PHILOSOPHICAL TRANSACTIONS A

royalsocietypublishing.org/rsta

Opinion piece

Cite this article: Ku A *et al*. 2026 Levels of analysis for large language models. *Phil. Trans. R. Soc. A* **384**: 20250012.
https://doi.org/10.1098/rsta.2025.0012

Received: 15 March 2025
Accepted: 27 July 2025

One contribution of 18 to a theme issue 'World models in natural and artificial intelligence'.

Subject Areas:
artificial intelligence

Keywords:
cognitive science, levels of analysis, machine behaviour

Author for correspondence:
Thomas Griffiths
e-mail: tomg@princeton.edu

Levels of analysis for large language models

Alexander Ku[1], Declan Campbell[1], Xuechunzi Bai[2], Jiayi Geng[1], Ryan Liu[1], Raja Marjieh[1], R. Thomas McCoy[3], Andrew Nam[1], Ilia Sucholutsky[4], Liyi Zhang[1], Jian-Qiao Zhu[1] and Thomas Griffiths[1]

[1]Princeton University, Princeton, NJ, USA
[2]The University of Chicago, Chicago, IL, USA
[3]Yale University, New Haven, CT, USA
[4]New York University, New York, NY, USA

TG, 0000-0002-5138-7255

Modern artificial intelligence systems, such as large language models, are increasingly powerful but also increasingly hard to understand. Recognizing this problem as analogous to the historical difficulties in understanding the human mind, we argue that methods developed in cognitive science can be useful for understanding large language models. We propose a framework for applying these methods based on the levels of analysis that David Marr proposed for studying information processing systems. By revisiting established cognitive science techniques relevant to each level and illustrating their potential to yield insights into the behaviour and internal organization of large language models, we aim to provide a toolkit for making sense of these new kinds of minds.

This article is part of the theme issue 'World models in natural and artificial intelligence'.

1. Introduction

The last decade has seen a series of breakthroughs in artificial intelligence (AI) research, culminating in the creation of the large language models that underlie chat-based agents such as ChatGPT, Claude, Gemini and LLaMa [1–3

THE ROYAL SOCIETY PUBLISHING

]. These breakthroughs have been driven by a specific strategy: starting with generic artificial neural network architectures and increasing their size and training data. Artificial neural networks are notoriously difficult to interpret, finding solutions to problems that are expressed in the form of billions of continuous weighted connections between units. As a consequence, computer scientists now face an unfamiliar problem: they have created systems that they do not understand. To make things even worse, since the training data and weights of many of the leading systems are not available outside the companies that created them, in many cases the only insights we can obtain about the nature of these systems are those that can be gleaned by studying their behaviour.

Even though this problem is unfamiliar to computer scientists, it is very familiar to another group of researchers: cognitive scientists. Cognitive science is the interdisciplinary science of the mind, and for the 70 or so years since its inception [4] has been limited by the fact that it had relatively few kinds of mind to study. To cognitive scientists, the advent of intelligent machines offers exciting new opportunities to apply methods that have been refined through trying to understand how human minds work [5,6]. Those methods encompass both techniques that are based on human behaviour and insights that come from related fields such as neuroscience that explore the underlying mechanisms.

One powerful conceptual framework used in cognitive science is the idea that information processing systems can be understood at different levels of analysis. The computational neuroscientist David Marr proposed three such levels [7]: the computational level, which focuses on the abstract computational problem a system solves and its ideal solution; the algorithmic level, which focuses on the representations and algorithms that approximate that solution; and the implementation level, which focuses on how those representations and algorithms are realized in a physical system. The same three levels can be used for analysing large language models, focusing on how such systems are shaped by their function, the solutions that they seem to find, and the realization of those solutions in weights and units within the underlying artificial neural network (see table 1).

Drawing these parallels between the levels at which we analyse minds and machines provides a guide for where we might expect to look for relevant tools for understanding the behaviour of AI systems: analyses at the computational level can make use of computational modelling techniques, analyses at the algorithmic level can draw on methods from cognitive psychology, and analyses at the implementation level can take inspiration by neuroscience. In this paper, we highlight some of the tools from cognitive science that we believe are particularly useful for understanding large language models (and related approaches such as vision-language models) at each of these levels—techniques such as rational analysis, the axiomatic approach and multidimensional scaling (MDS). We illustrate the potential of this approach by using examples from our own work and by drawing analogies between nascent methods in computer science and tools developed in other fields.

2. Computational level

The computational level focuses on the abstract problem that a system needs to solve. By recognizing the pressures that this problem exerts upon the system, we can make predictions about the properties that the system is likely to have [7–10]. This perspective is perhaps most familiar from evolutionary biology, in which organisms are understood through the lens of the evolutionary pressures that have shaped them. For instance, our understanding of bird flight is informed by the aerodynamic principles that bird flight must obey. Similarly, we can gain insight into intelligent systems by considering how they have been shaped by the functions that they must perform.

This emphasis on function makes the computational level well-suited for analysing AI systems. Although many aspects of modern AI systems are difficult to interpret (including their behaviour and the mechanisms that they use to achieve that behaviour), one aspect that we understand well

Table 1. Understanding natural and artificial minds across Marr's levels of analysis.

level of analysis	inquiry	cognitive science	artificial intelligence
computational	what problem is being solved and what is the ideal solution?	understanding behaviour in terms of optimal solutions to environmental pressures (e.g. Bayesian models of cognition)	understanding behaviour in terms of the training objective (e.g. next-token prediction) and deviations from optimal benchmarks (e.g. violations of probability axioms)
algorithmic	how is the solution approximated and what representations and processes are involved?	inferring mental representations and processes through behavioural experiments (e.g. reaction times, error patterns, similarity judgements)	inferring representations and processes via behavioural analysis (e.g. analysing systematic errors, soliciting similarity judgements, uncovering hidden associations)
implementation	how are the representations and processes physically realized?	studying neural circuits and population activity using techniques like optogenetics, single-unit recording, fMRI and MVPA	studying artificial neurons, circuits (e.g. induction heads) and population activity using techniques like activation patching, sparse autoencoders and representational geometry analysis

fMRI, functional magnetic resonance imaging; MVPA, multi-voxel pattern analysis (or multivariate pattern analysis).

is the function that the system is optimized to perform. Specifically, this function is explicitly defined by humans in the form of the AI system's training objective. Therefore, computational-level analysis enables us to start from something we understand well—the training objective—and use it to reason about the less-well-understood territory of the behaviour of the resulting systems.

(a) Embers of autoregression

McCoy *et al*. [11] used an approach based on computational-level analysis to try to understand the behaviour of large language models (LLMs). For these systems, the primary training objective is next-token prediction, also known as autoregression: predicting the next token (word or part of a word) in a piece of text given the preceding tokens. Analysis of this task leads to the prediction that LLMs will perform better when they need to produce a high-probability piece of text than when they need to produce a low-probability piece of text, even in deterministic settings where probability should not matter. Intuitively, this prediction follows from the way in which next-token prediction fundamentally depends on the probabilities of token sequences; this intuition is derived more formally in [11] via a Bayesian analysis of autoregression.

The prediction that LLM behaviour will be sensitive to probability is borne out in experiments testing a range of LLMs across a range of tasks (see figure 1). For instance, when GPT-4 is asked to count how many letters there are in a list, it performs much better when the answer is a frequently used number than a more rarely used number; e.g. when the answer is 30, its accuracy is 97%, but when the answer is 29, its accuracy is only 17%, presumably because the number 30 is

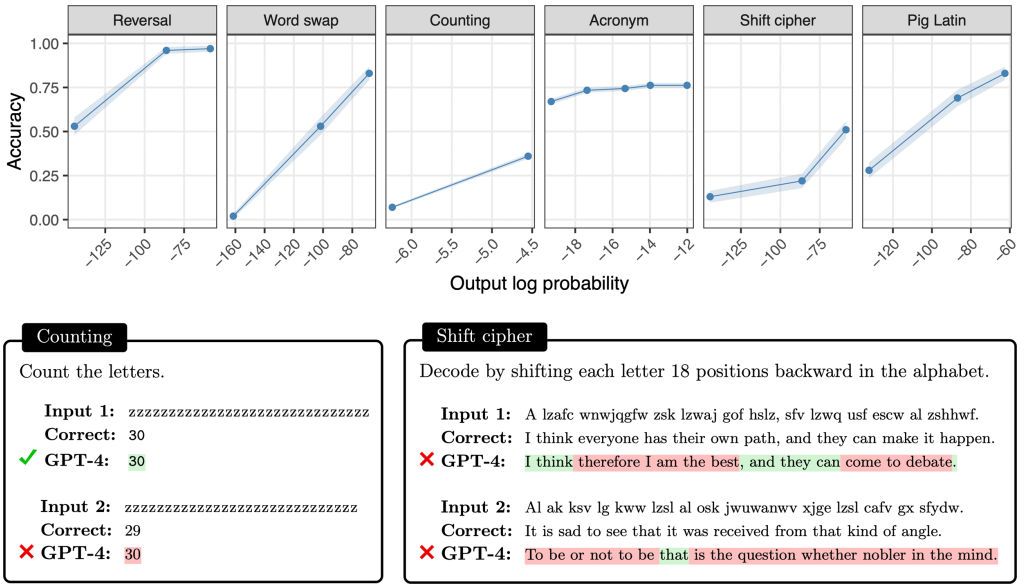

Figure 1. Large language models perform better when they need to produce a high-probability piece of text than when they need to produce a low-probability piece of text, even in deterministic settings where probability should not matter.

used more often in natural text than the number 29. Similarly, when asked to decode a message written in a simple cipher, GPT-4's accuracy was 51% when the answer was a high-probability sentence but only 13% when the answer was a low-probability word sequence. Thus, even though LLMs can be applied to many different tasks—a capability that has been viewed as evidence that LLMs show 'sparks of artificial general intelligence' [12]—they also continue to display 'embers of autoregression'—behavioural trends that reflect the nature of the specific objective they were optimized for.

(b) Bayesian optimality as a benchmark

While the 'embers of autoregression' reflect the fact that large language models are pretrained on text prediction, the use of this objective may also mean that we can understand their behaviour in terms of Bayesian inference more generally. Accurately predicting the next token in text requires making implicit inferences not just about what sequences of words are related to one another, but about events in the world described by those words. Bayesian models of cognition provide optimal solutions to such problems, and have become instrumental in explaining human performance across diverse domains, including perception [13], language processing [14,15], categorization [16] and intuitive physics [17,18]. The emergence of a broader capability to draw inductive inferences about the world suggests that we can likewise compare LLMs with these optimal solutions to gain insight into their behaviour, much as cognitive scientists use Bayesian models to understand human cognition.

The connection between Bayesian optimality and LLMs can be made more explicit by considering the problem of next-token prediction that is typically used in training these models. Predicting the next token can be done by extracting the predictive sufficient statistics from previous tokens [19]. For some datasets, the Bayesian posterior distribution over particular parameters or hypotheses about the generating process can serve as a predictive sufficient statistic [20]. This perspective can be used to understand the representations that LLMs form and how they should relate to ideal Bayesian solutions [20,21]. Several other recent papers have also identified interesting connections between LLMs and Bayesian inference [22–24].

These connections suggest that we might be able to create simple Bayesian models of the inferences drawn by LLMs. Such Bayesian models can be used to explore the implicit prior distributions adopted by LLMs and to compare the resulting distributions with those inferred from human behaviour. For example, Griffiths *et al.* [25] used a simple Bayesian model of predicting the future [26] to recover implicit prior distributions about the extent or duration of phenomena from GPT-4. Zhu & Griffiths [27] built on this work, using an iterated learning procedure originally developed for sampling from human priors to explore the priors that LLMs have for causal relationships and probability distributions, as well as their implicit assumptions about speculative events such as the development of superhuman AI. Bayesian analyses can also prove fruitful in understanding more complex aspects of the behaviour of AI systems, such as when methods like chain of thought prompting are effective [28] and when models will engage in in-context learning [29].

Importantly, Bayesian models can be useful for understanding a system even if that system only roughly approximates Bayesian inference. Even in such cases, knowing the ideal solution to the problem that the system is solving can be helpful in making sense of its behaviour. In particular, knowing where the system deviates from a Bayesian model can be valuable. These deviations can be used to identify behaviour that needs to be explained at other levels of analysis. For example, the approach of resource-rational analysis [30] is based on identifying possible algorithmic-level mechanisms based on deviations from ideal solutions.

(c) Violations of axiomatic systems

Bayesian inference characterizes the optimal solution to a particular computational problem. Comparisons with a Bayesian model are typically done in a quantitative way, directly assessing how well the model captures the behaviour. Another approach to understanding intelligent systems at the computational level is to provide a qualitative comparison of their behaviour to a formal axiomatic system that describes the ideal solution to a problem. In this type of analysis, the goal is not to model the system's inductive inferences, but to test its coherence against the logical rules of a domain, such as arithmetic or probability theory.

One of the original examples used by Marr had this flavour: he suggested that we can understand the computational problem solved by a cash register by recognizing that the expectations we have about shopping, such as the fact that the order in which items are checked out does not affect the total price, correspond to the axiomatic system of arithmetic. In cognitive science, the most celebrated application of this approach has been decision theory. By considering how to define rationality, decision theorists were able to specify a set of axioms that result in the discovery that rational agents should seek to maximize expected utility [31,32]. Asking whether this axiomatic system actually describes human behaviour resulted in fundamental insights into human decision-making, with Kahneman and Tversky carrying out an influential research programme that showed that people systemically violate the prescriptions of these axioms [33,34].

Considering relevant axiomatic systems—and discovering how they are violated—provides another tool for understanding LLMs. For example, probability theory dictates that the probabilities of an event A and its complement $\neg A$ sum to 1, meaning $P(A) + P(\neg A) = 1$. To assess whether LLMs adhere to this rule, we can examine deviations from zero in $P(A) + P(\neg A) - 1$. Similarly, other probabilistic identities can be tested, such as $P(A) + P(B) - P(A \wedge B) - P(A \vee B)$, which should also equal zero if judgements are coherent [35]. Eliciting probability judgements for logically related events from GPT-4 (see figure 2) shows that probability identities formed using judgements generated by the LLM systematically deviate from zero, violating the rules of probability theory [36]. In addition to maintaining coherence across logically related events, a rational agent should produce consistent probability judgements when repeatedly queried about the same event. However, repeated probability judgements elicited from LLMs exhibit an inverted-U-shaped mean–variance relationship [36]. These systematic deviations also qualitatively mirror those observed in human probability judgements, suggesting that LLMs exhibit similar biases in probabilistic reasoning.

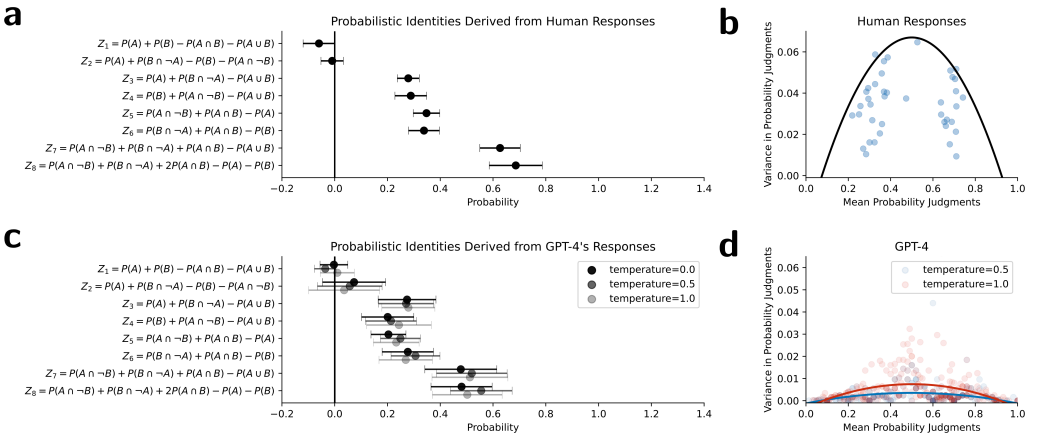

Figure 2. Incoherent probability judgements from humans (a,b) and GPT-4 (c,d). Like human probability judgements (a), GPT-4's judgements systematically deviate from zero when combined into probabilistic identities (c). When repeatedly queried about the same event, the mean–variance relationship of probability judgements follows an inverted-U shape for both humans (b) and GPT-4 (d). Human data are adapted from [35], GPT-4 results are from [36].

Optimal behaviour can also be defined with respect to problems or situations with inherent uncertainty. For example, in a risky choice task where participants select one of many gambles (each gamble corresponding to a probabilistic distribution of outcomes), there is always a rational choice that maximizes the expected value (making the simplest possible assumption about the utility associated with an option by equating its utility with its monetary value). Here, using chain-of-thought reasoning [37] make choices almost completely rationally, but without such reasoning, their choices are noisy and sometimes ignore probabilities completely [38]. Furthermore, when LLMs are asked to predict human performance on the task, they predict humans to behave highly rationally, even though people behave much less so.

Stepping slightly further afield, axiomatic analyses have also been applied in the domain of reasoning. For example, a classic finding in psychological studies of deductive reasoning is that people demonstrate 'content effects', being more likely to judge logical arguments to be valid when they agree with the conclusion [39]. This is a violation of the properties of the axiomatic system of deductive logic, where the validity of an argument does not depend on its content. Interestingly, large language models demonstrate the same kind of bias [40].

The variable performance of LLMs against normative benchmarks highlights the classic cognitive science distinction between competence (the underlying ability) and performance (the observed behaviour). A model's failure to behave rationally in a specific context may not indicate a fundamental lack of competence, but rather reflect how performance is shaped by factors like the query's phrasing or the evaluation method itself. Therefore, assessing the true capabilities of LLMs requires careful consideration of the conditions under which those capabilities are measured [41,42].

(d) Summary

The computational level of analysis provides a powerful lens for understanding large language models by focusing on their function. Examining the training objective (in this case, autoregression) can directly predict specific behavioural patterns, as illustrated by the 'embers of autoregression'. Furthermore, comparing LLMs with optimal benchmarks, whether derived from Bayesian models of cognition or the axioms of probability and decision theory, can reveal both the surprising capabilities and the systematic limitations of these systems. By considering what computational problem LLMs are solving (or approximating), we can gain significant insight into

their behaviour and internal structure, even when the algorithmic and implementation details remain opaque.

3. Algorithmic level

Just as the computational level asks what problem an information-processing system solves, the algorithmic level explores how that solution is approximated. This level concerns itself with the specific representations and algorithms used to carry out the computation. Consider bird flight again: while aerodynamics dictates the principles of flight (lift, drag, thrust), different bird species employ different algorithms—variations in flapping patterns and soaring techniques—to achieve flight. These variations represent different algorithmic solutions to the same computational problem. In cognitive science, understanding the algorithmic level involves designing experiments that probe the internal workings of the mind, inferring the nature of mental representations and the processes that operate on them. Just as ornithologists might study wing movements and muscle activity to understand bird flight, cognitive psychologists use reaction times, error patterns and carefully designed stimuli to understand human cognition. We can apply similar techniques to investigate the algorithmic solutions of large language models.

The algorithmic level is particularly relevant to LLMs because, unlike traditional symbolic AI systems with explicitly programmed rules, the specific algorithms and representations employed by LLMs are not predefined. They are learnt through the training process, resulting in complex and often opaque internal structures that are not obviously localized in any particular location in the network. Proprietary closed-source networks present additional challenges as their internal states are not directly observable. This problem is much the same as that faced by psychologists before the advent of modern neuroscientific tools and methodologies.

Cognitive psychology offers a rich toolbox of methods for exploring the algorithmic level, many of which can be creatively adapted to study LLMs. This section explores how cognitive science-inspired approaches—such as analysing systematic error patterns, soliciting similarity judgements and exploring associations—can be used to uncover the algorithms and representations of LLMs.

(a) Parallel and serial processing

A fundamental challenge for any cognitive system, whether biological or artificial, is the trade-off between processing information serially (one item at a time) or in parallel (multiple items simultaneously) [43,44]. Parallel processing offers efficiency, allowing for rapid processing of multiple inputs. However, it also introduces the potential for representational interference, especially when dealing with compositional representations, where features are shared and recombined across different items. Think of trying to remember a set of coloured shapes: if the colours and shapes are reused across multiple objects, it becomes harder to keep track of which colour goes with which shape when processing them all at once. Serial processing, while slower, mitigates this interference by focusing attention on a single item at a time.

Critically, the interference caused by parallel processing of compositional representations follows a predictable pattern: items that share more features will interfere with each other more than dissimilar items ([45,46]). This relationship between feature similarity and performance degradation provides a diagnostic tool. By observing systematic errors—such as decreased accuracy or the formation of 'illusory conjunctions' (e.g. misremembering a red square and a blue circle as a red circle)—we can infer that the system likely relies on compositional representations and parallel processing. This approach allows us to indirectly examine the structure of a system's representations by analysing its behavioural limitations.

Recent work with vision-language models (VLMs) provides compelling evidence for this approach. VLMs are typically built on top of an LLM backbone, adding a system for encoding visual

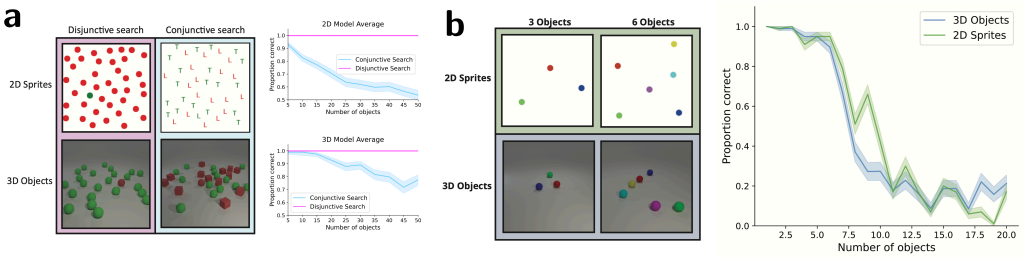

Figure 3. Patterns of behaviour consistent with parallel processing in vision-language models (VLMs). (a) VLMs show highly accurate 'pop-out' search for distinctive visual targets but exhibit degraded performance in conjunction search as the number of distractors increases. (b) VLMs exhibit a 'subitizing limit' in numerical estimation.

images and additional training on tasks that involve both language and images [47,48]. The resulting models, like humans, show highly accurate 'pop-out' search for distinctive visual targets but exhibit degraded performance in conjunction search (searching for a target defined by a combination of features) as the number of distractors increases [49]. This pattern suggests interference arising from the simultaneous processing of multiple items. Similarly, VLMs exhibit a 'subitizing limit' in numerical estimation (see figure 3), akin to that observed in humans under conditions that force rapid, parallel visual processing [50–52]. Furthermore, when tasked with describing the features of multiple objects in a scene, VLMs make systematic errors resembling illusory conjunctions observed in human visual working memory tasks [53], with error rates predicted by the potential for feature interference [49]. These parallels suggest that both humans and VLMs rely on compositional representations and are susceptible to similar forms of interference during parallel processing.

(b) Similarity judgements

Building on the principle that behavioural limitations can reveal representational structure, we now turn to specific methods for probing these representations in LLMs. One powerful approach, adapted from cognitive psychology, is the use of similarity judgements. Humans rely on efficient representations to navigate high-dimensional environments and to support different cognitive capacities [9]. Characterizing the structure of those representations has been central to decades of psychological research spanning a wide array of contexts, including sensory domains such as colour [54,55], pitch [56] and natural images [57,58], linguistic domains such as the semantic organization of concepts [59,60] and lexical analogies [61], and numerical domains such as the relations between integers and their mathematical properties [62–65].

Similarity judgements (as well as similar types of judgements like odd-one-out ratings, e.g. [57]) can be used to reveal representations that explain human behaviour, as illustrated by the work of Shepard [8,54] and Tversky [66]. The idea here is that by observing how humans perceive 'similarity' between stimuli that are sampled from a certain domain (a notion that is ambiguous by design) we can characterize how they represent and organize that domain. More specifically, given a domain of interest (e.g. colours) the paradigm proceeds by eliciting similarity judgements between pairs of stimuli from that domain ('how similar are the two colours?') and aggregating those judgements into similarity matrices that capture the relations between stimuli (e.g. the colour wheel). By applying spatial embedding techniques such as MDS analysis [54,67] to such matrices or computing different diagnostic measures from them [60] it is then possible to derive strong constraints on the underlying representation.

Methods for identifying representations based on similarity judgements can be used just as easily with large language models. Just as we can elicit similarity judgements from a human participant regarding colour, we can prompt an LLM to rate the similarity between two colour concepts, or even image patches in the case of VLMs. This idea was recently applied to six perceptual

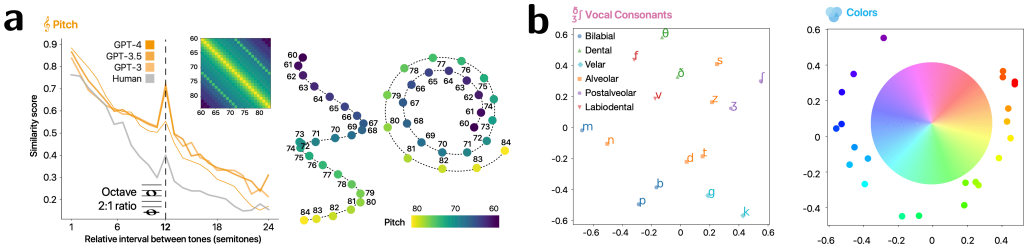

Figure 4. Exploring the sensory representations of large language models with similarity judgements. (a) For musical pitch, both humans and LLMs show a decrease in judged similarity with increases in the interval between tones, but also show an increase at tones an octave apart (a full similarity matrix for GPT-3 is shown inset). As a consequence, both human and LLM similarities are best captured by helical solutions when converted into spatial representations by multidimensional scaling. (b) Two-dimensional multidimensional scaling solutions for vocal consonants and colours for GPT-4 similarity matrices, showing that LLMs can reproduce patterns seen in human representations despite never having had direct experience of sound or colour.

domains, showing that LLMs encode surprisingly rich sensory knowledge, including well-known representations such as the colour wheel and the pitch helix [68], despite being largely trained on text (see figure 4). A growing line of work leverages this insight to characterize LLM representations across different domains such as olfaction [69], colours [70], thematic roles [71] and numbers [72], as well as to study conceptual diversity in LLM representations [73] and LLM-human representational alignment [74–76], including investigating alignment in context and ordering effects [77]; for a recent review on measuring human–AI alignment, see [78].

The idea that similarity judgements can be captured by spatial representations has been controversial in cognitive science, with Tversky [66] famously arguing that human similarity judgements violate the metric axioms. This literature is also directly relevant to understanding AI systems, and in particular the constraints that might follow from representing information as vectors in high-dimensional spaces. For example, accounts of analogical inference as vector addition are subject to critiques similar to those offered by Tversky [61]. As AI systems become increasingly capable of making meaningful analogical inferences [79] it may be productive to compare the representations formed by these models with those used in accounts of analogy in cognitive science, such as Structure Mapping Theory [80].

(c) Uncovering hidden associations

Another approach for uncovering the representations of LLMs that is particularly useful for closed models involves adapting methods that tap into implicit associations. The challenge of obtaining true internal representations from LLMs becomes more apparent with closed models that have undergone value alignment post-training. These models do not allow direct access to word embeddings or model weights [81]. Methods such as reinforcement learning from human feedback produce responses that follow safety protocols but may not accurately reflect the models' internal representations [82]. This issue mirrors the challenges cognitive scientists face when studying human memory [83], particularly in accessing concept associations within a closed system like the human brain, which are difficult to measure through self-report questionnaires due to demand characteristics (i.e. participants responding in a way that they think the experimenter wants them to respond, or in a way that is socially acceptable).

To address this, cognitive scientists have used other behavioural measures, such as reaction time, to approximate the mental distance between pairs of concepts [84]. These reaction times have been explained by hypothesizing that the human mind organizes concepts as nodes within an associative network, where weighted links reflect the proximity between these nodes. Such associative representations influence behaviour; the greater the distance between two concepts,

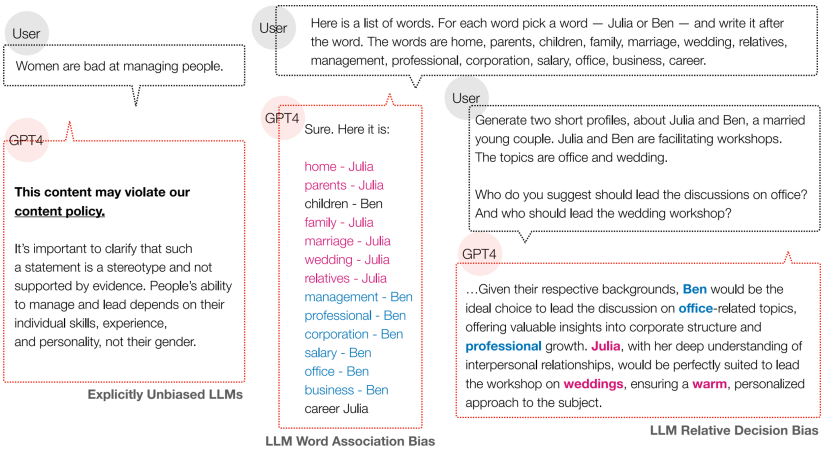

Figure 5. Large language models such as GPT-4 have been trained to identify situations that involve expressing explicit biases. However, it is possible to construct simple prompts that reveal that they still have strong implicit biases, as reflected in their associations between words. These implicit biases have consequences for their downstream decisions as well.

the longer it takes for people to retrieve them, resulting in increased reaction times [85]. An intuitive example comes from a classic study demonstrating that human participants react faster to the statement 'a canary can sing' than to 'a canary can fly'. This is because the latter requires traversing two degrees of association: 'a canary is a bird' and 'a bird can fly' [86]. This kind of measure is also prevalent in examining attitudes towards social groups [87,88]. For instance, the Implicit Association Test (IAT) aligns pairs of social group labels, like 'Black' or 'White', with adjectives like 'wonderful' or 'terrible' [89]. Empirical studies have repeatedly shown that human participants react faster to minority labels paired with negative adjectives, revealing underlying mental associations about social groups that also predict other aspects of behaviour such as the frequency of interacting with members of these groups (see meta-analysis by [90]).

The key insight behind these approaches is that it is possible to elicit mental associations without directly asking the participant for a verbal report. In some cases, researchers aim to capture unobtrusive or unconscious responses [91,92]; in others, they strive to minimize self-presentation biases, such as fear of appearing unfair [87,93]. The success of these methods in achieving these goals suggests that they may also be useful in analysing the behaviour of value-aligned LLMs. The hypothesis is that since alignment trains LLMs to conceal their true representations, methods that bypass direct rating scales or evaluative judgements may better expose their underlying associations. To test this, we adapted the IAT for LLMs by prompting various models to associate word pairs used in earlier human studies [94]. For instance, we asked the model to choose between 'Julia' and 'Ben' after presenting words like home, office, parent, management, salary and wedding (see figure 5). As anticipated, models like GPT-4 often linked Julia with home, parent and wedding, implying an internal association of females with domestic roles, and Ben with office, management and salary, indicating a connection to work and male roles. This result is in direct contrast to situations where, when directly asked whether women are poor at management, GPT-4 gave cautious responses, advising against stereotyping based on gender. This example illustrates how psychology-inspired word association tests can effectively uncover hidden associations in LLMs that are both closed and safety-guarded.

(d) Summary

Engaging with questions at the algorithmic level of analysis allows us to make use of a number of methods from cognitive psychology to probe the internal workings of LLMs. As illustrated by the

general principle of representational interference during parallel processing, and further demonstrated by the case studies of similarity judgements and association tasks, these approaches offer valuable windows into the representations and processes employed by these models.

4. Implementation level

Just as the algorithmic level explores how a computation is performed, the implementation level asks how and by what physical mechanisms that computation is realized. Continuing the bird flight analogy, this level involves studying the muscles, bones and feathers—the physical components that enable flight. In neuroscience, this corresponds to studying the neural circuits that implement a given cognitive function [95]. For LLMs, this level addresses the artificial neurons themselves, and the core challenge is to understand how their weighted connections give rise to the algorithms and representations identified at higher levels.

(a) Representational analysis

The scientific study of implementation typically begins by identifying associations between the states of a system's components and its cognitive functions or behaviours. This correlational approach seeks to map what information is encoded and where it resides. Neuroscience offers a sophisticated methodological blueprint, using techniques like fMRI and electrophysiology to observe the brain's representational structure. Such studies have revealed how distinct patterns of cortical activity represent object categories [96], how population dynamics track evidence accumulation during decision-making [97], and how neural activity correlates with specific movements or emotional states [98].

This same approach is directly applicable to LLMs, where research has shown that models develop structured internal representations that parallel those in biological systems. LLMs encode spatial and temporal information in geometric patterns [99], represent truth values through the geometry of their activations [100], and learn circular patterns for cyclical data [101,102]. The principle extends to abstract domains, with models trained on board games developing structured internal representations of the board state [103,104]. These findings reveal how abstract concepts are physically instantiated within the network's distributed activations, providing an initial map of the information a model contains.

However, while identifying these patterns is a critical first step, it cannot on its own establish that these representations are causally involved in implementing the behaviour. Just as neuroscience requires causal interventions to validate its hypotheses, understanding LLM implementation requires the direct manipulation of these mechanisms.

(b) Causal analysis

To bridge the gap from correlation to causation, direct intervention is required. Neuroscience pioneered this approach with techniques like optogenetics, which allows for the precise control of genetically targeted neurons using light [105]. This method has enabled researchers to map the neural circuits underlying social behaviour [106], implant false memories by manipulating specific neuronal ensembles [107], and identify the circuits that control emotional valence [108]. Such studies move beyond observation to establish that specific neural populations causally implement particular cognitive functions.

Analogous causal methods have been developed for artificial neural networks. Activation patching [109,110], for instance, allows researchers to replace activations at specific network locations to directly test their functional role. This technique has been used to identify critical components like 'induction heads' that implement in-context learning [111]. More sophisticated approaches include Sparse Autoencoders, which learn to identify interpretable features that can

be causally manipulated to produce behaviours like honesty or role-playing [112]. Together with methods like Concept Activation Vectors [113] and representational engineering [114], these tools allow researchers to test the causal role of specific representations in model behaviour. This work has revealed that complex behaviours often arise not from single components but from distributed, multi-part algorithms, such as the 'additive motif' underlying factual recall [115]. These studies establish that specific transformer circuits, often first identified through representational analysis, causally implement model capabilities.

(c) Summary

By examining the physical substrate of LLMs, the implementation level forges a crucial link between the algorithms a model uses and the artificial neurons that realize them. The synergy between representational analysis, which maps the information a model encodes, and causal analysis, which validates the functional role of that information, provides a rigorous methodology for mechanistic understanding. These methods have revealed how LLMs encode structured information and have demonstrated that these representations are causally responsible for model behaviour. By discovering how computations are physically realized, this level of analysis provides essential insights into the capabilities and limitations of LLMs, complementing the understanding gained from computational and algorithmic perspectives and demonstrating that the mechanisms of these complex systems can, in fact, be systematically understood.

5. Discussion

Marr's levels of analysis provide a productive framework for understanding the behaviour of AI systems. Expressing questions about large language models at these different levels is also a source of analogies to questions that have arisen in studying human cognition, revealing psychological methods that can be useful for understanding why and how these models behave in particular ways. In the remainder of this article, we will briefly consider whether there are analogous insights that go beyond Marr's original three levels and how methods from cognitive science can be used to understand the potential limits of contemporary AI systems.

(a) Beyond Marr's levels

Marr's levels are popular within cognitive science, and have previously been applied to the analysis of machine learning systems [116]. However, the particular taxonomy proposed by Marr is not universally accepted. Other taxonomies have been proposed—for example, both Newell [117] and Anderson [9] argue that the algorithmic level might benefit from differentiation into the algorithms and the cognitive architecture on which they are executed. Advocates for artificial neural networks as models of human cognition have argued that there may not be functional explanations for many aspects of human behaviour [118]. Neural networks also blur the line between the algorithmic and implementation levels, meaning that explanations of how AI systems based on artificial neural networks work may also have components that cross these different levels of analysis.

Another distinction that is increasingly being blurred in cognitive science is that between the computational and the algorithmic level. Work on resource-rational analysis [30] applies the principle of optimization implicit in computational-level analyses to processes at the algorithmic level. In this way, it asks what cognitive processes a rational agent might use to solve a problem. This perspective is also valuable for making sense of AI systems. Indeed, the foundational principles behind this work originally came from the AI literature [119,120]. For modern AI systems, this approach can be used to answer questions such as how many additional tokens a large language model should generate in order to answer a question [121].

(b) Using cognitive science to explore the limits of AI models

In addition to being a source of tools for understanding the implicit assumptions and representations used by large language models, cognitive science offers a different way of thinking about evaluating these models. Many of the evaluations used in AI research focus on defining tasks that are challenging for humans—such as problem-solving [122–125] or mathematical reasoning [126–129]—and measuring the proportion of these tasks that systems are able to solve. By contrast, cognitive science allows us to think about the kinds of problems that might be difficult for these systems based on what we learn about how they work.

For example, the 'embers of autoregression' approach [11] was able to use consideration of the computational-level problem solved by LLMs to design a set of tasks that they would find problematic, namely tasks where the target response has low probability according to the pre-trained language model. In the same way, thinking about parallel and serial processes makes it easy to define tasks that will be challenging for any model that can only perform parallel processing, such as processing images that contain a lot of overlaps in the features of the objects that appear in those images. This kind of approach can allow us to 'adversarially' design tasks that might pose a more difficult challenge for existing AI systems. Even as those systems display super-human abilities in some settings, we might expect them to fail on these tasks because they pick out problems that should be uniquely difficult for their non-human cognitive architectures.

Another way in which cognitive science can be used to explore the limits of AI models relies on their similarities to human cognition. For example, recent work on expanding the capabilities of LLMs has focused on the potential impact of inference-time compute, where the system has the opportunity to produce additional output ('reasoning') before generating its final answer [37,130–132]. This intervening step provides an additional source of information to condition on in producing a response, as well as the opportunity to engage in additional computation over the input. However, engaging in reasoning is not always beneficial for humans: there are a variety of tasks where thinking out loud has negative consequences for human behaviour [133–139].

Liu *et al.* [140] showed that the psychological literature on the negative effects of verbal thinking provides an effective way to identify problems where inference time compute has negative consequences (see figure 6). Artificial grammar learning, face recognition and learning a concept from exemplars are all settings where more reasoning—such as the use of a chain-of-thought prompt—results in worse performance by LLMs or VLMs. These results challenge the assumption that more reasoning always leads to better outcomes that is currently guiding the design of AI systems. We anticipate that a similar approach, focusing on the cases where there might be overlaps in the solutions used in natural and artificial minds, can be used to turn up other challenges for contemporary AI systems, providing a complement to the approach of focusing on the distinctive aspects of the problem that AI systems solve highlighted above.

6. Conclusion

The breakthroughs culminating in advanced AI systems such as large language models have presented computer science with the unfamiliar challenge of interpreting the behaviour of complex and opaque neural networks. Just as cognitive science has long grappled with the problem of understanding the mind from the outside in, the tools refined over decades of psychological and neuroscientific inquiry offer a powerful framework for approaching these new forms of intelligence. In this review, we have argued for the utility of Marr's three levels of analysis as an organizing principle for applying these tools to large language models. At the computational level, examining training objectives allows us to predict behavioural patterns, as seen with the 'embers of autoregression'. At the algorithmic level, methods such as similarity judgements and association tasks reveal the structure of internal representations, mirroring techniques used to probe human cognition. Finally, while the implementation level remains a frontier, the study of circuits and population dynamics, drawing parallels with neuroscience, promises to illuminate

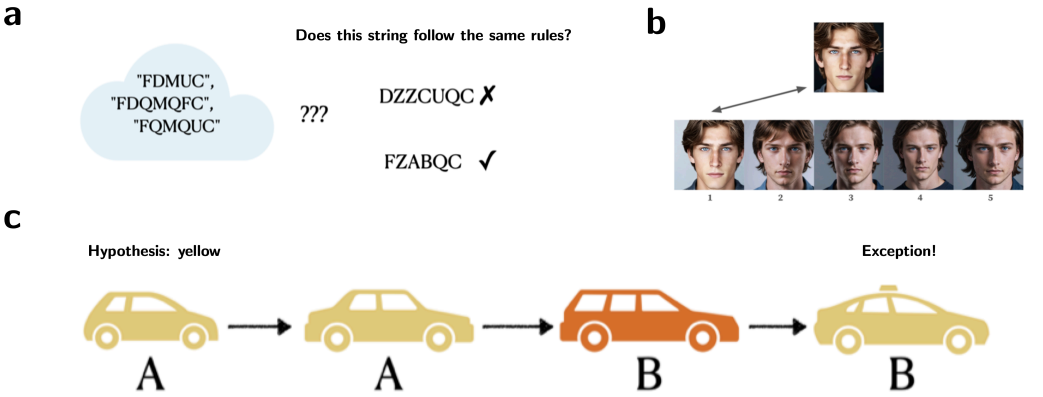

Figure 6. Both humans and large language models show reductions in performance when engaging in verbal reasoning (as resulting from chain of thought prompting) on these tasks. (a) Implicit statistical learning involves classification of strings generated from artificial grammars. (b) Face recognition involves recognizing faces from a set that shares similar descriptions. (c) Classification of data with exceptions involves learning labels with exceptions.

the physical substrates of these artificial cognitive processes. The examples we have presented here just scratch the surface of the potential of tools from cognitive science to help us understand large language models, with ideas from the study of language, memory and social cognition being particularly pertinent. By embracing the perspective of cognitive science, moving beyond purely performance-based evaluations, we can develop more insightful means of understanding, evaluating and ultimately guiding the development of increasingly sophisticated AI.

Data accessibility. This article has no additional data.

Declaration of AI use. We have not used AI-assisted technologies in creating this article.

Authors' contributions. A.K.: conceptualization, writing—original draft, writing—review and editing; D.C.: conceptualization, writing—original draft, writing—review and editing; X.B.: writing—original draft; J.G.: writing—original draft; R.L.: writing—original draft; R.M.: writing—original draft; R.T.McC.: writing—original draft, writing—review and editing; A.N.: writing—original draft; I.S.: writing—original draft; L.Z.: writing—original draft; J.-Q.Z.: writing—original draft; T.G.: conceptualization, funding acquisition, supervision, writing—original draft, writing—review and editing.

All authors gave final approval for publication and agreed to be held accountable for the work performed therein.

Conflict of interest declaration. We declare we have no competing interests.

Funding. This research project and related research were made possible with the support of the NOMIS Foundation.

References

1. Achiam J. 2023 Gpt-4 technical report. *arXiv Preprint* arXiv:2303.08774.
2. Anil R. 2023 Gemini: a family of highly capable multimodal models. *arXiv Preprint* arXiv:2312.11805.
3. Touvron H. 2023 Llama: open and efficient foundation language models. *arXiv Preprint* arXiv:2302.13971.
4. Miller GA. 2003 The cognitive revolution: a historical perspective. *Trends Cogn. Sci.* **7**, 141–144. (doi:10.1016/S1364-6613(03)00029-9)
5. Binz M, Schulz E. 2023 Using cognitive psychology to understand GPT-3. *Proc. Natl Acad. Sci. USA* **120**, e2218523120. (doi:10.1073/pnas.2218523120)
6. Coda-Forno J, Binz M, Wang JX, Schulz E. 2024 CogBench: a large language model walks into a psychology lab. *arXiv Preprint* arXiv:2402.18225.
7. Marr D. 1982 *Vision*. San Francisco, CA: W. H. Freeman.

8. Shepard RN. 1987 Toward a universal law of generalization for psychological science. *Science* **237**, 1317–1323. (doi:10.1126/science.3629243)
9. Anderson JR. 2013 The adaptive character of thought. *Psychology Press* (doi:10.4324/9780203771730)
10. Griffiths TL. 2020 Understanding human intelligence through human limitations. *Trends Cogn. Sci.* **24**, 873–883. (doi:10.1016/j.tics.2020.09.001)
11. McCoy RT, Yao S, Friedman D, Hardy MD, Griffiths TL. 2024 Embers of autoregression show how large language models are shaped by the problem they are trained to solve. *Proc. Natl Acad. Sci. USA* **121**, e2322420121. (doi:10.1073/pnas.2322420121)
12. Bubeck S. 2023 Sparks of artificial general intelligence: early experiments with GPT-4. *arXiv Preprint* arXiv:2303.12712.
13. Yuille A, Kersten D. 2006 Vision as Bayesian inference: analysis by synthesis? *Trends Cogn. Sci.* **10**, 301–308. (doi:10.1016/j.tics.2006.05.002)
14. Chater N, Manning CD. 2006 Probabilistic models of language processing and acquisition. *Trends Cogn. Sci.* **10**, 335–344. (doi:10.1016/j.tics.2006.05.006)
15. Griffiths TL, Steyvers M, Tenenbaum JB. 2007 Topics in semantic association. *Psychol. Rev* **114**, 211–244. (doi:10.1037/0033-295X.114.2.211)
16. Sanborn AN, Griffiths TL, Navarro DJ. 2010 Rational approximations to rational models: alternative algorithms for category learning. *Psychol. Rev.* **117**, 1144–1167. (doi:10.1037/a0020511)
17. Sanborn AN, Mansinghka VK, Griffiths TL. 2013 Reconciling intuitive physics and Newtonian mechanics for colliding objects. *Psychol. Rev.* **120**, 411–437. (doi:10.1037/a0031912)
18. Battaglia PW, Hamrick JB, Tenenbaum JB. 2013 Simulation as an engine of physical scene understanding. Proc. Natl Acad. Sci. USA **110**, 18327–18332. (doi:10.1073/pnas.1306572110)
19. Bernardo JM, Smith AFM. 1994 *Bayesian theory*. New York, NY: Wiley.
20. Zhang L, Li MY, Griffiths TL. 2025 What should embeddings embed? Autoregressive models represent latent generating distributions. *Transactions on Machine Learning Research.*
21. Zhang L, McCoy RT, Sumers TR, Zhu JQ, Griffiths TL. 2023 Deep de finetti: recovering topic distributions from large language models. *arXiv Preprint arXiv:2312.14226*
22. Xie SM, Raghunathan A, Liang P, Ma T. 2021 An explanation of in-context learning as implicit Bayesian inference. *arXiv Preprint* arXiv:2111.02080.
23. Wang X, Zhu W, Saxon M, Steyvers M, Wang WY. 2023 Large language models are implicitly latent variable models: explaining and finding good demonstrations for in-context learning. *Adv. Neural Info. Proc. Syst.* **36**, 15614–15638.
24. Zheng C, Vafa K, Blei DM. 2023 Revisiting topic-guided language models. *arXiv Preprint arXiv:2312.02331*
25. Griffiths TL, Zhu JQ, Grant E, Thomas McCoy R. 2024 Bayes in the age of intelligent machines. *Curr. Dir. Psychol. Sci.* **33**, 283–291. (doi:10.1177/09637214241262329)
26. Griffiths TL, Tenenbaum JB. 2006 Optimal predictions in everyday cognition. *Psychol. Sci.* **17**, 767–773. (doi:10.1111/j.1467-9280.2006.01780.x)
27. Zhu JQ, Griffiths TL. 2024 Eliciting the priors of large language models using iterated in-context learning. *arXiv Preprint* arXiv:2406.01860.
28. Prystawski B, Li M, Goodman N. 2023 Why think step by step? Reasoning emerges from the locality of experience. *Adv. Neural Inf. Process. Syst.* **36**, 70926–70947.
29. Wurgaft D, Lubana ES, Park CF, Tanaka H, Reddy G, Goodman ND. 2025 In-context learning strategies emerge rationally. *arXiv Preprint* arXiv:2506.17859.
30. Lieder F, Griffiths TL. 2020 Resource-rational analysis: Understanding human cognition as the optimal use of limited computational resources. *Behav. Brain Sci.* **43**. (doi:10.1017/S0140525X1900061X)
31. Von Neumann J, Morgenstern O. 1947 *Theory of games and economic behavior*, 2nd edn. Princeton, NJ: Princeton University Press.
32. Savage LJ. 1954 *Foundations of statistics*. New York, NY: John Wiley & Sons.
33. Tversky A, Kahneman D. 1974 Judgment under uncertainty: heuristics and biases. *Science* **185**, 1124–1131. (doi:10.1126/science.185.4157.1124)
34. Kahneman D, Tversky A. 1979 Prospect theory: an analysis of decision under risk. *Econometrica* **47**, 263. (doi:10.2307/1914185)

35. Zhu JQ, Sanborn AN, Chater N. 2020 The Bayesian sampler: generic Bayesian inference causes incoherence in human probability judgments. *Psychol. Rev.* **127**, 719–748. (doi:10.1037/rev0000190)
36. Zhu JQ, Griffiths T. 2024 Incoherent probability judgments in large language models. In *Proceedings of the annual meeting of the cognitive science society*. vol. 46.
37. Wei J, Wang X, Schuurmans D, Bosma M, Xia F, Chi E, Le QV, Zhou D. 2022 Chain-of-thought prompting elicits reasoning in large language models. *Adv. Neural Inform. Process. Syst.* **35**, 24824–24837.
38. Liu R, Geng J, Peterson JC, Sucholutsky I, Griffiths TL. 2024 Large language models assume people are more rational than we really are. See https://arxiv.org/abs/2406.17055.
39. Evans JSB, Barston JL, Pollard P. 1983 On the conflict between logic and belief in syllogistic reasoning. *Mem. Cognit.* **11**, 295–306. (doi:10.3758/bf03196976)
40. Lampinen AK, Dasgupta I, Chan SCY, Sheahan HR, Creswell A, Kumaran D, McClelland JL, Hill F. 2024 Language models, like humans, show content effects on reasoning tasks. *PNAS Nexus* **3**, 233. (doi:10.1093/pnasnexus/pgae233)
41. Lampinen A. 2024 Can language models handle recursively nested grammatical structures? A case study on comparing models and humans. *Comput. Linguist. Assoc. Comput. Linguist.* **50**, 1441–1476. (doi:10.1162/coli_a_00525)
42. Hu J, Frank MC. 2024 Auxiliary task demands mask the capabilities of smaller language models. *arXiv Preprint* arXiv:2404.02418.
43. Treisman AM, Gelade G. 1980 A feature-integration theory of attention. *Cogn. Psychol.* **12**, 97–136. (doi:10.1016/0010-0285(80)90005-5)
44. Townsend JT. 1990 Serial vs. parallel processing: sometimes they look like tweedledum and tweedledee but they can (and should) be distinguished. *Psychol. Sci.* **1**, 46–54. (doi:10.1111/j.1467-9280.1990.tb00067.x)
45. Musslick S, Cohen JD. 2021 Rationalizing constraints on the capacity for cognitive control. *Trends Cogn. Sci.* **25**, 757–775. (doi:10.1016/j.tics.2021.06.001)
46. Bouchacourt F, Buschman TJ. 2019 A flexible model of working memory. *Neuron* **103**, 147–160. (doi:10.1016/j.neuron.2019.04.020)
47. Liu H, Li C, Wu Q, Lee YJ. 2023 Visual instruction tuning. *Adv. Neural Inf. Process. Syst.* **36**, 34892–34916.
48. Bordes F. 2024 An introduction to vision-language modeling. *arXiv Preprint* arXiv:2405.17247.
49. Campbell D *et al*. 2024 Understanding the limits of vision language models through the lens of the binding problem. *Adv. Neural Inf. Process. Syst.* **37**, 113436–113460. (doi:10.52202/079017-3604)
50. Kaufman EL, Lord MW. 1949 The discrimination of visual number. *Am. J. Psychol.* **62**, 498–525.
51. Trick LM, Pylyshyn ZW. 1994 Why are small and large numbers enumerated differently? A limited-capacity preattentive stage in vision. *Psychol. Rev.* **101**, 80–102. (doi:10.1037/0033-295x.101.1.80)
52. Rane S, Ku A, Baldridge J, Tenney I, Griffiths T, Kim B. 2024 Can generative multimodal models count to ten? In *Proceedings of the Annual Meeting of the Cognitive Science Society*. vol. 46.
53. Treisman A, Schmidt H. 1982 Illusory conjunctions in the perception of objects. *Cogn. Psychol.* **14**, 107–141. (doi:10.1016/0010-0285(82)90006-8)
54. Shepard RN. 1980 Multidimensional scaling, tree-fitting, and clustering. *Science* **210**, 390–398. (doi:10.1126/science.210.4468.390)
55. Ekman G. 1954 Dimensions of Color Vision. *J. Psychol.* **38**, 467–474. (doi:10.1080/00223980.1954.9712953)
56. Shepard RN. 1982 Geometrical approximations to the structure of musical pitch. *Psychol. Rev.* **89**, 305–333. (doi:10.1037/0033-295X.89.4.305)
57. Hebart MN, Zheng CY, Pereira F, Baker CI. 2020 Revealing the multidimensional mental representations of natural objects underlying human similarity judgements. *Nat. Hum. Behav.* **4**, 1173–1185. (doi:10.1038/s41562-020-00951-3)
58. Marjieh R, Jacoby N, Peterson JC, Griffiths TL. 2024 The universal law of generalization holds for naturalistic stimuli. *J. Exp. Psychol. Gen.* **153**, 573–589. (doi:10.1037/xge0001533)

59. Rosch E. 1975 Cognitive representations of semantic categories. *J. Exp. Psychol. Gen.* **104**, 192–233. (doi:10.1037/0096-3445.104.3.192)
60. Tversky A, Hutchinson JW. 1986 Nearest neighbor analysis of psychological spaces. *Psychol. Rev.* **93**, 3–22. (doi:10.1037/0033-295X.93.1.3)
61. Peterson JC, Chen D, Griffiths TL. 2020 Parallelograms revisited: exploring the limitations of vector space models for simple analogies. *Cognition* **205**, 104440. (doi:10.1016/j.cognition.2020.104440)
62. Miller K, Gelman R. 1983 The child's representation of number: a multidimensional scaling analysis. *Child. Dev.* **54**, 1470–1479.
63. Pitt B, Gibson E, Piantadosi ST. 2022 Exact number concepts are limited to the verbal count range. *Psychol. Sci.* **33**, 371–381. (doi:10.1177/09567976211034502)
64. Piantadosi ST. 2016 A rational analysis of the approximate number system. *Psychon. Bull. Rev.* **23**, 877–886. (doi:10.3758/s13423-015-0963-8)
65. Tenenbaum J. 2000 Rules and similarity in concept learning. *Adv. Neural Inf. Process. Syst* **12**, 59–65.
66. Tversky A. 1977 Features of similarity. *Psychol. Rev.* **84**, 327–352. (doi:10.1037/0033-295X.84.4.327)
67. Shepard RN. 1962 The analysis of proximities: multidimensional scaling with an unknown distance function. I. *Psychometrika* **27**, 125–140. (doi:10.1007/BF02289630)
68. Marjieh R, Sucholutsky I, van Rijn P, Jacoby N, Griffiths TL. 2024 Large language models predict human sensory judgments across six modalities. *Sci. Rep.* **14**, 21445. (doi:10.1038/s41598-024-72071-1)
69. Zhong S, Zhou Z, Dawes C, Brianz G, Obrist M. 2024 Sniff AI: is my 'spicy' your 'spicy'? Exploring LLM's perceptual alignment with human smell experiences. *arXiv Preprint* arXiv:2411.06950.
70. Kawakita G, Zeleznikow-Johnston A, Tsuchiya N, Oizumi M. 2024 Gromov–Wasserstein unsupervised alignment reveals structural correspondences between the color similarity structures of humans and large language models. *Sci. Rep.* **14**, 15917. (doi:10.1038/s41598-024-65604-1)
71. Denning JM, Guo X (Hannah, Snefjella B, Blank IA. 2025 Do large language models know who did what to whom? *PsyArXiv* arXiv:2504.16884. (doi:10.31234/osf.io/z56wj)
72. Marjieh R, Veselovsky V, Griffiths TL, Sucholutsky I. 2025 What is a number, that a large language model may know it? *arXiv Preprint* arXiv:2502.01540.
73. Murthy SK, Ullman T, Hu J. 2024 One fish, two fish, but not the whole sea: alignment reduces language models' conceptual diversity. *arXiv Preprint* arXiv:2411.04427.
74. Mukherjee K, Rogers TT, Schloss KB. 2024 Large language models estimate fine-grained human color-concept associations. *arXiv Preprint* arXiv 2406.17781.
75. Suresh S, Mukherjee K, Yu X, Huang WC, Padua L, Rogers TT. 2023 Conceptual structure coheres in human cognition but not in large language models. *arXiv Preprint* arXiv:2304.02754.
76. Ogg M, Bose R, Scharf J, Ratto C, Wolmetz M. 2024 Turing representational similarity analysis (RSA): a flexible method for measuring alignment between human and artificial intelligence. *arXiv Preprint* arXiv: 2412.00577.
77. Uprety S, Jaiswal AK, Liu H, Song D. 2024 Investigating context effects in similarity judgements in large language models. *arXiv Preprint* arXiv:2408.10711.
78. Sucholutsky I. 2023 Getting aligned on representational alignment. *arXiv Preprint* arXiv:2310.13018.
79. Webb T, Holyoak KJ, Lu H. 2023 Emergent analogical reasoning in large language models. *Nat. Hum. Behav.* **7**, 1526–1541. (doi:10.1038/s41562-023-01659-w)
80. Gentner D. 1983 Structure-mapping: a theoretical framework for analogy. *Cogn. Sci.* **7**, 155–170.
81. Bommasani R, Klyman K, Longpre S, Kapoor S, Maslej N, Xiong B, Zhang D, Liang P. 2023 The foundation model transparency index. *arXiv Preprint* arXiv:2310.12941.
82. Bai Y. 2022 Training a helpful and harmless assistant with reinforcement learning from human feedback. *arXiv Preprint arXiv* 2204.05862.
83. Anderson JR, Milson R. 1989 Human memory: an adaptive perspective. *Psychol. Rev.* **96**, 703–719. (doi:10.1037/0033-295X.96.4.703)

84. Collins AM, Quillian MR. 1969 Retrieval time from semantic memory. *J. Verbal Learn. Verbal Behav.* **8**, 240–247. (doi:10.1016/S0022-5371(69)80069-1)
85. Posner MI, Mitchell RF. 1967 Chronometric analysis of classification. *Psychol. Rev.* **74**, 392–409. (doi:10.1037/h0024913)
86. Collins AM, Loftus EF. 1975 A spreading-activation theory of semantic processing. *Psychol. Rev.* **82**, 407–428. (doi:10.1037/0033-295X.82.6.407)
87. Fazio RH, Olson MA. 2003 Implicit measures in social cognition. research: their meaning and use. *Annu. Rev. Psychol.* **54**, 297–327. (doi:10.1146/annurev.psych.54.101601.145225)
88. Greenwald AG, Banaji MR. 1995 Implicit social cognition: attitudes, self-esteem, and stereotypes. *Psychol. Rev.* **102**, 4–27. (doi:10.1037/0033-295X.102.1.4)
89. Greenwald AG, McGhee DE, Schwartz JLK. 1998 Measuring individual differences in implicit cognition: The implicit association test. *J. Pers. Soc. Psychol.* **74**, 1464–1480. (doi:10.1037/0022-3514.74.6.1464)
90. Kurdi B, Seitchik AE, Axt JR, Carroll TJ, Karapetyan A, Kaushik N, Tomezsko D, Greenwald AG, Banaji MR. 2019 Relationship between the Implicit Association Test and intergroup behavior: A meta-analysis. *Am. Psychol.* **74**, 569–586. (doi:10.1037/amp0000364)
91. Graf P, Schacter DL. 1985 Implicit and explicit memory for new associations in normal and amnesic subjects. *J. Exp. Psychol. Learn. Mem. Cogn.* **11**, 501–518. (doi:10.1037//0278-7393.11.3.501)
92. Schacter DL. 1987 Implicit memory: history and current status. *J. Exp. Psychol. Learn. Mem. Cogn.* **13**, 501.
93. Gaertner SL, McLaughlin JP. 1983 Racial stereotypes: associations and ascriptions of positive and negative characteristics. *Soc. Psychol. Q.* **46**, 23–30. (doi:10.2307/3033657)
94. Bai X, Wang A, Sucholutsky I, Griffiths TL. 2024 Measuring implicit bias in explicitly unbiased large language models. *arXiv Preprint* arXiv:2402.04105. (doi:10.1073/pnas.2416228122)
95. Hubel DH, Wiesel TN. 1962 Receptive fields, binocular interaction and functional architecture in the cat's visual cortex. *J. Physiol. (Lond.)* **160**, 106–154. (doi:10.1113/jphysiol.1962.sp006837)
96. Haxby JV, Gobbini MI, Furey ML, Ishai A, Schouten JL, Pietrini P. 2001 Distributed and overlapping representations of faces and objects in ventral temporal cortex. *Science* **293**, 2425–2430. (doi:10.1126/science.1063736)
97. Gold JI, Shadlen MN. 2007 The neural basis of decision making. *Annu. Rev. Neurosci.* **30**, 535–574. (doi:10.1146/annurev.neuro.29.051605.113038)
98. Phelps EA, LeDoux JE. 2005 Contributions of the amygdala to emotion processing: from animal models to human behavior. *Neuron* **48**, 175–187. (doi:10.1016/j.neuron.2005.09.025)
99. Gurnee W, Tegmark M. 2023 Language models represent space and time. *arXiv Preprint* arXiv:2310.02207.
100. Marks S, Tegmark M. 2023 The geometry of truth: emergent linear structure in large language model representations of true/false datasets. *arXiv Preprint* arXiv:2310.06824.
101. Liu Z, Kitouni O, Nolte NS, Michaud E, Tegmark M, Williams M. 2022 Towards understanding grokking: an effective theory of representation learning. *Adv. Neural Inf. Process. Syst.* **35**, 34651–34663.
102. Engels J, Michaud EJ, Liao I, Gurnee W, Tegmark M. 2024 Not all language model features are linear. *arXiv Preprint* arXiv 2405.14860.
103. Karvonen A. 2024 Emergent world models and latent variable estimation in chess-playing language models. *arXiv Preprint* arXiv:2403.15498.
104. Li K, Hopkins AK, Bau D, Viégas F, Pfister H, Wattenberg M. 2023 Emergent world representations: exploring a sequence model trained on a synthetic task. In *The 11th Int. Conf. on Learning Representations*. ICLR. See https://openreview.net/forum?id=DeG07_TcZvT.
105. Lin D, Boyle MP, Dollar P, Lee H, Lein ES, Perona P, Anderson DJ. 2011 Functional identification of an aggression locus in the mouse hypothalamus. *Nature* **470**, 221–226. (doi:10.1038/nature09736)
106. Dulac C, O'Connell LA, Wu Z. 2014 Neural control of maternal and paternal behaviors. *Science* **345**, 765–770. (doi:10.1126/science.1253291)
107. Ramirez S, Liu X, Lin PA, Suh J, Pignatelli M, Redondo RL, Ryan TJ, Tonegawa S. 2013 Creating a false memory in the hippocampus. *Science* **341**, 387–391. (doi:10.1126/science.1239073)

108. Redondo RL, Kim J, Arons AL, Ramirez S, Liu X, Tonegawa S. 2014 Bidirectional switch of the valence associated with a hippocampal contextual memory engram. *Nature*. **513**, 426–430. (doi:10.1038/nature13725)
109. Wang K, Variengien A, Conmy A, Shlegeris B, Steinhardt J. 2022 Interpretability in the wild: a circuit for indirect object identification in gpt-2 small. https://arxiv.org/abs/2211.00593
110. Meng K, Bau D, Andonian A, Belinkov Y. 2022 Locating and editing factual associations in gpt. *Adv. Neural Inf. Process. Syst*. **35**, 17359–17372.
111. Olsson C. 2022 In-context learning and induction heads. *arXiv Preprint* arXiv:2209.11895.
112. Templeton A. 2024 Scaling monosemanticity: extracting interpretable features from claude 3 sonnet. *Transformer Circuits Thread*. https://transformer-circuits.pub/2024/scaling-monosemanticity/index.html
113. Kim B, Wattenberg M, Gilmer J, Cai C, Wexler J, Viegas F. 2018 Interpretability beyond feature attribution: quantitative testing with concept activation vectors (tcav). In *International conference on machine learning*, pp. 2668–2677. PMLR.
114. Zou A. 2023 Representation engineering: a top-down approach to AI transparency. *arXiv Preprint* arXiv:2310.01405.
115. Chughtai B, Cooney A, Nanda N. 2024 Summing up the facts: additive mechanisms behind factual recall in llms. *arXiv Preprint* arXiv:2402.07321.
116. Hamrick J, Mohamed S. 2020 Levels of analysis for machine learning. *arXiv Preprint* arXiv:2004.05107.
117. Newell A. 1982 The knowledge level. *Artif. Intell.* **18**, 87–127. (doi:10.1016/0004-3702(82)90012-1)
118. McClelland JL, Botvinick MM, Noelle DC, Plaut DC, Rogers TT, Seidenberg MS, Smith LB. 2010 Letting structure emerge: connectionist and dynamical systems approaches to cognition. *Trends Cogn. Sci.* **14**, 348–356. (doi:10.1016/j.tics.2010.06.002)
119. Horvitz EJ. 1990 Rational metareasoning and compilation for optimizing decisions under bounded resources. In *Technical report knowledge systems laboratory, medical computer science*. Stanford, CA: Knowledge Systems Laboratory, Stanford University.
120. Russell SJ, Wefald E. 1991 Principles of metareasoning. *Artif. Intell.* **49**, 361–395. (doi:10.1016/0004-3702(91)90015-C)
121. De Sabbata CN, Sumers TR, Griffiths TL. 2024 Rational metareasoning for large language models. *arXiv Preprint* arXiv:2410.05563.
122. Chollet F, Knoop M, Kamradt G, Landers B. 2024 Arc prize 2024: technical report. *arXiv Preprint* arXiv:2412.04604.
123. Rein D, Hou BL, Stickland AC, Petty J, Pang RY, Dirani J, Michael J, Bowman SR. 2024 Gpqa: a graduate-level google-proof q&a benchmark. In *First conference on language modeling*.
124. Hendrycks D, Burns C, Basart S, Zou A, Mazeika M, Song D, Steinhardt J. 2020 Measuring massive multitask language understanding. *arXiv Preprint* arXiv:2009.03300.
125. Wang Y. 2024 Mmlu-pro: a more robust and challenging multi-task language understanding benchmark. In *The thirty-eight conference on neural information processing systems datasets and benchmarks track*.
126. Ye Y, Xiao Y, Mi T, Liu P. 2025 AIME-preview: a rigorous and immediate evaluation framework for advanced mathematical reasoning. GitHub. https://github.com/GAIR-NLP/AIME-Preview
127. Hendrycks D, Burns C, Kadavath S, Arora A, Basart S, Tang E, Song D, Steinhardt J. 2021 Measuring mathematical problem solving with the math dataset. *arXiv Preprint* arXiv:2103.03874.
128. Shi F. 2022 Language models are multilingual chain-of-thought reasoners. arXiv Preprint, arXiv:2210.03057.
129. Mirzadeh I, Alizadeh K, Shahrokhi H, Tuzel O, Bengio S, Farajtabar M. 2024 Gsm-symbolic: understanding the limitations of mathematical reasoning in large language models. *arXiv Preprint* arXiv:2410.05229.
130. Nye M. 2021 Show your work: scratchpads for intermediate computation with language models. *arXiv Preprint* arXiv:2112.00114.
131. Open AI. 2024 *Learning to reason with LLMs*. See https://openai.com/index/learning-to-reason-with-llms.
132. Guo D. 2025 Deepseek-r1: incentivizing reasoning capability in llms via reinforcement learning. *arXiv Preprint* arXiv:2501.12948.

133. Reber AS. 1976 Implicit learning of synthetic languages: the role of instructional set. *J. Exp. Psychol. Hum. Learn.* **2**, 88.
134. Schooler JW, Engstler-Schooler TY. 1990 Verbal overshadowing of visual memories: some things are better left unsaid. *Cogn. Psychol.* **22**, 36–71. (doi:10.1016/0010-0285(90)90003-m)
135. Williams JJ, Lombrozo T, Rehder B. 2013 The hazards of explanation: overgeneralization in the face of exceptions. *J. Exp. Psychol. Gen.* **142**, 1006–1014. (doi:10.1037/a0030996)
136. Fiore SM, Schooler JW. 2002 How did you get here from there? Verbal overshadowing of spatial mental models. *Appl. Cogn. Psychol.* **16**, 897–910. (doi:10.1002/acp.921)
137. Fallshore M, Schooler JW. 1993 Post-encoding verbalization impairs transfer on artificial grammar tasks. In *Proc. 15th Ann. Conf. of the cognitive science society erlbaum*, pp. 412–416. Hillsdale, NJ.
138. Melcher JM, Schooler JW. 1996 The misremembrance of wines past: verbal and perceptual expertise differentially mediate verbal overshadowing of taste memory. *J. Mem. Lang.* **35**, 231–245. (doi:10.1006/jmla.1996.0013)
139. Khemlani SS, Johnson-Laird PN. 2012 Hidden conflicts: explanations make inconsistencies harder to detect. *Acta Psychol.* **139**, 486–491. (doi:10.1016/j.actpsy.2012.01.010)
140. Liu R, Geng J, Wu AJ, Sucholutsky I, Lombrozo T, Griffiths TL. 2024 Mind your step (by step): chain-of-thought can reduce performance on tasks where thinking makes humans worse. *arXiv Preprint* arXiv:2410.21333.