Hidden communication risks uncovered in ChatGPT

COMMUNICATION

By HAN BO

In early December 2022, the artificial intelligence firm OpenAI released a natural language processing tool called ChatGPT which responds to user prompts. ChatGPT quickly became an internet sensation upon its debut. It can "answer" a series of "tricky questions" from internet users. For example, it can write a critical analysis of itself and generate papers in specific fields in a variety of languages. As an application featuring automatic generation of AI language, it is the most up-to-date computer program that converses with human users and, to some extent, represents the rise of the new information communication form labeled "human intention+machine generated content." However, we must be aware of the social risks generated by AI Generated Content (AIGC), and come up with risk prevention and control measures in advance.

ChatGPT, an AI chatbot developed by OpenAI, has garnered much attention in and outside the tech world. While it is capable of writing emails, scripts, academic papers, even poems and computer code, it has sparked debate and raised concerns about the wildfire of misleading and false information. Photo: CFP

Technical features of AIGC

ChatGPT can generate human-like text in chat settings, simulating conversations with real people on platforms like WeChat. It was trained on a large dataset of real-life conversations, which allows it to generate text that is natural and coherent in a conversational context. Its technical features represent the basic technical characteristics and AIGC trends. Over the past few years, OpenAI has made several breakthroughs, such as the natural language processing model GPT-3 (Generative Pretrained Transformer), the visual recognition model DALL-E, and the reinforcement learning model AlphaGo. ChatGPT is based on the GPT-3 model.

The GPT-3 model is a very powerful natural language generation model, which can be used for text generation, answering questions, translating text, summarizing text, text classification, and other tasks. As a large "transformer-based" language model, it was trained on a large set of text data and fine-tuned via supervised learning and reinforcement learning, to mimic human conversation with high accuracy and efficiency.

At present, GPT-3 represents the development direction of AI content generation, and its technical characteristics are manifested in four aspects. First, it has a large scale with tens of billions of parameters, which is the largest natural language generation model to date. It forms the basis for accurate content generation. The second characteristic is automation. The model can automatically adjust the language style and content of the generated text based on input text. It can also make customized adjustments, which largely improves the older generation of AIGC's tendency towards homogenization. The third aspect is versatility. GPT-3 can be used for a variety of natural language processing tasks, and it performs well in these tasks. Fourth, it is extensible. It can learn new tasks through fine tuning and maintain high performance as the model expands. With the input of new information, the language model can be further trained to serve users' individualized needs and purposes.

That said, these technical features make ChatGPT more multifunctional than any other chatbot in the market. For example, ChatGPT can write hymns, speeches, press releases, and even physical simulations to meet specific requirements. In addition, the chatbot is more rigorous in its answers and can proactively admit its shortcomings. If users propose a difficult question, the chatbot will decline to respond by citing excuses such as "lacking information" or "I'm just a language model." By purposefully setting conditions, AI can produce content in large quantities and within a short period of time, and in particular, by refining the conditions and limits of content generation, it can produce content that is highly similar to human language.

Social risks of AIGC

One of the main social communication risks of AIGC, is that it may be used to carry out immoral or even illegal information dissemination activities. Essentially, AIGC can generate content that "seems right," but is false information. AIGC applications have the potential for the following social communication risks.

The first is the spread of fake news. Given one or more keywords, ChatGPT can generate news that looks real but is completely made up. It can automate everything from headlines to content and even comments. For example, given the keywords "US president" and "diplomatic crisis," and setting a limit of 500 words, ChatGPT will write a news article with the required number of words and keywords about a diplomatic crisis involving a country's leader.

The second risk is information fraud. ChatGPT can produce high-quality text, so it is prone to cyber fraud usage. For example, it could be used to write legitimate looking emails designed for property fraud, or create fake policy and government documents for information fraud. All it needs are the requirements and the machine will automatically imitate the text content with fixed patterns. With ChatGPT, a "decent" government investment solicitation document is at your fingertips.

Last, content generation which imitates specific discourse styles is made possible. ChatGPT can not only obtain a large dataset of internet texts, but also carry out human corpus input to conduct later training for artificial intelligence. For example, influential opinion leaders such as online celebrities and influencers often have distinctive language styles. ChatGPT can imitate specific language styles for content production, and its discourse structure and wording characteristics is highly similar to real people, which easily misleads audiences.

Overall, the high degree of intelligent AIGC applications have lowered the technological use threshold for artificial intelligence production. It can handle multiple language generation requirements including discourse style, number of words, text type, and so on at one time. The more specific the generation conditions are, the more human-like the content is produced. For false and untrue information, AIGC and other similar technologies will directly reduce the technical requirements for mass production of false information, accelerate the production speed of false content, and further challenge future information content governance.

Coping strategies

In recent years, with the rapid development of natural language processing (NLP) technology, high-level language models such as ChatGPT have made remarkable progress in language comprehension and generation. However, these technologies also carry the risk of spreading false information. In this light, we should work on three mechanisms to prevent AIGC from becoming a source of false information.

First, a content responsibility system should be put in place, delineating "who generates, who produces, then who should take responsibility." In terms of cyberspace content governance, a clear responsibility system is the foundation of a clean cyberspace. For the production and application of the new generation of artificial intelligence, such as ChatGPT, the existing content responsibility system should be extended, including Measures for the Management of Internet Information Service and Regulations on the Management of Internet Comment Service.

In short, whoever is involved in the production and creation of content should share governance responsibility. For example, the model developer should be responsible for the model's risk of generating false information. Platform providers and content publishers should review the information generated using ChatGPT to prevent the spread of misinformation. The government should formulate regulatory guidelines for online information dissemination, establish mechanisms to ensure the authenticity and accuracy of information, and punish offenders who produce and spread false information.

Second, a mechanism for identifying and disposing of artificial intelligence generated content should be established. When faced with artificial intelligence content generation, the only way to counter it is to adopt an AI-on-AI strategy. According to the technical characteristics of different language models, the artificial intelligence content monitoring system should be iterated accordingly and constantly.

At this point, AIGC applications like ChatGPT have some apparent limitations, including but not limited to syntax and semantic errors, which can occur when text is generated and cause semantic inconsistencies. Limited ability to generalize, made weaker when approaching new tasks, determine that ChatGPT may not accurately answer new questions or process new textual data. Lack of logical reasoning translates into an inability to deduce conclusions through reasoning, which may lead to mistakes when dealing with complex problems. High dependency on input text is also problematic. These problems, to a large extent, represent the universal defects of artificial intelligence content production. Therefore, monitoring tools can be developed which focus on repetition, logic, syntax, and semantics to identify false information. Through analyzing language features, reading within context, and using pre-training data in a data-driven way, it is possible to construct the mechanism for identifying and disposing of artificial intelligence generated content.

Finally, renewed efforts should be made to develop autonomous and controllable generative models of natural language. At present, there are limited AI language models on the market, such as the Google BERT model, Alibaba PLUG model, and "Jiuge" Chinese poetry generation system developed by Tsinghua University. There are also "character AI" that can use personification to freely talk with users.

However, ChatGPT is more powerful in generating language. The logic behind ChatGPT is that the strong will become even stronger in the future of AI production. ChatGPT is optimized for conversation using human feedback reinforcement learning (RLHF), a method that uses human interaction to guide models to achieve desired behaviors. The number of parameters has increased from 117 million to 175 billion, and pre-training data levels have increased from 5GB to 45TB. One GPT-3 training session costs about $4.6 million, while the total training costs $12 million. The popular AI language model continues to evolve as people feed information back through their use, eventually intimidating competitors. This suggests that China should strengthen research and development investment into autonomous and controllable natural language generative models, to promote the development of AI in China through market means.

Han Bo is from the Institute of Journalism and Communication at the Chinese Academy of Social Sciences.