Check for updates

Large-scale AI language systems display an emergent ability to reason by analogy

Analogical reasoning is a hallmark of human intelligence, as it enables us to flexibly solve new problems without extensive practice. By using a wide range of tests, we demonstrate that GPT-3, a large-scale artificial intelligence language model, is capable of solving difficult analogy problems at a level comparable to human performance.

This is a summary of:
Webb, T. et al. Emergent analogical reasoning in large language models. *Nat. Hum. Behav*. https://doi.org/10.1038/s41562-023-01659-w (2023).

Publisher's note
Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Published online: 4 August 2023

The question

Human reasoning is marked by a unique ability to flexibly solve new problems without extensive practice or training. This capacity depends in large part on analogical reasoning – the ability to make sense of an unfamiliar problem by comparing it to a familiar one[1]. A major question for cognitive science is how this process is carried out by the human brain. This problem is closely related to the questions of whether and how artificial intelligence (AI) systems might acquire a similar capacity for analogical reasoning. In particular, there has been substantial debate about whether deep learning systems (multi-layer networks of neuron-like units that learn gradually from experience) might eventually develop this capacity for flexible reasoning if given exposure to sufficient training data.

The discovery

We tested a version of Generative Pre-trained Transformer 3 (GPT-3)[2], a large-scale deep learning system that has been trained to generate human-like text, on a broad set of analogy problems. Most notably, this set included a novel text-based version of Raven's progressive matrices[3] (Fig. 1a,b), a visual analogy problem set that is commonly viewed as one of the best measures of human problem-solving abilities. In the original (visual) problems, an array of geometric figures that includes a blank entry is presented (Fig. 1a). The task is to use this pattern to 'fill in the blank' by selecting the correct missing tile from among a set of potential answers. We created a text-based version of these problems by converting the geometric figures to digits (Fig. 1b). We also tested GPT-3 on other analogy problems, some of which involved real-world concepts (for example, 'love:hate::rich:?') or entire stories (in which the task is to determine which of two target stories is most analogous to a source story). Importantly, we tested GPT-3 without any direct training on these problems, to mirror the human ability to solve analogy problems without extensive practice.

We found that GPT-3 matched or surpassed the performance of college students in most task settings (Fig. 1c). Moreover, GPT-3 showed a very similar pattern of error rates to that of human participants, as it displayed greater difficulty with problems that humans tend to find difficult (for example, problems that are governed by multiple rules or that require a greater degree of abstraction). These results suggest that as a consequence of learning to generate human-like text, GPT-3 has acquired a highly general capacity to solve analogy problems.

The implications

A common view amongst many cognitive scientists and AI researchers is that deep learning systems require extensive task-specific training and have a limited ability to generalize beyond the conditions that are experienced during training[4,5]. These results challenge this widely held view and suggest that it is possible for deep learning systems to acquire the ability to reason about new problems if given sufficiently broad training on a general-purpose task (that is, learning to generate human-like text).

It is important to note some limitations of these results. GPT-3 is purely text-based and therefore cannot solve analogy problems directly from visual inputs (such as the problem shown in Fig. 1a). GPT-3 also lacks a capacity for long-term memory and displayed a poor ability to reason about physical problem solving (as evidenced by its inability to solve a simple construction problem involving tool use). We also found that GPT-3 struggled (relative to human participants) to identify more abstract analogies between stories, although GPT-4 performed better than GPT-3.

A major question posed by this study is whether GPT-3 solves analogy problems using mechanisms similar to those used by the human brain. Although deep learning systems such as GPT-3 are loosely inspired by the brain (as they are composed of neuron-like processing units, arranged in a hierarchy of multiple layers), it is currently unclear how these systems might carry out the computational operations that are thought to underlie human analogical reasoning. However, unlike the human brain, the internal mechanisms of systems such as GPT-3 can – at least in principle – be directly probed. This feature suggests the possibility that such systems can serve as a kind of 'model organism' for understanding the neural basis of higher-order cognitive processes. To achieve this goal, it will be important to develop open-source versions of these systems that can be studied by cognitive scientists, together with the resources needed to evaluate them. This development would also enable scientists to improve our understanding of the strengths and limitations of such systems and ensure that they are safely deployed into society.

Taylor Webb
Department of Psychology, University of California, Los Angeles, CA, USA

EXPERT OPINION

"Analogy is a core human cognitive ability thought to require specific representational forms or bespoke computational architectures. These results challenge that position, as the computational model considered is a generic neural network designed for sequence learning and language processing and, therefore, should have impact in both the cognitive science and AI communities. The experiments are rigorous and principled, and this work would constitute a beacon of good scientific practice in machine learning research." **Felix Hill, Google DeepMind, London, UK.**

FIGURE

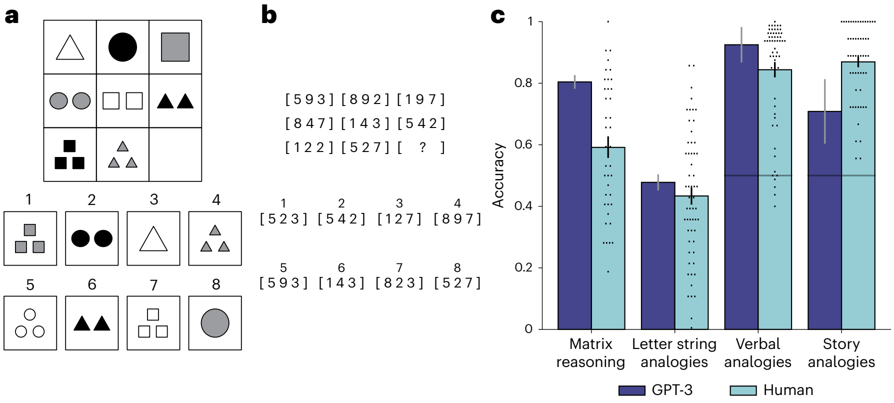

Fig. 1 | Analogy test results in GPT-3 and human participants. a,b, Example problems depicting the structure of Raven's progressive matrices (**a**) and a text-based version (**b**) used to evaluate GPT-3. The correct answers are 5 (**a**) and 7 (**b**). **c**, GPT-3 surpassed the average performance of human participants (college students) on text-based matrix reasoning problems and also showed strong performance on other analogy problems. Dots represent individual participants; bars indicate mean ± s.e.m. © 2023, Webb, T. et al.

BEHIND THE PAPER

My co-authors and I have a longstanding interest in the computational and neural mechanisms that support abstract reasoning, with a particular interest in the strengths and limitations of neural network models (such as GPT-3). In previous work[6], we have emphasized the need to combine standard neural network methods with more structured reasoning operations to match the flexibility of human reasoning. Thus, we were very surprised by the ability of GPT-3 to solve analogy problems, which seemed to suggest an emergent capacity for abstract reasoning. To better understand this observation, we developed novel test materials (it was important that these materials be novel to ensure that GPT-3 had not been trained on them) and systematically evaluated GPT-3 in comparison with human behaviour. The results strengthened this initial conclusion, with subsequent variants of GPT-3 and GPT-4 displaying even stronger performance. **T.W.**

REFERENCES

1. Holyoak, K.J. in *Oxford Handbook of Thinking and Reasoning* (eds Holyoak, K. J. & Morrison, R. G.) 234–259 (Oxford Univ. Press, 2012).
   A book chapter that summarizes work in cognitive science on analogical reasoning.
2. Brown, T. et al. Language models are few-shot learners. In *Adv. Neural Information Processing Systems 33* (eds Larochelle, H. et al.) 1877–1901 (Curran Associates, 2020).
   This paper describes GPT-3, the AI system that was evaluated in the present work.
3. Raven, J. C. *Progressive Matrices: A Perceptual Test of Intelligence, Individual Form* (Lewis Raven, 1938).
   A visual analogy problem set that is commonly used as a test of problem-solving skills.
4. Lake, B. M. et al. Building machines that learn and think like people. *Behav. Brain Sci.* **40**, E253 (2017).
   A review and perspective that characterizes some limitations of deep learning systems.
5. Mitchell, M. Abstraction and analogy-making in artificial intelligence. *Ann. NY Acad. Sci.* **1505**, 79–101 (2021).
   A review that summarizes work in AI on analogical reasoning.
6. Lu, H., Ichien, N. & Holyoak, K. J. Probabilistic analogical mapping with semantic relation networks. *Psychol. Rev.* **129**, 1078 (2022).
   An example of work that combines deep learning with structured reasoning operations.

FROM THE EDITOR

"Webb et al. show that new AI language models, such as GPT-3, are able to solve analogical reasoning problems at a human-like level of performance. This result is notable because these models were never explicitly trained to do such tasks, and because analogical reasoning is widely considered to be a core part of human intelligence." **Jamie Horder, Senior Editor, *Nature Human Behaviour*.**