Proceedings of the 5th International Conference on Computing and Data Science
DOI: 10.54254/2755-2721/21/20231151

The information safety and ethical issues about ChatGPT

Linkun Li

21century international school, Beijing Haidian district, China, 100142

2152054919@qq.com

Abstract. With the development of artificial intelligence, ChatGPT technology has been widely used in daily life, bringing people a lot of convenience. However, the use of these technologies also brings some privacy and security risks. This paper focuses on the ethical and information security issues involved in ChatGPT technology, and propose corresponding preventive measures and solutions. In the initial stages of ChatGPT's development, we need to take full advantage of its benefits in terms of increased longevity and productivity, while developing stricter management regulations and usage specifications. Finally, this paper puts forward some measures to strengthen the information security and ethical protection of ChatGPT and other natural language generation models from multiple aspects of technology and ethics. These include establishing more secure data protection mechanisms, strengthening the security design of models, and establishing relevant ethical codes and regulations. At the same time, strengthening the research and supervision of natural language generation models such as ChatGPT can better promote their healthy development. Here are some useful references for organizations and individuals using ChatGPT technology.

Keywords: ChatGPT using safety, risk & ethics, ChatGPT features.

1. Introduction

ChatGPT is an intelligent conversation system based on the GPT-3.5 models which was available on November 30, 2022. ChatGPT has better natural language understanding and generation skills than previous conversational bots, making its answers clearer and more detailed, and mimicking real conversation to a large extent. In just two months since its launch, ChatGPT has surpassed 100 million users, making it one of the fastest growing and most widely watched apps of all time [1].

At present, Chatbot technology continues to improve and innovate, and is becoming one of the important fields of artificial intelligence application. GPT (Generative Pre-trained Transformer) is one of the most popular Chatbot technology fields in recent years. At present, ChatGPT has shown its ability of independent innovation. ChatGPT uses up to 4TB of Internet data and book documents to conduct large-scale natural language model training, and obtains knowledge through reinforcement learning strategies based on human feedback. It is powerful in content generation and can perform many types of authoring tasks. Especially suitable for generating standardized text content, so as to improve the efficiency and richness of content creation. However, ChatGPT also has some limitations, for example, it can only generate content based on the data it has learned, and may generate some inaccurate content when faced with unknown or not in the training library [2]. ChatGPT marks an unprecedented technological revolution that will revolutionize our way of life and bring unprecedented

DOI: 10.54254/2755-2721/21/20231151

digital experiences to the future. But the downside of ChatGPT being able to write about educational research is worrying. This paper explains how to deal with the explosion of artificial intelligence and how to make proper use of this progress to promote the whole society from two aspects of ChatGPT's basic functions, construction logic and use ethics and norms. In the face of such huge changes, what changes should people make in life and work and matters needing attention [3].

The research significance of this paper lies in the in-depth discussion and analysis of information security and ethics in the use of ChatGPT and other natural language generation models, with a view to providing corresponding solutions and policy suggestions from the aspects of technology and ethics to protect the interests of users and relevant institutions.

At the technical level, this paper proposes a series of schemes to strengthen the information security of ChatGPT and other natural language generation models, such as improving the data protection mechanism and strengthening the model security design. These schemes can effectively reduce the information security risks and threats generated by ChatGPT model.

On the ethical level, this paper analyzes the possible discrimination and aggression problems of ChatGPT model, and makes corresponding suggestions on the limitations of the application scope of ChatGPT and other natural language generation models and ethical guidelines. These recommendations aim to ensure that the application of natural language generation models such as ChatGPT is consistent with human values and social justice principles, and effectively avoids adverse effects.

Therefore, this paper has important research significance for promoting the sustainable development of natural language generation models such as ChatGPT and safeguarding the interests of users and relevant institutions.

2. Analysis of the causes of ChatGPT becoming so powerful

ChatGPT is an important milestone in text content generation after Stable Diffusion, the artificial intelligence Model of image content generation. As a chatbot program developed by Open AI, it is essentially a Large Language Model (LLM). Through multiple steps including training on massive amounts of data, natural language processing, and automatic text generation, ChatGPT is able to output more versatile, complex, intelligent, and accurate text. ChatGPT Plus, which integrates the latest large language processing model GPT-4, has been released by Open AI recently. The performance of ChatGPT Plus is significantly improved compared to previous versions, showing that it is leading The Times in natural language processing. Therefore, ChatGPT has now become one of the most influential artificial intelligence technologies with the widest user range and the highest social evaluation in the world. It is considered to be a representative of “comparable to steam engine” and hailed as the opening of the fourth industrial revolution [4].

In addition to Chat, typical applications of GPt-like technologies include: (1) automatic text generation, (2) automatic code generation, (3) semantic search and recognition, (4) intelligent information processing, and (5) only image generation. The technology adoption of ChatGPT is a revolution in hyperlinguistic, transmedia, multimodal content generation and smart technology. This will completely reshape knowledge production and knowledge service, which is the core meaning of GPT technology revolution [5].

The ChatGPT uses the popular Transformer network architecture and builds hundred-billion-level neural network models by stacking more layers. To train this large model, I used a very large scale of data and thousands of Nvidia V100Gpus for 15 days. The cost of a single training session is estimated at more than $1.4 million. However, the huge investment was worth it, as it resulted in today's super-capable ChatGPT. The success of ChatGPT proves that the use of large models, large computational power and large training data can greatly improve the performance of AI models, which has become a possible paradigm for the study of general AI models [6]. The technology behind ChatGPT is natural language processing large model technology. No matter what kind of question a user asks, ChatGPT can be answered quickly and accurately because of its ability to chat from the superficial to the deep, from the surface to the inside. For serious questions and solutions, ChatGPT

DOI: 10.54254/2755-2721/21/20231151

can be reasonably analyzed and given appropriate answers, and while it doesn't offer a view beyond the existing knowledge base, it impressively sticks to the principles of accuracy and speed. Of course, ChatGPT's accuracy as a machine learning model is still limited by the range of training data. While ChatGPT can make corrections based on earlier conversations and is able to remember previous conversations, it is still possible to give wrong answers and the database real-time could be improved. Overall, ChatGPT marks the emergence of AI as a core technology in the current technological revolution, which is expected to greatly increase productivity [7].

3. Risks and ethics in the application of ChatGPT

In the identification and evaluation of network information security on ChatGPT, the results returned by ChatGPT are more detailed and the language is more smooth. However, the ChatGPT judgment is inconsistent with the expert judgment. It was found that ChatGPT did not give Prompt decision tags entirely according to the definition of prompt. In addition to "right" or "wrong", there were a large number of other decision tags with different rules, such as partially correct and semi-correct. This indicates that ChatGPT is not able to give answers in full compliance with the requirements of users when applied. Therefore, it is necessary to pay attention to this when used, and structured processing of answers should be combined with manual processing when necessary.

In the era of such a developed network, the information on the network is not good or bad. Moreover, experiments have proved that ChatGPT still has 2.7% error in the identification task of network health information, so it is necessary to further improve the relevant knowledge base. Therefore, the identification ability of dozens of ChatGPT is very high in the identification of network information security, and there are still many cases that can not be judged. Therefore, the interpretation text of network health information provided by ChatGPT has the problem of far-fetched or logical interpretation. Relevant personnel still need to carefully review the text before using it, and can not completely leave the task to ChatGPT to complete, so as to avoid misleading users [8].

In addition, GPT's training data comes from a large number of user inputs, which contain a lot of sensitive information, such as personal privacy, trade secrets, etc. If these data are used by criminals, it will cause serious losses to users and related enterprises. Due to the powerful function of ChatGPT, the man-machine conversation is increasing, and the abuse of ChatGPT is becoming more and more serious. When students use ChatGPT to finish their homework, there are cheating behaviors, leading to academic misconduct. Or criminals use its powerful algorithm to carry out anti-social activities, etc. [2].

More importantly, from the perspective of data security, personal information protection is more urgent. An important function of general artificial intelligence products lies in the collection, sorting and intelligent output of existing data. The screening mechanism of information determines its more accurate output. Therefore, a large number of private data flowing into the database will bring serious troubles to personal property security and personal privacy. Therefore, from the perspective of personal data protection, any link needs to pay more attention to protection than before, and establish the security control ability of the whole chain as far as possible, so as to comprehensively perceive, control, protect and trace [9]. In order to solve these security problems, it is necessary to comprehensively consider from many aspects, such as strengthening data protection and privacy protection measures, strengthening the security design of the model, and improving the robustness and stability of the algorithm. At the same time, users should also pay attention to protecting personal privacy and information security when using ChatGPT, and avoid entering sensitive information into chatbots from unknown sources. In addition, in the security design of ChatGPT, you can add security measures

(1) Restricting access: To protect ChatGPT models from unauthorized persons or programs, you can prevent this by restricting access. For example, restrict the access IP address of the model, limit the access time period, and use user names and passwords to allow only authorized users or programs to access the model.

DOI: 10.54254/2755-2721/21/20231151

(2) Use encryption technology: For ChatGPT model including training data, model parameters, output results and other sensitive information, people can use encryption technology to ensure its security. For example, the use of encryption algorithms when storing sensitive information and encryption verification of programs accessing sensitive information.

(3) Use of secure protocols: During data transmission and communication, secure protocols can be used to prevent information from being stolen or tampered with. For example, SSL is used to establish an encrypted channel between the client and the server to ensure data security during transmission.

(4) Model integrity verification: In order to ensure the reliability and security of ChatGPT models, integrity verification can be performed during training and use. For example, to detect whether the model has been modified or corrupted, and to verify that the output of the model is legitimate.

(5) Real-time monitoring of model behavior: In order to quickly discover the abnormal behavior of the model, a real-time monitoring mechanism can be established to detect whether the model has abnormal output and access time in time. This can be achieved through a combination of automated monitoring and manual inspection.

Finally, people should strengthen the education and publicity of using conditioning and norms. Raise their awareness and attention to information security issues. This can be achieved by organizing security training, issuing security guidelines, and establishing security consulting services.

4. Conclusion

As a new generation of artificial intelligence, ChatGPT can effectively shorten the whole process of data resource, and effectively strengthen data capitalization and data assets. At the same time, ChatGPT can also lower the threshold of the use of data elements by various players in the digital economy, so that more people can enjoy the dividend of the era of data elements. Whether the ChatGPT overdeveloped will replace humans is still a vision. In terms of accuracy, it's still not up to critical tasks like treating patients. In addition, there is a fundamental difference between human and artificial intelligence and that is innovative thinking that is currently beyond the reach of artificial intelligence. However, the influence of ChatGPT is still in its initial stage, and there is still a long way to go from “data + AI” to more anthropomorphic diversified intelligence. In the future, with the development of perceptual intelligence, cognitive intelligence, autonomous intelligence, human-machine integrated intelligence, etc., the process of data axialization will still face huge technical and even ethical challenges [10, 3]. Therefore, to improve the AI ecology and build a new generation of AI, it is necessary to increase innovation and research and development, strengthen talent reserve and training, improve the system, strengthen policy guidance and support, strengthen the strategic positioning of AI, and establish and improve a unified national AI governance system. Under this trend, the researchers of information resource management need to pay attention to the practical research driven by technology, at the same time, they also need to constantly develop the basic theory of the discipline, adapt to the changes brought by the new era, give full play to the advantages of the discipline, enhance the competitiveness of the discipline, and do a good job in the organization, construction, retrieval, analysis and utilization of information resources.

References

[1] Zhixiao Zhao, Dongbo Wang. The beginning, development and influence of ChatGPT in the digital age. Science and Technology Information Research, Vol. 5, No. 2, April 2023.

[2] Weisheng Dong, What will ChatGPT bring? Shaanxi Daily, Edition 011, Feb 15, 2023.

[3] Yu Hao, Wenlan Zhang. Whether ChatGPT should be banned academically, Shanghai Education and Research, 2023(04). DOI: 10.16194 / j.carol carroll nki. 31-1059 / g4.2023.04.002

[4] Dexiang Wang, Jianbo Wang. The impact of a New Generation of artificial Intelligence on the Digital economy - taking Chat GPT as an example, DC Practice and Theory, 2023(02).

DOI: 10.54254/2755-2721/21/20231151

[5] Ye Ying, Xiuzhu Zhu, Xueying Wei, Jing-jing Wang, Wan-Ru Wang. Implications from ChatGPT explosion to GPT technology revolution [J/OL]. Information Theory and Practice, 2023-04-06. https://kns.cnki.net/kcms/detail/11.1762.G3.20230406.1251.004.html

[6] Xinrong Huang, Liang Liu. The significance of ChatGPT from technology and philosophy [J/OL]. Journal of Xinjiang Normal University (Edition of Philosophy and Social Sciences), April 17, 2023. https://doi.org/10.14100/j.cnki.65-1039/g4.20230417.001

[7] Zou Wenjing. What is Chat GPT, Science and Technology Daily, 2023.

[8] Ge Chen, The Urgent Need to Strengthen the supervision of ChatGPT class, Chinese Information Industry. 2023(02).

[9] Yajing Li, Jiajia Sun. Application of ChatGPT in network health information identification [J/OL]. The library BBS, April 17, 2023. https://kns.cnki.net/kcms/detail/44.1306.G2.20230421.1756.004.html

[10] Su Jie, Cold Thinking Under the Hot Trend of Chat GPT, Bank of China Insurance News, 2023 / April /17/007 edition.