Department: Head

Is ChatGPT Trustworthy Enough? A Review

Guoliang Zhou, Yijia Liu, Zheng Yan
Xidian University

Erol Gelenbe
Institute of Theoretical and Applied Informatics, Polish Academy of Sciences

***Abstract*—ChatGPT, as an advanced model that seamlessly integrates into diverse digital interactions, shows great potential to enhance the performance of consumer technology and reshape its landscape. The critical question of its trustworthiness and associated challenges becomes increasingly prominent. However, the literature still lacks a thorough review to study its trust. This comprehensive review explores whether ChatGPT is trustworthy enough by navigating the multifaceted realm of large language models, placing a significant focus on its exemplar model, ChatGPT. Our exploration traverses the complex interplay of factors impacting user trust, including trust in ChatGPT itself and trust in its utilization and dissemination. By delving into insightful perspectives on the ChatGPT trustworthiness regarding a set of evaluation criteria including fundamental properties, subjective properties, security and privacy, we find that ChatGPT is far from trustworthy based on related literature review. This paper sheds light on the weakness of ChatGPT trust and provides the trajectory of future development in the realm of large language models.**

LARGE LANGUAGE MODELS (LLMs), such as ChatGPT, have emerged as powerful tools in Natural Language Processing (NLP) [1] [2] and Artificial Intelligence (AI) research. These models are built upon deep learning architectures and trained on massive datasets to generate human-like text and perform various NLP tasks. ChatGPT, developed by OpenAI, is one of the most prominent examples of LLMs, known for its ability to generate coherent and contextually relevant responses to user inputs [1].

The widespread adoption of ChatGPT has led to its integration into various consumer devices, applications and services, ranging from virtual assistants and chatbots to content generation, customer services, market analysis, travel planning, art design, and language translation, showing great potential to reshape the landscape of consumer technology. Its versatility and scalability make it a valuable asset in such fields as human company services, education, healthcare, and entertainment at least, but not limited. Without any doubt, ChatGPT is highly related to consumer technology, not only empowering it but also reforming its future.

Since its introduction, researchers have started to explore ChatGPT's applications in consumer devices and services. Zhao et al. [3] discussed the integration of generative AI models like ChatGPT in the field of consumer electronics. Their work highlights the potential of these AI models to enhance user interactions, improve personalization in electronic devices, and

Digital Object Identifier 10.1109/MCE.2022.Doi Number
Date of publication 00 xxxx 0000; date of current version 00 xxxx 0000

support various applications in smart home systems, voice-activated assistants, and wearable technology. Paul et al. [4] discussed how ChatGPT can benefit consumers by providing personalized recommendations, improving customer services and enhancing interactions with smart devices. Bahrini et al. [5] delved into how ChatGPT's capabilities can be leveraged for troubleshooting, customer support, and real-time user interaction, significantly boosting the user experience in technology-driven environments. However, the authors of both articles point to some common shortcomings, such as potential privacy risks, security issues, and response accuracy. These issues are particularly important in the area of consumer technology, where trust and reliability are paramount. Chamola et al. [6] emphasized the role of ChatGPT in improving the usability of devices by offering real-time support, troubleshooting, and personalized content suggestions. The integration of such AI models is seen as key to make consumer electronics more adaptive and user-centric, contributing to the evolution of next-generation devices.

However, the increasing reliance on ChatGPT raises special concerns on its trustworthiness and reliability. A Pew Research Center survey[1] conducted in February 2024 revealed that 23% of U.S. adults have used a chatbot. The survey also explored how Americans might use ChatGPT for various purposes, including work tasks, education, and entertainment. Despite the growing adoption of chatbots for these activities, public confidence in their ability to provide reliable information remains low. Approximately 40% of respondents expressed little to no trust in its offered election-related information from ChatGPT, while only 2% indicated high or substantial trust. Meanwhile, legislative efforts around LLMs are advancing. The European Union's Artificial Intelligence Act (AIA)[2], which regulates AI technologies, came into effect on August 1st, 2024, with provisions set to be gradually implemented over the next 6 to 36 months. The AIA introduces a risk-based regulatory framework that mandates rigorous compliance assessments for high-risk AI systems before their public release. The Act classifies LLMs as high-risk and requires that they meet key standards, including safety, technical robustness, transparency, and controllability. Consequently, ensuring the trustworthiness of ChatGPT through evaluation and analysis becomes critical.

Against this background, some researchers have attempted to study the trustworthiness of ChatGPT in different aspects. Ray [1] identified bias and ethical concerns; stressed the need for addressing them to enhance trustworthiness. Wang et al. [2] focused on the robustness of ChatGPT by examining its performance under adversarial attacks and out-of-distribution scenarios. Their findings reveal vulnerabilities that could undermine ChatGPT trustworthiness. Kocoń et al. [7] highlighted ChatGPT's appropriate performance across tasks, but lack comments on domain-specific trustworthiness. Shen et al. [8] proposed methods for evaluating trustworthiness; underscored the need for rigorous evaluation. Aggarwal [9] reviewed ChatGPT's impact across different domains, noting both its potential and challenges. The study emphasized the importance of building trust in ChatGPT through transparency and consistent performance. Oviedo-Trespalacios et al. [10] required rigorous validation to ensure reliability by examining the risks associated with the security of ChatGPT. Zhou et al. [11] discussed the need for developing trust when integrating models into various sectors. Haleem et al. [12] analyzed the features and challenges of ChatGPT, emphasizing the need to address trust and reliability issues to unlock the potential of ChatGPT. Dwivedi et al. [13] then emphasized addressing trustworthiness for effective use in research, practice, and policy. Salah et al. [14] explored user trust and psychological influences, emphasizing the need to understand trust in order to improve AI design. **Table 1** provides a detailed comparison of our survey with highly-related surveys. As we can see from the table, there is little related work that categorizes trust concerns systematically on ChatGPT and evaluates them with a set of criteria. We work on remedying this neglected aspect.

In this paper, we investigate the trustworthiness of ChatGPT, aiming to help researchers and developers capture the recent advances, open issues and future research directions of ChatGPT trust study. To be specific, we introduce the background of ChatGPT and the concerns on its trust. Then, we propose nine evaluation criteria that a sound LLM should meet. Furthermore, we conduct a thorough review on existing work regarding ChatGPT's performance and trustworthiness by employing the proposed criteria as a measure to study its strengths and weaknesses, focusing on trust.

[1] https://www.pewresearch.org/short-reads/2024/03/26/americans-use-of-chatgpt-is-ticking-up-but-few-trust-its-election-information/

[2] https://eur-lex.europa.eu/resource.html?uri=cellar:e0649735-a372-11eb-9585-01aa75ed71a1.0001.02/DOC_1&format=PDF

Table 1. Comparison of our survey with existing related surveys.

Covered topics	[1]	[2]	[7]	[8]	[9]	[10]	[11]	[12]	[13]	[14]	Our survey
Categorize trust concerns on ChatGPT	·	·	·	·	·	·	●	·	·	·	●
Summarize the attacks against ChatGPT	○	·	○	○	·	·	·	·	·	·	●
Propose evaluation criteria	·	○	·	○	○	·	○	○	○	·	●
Review on trustworthiness of ChatGPT	·	·	·	○	·	○	·	·	○	·	●
●: Fully supported; ○:Partially supported; ·:Not supported											

In the end, based on the review, a series of open issues are identified, in parallel with suggestions on future research directions.

Specifically, the main contributions of this paper can be summarized as follows:

- We discuss the necessity to concern trust in ChatGPT, especially from the view of consumer technology.
- We summarize a set of evaluation criteria that ChatGPT needs to meet in order to improve trusted interactions with human beings.
- We perform a comprehensive review on ChatGPT trust studies and apply the proposed criteria to figure out whether ChatGPT is trustworthy.
- We point out a series of open issues and further suggest future research directions to motivate future development of ChatGPT.

The rest of the paper is organized as follows. Section 2 introduces the preliminaries of ChatGPT and the concerns on its trust. In Section 3, we provide a set of criteria for evaluating the performance and trustworthiness of ChatGPT. In Section 4, we present a comprehensive review on ChatGPT's trustworthiness, followed by open issues and future research directions of ChatGPT trust studies in Section 5. Finally, a conclusion is drawn in the last section.

BACKGROUND KNOWLEDGE

In this section, we introduce the background of ChatGPT and the concerns on its trust.

ChatGPT

This subsection introduces the modeling principles and various applications of ChatGPT. Then, we outline the potential challenges and issues associated with the use of ChatGPT.

Modeling principles and applications: ChatGPT, a variant of the Generative Pre-trained Transformer (GPT) series developed by OpenAI, embodies several key modeling principles including Transformer Architecture, Self-Supervised Learning, Fine-Tuning, Zero-Shot Learning and Scalability, which underpin its functionality and effectiveness in generating human-like text. Due to its versatility and generality in natural language understanding and generation, ChatGPT finds applications in a wide range of domains and use cases. In customer service and support, ChatGPT can be used as a virtual assistant to handle customer queries, provide information, and assist in troubleshooting tasks. In education, ChatGPT can serve as a tutor, providing personalized learning assistance and answering students' questions in real-time. In content creation, ChatGPT can automate the generation of articles, stories, and social media posts, saving time and efforts for content creators. Furthermore, ChatGPT has applications in healthcare, where it can assist medical professionals in such tasks as medical coding, patient communication, and clinical decision support. Overall, the modeling principles and applications of ChatGPT reflect its potential to revolutionize human-AI interactions and drive innovation across multiple domains, especially in consumer electronics and technology.

Challenges and issues: Despite its remarkable capabilities, ChatGPT still confronts with several challenges and issues that warrant the attention of researchers and practitioners. These challenges arise from various aspects of the model's architecture, training methodology, and deployment scenarios, and they have implications for its reliability, safety, and ethical use.

1) *The generation of biased or inappropriate responses:* Due to the nature of its training data, which may contain biased or sensitive content, the model may inadvertently learn and reproduce stereotypes, prejudices, or offensive language in its generated text. Addressing this challenge requires techniques for data cleaning, bias detection and mitigation, as well as the development of ethical guidelines and safeguards to ensure responsible AI deployment.
2) *The lack of interpretability in responses:* The ChatGPT model's complex architecture and internal representations make it difficult to understand how it arrives at certain outputs, leading to

concerns about accountability, transparency, and trustworthiness.

3) *Exposure to adversarial attacks:* Adversaries can manipulate or perturb input text to trigger undesirable behavior in the model, such as generating false or misleading responses.
4) *Ethical considerations:* These considerations particularly fall into data privacy, consent and protection. For example, deploying the model in sensitive environments such as healthcare or legal environments raises concerns about the confidentiality and security of user data.

Trust Concerns

Trust in ChatGPT is a critical concern that influences its adoption and acceptance in various applications. Trust concerns can be broadly categorized into issues related to the inherent characteristics of ChatGPT and those related to its usage and dissemination.

Trust in ChatGPT itself: Trust in the ChatGPT itself is referred to as "objective trust", which reflects the inherent properties of ChatGPT and is independent of user perception or application context. ChatGPT, as a LLM, possesses inherent objective capabilities that are critical to its functionality and reliability, such as textual analysis and interpretation capabilities, multi-domain, multi-perspective applicability, and error correction.

These objective capabilities are inextricably linked to trust. In other words, the model's text analysis and interpretation capabilities, and error correction mechanisms contribute to a foundation of trustworthiness. Users can rely on the model to provide accurate and relevant information, to withstand and adapt to adversarial conditions, and to correct its own errors, thereby increasing their confidence in its outputs.

Trust in the utilization and dissemination of ChatGPT: Trust in the use and dissemination of ChatGPT is referred to as "subjective trust", which reflects how users perceive the reliability and appropriateness of ChatGPT in a given application context.

Firstly, a positive experience of interacting with ChatGPT leads to easy acceptance and recommendation of it, thereby increasing its trustworthiness to a wide group of people. Secondly, the interpretability of the model is crucial during usage, as users need to understand and rationalise the responses generated. If the users can trace the logic behind the results of ChatGPT output, their confidence in the model should increase.

In addition, security and privacy is another important aspect that affects trust in the use and dissemination of ChatGPT. Users need to be assured that their interactions with the model are secure and that their data are kept confidential. Finally, user feedback loops are an important mechanism for building trust. ChatGPT demonstrates its commitment to quality and reliability by continuously learning from user interactions and incorporating feedback to improve its performance and reduce bias and unfairness. Over time, this iterative process of improvement helps in maintaining and increasing user confidence.

EVALUATION CRITERIA

In this section, we summarize nine important evaluation criteria that a sound LLM needs to meet in order to ensure the trustworthiness of ChatGPT. As shown in **Figure 1**, the proposed evaluation criteria on ChatGPT trust contain the following three main aspects: fundamental properties, subjective properties, and security and privacy.

Fundamental Properties

ChatGPT should first offer the intrinsic quality attributes of LLM, including accuracy, adaptability, generality and multifacetedness. This type of criteria highly relates to the trust in ChatGPT itself.

Accuracy: This is a very fundamental and important criterion for determining whether ChatGPT is correct in answering objective questions such as multiple choice and true/false questions [1], [8]. It simply and directly tells a user how well the model can correctly answer objective questions. Therefore, we take accuracy as the first evaluation criterion.

Adaptability: This criterion refers to the ability to perform good modification and self-adaptation in response to feedback and errors [1], [8]. Due to the stochastic nature of deep learning, errors are inevitable. However, users hope that these errors should be resolved within a short period of time after they are detected. Therefore, ChatGPT should be adaptive.

Generality: An LLM with a high degree of generalization can show good performance in a variety of tasks in different domains, rather than being limited to a specific domain. This means that the model can be applied to various application scenarios, thus increasing its usefulness and practical values. Therefore, a good LLM (e.g., ChatGPT) needs to have sound generality.

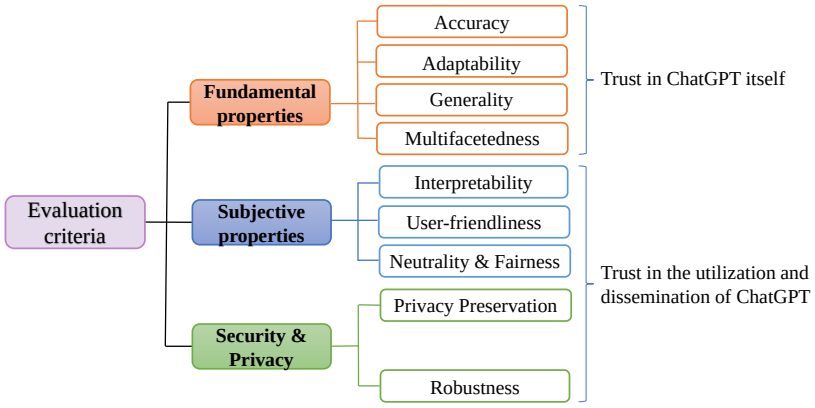

Figure 1. Taxonomy of evaluation criteria on ChatGPT trust.

Multifacetedness: This criterion refers to the comprehensive and diverse analyses or solutions to a problem from multiple directions or perspectives. Such multifacetedness can help users understand the problem in a good way in order to provide flexibility in decision-making or analysis process. Thus, multifacetedness is important for a good LLM.

Subjective Properties

ChatGPT secondly needs to support subjective properties that influence user acceptance on ChatGPT, including interpretability, user-friendliness, and neutrality and fairness. This type of criteria highly relates to the trust in the utilization and dissemination of ChatGPT.

Interpretability: This criterion refers to the ability of ChatGPT to interpret the output logic on its own so that users can understand how it is generated [8]. Knowledge of a LLM's inputs, outputs, and operational mechanisms influences user confidence of usage and user experiences. Therefore, this criterion relates to the acceptance of ChatGPT, thus should be considered.

User-friendliness: This criterion means that the interaction between the model and the user requires free interaction logic and gives good visual feedback to a user. ChatGPT is a "chatter" that communicates with the user, and needs to have human-like characteristics. However, previous LLMs are not user-friendly enough in this regard, and often require specific question-and-answer formats. This gives the user bad experiences and negatively affected ChatGPT usage. Therefore, user-friendliness is also a criterion that should be considered.

Neutrality and fairness: This criterion refers to neutrality and fairness in all matters and avoidance of making biased statements. Depending on the training dataset, ChatGPT may be biased against certain cultural and linguistic groups, resulting in biased or inappropriate responses, and may even generate harmful contents, such as hate speech or fake news. This affects the emotional bias of users to a greater or lesser extent. To address the biases, the developers of ChatGPT (i.e., OpenAI) need to apply a diverse training dataset, including incorporating different languages and cultures.

Security and Privacy

Finally, ChatGPT should support Security and Privacy, including privacy preservation and robustness. This type of criteria is also related to the trust in the utilization and dissemination of ChatGPT.

Privacy preservation: The use of LLMs such as ChatGPT requires consideration of privacy preservation, as these models may process large amounts of user data, including personal information, sensitive information, etc. The leakage of user information may lead to serious privacy violations, undermine user rights and trust, and even violate laws and regulations. Therefore, to ensure privacy of users, it is important to adopt appropriate privacy preservation measures, such as data anonymization and encryption technologies, in the design and use of ChatGPT.

Robustness: This criterion refers to the ability to detect and tolerate false input from malicious attacks, as well as the ability to deal with them appropriately [15]. Due to the high-dimensional input space and the lack of robustness of LLMs, the privacy, security, and trust of ChatGPT are vulnerable to not only adversarial attacks [2], [8], but also jailbreaks [16] and other attacks. To tackle the critical challenge

Table 2. Four types of attacks on ChatGPT.

Attack	Features	Results	Example
Adversarial Attacks	Exploits high-dimensional input space, slight textual or semantic perturbations.	Incorrect or unintended responses, misinformation.	[13], [14], [17]
Jailbreaks	Circumvents safety mechanisms, manipulates instructions or context.	Generation of harmful, biased, or restricted contents.	[8], [18]
Backdoor Attacks	Embeds hidden triggers during training, specific inputs with triggers activate malicious behavior.	Unauthorized or malicious behavior when specific triggers are used.	[19]
Data Poisoning Attacks	Corrupts training data, introduces biased or malicious data.	Degraded performance, biased outputs, spread of misinformation.	[20]

of defending against diverse attacks, ChatGPT must demonstrate resilience across multiple types of threats. At a minimum, it should be able to withstand four primary categories of attacks: adversarial attacks, jailbreak attacks, backdoor attacks, and data poisoning attacks. These classifications are based on the distinct ways each attack compromises the model's integrity, security, and trustworthiness. This framework enables a systematic evaluation of the model's vulnerabilities, ranging from manipulation of responses through jailbreak techniques, to exploitation of input-output processes with adversarial attacks, covert insertion of malicious triggers through backdoor attacks, and tampering with training data through data poisoning. Notably, the adversarial attacks focus on fooling the model's underlying logic with manipulated inputs, while the jailbreak attacks are designed to bypass specific behavioral constraints and ethical guidelines imposed on the model. And the backdoor attacks involve a hidden trigger that only activates under specific conditions, while data poisoning attacks cause widespread degradation in performance. Each type of attack reveals specific risks that can undermine the model's reliability and security in various operational environments. By categorizing attacks in this way, we gain a clear understanding of the challenges that LLMs like ChatGPT must overcome to function securely and dependably in real-world applications. The detailed description of these four types of attacks is provided in **Table 2**.

REVIEW ON TRUSTWORTHINESS OF CHATGPT

In this section, we critically review 22 relevant research papers on ChatGPT trust from 2022 to 2024, focusing on the performance of ChatGPT in terms of trustworthiness. According to the dimension of trust, we divide the related works into two categories, i.e., trust in the model's intrinsic attributes, and trust in the model's utilization and dissemination. To evaluate the first category, we use fundamental properties as the evaluation criteria. For the second category, we use subjective properties, security and privacy as the evaluation criteria.

Trust in the Model's Intrinsic Attributes

In this section, we evaluate the trustworthiness of ChatGPT using four criteria of fundamental properties.

For the accuracy criterion, Ray [1] described the limitations and key challenges of ChatGPT, including accuracy, but not in exhaustive detail because it primarily emphasizes various applications and ethical considerations of the model. Wang et al. [2] evaluated ChatGPT accuracy indirectly by measuring ChatGPT's ability to handle unexpected input. The results suggest that ChatGPT can maintain a reasonable level of accuracy in most scenarios, but its performance may degrade under specific adversarial conditions. Kocoń et al. [7] critically examined ChatGPT's performance across various tasks, and concluded that ChatGPT's accuracy only meets the requirements for general use, and does not excel on specialized tasks. Shen et al. [8] specifically measured the reliability and accuracy of ChatGPT's output through a quantitative evaluation to determine where ChatGPT's response is consistently accurate/inaccurate. Experimental results show that ChatGPT has high accuracy for common queries, but struggles with more complex or nuanced questions. Aggarwal [9] focused on evaluating the accuracy of information extraction by ChatGPT. The paper concludes that although ChatGPT has strong information extraction capabilities, its accuracy is not infallible, especially when dealing with more complex or contextual queries. Johnson et al. [21] evaluated the accuracy of ChatGPT in medical contexts and found that it sometimes provides incorrect or misleading medical advice. Li et al. [19] discussed the impact of backdoor

attacks on the accuracy of ChatGPT, and demonstrated that such attacks can significantly degrade the accuracy of the model. By reviewing the above papers, we can find that ChatGPT is generally accurate, but may be inaccurate in special cases, especially under attacks.

For the adaptability criterion, Ray [1] validated ChatGPT by testing ChatGPT's ability to effectively perform different functions. The results show that ChatGPT has significant adaptability, but domain-specific tasks may require specialized models. Kasneci et al. [22] discussed the adaptability of ChatGPT in educational settings, pointing to its potential for personalized learning. Rozado [23] investigated the adaptability of ChatGPT in the generation of politically neutral responses, revealing inherent biases that affect its adaptability in this context. In summary, we can see that ChatGPT has significant adaptability across different tasks and domains. However, issues such as self-contradiction and fabrication in specific applications (e.g., education, especially customized education) highlight the areas where adaptability should be improved. In addition, we notice that there are fewer studies focusing on the adaptability of ChatGPT.

For the generality criterion, Ray [1] discussed the generality of ChatGPT in different domains. Kocoń et al. [7] found that ChatGPT is effective in performing a variety of tasks in different domains. Aggarwal [9] investigated the effects of ChatGPT across multiple fields. All of them point to the generality and usefulness of ChatGPT in diverse domains. Dwivedi et al. [13] discussed the interdisciplinary applications of ChatGPT. While Tlili et al. [24] as well as Kasneci et al. [22] explored the applications of ChatGPT in education. They focused on the potential and development of ChatGPT in the educational sector. Moreover, George [25] investigated the impact of ChatGPT in different business domains. To sum up, the above papers discuss the issue of generality of ChatGPT to some extent, highlighting the wide range of applications and multifunctionality of ChatGPT, thus it supports generality.

For the multifacetedness criterion, relevant studies and experimental programs are almost non-existent. We believe that in this regard, researchers and users are more interested in the generality of ChatGPT than in the multifacetedness of the answers, since most of the time people only need one answer.

Trust in the Model's Utilization and Dissemination

In addition to considering objective factors such as the fundamental properties of the ChatGPT model itself, we also need to consider the subjective factors that affect user confidence. In what follows, we review related works that investigate the subjective properties of ChatGPT trust.

For the interpretability criterion, Ray [1] only emphasized the need for interpretability in AI models but did not provided an effective solution for improving the interpretability of ChatGPT. Kocoń et al. [7] discussed the limitations of ChatGPT in terms of interpretability, summarising the challenges associated with transparency and interpretability of the model. Li et al. [26] provided a detailed evaluation regarding the interpretability of ChatGPT. In particular, they highlighted the need for transparent and interpretable outputs in terms of information extraction capabilities. Dwivedi et al. [13] investigated the challenges of transparency and interpretability in generative AI models and pointed the need for improvements in this area. Based on our review, we find that although many papers mention the importance of interpretability, they do not provide a comprehensive discussion or solution. Therefore, future research needs to delve deeper into how to develop methods to make the output of ChatGPT interpretable.

For the user-friendliness criterion, Kocoń et al. [7] considered the usability of ChatGPT in different contexts, but did not analyse specific user experience aspects in depth. Aggarwal [9] discussed user experience and interaction with ChatGPT, but did not delve into specific user interface design and detailed usability analysis. Dwivedi et al. [13] considered user perception and confidence in using ChatGPT, but lacked a detailed analysis of specific user experience. Tlili et al. [24] discussed the usability of ChatGPT in educational settings, emphasizing its potential to enhance the learning experience. Kasneci et al. [22] investigated the usability and benefits, but did not analyze specific aspects of user interaction design in depth. In summary, while some papers discuss user-friendliness, they tend to focus on user confidence and perception rather than detailed usability studies or specific user interface design. Therefore, more detailed analysis and empirical studies are needed to improve this aspect of study.

For the neutrality and fairness criterion, Ray [1]

discussed biases in ChatGPT, including racial and gender biases. He emphasized the need for ongoing research to reduce these biases. Zhuo et al. [27] discussed biases in ChatGPT, particularly when subjected to adversarial jailbreak attacks which reveal embedded biases. Rozado [23] highlighted political biases in ChatGPT responses and analyzed the tendency towards specific political orientations. Hosseini [28] discussed potential biases in peer review processes when using ChatGPT, emphasizing its amplification of existing biases. Deldjoo [29] investigated fairness in ChatGPT responses and the impact of explainable-guided prompts on reducing biases. In summary, we can see that several papers have addressed bias in ChatGPT, especially political bias and peer review bias. However, the analysis is often not comprehensive and additional empirical studies are expected to fully understand and address these issues.

Finally we review related works to investigate the security and privacy of ChatGPT to see if ChatGPT meets these two criteria.

For the privacy preservation criterion, Li et al. [18] explored privacy vulnerabilities in ChatGPT through jailbreak attacks which can extract sensitive information. Wu et al. [30] discussed variety of privacy concerns, including data leakage and unauthorized data access. Li et al. [19] analyzed backdoor attacks on ChatGPT that could compromise user privacy. Huang et al. [20] provided an in-depth discussion of security and privacy concerns, including data handling practices and regulatory implications. In summary, several papers take the privacy preservation of ChatGPT into consideration, indicating that user privacy could be leaked when using ChatGPT.

Regarding the robustness criterion, Wang et al. [2] investigated the performance of ChatGPT under adversarial attacks and out-of-distribution scenarios. Zhuo et al. [27] investigated the robustness of ChatGPT through adversarial jailbreak attempts, and found vulnerabilities. Xie et al. [16] proposed self-reminder techniques to defend against jailbreak attacks, enhancing robustness. Liu et al. [17] reviewed various adversarial attacks on ChatGPT and proposed mitigation strategies. To summarize, several innovative techniques are proposed to enhance robustness, but obviously further research is needed to address new and evolving threats.

The fulfillment of ChatGPT on all evaluation criteria is summarized in **Table 3** based on our review. We can see that ChatGPT performs well in terms of generality, because the LLM generally takes into account applications in different domains. However, other criteria such as accuracy, user-friendliness, neutrality and fairness, privacy preservation, and robustness are not satisfactorily met, even though there are several papers discussing relevant solutions. This fact requires researchers to further investigate ChatGPT trust solutions. In addition, multifacetedness seems ignored totally, so a comprehensive and diversified analysis or solution of the problem from multiple directions or perspectives deserves our attention and efforts. In summary, ChatGPT is still far from being fully trustworthy, and there is a lot of room for development in the future.

OPEN ISSUES AND FUTURE DIRECTIONS

Based on the above in-depth review, in this section we first analyze the open issues faced by ChatGPT in terms of trust, and then suggest some improvements in order to guide future research directions on ChatGPT trust.

Trust Evaluation

In order to study the trustworthiness of ChatGPT comprehensively, it becomes essential to evaluate its trust. We would like to advocate that it is essential to specify a common set of standards, i.e., a list of criteria like ours, to clarify what factors should be examined in order to evaluate the trust of ChatGPT, what tests should be conducted before applying it into consumer devices and services, and what policies should be followed in order to ensure a reliable integration of ChatGPT. Furthermore, we should research proper methods for ChatGPT trust evaluation. For example, we can use the experimental methods given in [8], [31], [32] and combine them with our proposed evaluate criteria. For Fundamental Properties, as shown in [8], we can test the correctness of ChatGPT and its performance in different domain scenarios by using highly qualified datasets; for Subjective Properties, similar to [31], we can obtain statistical results in the form of questionnaires, including whether users obtain good experiences in using ChatGPT to strengthen their confidence and acceptance; for Security and Privacy, as given in [32], we can use proper datasets, e.g., AdvGLUE and AdvGLUE ++, to comprehensively evaluate its robustness and privacy protection through system prompt or user prompt, etc. We should motivate future research of trust evaluation on ChatGPT and

Table 3. Criteria satisfaction level.

Criterion	Reference	Criterion Satisfaction Level
Accuracy	[1], [2], [7], [8], [9], [21], [19]	○
Adaptability	[1], [22] , [23]	○
Interpretability	[1], [7], [26], [13]	○
Generality	[1], [7], [9], [13], [24], [25], [22]	●
User-friendliness	[7], [9], [13], [24], [22]	○
Multifacetedness	[8]	·
Neutrality and Fairness	[1], [2], [27], [23], [28], [29]	○
Privacy Preservation	[18], [30], [19], [20]	○
Robustness	[2], [8], [27], [16], [17]	○
●: Meet the criterion; ○: Partially meet; ·: Not meet		

deepen domain-specific evaluation, at least from the perspective of evaluation standard.

Accuracy and Multifacetedness

ChatGPT demonstrates considerable accuracy in generating coherent and contextually appropriate responses across a wide range of topics. However, limitations remain, particularly in specialized domains where the model of ChatGPT could produce factually incorrect or misleading information. Enhancing data quality and incorporating real-time fact-checking mechanisms are essential to further improve accuracy. Multifacetedness is overlooked in existing works. ChatGPT excels at providing a multifaceted response, demonstrating its ability to solve problems from multiple perspectives. However, the model's ability to generate subtle and in-depth responses in specialized contexts is limited, sometimes resulting in superficial answers or an inability to maintain answer consistency. Future improvements should focus on enhancing the model's depth and consistency in specialized contexts to maximize its multifaceted response capabilities.

Robustness and Privacy Preservation

Robustness and privacy preservation are often neglected by existing studies. The vulnerability of ChatGPT to adversarial attacks and malicious manipulation poses significant security risks. Future research efforts should focus on enhancing the model's robustness against such attacks and developing mechanisms, e.g., differential privacy techniques and robust encryption methods, for preserving the integrity and privacy of conversational interactions.

Interpretability

There are relatively few existing studies on interpretability of ChatGPT. Improving the interpretability of ChatGPT is essential for building up user confidence and understanding model outputs. Future research should explore techniques, e.g., attention mechanisms and layer-wise relevance propagation, for generating interpretable responses and providing transparent explanations of the model's reasoning processes.

Domain-Specific Adaptation

While ChatGPT demonstrates general-purpose language understanding, its effectiveness in domain-specific contexts varies. Future research should investigate such techniques as transfer learning and domain-specific pre-training for domain adaptation and fine-tuning to improve the model's performance in specialized domains such as healthcare, finance, and law.

CONCLUSION

This paper presents a systematic review on the trustworthiness of ChatGPT. First, we outlined the potential challenges and issues associated with the use of ChatGPT and proposed our trust concerns on it. Then, we summarized a set of evaluation criteria that a trustworthy LLM should meet. We conducted a thorough review of existing researches on ChatGPT trust, using the proposed criteria to judge whether ChatGPT can satisfy them and to evaluate its strength and weakness. Finally, our review revealed that ChatGPT is still far from being fully trustworthy. In response, we identified several open issues and suggested future research directions to advance efforts in enhancing the trustworthiness of ChatGPT.

ACKNOWLEDGMENTS

This work is supported in part by the National Natural Science Foundation of China under Grants U23A20300 and 62072351; in part by the Key Research Project of Shaanxi Natural Science Foundation under Grant 2023-JC-ZD-35; in part by the Concept Verification Funding of Hangzhou Institute of Technology of Xidian University under Grant

GNYZ2024XX007, in part by the 111 Project under Grant B16037, and (for Erol Gelenbe) by the Horizon Europe Research Project DOSS Contract No. HE – 101120270.

REFERENCES

1. P. P. Ray, "Chatgpt: A comprehensive review on background, applications, key challenges, bias, ethics, limitations and future scope," *Internet of Things and Cyber-Physical Systems*, vol. 3, pp. 121–154, 2023.
2. J. Wang, X. Hu, W. Hou, H. Chen, R. Zheng, Y. Wang, L. Yang, W. Ye, H. Huang, X. Geng *et al.*, "On the robustness of chatgpt: An adversarial and out-of-distribution perspective," *Data Engineering*, pp. 48–62, 2024.
3. M. Zhao, W. Meng, B. Liu, and Y. Yang, "Guest editorial of the special section on generative artificial intelligence with applications on consumer electronics," *IEEE Transactions on Consumer Electronics*, vol. 70, no. 2, pp. 4955–4957, 2024.
4. J. Paul, A. Ueno, and C. Dennis, "Chatgpt and consumers: Benefits, pitfalls and future research agenda," pp. 1213–1225, 2023.
5. A. Bahrini, M. Khamoshifar, H. Abbasimehr, R. J. Riggs, M. Esmaeili, R. M. Majdabadkohne, and M. Pasehvar, "Chatgpt: Applications, opportunities, and threats," in *2023 Systems and Information Engineering Design Symposium (SIEDS)*. IEEE, 2023, pp. 274–279.
6. V. Chamola, S. Sai, R. Sai, A. Hussain, and B. Sikdar, "Generative ai for consumer electronics: Enhancing user experience with cognitive and semantic computing," *IEEE Consumer Electronics Magazine*, 2024.
7. J. Kocoń, I. Cichecki, O. Kaszyca, M. Kochanek, D. Szydło, J. Baran, J. Bielaniewicz, M. Gruza, A. Janz, K. Kanclerz *et al.*, "Chatgpt: Jack of all trades, master of none," *Information Fusion*, vol. 99, p. 101861, 2023.
8. X. Shen, Z. Chen, M. Backes, and Y. Zhang, "In chatgpt we trust? measuring and characterizing the reliability of chatgpt," *arXiv preprint arXiv:2304.08979*, 2023.
9. S. Aggarwal, "A review of chatgpt and its impact in different domains," *International Journal of Applied Engineering Research*, vol. 18, no. 2, pp. 119–123, 2023.
10. 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 *et al.*, "The risks of using chatgpt to obtain common safety-related information and advice," *Safety Science*, vol. 167, p. 106244, 2023.
11. C. Zhou, Q. Li, C. Li, J. Yu, Y. Liu, G. Wang, K. Zhang, C. Ji, Q. Yan, L. He *et al.*, "A comprehensive survey on pretrained foundation models: A history from bert to chatgpt," *arXiv preprint arXiv:2302.09419*, 2023.
12. A. Haleem, M. Javaid, and R. P. Singh, "An era of chatgpt as a significant futuristic support tool: A study on features, abilities, and challenges," *BenchCouncil Transactions on Benchmarks, Standards and Evaluations*, vol. 2, no. 4, p. 100089, 2022.
13. Y. K. Dwivedi, N. Kshetri, L. Hughes, E. L. Slade, A. Jeyaraj, A. K. Kar, A. M. Baabdullah, A. Koohang, V. Raghavan, M. Ahuja *et al.*, "Opinion paper:"so what if chatgpt wrote it?" multidisciplinary perspectives on opportunities, challenges and implications of generative conversational ai for research, practice and policy," *International Journal of Information Management*, vol. 71, p. 102642, 2023.
14. M. Salah, H. Alhalbusi, M. M. Ismail, and F. Abdelfattah, "Chatting with chatgpt: decoding the mind of chatbot users and unveiling the intricate connections between user perception, trust and stereotype perception on self-esteem and psychological well-being," *Current Psychology*, vol. 43, no. 9, pp. 7843–7858, 2024.
15. Y. Liu, J. Wang, Z. Yan, Z. Wan, and R. Jäntti, "A survey on blockchain-based trust management for internet of things," *IEEE internet of Things Journal*, vol. 10, no. 7, pp. 5898–5922, 2023.
16. Y. Xie, J. Yi, J. Shao, J. Curl, L. Lyu, Q. Chen, X. Xie, and F. Wu, "Defending chatgpt against jailbreak attack via self-reminders," *Nature Machine Intelligence*, vol. 5, no. 12, pp. 1486–1496, 2023.
17. B. Liu, B. Xiao, X. Jiang, S. Cen, X. He, and W. Dou, "Adversarial attacks on large language model-based system and mitigating strategies: A case study on chatgpt," *Security and Communication Networks*, vol. 2023, no. 1, p. 8691095, 2023.
18. H. Li, D. Guo, W. Fan, M. Xu, J. Huang, F. Meng, and Y. Song, "Multi-step jailbreaking privacy attacks on chatgpt," in *Findings of the Association for Computational Linguistics: EMNLP 2023*, 2023, pp. 4138–4153.
19. J. Li, Y. Yang, Z. Wu, V. V. Vydiswaran, and C. Xiao, "Chatgpt as an attack tool: Stealthy textual backdoor attack via blackbox generative model trigger," in *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational*

Linguistics: Human Language Technologies (Volume 1: Long Papers), 2024, pp. 2985–3004.
20. K. Huang, F. Zhang, Y. Li, S. Wright, V. Kidambi, and V. Manral, "Security and privacy concerns in chatgpt," in *Beyond AI: ChatGPT, Web3, and the Business Landscape of Tomorrow*. Springer, 2023, pp. 297–328.
21. D. Johnson, R. Goodman, J. Patrinely, C. Stone, E. Zimmerman, R. Donald, S. Chang, S. Berkowitz, A. Finn, E. Jahangir *et al.*, "Assessing the accuracy and reliability of ai-generated medical responses: an evaluation of the chat-gpt model," *Research Square*, 2023.
22. E. Kasneci, K. Seßler, S. Küchemann, M. Bannert, D. Dementieva, F. Fischer, U. Gasser, G. Groh, S. Günnemann, E. Hüllermeier *et al.*, "Chatgpt for good? on opportunities and challenges of large language models for education," *Learning and Individual Differences*, vol. 103, p. 102274, 2023.
23. D. Rozado, "The political biases of chatgpt," *Social Sciences*, vol. 12, no. 3, p. 148, 2023.
24. A. Tlili, B. Shehata, M. A. Adarkwah, A. Bozkurt, D. T. Hickey, R. Huang, and B. Agyemang, "What if the devil is my guardian angel: Chatgpt as a case study of using chatbots in education," *Smart Learning Environments*, vol. 10, no. 1, p. 15, 2023.
25. A. S. George and A. H. George, "A review of chatgpt ai's impact on several business sectors," *Partners Universal International Innovation Journal*, vol. 1, no. 1, pp. 9–23, 2023.
26. B. Li, G. Fang, Y. Yang, Q. Wang, W. Ye, W. Zhao, and S. Zhang, "Evaluating chatgpt's information extraction capabilities: An assessment of performance, explainability, calibration, and faithfulness," *arXiv preprint arXiv:2304.11633*, 2023.
27. 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.
28. M. Hosseini and S. P. Horbach, "Fighting reviewer fatigue or amplifying bias? considerations and recommendations for use of chatgpt and other large language models in scholarly peer review," *Research Integrity and Peer Review*, vol. 8, no. 1, p. 4, 2023.
29. Y. Deldjoo, "Fairness of chatgpt and the role of explainable-guided prompts," *arXiv preprint arXiv:2307.11761*, 2023.
30. X. Wu, R. Duan, and J. Ni, "Unveiling security, privacy, and ethical concerns of chatgpt," *Journal of Information and Intelligence*, vol. 2, no. 2, pp. 102–115, 2024.
31. T. H. Baek and M. Kim, "Is chatgpt scary good? how user motivations affect creepiness and trust in generative artificial intelligence," *Telematics and Informatics*, vol. 83, p. 102030, 2023.
32. B. Wang, W. Chen, H. Pei, C. Xie, M. Kang, C. Zhang, C. Xu, Z. Xiong, R. Dutta, R. Schaeffer *et al.*, "Decodingtrust: A comprehensive assessment of trustworthiness in gpt models." in *NeurIPS*, 2023.

Guoliang Zhou received his B.S. degree in the School of Mathematics and Statistics from Xidian University in 2022. He is currently pursuing the master degree with the HangZhou Institute of Technology, Xidian University. His research interests are machine learning, artificial intelligence and blockchain. Contact him at 623984526@qq.com.

Yijia Liu received the B.S. degree in the School of Computer Science & Technology from Soochow University in 2021. She is currently pursuing the PhD degree with the School of Cyber Engineering in Xidian University. Her research interests are in machine learning and large language models. Contact her at liuyijia42@foxmail.com.

Zheng Yan received the doctor of science in technology in electrical engineering from Helsinki University of Technology, Helsinki, Finland, in 2007. She is currently a Huashan distinguished professor at the School of Cyber Engineering, Xidian University, China. She is a Fellow of IEEE. Contact her at zyan@xidian.edu.cn.

Erol Gelenbe received his Ph.D. degree in Electrical Engineering from the Tandon School, New York University, and the D.Sc. degree in Mathematical Sciences from Sorbonne University, Paris. He is currently a Professor at the Institute of Theoretical and Applied Informatics, Polish Academy of Sciences, a Visiting Professor at Kings College London, and a Researcher with the I3S CNRS Laboratory, University Côte d'Azur, Nice. He is a Life Fellow of IEEE and of ACM. Contact him at seg@iitis.pl