https://doi.org/10.1038/s41551-023-01012-6

Prepare for truly useful large language models

The consequences of their use will be far reaching.

The article discusses the rise of large language models, such as OpenAI's ChatGPT, which use machine learning to generate human-level prose. These models have learned structural, relational and semantic language patterns to create language that mimics human speech, but they do not have knowledge of logic, facts, the laws of the physical world, and the human sense of morality. The models generate text based on the textual patterns of human text they have been trained on, but they do not provide confidence for the accuracy of their output. Despite this, the author believes that these models can be useful for creative purposes and will have a significant impact on various industries, particularly those that rely on communication. They will also have a profound impact on science, making it easier and faster to generate and analyze information. The author argues that the rise of large language models will have significant consequences, both positive and negative, but it is inevitable that they will play a critical role in our future.

A summary of this Editorial, produced by OpenAI's ChatGPT.

The world has barely awakened to the fact that a relatively simple yet large neural network — with a feed-forward architecture and about 100 'attention blocks' and 200 billion parameters[1] — can generate new dialogue that passes the Turing test. Indeed, barring the use of advanced watermarking strategies[2], it is no longer possible to accurately distinguish text written by a human mind from that generated by a highly parallelizable artificial neural network with substantially fewer neural connections. Only a few years ago, most experts in machine learning and linguists would not have believed that human language could be mastered by a computing engine.

Yet mastering language doesn't imply broader conceptual understanding. Trained large language models have learnt structural, relational and semantic language patterns that make the generation of human-level prose possible. But they do not model logic, facts, the laws of the physical world, and morality. Still, because we use language to communicate knowledge and feelings, it is understandable — yet odd, if not downright inappropriate — to anthropomorphize and feel amazed or unsettled by uncanny dialogue with a language model when the only thing that it 'knows' is how to predict the best next word in a piece of text (or, more precisely, the next token). In fact, large language models such as OpenAI's ChatGPT have no knowledge of 'truth', and hence can fail at simple maths and logic. They can also write nonsense confidently.

Publicly available large language models do not provide a degree of confidence for the accuracy of their output. One main challenge is that they are not explicitly designed to provide truthful answers; rather, they are primarily trained to generate text that follows the patterns of human language. Because the models have been trained on huge swaths of text from the web and from digitized books, they can generate content that is as trustworthy or as misguided as pockets of information in their training datasets. And, when prompting longer dialogue, the algorithms may seem to have acquired 'personas' akin to those one can find in corners of the web.

Yet, current flaws and limitations neither imply that the models cannot be really useful, nor that they can't be used for creative purposes. New knowledge can arise from apparently disconnected ideas and concepts that language can help put into fertile use; hence, by ingesting corpuses, language models may unveil unapparent associations. In other words, that the models can 'hallucinate' can be a feature rather than a bug. The models are probabilistic; they are programmed to make use of a small degree of randomness, so that they can occasionally pick a lower-ranking token.

The unveiling of OpenAI's ChatGPT in late November 2022 may be seen as a watershed event. It is all but certain that general-purpose large language models will rapidly proliferate. OpenAI's ChatGTP, Microsoft's AI-powered Bing search, and Google's Bard will soon be competing for the public's attention (and for advertising money), and the quality of the models' output will improve as they are increasingly used. In particular, refining the models with reinforcement learning from human feedback can help align them with human preferences[3]. Other large language models will be trained for specific domains of knowledge by using smaller and higher-quality datasets. For example, large clinical language models with billions of parameters can leverage unstructured text in electronic health records to aid the extraction of medical concepts and answer medical questions[4], to predict disease or readmission risk and to summarize clinical text[5]. They can also be trained with protein sequences, rather than with strings of words, to generate candidate protein drugs[6]. Moreover, transfer learning helps to re-use datasets to train and retrain networks that can generalize and solve related tasks. And training the networks with diverse datasets — from electronic health records, laboratory tests, and wearables, in particular — is expected to boost the medical utility of the models[7]. Text-to-image models (such as DALL·E, Midjourney and Stable Diffusion) and upcoming large vision models[8] (also based on the transformer architecture) will be used to generate, classify and accurately describe images and videos.

Large language models and large vision models will have all sorts of profound consequences. It is a rather safe bet that they will change many industries over time, especially in domains highly reliant on the search, generation and analysis of written and visual communications. By making it easier, faster and cheaper to generate and analyse verbal and visual knowledge, the models will increase

productivity and efficiency. They may also precipitate job losses, especially for those who are unable or unwilling to embrace the new tools.

It is therefore inescapable that applications leveraging large machine-learning models may turbocharge science and the work of scientists. It has also not escaped our notice that they will change how editorial and publishing work is done. The current version of ChatGPT can already be used as a proficient line editor (indeed, it has helped us edit this Editorial), as a writer of summaries (pictured), as a trainer on editorial matters, as an editorial assistant, and as an efficient secretary for carrying out some administrative drudgery. We will need to give serious thought to the actual generation, quality and value of future research highlights and scholarly reviews. And we look forward to less tolerance for shoddily written text.

More consequentially, it is likely that large language and vision models that can digest the literature will be used to identify gaps in knowledge, help summarize and understand unfamiliar topics, and find the most relevant references, protocols, data and experts. They will also generate and explain complex graphs and schematics, and help write and edit routine computer code as well as scientific papers, reviews, grant applications, curriculum vitae and all sorts of reports. Producing content without assistance from machine-learning applications may soon be as rare as writing snail mail.

The stakes are high. Disruptive technology can democratize knowledge and power, yet also polarize, amplify biases and exacerbate inequalities, depending on how it is deployed and used. Businesses, research institutions and scientists with more resources, or with access to more useful datasets for training domain-specific models, may leverage a competitive advantage. Moreover, the models will make fraud easier to carry out at scale, and hence guidelines[9,10] and better safeguards will be needed to preserve the integrity of the content that the models will keep on ingesting. We may have to redefine plagiarism.

The advent of large language models will further blur the lines between truth and falsehood, especially at the forefront of knowledge when the evidence is weak, or when the information is scarce or under debate. Still, it may be possible to design models that alert of potential logical weaknesses, factual mistakes and fraud. Moreover, the limitations of the models will highlight the value and need of deep expertise, experience and sound judgement, and of knowledge of social and cultural contexts. That's also worth preparing for.

Published online: 7 March 2023

References

1. *Nat. Biomed. Eng.* **6**, 1319–1320 (2022).
2. Kirchenbauer, J. et al. Preprint at https://arxiv.org/abs/2301.10226 (2023).
3. Ouyang, L. et al. Preprint at https://doi.org/10.48550/arXiv.2203.02155 (2022).
4. Rasmy, L., Xiang, Y., Xie, Z., Tao, C. & Zhi, D. *NPJ Digital Med.* **4**, 86 (2021).
5. Zhang, L., Xing, L., Zou, J. & Wu, J. C. *Nat. Biomed. Eng.* **6**, 1330–1345 (2022).
6. Madani, A. et al. *Nat. Biotech.* https://doi.org/10.1038/s41587-022-01618-2 (2023).
7. Acosta, J. N., Falcone, G. J., Rajpurkar, P. & Topol, E. J. *Nat. Med.* **28**, 1773–1784 (2022).
8. Dehghani, M. et al. Preprint at https://doi.org/10.48550/arXiv.2302.05442 (2023).
9. *Nat. Mach. Intell.* **5**, 1 (2023).
10. *Nature* **613**, 612 (2023).