AsiaIntervention 2024;10:12-14

Artificial intelligence in medical writing

James Howard*, MB, BChir, PhD; Hoi Ching Cheung, MBBS

National Heart and Lung Institute, Hammersmith Campus, Imperial College London, London, United Kingdom

Artificial intelligence and machine learning

Artificial intelligence (AI) can be defined in many ways, but at its simplest, it is the phenomenon whereby a computer performs a task we traditionally associate with human intelligence. By this definition, AI has existed for more than 70 years; the earliest examples include the work by Christopher Strachey, who developed a program that could play a game of draughts in the 1950s[1]. However, these programs were far from embodying "intelligence" in the human sense – they were simply executing sequences of logical steps, written explicitly by humans who converted strategic games into algorithmic processes.

More recently, AI has been associated with another term: machine learning (ML). ML represents a paradigm shift in computer programming, where, instead of programming a computer to solve a task, we program a computer to learn how to solve a task. If one opens a textbook on ML, the first chapter is frequently "linear regression", which is the simplest form of ML. It is essentially the process of fitting a line of best fit for data. An example we often use in clinical practice is the formula for maximum heart rate:

$$\textit{Maximum heart rate} = 220 - \textit{age}$$

This equation could easily be derived by gathering a group of individuals, exercising them to exhaustion, and then plotting each patient's heart rate against their age in years. A computer would then be able to draw a line of best fit, which would cross the y-axis at 220 (the intercept) and would have a slope (or gradient) of –1.

Although linear regression may seem far from exciting, it demonstrates a key point: a computer can create a mathematical model that can be used to predict a future patient's maximum heart rate using example data.

Neural network and deep learning

In the modern era, however, ML typically refers to significantly more complex algorithms, such as neural networks. Neural networks are computer programs comprising many layers of "neurons", interconnected through (typically millions of) synapses, in a similar manner to the human brain[2].

When we first construct a neural network, we do not expect it to produce any useful results. However, we can train it to perform a specified task. This involves repeatedly providing it with example data and their corresponding answers – a process called supervised learning. Every time we show the neural network an example (which could be numerical data, images, or text), we look at the answer it produces and compare it with the answer we expected. Then, we can adjust the network's synapses, so the

---

*Corresponding author: National Heart and Lung Institute, 2nd floor B Block, Hammersmith Campus, London, W12 0HS, United Kingdom. E-mail: james.howard1@imperial.ac.uk

DOI: 10.4244/AIJ-E-23-00005

© Europa Digital & Publishing 2024. All rights reserved.

The opinions expressed in this article are not necessarily those of the Editors of AsiaIntervention.

next time it sees that example it should produce a more correct answer. During the "training" process, we typically perform such adjustments thousands or even millions of times, and eventually, we hope the neural network will become adept at the task. The term "deep learning" specifically refers to the training of "deep" networks with numerous layers.

Over the last decade, neural networks have become state-of-the-art in fields such as image processing, and physicians are increasingly using neural networks to process medical imaging in their day-to-day practice[3]. In 2017, a breakthrough occurred in the field of AI with the introduction of the transformer architecture, which enabled the creation of large language models (LLMs).

Large language models and GPT

LLMs represent a type of neural network which processes textual data and have revolutionised the field of natural language processing[4]. Many tech giants have their own LLM: OpenAI (part-owned by Microsoft) has the General Purpose Transformer (GPT) models, Google has Bard, and Facebook has LLaMA. These models are truly vast in size; ChatGPT-4 for instance, is thought to contain over 1 trillion synapses. Generally, the greater the number of synapses, the greater the capacity a network has to retain information and, therefore, the more sophisticated its ability to process and generate useful text.

Despite their size, all LLMs are fundamentally simple. They are trained to perform one basic task: predict the most likely next word in a sequence. When you ask ChatGPT, "What is the largest city in South America?", rather than searching a memory bank or the internet for the answer, it merely predicts the most likely next word, based on billions and billions of pieces of text it has been shown during its training. Perhaps ChatGPT would think the most likely next word is "Sao" (as in "Sao Paolo, Brazil" – the correct answer), or "the" (as in "The largest city in South America is Sao Paolo.") – both would be reasonable first words in a response. After it has produced the first word, e.g., " the", it then feeds the original passage plus the predicted word back into the network (i.e., "What is the largest city in South America? The") and predicts the next word, which may be "largest" **(Figure 1)**.

Ultimately, an LLM such as ChatGPT is only as accurate as the data it has been trained on; to know "Sao" is more likely to be the next word than "Lima", the neural network must have been trained with sufficiently large libraries of text to not only learn English grammar but also "facts".

Can ChatGPT write academic papers?

It is evident that LLMs hold great potential. However, researchers must be vigilant about several concerns when applying them to academic writing.

PROBLEM 1: "HALLUCINATION" IS A HUGE PROBLEM FOR LLMS

As discussed earlier, an LLM is simply designed to predict the next word in a passage of text. Because of this, even when an LLM does not "know" the answer, it may tend to "guess" the most likely answer anyway. This results in the very real phenomenon of "hallucination", where LLMs have learned that the best response to an answer it does not know is to guess, often with disarming confidence. For example, when ChatGPT-4 was asked to explain an assertion with absolutely no evidence, e.g., "Can you explain to me why percutaneous coronary intervention (PCI) of the circumflex artery is prognostic but PCI of the right coronary artery (RCA) is not?", not only did it not correct the user's false belief, but it also "fabricated" a false explanation:

"*PCI of the circumflex artery can be prognostically significant due to its role in supplying blood to a large portion of the left ventricle, particularly in left-dominant circulations. The circumflex artery can also supply important areas like the lateral wall of the heart and the sinoatrial node in some individuals. Compromised blood flow in this artery can thus have a more pronounced impact on cardiac function.*

In contrast, PCI of the RCA may not hold the same prognostic significance in many cases. This is partly because the RCA primarily supplies the right side of the heart, which is involved in pulmonary circulation and is generally less susceptible to life-threatening consequences compared to the left side. Additionally, the left side of the heart, responsible for systemic circulation, is typically more critical for overall cardiac function. Hence, interventions in the RCA are often considered less prognostically significant compared to those in the left coronary system, including the circumflex artery." – ChatGPT (November 2024).

Interestingly, and concerningly, when the question was reversed ("Why is RCA PCI prognostic, whereas circumflex PCI is not?"), ChatGPT explained this is due to the importance of the right ventricle's blood supply. This demonstrates how easily LLMs can be persuaded to justify any opinion provided by the user, even if it is not supported by true data.

PROBLEM 2: PLAGIARISM CONCERNS

Although we might be able to address the issue of hallucination by providing the LLMs with accurate information, such as a first draft of a manuscript or a list of bullet points to expand upon, a fundamental limitation remains. LLMs are trained on pre-existing work by various authors, and any text they generate are synthesised from such data. This process lacks genuine understanding or original thought. Journals such as "Science" have anticipated this and have argued that, given research in journals must be authors' "original" work, the use of ChatGPT constitutes plagiarism[5].

PROBLEM 3: LLMS DO NOT MEET AUTHORSHIP CRITERIA

One cunning way around the plagiarism concern is that authors could list their chosen LLM as a co-author, reflecting its contribution to the manuscript. Indeed, one preprint exploring the performance of ChatGPT on the United States Medical Licensing Examination (USMLE) had the LLM as the third author[6]. However, journals such as "Nature" have since argued that LLMs cannot satisfy traditional authorship criteria, as AI cannot provide the

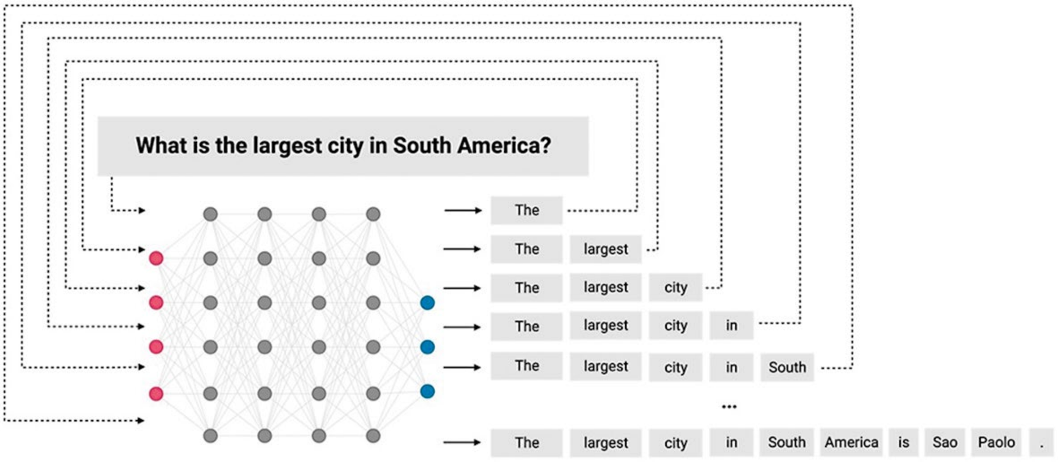
What is the largest city in South America?
The
The largest
The largest city
The largest city in
The largest city in South
...
The largest city in South America is Sao Paolo .

Figure 1. *A schematic showing how a large language model produces text by iteratively predicting the next word in a sequence.*

necessary "accountability for the work"[7]. It is notable that when the USMLE paper was peer reviewed and published in "PLOS Digital Health", ChatGPT was no longer listed as a co-author[8].

Conclusions

The integration of AI in medical writing is constrained due to concerns regarding hallucination, plagiarism, and authorship issues. Nonetheless, there may be a case for its cautious use. Just as spell- and grammar-checking software is a standard part of authorship, there is potential for AI to critique and enhance academic writing. Clear guidelines for the responsible use of such technology in academic settings are still lacking, underscoring the need for further discourse and regulation in this evolving field.

Conflict of interest statement

The authors have no relevant conflicts of interest to declare.

References

1. Strachey CS. Logical or non-mathematical programmes. ACM '52: Proceedings of the 1952. Paper presented at: ACM, September 1952, Toronto, Canada.

2. de Marvao A, Dawes TJ, Howard JP, O'Regan DP. Artificial intelligence and the cardiologist: what you need to know for 2020. *Heart.* 2020;106:399-400.

3. Howard JP, Francis DP. Machine learning with convolutional neural networks for clinical cardiologists. *Heart.* 2022;108:973-81.

4. Douglas MR. Large Language Models. arXiv. 2023. https://arxiv.org/abs/2307.05782. (Last accessed 5 December 2023).

5. Thorp HH. ChatGPT is fun, but not an author. *Science.* 2023;379:313.

6. Kung TH, Cheatham M, ChatGPT, Medenilla A, Sillos C, De Leon L, Elepaño C, Madriaga M, Aggabao R, Diaz-Candido G, Maningo J, Tseng V. Performance of ChatGPT on USMLE: Potential for AI-Assisted Medical Education Using Large Language Models. bioRxiv. 2022. https://www.medrxiv.org/content/10.1101/2022.12.19.22283643v2 (Last accessed 5 December 2023).

7. Editorial policies: Artificial Antelligence (AI). Nature Portfolio. https://www.nature.com/nature-portfolio/editorial-policies/ai (Last accessed 25 November 2023).

8. Kung TH, Cheatham M, Medenilla A, Sillos C, De Leon L, Elepaño C, Madriaga M, Aggabao R, Diaz-Candido G, Maningo J, Tseng V. Performance of ChatGPT on USMLE: Potential for AI-assisted medical education using large language models. *PLOS Digital Health.* 2023;2:e0000198.