https://doi.org/10.1038/s43018-023-00531-0

Not a generative AI–generated Editorial

What does the explosion of generative artificial intelligence tools mean for science?

Since Chat Generative Pre-trained Transformer (ChatGPT) was unveiled in November 2022, the world cannot seem to stop talking about generative artificial intelligence (AI) tools — the latest thing by which the internet, as a proxy of our collective expression, is enthralled and/or terrified.

Yet generative AI tools are far from new. ChatGPT itself is based on GPT-3, OpenAI's third-generation large language models (LLMs), which were presented in 2020. Other, less successful ventures into this space include Microsoft's infamous Tay chatbot, programmed to 'converse' with and continue learning from users, which was terminated 16 hours after its Twitter release in 2016, when users trained it to spout offensive abuse. Only 2 weeks before ChatGPT's release came the ill-fated launch of Meta's Galactica[1], an LLM that was trained on 48 million pieces of open scientific content, including papers, textbooks, websites, encyclopedias and molecule structures, with the goal of assisting researchers by answering science queries and writing scientific text and code, among other things. Within 3 days of its release, Galactica's public demonstration had been shut down after intense criticism over its frequently incorrect, fabricated, discriminatory and sometimes simply outlandish responses.

However, where Galactica fell flat, ChatGPT took off. Using a large swath of internet data and additional guardrails in model training, the creators of the latter controlled against (some of) the above-mentioned glitches, including whether and how the chatbot responds to offensive queries. This was followed by the limited preview release a few weeks ago of the new Microsoft Bing search engine[2], which is also powered by OpenAI's LLM technology and is stated to be more advanced than ChatGPT while also including safeguards against harmful content. In the short time since the public release of these tools, endless hours and countless opinion articles, news pieces and social media posts have been dedicated to discussing the potential and limitations of generative AI. As we grapple with the ethics of using such technologies, we continue to collectively participate in a large open experiment to identify these chatbots' weaknesses by testing and training them in real time.

For generative AI enthusiasts, the future is now! The chatbots can do everything from responding to science questions to taking school exams and writing all types of text. Indeed, ChatGPT has the potential to summarize the scientific literature (with the notable caveat of not citing sources) and produce text. Thus, in principle, it could be used to write, improve and proof parts of, or whole, scientific manuscripts. In the future, one could envision the integration of such tools into literature analysis, experimental design, the preparation of presentations and figures, manuscript writing and reviewing, and several parts of the editorial and publication process. Through such uses, AI-powered tools would also have the ability to level inequalities by eradicating language barriers and improving writing style and readability. Additional applications include aiding in research and medical training, disease diagnosis, drug development and therapy selection, to name a few.

Critics point to the well-known limitations of LLMs, including AI 'hallucination' phenomena — whereby chatbots provide spurious information as correct despite the existence of training data to the contrary. Several ChatGPT and Bing mistakes have already been reported, including stubbornly insisting that a falsehood is fact and presenting beautifully written, seemingly reasoned information that is completely made up. A relevant example is the production of plausible-sounding abstracts of non-existent scientific papers[3]. ChatGPT has no concept of science integrity and no qualms about fabricating research. As users have continued to test the abilities of these tools, they have come across unpredictable, troubling responses that in the case of Bing[4] have included declarations of love and forays into its own Jungian version of a 'shadow self'. Reports also abound of 'jailbreaks' that permit users to trick the chatbots into responding to problematic queries despite their own restrictions, thereby increasing the possibility of spreading offensive content and disinformation. Moreover, the wide adoption of such tools may actually exacerbate some inequities, most prominently those between well-resourced and poorly resourced research settings.

Another notable shortcoming is that an LLM is only as good as its training dataset. In ChatGPT's case, this ends somewhere around 2021, so one must look elsewhere for more-current information or wait for an updated version of the algorithm. The new Bing is more up to date and, in principle but not always in practice, able to process queries on recent events. An additional concern is whether the training data contain biases and inaccuracies that the chatbot will perpetuate and magnify. If trained on content espousing a flat-Earth view of the universe, an LLM will quickly be recognized as spouting flat-Earth gibberish. However, the presence of subtler biases and smaller inaccuracies may be harder to detect and therefore more dangerous, especially when someone engages such tools as a source of information and intends to use their inexpert, un-nuanced, potentially incorrect responses in their own communications. When such communications pertain to research and medicine, the consequences can be dire.

This leads to the contentious issue of whether generative AI models should be used in research communication and, if so, under what conditions. As noted in our January Editorial[5], AI tools cannot be credited as authors, as they cannot comply with the accountability that comes with authorship. However, other questions are more complex. Should generative AI be considered a tool that aids writing or one that enables plagiarism? Can the contributions of generative AI tools, the outputs of which may range from simple text proofing to new text synthesis, ever be fully disentangled for accurate attribution? Who holds the rights to content generated by AI tools, especially if the tools were trained on data that may have lacked permissions for use and reproduction? What is the place of generative AI tools trained on datasets that may not be well defined or publicly available in the era of open science and transparent reporting? At this time, the Nature journals mandate that use of any LLM be clearly documented in the manuscript, but publishers, academia and other stakeholders are deep in consideration of these matters, with policy development in progress.

For scientists and publishers, the broader aim is to engage positively and constructively with new AI-driven technologies while defining their limits and realizing protections against misuse. Tools are already being developed to detect text that has been generated by generative AI, with GPTZero being one such example. Other approaches include ongoing efforts to watermark AI-generated text. Neither approach would be infallible, as efforts to detect such text have already shown[3], but they may well become integrated in the publishing process much like other tools to ensure research integrity.

As generative AI technologies continue to develop in newer, more-interesting and hopefully more-reliable ways, the bigger question is to what degree they should be permitted to permeate scientific efforts. How reliant should research become on bots that ingest parts of the literature and regurgitate hopefully accurate and probably superficial text? How much human labor should be ceded to automation when it comes to curating and critically evaluating information to synthesize it in meaningful manners that lead to new ideas? To what extent are we open to supplanting the curiosity and creativity of the imperfect human mind to the time-saving but perhaps bland capabilities and opaque methods of an algorithm? Striking the right balance of embracing innovation while maintaining transparency and safeguarding the principles of the scientific endeavor will be key.

Much has been made of the ability of ChatGPT to do pretty much anything, including writing songs and poems — or sad simulacra of what a chatbot might be trained to recognize as a song or poem in "a grotesque mockery of what it is to be human," to quote the musician Nick Cave. Like art, science is inherently a creative endeavor. Like science, art is an exploration of the great mysteries of existence, an effort to make sense of the world around us. One uses reason; the other, feeling. Both require the experiences, flair and insight of a human being. What may be enticing in interacting with ChatGPT is right there in its name: it is a chatbot, programmed to respond to us in natural language. It lacks morality, consciousness and the ability of inspiration and original thought, yet it sounds as if it has all the above. Its neat and often lengthy responses have the feel of a dialog with a sentient and authoritative interlocutor, when behind the curtain is only a sophisticated model trained to emulate human responses on the basis of the statistical patterns of huge swaths of input text. It may be able to produce seemingly fluent responses, but it will not interpret experimental results or access the literature in a manner that goes beyond the surface of stringing probability-derived word combinations together — and it will certainly never write real poetry.

Published online: 28 February 2023

References

1. Taylor, R. et al. Preprint at https://arxiv.org/pdf/2211.09085v1.pdf (2022).
2. Mehdi, Y. *Microsoft* https://blogs.microsoft.com/blog/2023/02/07/reinventing-search-with-a-new-ai-powered-microsoft-bing-and-edge-your-copilot-for-the-web/ (7 February 2023).
3. Gao, C. A. et al. Preprint at *bioRxiv* https://www.biorxiv.org/content/10.1101/2022.12.23.521610v1 (2022).
4. Roose, K. *New York Times* https://www.nytimes.com/2023/02/16/technology/bing-chatbot-transcript.html (16 February 2023).
5. *Nat. Cancer* https://www.nature.com/articles/s43018-023-00517-y (2023).