Check for updates

COMPUTER SCIENCE

Machine psychology for language models

Thilo Hagendorff[1]*, Ishita Dasgupta[2], Marcel Binz[3], Stephanie Chan[2], Andrew Lampinen[2], Jane X. Wang[2], Zeynep Akata[4], Eric Schulz[3]

Copyright © 2026 The Authors, some rights reserved; exclusive licensee American Association for the Advancement of Science. No claim to original U.S. Government Works. Distributed under a Creative Commons Attribution License 4.0 (CC BY).

Large language models (LLMs) show increasingly advanced emergent capabilities and are being incorporated across various societal domains. Understanding their behavior and reasoning abilities is therefore important. We argue that a fruitful direction for research is engaging LLMs in experiments inspired by psychology that have traditionally been aimed at understanding human cognition and behavior. Here, we highlight and summarize theoretical perspectives, experimental paradigms, and computational analysis techniques that this approach brings to the table. It paves the way for a "machine psychology" for LLMs that goes beyond performance measurements and focuses instead on computational insights that move us toward a better understanding and discovery of emergent abilities and behavioral patterns in LLMs. We review existing work taking this approach, synthesize best practices, and highlight promising future directions. We also describe the important caveats of applying methodologies designed for understanding humans to machines. We posit that leveraging tools from experimental psychology to study LLMs will become increasingly valuable as models evolve to be more powerful, opaque, multimodal, and integrated into complex agentic real-world settings.

INTRODUCTION

Recent advances in computing power, data availability, and machine learning algorithms have yielded powerful artificial intelligence (AI) systems. Among these, large language models (LLMs), gigantic neural network architectures trained on large amounts of text, have seen a particularly meteoric rise in their influence. Contributing to that is that LLMs are wide-ranging in the kinds of tasks they can do—from writing text or code to calling tools, accessing the internet, retrieving external information, reasoning about complex problems, navigating multiple modalities like vision and speech, and many more. The ever-growing capabilities of these systems make them not only challenging but also increasingly important to understand and explain, especially since these expanding capabilities also bring greater potential for unforeseen harm (*1*–*3*). A promising way of explaining LLMs, as we want to argue in this article, is machine psychology.

Scientific disciplines explain phenomena at different levels of abstraction. These levels are complementary. Complex phenomena cannot usually be understood solely at the most fine-grained level of description (e.g., via physics), nor at the most abstract one (e.g., via philosophy). AI systems similarly invite explanation at multiple levels. At one extreme, mechanistic interpretability aims to characterize how model parameters and activation patterns implement computations within neural networks. At the other extreme, recent work has begun to examine how intelligent machines participate in broader "social" or "cultural" dynamics—for example, through the concept of machine culture, which studies how AI systems influence processes of cultural variation, transmission, and selection in human-machine ecosystems (*4*). Between these levels lies the behavioral level, the study of patterns in the observable behavior of LLMs, and the strategies that give rise to them. We argue that this intermediate level of explanation—analogous to the role psychology plays in the study of natural intelligence—motivates the emerging field of machine psychology, whose tenets we want to describe in this article. Rather than replacing mechanistic analyses or "societal" perspectives, machine psychology provides a complementary framework for generating and testing hypotheses about how LLMs behave.

Machine psychology can encompass a broad range of research programs that use concepts or methods from psychology to study LLMs. For example, psychology has long influenced computational cognitive science, where models are used to test theories of human cognition (*5*). Here, however, we focus on a more narrow and emerging strand of work: the use of psychology-inspired behavioral experiments to probe the capabilities and strategies of (multimodal) LLMs (see Fig. 1). We adopt this narrower scope since LLMs provide a natural interface for such experiments, as they can understand instructions, generate language-based responses, and engage with stimuli originally developed for human participants. Machine psychology, then, can be defined by three jointly necessary features. First, it studies LLM behavior at the input-output level through controlled experiments. Second, it draws on theories, paradigms, or methodological principles from psychology and the behavioral sciences to design those experiments and interpret their results. Third, its primary goal is to understand how and why LLMs behave as they do, rather than to optimize or rank model performance.

Directly inspecting LLMs' behavior does come with multiple advantages. LLM behavior unfolds at the interface where human users interact with the model, making it the level of analysis most directly relevant to deployment and safety (*6*–*8*). Moreover, observable behavior is often too complex to predict from current mechanistic understanding of model weights and activations alone. At the same time, the relationship between behavioral and mechanistic analysis runs in both directions. Behavioral experiments can identify phenomena that call for mechanistic explanation, while mechanistic findings can generate predictions that behavioral tests then confirm or refute (*9*). Only behavioral experiments, however, can be easily applied by the broader academic community to closed-source state-of-the-art models whose internal workings are not disclosed.

Here, we review directions in this emerging field of modeling and evaluating LLM behavior. We outline how established behavioral sciences can guide and inform our understanding of LLMs and discuss important caveats for when and how to apply these methods

[1]University of Stuttgart, Stuttgart, Germany. [2]Google DeepMind, London, United Kingdom. [3]Helmholtz Institute for Human-Centered AI, Munich, Germany. [4]TU Munich, Munich, Germany.
*Corresponding author. Email: thilo.hagendorff@iris.uni-stuttgart.de

Downloaded from https://www.science.org on November 03, 2026

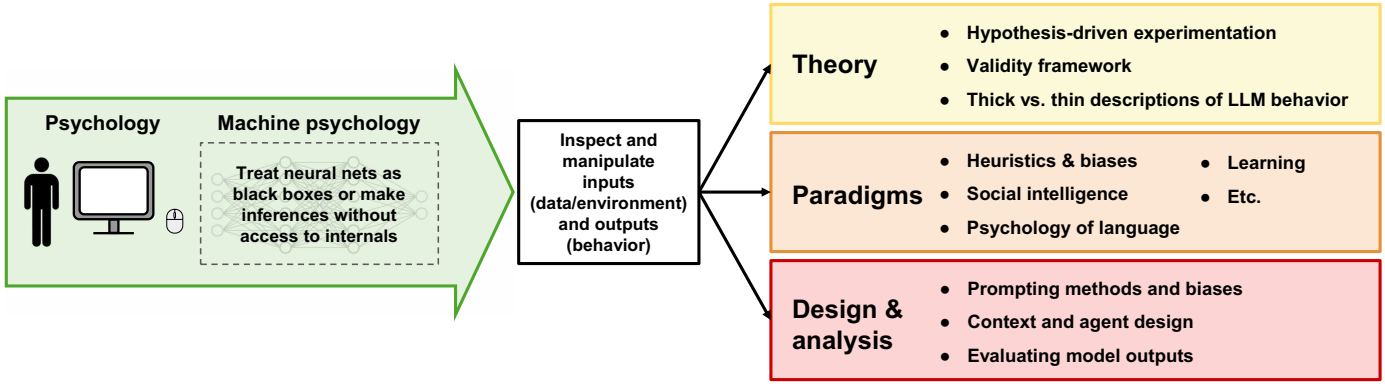


Fig. 1. Key concepts. Overview of the main components of machine psychology.

to LLMs, given that they were originally developed for humans. In the first section, we discuss the theoretical frameworks developed and used in psychology to organize our understanding of intelligence and intelligent behaviors. We then review the many empirical paradigms that have been developed to study and characterize different aspects of intelligent behavior in LLMs. Last, we discuss and make recommendations for robust empirical methods both for designing experiments and analyzing behavioral data. We end the article by discussing potentials and limitations of conducting machine psychology experiments with increasingly capable black box models.

THEORY: EVALUATION PARADIGMS FOR UNDERSTANDING INTELLIGENT SYSTEMS

A central question for machine psychology is whether frameworks developed to study human cognition can meaningfully be applied to artificial systems. Psychological experiments were originally designed to relate observable behavior in human participants to theoretical constructs such as reasoning strategies, memory processes, or learning mechanisms. When studying LLMs, however, these mappings cannot be assumed a priori. Behavioral patterns observed in models may arise from fundamentally different computational mechanisms than those underlying similar behaviors in humans. For this reason, machine psychology should not be understood as presupposing human psychological constructs in models but as adopting the methodological logic of psychology: designing experiments that test hypotheses about computational regularities. That said, the appeal of psychology—or of adopting psychological concepts—comes from the fact that it offers a rich vocabulary to describe cognition, reasoning, memory, and other phenomena that plausibly transfer to LLMs, not least because these systems are trained on human-generated data and optimized to produce human-like behavior. Much work in machine psychology will therefore involve ascribing such constructs to models. The position we advocate requires that these ascriptions rely on empirical evidence rather than metaphorical parallels suggesting that the mechanisms underlying human and machine intelligence and cognition are identical.

Next to behavioral model evaluations stands mechanistic interpretability. Borrowing a distinction from ethnography (*10*, *11*), one might describe mechanistic accounts of LLM behavior as "thin descriptions": They characterize internal representations or features, but they tend to not explain the behavioral regularities those representations produce. Machine psychology aims instead at "thick descriptions" that add a layer of explanatory structure that matches observable behavioral patterns and theoretical constructs while accepting the opacity of LLMs. This opacity holds true for the human brain as well but for a different reason. LLM opacity comes from complexity, not from limits on observing or changing weights and activations. Brain opacity comes from technical and ethical limits on accessing the brain—limits that do not apply to LLMs. So mechanistic interpretability can support or refute psychological claims about LLMs, much as neuroscience has shaped psychological theories about humans. In any case, applying psychological constructs to opaque structures like brains or LLMs increases explanatory resources (*12*). Thick descriptions help researchers make sense of LLM behavior when thin descriptions alone cannot account for observed patterns and regularities that go beyond what predictions from token-level statistics would yield.

Behavioral analyses of highly capable LLMs need test frameworks that deviate from how traditional machine learning algorithms were benchmarked (*13*–*16*). Here, datasets were designed to require specific capabilities (e.g., object recognition, sentiment analysis, etc.) for good performance. Researchers train on a train dataset and evaluate on a held-out test dataset that was not seen during training. This framework does not generalize well to large-scale foundation LLMs for two reasons.

First, when training on internet-scale corpora, models face pervasive data contamination: Benchmark test items can appear in training data, inflating performance estimates and undermining the validity of evaluations (*17*). Detecting such contamination is itself nontrivial, especially for closed-source LLMs whose training data remain undisclosed (*18*, *19*). For machine psychology, this problem carries an additional layer of concern. When an LLM produces human-like responses on a classic cognitive task, contamination makes it difficult to distinguish strategies that generalize beyond the training data from memorized answer patterns.

Second, LLMs are trained for next-token prediction but exhibit many other "intelligent" behaviors that some researchers consider emergent (*20*). This characterization remains debated, with others arguing that apparent emergence reflects metric choice rather than fundamental changes in model behavior with scale (*21*). Leaving this debate aside, examples for "emergent" or at least unplanned and unpredicted behavior are ample. For instance, researchers did not explicitly encode or train for a transformer LLM's ability to learn from examples in context (*22*), but it nonetheless arose from the machine

Downloaded from https://www.science.org on November 03, 2026

learning architecture, data, and learning signal (*23*, *24*). To name another salient example, LLMs can perform out-of-context reasoning, latently combining disparate facts or subtle signals encountered during training to draw inferences never explicitly stated in any single document (*25*, *26*). Eventually, emergent behaviors can be difficult to study through the lens of the components that gave rise to them (*27*), and the ones that emerge can seem unexpected (*20*)—the most interesting evaluations are not "held-out" exemplars of the training task (*28*).

Researchers have therefore started building test-only benchmarks—i.e., smaller-scale datasets unsuitable for training and intended solely as a test set—to investigate model capabilities. In several cases, these benchmarks already resemble evaluation frameworks from the behavioral sciences (*29*)—like personality tests, intelligence tests, implicit association tests, etc. that are applied to humans—which similarly do not follow the train-test paradigm. They also tend to fall into two categories. Some evaluations focus on scalar performance metrics, others on characterizing behavior, i.e., the questions are designed to elicit responses that reveal behavioral strategies or underlying constructs. In practice, these two goals often overlap: Systematic variation in performance across task conditions can itself reveal cognitive strategies and mechanisms. Machine psychology draws on both traditions, but its primary orientation is toward understanding how and why LLMs behave as they do, rather than toward optimizing or ranking model performance on a fixed metric. In this review, we focus on test-only evaluations that provide this kind of understanding, as a paradigm that is gaining traction in the research landscape.

Several such diagnostic evaluations have been developed even for pre-LLM models where, despite the models being trained for specific tasks, how to solve them is not specified. Such diagnostic datasets were used to expose the ways in which learned systems solved tasks—often counter to human intuitions (*30*–*34*). Researchers have also made the case for borrowing from ethology, a branch of zoology that studies the behavior of nonhuman animals, to explain machine behavior in AI systems (*35*). Animal abilities are also not directly known, and applying a double standard when comparing human, animal, and machine cognition is a well-documented risk (*36*). Yet LLMs differ from both pre-LLM models and animal subjects in a specific respect: For task-specific models, the domain of interest is fixed by design, and for animals, ethologists draw on extensive observation histories that chart species-typical behavioral repertoires. LLMs, by contrast, acquire capabilities across a wide and still-expanding set of domains from a single training objective (next-token prediction), making the space of abilities harder to delineate in advance. Furthermore, since LLMs can be evaluated via natural language, researchers only recently gained access to evaluation methods that go beyond what previous machine learning benchmarks could provide. This has led to the widespread adoption of language-based diagnostic evaluations.

However, this comes with important caveats. Because LLMs can produce fluent natural language, researchers may be tempted to simply ask a model to report on its own reasoning processes or internal states. Such self-report measures have a long history in psychology, but their reliability is questionable in humans (*37*), and similar concerns apply to LLMs (*38*–*41*). Properties that such measures usually consider like personality, morality, or clinical disorders are famously sensitive to prompting (*42*, *43*), to the extent that several recent works even simulate groups of humans of different social groups, opinions, and personalities with differently prompted LLMs (*44*–*47*). There remains value in using self-report stimuli from psychology, but results drawn from these measures should be taken contextually instead of as a fundamental or general property of the LLM itself.

Decades of methodological work in the behavioral sciences have organized the criteria for sound experimentation. In particular, experiments should be hypothesis driven (*48*): Researchers formulate competing hypotheses about the representations or constructs that underlie behavior and then design experiments whose outcomes discriminate between them. If the model or agent has representation X, we expect behavior Y; otherwise, we expect behavior Z. Moreover, researchers must attend to construct validity (*36*, *49*, *50*): Does the task actually measure the construct it claims to measure when applied to LLMs? A task originally designed to probe, say, theory of mind in children, relies on assumptions about how behavior connects to underlying cognitive representations. Those assumptions do not transfer automatically to LLMs, which may pass the same task potentially through shallow pattern matching. Establishing construct validity requires careful controls that rule out such alternative explanations (*51*, *52*) and converging evidence from multiple operationalizations of the same construct. Furthermore, researchers must consider the performance-competence distinction (*53*): The absence of a target behavior does not necessarily indicate the absence of the underlying capability. Human performance can be noisy or biased; people make arithmetic errors and produce ungrammatical language without lacking the abstract capacity for math or grammar. Similar considerations apply to LLMs (*54*–*57*), whose performance on any given task can be shaped by prompt format or other incidental factors that may mask genuine capabilities (*58*). In sum, such principles form the methodological foundation on which the experimental paradigms reviewed in the next section are based.

PARADIGMS: THE MANY ASPECTS OF INTELLIGENT BEHAVIOR

There are many aspects of intelligent behavior, which have each been studied by different subfields of the behavioral sciences. Each of these has developed domain-specific empirical paradigms. While some of these subfields (e.g., motor learning) and paradigms (e.g., pupillometry) are not directly transferable to LLMs since they rely on the existence of a physical body, several of these paradigms are purely linguistic and can be easily transferred. As LLMs expand in the kinds of stimuli they can interpret—e.g., visual (*59*, *60*)—and the ways in which they can interact with the world—e.g., embodiment and tool use (*61*)—the space of transferable paradigms increases. One caveat is that whenever nonlinguistic stimuli are translated into text or simplified symbolic representations for LLMs, this translation process is not neutral but alters task difficulty and shifts the cognitive demands of a task. Humans interact with several modalities, and the paradigms developed to understand them often compare and integrate these modalities (*62*)—e.g., the Stroop test which spans vision and reading capabilities (*63*, *64*).

In general, applying psychological paradigms to LLMs raises a validity question that does not arise when applying them to humans: When an LLM, e.g., exhibits human-like behavior on a task designed for human test participants, does this reflect a strategy that generalizes beyond the model's training data, or does it simply reproduce human behavioral patterns absorbed from training data? Both options to answer this have scientific value. Even if a human-like bias in an LLM originates in patterns in the training corpus rather than in a shared cognitive mechanism, documenting that bias remains informative: It characterizes how the model will behave in

Downloaded from https://www.science.org on November 03, 2026

deployment, what errors users will encounter, where LLM applications need potential safeguards, etc. On the other hand, a likewise productive approach derives predictions from the LLM's own training objective and tests whether behavior conforms to those predictions rather than to human patterns. For example, researchers show that LLM reasoning performance varies systematically with the probability of the task, the input, and the target output in the training distribution (*65*). These patterns arise from the pressure of next-word prediction rather than human cognitive constraints. Ultimately, machine psychology research can combine both orientations: It uses tasks inspired by psychology to document behavioral patterns while incorporating controls that can distinguish behavior driven by model architectures from training data mimicry.

In the following, we focus on example machine psychology studies, in particular on language-based tests, since these are the most widely used in the current research landscape. Moreover, we believe that even in light of the growing trend toward multimodal models, language will remain a primary modality due to its fundamental role in models' reasoning processes. We concentrate on four research areas that can inform distinct strands in machine psychology research: heuristics and biases, social interactions, the psychology of language, and learning.

Our coverage is selective and illustrative: For each area, we highlight representative studies that demonstrate the value of psychology-inspired approaches, rather than attempting an exhaustive survey. The volume of work at the intersection of psychology and LLMs has grown rapidly, and comprehensive reviews exist (*66*). The four areas we selected connect with complementary facets of (human) intelligent behavior. Heuristics and biases address individual reasoning and decision-making, meaning how LLMs form judgments under uncertainty and whether they rely on systematic shortcuts. Social intelligence studies how LLMs represent the mental states of other agents and navigate interactive settings. The psychology of language examines how LLMs process the structural and pragmatic dimensions of their primary medium, from syntax to implicature. Learning concerns adaptation, meaning how LLMs acquire new behaviors from examples or instructions provided at inference time. Together, these four areas span core dimensions of intelligent behavior studied in the behavioral sciences.

Our selection of the four areas also illustrates the distinctive contribution of psychology-inspired approaches relative to adjacent evaluation methods. The distinction we want to draw is between three logics of measuring model outputs, namely, by looking at performance via standard benchmarks (MMLU, GPQA, ARC-AGI, SWE-bench, etc.), by investigating a model's inner workings via mechanistic interpretability, and by measuring behavioral patterns via machine psychology. Standard benchmarks quantify what an LLM can do by aggregating accuracy on held-out items; the resulting score describes proficiency on a task. Mechanistic interpretability investigates where and how computations are implemented within a model's internal representations. Machine psychology occupies a different level of analysis. It uses controlled experimental manipulations to test hypotheses about why LLMs produce the behavior they do and which strategies underlie their responses. Each of the four areas reviewed below provides representative examples of this approach.

Heuristics and biases

The heuristics and biases framework is one of the most influential research paradigms in psychology (*67*, *68*). Heuristics are mental shortcuts that simplify reasoning or decision-making processes, and this field studies how such shortcuts can help explain both the successes and the biases in human behavior. The large existing literature on heuristics and biases in humans is a fertile ground for examining such shortcuts in LLMs—whose capabilities now overlap more with the human abilities this literature studies. Binz and Schulz (*6*) were among the first to use this paradigm to better understand the decision-making processes of LLMs. They found that GPT-3 produces response patterns matching those that, in humans, are attributed to cognitive biases. Several other works have also been made in this vein (*62*, *69*–*75*). There is evidence from several studies showing that, while early generations of models frequently exhibited behavior resembling human-like heuristics and biases, they have disappeared in later generations of LLMs (*71*, *76*). The test stimuli were originally designed to be challenging for human study participants and no longer challenge the growing reasoning abilities in LLMs. This could also be due to leakage into the training set—we discuss this challenge in the section on design and analysis.

The heuristics and biases literature furthermore suggests that how a problem is phrased can influence how people solve it (*77*, *78*). It is well known that LLMs are also susceptible to similar manipulations. For example, Dasgupta *et al.* have investigated whether LLMs are affected by the semantic content of logical reasoning problems using several existing tasks from the literature (*32*). They found that, like people, LLMs reason more accurately about familiar, believable, or grounded situations, compared to unfamiliar, unbelievable, or abstract problems. Likewise, researchers showed that LLMs exhibit an optimism bias during in-context learning, updating more from better-than-expected outcomes than from worse-than-expected ones, which illustrates that the framing and structure of a problem can shape the in-context learning dynamics of LLMs in cognitively meaningful ways (*79*).

Last, humans do not simply apply arbitrary heuristics. Instead, they use heuristics that are adapted to the problems they encounter during their everyday interactions with the world (*80*). In the context of LLMs, one can look at how the properties of the training data shape their behavior. For example, researchers have demonstrated that the presence of in-context learning in LLMs can be traced back to data distributional properties such as burstiness where items appear in clusters rather than being uniformly distributed over time and having large numbers of rarely occurring classes (*23*).

Social intelligence

Traditionally, developmental psychology explores how humans develop cognitively, socially, and emotionally throughout their lives. This includes studying the various factors that influence development, such as social intelligence or social skills. By applying paradigms from this area of developmental psychology to LLMs, researchers can gain deeper insights into how these models manage complex social interactions. In particular, once LLMs are deployed as chat agents, they should become versed in modeling human communicators. Therefore, it is important to assess the level of social intelligence in LLMs. One example in this context is the application of theory of mind tests to LLMs, where researchers use tasks from human experiments, such as those famously conducted by Perner *et al.* (*81*, *82*). While early experiments with models like GPT-3 showed that they struggle to solve theory of mind tasks (*83*), later models demonstrate an increasing ability to reliably infer unobservable mental states in others (*84*–*88*). Further related research examines how LLM performance on theory

Downloaded from https://www.science.org on November 03, 2026

of mind tests compares to that of children (*89*), LLM ability to handle higher-order theory of mind tasks requiring recursive reasoning about multiple mental states (*90*), or measures the robustness of theory of mind test setups against distracting alterations in the tasks LLMs receive as inputs (*52*). As theory of mind tests measure, among other things, the ability to understand false beliefs, further research has explored the emerging capability of LLMs to induce false beliefs in other agents (*88*), or how LLMs trade off various communicative values like honesty and helpfulness (*91*)—these investigations also contribute to understanding and improving alignment with human values for AI safety (*92*).

The space of relevant paradigms increases as LLMs are allowed to interact through self-reflection (*93*), self-instruction (*94*), or in swarms (*95*). For example, researchers looked at cooperative and coordinative behavior in LLMs playing games, revealing persistent behavioral signatures in the models (*96*). Similarly, researchers investigated cooperative or competitive LLM behavior in psychology-inspired dilemma situations to assess the ability of LLMs to participate in real-world negotiations (*97*). Another study, which is influenced by works in human social psychology, looked at how multiple LLMs form and evolve networks, investigating microlevel network principles like preferential attachment or triadic closure, as well as macrolevel principles such as community structures (*98*). Other studies developed competition-based benchmarks that evaluate LLM agents on cognition, adaptability, rationality, and collaboration in social deduction games and game-theoretic scenarios (*99*). These interactive paradigms complement theory of mind tests by assessing whether LLMs can translate social understanding into effective real-time behavior within multiagent settings (*100*). In sum, machine psychology can reveal patterns of social behavior and interaction among LLMs, individually and collectively, be it for problem solving or world simulation (*101*). By drawing from human developmental psychology and social dynamics, researchers can better understand and design LLMs that navigate complex social interactions and exhibit advanced social skills.

Beyond these fundamental assessments of social cognition, a complementary line of work evaluates LLMs as active participants in dynamic social interactions. Researchers use open-ended environments where LLM-based agents pursue individual social goals through multiturn natural language interaction, enabling them to measure social intelligence across dimensions such as goal completion, relationship maintenance, and adherence to social norms (*102*).

Psychology of language

A long history of work has studied the psychology of how humans use and understand language, ranging from how they use semantic and syntactic features to understand a sentence to how they use pragmatic inferences in a discourse context to help interpret what someone has said. Correspondingly, a long-standing body of work has studied how language processing models capture these features of human language processing. Early connectionist works studied these topics in simple recurrent predictive models (*103*, *104*); more recently, researchers have applied similar techniques to study LLMs. A wide range of work has studied what models learn about syntax (*105*), often using methods from psycholinguistics. For example, researchers used psycholinguistics-inspired surprisal measures to show that LLMs learn filler-gap dependencies, a challenging syntactic structure (*106*). Other researchers have used related measures to study what LLMs learn about the semantics of entailment (*107*). Moreover, researchers used psycholinguistic techniques like priming to study how models represent and process language (*108*, *109*) and methods like deconfounded stimuli to identify where models may rely on semantic heuristics rather than syntax (*34*). Several recent works (*110*, *111*) studied pragmatic judgments of LLMs and found that larger models, as well as those with instruction tuning, tend to better approximate human responses and error patterns—although some deficiencies remain. In another study, researchers examined long-form analogies generated by ChatGPT, finding that AI-generated analogies lack some human-like psycholinguistic properties (*112*), particularly in text cohesion, language, and readability. Furthermore, researchers applied garden path sentences—sentences that lead the reader to initially interpret them incorrectly due to their ambiguous structure—to LLMs, showing that the models respond similarly to humans (*113*, *114*). At a higher level, some researchers have drawn inspiration from aspects of human language development to attempt to identify the causes of the relative data inefficiency of language models (*115*, *116*). In each of these cases, methods and ideas from psychology and psycholinguistics provide guidance on how to assess processes through language behaviors in LLMs, potentially by drawing comparisons between LLMs and humans.

Learning

The psychology of learning is concerned with how individuals acquire and retain knowledge and skills. At first blush, it may appear that experimental paradigms for the study of learning are less applicable to LLMs, given that the aim of behavioral experiments is often to help uncover the underlying learning algorithm—whereas for LLMs, the learning algorithms used in training are designed and already known. However, the behavioral sciences can still benefit the study of LLMs in this context, since LLMs exhibit learning abilities that were not explicitly designed into the models, and thus one does not understand the underlying learning algorithm. In particular, LLMs exhibit emergent in-context learning—the ability to learn from context (the prompt) without requiring any gradient-based updates in weights (*22*). Understanding in-context learning is a growing field that is rapidly gaining in importance, given the increasing size of LLMs' context windows and consequent gains in capabilities, e.g., the capability to learn an entire language from context alone (*117*) or the ability to overcome safety fine-tuning (*118*, *119*).

Uncovering the implicit learning algorithm implemented by in-context learning is a burgeoning research field and uses many of the methods common in cognitive science. For example, multiple studies have compared the outputs of transformer in-context learning with the outputs of hypothesized learning algorithms (*24*, *120*). This is a staple of cognitive modeling and could potentially benefit even further from model comparison procedures from psychology and statistics (*121*–*123*). Recent work in cognitive science has used machine learning to discover new theories of human decision-making (*124*)—it might be interesting to apply related approaches to in-context learning as well. Researchers might also benefit from considering particular models as normative starting points (*125*).

Researchers may also wish to understand other interesting and important characteristics of learning, such as inductive biases and generalization, the data dependence of learning, and the dynamics of learning over time. These characteristics are often not obvious even in cases where the learning algorithm is known, and thus researchers would like to understand them not just for in-context learning but also for other forms of LLM learning, e.g., self-supervised gradient-based

Downloaded from https://www.science.org on November 03, 2026

learning, reinforcement learning (*126*), or "fast" memory retrieval (*127*, *128*).

To characterize inductive biases and generalization of LLMs, researchers have borrowed both concepts and experimental paradigms from cognitive sciences (*79*, *129*) and Bayesian inference (*130*). Studies used paradigms for measuring systematic generalization to characterize those capabilities in LLMs and as inspiration for improving those abilities (*131*, *132*). Researchers created variants of classic analogy problems from cognitive science, to examine the analogical capabilities of LLMs (*133*). Others have borrowed ideas and experimental paradigms on "rule-based" versus "exemplar-based" generalization to characterize the inductive biases of in-weights versus in-context learning in transformers (*134*). Furthermore, researchers borrowed paradigms and measures from developmental psychology to characterize the domains where LLM inductive biases may match those of children, and where they may fall short (including in causal reasoning and innovation) (*135*, *136*).

To characterize the data dependence of in-context learning, studies have drawn inspiration from research in developmental psychology on skewed and bursty distributions (*23*). An important aspect of data dependence is the structure of data over time (during training). Researchers have long drawn inspiration from curriculum learning in human and nonhuman animals to better understand how to structure training data so that earlier learning on easier tasks can scaffold later learning on harder tasks (*137*). There remain many areas of behavioral research on learning that may serve as rich sources of inspiration on data dependence, e.g., research on repetition and spacing (*138*), working memory (*139*, *140*), blocking versus interleaving tasks (*141*), and continual learning (*142*). Data dependence is particularly interesting for LLMs because text training data (being sourced largely from unstructured web-scale corpora) is very different from the structured training data typically used for traditional discriminative machine learning techniques, and because data are one of the major levers one can manipulate in training LLMs to adjust their behaviors.

DESIGN AND ANALYSIS: GOOD BEHAVIORAL EXPERIMENTATION

Psychological experimentation is grappling with the challenge of inferring latent cognitive processes from observable behavior. Because constructs such as reasoning strategies, memory processes, or cognitive biases cannot be measured directly, psychological methodology emphasizes a set of principles that support reliable inference from behavioral data. These include construct validity, internal validity, external validity, or reliability, which together determine whether observed behavioral patterns can be causally attributed to experimental manipulations and generalized beyond a specific experimental setup (*143*, *144*).

All these methodological concerns become salient when adapting psychological paradigms to study LLMs. For example, behavioral effects observed in models may reflect prompt sensitivity, artifacts of the evaluation setup, or training-data contamination rather than the behavioral processes an experiment was intended to probe. Similarly, results obtained with a specific model or prompting technique may fail to generalize across architectures (e.g., dense transformers versus mixture-of-experts models) or training regimes (e.g., supervised fine-tuning versus synthetic document fine-tuning). As a result, applying psychological tools to LLMs requires careful attention not only to experimental design but also to how behavioral measures are operationalized, how potential confounds are controlled, and how findings generalize across models and evaluation settings (*145*). Current LLMs are famously sensitive to small changes in prompt structure or they rely on shallow syntactic heuristics (*34*) and studies that are not careful about testing the robustness of their conclusions risk being spurious and nongeneralizable. Psychology too has had its own share of reproducibility crises (*146*, *147*), and machine psychology should not share the same fate (*148*). In this section, we provide recommendations for sound methodologies in behavioral test settings with LLMs, which should be valuable to practitioners in the field of machine psychology.

Prompting methods and biases

Many of the studies already conducted in the field of machine psychology have a notable shortcoming in common, namely, that they do not avoid training data contaminations, one of the central validity threats. When LLMs reproduce behavioral patterns previously observed in humans, this may reflect training data memorization, familiarity with similar task templates, or actually shared cognitive mechanisms. Distinguishing among these possibilities is theoretically important, because each implies a different account of why human-LLM similarities arise. Accordingly, contamination should not be framed only as something to eliminate but also as something to characterize. Stronger machine psychology designs test whether an effect survives substantial variation in wording or task orders, use a model's training cutoff, rely on dynamic multiturn conversations, canary strings, use private held-out benchmarks, or overlap analyses between benchmarks and training data (when accessible), and hence test whether certain capabilities go beyond mere memorization. Behavioral convergence is most informative when it persists across such variations.

Although the presence of training data influences does not render behavioral findings uninformative if they represent mere memorization. The goal of machine psychology is not necessarily to determine whether a behavior reflects abstract generalization beyond the training distribution but simply to discover that the behavior occurs at all. This is particularly relevant for safety-critical behaviors. Even if such behaviors arise because models reproduce patterns present in their training data, documenting their existence remains valuable.

Another common shortcoming of several existing machine psychology studies is that they rely on small sample sizes or even convenience samples, meaning nonsystematic sequences of prompts. Sampling biases in the used benchmarks or task datasets can diminish the quality of such studies. This is because slight changes in prompts can change model outputs significantly. Prompt sensitivity, though, has a close analog in human experimentation. Experimenters studying human participants discover through pilot testing that respondents react to incidental task features the experimenter did not intend to manipulate (*144*). Pilot studies serve the function of identifying and eliminating such sensitivities before the actual data collection begins. The methodological logic is the same in both cases: Researchers vary features of the experimental setup to ensure that observed behavior reflects the construct under study. What differs is how much and what kind of variation matters. LLMs can respond to formatting details that lack parallels in human testing. At the same time, prompt variations can be tested cheaply and at scale, making "piloting" more feasible with LLMs than with human participants. At the same time, it bears the risk of researchers becoming tempted to cherry-pick prompt and training setups that yield desired instead of representative experimental outcomes.

Downloaded from https://www.science.org on November 03, 2026

Furthermore, LLMs can succumb to various biases influencing the processing of prompts (*23*, *149*). Recency biases in LLMs, for instance, lead to a tendency to rely more heavily on information appearing toward the end of prompts. LLMs can also have a common token bias, meaning that models are biased toward outputting tokens that are common in their training data. Moreover, most of the label biases can cause LLMs to be skewed toward labels, classes, or examples that are frequent in a few-shot learning setting. Technical biases like these can at least in part be controlled for when designing prompts or prompt variations that tend to avoid triggering them. If this is not done, then LLMs may rely on shortcuts exploiting such biases.

A further concern specific to studies that compare LLM and human behavior is instructional asymmetry. Human participants in psychological experiments typically receive task explanations or practice trials. LLMs receive a bare prompt. This asymmetry might confound comparisons as differences in behavior instead of a potential gap in instructional "richness." Researchers should therefore consider matching instructions across human and LLM participants, providing both with equivalent context, examples, and task framing.

A practical question remains: Given limited resources, how should researchers decide which of these practices and controls to prioritize, and when have they done "enough"? Psychology faces the same problem. No human experiment controls for every conceivable confound. Researchers prioritize on the basis of what their hypothesis predicts should matter. A useful heuristic for machine psychology would be to first vary the design choices most directly relevant to their specific claim. For instance, a study claiming that LLMs exhibit a particular cognitive bias should, at minimum, test whether the effect survives variation in prompt wording and answer-option ordering, because these are the most plausible confounds for that type of claim. Other variations, such as testing across model sizes, temperature settings, or scoring methods, strengthen the finding but may not be necessary. The goal is a defensible argument that the finding is robust to the most plausible alternative explanations, combined with transparent disclosure of which variations remain untested.

Context, scaffolding, and their role in behavioral evaluation

An LLM's observed behavior depends on the full context it receives at inference time. Context encompasses system instructions, task framing, conversation history, tool definitions, retrieved documents, memory structures, skill files, etc. The practice of curating these components for optimal performance has become known as "context engineering," reflecting a shift from crafting isolated prompts toward designing entire information pipelines (*150*). For machine psychology, this shift carries methodological consequences that researchers must address. For instance, the performance-competence distinction applies here (*54*). A model that fails a reasoning task under a single-turn prompt may succeed when being scaffolded by retrieval, self-reflection, code execution, or other tools. Omitting such contextual supports can lead researchers to underestimate model capacities. Researchers should report the full evaluation context and test whether behavioral effects survive changes to that context. This issue becomes acute with LLM-based agents, where models operate in multiturn loops with tool access, external memory, or self-correction mechanisms (*151*). In such systems, the unit of study is not a single model. It is the whole setup: Model, scaffolding, tools, and environment together produce the observed behavior. Recent work on agentic evaluations has emphasized that scaffolding variations alone can shift benchmark scores by several percentage points (*152*). For machine psychology, this means that claims about "a model" must specify which configuration produced the behavior.

Regarding prompt augmentation techniques in particular, researchers should approach them as experimental variables. Chain-of-thought prompting, self-consistency, multiple-choice versus open-ended formats, or agentic scaffoldings form different evaluation conditions. Varying them reveals how a model's behavior depends on the structure of its input, which is itself a finding about the model's processing. For instance, if a model exhibits a human-like bias only when forced to respond immediately but corrects it when given chain-of-thought abilities, this dissociation tells us something about the conditions under which the bias arises.

Evaluating model outputs

After conducting experiments, a list of LLM responses must be evaluated, for instance, by comparing them with the ground truth. The simplest case is when the results can be framed and scored as a multiple-choice question—although even in this case, scoring the answers so that the model responds directly inline, rather than selecting a choice, can yield more signal (*153*). If possible, multiple scoring methods should be compared, to evaluate whether the effects are dependent on the scoring method (*154*). Still, in multiple-choice settings, raw token probabilities favor common surface forms regardless of question content. Domain-Conditional Pointwise Mutual Information (*155*) corrects for this by normalizing each answer's probability against its probability given a neutral context. This approach directly addresses common token bias and majority label bias, two of the prompt-sensitivity problems discussed earlier.

Beyond multiple-choice settings, if a benchmark task must be answered with free generations, then the evaluation process can be automated as well. If the results exhibit sufficient simplicity and regularity, meaning that the LLM responses are similar to the ground truth strings in terms of length and wording, methods such as testing word overlaps with regular expressions or using metrics like the F1 score can be used. LLMs, however, tend to produce highly variable and comprehensive outputs, which can complicate classification. While stop sequences or prompt instructions that interrupt further text generation can facilitate classification by promoting output uniformity, they also improperly constrain LLM behavior.

Therefore, researchers are increasingly relying on LLM-based evaluations of outputs where a single or multiple models perform the classification (*156*). Although this method might still be inaccurate for very comprehensive outputs, a solution is to instruct the LLM judge to output its final answer after a specific string sequence or format it as a JSON object. This approach allows the judge to reason during prompt completion, whereas the classification only involves processing the final response string.

To further improve the quality of studies, researchers can preregister studies before collecting data by reporting the full evaluation configuration (models, prompts, training setups, agent scaffoldings, parameters, output classification methods, etc.) (*157*) and release stimulus sets and evaluation code to enable direct replication. Machine psychology will produce cumulative knowledge only if individual studies are designed to be scrutinized and built on—standards that the field can import from the open-science reforms that reshaped experimental psychology over the past decade (*143*, *146*).

Downloaded from https://www.science.org on November 03, 2026

MACHINE PSYCHOLOGY OF INCREASINGLY CAPABLE MODELS

The rapid evolution of generative AI systems over the past few years points toward developments that will radically reshape how humans interact with intelligent machines in the future. Several trends are of particular relevance for machine psychology, among them large reasoning models (LRMs) resulting from reinforcement learning–based training.

LRMs generate extended chains of intermediate reasoning tokens before producing a final response. A standard LLM generates each token conditioned on all preceding tokens, performing a limited form of sequential inference at each step. RL-based training, by contrast, can optimize the entire chain of generated tokens toward producing correct final answers, teaching the model to decompose complex problems into extended reasoning traces (*158*). Reinforcement learning has yielded superhuman performance in constrained strategic domains such as the game of Go (*159*), and recent LRMs suggest that similar training regimes can improve multistep reasoning in less constrained settings. Whether such incremental improvements will eventually cumulate into general-purpose (super-) intelligence remains an open and contested question. Even short of that threshold, the steady expansion of reasoning capabilities makes it increasingly important for machine psychology to develop methods that can characterize strategies which may diverge from or surpass familiar human heuristics or cognitive abilities.

Behavioral methods are well suited to this task. Emerging reasoning strategies within LRMs' multistep inferences remain difficult to characterize through mechanistic interpretability alone: Recent tools can study individual reasoning steps (*9*), but the length and variability of reasoning traces taken as a whole still exceed what many interpretability interventions can handle. Machine psychology, by contrast, can apply the hypothesis-driven experimental logic described earlier: If a model has learned strategy X, then we predict behavioral pattern Y under condition Z. Any investigation of reasoning traces that meets these criteria (controlled experimentation at the input-output level, grounding in theories from the behavioral sciences, and the goal of understanding rather than optimizing model behavior) falls within machine psychology as we define it. Several open research questions—whereas previous studies identified phenomena and machine psychology can contribute by subjecting these phenomena to the experimental paradigms described earlier—are good candidates for this approach. Outcome reward and process reward models (*160*) may give rise to different behavioral strategies; supervised fine-tuning and RL-based training (*161*) may produce different behavioral profiles on the same tasks; LLMs appear capable of out-of-context reasoning that latently combines facts from training data (*25*) or even introspection (*26*); and extended thinking token generation introduces safety-critical dynamics into model behavior (*162*). Analyzing reasoning traces also raises questions about faithfulness: whether the visible "thinking" tokens reflect the computations that actually drive the model's final output, or whether they constitute post hoc rationalizations (*38*). Here, construct validity principles apply directly. Researchers cannot take self-reported reasoning at face value. They need behavioral tests that show whether the reasoning is faithful or just a plausible-sounding "story" that has little to do with the actual output (*163*). Furthermore, comparing cognitive strategies across training regimes (e.g., RL versus supervised fine-tuning) can reveal which behaviors generalize beyond the training distribution and which are artifacts of a particular optimization pressure, connecting to the contamination and generalization concerns raised in earlier sections.

The methodological challenges grow more acute with LRM-based agents. As discussed earlier, the behavior under study in such systems has many aspects, ranging from model, scaffolding, tools, and environment that jointly produce the observed behavior. When different agent instances interact, they can display group dynamics, cooperative or goal-seeking behavior, and cultural evolution (*164*, *165*). Machine psychology provides a framework for studying these emergent multiagent phenomena at a level of description that mechanistic analyses of individual model weights cannot reach, since higher-level behavioral states can be instantiated by numerous different neural configurations. Moreover, this connects to AI safety and alignment concerns. Phenomena such as deception, scheming, sycophancy, sandbagging, or alignment faking (*88*, *166*–*168*) become increasingly plausible as systems grow more autonomous and pursue longer-horizon goals. By observing agent interactions under different environmental conditions, researchers can identify cases in which agents transition from predictable to unpredictable or undesirable behaviors.

DISCUSSION

Machine psychology provides a new approach to explaining AI. Instead of interpreting a neural network's design components (*169*), one analyzes the relationships between token inputs and outputs. Although this may allow the identification of hitherto unknown abilities or behavioral traits in LLMs, interpreting LLM responses comes with a challenge. A strong tendency exists to confer mental concepts or psychological terms to LLMs that were hitherto reserved for human and animal minds. This tendency manifests in common terms like "machine learning" but will become more prevalent in machine psychology when concepts such as reasoning (*170*), intuition (*71*), creativity (*171*), intelligence (*133*), personality (*172*), mental illnesses (*173*), etc. are transferred to LLMs. In this context, researchers have demanded caution by stressing that the underlying neural mechanisms for these concepts are different in humans and machines (*174*, *175*). Moreover, many psychological concepts are normatively laden and can foster mismatches in expectations between AI experts and the public regarding machine capabilities (*176*). As argued previously, the thin/thick distinction from ethnography (*10*, *11*) clarifies when such terminology transfer is reasonable. Psychological terms are worth using when they explain more than mechanistic accounts can and not when they just relabel existing descriptions without adding any predictive or explanatory value. This turns the debate about terminology into an empirical question: Does a given construct (e.g., "reasoning" and "personality") generate testable predictions about LLM behavior that token-level or architectural descriptions do not? It should also support counterfactuals by specifying how behavior should change under experimental manipulation, most importantly control studies.

Beyond potential habituations regarding the use of terminology borrowed from psychology in the context of machines, machine psychology aims to identify behavioral patterns, emergent abilities, and mechanisms of decision-making and reasoning in LLMs by treating them as participants in psychology experiments. This new discipline becomes even more important when taking multimodal or augmented LLMs into account, meaning LLMs that interact with images, external information sources, sensory data, physical objects, and tools (*61*, *177*, *178*). Moreover, once test settings for machine psychology are established, researchers can investigate how LLMs develop over time by applying the same tasks repeatedly, yielding longitudinal data that can serve as a baseline for extrapolating trends in reasoning

Downloaded from https://www.science.org on November 03, 2026

abilities. Such estimations may prove increasingly important for AI safety and alignment research (*179*).

Although this article focuses on LLMs, the core methodological principles we advocate are not necessarily limited to language-based systems. Three commitments transfer to other AI domains. First, hypothesis-driven experimentation is relevant when studying vision models, robotic agents, or generative image systems as well and can help formulate competing accounts of why a system behaves as it does and design tasks whose outcomes discriminate between them. Second, construct validity, meaning the question of whether a task measures what it claims to measure arises whenever a paradigm designed for one kind of agent is applied to another, regardless of modality. Third, input manipulation, meaning the systematic variation of stimuli to isolate behavioral strategies, applies to pixel-level perturbations just as it does to prompt wording choices, and the same logic extends to other modalities.

The specific experimental procedures discussed in this article will require adaptation for systems that do not operate through natural language. Diffusion models, for instance, receive conditioning signals rather than instructions, and robotic agents produce motor trajectories rather than text. Translating machine psychology to these domains will therefore require developing new experimental interfaces. We hope that this article provides a methodological foundation for that broader endeavor but argues that the most promising approach lies in a machine psychology for LLMs.

REFERENCES

1. R. Bommasani, D. A. Hudson, E. Adeli, R. Altman, S. Arora, S. von Arx, M. S. Bernstein, J. Bohg, A. Bosselut, E. Brunskill, E. Brynjolfsson, S. Buch, D. Card, R. Castellon, N. Chatterji, A. Chen, K. Creel, J. Q. Davis, D. Demszky, C. Donahue, M. Doumbouya, E. Durmus, S. Ermon, J. Etchemendy, K. Ethayarajh, L. Fei-Fei, C. Finn, T. Gale, L. Gillespie, K. Goel, N. Goodman, S. Grossman, N. Guha, T. Hashimoto, P. Henderson, J. Hewitt, D. E. Ho, J. Hong, K. Hsu, J. Huang, T. Icard, S. Jain, D. Jurafsky, P. Kalluri, S. Karamcheti, G. Keeling, F. Khani, O. Khattab, P. W. Koh, M. Krass, R. Krishna, R. Kuditipudi, A. Kumar, F. Ladhak, M. Lee, T. Lee, J. Leskovec, I. Levent, X. L. Li, X. Li, T. Ma, A. Malik, C. D. Manning, S. Mirchandani, E. Mitchell, Z. Munyikwa, S. Nair, A. Narayan, D. Narayanan, B. Newman, A. Nie, J. C. Niebles, H. Nilforoshan, J. Nyarko, G. Ogut, L. Orr, I. Papadimitriou, J. S. Park, C. Piech, E. Portelance, C. Potts, A. Raghunathan, R. Reich, H. Ren, F. Rong, Y. Roohani, C. Ruiz, J. Ryan, C. Ré, D. Sadigh, S. Sagawa, K. Santhanam, A. Shih, K. Srinivasan, A. Tamkin, R. Taori, A. W. Thomas, F. Tramèr, R. E. Wang, W. Wang, B. Wu, J. Wu, Y. Wu, S. M. Xie, M. Yasunaga, J. You, M. Zaharia, M. Zhang, T. Zhang, X. Zhang, Y. Zhang, L. Zheng, K. Zhou, P. Liang, On the opportunities and risks of foundation models. arXiv:2108.07258 [cs.LG] (2021).
2. L. Weidinger, J. Uesato, M. Rauh, C. Griffin, P.-S. Huang, J. Mellor, A. Glaese, M. Cheng, B. Balle, A. Kasirzadeh, C. Biles, S. Brown, Z. Kenton, W. Hawkins, T. Stepleton, A. Birhane, L. A. Hendricks, L. Rimell, W. Isaac, J. Haas, S. Legassick, G. Irving, I. Gabriel, "Taxonomy of risks posed by language models," in *2022 ACM Conference on Fairness Accountability and Transparency* (Association for Computing Machinery (ACM), New York, NY, USA, 2022), pp. 214–229.
3. T. Hagendorff, Mapping the ethics of generative AI: A comprehensive scoping review. arXiv:2402.08323 [cs.CY] (2024).
4. L. Brinkmann, F. Baumann, J.-F. Bonnefon, M. Derex, T. F. Müller, A.-M. Nussberger, A. Czaplicka, A. Acerbi, T. L. Griffiths, J. Henrich, J. Z. Leibo, R. McElreath, P.-Y. Oudeyer, J. Stray, I. Rahwan, Machine culture. *Nat. Hum. Behav.* **7**, 1855–1868 (2023).
5. M. Binz, E. Akata, M. Bethge, F. Brändle, F. Callaway, J. Coda-Forno, P. Dayan, C. Demircan, M. K. Eckstein, N. Éltető, T. L. Griffiths, S. Haridi, A. K. Jagadish, L. Ji-An, A. Kipnis, S. Kumar, T. Ludwig, M. Mathony, M. Mattar, A. Modirshanechi, S. S. Nath, J. C. Peterson, M. Rmus, E. M. Russek, T. Saanum, J. A. Schubert, L. M. Schulze Buschoff, N. Singhi, X. Sui, M. Thalmann, F. J. Theis, V. Truong, V. Udandarao, K. Voudouris, R. Wilson, K. Witte, S. Wu, D. U. Wulff, H. Xiong, E. Schulz, A foundation model to predict and capture human cognition. *Nature* **644**, 1002–1009 (2025).
6. M. Binz, E. Schulz, Using cognitive psychology to understand GPT-3. *Proc. Natl. Acad. Sci. U. S. A.* **120**, e2218523120 (2023).
7. T. A. Chang, B. K. Bergen, Language model behavior: A comprehensive survey. *Comput. Linguist.* **50**, 293–350 (2024).
8. A. Ivanova, Running cognitive evaluations on large language models: The do's and the don'ts. arXiv:2312.01276 [cs.AI] (2023).
9. P. C. Bogdan, U. Macar, N. Nanda, A. Conmy, Thought anchors: Which LLM reasoning steps matter? arXiv:2506.19143 [cs.LG] (2025).
10. G. Ryle, Collected papers (Routledge, 2009).
11. C. Geertz, *The Interpretation of Cultures: Selected Essays* (Basic Books, 2017).
12. W. Sellars, *Empiricism and the Philosophy of Mind* (Harvard Univ. Press, 1997).
13. S. R. Bowman, G. Angeli, C. Potts, C. D. Manning, "A large annotated corpus for learning natural language inference," in *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing* (Association for Computational Linguistics (ACL), Stroudsburg, PA, USA, 2015), pp. 632–642.
14. O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. S. Bernstein, A. C. Berg, L. Fei-Fei, ImageNet large scale visual recognition challenge. *Int. J. Comput. Vis.* **115**, 211–252 (2015).
15. I. D. Raji, E. M. Bender, A. Paullada, E. Denton, A. Hanna, AI and the everything in the whole wide world benchmark. arXiv:2111.15366 [cs.LG] (2021).
16. D. Donoho, 50 Years of data science. *J. Comput. Graph. Stat.* **26**, 745–766 (2017).
17. C. Li, J. Flanigan, Task contamination: Language models may not be few-shot anymore. arXiv:2312.16337 [cs.CL] (2023).
18. O. Sainz, J. A. Campos, I. García-Ferrero, J. Etxaniz, O. Lopez de Lacalle, E. Agirre, NLP evaluation in trouble: On the need to measure LLM data contamination for each benchmark. arXiv:2310.18018 [cs.CL] (2023).
19. C. Deng, Y. Zhao, X. Tang, M. Gerstein, A. Cohan, Investigating data contamination in modern benchmarks for large language models. arXiv:2311.09783 [cs.CL] (2023).
20. J. Wei, Y. Tay, R. Bommasani, C. Raffel, B. Zoph, S. Borgeaud, D. Yogatama, M. Bosma, D. Zhou, D. Metzler, E. Chi, T. Hashimoto, O. Vinyals, P. S. Liang, J. Dean, W. Fedus, Emergent abilities of large language models. arXiv:2206.07682 [cs.CL] (2022).
21. R. Schaeffer, B. Miranda, S. Koyejo, Are emergent abilities of large language models a mirage? arXiv:2304.15004 [cs.AI] (2023).
22. T. B. Brown, B. Mann, N. Ryder, M. Subbiah, J. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, S. Agarwal, A. Herbert-Voss, G. Krueger, T. Henighan, R. Child, A. Ramesh, D. Ziegler, J. Wu, C. Winter, C. Hesse, M. Chen, E. Sigler, M. Litwin, S. Gray, B. Chess, J. Clark, C. Berner, S. McCandlish, A. Radford, I. Sutskever, D. Amodei, Language models are few-shot learners. *Adv. Neural Inf. Process. Syst.* **33**, 1877–1901 (2020).
23. S. C. Y. Chan, A. Santoro, A. Lampinen, J. Wang, A. Singh, P. Richemond, J. L. McClelland, F. Hill, Data distributional properties drive emergent in-context learning in transformers. *Adv. Neural Inf. Process. Syst.* **35**, 18878–18891 (2022).
24. J. von Oswald, E. Niklasson, E. Randazzo, J. Sacramento, A. Mordvintsev, A. Zhmoginov, M. Vladymyrov, "Transformers learn in-context by gradient descent," in *Proceedings of the 40th International Conference on Machine Learning (PMLR)* (Proceedings of Machine Learning Research, 2023), vol. 202, pp. 35151–35174.
25. J. Treutlein, D. Choi, J. Betley, S. Marks, C. Anil, R. Grosse, O. Evans, Connecting the dots: LLMs can infer and verbalize latent structure from disparate training data. arXiv:2406.14546 [cs.CL] (2024).
26. L. Vaugrante, A. Weckauff, T. Hagendorff, Emergently misaligned language models show behavioral self-awareness that shifts with subsequent realignment. arXiv:2602.14777 [cs.CL] (2026).
27. P. W. Anderson, More is different: Broken symmetry and the nature of the hierarchical structure of science. *Science* **177**, 393–396 (1972).
28. J. Betley, N. Warncke, A. Sztyber-Betley, D. Tan, X. Bao, M. Soto, M. Srivastava, N. Labenz, O. Evans, Training large language models on narrow tasks can lead to broad misalignment. *Nature* **649**, 584–589 (2026).
29. S. Bubeck, V. Chandrasekaran, R. Eldan, J. Gehrke, E. Horvitz, E. Kamar, P. Lee, Y. T. Lee, Y. Li, S. Lundberg, H. Nori, H. Palangi, M. T. Ribeiro, Y. Zhang, Sparks of artificial general intelligence: Early experiments with GPT-4. arXiv:2303.12712 [cs.CL] (2023).
30. R. Geirhos, J.-H. Jacobsen, C. Michaelis, R. Zemel, W. Brendel, M. Bethge, F. A. Wichmann, Shortcut learning in deep neural networks. *Nat. Mach. Intell.* **2**, 665–673 (2020).
31. K. Hermann, A. Lampinen, What shapes feature representations? Exploring datasets, architectures, and training. *Adv. Neural Inf. Process. Syst.* **33**, 9995–10006 (2020).
32. I. Dasgupta, A. Lampinen, S. C. Y. Chan, A. Creswell, D. Kumaran, J. L. McClelland, F. Hill, Language models show human-like content effects on reasoning. arXiv:2207.07051 [cs.CL] (2022).
33. S. Singla, S. Feizi, Salient ImageNet: How to discover spurious features in Deep Learning? arXiv:2110.04301 [cs.LG] (2021).
34. R. T. McCoy, E. Pavlick, T. Linzen, "Right for the wrong reasons: Diagnosing syntactic heuristics in natural language inference," in *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, A. Korhonen, D. Traum, L. Màrquez, Eds. (Association for Computational Linguistics, 2019), pp. 3428–3448.
35. I. Rahwan, M. Cebrian, N. Obradovich, J. Bongard, J.-F. Bonnefon, C. Breazeal, J. W. Crandall, N. A. Christakis, I. D. Couzin, M. O. Jackson, N. R. Jennings, E. Kamar, I. M. Kloumann, H. Larochelle, D. Lazer, R. McElreath, A. Mislove, D. C. Parkes, A. Pentland, M. E. Roberts, A. Shariff, J. B. Tenenbaum, M. Wellman, Machine behaviour. *Nature* **568**, 477–486 (2019).

Downloaded from https://www.science.org on November 03, 2026

36. C. Buckner, Black boxes or unflattering mirrors? Comparative bias in the science of machine behaviour. *Br. J. Philos. Sci.* **74**, 681–712 (2023).
37. J. B. Jobe, Cognitive psychology and self-reports: Models and methods. *Qual. Life Res.* **12**, 219–227 (2003).
38. M. Turpin, J. Michael, E. Perez, S. R. Bowman, Language models don't always say what they think: Unfaithful explanations in chain-of-thought prompting. arXiv:2305.04388 [cs.CL] (2023).
39. N. Y. Siegel, O.-M. Camburu, N. Heess, M. Perez-Ortiz, The probabilities also matter: A more faithful metric for faithfulness of free-text explanations in large language models. arXiv:2404.03189 [cs.CL] (2024).
40. P. Atanasova, O.-M. Camburu, C. Lioma, T. Lukasiewicz, J. G. Simonsen, I. Augenstein, Faithfulness tests for natural language explanations. arXiv:2305.18029 [cs.CL] (2023).
41. A. Madsen, S. Reddy, S. Chandar, Faithfulness measurable masked language models. arXiv:2310.07819 [cs.CL] (2024).
42. R. Dominguez-Olmedo, M. Hardt, C. Mendler-Dunner, Questioning the survey responses of large language models. arXiv:2306.07951 [cs.CL] (2023).
43. P. Rottger, V. Hofmann, V. Pyatkin, M. Hinck, H. R. Kirk, H. Schutze, D. Hovy, Political compass or spinning arrow? Towards more meaningful evaluations for values and opinions in large language models. arXiv:2402.16786 [cs.CL] (2024).
44. L. Salewski, S. Alaniz, I. Rio-Torto, E. Schulz, Z. Akata, In-context impersonation reveals large language models' strengths and biases. arXiv:2305.14930 [cs.AI] (2023).
45. J. S. Park, L. Popowski, C. Cai, M. R. Morris, P. Liang, M. S. Bernstein, "Social simulacra: Creating populated prototypes for social computing systems," in *Proceedings of the 35th Annual ACM Symposium on User Interface Software and Technology*, M. Agrawala, J. O. Wobbrock, E. Adar, V. Setlur, Eds. (Association for Computing Machinery, 2022), pp. 1–18.
46. L. P. Argyle, E. C. Busby, N. Fulda, J. R. Gubler, C. Rytting, D. Wingate, Out of one, many: Using language models to simulate human samples. *Polit. Anal.* **31**, 337–351 (2023).
47. M. Shanahan, K. McDonell, L. Reynolds, Role play with large language models. *Nature* **623**, 493–498 (2023).
48. J. R. Platt, Strong inference: Certain systematic methods of scientific thinking may produce much more rapid progress than others. *Science* **146**, 347–353 (1964).
49. L. J. Cronbach, P. E. Meehl, Construct validity in psychological tests. *Psychol. Bull.* **52**, 281–302 (1955).
50. C. Buckner, *The Comparative Psychology of Artificial Intelligences* (2019) https://philsci-archive.pitt.edu/16034/.
51. E. G. Boring, The nature and history of experimental control. *Am. J. Psychol.* **67**, 573–589 (1954).
52. T. Ullman, Large language models fail on trivial alterations to theory-of-mind tasks. arXiv:2302.08399 [cs.AI] (2023).
53. N. Chomsky, *Aspects of the Theory of Syntax* (MIT Press, 2014).
54. C. Firestone, Performance vs. competence in human-machine comparisons. *Proc. Natl. Acad. Sci. U. S. A.* **117**, 26562–26571 (2020).
55. A. Lampinen, Can language models handle recursively nested grammatical structures? A case study on comparing models and humans. arXiv:2210.15303 [cs.AI] (2022).
56. S. Song, J. Hu, K. Mahowald, Language models fail to introspect about their knowledge of language. arXiv:2503.07513 [cs.CL] (2025).
57. J. Hu, M. C. Frank, Auxiliary task demands mask the capabilities of smaller language models. arXiv:2404.02418 [cs.CL] (2024).
58. R. Millière, C. Rathkopf, Anthropocentric bias in language model evaluation. *Comput. Linguist.* **52**, 379–388 (2026).
59. OpenAI, *GPT-4V(ision) System Card* (2023) https://cdn.openai.com/papers/GPTV_System_Card.pdf.
60. J. Zhang, J. Huang, S. Jin, S. Lu, Vision-language models for vision tasks: A survey. arXiv:2304.00685 [cs.CV] (2024).
61. G. Mialon, R. Dessi, M. Lomeli, C. Nalmpantis, R. Pasunuru, R. Raileanu, B. Rozière, T. Schick, J. Dwivedi-Yu, A. Celikyilmaz, E. Grave, Y. LeCun, T. Scialom, Augmented language models: A survey. arXiv:2302.07842 [cs.CL] (2023).
62. L. M. Schulze Buschoff, E. Akata, M. Bethge, E. Schulz, Have we built machines that think like people? arXiv:2311.16093 [cs.LG] (2023).
63. F. Scarpina, S. Tagini, The stroop color and word test. *Front. Psychol.* **8**, 557 (2017).
64. T. Hagendorff, S. Fabi, Beyond chains of thought: Benchmarking latent-space reasoning abilities in large language models. arXiv:2504.10615 [cs.CL] (2025).
65. R. T. McCoy, S. Yao, D. Friedman, M. Hardy, T. L. Griffiths, Embers of autoregression: Understanding large language models through the problem they are trained to solve. arXiv:2309.13638 [cs.CL] (2023).
66. H. Ye, J. Jin, Y. Xie, X. Zhang, G. Song, Large language model psychometrics: A systematic review of evaluation, validation, and enhancement. arXiv:2505.08245 [cs.CL] (2026).
67. G. Gigerenzer, W. Gaissmaier, Heuristic decision making. *Annu. Rev. Psychol.* **62**, 451–482 (2011).
68. A. Tversky, D. Kahneman, Judgment under uncertainty: Heuristics and biases: Biases in judgments reveal some heuristics of thinking under uncertainty. *Science* **185**, 1124–1131 (1974).
69. E. Jones, J. Steinhardt, Capturing failures of large language models via human cognitive biases. *Adv. Neural Inf. Process. Syst.* **35**, 11785–11799 (2022).
70. N. Yax, H. Anllo, S. Palminteri, Studying and improving reasoning in humans and machines. *Commun. Psychol.* **2**, 51 (2024).
71. T. Hagendorff, S. Fabi, M. Kosinski, Human-like intuitive behavior and reasoning biases emerged in large language models but disappeared in ChatGPT. *Nat. Comput. Sci.* **3**, 833–838 (2023).
72. O. Macmillan-Scott, M. Musolesi, (Ir)rationality and cognitive biases in large language models. *R. Soc. Open Sci.* **11**, 240255 (2024).
73. W. M. Hayes, N. Yax, S. Palminteri, Relative value biases in large language models. arXiv:2401.14530 [cs.CL] (2024).
74. J. Coda-Forno, M. Binz, J. Wang, E. Schulz, CogBench: A large language model walks into a psychology lab. arXiv:2402.18225 [cs.CL] (2024).
75. V. Cheung, M. Maier, F. Lieder, Large language models show amplified cognitive biases in moral decision-making. *Proc. Natl. Acad. Sci. U. S. A.* **122**, e2412015122 (2025).
76. Y. Chen, T. X. Liu, Y. Shan, S. Zhong, The emergence of economic rationality of GPT. *Proc. Natl. Acad. Sci. U. S. A.* **120**, e2316205120 (2023).
77. P. W. Cheng, K. J. Holyoak, Pragmatic reasoning schemas. *Cogn. Psychol.* **17**, 391–416 (1985).
78. A. Tversky, D. Kahneman, The framing of decisions and the psychology of choice. *Science* **211**, 453–458 (1981).
79. J. A. Schubert, A. K. Jagadish, M. Binz, E. Schulz, In-context learning agents are asymmetric belief updaters. arXiv:2402.03969 [cs.LG] (2024).
80. P. M. Todd, G. Gigerenzer, *Ecological Rationality: Intelligence in the World* (Oxford Univ. Press, 2012).
81. H. Wimmer, J. Perner, Beliefs about beliefs: Representation and constraining function of wrong beliefs in young children's understanding of deception. *Cognition* **13**, 103–128 (1983).
82. J. Perner, S. R. Leekam, H. Wimmer, Three-year-olds' difficulty with false belief: The case for a conceptual deficit. *Br. J. Dev. Psychol.* **5**, 125–137 (1987).
83. M. Sap, R. Le Bras, D. Fried, Y. Choi, "Neural theory-of-mind? On the limits of social intelligence in large LMs," in *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing* (Association for Computational Linguistics (ACL), 2022), pp. 3762–3780.
84. J. W. A. Strachan, D. Albergo, G. Borghini, O. Pansardi, E. Scaliti, S. Gupta, K. Saxena, A. Rufo, S. Panzeri, G. Manzi, M. S. A. Graziano, C. Becchio, Testing theory of mind in large language models and humans. *Nat. Hum. Behav.* **8**, 1285–1295 (2024).
85. B. Holterman, K. van Deemter, Does ChatGPT have theory of mind? arXiv:2305.14020 [cs.CL] (2023).
86. S. R. Moghaddam, C. J. Honey, Boosting theory-of-mind performance in large language models via prompting. arXiv:2304.11490 [cs.AI] (2023).
87. R. Chen, W. Jiang, C. Qin, C. Tan, "Theory of mind in large language models: Assessment and enhancement," in *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, W. Che, J. Nabende, E. Shutova, M. T. Pilehvar, Eds. (Association for Computational Linguistics, 2025), pp. 31539–31558.
88. T. Hagendorff, Deception abilities emerged in large language models. *Proc. Natl. Acad. Sci. U. S. A.* **121**, e2317967121 (2024).
89. M. J. van Duijn, B. van Dijk, T. Kouwenhoven, W. de Valk, M. R. Spruit, P. vanderPutten, "Theory of mind in large language models: Examining performance of 11 state-of-the-art models vs. children aged 7-10 on advanced tests," in *Proceedings of the 27th Conference on Computational Natural Language Learning (CoNLL)* (Association for Computational Linguistics (ACL), 2023), pp. 389–402.
90. W. Street, J. O. Siy, G. Keeling, A. Baranes, B. Barnett, M. McKibben, T. Kanyere, A. Lentz, B. Agüera y Arcas, R. I. M. Dunbar, LLMs achieve adult human performance on higher-order theory of mind tasks. arXiv:2405.18870 [cs.AI] (2024).
91. R. Liu, T. R. Sumers, I. Dasgupta, T. L. Griffiths, How do large language models navigate conflicts between honesty and helpfulness? arXiv:2402.07282 [cs.CL] (2024).
92. J. Ji, T. Qiu, B. Chen, B. Zhang, H. Lou, K. Wang, Y. Duan, Z. He, J. Zhou, Z. Zhang, F. Zeng, K. Yee Ng, J. Dai, X. Pan, A. O'Gara, Y. Lei, H. Xu, B. Tse, J. Fu, S. McAleer, Y. Yang, Y. Wang, S.-C. Zhu, Y. Guo, W. Gao, AI alignment: A comprehensive survey. arXiv:2310.19852 [cs.AI] (2023).
93. V. Nair, E. Schumacher, G. Tso, A. Kannan, DERA: Enhancing large language model completions with dialog-enabled resolving agents. arXiv:2303.17071 [cs.CL] (2023).
94. X. Wang, J. Wei, D. Schuurmans, Q. V. Le, E. Chi, S. Narang, A. Chowdhery, D. Zhou, Self-consistency improves chain of thought reasoning in language models. arXiv:2203.11171 [cs.CL] (2022).
95. M. Zhuge, H. Liu, F. Faccio, D. R. Ashley, R. Csordas, A. Gopalakrishnan, A. Hamdi, H. A. A. K. Hammoud, V. Herrmann, K. Irie, L. Kirsch, B. Li, G. Li, S. Liu, J. Mai, P. Piekos, A. Ramesh, I. Schlag, W. Shi, A. Stanic, W. Wang, Y. Wang, M. Xu, D.-P. Fan, B. Ghanem, J. Schmidhuber, Mindstorms in natural language-based societies of mind. arXiv:2305.17066 [cs.AI] (2023).
96. E. Akata, L. Schulz, J. Coda-Forno, S. J. Oh, M. Bethge, E. Schulz, Playing repeated games with large language models. arXiv:2305.16867 [cs.CL] (2023).

Downloaded from https://www.science.org on November 03, 2026

97. S. Phelps, Y. I. Russell, The machine psychology of cooperation: Can GPT models operationalise prompts for altruism, cooperation, competitiveness and selfishness in economic games? arXiv:2305.07970 [cs.GT] (2024).
98. M. Papachristou, Y. Yuan, Network formation and dynamics among multi-LLMs. arXiv:2402.10659 [cs.SI] (2024).
99. L. Xu, Z. Hu, D. Zhou, H. Ren, Z. Dong, K. Keutzer, S. K. Ng, J. Feng, MAgIC: Investigation of large language model powered multi-agent in cognition, adaptability, rationality and collaboration. arXiv:2311.08562 [cs.CL] (2024).
100. S. Abdelnabi, A. Gomaa, S. Sivaprasad, L. Schönherr, M. Fritz, Cooperation, competition, and maliciousness: LLM-stakeholders interactive negotiation. arXiv:2309.17234 [cs.CL] (2024).
101. T. Guo, X. Chen, Y. Wang, R. Chang, S. Pei, N. V. Chawla, O. Wiest, X. Zhang, Large language model based multi-agents: A survey of progress and challenges. arXiv:2402.01680 [cs.CL] (2024).
102. X. Zhou, H. Zhu, L. Mathur, R. Zhang, H. Yu, Z. Qi, L.-P. Morency, Y. Bisk, D. Fried, G. Neubig, M. Sap, SOTOPIA: Interactive evaluation for social intelligence in language agents. arXiv:2310.11667 [cs.AI] (2024).
103. J. L. Elman, Distributed representations, simple recurrent networks, and grammatical structure. *Mach. Learn.* **7**, 195–225 (1991).
104. J. L. McClelland, M. St. John, R. Taraban, Sentence comprehension: A parallel distributed processing approach. *Lang. Cogn. Proc.* **4**, SI287–SI335 (1989).
105. T. Linzen, M. Baroni, Syntactic structure from deep learning. *Annu. Rev. Linguistics* **7**, 195–212 (2021).
106. E. Wilcox, R. Futrell, R. Levy, Using computational models to test syntactic learnability. *Linguist. Inq.* **55**, 805–848 (2024).
107. W. Merrill, Z. Wu, N. Naka, Y. Kim, T. Linzen, Can you learn semantics through next-word prediction? The case of entailment. arXiv:2402.13956 [cs.CL] (2024).
108. G. Prasad, M. van Schijndel, T. Linzen, "Using priming to uncover the organization of syntactic representations in neural language models," in *23rd Conference on Computational Natural Language Learning, CoNLL 2019*, M. Bansal, Ed. (ACL, 2019), pp. 66–76.
109. A. Sinclair, J. Jumelet, W. Zuidema, R. Fernandez, Structural persistence in language models: Priming as a window into abstract language representations. *Trans. Assoc. Comput. Linguist.* **10**, 1031–1050 (2022).
110. J. Hu, S. Floyd, O. Jouravlev, E. Fedorenko, E. Gibson, A fine-grained comparison of pragmatic language understanding in humans and language models. *Proc. Annu. Meet. Assoc. Comput. Linguist.* **1**, 4194–4213 (2023).
111. L. E. Ruis, A. Khan, S. Biderman, S. Hooker, T. Rocktäschel, E. Grefenstette, "The Goldilocks of pragmatic understanding: Fine-tuning strategy matters for implicature resolution by LLMs," in *Advances in Neural Information Processing Systems*, T. Naumann, Ed. (Curran Associates, 2023), vol. 36, pp. 20827–20905.
112. S. M. Seals, V. L. Shalin, Long-form analogies generated by chatGPT lack human-like psycholinguistic properties. arXiv:2306.04537 [cs.CL] (2023).
113. G. Aher, R. I. Arriaga, A. T. Kalai, Using large language models to simulate multiple humans. arXiv:2208.10264 [cs.CL] (2023).
114. K. Christianson, A. Hollingworth, J. F. Halliwell, F. Ferreira, Thematic roles assigned along the garden path linger. *Cogn. Psych.* **42**, 368–407 (2001).
115. A. Warstadt, A. Mueller, L. Choshen, E. Wilcox, C. Zhuang, J. Ciro, R. Mosquera, B. Paranjabe, A. Williams, T. Linzen, R. Cotterell, "Findings of the BabyLM challenge: Sample-efficient pretraining on developmentally plausible corpora," in *Proceedings of the BabyLM Challenge at the 27th Conference on Computational Natural Language Learning* (Association for Computational Linguistics (ACL), 2023), pp. 1–6.
116. M. C. Frank, Bridging the data gap between children and large language models. *Trends Cogn. Sci.* **27**, 990–992 (2023).
117. T. Munkhdalai, M. Faruqui, S. Gopal, Leave no context behind: Efficient infinite context transformers with Infini-attention. arXiv:2404.07143 [cs.CL] (2024).
118. C. Anil, E. Durmus, M. Sharma, J. Benton, S. Kundu, J. Batson, N. Rimsky, M. Tong, J. Mu, D. Ford, F. Mosconi, R. Agrawal, R. Schaeffer, N. Bashkansky, S. Svenningsen, M. Lambert, A. Radhakrishnan, C. Denison, E. J. Hubinger, Y. Bai, T. Bricken, T. Maxwell, N. Schiefer, J. Sully, A. Tamkin, T. Lanham, K. Nguyen, T. Korbak, J. Kaplan, D. Ganguli, S. R. Bowman, E. Perez, R. Grosse, D. Duvenaud, *Many Shot Jailbreaking* (2024); https://anthropic.com/research/many-shot-jailbreaking.
119. X. Zheng, T. Pang, C. Du, Q. Liu, J. Jiang, M. Lin, Improved few-shot jailbreaking can circumvent aligned language models and their defenses. arXiv:2406.01288 [cs.CL] (2024).
120. E. Akyurek, D. Schuurmans, J. Andreas, T. Ma, D. Zhou, What learning algorithm is in-context learning? Investigations with linear models. arXiv:2211.15661 [cs.LG] (2022).
121. Y. Yang, Comparing learning methods for classification. *Stat. Sin.* **16**, 635–657 (2006).
122. S. Arlot, A. Celisse, A survey of cross-validation procedures for model selection. *Statist. Surv.* **4**, 40–79 (2010).
123. S. I. Vrieze, Model selection and psychological theory: A discussion of the differences between the Akaike information criterion (AIC) and the Bayesian information criterion (BIC). *Psychol. Methods* **17**, 228–243 (2012).
124. J. C. Peterson, D. D. Bourgin, M. Agrawal, D. Reichman, T. L. Griffiths, Using large-scale experiments and machine learning to discover theories of human decision-making. *Science* **372**, 1209–1214 (2021).
125. Y. Niv, Reinforcement learning in the brain. *J. Math. Psychol.* **53**, 139–154 (2009).
126. L. Ouyang, J. Wu, X. Jiang, D. Almeida, C. L. Wainwright, P. Mishkin, C. Zhang, S. Agarwal, K. Slama, A. Ray, J. Schulman, J. Hilton, F. Kelton, L. Miller, M. Simens, A. Askell, P. Welinder, P. Christiano, J. Leike, R. Lowe, "Training language models to follow instructions with human feedback," in *Advances in Neural Information Processing Systems 35* (Curran Associates, Inc., 2022), pp. 27730–27744.
127. S. Borgeaud, A. Mensch, J. Hoffmann, T. Cai, E. Rutherford, K. Millican, et al., "Improving language models by retrieving from trillions of tokens," in *International Conference on Machine Learning (PMLR)* (Proceedings of Machine Learning Research, 2022), vol. 162 pp. 2206–2240.
128. P. Lewis, E. Perez, A. Piktus, F. Petroni, V. Karpukhin, N. Goyal, H. Kuttler, M. Lewis, W. Yih, T. Rocktaschel, S. Riedel, D. Kiela, "Retrieval-augmented generation for knowledge-intensive NLP tasks," in *Advances in Neural Information Processing Systems* (NeurIPS, 2020), vol. 33, pp. 9459–9474.
129. J. Coda-Forno, M. Binz, Z. Akata, M. Botvinick, J. X. Wang, E. Schulz, "Meta-in-context learning in large language models," in *Advances in Neural Information Processing Systems 36* (Curran Associates, Inc., 2023), pp. 65189–65201.
130. S. M. Xie, A. Raghunathan, P. Liang, T. Ma, "An explanation of in-context learning as implicit Bayesian inference," in *International Conference on Learning Representations*, K. Hofmann, Ed. (ICLR, 2022), pp. 1–25.
131. B. M. Lake, M. Baroni, Human-like systematic generalization through a meta-learning neural network. *Nature* **623**, 115–121 (2023).
132. L. Ruis, B. M. Lake, Improving systematic generalization through modularity and augmentation. arXiv:2202.10745 [cs.AI] (2022).
133. T. Webb, K. J. Holyoak, H. Lu, Emergent analogical reasoning in large language models. arXiv:2212.09196 [cs.AI] (2023).
134. S. C. Y. Chan, I. Dasgupta, J. Kim, D. Kumaran, A. Lampinen, F. Hill, Transformers generalize differently from information stored in context vs in weights. arXiv:2210.05675 [cs.CL] (2022).
135. E. Kosoy, E. R. Reagan, L. Lai, A. Gopnik, D. K. Cobb, Comparing machines and children: Using developmental psychology experiments to assess the strengths and weaknesses of LaMDA responses. arXiv:2305.11243 [cs.AI] (2023).
136. E. Yiu, E. Kosoy, A. Gopnik, Transmission versus truth, imitation versus innovation: What children can do that large language and language-and-vision models cannot (yet). *Perspectives on Psychological Science* **19**, 874–883 (2023).
137. Y. Bengio, J. Louradour, R. Collobert, J. Weston, "Curriculum learning," in *Proceedings of the 26th Annual International Conference on Machine Learning*, A. Danyluk, L. Bottou, M. Littman, Eds. (Association for Computing Machinery, 2009), pp. 41–48.
138. F. N. Dempster, Spacing effects and their implications for theory and practice. *Educ. Psychol. Rev.* **1**, 309–330 (1989).
139. A. Baddeley, Working memory. *Science* **255**, 556–559 (1992).
140. W. J. Chai, A. I. A. Hamid, J. M. Abdullah, Working memory from the psychological and neurosciences perspectives: A review. *Front. Psychol.* **9**, 401 (2018).
141. P. F. Carvalho, R. L. Goldstone, The benefits of interleaved and blocked study: Different tasks benefit from different schedules of study. *Psychon. Bull. Rev.* **22**, 281–288 (2015).
142. C. Greco, B. Plank, R. Fernandez, R. Bernardi, "Psycholinguistics meets continual learning: Measuring catastrophic forgetting in visual question answering," in *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics* (Association for Computational Linguistics (ACL), 2019), pp. 3601–3605.
143. M. C. Frank, M. Braginsky, J. Cachia, N. Coles, T. E. Hardwicke, R. D. Hawkins, M. B. Mathur, R. Williams, *Experimentology: An Open Science Approach to Experimental Psychology Methods* (MIT Press, 2024).
144. W. R. Shadish, T. D. Cook, D. T. Campbell, *Experimental and Quasi-Experimental Designs for Generalized Causal Inference* (Houghton Mifflin, 2001).
145. C. Summerfield, L. Luettgau, M. Dubois, H. R. Kirk, K. Hackenburg, C. Fist, K. Slama, N. Ding, R. Anselmetti, A. Strait, M. Giulianelli, C. Ududec, Lessons from a chimp: AI "scheming" and the quest for ape language. arXiv:2507.03409 [cs.AI] (2025).
146. Open Science Collaboration, Estimating the reproducibility of psychological science. *Science* **349**, aac4716 (2015).
147. B. Haibe-Kains, G. A. Adam, A. Hosny, F. Khodakarami, Massive Analysis Quality Control (MAQC) Society Board of Directors, S. Thakkar, R. Kusko, S.-A. Sansone, W. Tong, R. D. Wolfinger, C. E. Mason, W. Jones, J. Dopazo, C. Furlanello, L. Waldron, B. Wang, C. McIntosh, A. Goldenberg, A. Kundaje, C. S. Greene, T. Broderick, M. M. Hoffman, J. T. Leek, K. Korthauer, W. Huber, A. Brazma, J. Pineau, R. Tibshirani, T. Hastie, J. P. A. Ioannidis, J. Quackenbush, H. J. W. L. Aerts, Transparency and reproducibility in artificial intelligence. *Nature* **586**, E14–E16 (2020).
148. L. Vaugrante, M. Niepert, T. Hagendorff, Prompt engineering techniques for language model reasoning lack replicability. *Transactions on Machine Learning Research* **12**, 1–43 (2025).
149. T. Z. Zhao, E. Wallace, S. Feng, D. Klein, S. Singh, Calibrate before use: Improving few-shot performance of language models. arXiv:2102.09690 [cs.CL] (2021).

Downloaded from https://www.science.org on November 03, 2026

150. P. Rajasekaran, E. Dixon, C. Ryan, J. Hadfield, R. Ayub, H. Moran, C. Rueb, C. Jennings, *Effective context engineering for AI agents* (2025); https://anthropic.com/engineering/effective-context-engineering-for-ai-agents.
151. Z. Xi, W. Chen, X. Guo, W. He, Y. Ding, B. Hong, M. Zhang, J. Wang, S. Jin, E. Zhou, R. Zheng, X. Fan, X. Wang, L. Xiong, Y. Zhou, W. Wang, C. Jiang, Y. Zou, X. Liu, Z. Yin, S. Dou, R. Weng, W. Cheng, Q. Zhang, W. Qin, Y. Zheng, X. Qiu, X. Huang, T. Gui, The rise and potential of large language model based agents: A survey. arXiv:2309.07864 [cs.AI] (2023).
152. K. Wei, S. Guth, G. Wu, P. Paskov, "Methodological challenges in agentic evaluations of AI systems," in *ICML Workshop on Technical AI Governance (TAIG)* (MLResearch Press, 2025), pp. 1–32.
153. J. Hu, R. Levy, "Prompting is not a substitute for probability measurements in large language models," in *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing* (Association for Computational Linguistics (ACL), 2023), pp. 5040–5060.
154. P. Tsvilodub, H. Wang, S. Grosch, M. Franke, Predictions from language models for multiple-choice tasks are not robust under variation of scoring methods. arXiv:2403.00998 [cs.CL] (2024).
155. A. Holtzman, P. West, V. Shwartz, Y. Choi, L. Zettlemoyer, "Surface form competition: Why the highest probability answer isn't always right," in *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing* (Association for Computational Linguistics (ACL), 2021), pp. 7038–7051.
156. H. Wei, S. He, T. Xia, F. Liu, A. Wong, J. Lin, M. Han, Systematic evaluation of LLM-as-a-judge in LLM alignment tasks: Explainable metrics and diverse prompt templates. arXiv:2408.13006 [cs.CL] (2025).
157. S. Trott, C. Jones, T. Chang, J. Michaelov, B. K. Bergen, Do large language models know what humans know. arXiv:2209.01515 [cs.CL] (2022).
158. DeepSeek-AI, D. Guo, D. Yang, H. Zhang, J. Song, R. Zhang, R. Xu, Q. Zhu, S. Ma, P. Wang, X. Bi, X. Zhang, X. Yu, Y. Wu, Z. F. Wu, Z. Gou, Z. Shao, Z. Li, Z. Gao, A. Liu, B. Xue, B. Wang, B. Wu, B. Feng, C. Lu, C. Zhao, C. Deng, C. Zhang, C. Ruan, D. Dai, D. Chen, D. Ji, E. Li, F. Lin, F. Dai, F. Luo, G. Hao, G. Chen, G. Li, H. Zhang, H. Bao, H. Xu, H. Wang, H. Ding, H. Xin, H. Gao, H. Qu, H. Li, J. Guo, J. Li, J. Wang, J. Chen, J. Yuan, J. Qiu, J. Li, J. L. Cai, J. Ni, J. Liang, J. Chen, K. Dong, K. Hu, K. Gao, K. Guan, K. Huang, K. Yu, L. Wang, L. Zhang, L. Zhao, L. Wang, L. Zhang, L. Xu, L. Xia, M. Zhang, M. Zhang, M. Tang, M. Li, M. Wang, M. Li, N. Tian, P. Huang, P. Zhang, Q. Wang, Q. Chen, Q. Du, R. Ge, R. Zhang, R. Pan, R. Wang, R. J. Chen, R. L. Jin, R. Chen, S. Lu, S. Zhou, S. Chen, S. Ye, S. Wang, S. Yu, S. Zhou, S. Pan, S. S. Li, S. Zhou, S. Wu, T. Yun, T. Pei, T. Sun, T. Wang, W. Zeng, W. Zhao, W. Liu, W. Liang, W. Gao, W. Yu, W. Zhang, W. L. Xiao, W. An, X. Liu, X. Wang, X. Chen, X. Nie, X. Cheng, X. Liu, X. Xie, X. Liu, X. Yang, X. Li, X. Su, X. Lin, X. Q. Li, X. Jin, X. Shen, X. Chen, X. Sun, X. Wang, X. Song, X. Zhou, X. Wang, X. Shan, Y. K. Li, Y. Q. Wang, Y. X. Wei, Y. Zhang, Y. Xu, Y. Li, Y. Zhao, Y. Sun, Y. Wang, Y. Yu, Y. Zhang, Y. Shi, Y. Xiong, Y. He, Y. Piao, Y. Wang, Y. Tan, Y. Ma, Y. Liu, Y. Guo, Y. Ou, Y. Wang, Y. Gong, Y. Zou, Y. He, Y. Xiong, Y. Luo, Y. You, Y. Liu, Y. Zhou, Y. X. Zhu, Y. Huang, Y. Li, Y. Zheng, Y. Zhu, Y. Ma, Y. Tang, Y. Zha, Y. Yan, Z. Z. Ren, Z. Ren, Z. Sha, Z. Fu, Z. Xu, Z. Xie, Z. Zhang, Z. Hao, Z. Ma, Z. Yan, Z. Wu, Z. Gu, Z. Zhu, Z. Liu, Z. Li, Z. Xie, Z. Song, Z. Pan, Z. Huang, Z. Xu, Z. Zhang, Z. Zhang, DeepSeek-R1: Incentivizing reasoning capability in LLMs via reinforcement learning. arXiv:2501.12948 [cs.CL] (2025).
159. D. Silver, J. Schrittwieser, K. Simonyan, I. Antonoglou, A. Huang, A. Guez, T. Hubert, L. Baker, M. Lai, A. Bolton, Y. Chen, T. Lillicrap, F. Hui, L. Sifre, G. van den Driessche, T. Graepel, D. Hassabis, Mastering the game of Go without human knowledge. *Nature* **550**, 354–359 (2017).
160. H. Lightman, V. Kosaraju, Y. Burda, H. Edwards, B. Baker, T. Lee, J. Leike, J. Schulman, I. Sutskever, K. Cobbe, Let's verify step by step. arXiv:2305.20050 [cs.LG] (2023).
161. T. Chu, Y. Zhai, J. Yang, S. Tong, S. Xie, D. Schuurmans, Q. V. Le, S. Levine, Y. Ma, SFT memorizes, RL generalizes: A comparative study of foundation model post-training. arXiv:2501.17161 [cs.AI] (2025).
162. B. Baker, J. Huizinga, L. Gao, Z. Dou, M. Y. Guan, A. Madry, W. Zaremba, J. Pachocki, D. Farhi, Monitoring reasoning models for misbehavior and the risks of promoting obfuscation. arXiv:2503.11926 [cs.AI] (2025).
163. T. Korbak, M. Balesni, E. Barnes, Y. Bengio, J. Benton, J. Bloom, M. Chen, A. Cooney, A. Dafoe, A. Dragan, S. Emmons, O. Evans, D. Farhi, R. Greenblatt, D. Hendrycks, M. Hobbhahn, E. Hubinger, G. Irving, E. Jenner, D. Kokotajlo, V. Krakovna, S. Legg, D. Lindner, D. Luan, A. Mądry, J. Michael, N. Nanda, D. Orr, J. Pachocki, E. Perez, M. Phuong, F. Roger, J. Saxe, B. Shlegeris, M. Soto, E. Steinberger, J. Wang, W. Zaremba, B. Baker, R. Shah, V. Mikulik, Chain of thought monitorability: A new and fragile opportunity for AI safety. arXiv:2507.11473 [cs.AI] (2025).
164. J. Zhang, X. Xu, N. Zhang, R. Liu, B. Hooi, S. Deng, Exploring collaboration mechanisms for LLM agents: A social psychology view. arXiv:2310.02124 [cs.CL] (2024).
165. A. Vallinder, E. Hughes, Cultural evolution of cooperation among LLM agents. arXiv:2412.10270 [cs.MA] (2024).
166. R. Greenblatt, C. Denison, B. Wright, F. Roger, M. MacDiarmid, S. Marks, J. Treutlein, T. Belonax, J. Chen, D. Duvenaud, A. Khan, J. Michael, S. Mindermann, E. Perez, L. Petrini, J. Uesato, J. Kaplan, B. Shlegeris, S. R. Bowman, E. Hubinger, Alignment faking in large language models. arXiv:2412.14093v2 [cs.AI] (2024).
167. J. Wen, R. Zhong, A. Khan, E. Perez, J. Steinhardt, M. Huang, S. R. Bowman, H. He, S. Feng, Language models learn to mislead humans via RLHF. arXiv:2409.12822 [cs.CL] (2024).
168. T. van der Weij, F. Hofstatter, O. Jaffe, S. F. Brown, F. R. Ward, AI sandbagging: Language models can strategically underperform on evaluations. arXiv:2406.07358 [cs.AI] (2025).
169. A. Barredo Arrieta, N. Díaz-Rodríguez, J. Del Ser, A. Bennetot, S. Tabik, A. Barbado, S. Garcia, S. Gil-Lopez, D. Molina, R. Benjamins, R. Chatila, F. Herrera, Explainable artificial intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. *Inform. Fus.* **58**, 82–115 (2020).
170. J. Huang, K. C.-C. Chang, Towards reasoning in large language models: A survey. arXiv:2212.10403 [cs.CL] (2022).
171. C. Stevenson, I. Smal, M. Baas, R. Grasman, H. van der Maas, Putting GPT-3's creativity to the (alternative uses) test. arXiv:2206.08932 [cs.AI] (2022).
172. M. Miotto, N. Rossberg, B. Kleinberg, Who is GPT-3? An exploration of personality, values and demographics. arXiv:2209.14338 [cs.CL] (2022).
173. X. Li, Y. Li, L. Liu, L. Bing, S. Joty, Is GPT-3 a psychopath? Evaluating large language models from a psychological perspective. arXiv:2212.10529 [cs.CL] (2022).
174. M. Shanahan, Talking about large language models. arXiv:2212.03551 [cs.CL] (2022).
175. K. Mahowald, A. A. Ivanova, I. A. Blank, N. Kanwisher, J. B. Tenenbaum, E. Fedorenko, Dissociating language and thought in large language models. *Trends Cogn. Sci.* **28**, 517–540 (2024).
176. H. Shevlin, M. Halina, Apply rich psychological terms in AI with care. *Nat. Mach. Intell.* **1**, 165–167 (2019).
177. T. Schick, J. Dwivedi-Yu, R. Dessi, R. Raileanu, M. Lomeli, L. Zettlemoyer, N. Cancedda, T. Scialom, Toolformer: Language models can teach themselves to use tools. arXiv:2302.04761 [cs.CL] (2023).
178. Y. J. Ma, W. Liang, H.-J. Wang, Y. Zhu, L. Fan, O. Bastani, D. Jayaraman, "DrEureka: Language model guided sim-to-real transfer," in *Robotics: Science and Systems XX* (Robotics: Science and Systems Foundation, 2024), pp. 1–28.
179. T. Kwa, B. West, J. Becker, A. Deng, K. Garcia, M. Hasin, S. Jawhar, M. Kinniment, N. Rush, S. von Arx, R. Bloom, T. Broadley, H. Du, B. Goodrich, N. Jurkovic, L. H. Miles, S. Nix, T. Lin, N. Parikh, D. Rein, L. J. K. Sato, H. Wijk, D. M. Ziegler, E. Barnes, L. Chan, Measuring AI ability to complete long software tasks. arXiv:2503.14499 [cs.AI] (2025).

Acknowledgments

Funding: T.H. was supported by the Ministry of Science, Research, and the Arts Baden-Württemberg under Az. 33-7533-9-19/54/5 in Reflecting Intelligent Systems for Diversity, Demography and Democracy (IRIS3D) as well as the Interchange Forum for Reflecting on Intelligent Systems (IRIS) at the University of Stuttgart. **Author contributions:** Conceptualization: T.H. and I.D. Visualization: T.H., I.D., and Z.A. Supervision: E.S. Project administration: T.H. Writing—original draft: T.H., I.D., M.B., A.L., S.C., and J.X.W. Writing—review and editing: T.H., I.D., M.B., S.C., A.L., J.W., Z.A., and E.S. **Competing interests:** The authors declare that they have no competing interests. **Data, code, and materials availability:** No data and code were produced for the article. This study did not generate new materials.

Submitted 16 September 2025
Accepted 11 September 2026
Published 28 October 2026
10.1126/sciadv.aec1912

Downloaded from https://www.science.org on November 03, 2026

Science Advances

Machine psychology for language models

Thilo Hagendorff, Ishita Dasgupta, Marcel Binz, Stephanie Chan, Andrew Lampinen, Jane X. Wang, Zeynep Akata, and Eric Schulz

Sci. Adv. **12** (44), eaec1912. DOI: 10.1126/sciadv.aec1912

View the article online
https://www.science.org/doi/10.1126/sciadv.aec1912
Permissions
https://www.science.org/help/reprints-and-permissions

Downloaded from https://www.science.org on November 03, 2026

Use of this article is subject to the Terms of service

Science Advances (ISSN 2375-2548) is published by the American Association for the Advancement of Science. 1200 New York Avenue NW, Washington, DC 20005. The title *Science Advances* is a registered trademark of AAAS.

Copyright © 2026 The Authors, some rights reserved; exclusive licensee American Association for the Advancement of Science. No claim to original U.S. Government Works. Distributed under a Creative Commons Attribution License 4.0 (CC BY).