https://doi.org/10.1038/s42256-023-00655-z

What's the next word in large language models?

We are trying to keep up with the torrent of developments and discussions in AI and language models since ChatGPT was unleashed on the world.

The past two decades have seen a steady rise in the adoption of machine learning tools in everyday applications, such as in search engines, recommender systems, language translation tools, image editing apps, health applications and many more. A new phase may be starting with the advent of AI generative tools that are powered by large language models (LLMs), such as ChatGPT for text and DALL-E or Stable Diffusion for images, which give millions of people direct access to powerful creative applications. Many news articles and commentaries have been written to debate the opportunities, disruptive societal impact and ethical concerns of LLMs and their downstream applications. A Correspondence in this issue, for instance, discusses the dilemma that is faced by higher education in allowing or banning the use of ChatGPT and related tools by students.

Keeping up with developments in this area is challenging, as tech companies race to compete in developing new, more powerful and versatile versions of LLMs. Just in recent weeks, Meta reported their LLaMA model on 24 February; Google announced PaLM-E on 10 March, a multimodal version of the PaLM language model, which incorporates robot sensor data; Baidu introduced their LLM-based Chatbot ERNIE on 15 March; OpenAI revealed their next GPT version – GPT-4 – on 14 March; and GitHub announced Copilot X on 22 March, which adopts GPT-4 and Chatbot features to support code developers.

Then there are the policy and ethics responses. Getty Images are suing Stability AI – the creators of Stable Diffusion – for copyright infringement; Italy is banning ChatGPT; Canada's federal privacy watchdog has launched a probe into privacy concerns over ChatGPT; and, as widely reported, an Open Letter from the Future of Life Institute calling for a pause on 'giant AI' for at least 6 months has been signed by thousands, including well-known AI researchers and commentators. Within a few days, a response from AI ethics experts appeared to criticize the Open Letter for fuelling hype and ignoring ongoing societal harms from AI systems, which will not be solved by a 6-month pause.

The scale of developments and the unprecedented level of continuing wide public interest have made it difficult for both experts and interested parties to make sense of the latest AI breakthroughs. It may be surprising to many, perhaps, that the connection between LLMs and human language understanding is heavily debated by researchers[1]. A conservative view is that LLMs are just very good at next-word prediction, unrelated to any real understanding of language. A chatbot like ChatGPT may seem to have a confident answer to everything, but it also makes simple factual and conceptual mistakes. This is arguably because LLMs have no real experiences and no understanding of the real world, in a non-linguistic way. They learn 'form' of language but no meaning, as argued in an influential paper from 2020 by Emily Bender and Alexander Koller[2]. On the other hand, the way language is handled in human brains will incorporate at least some sort of next-word prediction and there may be shared computational principles between LLMs and human language[3].

It is often pointed out in this debate on 'understanding' and LLMs that the models lack grounding in the physical world. But is sensory grounding really needed for meaning and understanding? This fundamental question was debated by six experts in machine learning, cognitive science, neuroscience, philosophy and linguistics at a recent conference on the philosophy of deep learning. The answer was, of course, far from straightforward. One of the panelists, Ellie Pavlick from Brown University and Google AI, pointed out that much of human understanding and knowledge is transferred by language alone and it may be possible to have a good understanding of the world without sensory grounding. Her group published a study in 2021 reporting that GPT-3 can learn concepts such as 'north' and 'left' in a grid world[4]. They reasoned that it is possible for a model to devise a conceptual structure from text alone that looks like what a model would learn when it could interact in a grounded world.

A next step in the development of LLMs is to combine them with multimodal capabilities, including sensory input. OpenAI's GPT-4 has been trained as a multimodal model, but at the time of writing, the ability to analyse or even generate images has not been shown outside of the launch demo and is not available for the general public to use. Training on images in addition to text could either be seen as the solution to ground text more firmly in human experience, or it could just be seen as adding more ungrounded data. Adding sensory data such as in Google's PaLM-E model could bring a new level of grounding for LLMs.

These are clearly exciting times for large language models. The underlying approach – the combination of pre-training with transformer architecture – is a game changer for applications in many scientific research areas such as materials discovery[5], molecular property predictions[6] and protein design[7]. Other interesting developments are in improving the efficiency of LLMs by careful parameter tuning[8] or, rather than scaling the models up further, making them smaller while preserving similar capabilities; researchers from Stanford University developed the Alpaca model, a fine-tuned version of LLaMA that is trained with text that is generated by GPT-3, and that, the authors say, costs only US$600 to reproduce. A potential advantage of smaller models with explicit internal dialogues is that the reasoning to reach the output can be more easily explained.

With the whirlwind of developments that have both scientific and societal impact, it is challenging to see through the hype. In a recent preprint, Microsoft researchers reported on a range of experiments to demonstrate the powerful performance of GPT-4 and were sufficiently impressed to conclude that there are 'sparks of artificial general intelligence'[9]. The paper quickly came under fire by experts. LLMs are clearly capable of tackling a range of complex tasks, and the widely demonstrated possibility of harnessing the power of language provides exciting, surprising scientific opportunities – without reaching for the elusive idea of artificial general intelligence.

Published online: 24 April 2023

References

1. Mitchell, M. & Krakauer, D. C. *Proc. Natl Acad. Sci. USA* **120**, e2215907120 (2023).
2. Bender, E. M. & Koller, A. in *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics* 5185–5198 (2020).
3. Goldstein, A. et al. *Nat. Neurosci.* **25**, 369–380 (2022).
4. Patel, R. & Pavlick, E. in *International Conference on Learning Representations* (2022).
5. Kang, Y., Park, H. & Smit, B. et al. *Nat. Mach. Intell.* **5**, 309–318 (2023).
6. Ross, J., Belgodere, B. & Chenthamarakshan, V. et al. *Nat. Mach. Intell.* **4**, 1256–1264 (2022).
7. Castro, E., Godavarthi, A. & Rubinfien, J. et al. *Nat. Mach. Intell.* **4**, 840–851 (2022).
8. Ding, N., Qin, Y. & Yang, G. et al. *Nat. Mach. Intell.* **5**, 220–235 (2023).
9. Bubeck, S. et al. Preprint at *arXiv* https://doi.org/10.48550/arXiv.2303.12712 (2023).