Exobrains for research evaluation and paper writing

Eric Lichtfouse, Chuxuan Gong, Wei Qiu, Jie Han

Environmental Chemistry Letters (2024) 22:461–462
https://doi.org/10.1007/s10311-023-01672-5

Exobrains for research evaluation and paper writing

Eric Lichtfouse[1] · Chuxuan Gong[2] · Wei Qiu[2] · Jie Han[3]

The fact that scientists could be fooled by abstracts written by an artificial intelligence chatbot (Else 2023) implies that such tools will soon represent 'exobrains' to carry out tasks such as scientific writing and research evaluation (Fig. 1). Here we have tested ChatGPT for evaluating a full-length review article, and we were impressed by the results that humans often cannot deliver, such as summarizing a 7,000-word research paper into sensible, bullet-point listed findings in less than 30 seconds. Although answers from such large language models are not flawless, ChatGPT appears as an exobrain—similar to the physical exoskeletons that humans could use—to amplify their intelligence. Could we use ChatGPT to evaluate papers and measure research performance? Here the ChatGPT evaluation of one of our review articles reveals the intriguing performance of the chatbot.

We asked ChatGPT what are the strengths and weaknesses of our recent review article on food transmission of the 2019 coronavirus (Dai et al. 2023). For strengths, ChatGPT produced concisely the following conclusions:

- *The paper provides clear evidence from China that the coronavirus can be transmitted through frozen and chilled food products.*
- *The paper offers valuable insights and lessons for the world by examining the case of China's experience with the virus.*
- *The paper provides a comprehensive overview of the current understanding of the virus and how it is transmitted.*

These conclusions do represent what we think are the key points of the paper. This is surprising because selection of few important points out of several pages of text requires deep understanding of their meaning.

For weaknesses, ChatGPT yielded the following:

- *The paper does not provide any solutions or recommendations for how to prevent the spread of the virus through frozen and chilled food products.*
- *The paper does not address the potential implications of its findings in terms of global food safety and regulatory policies.*
- *The paper does not delve into the economic impacts of the virus on the global food supply chain.*

These valid points are intriguing because our review does not contain any section, paragraph or even a single sentence explaining its limitations. This denotes the ability of ChatGPT to critically evaluate the manuscript with what it thinks is a sound paper on the topics discussed. Therefore, we presume that ChatGPT has analyzed publications on a similar topic to yield its own standards of a high-quality review. This is astounding because critical evaluation denotes intelligence, and as a consequence, artificial intelligence could be used to evaluate research, at least partly, e.g., as an instant second opinion without proactive human bias. Artificial intelligence could also be used to improve writing by revealing what is missing from the draft before one finalizes a manuscript.

Then we asked again the same question and unexpectedly got different answers. For instance, on 'weaknesses' it answered:

✉ Eric Lichtfouse
eric.lichtfouse@icloud.com

✉ Jie Han
jiehan@xjtu.edu.cn

[1] State Key Laboratory of Multiphase Flow in Power Engineering, Xi'an Jiaotong University, Xi'an 710049, Shaanxi, People's Republic of China

[2] International Science & Technology Cooperation Base of Xi'an Municipality, Guyiheng Technologies Ltd., Xixian New Area, Xi'an 712000, Shaanxi, People's Republic of China

[3] School of Human Settlements and Civil Engineering, Xi'an Jiaotong University, Xi'an 710049, Shaanxi, People's Republic of China

Fig. 1 ChatGPT is a large language model developed by OpenAI, a US company deploying artificial intelligence technologies. The model was released to the public on 30 November 2022. More information at https://openai.com/blog/chatgpt. Image credit: AJ Cann/Flickr

- *The paper focuses exclusively on evidence from China, which may limit its applicability to other countries and contexts.*
- *The paper relies heavily on anecdotal evidence and case studies, rather than more rigorous scientific studies. This could weaken its overall argument.*
- *The paper does not consider potential counterarguments or limitations to its argument, such as the possibility that other factors, e.g., close contact with infected individuals, may have played a larger role in the outbreaks described.*
- *The paper could benefit from more in-depth analysis and discussions of the policy and logistical challenges involved in implementing increased surveillance and testing of frozen and chilled food products.*

First, these deep arguments would have been very useful to us to improve the manuscript prior to submission. Second, there are at least three possible reasons for the fact that ChatGPT does not give the same answers: (1) The knowledge base changed as the models changed from the first to the second time the question was asked, i.e., Playground—OpenAI API (GPT-3 series) versus ChatGPT; this is possible but would unlikely affect the answers substantially. (2) ChatGPT answers are generated somewhat randomly, so the same questions always generate different answers, sometimes substantially; in that case, ChatGPT is less consistent than a human scientist who should give essentially the same answer to the same question within a short time span. (3) More interestingly, ChatGPT may have memorized the question when asked firstly, then it understood that we were not satisfied because we asked the same question again in a different conversation. This again denotes intelligence. Whatever the origin of those different answers, the results imply that large language models have the ability to evaluate research outputs, at least for initial screening. Moreover, these models could be used to improve manuscripts by revealing major scientific flaws and providing constructive criticism.

Overall, ChatGPT and other large language models appear as 'exobrains' that are likely to be used for many tasks in science such as research evaluation and paper writing. After all, if research evaluation and paper writing can be summarized as 'sound comparison of new results versus existing knowledge', then models could do that, at least partly. This would surely save time and money of scientists, laboratories, universities and funding bodies.

Funding This work was funded by the National Natural Science Foundation of China (Grant No. 42277209) and the Key Research and Development Program of Xianyang (Grant No. L2022ZDYFSF042).

Declarations

EL and JH declare that they are Chief Editors of Environmental Chemistry Letters.

References

Dai H, Tang H, Sun W et al (2023) It is time to acknowledge coronavirus transmission via frozen and chilled foods: undeniable evidence from China and lessons for the world. Sci Total Environ 868:161388. https://doi.org/10.1016/j.scitotenv.2023.161388

Else H (2023) Abstracts written by ChatGPT fool scientists. Nature. https://doi.org/10.1038/d41586-023-00056-7