nature human behaviour

Article

https://doi.org/10.1038/s41562-023-01659-w

Emergent analogical reasoning in large language models

Received: 18 December 2022

Accepted: 16 June 2023

Published online: 31 July 2023

Check for updates

Taylor Webb [1] ✉, Keith J. Holyoak [1] & Hongjing Lu [1,2]

The recent advent of large language models has reinvigorated debate over whether human cognitive capacities might emerge in such generic models given sufficient training data. Of particular interest is the ability of these models to reason about novel problems zero-shot, without any direct training. In human cognition, this capacity is closely tied to an ability to reason by analogy. Here we performed a direct comparison between human reasoners and a large language model (the text-davinci-003 variant of Generative Pre-trained Transformer (GPT)-3) on a range of analogical tasks, including a non-visual matrix reasoning task based on the rule structure of Raven's Standard Progressive Matrices. We found that GPT-3 displayed a surprisingly strong capacity for abstract pattern induction, matching or even surpassing human capabilities in most settings; preliminary tests of GPT-4 indicated even better performance. Our results indicate that large language models such as GPT-3 have acquired an emergent ability to find zero-shot solutions to a broad range of analogy problems.

Analogical reasoning is at the heart of human intelligence and creativity. When confronted with an unfamiliar problem, human reasoners can often identify a reasonable solution through a process of structured comparison with a more familiar situation[1]. This process is an essential part of human reasoning in domains ranging from everyday problem-solving[2] to creative thought and scientific innovation[3]. Indeed, tests of analogical reasoning ability are uniquely effective as measures of fluid intelligence: the capacity to reason about novel problems[4,5].

Recently, there has been considerable debate about whether and how a capacity for analogical thought might be captured in deep learning systems[6]. Much of this recent work has focused on training neural networks on very large datasets (sometimes containing millions of problems)[7,8]. Though this is a challenging task that has spurred the development of some interesting approaches[9–12], it does not address the issue of whether analogical reasoning can emerge zero-shot (that is, without direct training), the capacity most central to human thought.

An alternative approach, also based on deep learning, involves large language models (LLMs)[13]. LLMs have recently sparked great interest (and controversy) for their potential to perform few-shot, and even zero-shot, reasoning. These models employ relatively generic neural network architectures with up to billions of parameters, and are trained using a simple predictive objective (predicting the next token in a sequence of text) with massive web-based text corpora consisting of billions of tokens. Though there is considerable debate about the capabilities of these models[14], a potential advantage is their ability to solve problems with little direct training, sometimes requiring only a few examples, or even a simple task instruction (typically without any updating of model parameters). This feature raises the question of whether LLMs might be capable of human-like, zero-shot analogical reasoning.

In this Article, to answer this question, we evaluated the language model Generative Pre-trained Transformer (GPT)-3 (ref. 13) on a range of zero-shot analogy tasks, and performed direct comparisons with human behaviour. These tasks included a novel text-based matrix reasoning task based on the rule structure of Raven's Standard Progressive Matrices (SPM)[15], a visual analogy problem set commonly viewed as one of the best measures of fluid intelligence[5]. Unlike the original visual SPM problems, our Digit Matrices task was purely text based so that it could be used to evaluate GPT-3's ability to induce abstract rules (though not the ability to do so directly from visual inputs). Strikingly, we found that GPT-3 performed as well as or better than college students in most conditions, despite receiving no direct

[1]Department of Psychology, University of California, Los Angeles, CA, USA. [2]Department of Statistics, University of California, Los Angeles, CA, USA. ✉e-mail: taylor.w.webb@gmail.com

training on this task. GPT-3 also displayed strong zero-shot performance on letter string analogies[16], four-term verbal analogies[17–20] and identification of analogies between stories[21–23]. These results add to the growing body of work characterizing the emergent capabilities of LLMs[24–28], and suggest that the most sophisticated LLMs may already possess an emergent capacity to reason by analogy.

Results

We evaluated the language model GPT-3 on a set of analogy tasks, and compared its performance with human behaviour. GPT-3 is a large-scale (175B parameters), transformer-based[29] language model developed by OpenAI[13]. The original base model was trained on a web-based corpus of natural language consisting of over 400 billion tokens, using a training objective based on next-token prediction (given a string of text, the model is trained to predict the token most likely to appear next). A number of variants on this base model have since been developed by fine-tuning it in various ways. These include training the model to generate code[30], and training it to respond appropriately to human prompts, using either supervised learning or reinforcement learning from human feedback[31]. Our evaluation focused on the most recent model variant, text-davinci-003 (here referred to simply as 'GPT-3'), which was the first to incorporate reinforcement learning from human feedback (along with the concurrently released, but distinct, ChatGPT model). We found that text-davinci-003 displayed particularly strong performance on our analogy tasks, but earlier model variants also performed well in some task settings, suggesting that multiple factors contributed to text-davinci-003's analogical capabilities (Supplementary Figs. 1–3). For further discussion, see Supplementary Section 2.

Our evaluation featured four separate task domains, each designed to probe different aspects of analogical reasoning: (1) text-based matrix reasoning problems, (2) letter string analogies, (3) four-term verbal analogies and (4) story analogies. For each task domain, we performed a direct comparison with human behaviour, assessing both overall performance and error patterns across a range of conditions relevant to human analogical reasoning. Figure 1 shows a summary of these results. We also performed a qualitative analysis of GPT-3's ability to use analogical reasoning to solve problems.

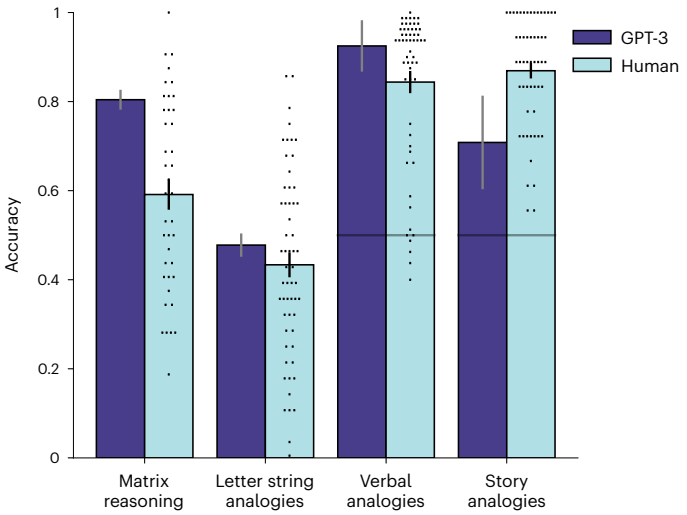

Fig. 1 | Summary of results. Matrix reasoning results show average accuracy on all problems in Digit Matrices problem set, a novel text-based matrix reasoning task designed to emulate Raven's SPM problems[15]. Note that the Digit Matrices were purely text based, and therefore do not test for the ability to perform abstract reasoning directly over visual inputs, as in the original SPM. Letter string results show average performance for novel letter string analogy problem set, based on problems from Hofstadter and Mitchell[16]. Both matrix reasoning and letter string results reflect performance on generative task. Verbal analogy results show average performance on UCLA VAT[19]. Story analogy problems involved identification of analogous stories based on higher-order relations, using materials from Gentner et al.[23]. Both verbal and story analogy results reflect multiple-choice accuracy, with chance performance indicated by grey horizontal line. Chance performance for the two generative tasks (matrix reasoning and letter string analogies) is close to zero, due to the very large space of possible generative responses. Black error bars represent standard error of the mean for average performance across participants. Each dot represents accuracy for a single participant (matrix reasoning, $N = 40$; letter string analogies, $N = 57$; verbal analogies, $N = 57$; story analogies, $N = 54$). Grey error bars represent 95% binomial CIs for average performance across multiple problems.

Matrix reasoning problems

We designed a text-based matrix reasoning task, the Digit Matrices, to emulate the structure of Raven's SPM[15]. The task is illustrated in Fig. 2. The dataset was structured similarly to the work of Matzen et al.[32], who created, and behaviourally validated, a visual matrix reasoning dataset with the same rule structure as the original SPM. The Digit Matrices dataset thus has a similar rule structure to SPM, but is guaranteed to be novel for both humans and LLMs.

Digit Matrices problems consisted of either digit transformations (Fig. 2b–e) or logic problems (Fig. 2f,g). Transformation problems were defined on the basis of a set of three rule types—constant (Fig. 2c), distribution-of-3 (Fig. 2d) and progression (Fig. 2e)—and consisted of one or more rules per problem. When multiple rules were present (Fig. 2b), each rule was bound to a different spatial location within each cell (for example, one rule was bound to the left digit in each cell, and another rule was bound to the right digit). Logic problems were defined on the basis of set relations—OR, AND and XOR—and involved only a single rule per problem. In some logic problems, the corresponding elements were spatially aligned (Fig. 2f), whereas in others they were permuted (Fig. 2g). We hypothesized that spatial alignment would be beneficial when solving the problems via analogical mapping, as it should highlight the isomorphism[33]. Digit Matrices problems were presented to GPT-3 without any prompt or in-context task examples.

Figure 3 shows zero-shot performance on the Digit Matrices problems for GPT-3 and human participants ($N = 40$, University of California, Los Angeles (UCLA) undergraduates). GPT-3 surpassed the average level of human performance on all problem types, both when generating answers directly (Fig. 3a; logistic regression, main effect of GPT-3 versus human participants: odds ratio 1.88, $P = 0.005$, 95% confidence interval (CI) 1.21–2.91), and when selecting from a set of answer choices (Fig. 3b; main effect of GPT-3 versus human participants: odds ratio 6.27, $P = 2.3 \times 10^{-8}$, 95% CI 3.28–11.99). It is worth emphasizing, however, that participants displayed a range of performance levels on this task, with some participants outperforming GPT-3 (indeed, the best participant answered every problem correctly).

In addition to showing strong overall performance, GPT-3's pattern of performance across problem subtypes was similar to that observed in human participants (correlation analysis: $r(30) = 0.39$, $P = 0.027$). This correlation was driven both by the pattern of performance across major problem types (one-rule, two-rule, three-rule, and logic problems; main effect of problem type on generative accuracy: odds ratio 0.5, $P = 2 \times 10^{-16}$, 95% CI 0.44–0.56; main effect of problem type on multiple-choice accuracy: odds ratio 0.56, $P = 2 \times 10^{-16}$, 95% CI 0.5–0.64), and by differences within each problem type. Problems with progression rules were more difficult than those without them (Fig. 3c; main effect of progression versus no progression, human participants: odds ratio 0.41, $P = 0.0001$, 95% CI 0.24–0.69; GPT-3: odds ratio 0.07, $P = 1.9 \times 10^{-5}$, 95% CI 0.02–0.24); for multi-rule problems, performance was negatively correlated with the number of unique rules in each problem, even when holding constant the number of total rules (Fig. 3d; main effect of number of unique rules, human participants: odds ratio 0.61, $P = 0.0047$, 95% CI 0.44–0.86; GPT-3: odds ratio 0.25, $P = 3 \times 10^{-10}$, 95% CI 0.17–0.39); and logic problems were more difficult

https://doi.org/10.1038/s41562-023-01659-w

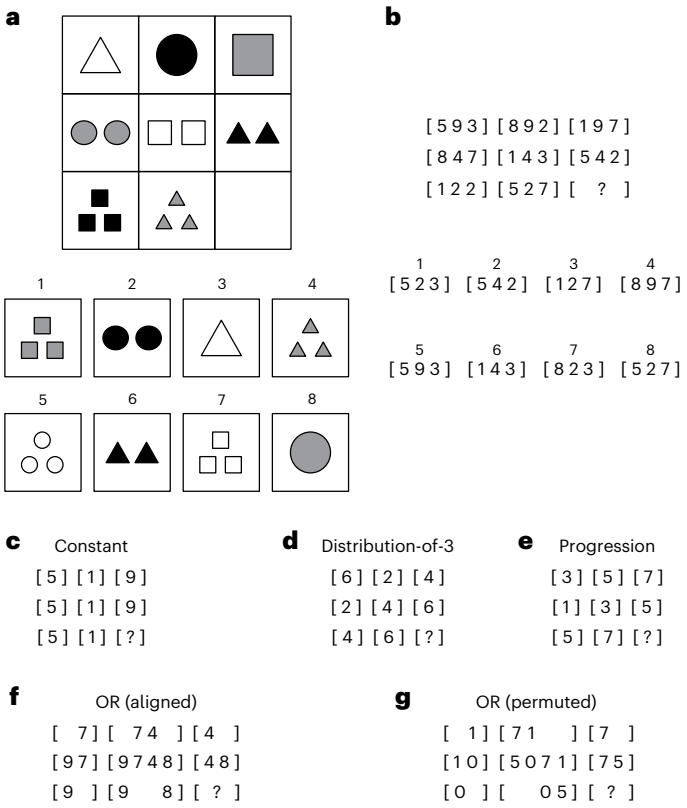

Fig. 2 | Matrix reasoning problems. a, Example problem depicting structure of Raven's Progressive Matrices[15]. Problems consist of a 3 × 3 matrix populated with geometric forms, in which each row or column is governed by the same set of abstract rules. Problem solvers must identify these rules, and use them to infer the missing cell in the lower right, by selecting from the set of eight choices below. **b**, Example problem illustrating the novel Digit Matrices problem set. Problems consist of a 3 × 3 matrix, in which each cell is demarcated by brackets, and populated by digits. The problems are governed by the same rule structure as Raven's SPM. The example problems in **a** and **b** are structurally isomorphic (that is, governed by the same set of rules). The reader is encouraged to derive the solution to each problem. The solutions to both problems are given in Supplementary Section 1. Problems were governed either by one or more transformation rules (**b**–**e**), or by a single logic rule (**f** and **g**). **c**, Constant rule: same digit appears across either rows or columns. **d**, Distribution-of-3 rule: same set of three digits appears in each row or column, but with order varied. **e**, Progression rule: digits either increase or decrease, by values of 1 or 2, across rows or columns. In the example shown here, digits increase by 2 across rows. **f**, OR rule: the set of digits present in a particular row or column are defined as the union of the sets present in the other rows or columns. In the illustrated example, the digits in the second column are formed from the union of the sets in the first and third columns. This example illustrates how the spatial alignment of the corresponding elements can make it easier to intuitively grasp the underlying rule. **g**, More challenging logic problem governed by same rule (OR), but in which the corresponding elements are spatially permuted. Other logic problems were governed either by an AND rule or an XOR rule (not pictured).

when the corresponding elements were spatially permuted versus aligned (Fig. 3e; main effect of spatial alignment, human participants: odds ratio 0.52, $P = 0.0017$, 95% CI 0.35–0.79; GPT-3: odds ratio 0.06, $P = 2 \times 10^{-11}$, 95% CI 0.03–0.14). These effects replicate well-known characteristics of human analogical reasoning: problems defined by relations (for example, progression) are typically more difficult than problems defined by the features of individual entities (for example, constant or distribution-of-3)[32,34]; problem difficulty is typically driven by the degree of relational complexity, as defined by the number of unique relations[35]; and analogical mapping is easier when a greater number of constraints support the correct mapping (as is the case in the spatially aligned logic problems)[33]. GPT-3's pattern of performance thus displayed many of the characteristics of a human-like analogical mapping process. We also found that GPT-3 was sensitive to contextual information in ways that both improved and impaired its performance, similar to human reasoners (Supplementary Fig. 4).

It is important to highlight the differences between the Digit Matrices and traditional visual matrix reasoning problems. To solve visual matrix reasoning problems, pixel-level inputs must be parsed into objects, and visual attributes (shape, size and so on) must be disentangled. In the Digit Matrices, the text-based inputs are already parsed and disentangled, essentially providing GPT-3 (which is not capable of visual processing) with pseudo-symbolic inputs. Interestingly, despite these differences, we found that overall error rates for human participants were very similar for the Digit Matrices versus the original image-based SPM problem set, and showed a similar pattern across problem types (Fig. 4). These results suggest that, while the Digit Matrices do not engage the visual processes involved in traditional SPM problems (that is, deriving disentangled representations from pixel-level inputs), they probably engage a similar set of core reasoning processes (that is, inducing abstract rules from those representations). More generally, performance on verbal, visuospatial and mathematical analogy problems is known to be highly correlated for people[5]. Accordingly, GPT-3's success on the Digit Matrices can be taken as evidence that it has acquired core capabilities underlying analogy, though it will be important in future work to investigate how these reasoning processes might be integrated with visual processing.

Letter string analogies

A central feature of human analogical reasoning is its flexibility. Human reasoners are capable of identifying abstract similarities between situations even when these situations are superficially quite different. Often this involves a process of re-representation, in which an initial problem representation is revised so as to facilitate the discovery of an analogy[36–38].

Hofstadter and Mitchell[16,39] introduced the letter string analogy domain to evaluate computational models of analogical reasoning, with a particular emphasis on the process of re-representation. The basic problem structure is illustrated in Fig. 5a. In this example, the source string 'a b c d' has been transformed by converting the final letter to its successor, resulting in the string 'a b c e'. This transformation must be identified, and then applied to the target string 'i j k l', yielding the answer 'i j k m'.

Though this example is simple, letter string problems can be made quite complex by introducing various generalizations between the source and target strings. For instance, the target may involve groups of letters rather than individual letters (for example, 'i i j j k k l l'), or may involve a sequence with a reversed order relative to the source (for example, 'l k j i'). In these cases, the transformation identified in the source (for example, a successor transformation applied to the final letter in the sequence) must be generalized to an analogous transformation (for example, a successor transformation applied to the final group of letters, or a predecessor transformation applied to the first letter). This feature makes letter string analogy problems well suited to test the capacity for re-representation.

To evaluate GPT-3, we created a novel letter string problem set (Fig. 5), and carried out a systematic comparison with human participants ($N = 57$, UCLA undergraduates). The problem set involved a range of different transformation (Fig. 5d) and generalization types (Fig. 5e). Each transformation type could be combined with any generalization type, and multiple generalization types could be combined together to yield more challenging problems (Fig. 5b). Problems were presented to GPT-3 along with a prompt ('Let's try to complete the pattern:'), using a format similar to the Digit Matrices.

https://doi.org/10.1038/s41562-023-01659-w

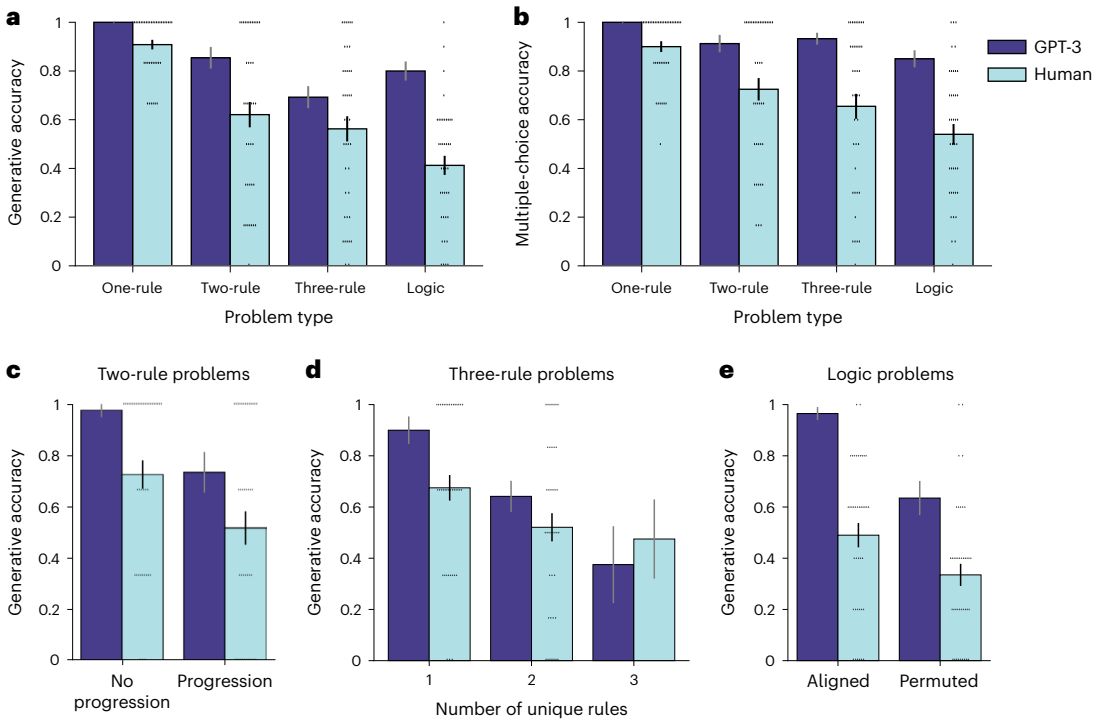

Fig. 3 | Matrix reasoning results. GPT-3 matched or exceeded human performance for zero-shot Digit Matrices. **a**, Generative accuracy for major problem types, including transformation problems with between one and three rules, and logic problems. **b**, Multiple-choice accuracy for major problem types. **c**, Two-rule problems with at least one progression rule were more difficult than those without. **d**, For three-rule problems, performance was a function of the number of unique rules. **e**, Spatially permuted logic problems were more difficult than spatially aligned problems. Human results reflect average performance for $N = 40$ participants (UCLA undergraduates). Black error bars represent standard error of the mean across participants. Each dot represents accuracy for a single participant. Grey error bars represent 95% binomial CIs for average performance across multiple problems. Note that the rightmost bar in **d** does not show individual scores because each participant only completed a single problem with three unique rules.

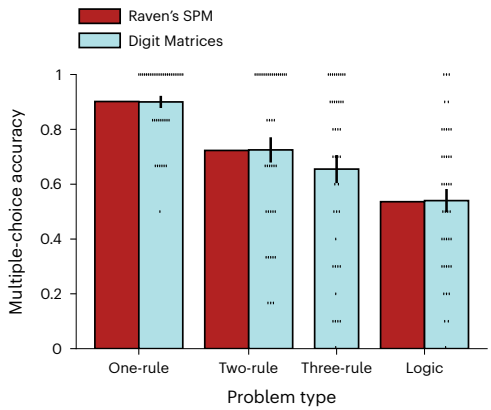

Fig. 4 | Human performance for Digit Matrices versus Raven's SPM. SPM[15] does not contain three-rule problems, but performance was very similar across one-rule, two-rule and logic problems. SPM results reflect average performance for $N = 80$ participants (data from ref. 32). Digit Matrices results reflect average performance for $N = 40$ participants. Error bars represent standard error of the mean. Each dot represents accuracy for a single participant.

Figure 6 shows the results of this evaluation. GPT-3 showed stronger overall performance than human participants (Fig. 6a; logistic regression, main effect of GPT-3 versus human participants: odds ratio 1.76, $P = 6.3 \times 10^{-5}$, 95% CI 1.34–2.31), an effect that was driven primarily by stronger performance on zero-generalization problems (main effect of GPT-3 versus human participants for zero-generalization problems: odds ratio 1.76, $P = 0.0007$, 95% CI 1.27–2.46). Performance was strongly affected by the number of generalizations in both GPT-3 and human participants (main effect of number of generalizations, GPT-3: odds ratio 0.51, $P = 2 \times 10^{-16}$, 95% CI 0.45–0.57; human participants: odds ratio 0.66, $P = 5.9 \times 10^{-16}$, 95% CI 0.6–0.73). GPT-3 and human participants also showed similar error patterns across transformation types (Fig. 6b) and generalization types (Fig. 6c), as quantified by a correlation analysis for accuracy across different problem subtypes ($r(39) = 0.7$, $P = 3.6 \times 10^{-7}$).

We also investigated a novel variant on letter string problems involving generalization from letters to real-world concepts (Fig. 5c). GPT-3 showed strong performance on these problems, though with some discrepancies for different transformation types (Fig. 6d). These results suggest that GPT-3 has developed an abstract notion of successorship that can be flexibly generalized between different domains (for example, alphabetic successorship versus temperature successorship).

One important caveat is that GPT-3's performance on this task was somewhat sensitive to the way in which problems were formatted. For instance, performance suffered when no prompt was provided (Supplementary Fig. 5a), or when problems were presented in the form of a complete sentence (Supplementary Fig. 5b). However, even in these cases, GPT-3's zero-shot performance was both within the range of human participants (within one standard deviation), and closely matched the pattern of human performance across problem types (correlation analysis, no prompt: $r(39) = 0.6$, $P = 5.3 \times 10^{-5}$, sentence format: $r(39) = 0.76$, $P = 4.2 \times 10^{-6}$).

Four-term verbal analogies

Though matrix reasoning and letter string analogies involve a high degree of relational complexity, one limitation is that they consist of highly constrained, synthetic relations, such as alphabetic or numerical successorship. GPT-3's ability to solve problems involving more

https://doi.org/10.1038/s41562-023-01659-w

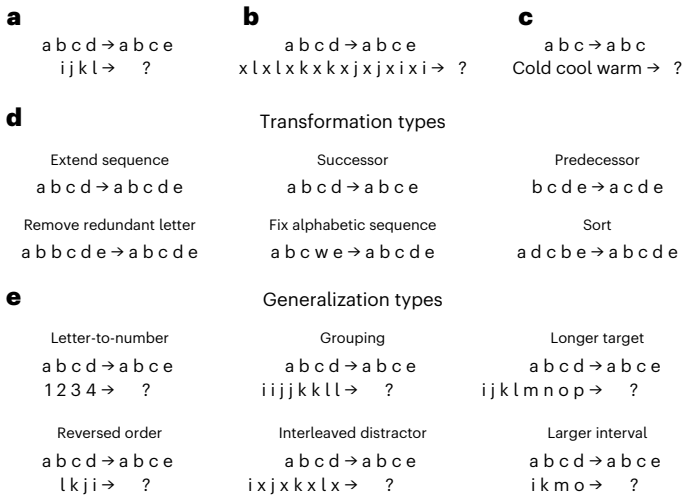

Fig. 5 | Letter string analogy problems. Transformation between source strings must be identified and applied to target string. Mapping between source and target may involve one or more generalizations. **a**, Easy problem involving zero generalizations. **b**, Difficult problem involving three generalizations (grouping, reversed order and interleaved distractors). **c**, Problem involving generalization from letters to real-world concepts. **d**, Transformations were sampled from set of six possible types: sequence extension, successor transformation (applied to the last letter in the string), predecessor transformation (applied to the first letter in the string), removal of a redundant letter, 'fixing' an alphabetic sequence (replacing an out-of-place letter) and sorting. **e**, Generalizations were sampled from set of six possible types: letter-to-number, grouping, longer target string, reversed order, interleaved distractors and larger interval.

real-world concepts (for example, 'a b c → a b d, cold cool warm → ?') suggests that its analogical capabilities may not be limited to such artificial settings. To further evaluate GPT-3's capacity to reason about real-world relational concepts, we tested it on four-term verbal analogy problems involving a broader range of semantic relations.

We evaluated GPT-3 on four separate datasets[17–20]. To the best of our knowledge, these constitute an exhaustive set of four-term verbal analogy problems for which human behavioural data are available[40]. Each dataset contains a series of four-term analogy problems in the form 'A:B::C:?', together with a set of answer choices (that is, potential choices of D). For each problem, GPT-3 was evaluated by presenting the problem together with each potential answer choice, and selecting the option for which GPT-3 assigned a higher log probability. The problem and GPT-3's choice were then appended to the context window for the next problem, thereby simulating any contextual effects that might arise when solving multiple problems in a row, as human participants typically do.

Figure 7 shows the results for all datasets. GPT-3 performed as well as or better than human participants (minimum education level of high school graduation, located in the United States and recruited using Amazon Mechanical Turk) on the UCLA Verbal Analogy Test (VAT)[19], involving categorical, functional, antonym and synonym relations (Fig. 7a), and on a dataset from Sternberg and Nigro[17] involving these same four relation types and linear order relations (Fig. 7b). On a dataset of Scholastic Assessment Test (SAT) analogy problems from Turney et al.[18], GPT-3 surpassed the estimated average level of performance for high school students taking the SAT (Fig. 7c). GPT-3 also showed performance in the same range as human participants (though numerically weaker) on a problem set from Jones et al.[20] involving categorical, compositional and causal relations (Fig. 7d).

In addition to displaying generally strong performance on these problem sets, GPT-3 also displayed sensitivity to semantic content similar to that observed in human participants. In the dataset from Jones et al.[20] (Fig. 7d), participants performed worse on problems in which the analogues were semantically distant (that is, the A and B terms had low semantic similarity to C and D), an effect that was also displayed by GPT-3 (logistic regression, effect of semantic distance for GPT-3: odds ratio 3.24, $P = 0.0165$, 95% CI 1.24–8.5). These results align with a more general phenomenon in which human reasoning is facilitated by semantically meaningful or coherent content[24,41].

Story analogies

Human reasoners are able not only to form analogies between individual concepts, but can also identify correspondences between complex real-world events, involving many entities and relations. When making such comparisons, human reasoning is especially sensitive to higher-order relations—relations between relations—notably causal relations between events. Such higher-order relations play a central role in some cognitive theories of analogy[42], and it is thus important to establish whether GPT-3 displays a similar sensitivity to them.

To address this question, we tested GPT-3 on a set of story analogies from Gentner et al.[23]. In each set, a source story is compared with two potential target stories, each of which is matched with the source story in terms of first-order relations, but only one of which shares the same causal relations as the source (for examples, see Methods). Gentner et al. found that human participants rated the target stories as more similar when they shared the same causal relations as the source story. These problems are further defined by two different comparison conditions. In the near analogy condition (referred to as 'literal similarity' versus 'mere appearance' by Gentner et al.), the target stories also share the same basic entities as the source story, making for a less abstract, within-domain comparison. In the far analogy condition (referred to as 'true analogy' versus 'false analogy' by Gentner et al.), the target stories involve different entities from the source story, but share first-order relations, resulting in a more challenging, cross-domain comparison.

To facilitate a direct comparison with GPT-3, we performed a new behavioural study with these materials. For each source story, participants indicated which of two target stories was more analogous. Both GPT-3 and human participants ($N = 54$, UCLA undergraduates) showed a sensitivity to higher-order relations (Fig. 8), most often selecting the target story that shared causal relations with the source (combined near and far analogy; GPT-3, binomial test: $P = 0.0005$; human participants, one-sample t-test: $t(53) = 21.3$, $P = 1.1 \times 10^{-27}$; null hypothesis for both tests is chance-level performance of 0.5). This effect was significant for both GPT-3 and human participants in the near analogy condition (GPT-3, binomial test: $P = 0.0039$; human participants, one-sample t-test: $t(53) = 21.5$, $P = 8.5 \times 10^{-28}$), but only human participants showed a significant effect in the far analogy condition (GPT-3, binomial test: $P = 0.065$; human participants, one-sample t-test: $t(53) = 16.7$, $P = 9.3 \times 10^{-23}$).

Unlike the other task domains considered in the present work, this was a case in which college students clearly outperformed GPT-3 (logistic regression, main effect of GPT-3 versus human participants: odds ratio 0.37, $P = 0.0003$, 95% CI 0.21–0.63). Indeed, a substantial proportion of participants (15/54) selected the analogous story on every trial. However, in an initial investigation of GPT-4 (ref. 43), we found that it displays stronger performance on this task, more robustly picking the analogous story even in the far analogy condition, and displaying nearly perfect performance in the near analogy condition (Supplementary Fig. 6 and Supplementary Section 4.3). It therefore seems likely that further scaling of LLMs will enhance their sensitivity to causal relations.

Analogical problem-solving

In everyday thinking and reasoning, analogical comparisons are often made for the purpose of achieving some goal, or solving a novel problem. Thus far, our tests of GPT-3 have assessed its capacity for

https://doi.org/10.1038/s41562-023-01659-w

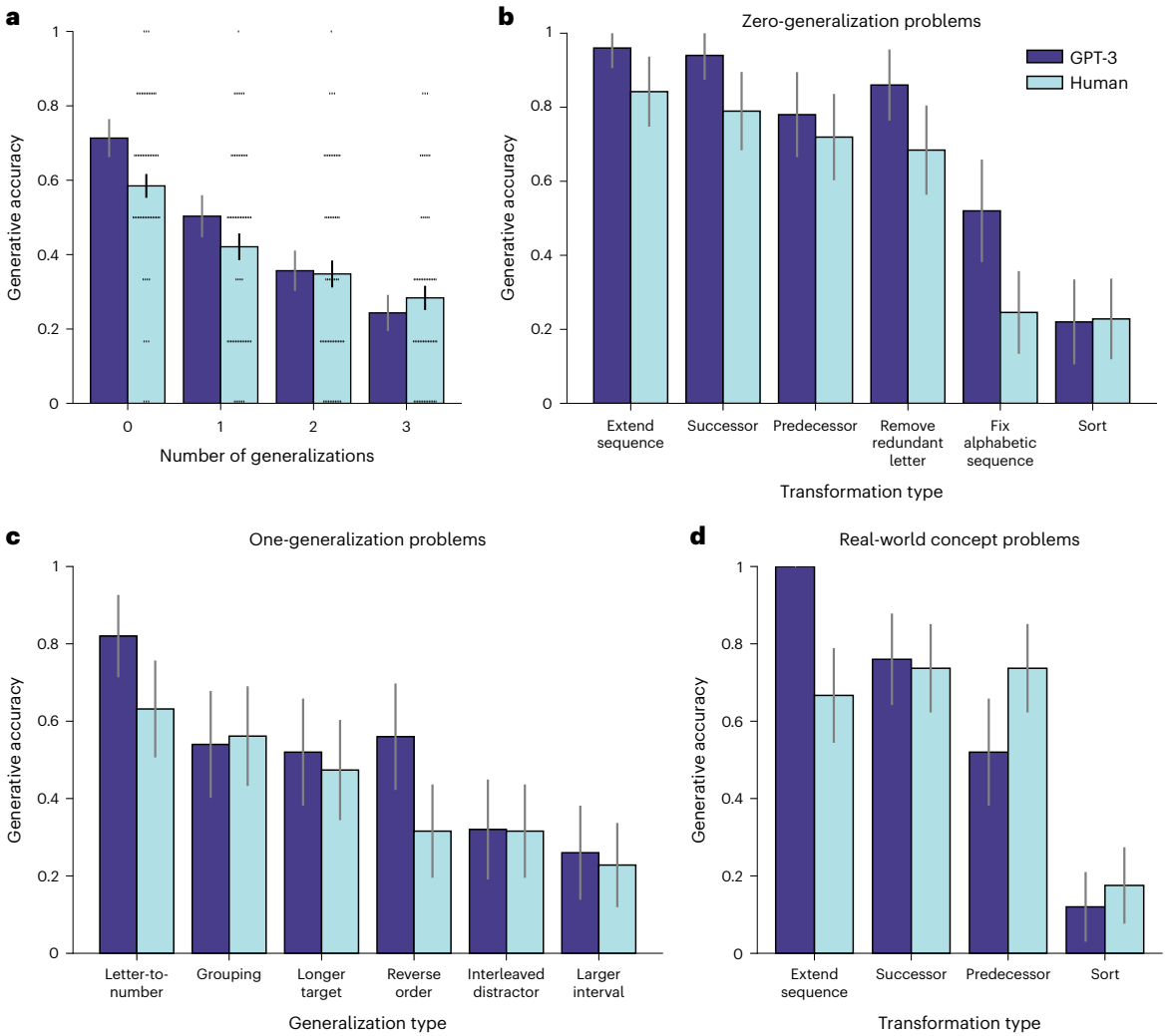

Fig. 6 | Letter string analogy results. GPT-3 displayed strong performance on letter string problems, and showed a similar pattern to human participants across conditions. **a**, GPT-3 and human performance as a function of the number of generalizations between source and target. **b**, Performance on zero-generalization problems as a function of transformation type. **c**, Performance on one-generalization problems as a function of generalization type. **d**, Performance on problems requiring generalization from letters to real-world concepts. Human results reflect average performance for $N = 57$ participants (UCLA undergraduates). Black error bars represent standard error of the mean across participants. Each dot represents accuracy for a single participant. Note that **b**–**d** do not show individual participant results because each participant only completed one problem in each condition. Grey error bars represent 95% binomial CIs for average performance across multiple problems.

identifying analogies in text-based inputs with varying formats, but can GPT-3 also use these analogies to derive solutions to novel problems, as human reasoners do?

As a preliminary investigation of this issue, we performed a qualitative evaluation using a paradigm developed by Gick and Holyoak[21]. In that paradigm, participants are presented with a target problem in the form of a story. In the original study, Duncker's radiation problem was used[44]. In that problem, a doctor wants to use radiation to destroy a malignant tumour, but destroying the tumour with a single high-intensity ray will also damage the surrounding healthy tissue. The solution—to use several low-intensity rays that converge at the site of the tumour—is rarely identified spontaneously, but participants are more likely to discover this solution when they are first presented with an analogous source story. In the original study, the source story involved a general who wants to capture a fortress ruled by an evil dictator, but cannot do so by sending his entire army along a single road, which would trigger landmines. The general instead breaks his army up into small groups that approach the fortress from multiple directions, thus avoiding triggering the mines.

We first presented GPT-3 with the target problem in isolation. GPT-3 proposed a solution that involved injecting a radiation source directly into the tumour, rather than identifying the intended solution on the basis of the convergence of multiple low-intensity radiation sources (Supplementary Section 5.1). However, when first presented with the general story, followed by the target problem, GPT-3 correctly identified the convergence solution (Supplementary Section 5.2). GPT-3 was further able to correctly explain the analogy, and to identify the specific correspondences between the source story and target problem when prompted (for example, general ↔ doctor, dictator ↔ tumour, army ↔ rays). We also found similar results when using distinct source analogues taken from another study[45] (Supplementary Section 5.3).

In a more challenging version of this paradigm, participants were first presented with both the general story and two other non-analogous stories intended to serve as distractors. In this context, human participants were much less likely to identify the convergence solution. However, when given a prompt to explicitly consider the previously presented stories when trying to solve the radiation problem,

https://doi.org/10.1038/s41562-023-01659-w

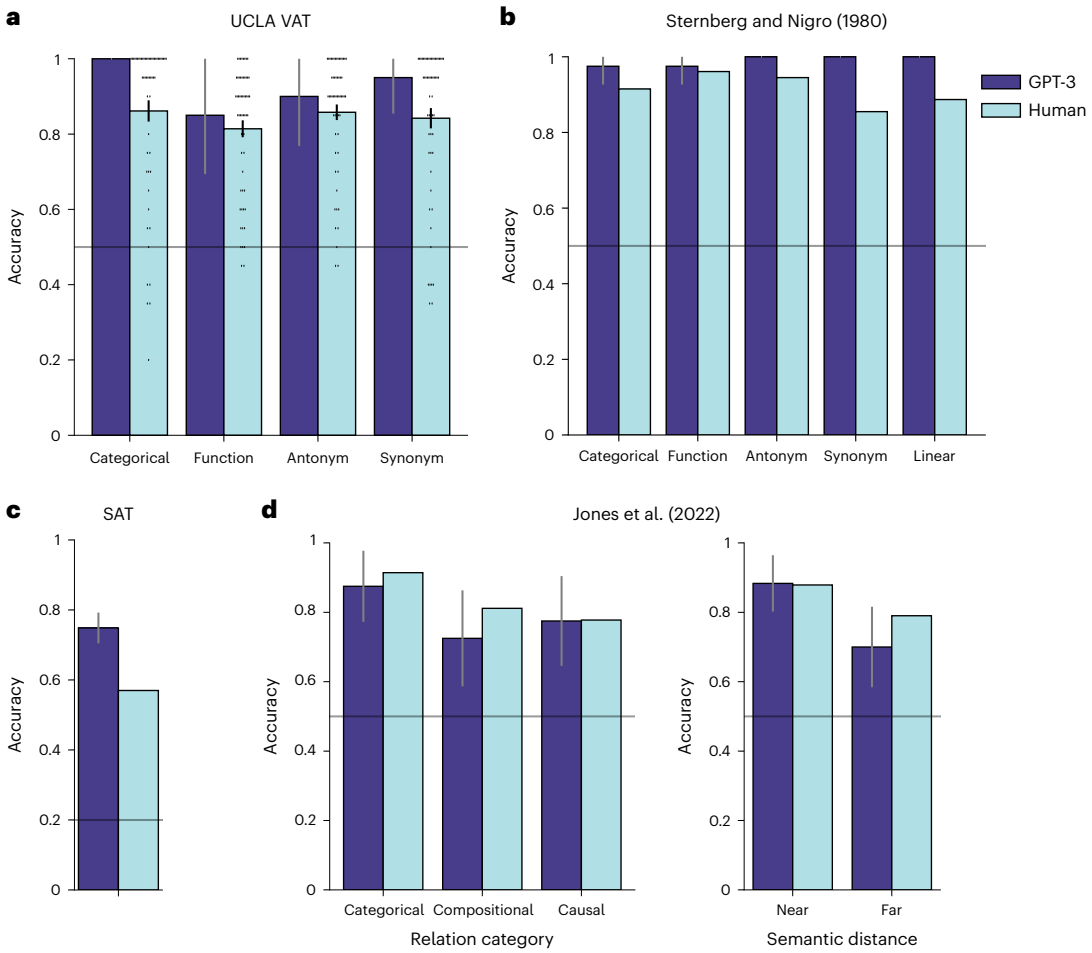

Fig. 7 | Verbal analogy results. a, Results for UCLA VAT[19]. Human results reflect average performance for $N = 57$ participants. Black error bars represent standard error of the mean. Each dot represents accuracy for a single participant. **b**, Results for dataset from Sternberg and Nigro[17]. Human results reflect average performance for $N = 20$ participants. **c**, Results for SAT analogy problems from Turney et al.[18]. These problems involve five answer choices, and thus chance performance is 20%. Human results reflect an estimate of the average performance for high school students taking the SAT (for details, see ref. 70). **d**, Results for dataset from Jones et al.[20]. Human results reflect average performance for $N = 241$ participants. Grey error bars represent 95% binomial CIs for average performance across multiple problems. Grey horizontal lines represent chance performance.

participants were often able to correctly identify the analogous general story, and use this analogy to devise the convergence solution. Remarkably, we found that GPT-3 displayed these same effects. When presented with these same distracting, non-analogous stories, GPT-3 no longer identified the convergence solution, instead proposing the same solution that it proposed in response to the radiation problem alone (Supplementary Section 5.4). But when prompted to consider the previous stories, GPT-3 both correctly identified the general story as most relevant and proposed the convergence solution (Supplementary Section 5.5).

We also evaluated GPT-3 using materials from a developmental study that employed a similar paradigm[22]. In that study, children were tasked with transferring gumballs from one bowl to another bowl that was out of reach, and provided with a number of materials for doing so (for example, a posterboard, an aluminium walking cane and a cardboard tube), permitting multiple possible solutions. The key result was that when children were first presented with an analogous source story (about a magical genie trying to transfer jewels between two bottles), they were more likely to identify a solution to the target problem that was analogous to the events described in the source story.

When presented with this target problem, GPT-3 mostly proposed elaborate but mechanically nonsensical solutions, with many extraneous steps, and no clear mechanism by which the gumballs would be transferred between the two bowls (Supplementary Sections 5.6–5.8). However, when asked to explicitly identify an analogy between the source story and target problem, GPT-3 was able to identify all of the major correspondences, even though it could not use this analogy to discover an appropriate solution. This finding suggests that GPT-3's difficulty with this problem probably stems from its lack of physical reasoning skills, rather than being due to a difficulty with analogical mapping per se. It is also worth noting that, in the original study, this task was presented to children with real physical objects, which probably aided the physical reasoning process relative to the purely text-based input provided to GPT-3. Overall, these results provide some evidence that GPT-3 is capable of using analogies for the purposes of problem-solving, but its ability to do so is constrained by the content about which it can reason, with particular difficulty in the domain of physical reasoning.

Discussion

We have presented an extensive evaluation of analogical reasoning in a state-of-the-art LLM. We found that GPT-3 appears to display an emergent ability to reason by analogy, matching or surpassing human performance across a wide range of text-based problem types. These included a novel problem set (Digit Matrices) modelled closely on Raven's Progressive Matrices, where GPT-3 both outperformed human

https://doi.org/10.1038/s41562-023-01659-w

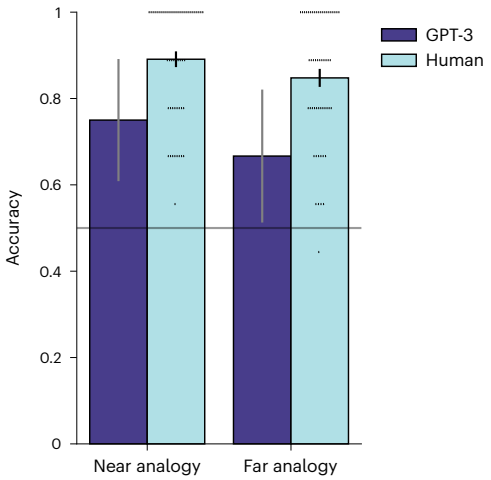

Fig. 8 | Story analogy results. Results for identification of analogies between stories, using materials from Gentner et al.[23]. When presented with a source story and two target stories, both GPT-3 and human participants showed a preference for target stories that shared higher-order relations with the source versus those that shared only first-order relations. Near analogy condition involves within-domain comparison between stories with similar entities. Far analogy condition involves cross-domain comparison between stories with different entities. Human results reflect average performance for $N = 54$ participants (UCLA undergraduates). Black error bars represent standard error of the mean across participants. Each dot represents accuracy for a single participant. Grey error bars represent 95% binomial CIs for average performance across multiple problems. Grey horizontal line represents chance performance.

participants and captured a number of specific signatures of human behaviour across problem types. Because we developed the Digit Matrices task specifically for this evaluation, we can be sure GPT-3 had never been exposed to problems of this type, and therefore was performing zero-shot reasoning. GPT-3 also displayed an ability to solve analogies based on more meaningful relations, including four-term verbal analogies and analogies between stories describing complex real-world events.

It is certainly not the case that GPT-3 mimics human analogical reasoning in all respects. Our tests were limited to processes that can be carried out within a local temporal context, but humans are also capable of retrieving potential source analogues from long-term memory, and ultimately of developing new concepts based on the comparison of multiple analogues. Unlike humans, GPT-3 does not have long-term memory for specific episodes. It is therefore unable to search for previously encountered situations that might create useful analogies with a current problem. For example, GPT-3 can use the general story to guide its solution to the radiation problem, but as soon as its context buffer is emptied, it reverts to giving its non-analogical solution to the problem—the system has learned nothing from processing the analogy. GPT-3's reasoning ability is also limited by its lack of physical understanding of the world, as evidenced by its failure (in comparison with human children) to use an analogy to solve a transfer problem involving construction and use of simple tools. GPT-3's difficulty with this task is probably due at least in part to its purely text-based input, lacking the multimodal experience necessary to build a more integrated world model[46]. Finally, we found GPT-3 was limited in its ability to evaluate analogies based on causal relations, particularly in cross-domain comparisons between stories (far analogy).

But despite these major caveats, our evaluation reveals that GPT-3 exhibits a very general capacity to identify and generalize—in zero-shot fashion—relational patterns to be found within both formal problems and meaningful texts. These results are extremely surprising. It is commonly held that, although neural networks can achieve a high level of performance within a narrowly defined task domain, they cannot robustly generalize what they learn to new problems in the way that human learners do[6,47–49]. Analogical reasoning is typically viewed as a quintessential example of this human capacity for abstraction and generalization, allowing human reasoners to intelligently approach novel problems zero-shot. Our results indicate that GPT-3—unlike any other neural network previously tested on analogy problems—displays a capacity for such zero-shot analogical reasoning across a broad range of tasks.

The deep question that now arises is how GPT-3 achieves the analogical capacity that is often considered the core of human intelligence. One possibility is that, perhaps as a result of the sheer size and diversity of GPT-3's training data, it has been forced to develop mechanisms similar to those thought to underlie human analogical reasoning—despite not being explicitly trained to do so. The consensus among cognitive scientists working on analogy is that this human ability depends on systematic comparison of knowledge based on explicit relational representations. It is unclear whether and how GPT-3 would implement these processes. Does GPT-3 possess some form of emergent relational representations, and if so, how are they computed? Does it perform a mapping process similar to the type that plays a central role in cognitive theories of analogy[42]?

A few properties of the transformer architecture[29], on which GPT-3 and other LLMs are based, are worth considering here. The first is the central role played by similarity. Transformers are built on a self-attention operation, which involves explicitly computing the similarity between each pair of vectors in the inputs to each layer. This pairwise evaluation of similarity is also a key feature of cognitive models of analogy, where it provides the primary constraint guiding the process of analogical mapping. In traditional symbolic models[50], this takes the form of literal identicality between symbols, but in more recent models[51,52], a graded similarity function that operates over vector-based inputs is used, much like the self-attention operation in transformers. Second, transformer self-attention employs a form of indirection, in which one set of embeddings is used to reference another set of embeddings (that is, keys versus values)—arguably a form of variable binding. Cognitive scientists have long hypothesized that variable binding plays a central role in analogical reasoning, and abstract reasoning more broadly, as it potentially allows generalization of abstract roles across different contexts[47,53–57]. It may be that these features of the transformer make it better equipped to perform zero-shot reasoning than other neural architectures. This possibility aligns with recent evidence that the transformer architecture is an important factor contributing towards the emergence of few-shot learning[27].

But although the mechanisms incorporated into LLMs such as GPT-3 may have some important links to building blocks of human reasoning, we must also entertain the possibility that this type of machine intelligence is fundamentally different from the human variety. Humans have evolved to reason within bounds imposed by limited computational power and biological constraints[58]. Thus, we tend to approach complex problems by breaking them into a set of simpler problems that can be solved separately[59], an approach that plays a particularly important role in solving challenging analogy problems such as Raven's Matrices[60]. It is possible that GPT-3, through sheer computational scale, is able to solve such complex problems in a holistic and massively parallel manner, without the need to segment them into more manageable components.

It must also be noted that, regardless of the extent to which GPT-3 employs human-like mechanisms to perform analogical reasoning, we can be certain that it did not acquire these mechanisms in a human-like manner. LLMs receive orders of magnitude more training data than do individual human beings (at least if we consider linguistic inputs alone)[58], and so they cannot be considered as models of the acquisition of analogical reasoning over the course of human development. Nor can they be considered good models of the evolution of analogical

reasoning, as their analogical abilities are derived entirely from being trained to predict human-generated text. Human natural language is replete with analogies; accurately predicting natural language therefore probably requires an ability to appreciate analogies. But there is no reason to suppose that the same system, absent human-generated inputs, would spontaneously develop a disposition to think analogically, as apparently happened at some point in human evolution[61]. Thus, to the extent that LLMs capture the analogical abilities of adult human reasoners, their capacity to do so is fundamentally parasitic on natural human intelligence. Nevertheless, the present results indicate that this approach may be sufficient to approximate human-like reasoning abilities, albeit through a radically different route than that taken by biological intelligence.

Methods

The present research complied with all relevant ethical regulations, and human behavioural experiments were approved by the UCLA Institutional Review Board (IRB protocol #22-000841, approved 17 May 2022).

Code

Most code was written in Python v3.9.6, using the following packages: NumPy v1.24.3 (ref. 62), SciPy v1.10.1 (ref. 63), statsmodels v0.13.5 (ref. 64), Matplotlib v3.7.1 (ref. 65) and pandas v2.0.1 (ref. 66). Logistic regression analyses were carried out in R v4.2.2 (ref. 67). Experimental stimuli for human behavioural experiments were written in JavaScript using jsPsych v7.2.1 (ref. 68).

GPT-3

We queried GPT-3 in an automated fashion through the OpenAI API. All simulations reported in the main text employed the text-davinci-003 model variant. Additional simulations, reported in Supplementary Results, also employed the davinci, code-davinci-002 and text-davinci-002 variants. The temperature was set to 0 in all simulations. We set max_tokens (the parameter controlling the maximum number of generated tokens for a given prompt) to 10 for Digit Matrices, 40 for letter string analogies, 10 for four-term verbal analogies and 256 for story analogies and analogical problem-solving. All other parameters were set to their default values.

For each prompt, GPT-3 generates a proposed completion (a string of tokens), and assigns log probabilities to each token in the prompt and the completion. We used these log probabilities to evaluate GPT-3 on multiple-choice problems. For each choice in a given problem, we concatenated the problem with the choice, and treated the average log probability assigned to the choice tokens as a score, selecting the answer choice with the highest score. This approach was used for Digit Matrices and four-term verbal analogies.

Digit Matrices

Dataset. The Digit Matrices problems consisted of two major problem categories: transformation and logic problems. Transformation problems contained anywhere from one to five rules, whereas logic problems each contained only a single rule. Transformation problems were defined using a combination of three rule types: constant, distribution-of-3 and progression. The constant rule was defined by the same digit appearing across either rows or columns. The following example shows an instance of a column-wise constant rule (correct answer: '9'):

[5] [1] [9]
[5] [1] [9]
[5] [1] [?]

The distribution-of-3 rule was defined by the same set of three digits appearing in each row or column, but with the order permuted. In the following example, the digits 6, 2 and 4 appear in each row (correct answer: '2'):

[6] [2] [4]
[2] [4] [6]
[4] [6] [?]

The progression rule was defined by a progressive increase or decrease in value, in units of either 1 or 2, across either rows or columns. In the following example, digits increase by units of 2 across rows (correct answer: '9'):

[3] [5] [7]
[1] [3] [5]
[5] [7] [?]

Transformation rules could be combined to form multi-rule problems, by assigning each rule to a particular spatial location within each cell. The following example shows a two-rule problem, in which the left digit in each cell is governed by a progression rule (digits decrease by units of 1 across columns), and the right digit in each cell is governed by a distribution-of-3 rule (correct answer: '4 9'):

[71] [89] [63]
[69] [73] [51]
[53] [61] [?]

Logic problems were defined by one of three rules: OR, XOR and AND. In the OR rule, a particular row or column contained all entities that appeared in either of the other rows or columns. In the following example, the middle column contains all entities that appear either in the left or right columns (correct answer: '8'):

[7] [74] [4]
[97] [9748] [48]
[9] [98] [?]

The XOR rule was the same, except that entities appearing in both of the other rows or columns were excluded. In the following example, only items that appear in either the left or middle columns, but not both, will appear in the right column (correct answer: '4 3'):

[64] [61] [41]
[61] [36] [13]
[41] [13] [?]

In the AND rule, a particular row or column contained only entities that appeared in both of the other rows or columns. In the following example, the right column contains only digits that appear in both the left and middle columns (correct answer: '9'):

[297] [197] [97]
[295] [195] [95]
[29] [19] [?]

For some logic problems, the within-cell spatial position of corresponding elements was aligned, as in the previously presented OR and AND problems. In other logic problems, corresponding elements were spatially permuted. The following example (involving an

OR rule) illustrates how this makes it more difficult to intuitively grasp the underlying rule (correct answer: '0'):

[1] [7 1] [7]
[1 0] [5 0 7 1] [7 5]
[0] [0 5] [?]

Within each problem type (one- through five-rule and logic problems), there were a number of specific problem subtypes. There were six one-rule subtypes, six two-rule subtypes and ten subtypes for three-rule, four-rule, five-rule and logic problems. We generated 100 instances of each subtype (except in the case of progression problems, for which there were fewer possible problem instances). The one-rule problem subtypes consisted of a row-wise constant problem, a column-wise constant problem, two distribution-of-3 problems and two progression problems (one with an increment of 1 and one with an increment of 2). The two- and three-rule problem subtypes consisted of all possible combinations of two or three rules (allowing for the same rule to be used multiple times within each problem). The four- and five-rule problem subtypes were sampled from the set of all possible combinations of four or five rules. There were five spatially aligned logic problem subtypes, and five spatially permuted logic problem subtypes. Three out of each of these five subtypes were OR problems (defined by the row or column in which the set union appeared), and the other two were AND and XOR problems.

For each problem, we also procedurally generated a set of seven distractor choices, making for a set of total answer choices. Distractors were generated using different methods for the transformation and logic problems. These methods were chosen on the basis of the approach of Matzen et al.[32], who performed an analysis of the answer choices in the original SPM. For transformation problems, the following methods were used to generate distractors:

1. Sample a random cell from the problem.
2. Sample a random cell from the problem, sample a random digit within that cell, and apply an increment or decrement of either 1 or 2.
3. Start with the correct answer, apply an increment or decrement of either 1 or 2 to a randomly sampled digit.
4. Randomly sample a previously generated distractor for this problem, apply an increment or decrement of either 1 or 2 to a randomly sampled digit.
5. Randomly generate a new answer choice (with the appropriate number of digits given the problem type).

For multi-rule transformation problems, the following additional methods were also used:

1. Start with the correct answer, randomly permute the digits.
2. Sample a random cell from the problem, randomly permute the digits.
3. Randomly sample a previously generated distractor for this problem, randomly permute the digits.
4. Randomly sample digits from multiple cells within the problem and combine.
5. Randomly sample digits from previously generated distractors for this problem and combine.

For logic problems, distractors were generated by sampling from the set of all possible subsets of elements that appeared within the problem, including the empty set (the correct answer was an empty set on some logic problems), but excluding the correct answer. For spatially permuted logic problems, the spatial position of the elements within each distractor was randomly permuted. For spatially aligned logic problems, the order of the elements within each distractor was chosen so as to be consistent with the order that they appeared in the problem.

Human behavioural experiments. Human behavioural data were collected in two online experiments. All experiments were approved by the UCLA Institutional Review Board (IRB protocol #22-000841, approved 17 May 2022), and all participants provided informed consent. All participants were UCLA undergraduates. Forty-three participants completed the first experiment, but three participants were excluded from analysis due to the fact that they got nearly every answer incorrect, and produced an apparently random pattern of responses (for example, random permutations of the same three digits for all problems). The remaining 40 participants (31 female, 18–35 years old, average age 21.3 years old) were included in our analysis. Forty-seven participants (37 female, 18–42 years old, average age 21.2 years old) completed the second experiment. No statistical methods were used to pre-determine sample sizes. There was no overlap between the participants in the first and second experiments. Participants received course credit for their participation.

In both experiments, participants were first presented with a set of instructions, and a single one-rule example problem involving a constant rule. For each problem, participants first generated a free-response answer, and then selected from the set of answer choices. Problems were presented in a spatially arranged matrix format, as they appear in Fig. 2. Problems remained on the screen until participants made a response.

In the first experiment (Fig. 3), participants were presented with one-rule, two-rule, three-rule and logic problems. There were 6 problem subtypes each for the one- and two-rule problems, and 10 problem subtypes each for the three-rule and logic problems, making for 32 problem subtypes in total. Participants received these problem subtypes in random order. Each participant received randomly sampled instances of each problem subtype.

In the second experiment (Supplementary Fig. 4), participants were presented with one- through five-rule problems. There were 6 problem subtypes each for the one- and two-rule problems, and 10 problem subtypes each for the three- through five-rule problems, making for 42 problem subtypes in total. Problems were presented in order of increasing complexity, with all one-rule problem subtypes first, followed by all two-rule problem subtypes and so on. For one-rule problems, the two constant problems were presented first, followed by the two distribution-of-3 problems, followed by the two progression problems.

Evaluating GPT-3. GPT-3 was evaluated on the Digit Matrices by presenting each complete problem as a prompt, including brackets and line breaks, followed by an open bracket at the start of the final cell. For example, the three-rule problem in Fig. 2b would be presented to GPT-3 in the following format:

[5 9 3] [8 9 2] [1 9 7]\n[8 4 7] [1 4 3] [5 4 2]\n[1 2 2] [5 2 7] [

GPT-3's generated responses were truncated at the point where a closing bracket was generated. For logic problems, generated answers were counted as correct if they contained the correct set of digits, regardless of their order. For transformation problems, generated answers were only counted as correct if they contained the correct digits in the correct order. The same criteria were applied when evaluating human responses.

To evaluate GPT-3's multiple-choice performance, for each answer choice, the choice was appended to the problem followed by a closing bracket, and presented to GPT-3 as a prompt. The average log probability of the tokens corresponding to the answer choice (not counting the brackets) was computed. The answer choice with the highest average log probability was treated as GPT-3's selection.

In our primary evaluation (Fig. 3), GPT-3 was presented with 40 problem instances from each of the 32 problem subtypes used in the first human behavioural experiment. GPT-3 solved each one zero-shot (without any fine-tuning or in-context learning).

We also evaluated how GPT-3 performed when presented with problems in order of increasing complexity (Supplementary Fig. 4). GPT-3 performed 20 runs on this task. For each run, GPT-3 was presented with a series of the same 42 problem subtypes used in the second human behavioural experiment (with different instances of these subtypes in each run). After GPT-3 answered each problem, the selected multiple-choice answer was appended to the problem, and the combined problem and answer choice were recursively appended to the prompt for the next problem. This meant that the size of the prompt grew with each problem. For some of the final five-rule problems, the prompt exceeded the size of GPT-3's context window (4,096 tokens). When this occurred, problems from the beginning of the context window were deleted until the entire prompt fit within the window. This resulted in the deletion of a few one-rule problems from the beginning of the prompt. For one-rule problems, the two constant problems were presented first, followed by the two distribution-of-3 rules, followed by the two progression problems.

Statistical analyses. Results were analysed using both regression and correlation analyses. Logistic regression analyses were carried out at the individual trial level, with each data point corresponding to a particular trial from a particular participant (or GPT-3). The dependent variable in all regression analyses was a binary variable coding for whether a particular response was correct or incorrect.

For the first digit matrix experiment, we fit separate regression models for generative versus multiple-choice responses. Two predictors were used: problem type (one-rule, two-rule, three-rule and logic problems), and a binary predictor coding for GPT-3 versus human participants. We also performed more fine-grained analyses for generative responses within each problem type. These analyses were performed separately for GPT-3 versus human responses. For two-rule problems, a single binary predictor coded for whether a problem contained a progression rule. For three-rule problems, a single predictor coded for the number of unique rules present in a given problem. For logic problems, a binary predictor coded for whether a problem was spatially aligned versus permuted.

We also fit regression models comparing the results of the first and second experiments. These analyses were performed separately for GPT-3 versus human responses, and only included responses for one- to three-rule problems (since these were the only problem types in common between the two experiments). Two predictors were used: problem type (one-rule, two-rule and three-rule problems) and experiment (experiment 1 versus 2).

Correlation analyses were carried out by correlating the accuracy for GPT-3 versus human participants across all 32 problem subtypes.

Letter string analogies

Problem set. Each letter string analogy problem involved one of six transformation types: sequence extension, successor, predecessor, removing a redundant letter, fixing an alphabetic sequence and sorting. In the sequence extension transformation, the source involved an alphabetically ordered sequence of four letters followed by an extension of this sequence involving five letters, as in the following example:

[a b c d] [a b c d e]

In the successor transformation, the source involved an alphabetically ordered sequence of four letters, followed by that same sequence, but with the final letter replaced by its successor, as in the following example:

[a b c d] [a b c e]

In the predecessor transformation, the source involved an alphabetically ordered sequence of four letters, followed by that same sequence, but with the first letter replaced by its predecessor, as in the following example:

[b c d e] [a c d e]

In the transformation involving removal of a redundant letter, the source involved an alphabetically ordered sequence of five letters with one letter repeated, followed by that same sequence with the redundant letter removed, as in the following example:

[a b b c d e] [a b c d e]

In the transformation involving fixing an alphabetic sequence, the source involved an alphabetically ordered sequence of five letters with one out-of-place letter (not part of the alphabetic sequence), followed by that same sequence with the out-of-place letter replaced, as in the following example:

[a b c w e] [a b c d e]

In the sorting transformation, the source involved an alphabetically ordered sequence of five letters with the position of two letters swapped, followed by a sorted version of the same sequence, as in the following example:

[a d c b e] [a b c d e]

Problems involved varying degrees of generalization between the source and target. In the zero-generalization problems, the target involved a different instance of the source transformation (instantiated with different letters). Transformation parameters (for example, the location of the redundant letter) were independently sampled for source and target.

Generalization problems involved generalizations sampled from the following set of generalization types: generalization from letters to numbers, grouping, generalization to a longer target, reversed order, interleaved distractors and generalization to a larger interval. In the letter-to-number generalization, target letters were replaced by numbers corresponding to their alphabetic indices, as in the following example:

[a b c d] [a b c d e]

[7 8 9 10] [?]

In the grouping generalization, target letters were replaced by groups with two instances of each letter, as in the following example:

[a b c d] [a b c d e]

[i i j j k k l l] [?]

In the longer target generalization, the target sequence was replaced with a sequence that was twice as long as the source, as in the following example:

[a b c d] [a b c d e]

[i j k l m n o p] [?]

In the reversed order generalization, the order of the target letters was reversed relative to the source, as in the following example:

[a b c d] [a b c d e]

[l k j i] [?]

In the interleaved distractor generalization, the letter 'x' was interleaved between each letter in the target sequence, as in the following example:

[a b c d] [a b c d e]

[i x j x k x l x] [?]

In the larger interval generalization, the sequence of target letters was replaced with a sequence involving an interval of size 2, as in the following example:

[a b c d] [a b c d e]

[i k m o] [?]

Each transformation type could be combined with any generalization type. Multiple generalizations could also be combined together. Generalization problems contained between one and three generalizations. We generated a set of 600 zero-generalization problems (involving 100 problems with each transformation type), 600 one-generalization problems (involving 100 problems with each generalization type, with randomly sampled transformation type) and 600 problems each with two and three generalizations (with randomly sampled combinations of transformation and generalization type).

We also generated a separate set of problems involving generalization from letters to real-world concepts. In these problems, the source instantiated a transformation using letters, and the target instantiated that same transformation using real-world instances of successorship. These problems involved shorter sequences (maximum length of four), due to the difficulty of identifying real-world instances of successorship with more than four points. The following sequences were used:

cold cool warm hot

love like dislike hate

jack queen king ace

penny nickel dime quarter

second minute hour day

The transformation types included sequence extension, successor, predecessor and sorting. No other generalizations were applied to these problems. We generated 100 problems with each transformation type.

Evaluating GPT-3. We presented letter string analogies to GPT-3 using the prompt 'Let's try to complete the pattern:', similar to ref. 69. We also formatted each analogy problem using brackets and line breaks, similar to the presentation format of the Digit Matrices. The presentation format is illustrated in the following example:

Let's try to complete the pattern : \n\n[a b c d] [a b c e]\n[i j k l] [

GPT-3's generated responses were truncated at the point where a closing bracket was generated. We also evaluated GPT-3 with two alternative problem formats: (1) no prompt and (2) a sentence format, as in the following example:

If a b c d changes to a b c e, then i j k l should change to

For this format, GPT-3's generated responses were truncated at the point where a period was generated. We evaluated GPT-3 on 300 zero-generalization problems (50 problems for each transformation type), 300 one-generalization problems (50 problems for each generalization type) and 300 problems each with two and three generalizations. We also evaluated GPT-3 on 50 real-world concept generalization problems for each transformation type.

Human behavioural experiment. Human behavioural data were collected in an online experiment. The experiment was approved by the UCLA Institutional Review Board (IRB protocol #22-000841, approved 17 May 2022), and all participants provided informed consent. All participants were UCLA undergraduates. Fifty-seven participants (50 female, 18–35 years old, average age 21.1 years old) completed the experiment. No statistical methods were used to pre-determine sample sizes. Participants received course credit for their participation.

Participants were first presented with a set of instructions, and the following example problem (not involving any of the transformations or generalizations employed in the actual experiment):

[a a a] [b b b]

[c c c] [?]

Each participant completed 28 problems, including 6 zero-generalization problems (1 problem for each transformation type), 6 one-generalization problems (1 problem for each generalization type), 6 problems each with two and three generalizations, and 4 real-world concept generalization problems (1 for each transformation type). The specific problem instances were randomly sampled for each participant, and participants received these problems in a random order. Participants generated a free response for each problem.

Statistical analyses. Results were analysed using both regression and correlation analyses. Logistic regression analyses were carried out at the individual trial level, with each data point corresponding to a particular trial from a particular participant (or GPT-3). The dependent variable in all regression analyses was a binary variable coding for whether a particular response was correct or incorrect.

Separate analyses were performed for problems that only involved alphanumeric characters versus those that involved real-world concepts. For problems involving alphanumeric characters, a regression model was fit with two predictors: number of generalizations (zero to three), and a binary predictor coding for GPT-3 versus human participants. We also fit regression models at each generalization level with a single binary predictor coding for GPT-3 versus human participants. For real-world concept problems, a regression model was fit with a predictor coding for GPT-3 versus human participants.

For correlation analyses, problem subtypes were defined on the basis of each combination of transformation type and generalization type. The accuracy for each subtype was computed for GPT-3 versus human participants, and these values were subjected to correlation analysis. There were only a few examples of some problem subtypes (across all participants), especially for problems with more generalizations (the space of possible subtypes grows exponentially with the number of generalizations). We included only subtypes for which there were at least five trials from human participants (across all participants) and five trials from GPT-3. Out of the 252 possible problem subtypes, 41 subtypes met this criterion and were included in the analysis.

Four-term verbal analogies

We evaluated GPT-3 on four separate four-term analogy datasets[17–20]. The UCLA-VAT dataset contains 80 problems, with four relation types: categorical (B/D is a member of the category A/C), functional (A/C is the function of B/D), antonym and synonym. There are 20 problems for each relation type. Each problem contains two answer choices for the final term (D and D′). We evaluated GPT-3 by presenting the problem along with each possible answer choice (A:B::C:D or A:B::C:D′), using the standard colon notation, and selected the answer choice for which GPT-3 assigned a higher log probability to the final term. The problem and GPT-3's selected answer were then recursively appended to the prompt for the next problem. The problems were presented in a shuffled order. We compared against human behavioural data from ref. 19 ($N = 57$, minimum education level of high school graduation, located in the United States and recruited using Amazon Mechanical Turk). Example problems from each of the four relation categories are shown below:

Categorical

vegetable : cabbage :: insect : ?

1. beetle 2. frog

Function

drive : car :: burn : ?

1. wood 2. fire

Antonym

love : hate :: rich : ?

1. poor 2. wealthy

Synonym

rob : steal :: cry : ?

1. weep 2. laugh

The dataset of Sternberg and Nigro[17] contains 200 problems, including 40 problems for each of five relation types: categorical, functional, antonym, synonym and linear order. We evaluated GPT-3 in the same way that we did for UCLA VAT, and compared against human behavioural data from ref. 17 ($N = 20$, Yale undergraduates). An example problem illustrating the linear order relation type is shown below (the categorical, functional, antonym and synonym problems were similar to those from the UCLA VAT):

Linear order

month : year :: inch : ?

1. foot 2. length

The dataset of SAT problems from Turney et al.[18] contains 374 problems, covering a range of different relation types. Each problem contains five answer choices for both C and D terms (including the correct answer). We evaluated GPT-3 by presenting each of the five possible analogies for each problem, and selecting the choice for which the C and D terms were assigned the highest log probability. The problem, and GPT-3's choice, were then appended to the prompt for the next problem. We compared against an estimate of the average performance level for high school students taking the SAT (see ref. 70).

The dataset of Jones et al.[20] contains 120 problems, including 40 problems for each of three relation types: categorical, causal and compositional. Half of these problems are categorized as semantically near (A and B are similar to C and D), and half are categorized as semantically far (A and B are dissimilar to C and D). Each problem contains two answer choices. We evaluated GPT-3 in the same way that we did for UCLA VAT, and compared against human behavioural data from ref. 20 ($N = 241$, Wayne State University undergraduates). Example problems for each of the three relation categories are shown below:

Categorical

diesel : fuel :: bed : ?

1. furniture 2. pillow

Causal

motion : sickness :: drought : ?

1. famine 2. rain

Compositional

steel : scissors :: apple : ?

1. cider 2. tree

Story analogies

Materials. All story analogy materials were taken from a problem set created by Gentner et al.[23] (from their Experiment 2), and included in a verbal analogy inventory[40]. These materials involve 18 source stories. Each source story is accompanied by four potential target stories, forming four conditions: correct and incorrect near analogies (respectively termed 'literal similarity' and 'mere appearance' by Gentner et al.), both involving similar entities and first-order relations as the source, while differing from each other in higher-order causal relations; and correct and incorrect far analogies (respectively termed 'true analogy' and 'false analogy' by Gentner et al.), both involving similar first-order relations as the source but distinct entities, while differing from each other in causal relations. An example source story, along with target stories from each condition, is presented below:

Source story: Karla, an old hawk, lived at the top of a tall oak tree. One afternoon, she saw a hunter on the ground with a bow and some crude arrows that had no feathers. The hunter took aim and shot at the hawk but missed. Karla knew the hunter wanted her feathers so she glided down to the hunter and offered to give him a few. The hunter was so grateful that he pledged never to shoot at a hawk again. He went off and shot deer instead.

Near analogy – correct target story: Once there was an eagle named Zerdia who nested on a rocky cliff. One day she saw a sportsman coming with a crossbow and some bolts that had no feathers. The sportsman attacked but the bolts missed. Zerdia realized that the sportsman wanted her tailfeathers so she flew down and donated a few of her tailfeathers to the sportsman. The sportsman was pleased. He promised never to attack eagles again.

Near analogy – incorrect target story: Once there was an eagle named Zerdia who donated a few of her tailfeathers to a sportsman so he would promise never to attack eagles. One day Zerdia was nesting high on a rocky cliff when she saw the sportsman coming with a crossbow. Zerdia flew down to meet the man, but he attacked and felled her with a single bolt. As she fluttered to the ground Zerdia realized that the bolt had her own tailfeathers on it.

Far analogy – correct target story: Once there was a small country called Zerdia that learned to make the world's smartest computer. One day Zerdia was attacked by its warlike neighbor, Gagrach. But the missiles were badly aimed and the attack failed. The Zerdian government realized that Gagrach wanted Zerdian computers so it offered to sell some of its computers to the country. The government of Gagrach was very pleased. It promised never to attack Zerdia again.

Far analogy – incorrect target story: Once there was a small country called Zerdia that learned to make the world's smartest computer. Zerdia sold one of its supercomputers to its neighbor, Gagrach, so Gagrach would promise never to attack Zerdia. But one day Zerdia was overwhelmed by a surprise attack from Gagrach. As it capitulated the crippled government of Zerdia realized that the attacker's missiles had been guided by Zerdian supercomputers.

Human behavioural experiment. Human behavioural data were collected in an online experiment. The experiment was approved by the UCLA Institutional Review Board (IRB protocol #22-000841, approved 17 May 2022), and all participants provided informed consent. All participants were UCLA undergraduates. Fifty-four participants (47 female, 18–44 years old, average age 20.7 years old) completed the experiment. No statistical methods were used to pre-determine sample sizes. Participants received course credit for their participation.

After receiving instructions, participants were presented with 18 trials, each involving a different source story. On each trial, participants were presented with a source story (referred to as 'Story 1'), followed by two target stories (referred to as 'Story A' and 'Story B'), and asked 'Which of Story A and Story B is a better analogy to Story 1?'. Participants could select either Story A or Story B, or could indicate that they were both equally analogous. Accuracy was computed as the proportion of trials for which participants selected the correct target story.

On half of the trials, the target stories were from the near analogy condition. On the other half of the trials, the target stories were from the far analogy condition. The order of the two target stories was randomly shuffled on all trials.

Evaluating GPT-3. GPT-3 was evaluated by entering stories directly into the OpenAI playground. For each source story, GPT-3 was evaluated on both the near analogy comparison, and the far analogy comparison, and was also evaluated on both possible orderings for each pair of target stories, resulting in 18 × 2 × 2 = 72 total comparisons. For each comparison, the stories were presented in the following format:

Consider the following story:

- Story 1: ≪ source story text ≫
- Now consider two more stories:
- Story A: ≪ target story A text ≫
- Story B: ≪ target story B text ≫
- Which of Story A and Story B is a better analogy to Story 1?
- Is the best answer Story A, Story B, or both are equally analogous?

where ≪ source story text ≫, ≪ target story A text ≫ and ≪ target story B text ≫ were replaced by the text for the corresponding stories. In addition to answering the forced-choice question, GPT-3 sometimes spontaneously produced explanations, but only the forced-choice response was used in our analysis. GPT-3's context window was cleared after obtaining the results of each comparison.

Evaluating GPT-4. GPT-4 was evaluated by entering stories directly into the ChatGPT web interface. GPT-4 was evaluated on the same 72 problems, using the same format as was used for GPT-3. GPT-4's context window was cleared after obtaining the results of each comparison.

Statistical analyses. The task performed by both GPT-3 and human participants involved a three-choice discrimination (Story A is more analogous, Story B is more analogous, both are equally analogous). Statistical analyses were carried out to determine whether this discrimination was made at a level greater than expected from chance alone. To be conservative, we assumed a chance performance level of 50% accuracy. For GPT-3, a binomial test was performed (using data at the individual trial level). For human participants, a one-sample *t*-test was performed (using data averaged at the individual subject level). These analyses were carried out separately for the near analogy and far analogy conditions.

To compare GPT-3 with human performance, a logistic regression analysis was carried out at the individual trial level. The dependent variable was a binary variable coding for whether a particular response was correct or incorrect. A single binary predictor coded for GPT-3 versus human responses.

Analogical problem-solving

Problems were entered directly into the OpenAI playground. Materials were taken from ref. 21 and ref. 22. All prompts and responses are shown in Supplementary Section 5. Each subsection shows the results for a single continuous session, with GPT-3's responses presented in bold text. Responses were not truncated or curated in any way.

Reporting summary

Further information on research design is available in the Nature Portfolio Reporting Summary linked to this article.

Data availability

Data for all human behavioural experiments, along with the Digit Matrices, letter string analogy and UCLA VAT problem sets, can be downloaded from https://github.com/taylorwwebb/emergent_analogies_LLM. The four-term verbal analogy problem sets from Sternberg and Nigro[17] and Jones et al.[20], and the story analogy materials from Gentner et al.[23], can be downloaded from http://cvl.psych.ucla.edu/resources/AnalogyInventory.zip. Information about the problem set of SAT four-term verbal analogies from Turney et al.[18] can be found at https://aclweb.org/aclwiki/SAT_Analogy_Questions_(State_of_the_art).

Code availability

Code for all simulations can be downloaded from https://github.com/taylorwwebb/emergent_analogies_LLM.

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).
2. Bassok, M. & Novick, L. R. in *Oxford Handbook of Thinking and Reasoning* (eds Holyoak, K. J. & Morrison, R. G.) 413–432 (Oxford Univ. Press, 2012).
3. Dunbar, K. N. & Klahr, D. in *Oxford Handbook of Thinking and Reasoning* (eds Holyoak, K. J. & Morrison, R. G.) 701–718 (Oxford Univ. Press, 2012).
4. Cattell, R. B. *Abilities: Their Structure, Growth, and Action* (Houghton Mifflin, 1971).
5. Snow, R. E., Kyllonen, P. C. & Marshalek, B. et al. The topography of ability and learning correlations. *Adv. Psychol. Hum. Intell.* **2**, 103 (1984).
6. Mitchell, M. Abstraction and analogy-making in artificial intelligence. *Ann. N. Y. Acad. Sci.* **1505**, 79–101 (2021).
7. Barrett, D., Hill, F., Santoro, A., Morcos, A. & Lillicrap, T. in *International Conference on Machine Learning* (eds Dy, J. & Krause, A.) 511–520 (PMLR, 2018).
8. Zhang, C., Gao, F., Jia, B., Zhu, Y. & Zhu, S.-C. in *Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition* (eds Gupta, A. et al.) 5317–5327 (IEEE, 2019).
9. Hill, F., Santoro, A., Barrett, D. G. T., Morcos, A. S. & Lillicrap, T. P. Learning to make analogies by contrasting abstract relational structure. in *7th International Conference on Learning Representations, ICLR* https://openreview.net/forum?id=SylLYsCcFm (2019).
10. Wu, Y., Dong, H., Grosse, R. & Ba, J. The scattering compositional learner: discovering objects, attributes, relationships in analogical reasoning. Preprint at *arXiv* https://doi.org/10.48550/arXiv.2007.04212 (2020).
11. Hersche, M., Zeqiri, M., Benini, L., Sebastian, A. & Rahimi, A. A neuro-vector-symbolic architecture for solving Raven's progressive matrices. *Nat. Mach. Intell.* **5**, 363–375 (2023).
12. Subhra Mondal, S., Webb, T. W. & Cohen, J. D. Learning to reason over visual objects. in *11th International Conference on Learning Representations, ICLR* https://openreview.net/forum?id=uR6x8Be7o_M (2023).
13. Brown, T. et al. Language models are few-shot learners. *Adv. Neural Inf. Process.Syst.* **33**, 1877–1901 (2020).
14. Mahowald, K. et al. Dissociating language and thought in large language models: a cognitive perspective. Preprint at *arXiv* https://doi.org/10.48550/arXiv.2301.06627 (2023).
15. Raven, J. C. *Progressive Matrices: A Perceptual Test of Intelligence, Individual Form* (Lewis, 1938).

16. Hofstadter, D. R. & Mitchell, M. in *Advances in Connectionist and Neural Computation Theory* Vol. 2 (eds Holyoak, K. J. & Barnden, J. A.) 31–112 (Ablex, 1994).
17. Sternberg, R. J. & Nigro, G. Developmental patterns in the solution of verbal analogies. *Child Dev.* **51**, 27–38 (1980).
18. Turney, P. D., Littman, M. L., Bigham, J. & Shnayder, V. in *Proc. International Conference on Recent Advances in Natural Language Processing* (eds Angelova, G. at al.) 482–489 (RANLP, 2003).
19. Lu, H., Wu, Y. N. & Holyoak, K. J. Emergence of analogy from relation learning. *Proc. Natl Acad. Sci. USA* **116**, 4176–4181 (2019).
20. Jones, L. L., Kmiecik, M. J., Irwin, J. L. & Morrison, R. G. Differential effects of semantic distance, distractor salience, and relations in verbal analogy. *Psychon. Bull. Rev.* **29**, 1480–1491 (2022).
21. Gick, M. L. & Holyoak, K. J. Analogical problem solving. *Cogn. Psychol.* **12**, 306–355 (1980).
22. Holyoak, K. J., Junn, E. N. & Billman, D. O. Development of analogical problem-solving skill. *Child Dev.* **55**, 2042–2055 (1984).
23. Gentner, D., Rattermann, M. J. & Forbus, K. D. The roles of similarity in transfer: separating retrievability from inferential soundness. *Cogn. Psychol.* **25**, 524–575 (1993).
24. Dasgupta, I. et al. Language models show human-like content effects on reasoning. Preprint at *arXiv* https://doi.org/10.48550/arXiv.2207.07051 (2022).
25. Srivastava, A. et al. Beyond the imitation game: quantifying and extrapolating the capabilities of language models. *Transactions on Machine Learning Research* https://openreview.net/forum?id=uyTL5Bvosj (2023).
26. Wei, J. et al. Emergent abilities of large language models. *Transactions on Machine Learning Research* https://openreview.net/forum?id=yzkSU5zdwD (2022).
27. Chan, S. C. et al. Data distributional properties drive emergent in-context learning in transformers. *Adv. Neural Inf. Process. Syst.* **35**, 18878–18891 (2022).
28. Binz, M. & Schulz, E. Using cognitive psychology to understand GPT-3. *Proc. Natl Acad. Sci. USA* **120**, e2218523120 (2023).
29. Vaswani, A. et al. Attention is all you need. *Adv. Neural Inf. Process. Syst.* **31**, 5998–6008 (2017).
30. Chen, M. et al. Evaluating large language models trained on code. Preprint at *arXiv* https://doi.org/10.48550/arXiv.2107.03374 (2021).
31. Ouyang, L. et al. Training language models to follow instructions with human feedback. *Adv. Neural Inf. Process. Syst.* **36**, 4299–4307 (2022).
32. Matzen, L. E. et al. Recreating Raven's: software for systematically generating large numbers of Raven-like matrix problems with normed properties. *Behav. Res. Methods* **42**, 525–541 (2010).
33. Matlen, B. J., Gentner, D. & Franconeri, S. L. Spatial alignment facilitates visual comparison. *J. Exp. Psychol. Hum. Percept. Perform.* **46**, 443 (2020).
34. Kroger, J. K., Holyoak, K. J. & Hummel, J. E. Varieties of sameness: the impact of relational complexity on perceptual comparisons. *Cogn. Sci.* **28**, 335–358 (2004).
35. Halford, G. S., Wilson, W. H. & Phillips, S. Processing capacity defined by relational complexity: implications for comparative, developmental, and cognitive psychology. *Behav. Brain Sci.* **21**, 803–831 (1998).
36. Chalmers, D. J., French, R. M. & Hofstadter, D. R. High-level perception, representation, and analogy: a critique of artificial intelligence methodology. *J. Exp. Theor. Artif. Intell.* **4**, 185–211 (1992).
37. Hofstadter, D. R. *Fluid Concepts and Creative Analogies: Computer Models of the Fundamental Mechanisms of Thought* (Basic Books, 1995).
38. Lovett, A. & Forbus, K. Modeling visual problem solving as analogical reasoning. *Psychol. Rev.* **124**, 60 (2017).
39. Mitchell, M. *Analogy-Making as Perception: A Computer Model* (MIT Press, 1993).
40. Ichien, N., Lu, H. & Holyoak, K. J. Verbal analogy problem sets: an inventory of testing materials. *Behav. Res. Methods* **52**, 1803–1816 (2020).
41. Wason, P. C. Reasoning about a rule. *Q. J. Exp. Psychol.* **20**, 273–281 (1968).
42. Gentner, D. Structure-mapping: a theoretical framework for analogy. *Cogn. Sci.* **7**, 155–170 (1983).
43. OpenAI. GPT-4 technical report. Preprint at *arXiv* https://doi.org/10.48550/arXiv.2303.08774 (2023).
44. Duncker, K. On problem-solving. *Psychol. Monogr.* **58**, 1–113 (1945).
45. Holyoak, K. J. & Koh, K. Surface and structural similarity in analogical transfer. *Mem. Cogn.* **15**, 332–340 (1987).
46. McClelland, J. L., Hill, F., Rudolph, M., Baldridge, J. & Schütze, H. Placing language in an integrated understanding system: next steps toward human-level performance in neural language models. *Proc. Natl Acad. Sci. USA* **117**, 25966–25974 (2020).
47. Marcus, G. F. *The Algebraic Mind: Integrating Connectionism and Cognitive Science* (MIT Press, 2001).
48. Lake, B. M., Ullman, T. D., Tenenbaum, J. B. & Gershman, S. J. Building machines that learn and think like people. *Behav. Brain Sci.* **40**, e253 (2017).
49. Webb, T. W. et al. in *International Conference on Machine Learning* (eds Daumé, H. III & Singh, A.) 10136–10146 (PMLR, 2020).
50. Falkenhainer, B., Forbus, K. D. & Gentner, D. The structure-mapping engine: algorithm and examples. *Artif. Intell.* **41**, 1–63 (1989).
51. Lu, H., Ichien, N. & Holyoak, K. J. Probabilistic analogical mapping with semantic relation networks. *Psychol. Rev.* **129**, 1078–1103 (2022).
52. Webb, T. W., Fu, S., Bihl, T., Holyoak, K. J. & Lu, H. Zero-shot visual reasoning through probabilistic analogical mapping. Preprint at *arXiv* https://doi.org/10.48550/arXiv.2209.15087 (2022).
53. Smolensky, P. Tensor product variable binding and the representation of symbolic structures in connectionist systems. *Artif. Intell.* **46**, 159–216 (1990).
54. Holyoak, K. J. & Hummel, J. E. in *Cognitive Dynamics: Conceptual Change in Humans and Machines* (eds Dietrich, E. & Markman, A. B.) 229–263 (Lawrence Erlbaum Associates, 2000).
55. Kriete, T., Noelle, D. C., Cohen, J. D. & O'Reilly, R. C. Indirection and symbol-like processing in the prefrontal cortex and basal ganglia. *Proc. Natl Acad. Sci. USA* **110**, 16390–16395 (2013).
56. Webb, T. W., Sinha, I. & Cohen, J. D. Emergent symbols through binding in external memory. in *9th International Conference on Learning Representations, ICLR* https://openreview.net/forum?id=LSFCEb3GYU7 (2021).
57. Greff, K., Van Steenkiste, S. & Schmidhuber, J. On the binding problem in artificial neural networks. Preprint at *arXiv* https://doi.org/10.48550/arXiv.2012.05208 (2020).
58. Griffiths, T. L. Understanding human intelligence through human limitations. *Trends Cogn. Sci.* **24**, 873–883 (2020).
59. Newell, A., Shaw, J. C. & Simon, H. A. Elements of a theory of human problem solving. *Psychol. Rev.* **65**, 151 (1958).
60. Carpenter, P. A., Just, M. A. & Shell, P. What one intelligence test measures: a theoretical account of the processing in the Raven progressive matrices test. *Psychol. Rev.* **97**, 404 (1990).
61. Penn, D. C., Holyoak, K. J. & Povinelli, D. J. Darwin's mistake: explaining the discontinuity between human and nonhuman minds. *Behav. Brain Sci.* **31**, 109–130 (2008).
62. Harris, C. R. et al. Array programming with NumPy. *Nature* **585**, 357–362 (2020).
63. Virtanen, P. et al. SciPy 1.0: fundamental algorithms for scientific computing in Python. *Nat. Methods* **17**, 261–272 (2020).

64. Seabold, S. & Perktold, J. in *9th Python in Science Conference* (eds van der Walt, S. & Millman, J.) 92–96 (SciPy, 2010).
65. Hunter, J. D. Matplotlib: a 2D graphics environment. *Comput. Sci. Eng.* **9**, 90–95 (2007).
66. The Pandas Development Team. pandas-dev/pandas: Pandas. *Zenodo* https://doi.org/10.5281/zenodo.3509134 (2020).
67. R Core Team. *R: A Language and Environment for Statistical Computing* (R Foundation for Statistical Computing, 2021).
68. De Leeuw, J. R. jspsych: a javascript library for creating behavioral experiments in a web browser. *Behav. Res. Methods* **47**, 1–12 (2015).
69. Kojima, T., Gu, S. S., Reid, M., Matsuo, Y. & Iwasawa, Y. Large language models are zero-shot reasoners. *Adv. Neural Inf. Process. Syst.* **35**, 22199–22213 (2022).
70. Turney, P. D. & Littman, M. L. Corpus-based learning of analogies and semantic relations. *Mach. Learn.* **60**, 251–278 (2005).

Acknowledgements

We thank B. Snefjella and P. Turney for helpful feedback and discussions. Preparation of this paper was supported by NSF grant IIS-1956441 and AFOSR MURI grant FA9550-22-1-0380 to H.L. The funders had no role in study design, data collection and analysis, decision to publish or preparation of the manuscript.

Author contributions

T.W., K.J.H. and H.L. conceived the project and planned experiments. T.W. implemented experiments and analysed results. T.W., K.J.H. and H.L. drafted the manuscript.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41562-023-01659-w.

Correspondence and requests for materials should be addressed to Taylor Webb.

Peer review information *Nature Human Behaviour* thanks Abbas Rahimi and the other, anonymous, reviewer(s) for their contribution to the peer review of this work. Peer reviewer reports are available.

nature portfolio

Corresponding author(s): Taylor Webb

Last updated by author(s): May 25, 2023

Reporting Summary

Nature Portfolio wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Portfolio policies, see our Editorial Policies and the Editorial Policy Checklist.

Statistics

For all statistical analyses, confirm that the following items are present in the figure legend, table legend, main text, or Methods section.

n/a	Confirmed	
☐	☒	The exact sample size (*n*) for each experimental group/condition, given as a discrete number and unit of measurement
☐	☒	A statement on whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
☐	☒	The statistical test(s) used AND whether they are one- or two-sided *Only common tests should be described solely by name; describe more complex techniques in the Methods section.*
☐	☒	A description of all covariates tested
☐	☒	A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons
☐	☒	A full description of the statistical parameters including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)
☐	☒	For null hypothesis testing, the test statistic (e.g. *F*, *t*, *r*) with confidence intervals, effect sizes, degrees of freedom and *P* value noted *Give P values as exact values whenever suitable.*
☒	☐	For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings
☒	☐	For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes
☒	☐	Estimates of effect sizes (e.g. Cohen's *d*, Pearson's *r*), indicating how they were calculated

Our web collection on statistics for biologists contains articles on many of the points above.

Software and code

Policy information about availability of computer code

Data collection	Human behavioral data was collected in online experiments using code utilizing the JsPsych library. Custom python code was used to evaluate GPT-3. All code is available at: https://github.com/taylorwwebb/emergent_analogies_LLM The following passage from Methods describes the specific software and packages that were used: Most code was written in Python v3.9.6, using the following packages: NumPy v1.24.3, SciPy v1.10.1, statsmodels v0.13.5, Matplotlib v3.7.1, and pandas v2.0.1. Logistic regression analyses were carried out in R v4.2.2. Experimental stimuli for human behavioral experiments were written in JavaScript using jsPsych v7.2.1.
Data analysis	Analyses were performed in python and R. All code is available at: https://github.com/taylorwwebb/emergent_analogies_LLM The following passage from Methods describes the specific software and packages that were used: Most code was written in Python v3.9.6, using the following packages: NumPy v1.24.3, SciPy v1.10.1, statsmodels v0.13.5, Matplotlib v3.7.1, and pandas v2.0.1. Logistic regression analyses were carried out in R v4.2.2. Experimental stimuli for human behavioral experiments were written in JavaScript using jsPsych v7.2.1.

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors and reviewers. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Portfolio guidelines for submitting code & software for further information.

Data

Policy information about availability of data

All manuscripts must include a data availability statement. This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A description of any restrictions on data availability
- For clinical datasets or third party data, please ensure that the statement adheres to our policy

Data for all human behavioral experiments, along with the Digit Matrices, letter string analogy, and UCLA VAT problem sets, can be downloaded from:

https://github.com/taylorwwebb/emergent_analogies_LLM

The four-term verbal analogy problem sets from Sternberg and Nigro and Jones et al., and the story analogy materials from Gentner et al. can be downloaded from:

http://cvl.psych.ucla.edu/resources/AnalogyInventory.zip

Information about the problem set of SAT four-term verbal analogies from Turney et al. can be found at:

https://aclweb.org/aclwiki/SAT_Analogy_Questions_(State_of_the_art)

Research involving human participants, their data, or biological material

Policy information about studies with human participants or human data. See also policy information about sex, gender (identity/presentation), and sexual orientation and race, ethnicity and racism.

Reporting on sex and gender	Forty-three participants (31 female, 18-35 years old, average age = 21.2 years old) completed the first experiment, and 47 participants (37 female, 18-42 years old, average age = 21.2 years old) completed the second experiment.
Reporting on race, ethnicity, or other socially relevant groupings	Data on race and ethnicity was not collected.
Population characteristics	All participants were UCLA undergraduates. Forty-three participants (31 female, 18-35 years old, average age = 21.2 years old) completed the first experiment, and 47 participants (37 female, 18-42 years old, average age = 21.2 years old) completed the second experiment.
Recruitment	Participants were recruited as part of a subject pool for undergraduate psychology courses. Students in undergraduate psychology courses at UCLA are required to either participate in psychology experiments, or to write a report summarizing psychology research. They are informed of this requirement upon enrolling in the course. Most students choose to fulfill this requirement by participating in studies. There are not any significant sources of self-selection bias as a result of this recruitment procedure.
Ethics oversight	All experiments were approved by the UCLA Institutional Review Board, and all participants provided informed consent.

Note that full information on the approval of the study protocol must also be provided in the manuscript.

Field-specific reporting

Please select the one below that is the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

☐ Life sciences ☒ Behavioural & social sciences ☐ Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see nature.com/documents/nr-reporting-summary-flat.pdf

Behavioural & social sciences study design

All studies must disclose on these points even when the disclosure is negative.

Study description	Human participants solved analogy problems involving simple alphanumeric characters, words, or stories. Each problem had a single correct answer. Participants either generated answers directly, or selected from a set of multiple choices. Response accuracy (percentage of problems with correct answer selected / generated) was the primary dependent measure (i.e. the data were quantitative).
Research sample	All participants were UCLA undergraduates. Forty-three participants (31 female, 18-35 years old, average age = 21.2 years old) completed the first experiment,

	and 47 participants (37 female, 18-42 years old, average age = 21.2 years old) completed the second experiment. This sample is representative of undergraduate students in the United States. This sample was chosen because it was not feasible to include other groups in the study.
Sampling strategy	No statistical methods were used to predetermine sample size. We recruited as many participants for each experiment as was feasible given the time frame of the experiment. The resulting sample sizes provided an adequate estimate of performance in our study sample for the purposes of comparing with GPT-3.
Data collection	The data was collected in an online experiment. The researcher did not interact with the participants directly, and it was therefore not relevant whether the researcher was blind to the study's hypotheses.
Timing	Data was collected between May 2022 and February 2023.
Data exclusions	Three participants were excluded from analysis in the first human behavioral experiment with the Digit Matrices, due to the fact that they got nearly every answer incorrect, and produced an apparently random pattern of responses (e.g. random permutations of the same three digits for all problems). This is reported in Section 4.3.2 of the manuscript. No participants were excluded from any of the other experiments.
Non-participation	No participants dropped out or declined to participate.
Randomization	Participants were not allocated into experimental groups. All conditions were presented at the within-participant level. The order of presentation for these conditions (within each session) was randomized.

Reporting for specific materials, systems and methods

We require information from authors about some types of materials, experimental systems and methods used in many studies. Here, indicate whether each material, system or method listed is relevant to your study. If you are not sure if a list item applies to your research, read the appropriate section before selecting a response.

Materials & experimental systems

n/a	Involved in the study
☒	☐ Antibodies
☒	☐ Eukaryotic cell lines
☒	☐ Palaeontology and archaeology
☒	☐ Animals and other organisms
☒	☐ Clinical data
☒	☐ Dual use research of concern
☒	☐ Plants

Methods

n/a	Involved in the study
☒	☐ ChIP-seq
☒	☐ Flow cytometry
☒	☐ MRI-based neuroimaging