4SE). He is broadly interested in applying static code analysis, dynamic program testing, and machine/deep learning techniques to enhance the security and reliability of software systems. His work has been published in top-tier conferences and journals, including ESEC/FSE, ASE, ISSTA, TheWebConf, TOSEM, TSE, TDSC, and TIFS.

Sheng Wen (Senior Member, IEEE) received the Ph.D. degree from Deakin University, Melbourne, in October 2014. Now, he is working full-time as an Associate Professor in Swinburne University of Technology. He managed several research projects in the last few years. He is now the Director of Blockchain Innovation Lab and the Deputy Director of Cybersecurity Lab in Swinburne University. He is also leading a medium-sized research team with co-/supervised Ph.D. students in the system security area. In addition, he has published over 100 high-quality papers, including top conference papers such as papers in IEEE S&P, ACM CCS, NDSS, ICSE and FSE, as well as many papers in IEEE/ACM transactions series journals.

Yang Xiang (Fellow, IEEE) received the Ph.D. degree in computer science from Deakin University, Australia. He is currently a Full Professor and the Dean of Digital Research, Swinburne University of Technology, Australia. His research interests include cyber security, which covers network and system security, data analytics, distributed systems, and networking. In the past 20 years, he has published more than 300 research papers in many international journals and conferences. He is the Editor-in-Chief of the SpringerBriefs on *Cyber Security Systems and Networks*. He serves as the Associate Editor of *IEEE Transactions on Dependable and Secure Computing*, *IEEE Internet of Things Journal*, and *ACM Computing Surveys*. He is the Coordinator, Asia for IEEE Computer Society Technical Committee on Distributed Processing (TCDP), and the Chair of the Australia and New Zealand, IEEE Blockchain Technical Community.