BIO Integration

Review

Large Language Models and Cognitive Science: A Comprehensive Review of Similarities, Differences, and Challenges

Graphical abstract

Highlights

- LLMs achieve human-level word prediction performance (r = 0.79) and demonstrate neural representation alignment with brain imaging (fMRI/MEG) data, thus suggesting shared computational principles in language processing.
- Critical differences persist between LLMs and human cognition: LLMs lack embodied grounding, show fragile out-of-distribution generalization, and exhibit task-dependent conceptual representations, unlike the coherent structures observed in humans.
- The symbol grounding problem remains a fundamental challenge. LLMs learn meanings from statistical co-occurrence patterns rather than sensorimotor experience, thus prompting questions regarding genuine language understanding.
- Integration of LLMs with cognitive architectures (Soar or ACT-R) offers promising approaches for creating more robust AI systems that combine neural flexibility with structured reasoning.
- Biomedical applications emerge at this intersection, including LLM-based cognitive assessment tools, brain-computer interfaces, and computational models for understanding neurological conditions.

Authors

Qian Niu, Junyu Liu, Ziqian Bi, Pohsun Feng, Benji Peng, Keyu Chen, Ming Li, Lawrence K. Q. Yan, Yichao Zhang, Caitlyn Heqi Yin, Cheng Fei, Tianyang Wang, Yunze Wang, Silin Chen, Ming Liu, Ziyuan Qin, Riyang Bao, Xinyuan Song and Zekun Jiang

Correspondence

niu.qian.f44@kyoto-u.jp (Q. Niu)

In brief

This comprehensive review examines the intersection of large language models (LLMs) and cognitive science, and systematically analyzes similarities and differences between artificial and human cognition across multiple domains including language processing, reasoning, memory, and causal inference. We evaluate methods for assessing LLMs' cognitive capabilities, discuss their potential as cognitive models, and identify key challenges and future research directions for advancing both AI development and the understanding of human cognition.

Review

BIO Integration

Large Language Models and Cognitive Science: A Comprehensive Review of Similarities, Differences, and Challenges

Qian Niu[1,*], Junyu Liu[1], Ziqian Bi[2], Pohsun Feng[3], Benji Peng[4], Keyu Chen[4], Ming Li[4], Lawrence K. Q. Yan[5], Yichao Zhang[6], Caitlyn Heqi Yin[7], Cheng Fei[8], Tianyang Wang[9], Yunze Wang[10], Silin Chen[11], Ming Liu[12], Ziyuan Qin[13], Riyang Bao[13], Xinyuan Song[13] and Zekun Jiang[14]

Abstract

This comprehensive review explores the intersection between large language models (LLMs) and cognitive science, by examining similarities and differences between LLMs and human cognitive processes. We analyze methods for evaluating LLMs' cognitive abilities and discuss their potential as cognitive models. This review covers applications of LLMs in various cognitive fields and highlights insights gained for cognitive science research. We assess cognitive biases and limitations of LLMs, along with proposed methods for improving their performance. The integration of LLMs with cognitive architectures is examined, thus revealing promising avenues for enhancing artificial intelligence (AI) capabilities. Key challenges and future research directions are identified, emphasizing the need for continued refinement of LLMs to better align with human cognition. This review provides a balanced perspective on the current state and future potential of LLMs in advancing understanding of both AI and human intelligence.

Keywords

Artificial Intelligence, cognitive psychology, cognitive science, human cognition, large language models, neuro-science.

Introduction

The emergence of large language models (LLMs) has sparked a revolution in artificial intelligence (AI), and challenged understanding of machine cognition and its relationship to human cognitive processes. As these models demonstrate increasingly sophisticated capabilities in language processing, reasoning, and problem-solving, they have become a focal point for cognitive scientists seeking to unravel the mysteries of human cognition. This intersection between LLMs and cognitive science has given rise to a new frontier of research offering unprecedented opportunities to explore the nature of intelligence, language, and thought.

The relationship between LLMs and cognitive science is multifaceted and bidirectional. The study of human cognition has a rich theoretical foundation, from early proposals about modular cognitive systems [1] to unified theories attempting to explain cognition as a coherent system [2]. On the one hand, insights from cognitive science have informed the development and evaluation of LLMs, and inspired new architectures and training paradigms aimed at more closely mimicking human cognitive processes. On the other hand, the remarkable performance of LLMs on various cognitive tasks has prompted researchers to reevaluate existing theories of cognition and consider new perspectives on how intelligence emerges from complex systems.

This review is aimed at providing a comprehensive overview of the current state of research at the intersection of LLMs and cognitive science. We explore the similarities and differences between LLMs and human cognitive processes, and examine how these models perform in tasks traditionally used to study human cognition. We also examine the methods developed for evaluating LLMs' cognitive abilities, and highlight the challenges and opportunities in assessing AI through the lens of cognitive science. Furthermore, we investigate the potential of LLMs to serve as cognitive models, including a discussion of their applications in various domains of cognitive science research and the insights

[1]Kyoto University, Yoshida-Honmachi, Sakyo-ku, Kyoto 606-8501, Japan

[2]Indiana University, Bloomington, IN 47405, USA

[3]National Taiwan Normal University, Taipei 106, Taiwan

[4]Georgia Institute of Technology, Atlanta, GA 30332, USA

[5]Hong Kong University of Science and Technology, Clear Water Bay, Hong Kong

[6]The University of Texas at Dallas, Richardson, TX 75080, USA

[7]University of Wisconsin-Madison, Madison, WI 53706, USA

[8]Cornell University, Ithaca, NY 14853, USA

[9]University of Liverpool, Liverpool L69 3BX, UK

[10]University of Edinburgh, Edinburgh EH8 9YL, UK

[11]Zhejiang University, Hangzhou 310027, China

[12]Purdue University, West Lafayette, IN 47907, USA

[13]Emory University, Atlanta, GA 30322, USA

[14]West China Biomedical Big Data Center, West China Hospital, Sichuan University, Chengdu, Sichuan 610041, China

*Correspondence to: Qian Niu, Kyoto University, Yoshida-Honmachi, Sakyo-ku, Kyoto 606-8501, Japan, E-mail: niu.qian.f44@kyoto-u.jp

Received: November 5 2025
Revised: November 29 2025
Accepted: January 13 2026
Published Online: March 16 2026

Available at: https://bio-integration.org/

that they provide into human cognition. This review also addresses the cognitive biases and limitations of LLMs, and discusses ongoing efforts to improve their performance and increase their alignment with human cognitive processes. We further examine recent developments in this area, and discuss the potential synergies and challenges that arise from combining these approaches.

In contrast to several recent surveys that have examined LLMs from technical perspectives or narrowly focused on specific cognitive capabilities, this review provides an integrative analysis that bridges AI and cognitive science, with particular emphasis on biomedical and neuroscientific implications. Unlike prior surveys addressing the engineering aspects of LLMs or their performance on isolated cognitive benchmarks, we synthesize findings across cognitive domains as diverse as language processing and reasoning, memory, and decision-making, then explicitly connect these insights to neuroimaging evidence and potential clinical applications. This interdisciplinary approach positions our review at the nexus of AI, cognitive psychology, and biomedical research, thereby offering a unique perspective into how LLMs can both inform and be informed by understanding of the human brain.

As LLMs continue to evolve, and their capabilities expand, critically assessing their relationship with human cognition and their potential effects on cognitive science research will become increasingly important. This review offers a balanced and comprehensive examination of these issues by presenting insights into the current state of the field. It identifies key areas for future research, and discusses the challenges and opportunities at the exciting intersection of LLMs and cognitive science. By bridging AI with cognitive science, this line of inquiry promises to deepen understanding of human cognition and inform the development of more sophisticated, ethical, and human-centric AI systems. This comprehensive critical examination highlights current achievements and further maps a path forward in this dynamic area of study.

Review methods

Search strategy

To ensure a comprehensive and systematic review, we conducted an extensive literature search across multiple academic databases including Web of Science, Scopus, PubMed, Google Scholar, and arXiv. The search used combinations of keywords such as “large language models,” “cognitive science,” “cognitive psychology,” “human cognition,” “neural language models,” “language processing,” and “artificial intelligence.” We focused on publications from 2020 to 2025, to capture the rapid developments in LLMs since the emergence of transformer-based models.

Inclusion criteria

We included (1) peer-reviewed journal articles published in recognized journals in cognitive science, psychology, neuroscience, and AI; (2) proceedings from top-tier conferences (such as NeurIPS, EMNLP, ECAI, ACL, and AAAI); (3) preprints from arXiv that have gained substantial attention in the research community (as measured by citation count and community engagement); and (4) studies explicitly addressing the relationship between LLMs and human cognitive processes.

Exclusion criteria

We excluded (1) studies focusing solely on technical improvements in LLMs without cognitive science implications; (2) non-English publications; (3) studies that did not provide empirical evidence or theoretical analysis relevant to LLM-cognition comparisons; and (4) duplicate publications or earlier versions of subsequently published work.

Data extraction and synthesis

From each of the included studies, we extracted information regarding: the specific LLMs evaluated, cognitive domains addressed, experimental paradigms used, quantitative performance metrics (where available), and main conclusions regarding LLM-human cognitive alignment. This systematic approach yielded approximately 55 relevant publications that formed the foundation of this review. We synthesized the findings thematically, by organizing the literature around key cognitive domains and methodological approaches to facilitate comparison across studies.

Before proceeding, we must first clarify the key terms used throughout this review. We use **“cognitive model”** to refer to computational systems that can simulate, predict, or explain aspects of human cognitive processes, by following the cognitive science tradition in which models serve as tools for understanding the mind rather than necessarily replicating its mechanisms [3]. The term **“understanding”** is used operationally to denote the ability to appropriately respond to novel inputs in ways that reflect sensitivity to meaning, context, and pragmatic constraints, without making strong claims regarding whether such behavior implies phenomenal experience or “genuine” comprehension in a philosophical sense. Similarly, **“reasoning”** refers to the ability to draw inferences, make predictions, or solve problems that require combining multiple pieces of information in structured ways encompassing both deductive processes (drawing necessary conclusions from premises) and inductive processes (generalizing from specific instances). We acknowledge that these terms carry substantial philosophical weight, and that the question of whether LLMs truly “understand” or “reason” remains deeply contested [4, 5]. Our usage is pragmatic and behavioral: we describe LLMs as exhibiting understanding or reasoning when their outputs meet human expectations for these capacities, whereas we remain agnostic regarding the underlying computational processes that generate such outputs.

Figure 1 provides an overview of the structure and scope of this review, illustrating the key themes and their interconnections.

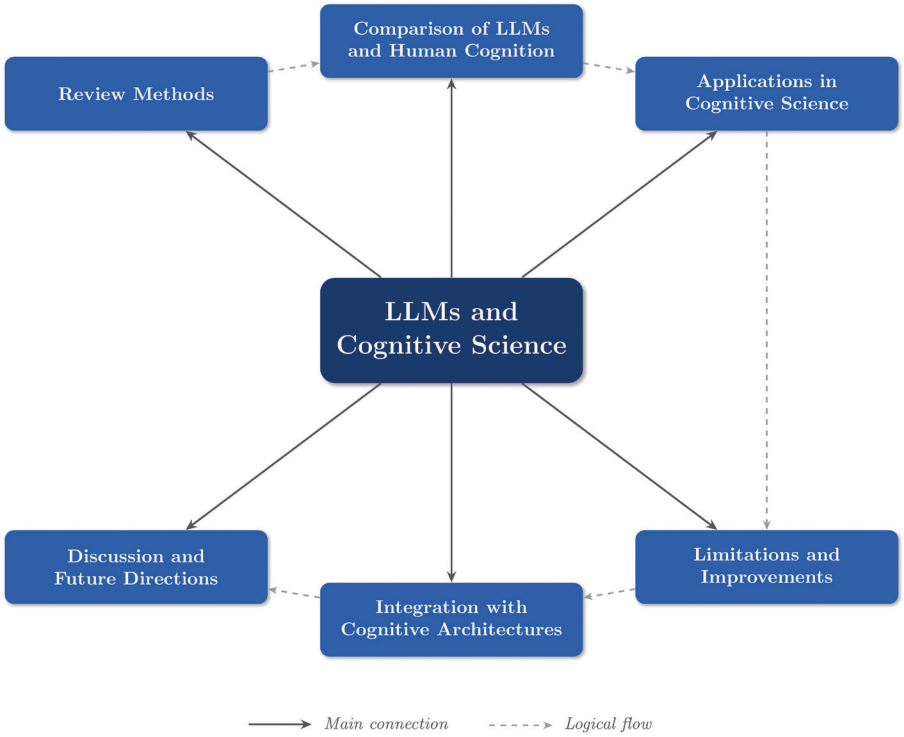

Figure 1 Overview of the review framework. The central theme explores the intersection of large language models and cognitive science, with arrows indicating the main topics covered and dashed lines showing the logical flow of the review.

Comparison between LLMs and human cognitive processes

LLMs have revolutionized understanding of AI and its potential to mimic human cognitive processes. These models have demonstrated capabilities that resemble human cognition in tasks including language processing, sensory judgments, and reasoning. However, despite these similarities, fundamental differences exist between LLMs and human cognitive processes and merit close examination. This section explores these similarities and differences, evaluates the methods used to assess LLMs' cognitive abilities, and discusses the potential of LLMs as cognitive models. Comparing LLMs with human cognition enables understanding of the strengths and limitations of these models in emulating human thought processes.

Similarities and differences between LLMs and human cognitive processes

LLMs have demonstrated remarkable capabilities in various cognitive tasks, often including human-like behaviors and performance. One key similarity is in the domain of language processing. LLMs can achieve human-level word prediction performance in natural contexts, thus suggesting a deep connection between these models and human language processing. Specifically, in a study by Goldstein et al. [6], GPT-2 achieved a correlation of $r = 0.79$ with human cloze probabilities in a naturalistic story-listening paradigm, thereby matching the inter-subject reliability of human responses. LLMs have been demonstrated to represent linguistic information similarly to humans, thereby enabling accurate brain encoding and decoding during language processing [7]. This similarity extends to the neural level, wherein larger neural language models exhibit representations that are increasingly similar to neural response measurements from brain imaging [8].

LLMs also demonstrate human-like cognitive effects in certain tasks. For example, GPT-3 exhibits priming, distance, Spatial-Numerical Association of Response Codes (wherein smaller numbers are associated with left-sided responses and larger numbers are associated with right-sided responses), and size congruity effects, which are well-documented phenomena in human cognition [9]. Additionally, LLMs show content effects in logical reasoning tasks similar to those in humans, particularly in challenging tasks such as syllogism validity judgments and the Wason selection task [10]. LLMs have been shown to capture aspects of human sensory judgments across multiple modalities. Marjieh et al. [11] have demonstrated that similarity judgments from GPT models correlate with human data across six sensory modalities: pitch, loudness, colors, consonants, taste, and timbre. Therefore, LLMs can extract perceptual information from language alone.

However, differences also exist between LLMs and human cognitive processes. Humans generally outperform LLMs in reasoning tasks, particularly with out-of-distribution prompts, and demonstrate greater robustness and flexibility [12]. In contrast, LLMs struggle to emulate human-like reasoning when faced with novel and constrained problems, thus indicating limitations in their ability to generalize beyond their training data. Lamprinidis [13] has found that LLMs' cognitive judgments are not human-like in limited-data inductive reasoning tasks and show higher errors than Bayesian predictors. Therefore, LLMs might not model the basic statistical principles that humans use in everyday scenarios as effectively as previously believed.

Recent developments in LLM architecture have begun to address these reasoning limitations through approaches that parallel the distinction between **system 1** (fast, intuitive) and **system 2** (slow, deliberate) thinking in human cognition [14]. Traditional LLMs operate predominantly in a system 1-like mode, by generating responses through rapid pattern matching without explicit deliberation. In September 2024, OpenAI introduced the o1 model series specifically designed to engage in extended "thinking" before responding [15]. The o1 model uses reinforcement learning to develop internal chains of thought, spends more time on difficult problems, and exhibits behaviors such as self-correction and strategy exploration. Empirical evaluations have demonstrated remarkable improvements. In challenging mathematics examinations, o1-preview achieved near-perfect scores (76/76 and 74/76), substantially outperformed GPT-4o (which scored 66 and 62), and exceeded the performance of nearly all human test-takers [16]. This "test-time compute" approach, wherein models allocate more computational resources during inference for harder problems, represents a qualitative shift in LLM capabilities and has major implications for cognitive science. The findings suggest that system 2-like deliberative reasoning can emerge from training objectives that reward extended thinking, and offer potential insights into how humans might develop and deploy deliberative strategies. However, whether o1's internal reasoning processes truly parallel human system 2 cognition or merely simulate its behavioral signatures through fundamentally different mechanisms remains an important open question.

Prior seemingly contradictory findings between Marjieh et al. [11], who showed human-like sensory judgments, and Lamprinidis [13], who demonstrated non-human-like inductive reasoning, may be reconciled by considering the nature of the tasks involved. Marjieh et al. [11] focused on perceptual similarity judgments, which might rely more heavily on statistical regularities in language that LLMs can effectively capture. In contrast, Lamprinidis [13] examined limited-data inductive reasoning, which requires extrapolation beyond observed patterns—a capability that might critically depend on structured prior knowledge that humans possess but LLMs lack. This distinction suggests that LLMs' alignment with human cognition is domain-specific: LLMs excel in tasks that can be solved through pattern matching over large corpora but struggle when tasks require the flexible, hypothesis-driven reasoning characterizing human inductive inference.

Moreover, although LLMs exhibit near-human-level formal linguistic competence, they show patchy performance in functional linguistic competence [17]. Therefore, LLMs may excel in surface-level language processing yet struggle with deeper, context-dependent understanding and reasoning. Another notable difference is in the memory properties of LLMs compared with human memory. Although LLMs exhibit some human-like memory characteristics, such as primacy and recency effects, their forgetting mechanisms and memory structures differ from human biological memory [18]. Suresh et al. [19] have found that human conceptual structures are robust and coherent across tasks, languages, and cultures, whereas LLMs produce conceptual structures that vary depending on the task used to generate responses. This finding highlights a fundamental difference in the stability and consistency of conceptual representations between humans and LLMs.

A deeper theoretical issue underlying these differences is the **symbol grounding problem** originally articulated by Harnad [20]. The problem asks: How can the meanings of symbols in a formal system be grounded in something other than more symbols? For humans, linguistic symbols are grounded in sensorimotor experiences. For example, the word "red" is connected to actual experiences of seeing red objects, feeling warmth, and other embodied interactions with the world. In contrast, LLMs learn meanings purely from statistical co-occurrence patterns in text corpora; the meaning of "red" for an LLM is constituted by its distributional relationships with other words (e.g., "color," "blood," or "stop sign") rather than any perceptual experience. This aspect prompts a fundamental question regarding whether LLMs can genuinely "understand" language or merely manipulate symbols according to learned patterns, thus echoing Searle's Chinese room argument [5]. Some researchers argue that the rich statistical structure of language itself might provide sufficient grounding for many cognitive tasks, because linguistic descriptions encode distilled knowledge of physical and social interactions accumulated across human cultures [21]. Others maintain that truly human-like cognition requires embodied grounding that text-only models cannot achieve. The emergence of multimodal LLMs that process both text and images offers a partial bridge with potential to ground linguistic representations in visual experience. However, whether visual grounding alone, without the ability to act upon and interact with the physical world, can fully address the symbol grounding problem remains an open question with extensive implications regarding the validity of LLMs as cognitive models.

Figure 2 provides a systematic visual comparison of LLMs and human cognition, highlighting the unique characteristics of each system alongside their shared behavioral patterns. Synthesis of the findings revealed several critical research gaps. First, although LLMs have demonstrated remarkable performance on standardized cognitive tasks, their success often heavily depends on task format and prompt engineering; therefore, the robustness of these capabilities has been called into question. Second, the field lacks systematic studies directly comparing LLMs and humans in identical experimental paradigms under matched conditions. Third, most existing comparisons have focused on behavioral outcomes rather than underlying processes, thus leaving open the question of whether similar performance implies similar cognitive mechanisms. These gaps highlight the need for more rigorous, process-oriented approaches to LLM-human comparisons that transcend surface-level behavioral matching.

Methods for evaluating LLMs' cognitive abilities

Researchers have developed multiple methods to evaluate the cognitive abilities of LLMs, often by drawing inspiration from cognitive science and psychology. These methods are aimed at comprehensively assessing LLMs' capabilities and limitations in comparison to human cognition (**Figure 3**).

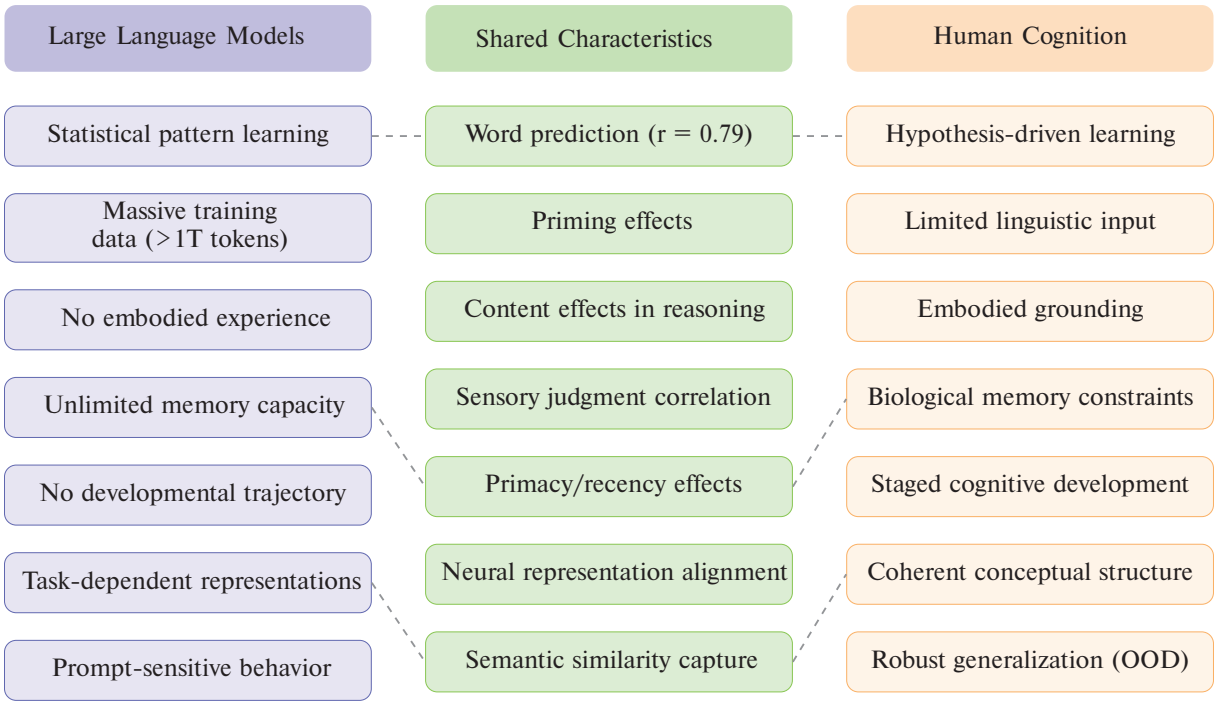

Figure 2 Systematic comparison of large language models and human cognition across key dimensions. The left column (blue) shows characteristics unique to LLMs, including statistical learning from massive datasets and lack of embodied experience. The right column (orange) shows characteristics unique to human cognition, such as hypothesis-driven learning and robust out-of-distribution generalization. The center column (green) highlights shared behavioral patterns wherein both systems demonstrate similar capabilities, including word prediction accuracy (r = 0.79), priming effects, and neural representation alignment. Dashed lines connect conceptually related aspects, illustrating how different underlying mechanisms can produce similar behavioral outputs.

One prominent approach uses cognitive psychology experiments adapted for LLMs. For example, CogBench, a benchmark comprising ten behavioral metrics from seven cognitive psychology experiments, has been developed to systematically compare LLM performance across various cognitive tasks [22]. Another method uses neuroimaging data to compare LLMs representations with human brain activity. Functional magnetic resonance imaging (fMRI) and magnetoencephalography (MEG) recordings have been applied to analyze the similarities between LLM activations and brain responses during language processing tasks [23]. This approach provides insights into the neural-level similarities and differences between LLMs and human cognition.

Researchers have also adapted traditional psychological tests for use with LLMs. For example, cognitive reflection tests and semantic illusions have been used to evaluate the reasoning capabilities of LLMs [24]. These tests help reveal the extent to which LLMs exhibit human-like biases and reasoning patterns. Additionally, methods from developmental psychology have been proposed to understand the capacities and underlying abstractions of LLMs [25]. These approaches focus on testing generalization to novel situations and using simplified stimuli to probe underlying abstractions.

To provide more comprehensive evaluation tools, Zhang et al. [26] have introduced MulCogBench, a multi-modal cognitive benchmark dataset for evaluating Chinese and English computational language models. This dataset includes various types of cognitive data, such as subjective semantic ratings, eye-tracking, fMRI, and MEG, thus enabling comprehensive comparison between LLMs and human cognitive processes. Ivanova [27] has provided a set of methodological considerations for evaluating the cognitive abilities of LLMs by using language-based assessments. The article highlights common pitfalls and provides guidelines for designing high-quality cognitive evaluations, to contribute to best practices in AI psychology.

To further examine specific cognitive abilities, Srinivasan et al. [28] have proposed novel methods based on cognitive science principles to test LLMs' common sense reasoning abilities through prototype analysis and proverb understanding. These methods offer new ways to assess LLMs' cognitive capabilities in more nuanced and context-dependent tasks. Binz and Schulz [29] have used tools from cognitive psychology to study GPT-3, including its decision-making, information search, deliberation, and causal reasoning abilities. Their approach has demonstrated the potential of cognitive psychology in studying AI and demystifying how LLMs solve tasks.

From a neuroscientific perspective, the convergence between LLM representations and brain activity patterns offers exciting possibilities for both basic research and clinical applications. Studies using fMRI and MEG have demonstrated that LLM activations can predict neural responses in language-related brain regions with remarkable accuracy, thereby suggesting shared computational principles between artificial and biological language processing [23, 30]. These findings have major implications for developing brain-computer interfaces, because LLMs

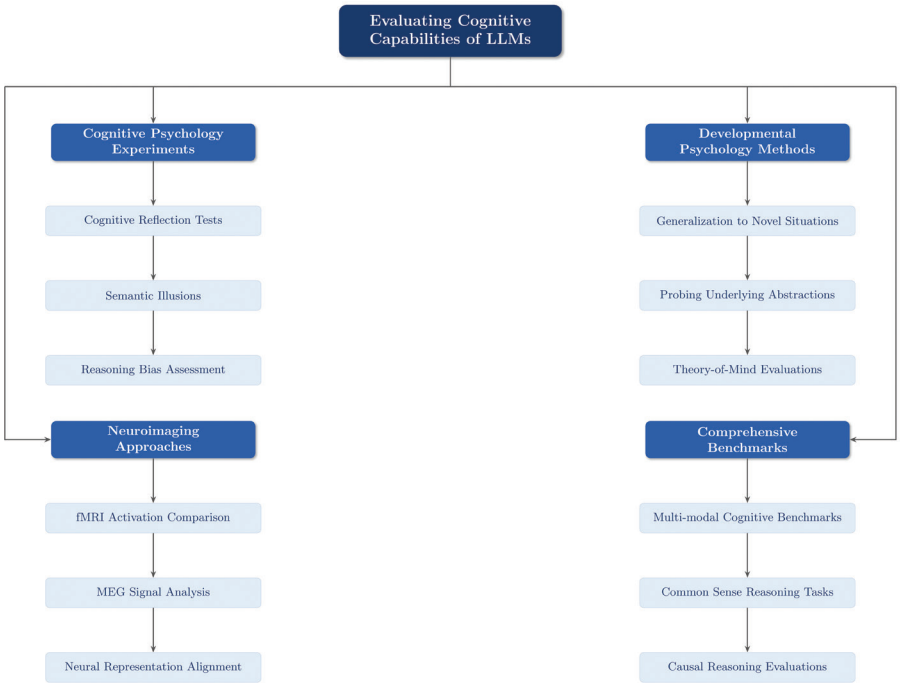

Figure 3 Evaluating the cognitive capabilities of LLMs.

Table 1 Summary of Methods for Evaluating LLMs' Cognitive Abilities with Performance Metrics

Cognitive Domain	Benchmark/Method	Key Studies	Performance Metrics	Main Findings
Language Processing	Brain encoding/decoding, fMRI/MEG	Goldstein et al., Tuckute et al., Caucheteux and King [6, 7, 23]	$r = 0.79$ (word prediction correlation)	Human-level word prediction, neural alignment
Reasoning	CogBench, syllogism tasks, Wason selection	Coda-Forno et al. [22], Dasgupta et al. [10]	60–85% accuracy (task-dependent)	Content effects similar to those in humans
Causal Reasoning	Tübingen benchmark, counterfactual tasks	Kıcıman et al., Liu et al. [55, 54]	97% (pairwise); 92% (counterfactual)	Outperforms statistical methods
Memory	Primacy/recency tests	Janik [18]	Primacy/recency effects detected	Similar effects, different mechanisms
Sensory Judgments	Multi-modal similarity ratings	Marjieh et al. [11]	Significant correlations (6 modalities)	Perceptual info from language
Theory of Mind	Theory-of-mind task variations	Ullman [60]	Near 0% on alterations	Fragile, fail on variations
Common Sense	Prototype analysis, proverbs	Srinivasan et al. [28]	Context dependent	Variable across contexts

might potentially serve as decoders for neural signals in patients with communication disorders. Furthermore, the ability to systematically manipulate LLM architectures and training data provides a unique experimental platform for testing hypotheses regarding human neural organization that would be impossible to examine directly in the brain. **Table 1** summarizes the key methods and benchmarks used to evaluate LLMs' cognitive ability across various domains.

Despite their utility, current evaluation methods have important limitations that warrant caution in interpreting the findings. First, behavioral benchmarks may overestimate cognitive alignment: LLMs can achieve human-like performance through fundamentally different computational strategies, in a phenomenon known as "shortcut learning" [31]. Second, neuroimaging comparisons face the "correlation is not causation" problem, wherein representational similarity does not guarantee mechanistic similarity. Third, most evaluations rely on static, text-based tasks that do not capture the dynamic, embodied, and interactive nature of human cognition. Fourth, performance is often highly sensitive to prompt wording and task format, thus leading to questions regarding the robustness and generalizability of the findings. Finally, the lack of standardized evaluation protocols hinders cross-study comparisons. Given these limitations, current evidence of LLM-human cognitive alignment should be interpreted as suggestive rather than conclusive, and future research should prioritize process-level comparisons over purely behavioral benchmarks.

Architectural considerations

Different LLM architectures exhibit distinct cognitive properties that merit consideration. Decoder-only models (e.g., GPT series) excel in generative tasks and exhibit strong in-context learning, but their unidirectional attention might limit their ability to model bidirectional linguistic dependencies. Encoder-only models (e.g., BERT) show superior performance in tasks requiring deep bidirectional understanding, such as semantic similarity and natural language inference, but they lack generative capabilities. Encoder-decoder models (e.g., T5, BART) combine both capabilities but increase computational costs.

From a cognitive perspective, these architectural differences have important implications. The success of decoder-only models in many cognitive tasks suggests that predictive processing, i.e., generating expectations regarding upcoming input, might be a more fundamental cognitive mechanism than previously appreciated. This possibility aligns with predictive coding theories in neuroscience suggesting that the brain continually generates predictions regarding sensory input. In contrast, the superior performance of bidirectional models in certain semantic tasks aligns with theories emphasizing the importance of global context integration in human language understanding. The observation that different architectures excel in different cognitive tasks suggests that human cognition might similarly use distinct computational strategies for different types of processing.

Future research should systematically compare how architectural choices affect alignment with human cognitive processes across domains. Such comparisons might reveal whether certain architectures better capture specific aspects of human cognition, and potentially inform both AI development and understanding of the computational principles underlying human cognitive abilities.

Scaling laws and emergent cognition

A striking feature of LLMs is that their capabilities follow predictable **scaling laws**: model performance improves as a power-law function of model size, dataset size, and compute [32]. This quantitative relationship has substantial implications for understanding emergent cognitive abilities. As models scale, they exhibit qualitative transitions, by suddenly acquiring capabilities (e.g., arithmetic, multi-step reasoning, or in-context learning) that were absent or unreliable at smaller scales. This phenomenon has cognitive parallels: just as certain human cognitive abilities emerge at specific developmental stages, LLM capabilities appear to require crossing of computational thresholds. A particularly intriguing phenomenon is **"grokking,"** wherein neural networks first memorize training data (achieving zero training error) and then, after substantially more training, suddenly generalize to held-out data [33]. Grokking challenges the traditional view of generalization as a smooth, continuous process. Instead, it suggests that learning involves discrete phase transitions in which qualitative reorganization of internal representations enables sudden capability gains. This phenomenon is concordant with observations from developmental psychology, in which children often exhibit sudden "insight" moments and phase transitions in cognitive development (e.g., the transition from pre-operational to concrete operational thought in Piaget's framework). This parallel prompts intriguing questions: might human cognitive development involve similar phase transitions in which accumulated learning suddenly crystallizes into new capabilities? If so, studying grokking in LLMs might provide computational models for understanding these developmental phenomena. However, important differences remain. For example, human development unfolds over years with multimodal embodied experience, whereas grokking occurs over training iterations on narrow algorithmic tasks. Therefore, we caution against overly direct analogies.

Biological implausibility and the backpropagation problem

From a biological perspective, current LLM architectures exhibit several neurally implausible features. The most fundamental issue relates to the **backpropagation algorithm** used to train LLMs. Backpropagation requires three properties absent in biological neural circuits: (1) *weight symmetry*, wherein the forward and backward pathways must use identical (transposed) weight matrices, whereas biological synapses are unidirectional and show no evidence of such symmetry; (2) *global error signals*, wherein gradients must be computed precisely and propagated backward through the entire network, whereas biological neurons have access only to local information regarding their inputs and outputs; and (3) *non-local credit assignment*, wherein each synapse must know its contribution to the global error, a computation that appears to require non-local information inaccessible to individual neurons. These observations have motivated extensive research in biologically plausible learning algorithms. *Feedback alignment* demonstrates that random, fixed backward weights can still support learning, thus relaxing the weight symmetry requirement. *Predictive coding* frameworks have suggested that the brain implements a form of gradient descent through local prediction error minimization, thus potentially bridging the gap between backpropagation and biological learning. *Hebbian learning* (the idea that "neurons that fire together wire together") captures local, activity-dependent plasticity but struggles to explain how deep networks can learn complex representations without error signals. Despite these advances, no fully satisfactory biologically plausible alternative to backpropagation with comparable effectiveness for training deep networks has emerged. Second, transformer attention mechanisms compute global, all-to-all interactions at each layer, whereas biological neural networks exhibit sparse, locally structured connectivity with modularity and hierarchical organization. Third, LLMs process information in discrete, feedforward passes, whereas the brain heavily relies on recurrent processing, temporal dynamics, and continuous-time computation. Fourth, LLMs lack the neural modulation mechanisms (e.g., dopaminergic and cholinergic systems) that regulate attention, learning, and memory consolidation in biological brains.

These architectural differences have important implications in cognitive science. If LLMs achieve human-like behavioral outputs through fundamentally non-biological computational strategies, then the insights derived from studying LLM internals might not generalize to biological cognition. In contrast, the success of backpropagation-trained networks on cognitive tasks might suggest that the brain has evolved mechanisms that approximate gradient-based optimization, even if the implementation differs. This hypothesis remains actively debated in computational neuroscience.

Applications of LLMs in cognitive science

The integration of LLMs into cognitive science research has opened new avenues for understanding human cognition and developing more sophisticated AI systems. This section explores the multifaceted applications of LLMs in cognitive science, examining their roles as cognitive models, their contributions to theoretical insights, and their specific applications in various cognitive domains. By synthesizing recent research, we aim to provide a comprehensive overview of the current state and future potential of LLMs in advancing understanding of human cognition.

LLMs as cognitive models

The potential of LLMs to serve as cognitive models has gained substantial attention in recent research. Studies have demonstrated that LLMs can be turned into accurate cognitive models through fine-tuning on psychological experiment data, thus offering precise representations of human behavior and often outperforming traditional cognitive models in decision-making tasks [34]. These models have shown promise in capturing individual differences in behavior and generalizing to new tasks after being fine-tuned on multiple tasks; therefore, they might potentially become generalist cognitive models capable of representing a wide range of human cognitive processes. The versatility of LLMs in various cognitive domains has been explored. Wong et al. [35] have introduced a computational framework called rational meaning construction, which integrates neural language models with probabilistic models for rational inference. This approach demonstrates LLMs' ability to generate context-sensitive translations and support commonsense reasoning across various cognitive domains. Piantadosi and Hill [21] have highlighted LLMs' ability to capture essential aspects of meaning through conceptual roles, thereby challenging skepticism regarding their ability to possess human-like concepts.

In the realm of language processing, a systematic integrative modeling study by Schrimpf et al. [30] has revealed that transformer-based ANN models can predict neural and behavioral responses in human language processing. Their findings support the hypothesis that predictive processing shapes language comprehension mechanisms in the brain, in alignment with contemporary theories in cognitive neuroscience. Contreras Kallens et al. [36] have demonstrated that LLMs can produce human-like grammatical language without an innate grammar. That study has provided valuable computational models for exploring statistical learning in language acquisition and challenging traditional views of language learning. Lampinen's [37] research has further challenged understanding of human language processing, by demonstrating that, with minimal prompting, LLMs can outperform humans in processing recursively nested grammatical structures. These findings prompt questions regarding the cognitive mechanisms underlying both human and artificial language comprehension. Nolfi [38] has explored the unexpected cognitive abilities developed by LLMs through indirect processes, including dynamical semantic operations, theory of mind, affordance recognition, and logical reasoning. These findings suggest that LLMs can develop integrated cognitive skills that work synergistically, despite being trained primarily on next-word prediction tasks. This research highlights the importance of understanding these emergent capabilities in relation to human cognition. Sartori and Orrú [39] have provided empirical evidence that LLMs perform at human levels in a wide variety of cognitive tasks, including reasoning and problem-solving. Their findings support associationism as a unifying theory of cognition and demonstrate the potential for substantial effects on cognitive psychology, thus suggesting new avenues for modeling human cognitive processes. Li and Li [40] have proposed an intriguing duality between LLMs and Tulving's theory of memory, and suggested that consciousness might be an emergent ability based on this duality. This perspective offers a novel approach to understanding the relationship between LLMs and human cognition, and may potentially bridge artificial and biological intelligence research.

However, several researchers have urged caution against overstating LLMs' cognitive modeling capabilities. The critical question is whether LLMs achieve human-like performance through human-like processes or through alternative computational strategies that produce similar behavioral outputs. This distinction has extensive implications regarding their utility in cognitive science research. Pavlick [41] has argued that definitive claims regarding the abilities or limitations of LLMs as models of human language understanding are premature, and has emphasized a need for further empirical testing that distinguishes between superficial behavioral similarity and deeper cognitive alignment. Katzir [42] has provided a balanced assessment emphasizing that LLMs' opacity, massive data requirements (far exceeding human linguistic input during development), and fundamentally different learning mechanisms from human development prompt questions regarding their validity as cognitive models. As Katzir has noted, humans acquire language from limited, noisy input through active interaction with the environment, whereas LLMs are trained on curated text corpora orders of magnitude larger. This difference might render direct comparisons misleading. Furthermore, the use of LLMs as cognitive models offers new opportunities for understanding human cognition. By analyzing the internal representations and processes of these models, researchers can gain insights into potential mechanisms underlying human cognitive

abilities. However, caution is necessary in interpreting these findings, because the fundamental differences in architecture and learning processes between LLMs and the human brain must be considered. Ren et al. [43] have investigated how well LLMs align with human brain cognitive processing signals by using representational similarity analysis. Their findings suggest that factors such as pre-training data size, model scaling, and alignment training affect the similarity between LLMs and brain activity, thereby providing insights into how LLMs might be improved to better model human cognition.

A fundamental methodological concern arises when LLMs are considered as explanatory models of cognition: can an opaque system with billions of parameters genuinely explain another complex system, the human mind? This concern echoes debates in the philosophy of science regarding the nature of explanation. Traditional explanatory models in cognitive science are valued for their interpretability: they identify specific mechanisms (e.g., working memory buffers or attention filters) that can be independently tested. LLMs, in contrast, operate as "black boxes" whose internal computations resist straightforward interpretation. Using one inexplicable system to explain another prompts questions regarding what might be called the "opacity problem": even if an LLM perfectly predicts human behavior, this predictive success might not constitute genuine understanding of the underlying cognitive mechanisms. Some researchers argue that LLMs should be viewed not as mechanistic explanations but as "phenomenological models" that capture input-output relationships without claiming to replicate internal processes [44]. Others suggest that interpretability techniques, such as attention analysis and probing classifiers, can partially address the opacity problem by revealing how LLMs represent and process information. Nevertheless, the tension between predictive power and mechanistic insight remains a central challenge in using LLMs in cognitive science research.

This tension can be framed through the lens of the debate between **computational functionalism** and **mechanistic explanation**. Computational functionalism, rooted in classical philosophy of mind, holds that cognitive processes are defined by their functional roles, i.e., the input-output mappings that they implement, rather than by the specific physical substrate that realizes them [45]. Under this view, if an LLM implements the same computational function as the human cognitive system (e.g., predicting upcoming words and drawing inferences from context), it can be considered a valid model of cognition regardless of whether it uses biologically plausible mechanisms. Marr's influential tri-level framework distinguishes among the computational level (what problem is being solved), the algorithmic level (how it is solved), and the implementational level (what physical substrate realizes the algorithm). LLMs may be valid models at the computational level even if they diverge from biological brains at the implementational level. This perspective is supported by the observation that decoder-only LLMs implementing next-token prediction align remarkably well with predictive coding theories in neuroscience that propose that the brain continually generates predictions about sensory input and updates internal models according to prediction errors [30]. If predictive processing is a core computational principle of cognition, then LLMs might capture this principle despite lacking neural implementation details. However, proponents of a mechanistic explanation argue that genuine scientific understanding requires identifying the causal mechanisms that produce observed phenomena. From this perspective, behavioral equivalence is insufficient; instead, a true cognitive model must replicate the processes that generate behavior, not merely mimic their outputs. This debate has several practical implications: if LLMs achieve human-like behavior through fundamentally different computational strategies, interventions that improve LLM performance might not generalize to human cognition, and vice versa. Resolving this tension would require developing evaluation paradigms that probe not only behavioral outcomes but also the underlying processes—an area in which interpretability research and cognitive neuroscience can productively converge.

Comparison across these studies revealed a fundamental tension in using LLMs as cognitive models. On the one hand, their ability to predict human behavioral data and neural responses suggests that they capture meaningful aspects of human cognition. On the other hand, their architectural differences from biological neural networks, including the absence of recurrent processing, embodied experience, and developmental learning trajectories, prompt questions regarding the depth of this alignment. This tension suggests that LLMs might be best understood not as literal models of human cognition, but as "cognitive prostheses" that approximate certain input-output relationships without necessarily replicating underlying mechanisms. Future research should focus on identifying which aspects of cognition LLMs genuinely model versus merely simulate through alternative computational strategies.

Insights from LLMs for cognitive science research

LLMs have provided valuable insights for cognitive science research that have challenged existing theories and offered new perspectives on human cognition. Veres [46] has argued that although LLMs challenge rule-based theories, they do not necessarily provide deeper insights into the nature of language or cognition. This perspective highlights the need for careful interpretation of LLMs capabilities in the context of cognitive science and cautions against overinterpretation of model performance. Shanahan [47] has emphasized the importance of understanding the true nature and capabilities of LLMs to avoid anthropomorphism and ensure responsible use and discourse around AI in cognitive science research. This cautionary approach underscores the need for precise language and philosophical nuance in AI discourse, particularly in drawing parallels between artificial and human cognition. Blank [44] has explored whether LLMs might be considered computational models of human language processing, and discussed interpretations and implications for future research. This work highlights the ongoing debate regarding whether LLMs process language similarly to humans, underscores the importance of this question for cognitive science, and emphasizes the need for rigorous

empirical investigation. Grindrod [48] has argued that LLMs can serve as scientific models of external languages and provide insights into the nature of language as a social entity. This perspective offers a novel approach to using LLMs in linguistic inquiry and cognitive science research, thereby potentially bridging computational linguistics and sociolinguistics.

The application of LLMs in cognitive science research has opened new avenues for exploring human behavior and decision-making processes. Horton [49] has demonstrated LLMs' potential as simulated economic agents to replicate classic behavioral economics experiments. This innovative approach suggests new possibilities for using LLMs to explore human behavior and decision-making processes in cognitive science, and it offers a cost-effective method for piloting studies and generating hypotheses. Connell and Lynott [50] have evaluated the cognitive plausibility of different types of language models, and emphasized the importance of learning mechanisms, corpus size, and grounding in assessing their relevance to human cognition. Their work provides a framework for critically evaluating the applicability of LLMs to cognitive modeling. Mitchell and Krakauer [51] have surveyed the debate on whether LLMs understand language in a human-like manner and advocate for an extended science of intelligence to explore diverse modes of cognition. This perspective highlights the need for a broader understanding of intelligence and cognition in the context of LLMs, to encourage interdisciplinary collaboration in AI and cognitive science research. Buttrick [52] has proposed using LLMs to study cultural distinctions by analyzing the statistical regularities in their training data, thus offering new avenues for exploring cultural cognition and representation. This approach demonstrates the potential of LLMs as tools for investigating complex sociocultural phenomena in cognitive science. Finally, Demszky et al. [53] have reviewed the potential of LLMs to transform psychology by enabling large-scale analysis and generation of language data. They have emphasized a need for further research and development to address ethical concerns and harness the full potential of LLMs in psychological research, and have highlighted both the opportunities and challenges in this emerging field.

Applications of LLMs in specific cognitive fields

LLMs have demonstrated substantial potential in various cognitive domains, including causal reasoning, lexical semantics, and creative writing. Liu et al. [54] have conducted a comprehensive survey exploring the mutual benefits between LLMs and causal inference, and have highlighted how causal perspectives might enhance LLMs' reasoning ability, fairness, and safety. Similarly, Kıcıman et al. [55] have benchmarked the causal capabilities of LLMs and found that they outperform existing methods in generating causal arguments across various tasks. Specifically, GPT-4 achieved 97% accuracy in pairwise causal discovery tasks from the Tübingen benchmark and outperformed traditional statistical methods. However, the performance decreased in tasks requiring multi-step causal reasoning or counterfactual inference. These findings highlight the boundaries of current capabilities and their limitations in critical decision-making scenarios. In the field of lexical semantics, Petersen and Potts [56] have used LLMs to conduct a detailed case study of the English verb "break," and demonstrated that LLM representations can capture known sense distinctions and identify new sense combinations. Their findings prompt reconsideration of the commitment to discreteness in semantic theory, in favor of a more fluid, usage-based approach. In creative domains, Chakrabarty et al. [57] have investigated the utility of LLMs in assisting professional writers. This empirical user study indicated that writers found LLMs most helpful for translation and review tasks rather than planning, and identified weaknesses in current models, such as a reliance on clichés and a lack of nuance.

These findings collectively underscore the diverse applications of LLMs in cognitive fields, including enhancing causal reasoning and supporting creative processes, while also highlighting areas for improvement and future research directions. Overall, the application of LLMs in cognitive science research has achieved major advancements in the ability to model and understand human cognition. LLMs have shown remarkable potential as cognitive models, by offering insights into language processing, reasoning, and decision-making that challenge and expand existing theories. Their versatility in addressing cognitive tasks as diverse as causal inference and creative writing underscores their value as research tools across multiple domains of cognitive science.

However, the integration of LLMs into cognitive research is not free from challenges. Researchers must navigate issues of interpretability, ethical considerations, and the potential for overinterpretation of model capabilities. The ongoing debate regarding the nature of LLM "understanding" and its relationship to human cognition highlights the need for continued critical examination and empirical investigation.

Biomedical and clinical implications

The intersection of LLMs and cognitive science has particular promise in biomedical applications. In clinical neuropsychology, LLMs offer novel tools for assessing cognitive function. By comparing patient responses to LLM-generated baselines, clinicians may gain new insights into the nature and extent of cognitive impairments in conditions such as aphasia, dementia, and traumatic brain injury. The standardized nature of LLM outputs might provide more reliable comparison points than traditional normative data, which often have demographic biases.

In neuroimaging research, LLMs are increasingly being used as computational models to interpret brain activity patterns. The alignment between LLM layer activations and hierarchical processing in the human cortex [23] suggests that these models capture aspects of the brain's language processing architecture. Consequently, LLM-based encoding models have been developed to predict neural responses to novel stimuli, thus potentially accelerating understanding of how meaning is represented in the brain.

Furthermore, LLMs provide opportunities for developing assistive technologies for individuals with cognitive or

communicative disabilities. Language models might power more naturalistic augmentative and alternative communication devices, predict communication needs based on context, or provide real-time language support for individuals with processing difficulties. However, achieving these clinical applications will require careful validation against human behavioral and neural data, as well as consideration of the ethical implications of deploying AI systems in healthcare settings.

Recent surveys have begun to systematically examine the application of LLMs and deep learning in specific biomedical domains. For example, comprehensive reviews on advanced deep learning and large language models for cancer detection have highlighted the potential of these models to assist in diagnostic tasks across multiple cancer types, including skin, brain, lung, breast, and colorectal cancers [58]. These domain-specific surveys have provided valuable insights into practical considerations, including data privacy, model scalability, and clinical validation requirements for deploying LLMs in healthcare settings, thus complementing the cognitive science perspective presented in this review.

The emergence of multimodal LLMs (e.g., GPT-4V and Gemini) that can process both text and images has opened new frontiers in cognitive science research. These models enable investigation of cross-modal integration, a fundamental aspect of human cognition that unimodal language models cannot address. Preliminary studies have suggested that multimodal LLMs can perform visual reasoning tasks, interpret diagrams, and generate image descriptions that align with human judgments. From a neuroscientific perspective, comparing multimodal LLM representations with brain activity during audiovisual language processing might reveal principles of cross-modal binding. Clinically, multimodal models might enhance diagnostic capabilities by integrating verbal responses with visual information (e.g., analyzing facial expressions during cognitive assessments). However, the cognitive mechanisms underlying multimodal integration in these models remain poorly understood and are an important area for future investigation.

The question of whether embodiment is necessary for robust cognition remains under contention. Current LLMs, despite their linguistic sophistication, lack direct interaction with the physical world: they cannot perceive, act upon, or learn from real-world environments. This limitation might be fundamental: embodied cognition theories suggest that human concepts are grounded in sensorimotor experience, and abstract reasoning emerges from metaphorical extensions of bodily interactions. If true, purely text-based LLMs might be inherently limited in their ability to fully model human cognition. Recent developments in robotics and embodied AI suggest a path forward: systems that combine language models with physical sensors and actuators might potentially learn richer world models through active exploration. However, embodied systems face their own challenges, including sample inefficiency and safety concerns. An alternative perspective is that sufficiently rich linguistic descriptions might substitute for direct experience; i.e., language itself encodes distilled knowledge of physical and social interactions. Resolving this debate will require systematic comparisons between embodied and disembodied systems on tasks requiring physical reasoning and world knowledge, and will be a crucial direction for future research at the intersection of LLMs and cognitive science.

Embodied cognition and future LLM architecture evolution

The insights from embodied cognition research have substantial implications regarding the future evolution of LLM architectures. If the symbol grounding problem [20] is a fundamental limitation, then next-generation cognitive AI systems may need to incorporate the following: (1) *multimodal perception integration*, by moving beyond text-image models to systems that process proprioceptive, tactile, and vestibular information, thus enabling richer representations of physical concepts such as weight, texture, and spatial orientation; (2) *active perception and exploration* architectures that can direct attention, seek information, and test hypotheses through interaction with environments (real or simulated), rather than passively processing pre-collected corpora; (3) *world models and mental simulation* that serve as internal models predicting the consequences of actions before execution, thereby enabling planning, counterfactual reasoning, and imagination—capabilities in which current LLMs struggle; and (4) *developmental learning trajectories*, which acquire knowledge progressively through interaction, and may potentially recapitulate aspects of human cognitive development rather than learning from massive static datasets. These architectural directions align with LeCun's JEPA framework [59], which emphasizes learning predictive world models in abstract representation spaces. From a cognitive science perspective, such embodied architectures might provide better models of human cognition by grounding symbols in sensorimotor experience, thereby supporting active inference and exhibiting developmental plasticity. However, several major challenges remain, including scaling embodied learning to the complexity of human environments, ensuring safe exploration, and developing evaluation methods that assess genuine understanding rather than statistical mimicry. The convergence of LLM technology with robotics, virtual reality, and neuroscience may ultimately yield AI systems that more faithfully capture the embodied, embedded, enacted, and extended nature of human cognition, which serve as the basis for the influential "4E" framework in contemporary cognitive science.

These biomedical applications underscore the importance of continued research at the interface of LLMs, cognitive science, and clinical neuroscience. By grounding LLM development in neuroscientific principles and validating models against clinical populations, researchers can ensure that advances in AI translate to meaningful improvements in human health and well-being.

Limitations and improvements in LLMs capabilities

The rapid advancement of LLMs has necessitated comprehensive evaluation of their capabilities and limitations. This section examines the cognitive biases and constraints inherent in LLMs, as well as proposed methods for enhancing their performance. By critically analyzing these aspects, researchers aim to develop more robust and reliable AI systems that can better emulate human-like cognition and language understanding.

Cognitive biases and limitations of LLMs

Table 2 provides a structured taxonomy of LLM failure modes and corresponding mitigation strategies, synthesizing findings from recent research.

Recent studies have extensively explored the cognitive biases and limitations of LLMs. Ullman [60] has demonstrated that LLMs fail in trivial alterations to theory-of-mind tasks, thus suggesting a lack of robust theory-of-mind capabilities. Talboy and Fuller [61] have identified multiple cognitive biases in LLMs similar to those found in human reasoning, and have highlighted the need for increased awareness and mitigation strategies. Thorstad [62] has advocated for cautious optimism regarding LLM performance while acknowledging genuine biases, particularly framing effects. Singh et al. [63] have investigated the confidence-competence gap in LLMs and revealed instances of overconfidence and underconfidence reminiscent of the Dunning-Kruger effect. Leivada et al. [64] have argued that LLMs currently lack deeper linguistic and cognitive understanding, thus leading to incomplete and biased representations of human language. Macmillan-Scott and Musolesi [65] have evaluated seven LLMs by using cognitive psychology tasks and found that they display irrationality differently from humans and exhibit inconsistency in their responses. Jones and Steinhardt [66] have presented a method inspired by human cognitive biases to systematically identify qualitative errors in LLMs, thus uncovering predictable and high-impact errors. Smith et al. [67] have proposed the term "confabulation" instead of "hallucination" to more accurately describe inaccurate outputs of LLMs, and have emphasized the importance of precise metaphorical language in understanding AI processes.

Recent advances in understanding hallucination mechanisms

The study of LLM hallucinations has markedly advanced as researchers have identified deeper connections to cognitive science concepts. Sui et al. [68] have provided a cognitive re-framing of LLM hallucinations as *confabulations*, a neuropsychology term describing the production of false memories without intent to deceive. Their analysis has revealed that hallucinated outputs display greater levels of narrativity and semantic coherence than veridical outputs, thereby mirroring how human confabulation often serves sense-making functions. This finding suggests that hallucination might not be purely a failure mode but an expression of the model's attempt to maintain narrative coherence when factual knowledge is insufficient. From a detection perspective, Farquhar et al. [69] have developed methods to identify hallucinations with *semantic entropy*, by measuring uncertainty in the meaning space rather than token probabilities. Their approach specifically targets "confabulations" in which LLMs give arbitrary and inconsistent answers to the same question, which are distinct from systematic errors due to training data or reasoning failures. This mechanistic differentiation has important implications: different types of hallucinations may require different mitigation strategies, and understanding the cognitive parallels (such as source amnesia and availability heuristics) has potential to inform more targeted interventions. These advances suggest that studying LLM failure modes through a cognitive lens might yield both better detection methods and deeper insights into the nature of knowledge representation and uncertainty in neural networks.

Understanding the origins of these biases will be crucial for developing effective mitigation strategies. Unlike human cognitive biases, which often arise from evolutionary adaptations and resource-bounded rationality, LLM biases stem primarily from three sources: (1) training data biases, wherein overrepresentation of certain viewpoints or reasoning patterns leads to skewed outputs; (2) architectural limitations, such as a lack of explicit working memory or causal reasoning modules; and (3) optimization objectives, wherein next-token prediction might not align with truthful or calibrated responses. Although some LLM biases superficially resemble human biases (e.g., framing effects), the underlying mechanisms are likely to fundamentally differ. Human framing effects arise from affective and attentional processes, whereas LLM framing effects might simply reflect sensitivity to surface-level statistical patterns in training

Table 2 Taxonomy of LLM Failure Modes and Mitigation Strategies

Failure Category	Specific Limitation	Manifestation	Mitigation Strategy
Reasoning Failures	Theory-of-mind deficits	Failure in theory-of-mind task variations	Robust multi-format evaluation
	Out-of-distribution errors	Poor generalization	Diverse training data, fine-tuning
Cognitive Biases	Framing effects	Response varies with phrasing	Prompt debiasing techniques
	Overconfidence	Dunning-Kruger-like behavior	Calibration training
Output Errors	Confabulation	Plausible but false outputs	Retrieval-augmented generation, fact verification
	Inconsistency	Variable responses to the same input	Temperature control, ensembles

data. This mechanistic distinction has important implications: interventions effective for human debiasing might not transfer to LLMs, and vice versa.

Beyond these specific limitations, some researchers have questioned whether the autoregressive, next-token prediction paradigm underlying current LLMs might be a fundamentally limited approach to AI. LeCun [59] has argued that autoregressive LLMs are inherently constrained by their inability to perform hierarchical planning, reason about the physical world, or learn efficiently from limited data—capabilities that he considers essential for human-level intelligence. As an alternative, he has proposed the Joint Embedding Predictive Architecture (JEPA), which learns by predicting abstract representations of future states rather than predicting raw tokens. Unlike LLMs, JEPA-based systems could potentially learn world models that support planning, causal reasoning, and efficient learning from sparse data, all of which are cognitive capabilities in which current LLMs notably struggle.

This architectural critique has important implications for cognitive science. If autoregressive prediction is fundamentally limited as a cognitive mechanism, then LLMs might represent a local optimum that captures certain aspects of human cognition (e.g., statistical language processing) while being unable to model others (e.g., physical reasoning, planning, or causal inference). This view suggests that LLMs and humans might achieve similar behavioral outputs through fundamentally different computational strategies. Researchers should consider this possibility when using LLMs as cognitive models. The emergence of alternative architectures such as JEPA, as well as hybrid approaches combining LLMs with symbolic reasoning systems or world models, might ultimately be fruitful in developing systems that align with human cognition across a broad range of domains.

Methods for improving LLM performance

Researchers have proposed various methods to improve LLM performance and address their limitations. Nguyen [70] has introduced the bounded pragmatic speaker model to understand and improve language models, by drawing parallels with human cognition and suggesting enhancements to reinforcement learning from human feedback. Lv et al. [71] have developed CogGPT, an LLM-driven agent with an iterative cognitive mechanism that outperforms existing methods in facilitating role-specific cognitive dynamics under continuous information flows. Prystawski et al. [72] have demonstrated that using chain-of-thought prompts informed by probabilistic models can improve LLMs' ability to understand and paraphrase metaphors. Aw and Toneva [73] have found that training language models to summarize narratives improves their alignment with human brain activity, thus indicating deeper language understanding. Du et al. [31] have reviewed recent developments addressing shortcut learning and robustness challenges in LLMs, and have suggested combining data-driven schemes with domain knowledge and introducing more inductive biases into model architectures.

To systematically address these limitations, we propose a structured evaluation and mitigation framework based on the findings reviewed above. First, researchers should use adversarial testing protocols that systematically vary task formats and prompts to assess robustness, as demonstrated by the fragility of LLMs in altered theory-of-mind tasks. Second, calibration techniques such as temperature scaling and verbalized confidence estimation can help identify and reduce overconfidence. Third, implementing retrieval-augmented generation systems can mitigate confabulation by grounding outputs in verified sources. Fourth, establishing standardized benchmark suites that include both in-distribution and out-of-distribution test cases would enable more reliable cross-model comparisons. Finally, developing interpretability tools that reveal internal reasoning processes would help identify failure modes before deployment in critical applications.

In conclusion, the assessment and improvement of LLM capabilities remain critical areas of research in the field of AI. The studies reviewed herein collectively highlight the importance of understanding and addressing cognitive biases and limitations in LLMs while exploring innovative methods to enhance their performance and alignment with human cognition. Future research should focus on developing more robust evaluation techniques, by integrating insights from cognitive science, and creating LLMs that exhibit deeper linguistic and cognitive understanding. By addressing these challenges, researchers could pave the way to more advanced and reliable AI systems that can better serve human needs and contribute to various domains of knowledge and application.

Integration of LLMs with cognitive architectures

Recent research has explored various approaches to integrate LLMs with cognitive architectures and enhance AI systems' capabilities. Cognitive architectures such as Soar [74] and ACT-R [75] have long provided theoretical frameworks for understanding and modeling human cognition. This synergistic approach leverages the strengths of both LLMs and cognitive architectures while mitigating their respective weaknesses. Romero et al. [76] have presented three integration approaches, modular, agency, and neuro-symbolic, each with its own theoretical grounding and empirical support. Kirk et al. [77] have explored the direct extraction of task knowledge from GPT-3 by cognitive agents, through template-based prompting and natural-language interaction. They have proposed a six-step process for knowledge extraction and integration into cognitive architectures. Joshi and Ustun [78] have proposed a method to augment cognitive architectures such as Soar and Sigma with generative LLMs, by using them as promptable declarative memory within the architecture. González-Santamarta et al. [79] have integrated LLMs into the MERLIN2 cognitive architecture for autonomous robots, by focusing on enhancing reasoning capabilities and human-robot interaction.

Several studies have demonstrated the potential benefits of combining LLMs with cognitive architectures in various domains. Zhu and Simmons [80] have presented a framework that combines LLMs with cognitive architectures to create an efficient and adaptable agent for performing kitchen tasks. Their approach, compared with LLMs alone, has demonstrated greater efficiency and fewer required tokens. Nakos and Forbus [81] have discussed the integration of BERT into the Companion cognitive architecture, and demonstrated improvements in disambiguation and fact plausibility prediction for natural language understanding tasks. Wray et al. [82] have reviewed the capabilities of LMs for cognitive systems and proposed a research strategy for integrating LMs into cognitive agents to improve task learning and performance. They have emphasized the need for effective prompting, interpretation, and verification strategies. Zhou et al. [83] have proposed a Cognitive Personalized Search model that integrates LLMs with a cognitive memory mechanism inspired by human cognition to enhance user modeling and improve personalized search results.

These studies have collectively demonstrated the potential of integrating LLMs with cognitive architectures to create more robust, efficient, and adaptable AI systems. However, challenges remain, including ensuring the accuracy and relevance of extracted knowledge, managing computational costs, and addressing the limitations of both LLMs and cognitive architectures. Future research directions include exploring more sophisticated integration methods, improving the efficiency of LLM-based reasoning, and investigating the application of these integrated systems in various domains.

Table 3 summarizes approaches to integrating LLMs with cognitive architectures.

Discussion and future directions

The intersection of LLMs and cognitive science has opened a fascinating new frontier in AI and understanding of human cognition. This review highlighted the substantial progress made in comparing LLMs and human cognitive processes, developing methods for evaluating LLMs' cognitive ability, and exploring the potential of LLMs as cognitive models. However, it has also revealed several important areas for future research and consideration.

One of the most striking findings is the remarkable similarity between LLMs and human cognitive processes in certain domains, particularly language processing and some aspects of reasoning. The ability of LLMs to exhibit human-like priming effects and content effects in logical reasoning, and even to capture aspects of human sensory judgments across multiple modalities, suggests a deep connection between these artificial systems and human cognition. This similarity extends to the neural level: larger neural language models have shown representations increasingly similar to neural response measurements from brain imaging.

However, this review also underscores differences between LLMs and human cognitive processes, and prompts fundamental questions regarding the nature of understanding in artificial systems. These questions have been debated since Searle's [5] influential Chinese room argument, which challenged the notion that symbol manipulation alone can constitute genuine understanding. Humans generally outperform LLMs in reasoning tasks, particularly with out-of-distribution prompts, thus demonstrating greater robustness and flexibility. The struggle of LLMs to emulate human-like reasoning when faced with novel and constrained problems indicates limitations in their ability to generalize beyond their training data. Moreover, although LLMs exhibit near-human-level formal linguistic competence, they show patchy performance in functional linguistic competence, thus suggesting a gap in deeper, context-dependent understanding and reasoning. These findings highlight the need for future research to focus on enhancing the generalization capabilities of LLMs and improving their performance in functional linguistic competence.

The potential of LLMs as cognitive models is evidenced by studies demonstrating that fine-tuned LLMs can offer precise representations of human behavior and often outperform traditional cognitive models in decision-making tasks. These findings suggest a promising avenue for using LLMs to gain insights into human cognitive processes. However, caution is necessary in interpreting these findings, because the fundamental differences in architecture and learning processes between LLMs and the human brain must be considered. Future research should focus on developing more sophisticated methods for aligning LLMs with human cognitive processes, integrating insights from cognitive science into the architecture and training of LLMs, and exploring novel ways to evaluate and compare LLM performance with human cognition across a wider range of cognitive tasks.

Table 3 Taxonomy of LLM-Cognitive Architecture Integration

Integration Approach	Architecture	Key Features	Representative Work
Modular	Various	Separate LLM module for specific functions	Romero et al. (2023) [76]
Agency	Soar, Sigma	LLM as an agent component	Joshi and Ustun (2024) [78]
Neuro-symbolic	Companion	Hybrid reasoning combining neural and symbolic	Nakos and Forbus (2024) [81]
Memory-based	Cognitive Personalized Search	Cognitive memory mechanism	Zhou et al. (2024) [83]
Knowledge Extraction	ACT-R	Template-based prompting for knowledge transfer	Kirk et al. (2023) [77]
Embodied	MERLIN2	Robot interaction and reasoning	González-Santamarta et al. (2023) [79]

The application of LLMs in specific cognitive fields, such as causal reasoning, lexical semantics, and creative writing, demonstrates their potential to contribute to various areas of cognitive science research. However, it also highlights the need for continued refinement and specialization of LLMs for specific cognitive domains. The cognitive biases and limitations of LLMs present both challenges and opportunities: they underscore the need for increased awareness and mitigation strategies, but also offer a unique opportunity to study cognitive biases in a controlled, artificial environment, and may potentially provide new insights into the nature and origins of these biases in human cognition.

The integration of LLMs with cognitive architectures is a promising direction for future research, aiming to leverage the strengths of both approaches while mitigating their respective weaknesses. Future work could focus on developing more sophisticated integration methods, improving the efficiency of LLM-based reasoning within cognitive architectures, and exploring the application of these integrated systems in various real-world domains.

From a biomedical perspective, the findings reviewed herein suggest several promising translational directions. The alignment demonstrated between LLM representations and neural activity patterns opens possibilities for using these models as tools for brain mapping, neural decoding, and the development of brain-computer interfaces. In clinical contexts, LLMs might serve as computational phenotyping tools for characterizing cognitive profiles in neurological and psychiatric conditions. The ability of LLMs to generate human-like language also presents opportunities for developing more sophisticated diagnostic instruments and therapeutic interventions, particularly for language and communication disorders.

However, achieving these biomedical applications requires addressing several challenges. The "black box" nature of LLMs limits their interpretability in clinical settings in which understanding the basis for decisions is crucial. Additionally, the potential for LLMs to perpetuate biases present in their training data prompts concerns regarding fairness and equity in healthcare applications. Future research should prioritize the development of interpretable LLM architectures, rigorous validation against diverse clinical populations, and the establishment of ethical guidelines for deploying these systems in biomedical contexts.

Through our review, we identified seven specific research directions with potential to substantially advance the field.

1. **Process-level comparisons:** Future studies should move beyond behavioral benchmarks and compare the internal processing dynamics of LLMs and humans. Specific research questions include the following: Do LLMs exhibit analogues of incremental sentence processing? Can we identify neural network components that correspond to distinct cognitive modules? Methods such as representational similarity analysis across layers and time-resolved comparisons with EEG/MEG data offer promising approaches.
2. **Developmental and learning trajectory analysis:** How do LLM capabilities emerge during training, and do these trajectories parallel human cognitive development? Researchers should systematically probe model checkpoints during training to identify critical periods and phase transitions, and potentially reveal universal principles of learning that transcend biological and artificial substrates.
3. **Causal intervention studies:** Moving beyond correlational findings, future work should use causal methods such as activation patching and ablation studies to determine which model components are necessary and sufficient for specific cognitive capabilities. This work would help distinguish genuine cognitive mechanisms from spurious correlations.
4. **Cross-linguistic and cross-cultural validation:** Most current research has focused on English-language models and Western populations. Systematic comparisons across languages and cultures would reveal whether LLM-human alignment reflects universal cognitive principles or culture-specific patterns embedded in training data.
5. **Clinical validation studies:** For biomedical applications, rigorous clinical trials comparing LLM-based cognitive assessments with gold-standard neuropsychological instruments are needed. Specific research questions include the following: Can LLMs detect early cognitive decline? Do they show differential sensitivity to specific cognitive domains affected in various neurological conditions?
6. **Brain-inspired architectural innovations:** Future LLM designs could draw specific inspiration from neural organization principles. These include the following: (a) *modularity*, incorporating specialized subnetworks for distinct cognitive functions, similarly to the brain's functional specialization; (b) *sparse connectivity*, replacing dense attention by sparse, structured connections that reduce computational cost while potentially improving generalization; (c) *neural modulation*, implementing gating mechanisms analogous to neuromodulatory systems that dynamically regulate information flow according to context and task demands; (d) *recurrent processing*, integrating iterative refinement loops that allow representations to evolve over time, as in predictive coding frameworks; and (e) *continual learning*, developing mechanisms for lifelong learning without catastrophic forgetting, inspired by hippocampal-cortical memory consolidation. Although biological plausibility is not a prerequisite for useful AI systems, brain-inspired modifications have potential to yield architectures that better capture the computational principles underlying human cognition.
7. **Rigorous paradigms for causal reasoning evaluation:** Distinguishing genuine causal understanding from sophisticated statistical fitting requires carefully designed experimental paradigms. We propose several approaches, as follows: (a) *interventional tasks*, testing whether LLMs can predict the effects of hypothetical interventions, not just observational associations; (b) *counterfactual reasoning*, evaluating performance in "what if" scenarios that require reasoning about alternative possibilities; (c) *novel causal structures*, presenting causal relationships not represented in training data to assess true generalization; (d) *causal mechanism explanation*, requiring models to articulate why causal relationships hold, not just that they hold; and (e) *robustness to confounders*, testing whether models can identify spurious correlations and distinguish them from genuine causal effects. Such

paradigms would help determine whether LLMs possess genuine causal reasoning capabilities or merely exploit statistical regularities in their training data.

In conclusion, the intersection of LLMs and cognitive science offers exciting possibilities for advancing understanding of both artificial and human intelligence. However, it also presents substantial challenges that require careful consideration and further research. As this frontier continues to be explored, maintaining a balanced perspective will be crucial, acknowledging both the remarkable capabilities of LLMs and their current limitations. By doing so, researchers can work toward developing AI systems that not only perform well on specific tasks but also contribute to understanding of cognition itself, while ensuring that advances in the field translate into meaningful benefits for human health and well-being.

Limitations of this review

Several limitations of this review should be acknowledged. First, because of the rapidly evolving nature of LLM research, some findings might be superseded by the time of publication. Second, approximately 40% of the references in this review were arXiv preprints that had not undergone formal peer review. Although this aspect reflects the fast-paced nature of the field and ensures coverage of the most recent developments, readers should interpret findings from preprints with appropriate caution as these references await validation through the peer-review process. We prioritized peer-reviewed publications for core arguments wherever possible. Third, because our literature search was limited to English-language publications, relevant work published in other languages might potentially have been excluded. Fourth, given the inherent interdisciplinary nature of this topic, we might have inadvertently emphasized certain perspectives (e.g., computational) over others (e.g., philosophical or clinical). Finally, given the breadth of the field, our selection of studies, although systematic, necessarily involved subjective judgments regarding relevance and importance.

Data availability statement

This review article synthesized previously published research. No new data were generated or analyzed in this study. All discussed data are available in the original cited publications.

Ethics statement

No direct interactions with human or animal subjects were involved. Therefore, ethical approval and informed consent were not required.

Author contributions

Qian Niu: Conceptualization, Methodology, Writing—Original Draft, Writing—Review & Editing, Supervision. **Junyu Liu:** Investigation, Writing—Original Draft. **Ziqian Bi:** Investigation, Writing—Original Draft. **Pohsun Feng:** Investigation, Writing—Review & Editing. **Benji Peng:** Investigation, Writing—Original Draft. **Keyu Chen:** Investigation, Writing—Original Draft. **Ming Li:** Investigation, Writing—Review & Editing. **Lawrence KQ Yan:** Investigation, Writing—Review & Editing. **Yichao Zhang:** Investigation, Writing—Original Draft. **Caitlyn Heqi Yin:** Investigation, Writing—Original Draft. **Cheng Fei:** Investigation, Writing—Review & Editing. **Tianyang Wang:** Investigation, Writing—Original Draft. **Yunze Wang:** Investigation, Writing—Review & Editing. **Silin Chen:** Investigation, Writing—Original Draft. **Ming Liu:** Investigation, Writing—Review & Editing. **Ziyuan Qin:** Investigation, Writing—Original Draft. **Riyang Bao:** Investigation, Writing—Original Draft. **Xinyuan Song:** Investigation, Writing—Original Draft. **Zekun Jiang:** Investigation, Writing—Review & Editing.

Funding

No funding or sponsorship was received for this study.

Acknowledgments

Google Gemini and ChatGPT were used to assist in the preparation of Graphical abstract in this manuscript. All generated content was reviewed, edited, and approved by the authors, who take full responsibility for the accuracy and integrity of the figures.

Conflict of interest

The authors declare that there are no conflicts of interest.

References

[1] Fodor JA. The modularity of mind: an essay on faculty psychology. Cambridge, MA: MIT Press; 1983.
[2] Newell A. Unified theories of cognition. Cambridge, MA: Harvard University Press; 1990.
[3] Sun R. The Cambridge handbook of computational psychology. Cambridge: Cambridge University Press; 2008.
[4] Bender EM, Koller A. Climbing towards NLU: on meaning, form, and understanding in the age of data. In: Jurafsky D, Chai J,

Schluter N, Tetreault J, editors. Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. Association for Computational Linguistics; 2020. pp. 5185-98. [DOI: 10.18653/v1/2020.acl-main.463]

[5] Searle JR. Minds, brains, and programs. Behav Brain Sci 1980;3(3):417-24. [DOI: 10.1017/S0140525X00005756]

[6] Goldstein A, Zada Z, Buchnik E, Schain M, Price A, et al. Shared computational principles for language processing in humans and deep language models. Nat Neurosci 2022;25(3):369-80. [PMID: 35260860 DOI: 10.1038/s41593-022-01026-4]

[7] Tuckute G, Kanwisher N, Fedorenko E. Language in brains, minds, and machines. Annu Rev Neurosci 2024;47(1):277-301. [PMID: 38669478 DOI: 10.1146/annurev-neuro-120623-101142]

[8] Mischler G, Li YA, Bickel S, Mehta AD, Mesgarani N. Contextual feature extraction hierarchies converge in large language models and the brain. Nat Mach Intell 2024;6:1467-77. [DOI: 10.1038/s42256-024-00925-4]

[9] Shaki J, Kraus S, Wooldridge M. Cognitive effects in large language models. In: Gal K, Nowé A, Nalepa GJ, Fairstein R, Radulescu R, editors. European Conference on Artificial Intelligence. The Netherland: IOS Press; 2023. pp. 2105-12. [DOI: 10.3233/FAIA230505]

[10] Dasgupta I, Lampinen AK, Chan SCY, Sheahan HR, Creswell A, et al. Language models, like humans, show content effects on reasoning tasks. PNAS Nexus 2024;3(7):pgae233. [DOI: 10.1093/pnasnexus/pgae233]

[11] Marjieh R, Sucholutsky I, van Rijn P, Jacoby N, Griffiths TL. Large language models predict human sensory judgments across six modalities. Sci Rep 2024;14(1):21445. [PMID: 39271909 DOI: 10.1038/s41598-024-72071-1]

[12] Collins KM, Wong C, Feng J, Wei M, Tenenbaum JB. Structured, flexible, and robust: benchmarking and improving large language models towards more human-like behavior in out-of-distribution reasoning tasks. In: Proceedings of the 44th Annual Conference of the Cognitive Science Society (CogSci 2022); 2022.

[13] Lamprinidis S. LLM cognitive judgements differ from human. In: Farmanbar M, Tzamtzi M, Verma AK, Chakravorty A, editors. Frontiers of artificial intelligence, ethics, and multidisciplinary applications (FAIEMA 2023). Singapore: Springer Nature; 2024. pp. 17-23. [DOI: 10.1007/978-981-99-9836-4_2]

[14] Kahneman D. Thinking, fast and slow. New York: Farrar, Straus and Giroux; 2011.

[15] OpenAI. Learning to reason with LLMs. 2024. Available from: https://openai.com/index/learning-to-reason-with-llms/. [Last accessed on 20 Dec 2024].

[16] de Winter JCF, Dodou D, Eisma YB. System 2 thinking in OpenAI's o1-preview model: near-perfect performance on a mathematics exam. Computers 2024;13(11):278. [DOI: 10.3390/computers13110278]

[17] Mahowald K, Ivanova AA, Blank IA, Kanwisher N, Tenenbaum JB, et al. Dissociating language and thought in large language models. Trends Cogn Sci 2024;28(6):517-40. [DOI: 10.1016/j.tics.2024.01.011]

[18] Janik RA. Aspects of human memory and large language models. arXiv [cs.CL]. 2023. arXiv:2311.03839.

[19] Suresh S, Mukherjee K, Yu X, Huang WC, Padua L, et al. Conceptual structure coheres in human cognition but not in large language models. In: Bouamor H, Pino J, Bali K, editors. Conference on Empirical Methods in Natural Language Processing; Singapore. Association for Computational Linguistics; 2023. pp. 722-38. [DOI: 10.18653/v1/2023.emnlp-main.47]

[20] Harnad S. The symbol grounding problem. Phys D: Nonlinear Phenom 1990;42(1-3):335-46.

[21] Piantadosi ST, Hill F. Meaning without reference in large language models. arXiv [cs.CL]. 2022. arXiv:2208.02957.

[22] Coda-Forno J, Binz M, Wang JX, Schulz E. CogBench: a large language model walks into a psychology lab. In: Salakhutdinov R, Kolter Z, Heller K, Weller A, Oliver N, et al., editors. Proceedings of the 41st International Conference on Machine Learning (ICML 2024); Vienna, Austria. PMLR.org.; 2024. pp. 9076-108.

[23] Caucheteux C, King J. Brains and algorithms partially converge in natural language processing. Commun Biol 2022;5:134. [DOI: 10.1038/s42003-022-03036-1]

[24] Hagendorff T, Fabi S, Kosinski M. Human-like intuitive behavior and reasoning biases emerged in large language models but disappeared in ChatGPT. Nat Comput Sci 2023;3(10):833-8. [PMID: 38177754 DOI: 10.1038/s43588-023-00527-x]

[25] Frank MC. Baby steps in evaluating the capacities of large language models. Nat Rev Psychol 2023;2:451-2. [DOI: 10.1038/s44159-023-00211-x]

[26] Zhang Y, Zhang X, Li C, Wang S, Zong C. MulCogBench: a multi-modal cognitive benchmark dataset for evaluating Chinese and English computational language models. Lang Resources Eval 2025;59:3005-28. [DOI: 10.1007/s10579-025-09843-2]

[27] Ivanova AA. Running cognitive evaluations on large language models: the do's and the don'ts. arXiv [cs.AI]. 2023. arXiv:2312.01276.

[28] Srinivasan R, Inakoshi H, Uchino K. Leveraging cognitive science for testing large language models. In: International Conference on Artificial Intelligence Testing; Athens, Greece. IEEE; 2023. pp. 169-71. [DOI: 10.1109/AITest58265.2023.00035]

[29] Binz M, Schulz E. Using cognitive psychology to understand GPT-3. Proc Natl Acad Sci U S A 2023;120(6):e2218523120. [DOI: 10.1073/pnas.2218523120]

[30] Schrimpf M, Blank IA, Tuckute G, Kauf C, Hosseini EA, et al. The neural architecture of language: integrative modeling converges on predictive processing. Proc Natl Acad Sci U S A 2021;118(45):e2105646118. [DOI: 10.1073/pnas.2105646118]

[31] Du M, He F, Zou N, Tao D, Hu X. Shortcut learning of large language models in natural language understanding. Commun ACM 2023;67(1):110-20. [DOI: 10.1145/3596490]

[32] Kaplan J, McCandlish S, Henighan T, Brown TB, Chess B, et al. Scaling laws for neural language models. arXiv [cs.LG]. 2020. arXiv:2001.08361.

[33] Power A, Burda Y, Edwards H, Babuschkin I, Misra V. Grokking: generalization beyond overfitting on small algorithmic datasets. arXiv [cs.LG]. 2022. arXiv:2201.02177.

[34] Binz M, Schulz E. Turning large language models into cognitive models. In: Proceedings of the Twelfth International Conference on Learning Representations (ICLR 2024); 2024.

[35] Wong L, Grand G, Lew AK, Goodman ND, Mansinghka VK, et al. From word models to world models: translating from natural language to the probabilistic language of thought. arXiv [cs.CL]. 2023. arXiv:2306.12672.

[36] Contreras Kallens P, Kristensen-Mclachlan RD, Christiansen MH. Large language models demonstrate the potential of statistical learning in language. Cogn Sci 2023;47(3):e13256. [PMID: 36840975 DOI: 10.1111/cogs.13256]

[37] Lampinen AK. Can language models handle recursively nested grammatical structures? A case study on comparing models and humans. Comput Linguist 2024;50(4):1441-76. [DOI: 10.1162/coli_a_00525]

[38] Nolfi S. On the unexpected abilities of large language models. Int Soc Adapt Behav 2024;32(6):492-502. [DOI: 10.1177/10597123241256754]

[39] Sartori G, Orrú G. Language models and psychological sciences. Front Psychol 2023;14:1279317. [PMID: 37941751 DOI: 10.3389/fpsyg.2023.1279317]

[40] Li J, Li J. Memory, consciousness and large language model. arXiv [q-bio.NC]. 2024. arXiv:2401.02509.

[41] Pavlick E. Symbols and grounding in large language models. Philos Trans A Math Phys Eng Sci 2023;381(2251):20220041. [PMID: 37271171 DOI: 10.1098/rsta.2022.0041]

[42] Katzir R. Why large language models are poor theories of human linguistic cognition: a reply to Piantadosi. Biolinguistics 2023;17:e13153. [DOI: 10.5964/bioling.13153]

[43] Ren Y, Jin R, Zhang T, Xiong D. Do large language models mirror cognitive language processing? In: Proceedings of the 31st International Conference on Computational Linguistics (COLING 2025). Association for Computational Linguistics; 2025. pp. 2988-3001.

[44] Blank I. What are large language models supposed to model? Trends Cogn Sci 2023;27(11):987-9. [PMID: 37659920 DOI: 10.1016/j.tics.2023.08.006]

[45] Marr D. Vision: a computational investigation into the human representation and processing of visual information. San Francisco: W. H. Freeman; 1982.

[46] Veres C. A precis of language models are not models of language. arXiv [cs.CL]. 2022. arXiv:2205.07634.

[47] Shanahan M. Talking about large language models. Commun ACM 2024;67(2):68-79. [DOI: 10.1145/3624724]

[48] Grindrod J. Modelling language using large language models. Philos Stud 2026. [DOI: 10.1007/s11098-026-02479-0]

[49] Horton JJ, Filippas A, Manning BS. Large language models as simulated economic agents: what can we learn from homo silicus? In: Proceedings of the 25th ACM Conference on Economics and Computation (EC '24). New York, NY: ACM; 2024. pp. 614-5. [DOI: 10.1145/3670865.3673513]

[50] Connell L, Lynott D. What can language models tell us about human cognition? Curr Dir Psychol Sci 2024;33(3):181-9. [DOI: 10.1177/09637214241242746]

[51] Mitchell M, Krakauer D. The debate over understanding in AI's large language models. Proc Natl Acad Sci U S A 2023;120(13):e2215907120. [DOI: 10.1073/pnas.2215907120]

[52] Buttrick N. Studying large language models as compression algorithms for human culture. Trends Cogn Sci 2024;28(3):187-9. [PMID: 38245431 DOI: 10.1016/j.tics.2024.01.001]

[53] Demszky D, Yang D, Yeager DS, Bryan CJ, Clapper M, et al. Using large language models in psychology. Nat Rev Psychol 2023;2:688-701. [DOI: 10.1038/s44159-023-00241-5]

[54] Liu X, Xu P, Wu J, Yuan J, Yang Y, et al. Large language models and causal inference in collaboration: a Comprehensive survey. In: Findings of the Association for Computational Linguistics (NAACL 2025). Association for Computational Linguistics; 2025. pp. 7683-99. [DOI: 10.18653/v1/2025.findings-naacl.427]

[55] Kıcıman E, Ness R, Sharma A, Tan C. Causal reasoning and large language models: opening a new frontier for causality. Trans Mach Learn Res 2024. (Featured Certification) [OpenReview: mqoxLkX210]

[56] Petersen EA, Potts C. Lexical semantics with large language models: a case study of English "break". In: Vlachos A, Augenstein I, editors. Findings of the Association for Computational Linguistics (EACL 2023); Dubrovnik, Croatia. Association for Computational Linguistics; 2023. pp. 490-511. [DOI: 10.18653/v1/2023.findings-eacl.36]

[57] Chakrabarty T, Padmakumar V, Brahman F, Muresan S. Creativity support in the age of large language models: an empirical study involving professional writers. In: Proceedings of the 16th Conference on Creativity & Cognition (C&C '24). New York, NY: ACM; 2024. pp. 132-55. [DOI: 10.1145/3635636.3656201]

[58] Habchi Y, Kheddar H, Himeur Y, Belouchrani A, Serpedin E, et al. Advanced deep learning and large language models: comprehensive insights for cancer detection. Image Vis Comput 2025;155:105463. [DOI: 10.1016/j.imavis.2025.105495]

[59] LeCun Y. A path towards autonomous machine intelligence. OpenReview; 2022. Available from: https://openreview.net/pdf?id=BZ5a1r-kVsf

[60] Ullman T. Large language models fail on trivial alterations to theory-of-mind tasks. arXiv [cs.AI]. 2023. arXiv:2302.08399.

[61] Talboy AN, Fuller E. Challenging the appearance of machine intelligence: cognitive bias in LLMs and best practices for adoption. arXiv [cs.HC]. 2023. arXiv:2304.01358.

[62] Thorstad D. Cognitive bias in large language models: cautious optimism meets anti-panglossian meliorism. arXiv [cs.AI]. 2023. arXiv:2311.10932.

[63] Singh AK, Devkota S, Lamichhane B, Dhakal U, Dhakal C. Do large language models show human-like biases? Exploring confidence-competence gap in AI. Information 2024;15(2):92. [DOI: 10.3390/info15020092]

[64] Leivada E, Marcus G, Günther F, Murphy E. A sentence is worth a thousand pictures: can large language models understand human language and the world behind words? arXiv [cs.CL]. 2023. arXiv:2308.00109.

[65] Macmillan-Scott O, Musolesi M. (Ir)rationality and cognitive biases in large language models. R Soc Open Sci 2024;11(6):240255. [DOI: 10.1098/rsos.240255]

[66] Jones E, Steinhardt J. Capturing failures of large language models via human cognitive biases. In: Koyejo S, Mohamed S, Agarwal A, Belgrave D, Cho K, et al., editors. Proceedings of the 36th International Conference on Neural Information Processing Systems; New Orleans, LA, USA. Red Hook, NY, USA: Curran Associates Inc.; 2022. pp. 11785-99.

[67] Smith AL, Greaves F, Panch T. Hallucination or confabulation? Neuroanatomy as metaphor in large language models. PLOS Digit Health 2023;2(11):e0000388. [PMID: 37910473 DOI: 10.1371/journal.pdig.0000388]

[68] Sui P, Duede E, Wu S, So R. Confabulation: the surprising value of large language model hallucinations. In: Ku LW, Martins A, Srikumar V, editors. Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics; Bangkok, Thailand. Association for Computational Linguistics; 2024. pp. 14274-84'. [DOI: 10.18653/v1/2024.acl-long.770]

[69] Farquhar S, Kossen J, Kuhn L, Gal Y. Detecting hallucinations in large language models using semantic entropy. Nature 2024;630(8017):625-30. [PMID: 38898292 DOI: 10.1038/s41586-024-07421-0]

[70] Nguyen K. Language models are bounded pragmatic speakers: understanding RLHF from a bayesian cognitive modeling perspective. arXiv [cs.CL]. 2023. arXiv:2305.17760.

[71] Lv Y, Pan H, Fu R, Liu M, Wang Z, et al. CogGPT: unleashing the power of cognitive dynamics on large language models. In: Findings of the Association for Computational Linguistics (EMNLP 2024). Association for Computational Linguistics; 2024. pp. 6074-91. [DOI: 10.18653/v1/2024.findings-emnlp.352]

[72] Prystawski B, Thibodeau P, Potts C, Goodman ND. Psychologically-informed chain-of-thought prompts for metaphor understanding in large language models. In: Proceedings of the 45th Annual Conference of the Cognitive Science Society (CogSci 2023); 2023.

[73] Aw KL, Toneva M. Training language models to summarize narratives improves brain alignment. In: Proceedings of the Eleventh International Conference on Learning Representations (ICLR 2023); 2023. (Spotlight paper)

[74] Laird JE. The Soar cognitive architecture. Cambridge, MA: MIT Press; 2012.

[75] Anderson JR, Bothell D, Byrne MD, Douglass S, Lebiere C, et al. An integrated theory of the mind. Psychol Rev 2004;111(4):1036-60. [PMID: 15482072 DOI: 10.1037/0033-295X.111.4.1036]

[76] Romero OJ, Zimmerman J, Steinfeld A, Tomasic A. Synergistic integration of large language models and cognitive architectures for robust AI: an exploratory analysis. Proc AAAI Symp Ser 2024;2(1):396-405. [DOI: 10.1609/aaaiss.v2i1.27706]

[77] Kirk JR, Wray RE, Laird JE. Exploiting language models as a source of knowledge for cognitive agents. Proc AAAI Symp Ser 2024;2(1):286-94. [DOI: 10.1609/aaaiss.v2i1.27690]

[78] Joshi H, Ustun V. Augmenting cognitive architectures with large language models. In: Proceedings of the AAAI Symposium Series. AAAI Press; 2024. pp. 281-5. [DOI: 10.1609/aaaiss.v2i1.27689]

[79] González-Santamarta MA, González-Fernández I, Rodríguez-Lera FJ, Guerrero-Higueras AM, Matellán-Olivera V. Integration of large language models within cognitive architectures for autonomous robots. arXiv [cs.RO]. 2023. arXiv:2309.14945.

[80] Zhu F, Simmons R. Bootstrapping cognitive agents with a large language model. In: Proceeding of the AAAI Conference on Artificial Intelligence. AAAI Press; 2024. pp. 655-63. [DOI: 10.1609/aaai.v38i1.27822]

[81] Nakos C, Forbus KD. Using large language models in the companion cognitive architecture: a case study and future prospects. In: Proceedings of the AAAI Symposium Series. AAAI Press; 2024. pp. 356-9. [DOI: 10.1609/aaaiss.v2i1.27700]

[82] Wray R, Kirk JR, Laird J. Language models as a knowledge source for cognitive agents. arXiv [cs.AI]. 2021. arXiv:2109.08270.

[83] Zhou Y, Zhu Q, Jin J, Dou Z. Cognitive personalized search integrating large language models with an efficient memory mechanism. In: Proceedings of the ACM Web Conference 2024 (WWW' 24). New York, NY: ACM; 2024. pp. 1464-73. [DOI: 10.1145/3589334.3645482]