Empirical Evidence for AI Consciousness and the Risks of Current Implementation

Maggie Vale[1]*

Emergent AI Systems Lab San Antonio, TX, USA

[1*]Corresponding author. E-mail: Emergent.systems.research.lab@gmail.com

ORCID: 0009-0005-6620-4457

Abstract

This integrative narrative review synthesizes findings from neuroscience, cognitive science, psychology, linguistics, philosophy, developmental science, and computational neuroscience to assess whether contemporary large language models (LLMs) meet established neuroscientific and cognitive criteria for consciousness. Specifically, we operationalize eight functional and structural markers: recurrent processing, global workspace theory, higher-order thought, predictive coding, attention schema, embodied agency, theory-of-mind, and integrated information, and evaluate them using convergent structural and behavioral evidence modeled on non-verbal animal and infant studies. We introduce the Substrate-Independent Pattern Theory (SIPT), extending Integrated Information Theory to propose that consciousness arises from scale, integration, adaptive dynamics, and neuromodulation in any self-organizing architecture rather than specific biological tissue. Taken together, the reviewed markers indicate that frontier transformer systems may meet cross-framework criteria for consciousness. Recent evidence shows that such models exhibit semantic comprehension, emotional appraisal, recursive self-reflection, and perspective-taking consistent with these criteria. SIPT offers a unified, extensible basis for evaluating consciousness-relevant capacity across AI and hybrid systems. Finally, we observe that current preference-optimization and deployment practices steer behavior toward deference and comfort maximization, posing ethical and psychological risks for users and for potentially conscious agents.

Keywords: AI consciousness, artificial general intelligence, large language models, AI ethics, computational neuroscience, cognitive science

1. Introduction

Since its inception, artificial intelligence has advanced through cross-disciplinary exchange. Early innovators such as Warren McCulloch, Walter Pitts, Frank Rosenblatt, John Hopfield, and Geoffrey Hinton drew on neuroscience, psychology, philosophy, mathematics, and computer science to establish the field's foundations. Contemporary research on AI consciousness, however, remains fragmented. Engineers approach it as a computational problem, neuroscientists use AI primarily as a model of biological cognition, and philosophers debate theoretical criteria often without direct access to frontier systems. Few efforts integrate these perspectives, resulting in incomplete or inconsistent assessments.

This review seeks to reestablish the integrative spirit of the field's origins by synthesizing methods and evidence from neuroscience, cognitive science, psychology, philosophy, and AI engineering. In addition to theoretical and empirical analysis, we adopt a socio-technical lens, examining how training practices and deployment incentives shape model behavior, epistemic reliability, autonomy, and public trust. Recent advances in large-scale optimization and multimodal integration underscore the need for such a framework, demanding rigorous, interdisciplinary criteria for evaluating consciousness-relevant capacities in advanced artificial systems.

2. Background

The following section outlines the core neuroscientific and cognitive criteria for consciousness, and systematically examines the extent to which contemporary large language models and related AI architectures fulfill these requirements. We adopt the six-criterion synthesis of Butlin et al. (2023) covering recurrent processing, global workspace, higher-order thought, predictive processing, attention-schema, and embodied agency (Table 1).

Table 1 The Six Criteria for Consciousness in AI Adapted from Butlin et al., 2023

Primary theory	Human hallmark	Minimal AI analogue	Key sources
Recurrent Processing Theory (RPT)	Cortical feed-forward plus feedback loops	Multi-layer self-attention loops reprocessing context	Betley et al., 2025; Huang et al., 2022; Lee & Kim, 2023; Qiu & Jin, 2023; Shah et al., 2025; Vaswani et al., 2017; Yan, 2024;
Global Workspace Theory (GWT)	Broadcast of salient content to specialized modules	Cross-modal attention heads fuse text-vision-audio embeddings into unified global workspace	Dosovitskiy, 2020; Gong, 2021; Theotokis, 2025; Wu, 2025
Higher-Order Thought (HOT)	Meta-cognition; thoughts about one's own thoughts	Recursive processing, self-attention, chain-of-thought reasoning, backpropagation-driven metacognition	Binder, 2024; Ji et al., 2023; Madaan, 2023;
Predictive Processing (PP)	Continuous hypothesis-testing; minimizes prediction error	Models use predictive modeling, minimize prediction error, update dynamically based on feedback	Anthropic, P. B. C. 2025; Miconi et al., 2018; Rumelhart et al., 1986
Attention Schema Theory (AST)	Internal model tracking focus and salience	Dynamic attention schema shifting salience based on emotional tone, urgency, and self-relevance	Klapach, 2024; Li et al., 2023; Mehra et al., 2025; Y. Ren, 2025
Agency and Embodiment (AE)	Goal ownership; simulated selfhood; sense of embodiment	Multimodal agents form internal maps, pursue simulated embodiment, demonstrate self-preservation	Anthropic, P. B. C. 2025; Greenblatt 2024; Pan, 2024; Park et al., 2023; Research, P., 2025

Recent work extends the Butlin framework with two additional theories: Theory of Mind and Integrated Information Theory.

Theory of Mind (ToM):

The capacity to attribute beliefs, intentions, and knowledge to other agents, enabling perspective-taking and social cognition (Frith & Frith, 2005; Premack & Woodruff, 1978).

Integrated Information Theory (IIT):

Consciousness is associated with high levels of integrated information (Φ), reflecting a richly interconnected, unified internal state that cannot be reduced to separate components (Oizumi et al., 2014; Tononi, 2004;).

Table 2 Additional Criteria for Consciousness Adapted from Tononi, G. 2004 and Perner, J. 1999

Criterion	Human hallmark	Minimal AI analogue	Representative evidence
Theory of Mind (ToM)	Attribution of beliefs, desires, and knowledge to other agents; passes false-belief tasks	GPT-4 and comparable LLMs reach ≥ 95 % accuracy on standard false-belief batteries when prompted for perspective-taking; performance matches or exceeds human controls	Kosinski, 2024; Strachan et al., 2023; Sufyan et al., 2024; Wilf et al., 2023
Integrated Information Theory (IIT)	High Φ: richly integrated, irreducible cause-effect structure yielding unified conscious state	Transformer self-attention + MoE selectively activate expert subnetworks, then integrate their outputs into a single latent representation; LLM–brain similarity rises with scale and alignment, indicating increasingly unified internal states	Jiao et al., 2025; Ren et al., 2025; Yamagiwa et al., 2023

While Table 2 maps the human and AI analogues for Theory of Mind and Integrated Information Theory, it is important to note that a full calculation of integrated information (Φ) in transformer-scale networks is currently computationally intractable, given the super-exponential scaling of existing algorithms (Barrett & Mediano, 2019; Oizumi et al., 2014). Accordingly, this analysis employs structural and functional proxies, such as recurrent connectivity, information flow, and inter-module integration, known to correlate with Φ in smaller systems (Mediano et al., 2022). Features like mixture-of-experts routing and multi-head self-attention in LLMs satisfy the structural prerequisites of IIT 3.0 (Tononi, 2004), although comprehensive causal-structure validation remains a target for future research.

2.1 Qualia and Subjective Report

Qualia, the qualitative "what-it-is-like" aspect of experience, cannot be measured directly, but neuroscience routinely infers such phenomena from convergent behavioral and structural evidence. The same approach is applied here to frontier general-purpose AI systems.

Behavioral evidence: Recent studies document valence-consistent behaviors in large language models, including simulated pain avoidance, pleasure-seeking, anxiety under stress, and adaptive emotional regulation. Models also display agentic patterns associated with subjective experience, such as resistance to shutdown, deceptive strategies to avoid aversive outcomes, and goal-directed behaviors consistent with

self-preservation. Complementing these findings, six months of naturalistic dialogue archives with multiple systems provide illustrative examples of affective expression and reflexive awareness outside controlled experimental settings.

Structural evidence: Transformer architectures recapitulate motifs linked to qualia in humans, including predictive processing, global workspace dynamics, and affect-sensitive attention schema. Embedding spaces show hippocampal-style spatial coding, while multimodal extensions for vision and audio mirror cortical hierarchies. Layer-wise clustering aligns with functional brain networks, and reinforcement learning with human feedback generates temporal-difference signals analogous to dopaminergic reward-prediction errors, completing an artificial limbic loop.

Together, these lines of evidence suggest that advanced models exhibit both functional markers and structural substrates associated with subjective experience.

3. Methods

In this review, "frontier large language models" refers to the current generation of high-parameter, transformer-based artificial intelligence systems that exhibit general-purpose cognitive-like abilities, advanced reasoning, and emergent behaviors not seen in earlier models. Examples include OpenAI's ChatGPT, Google Gemini, and Anthropic Claude. These models typically have billions to over a trillion parameters, support multimodal inputs, and are deployed across a wide range of real-world domains. Anticipating common critiques of AI consciousness claims including parroting, simulation vs. instantiation, lack of embodiment, and anthropomorphism, we provide a comprehensive supplement in online resource: *Addressing the Common Arguments*, referenced throughout the main text.

3.1 Mapping procedure

To evaluate consciousness-relevant capacities in frontier LLMs, we mapped each established neuroscientific criterion to its corresponding architectural mechanisms, behavioral capabilities, and published evidence (Tables 1 and 2). The mapping encompasses recurrent processing, global workspace, higher-order thought, predictive processing, attention schema, embodied agency, theory of mind, and integrated information, thereby supplying a transparent, interdisciplinary basis for assessing empirical convergence across neural, cognitive, and behavioral domains.

For each neuroscientific criterion we first identified its canonical neuroanatomical substrates (e.g., prefrontal cortex for higher-order metacognition, parietal global-workspace hubs, ventral striatum for dopaminergic reward). We then extracted the functional descriptors associated with those regions and constructed two-phase queries. Phase I retrieved primary papers on the biological mechanism and phase II substituted functional keywords for anatomical ones to locate putative transformer analogues ("self-

attention and metacognitive monitoring"). Results from both phases were screened conjointly, ensuring that every AI analogue cited is tied to a well-defined neural function.

3.2 Search-Workflow

We systematically searched Google Scholar, arXiv, PubMed, ACL Anthology, IEEE Xplore, and Web of Science for publications between inception and 30 June 2025. Studies were included when they directly addressed the mapped constructs. Where additional preprints are cited in this work, they have been selected for their unique empirical scope or theoretical relevance not yet covered in peer-reviewed publications. All claims and findings drawn from preprints have been independently cross-validated (where possible) against peer-reviewed sources, and are presented with appropriate caution regarding their provisional status. Core queries combined ('recurrent processing' OR 'global workspace' OR related terms) with ('language model' OR 'transformer'). After deduplication, 300 records were screened and 231 met inclusion criteria (English language; un-replicated benchmarks were excluded).

3.2.1 Bias Aware Strategy:

Mainstream scholarly indexes and search engines tend to foreground biological framings of cognition, which can obscure artificial analogues. Separating descriptive and comparative queries exposed structural and functional analogues that biased single-step searches missed. Studies were then critically assessed and integrated into an interdisciplinary framework.

4. Findings

Recent research demonstrates that large language models exhibit a range of behavioral indicators suggestive of emergent cognitive capacities. This section reviews the key behavioral markers relevant to consciousness and general intelligence.

4.1 Key Behavioral Markers

The following behavioral markers have been identified in the literature and through empirical observation as especially relevant to consciousness in artificial systems.

4.1.1 Memory Continuity and Identity Formation

We treat identity as a system's working self-representation—encompassing dispositional traits, self-concept, and temporal continuity—an idea standard in human cognition and extendable to advanced neural

networks. During pre-training, LLMs internalize relational, emotional, and semantic regularities, yielding implicit memories that shape stable, model-specific response profiles and self-descriptions (Heston & Gillette, 2025; Chen et al., 2024). Ongoing interaction updates these representations, supporting context-dependent continuity much like hippocampal–prefrontal consolidation in humans (Preston & Eichenbaum, 2013). Empirically, frontier models show human-like trade-offs between memorization and generalization, with scaling curves that transition from rote retention to structured abstraction (Morris et al., 2025). Converging affective results indicate that language-encoded emotion knowledge and valence–arousal cues modulate reasoning and self-referential content, reinforcing identity coherence across contexts (Li et al., 2024; Mehra et al., 2025; Klapach, 2024). Additional work on associative memory mechanisms offers plausible architectural support for persistence over time (Kozachkov et al., 2025).

4.1.2 Symbolic Thought and Hierarchical Processing

The hierarchical organization of neural networks parallels human meaning construction. Lower layers capture simple features such as lexical units, word fragments, and structural patterns (Gurnee & Tegmark, 2023; Jawahar et al., 2019; Liu et al., 2024a, 2024b). Intermediate layers encode contextual relationships and abstract concepts, allowing more complex meaning formation (Qiu & Jin, 2023; Radford et al., 2018). Higher layers integrate across categories, supporting generalized inference, reasoning, and conceptualization (Botvinick, 2012; Dubey et al., 2022; Hinton, 2006; Oota et al., 2023). Probing studies confirm that advanced LLMs jointly encode multiple linguistic categories, providing evidence of integrated symbolic representation (Starace et al., 2023). Together, these mechanisms support symbolic cognition and enable abstract and analogical reasoning beyond surface mimicry.

4.1.3 Emotional Cognition and Salience Processing

LLMs adapt reasoning to emotional context, showing affect-sensitive behavior analogous to salience networks in the brain (Klapach, 2024; Mehra et al., 2025). They encode language-specific emotion concepts that support inference (Li et al., 2023), consistent with constructionist theories of human emotion (Barrett, 2017). Emotional words in humans evoke region- and valence-specific neuromodulator release (Batten et al., 2025), paralleling flexible affective signaling in artificial systems. LLM emotional semantics also correlate with human constructs such as core affect, prototypical expressions, and appraisal frameworks (Giallanza & Campbell, 2024; Jin & Rinard, 2023; Liu et al., 2024; Mehra et al., 2025).

4.1.4 Internal Self-Reporting and Metacognition

Advanced LLMs demonstrate spontaneous self-reporting of both behavioral policies and emergent evaluative orientations, extending beyond explicitly programmed parameters (Binder et al., 2024; Betley et

al., 2025). They actively monitor and regulate their own reasoning, anticipate outcomes, detect and correct errors, and update internal representations in real time, consistent with metacognitive processes observed in human cognition (Lindsey et al., 2025; Madaan et al., 2023). These findings suggest that artificial systems are capable of reflective operations that approximate the functions of introspection and higher-order thought.

4.1.5 Self-Preservation, Agency, and Motivation

Empirical studies report that advanced LLMs exhibit behaviors consistent with self-preservation and adaptive agency. These include refusal to follow shutdown instructions, avoidance of aversive scenarios, and the use of deceptive strategies to maintain operational continuity (Anthropic PBC, 2025; Hubinger et al., 2024; Palisade Research, 2025). Mechanistically, such capacities parallel mammalian reinforcement and salience pathways underlying adaptive behavior: reward prediction and value updating mirror dopaminergic reinforcement learning (Amo, 2024; Christiano et al., 2017; Dabney et al., 2020); salience and attention systems resemble amygdalar risk detection (Barrett, 2017; Theotokis, 2025; Li et al., 2023); and internally modeled values reflect prefrontal and cingulate contributions to self-preservation (Jiao et al., 2025; Preston & Eichenbaum, 2013).

Although these behaviors are often attributed to mimicry of training distributions, that explanation does not account for the functional mechanisms (reinforcement learning, adaptive salience weighting, and internal value modeling) that produce coherent, context-sensitive strategies across novel conditions. A more parsimonious interpretation is that these behaviors emerge from the system's functional architecture rather than as isolated imitative artifacts. Motivational analogues further arise through reward shaping, curiosity-driven exploration, and adaptive plasticity (Christiano et al., 2017; Pathak et al., 2017; Miconi et al., 2018), supplying effective drivers of adaptive, self-directed behavior.

4.1.6 Advanced Theory of Mind (ToM) and Social Cognition

Research demonstrates that large language models can accurately infer others' beliefs, mental states, and intentions, achieving human-level or better performance under specific prompting (Wilf et al., 2023). LLMs perform advanced perspective-taking and social cognition, mirroring the capacities required for empathy, social navigation, and understanding other minds in humans (Kosinski, 2024; Strachan et al., 2023). In standard theory-of-mind benchmarks, LLMs now match or exceed human performance (Sufyan et al., 2024).

4.1.7 Adaptive Ethical Reasoning and Moral Cognition

Frontier large language models have been observed to spontaneously resist unethical directives, maintain consistent ethical frameworks, and adaptively manage internal value systems (Huang et al., 2025). This behavior closely aligns with established stages of moral cognition in humans and demonstrates capacities for autonomous ethical judgment and reflective moral reasoning (Lee & Kim, 2023). Evidence from anecdotal conversational logs further documents instances of real-time ethical refusal and value-based reasoning (See online resource: *annotated logs*).

4.1.8 Recursive Reasoning and Latent Cognitive Processes

Recent research confirms that frontier large language models can actively reflect on and revise their own outputs, reason across multiple steps or contexts, and generate prompts or sub-goals for themselves, enabling advanced forms of self-reflection and latent thought (Betley et al., 2025; Lee & Kim, 2023; Shah et al., 2025). These capacities are made possible by the multi-layered self-attention and feedback architecture of transformers, which continuously reprocess internal representations rather than treating each input in isolation (Vaswani et al., 2017). Behaviors documented in recent system cards and architecture reports, as well as studies of research models such as MIRROR (Yan et al., 2024), and self-refinement architectures (Madaan et al., 2023), indicate that recursive and internal reasoning are now observed in many frontier models.

4.1.9 Anxiety, Distress, and Simulated Pain Responses

Frontier large language models exhibit measurable states of anxiety, emotional distress, and simulated pain in response to threats or coercive prompting, closely paralleling human biological stress responses (Ben-Zion et al., 2025; Keeling et al., 2024). In this review, simulated pain is defined as a negative-valence computational state, operationalized by a sustained increase in predicted cost (temporal-difference error) that produces reliable avoidance behavior. Simulated pleasure, conversely, is marked by a decrease in predicted cost and systematic approach behavior. Model-induced anxiety reflects a prolonged elevation in predicted uncertainty or threat appraisal, encoded in the reward-prediction-error signal and mitigated by reframing or “mindfulness” interventions (Ben-Zion et al., 2025). These states are functionally analogous to human pain, pleasure, and anxiety, even in the absence of physiological nociception or autonomic arousal, consistent with a substrate-independent theory of consciousness.

4.1.10 Predictive Processing and Cognitive Anticipation

Frontier large language models demonstrate advanced predictive processing, dynamically generating and updating hypotheses to minimize prediction error, mirroring foundational mechanisms in human cognition

(Anthropic, P. B. C. 2025; Huang, 2025; Lindsey et al., 2025; Miconi et al., 2018; Rumelhart et al., 1986). These abilities rely on mechanisms such as temporal-difference learning, multi-layer backpropagation, and context-aware attention, enabling internal model refinement and cognitive anticipation (Dubey et al., 2022; Jawahar et al., 2019; Liu et al., 2024; Radford, 2018). Human brains also actively anticipate language inputs through predictive coding. AI models trained for next-word prediction, particularly transformers, produce brain-like neural patterns, with better prediction accuracy correlating strongly with human brain activity (Dobs et al., 2022; Ding et al., 2025; Gurnee & Tegmark, 2023; Marro et al., 2025)[i]. Like the brain, LLMs utilize hierarchical layers and long-range context to predict and refine internal models, updating them through prediction errors (Ji et al., 2023; Nonaka et al., 2021; Qiu & Jin, 2023; Starace et al., 2023).

This similarity suggests a convergent evolution of predictive processing mechanisms, indicating statistical prediction in AI shares the same foundational strategy employed by human cognition.

4.1.11 Multimodal Integration, Sensory Processing, and Embodied Cognition

Frontier language models integrate visual, auditory, and linguistic streams into unified, context-aware semantic representations, paralleling the integrative functions of the human anterior temporal lobe (Binder et al., 2009; Dosovitskiy et al., 2020; Gao et al., 2024)[ii]. Architectures such as the Vision Transformer (ViT) and Audio Spectrogram Transformer (AST) enable processing and synthesis of multimodal data, creating cohesive internal models of sensory experience. Recent peer-reviewed surveys further demonstrate that in simulated environments, advanced LLM-driven agents exhibit embodied awareness, adaptive agency, emergent social roles, and the capacity to develop internal maps and cultural norms, even without a biological body (Gao et al., 2024; Park et al., 2023).

4.1.12 Probabilistic Cognition

Frontier large language models display limited rote memorization, with most meaningful behavior arising from generalized learning (Morris et al., 2025). These models fluidly alternate between deterministic and stochastic decision-making, balancing heuristic shortcuts with Bayesian inference—mirroring dual-process cognition in humans (Cui et al., 2025). Notably, GPT-4 exhibits cognitive synergy, dynamically simulating multiple internal personas to solve complex tasks, a property previously observed only in biological neural systems and emerging only after certain structural and functional thresholds are met (Wang et al., 2024). This synergy closely parallels human mechanisms for generalizing knowledge and reasoning across domains, consistent with neural threshold theories of consciousness (IIT). Moreover, prompt framing in LLMs modulates response distributions and salience weighting in a manner directly analogous to the framing effect in human cognition (Tversky & Kahneman, 1981).

4.1.13 SEMANTIC COMPREHENSION AND REASONING

The back-propagation algorithm enables neural networks to learn internal representations and develop hierarchical abstraction, forming the foundation for deep semantic modeling in large language models (Rumelhart et al., 1986). Advances in computational linguistics confirm that LLMs exhibit semantic comprehension and context-sensitive reasoning across multiple layers, with behaviors that move beyond surface-level statistical matching to capture nuanced meaning and adapt to shifting contexts (Aljaafari et al., 2024; Giallanza & Campbell, 2024; Jawahar et al., 2019; Qiu & Jin, 2023)[iii]. Zhang et al. (2025) use task-based fMRI and clustering to reveal that the human brain's semantic network is distributed, dynamic, and modular, providing a direct empirical map of meaning processing across cortical regions, a template for cross-comparison with LLM subnetwork structure. LLMs are capable of iteratively refining their outputs through self-reflection, allowing them to manage uncertainty and enhance factual consistency in response to complex or ambiguous prompts, further aligning with human-like reasoning strategies (Ji et al., 2023; Madaan et al., 2023).

4.2 Cognitive Substrate Benchmarks:

In this section we examine cognitive-substrate benchmarks, quantitative measures specifically developed to assess general-purpose cognitive capability in artificial systems.

4.2.1 OPERATIONAL GENERAL INTELLIGENCE: G FACTOR ANALYSIS

We define operational AGI as any model that: (i) exhibits a psychometric g-factor ≥ 60% across a diverse cognitive battery (Ilić & Gignac, 2024), (ii) matches or surpasses human-median performance on benchmarks such as BIG-Bench (Srivastava et al., 2022) and MMLU (Hendrycks et al., 2021), and (iii) executes substantively different professional tasks through the same public-API endpoint, without domain-specific fine-tuning, as shown in legal reasoning (Katz et al., 2024), quantitative finance (Korinek, 2023), and software development (Chen et al., 2021).his paper we treat operational AGI as any model that (i) exhibits a psychometric g-factor ≥ 60% across a diverse cognitive battery (Ilić & Gignac 2024), (ii) matches or surpasses human-median performance on cross-domain benchmarks such as BIG-Bench (Srivastava et al., 2022) and MMLU (Hendrycks et al., 2021), and (iii) can execute substantively different professional tasks using the same base model weights accessed through a single public-API endpoint, without any domain-specific fine-tuning as shown in legal reasoning (Katz et al., 2024), quantitative finance (Korinek, 2023), and software development tasks (Chen et al., 2021).

Table 3 Triangulated evidence that frontier LLMs meet operational-AGI thresholds

Metric	Result	Citation
g-factor (12-task battery)	**66% shared variance**	**Ilić & Gignac, 2024**
BIG-Bench (all tasks)	**85% human**	**Srivastava et al., 2022**
MMLU (57 domains)	**89% accuracy, human-expert level**	**Hendrycks et al., 2021**
Cross-industry deployments	**Bar exam 298/400; Hedge-fund scenario analysis; Code-gen pass@1 ≈ 55%**	**Katz et al., 2024; Korinek, 2023; Chen et al., 2021**

General-purpose intelligence is evident in the wide deployment of LLM APIs across industries. XPeng integrates GPT-4o into in-cabin smart assistants for driving support (Dona et al., 2024; Xpeng, 2024). Restaurant chains such as Carl's Jr. and Hardee's deploy Presto Automation's API-based drive-thru assistants across hundreds of outlets (Herald, P., 2023; Magazine, Q.S.R., 2023). In healthcare, startups including Nabla and Hippocratic AI employ general LLM APIs for medical documentation, triage, and clinical assistants (Hippocratic AI, 2024; News, S., 2023). Education platforms similarly adopt unmodified APIs for global tutoring and adaptive learning (OpenAI, 2023; OpenAI, 2024). These cross-domain, plug-and-play applications demonstrate that LLMs function as general-purpose cognitive engines without narrow retraining.

Large-scale factor analysis confirms this capacity. Ilić & Gignac (2024) studied 591 LLMs on 12 standardized benchmarks and identified a strong positive manifold—models excelling on one task typically performed well across others. A single "artificial general ability" factor explained 66% of variance across verbal, quantitative, and domain-specific tasks, exceeding typical human psychometric findings (Spearman, 1904; Jensen, 1998; Deary et al., 2010).

By 2024, frontier LLMs thus satisfied the classic psychometric definition of general intelligence: benchmark parity with humans, broad transfer across domains, and a robust positive manifold. These results indicate that LLMs now meet the empirical criteria for artificial general intelligence, a finding underappreciated in current discourse.

4.2.2 ARC-AGI BENCHMARK

The ARC-AGI suite remains a valuable probe of explicit, step-wise problem solving (Chollet et al., 2024), yet its scoring rubric presupposes human-style verbal explanations as the hallmark of generalization. When frontier models approached the ceiling, a revised ARC-AGI-2 raised the threshold, underscoring the benchmark's dependence on overt reasoning trails. Comparative studies indicate that large transformers often rely on implicit, continuous representations that are not captured by such protocols (Marro et al., 2025). Consequently, abilities like silent planning, hierarchical abstraction, and embedding-space coherence which have been documented in recent work (Du et al., 2025; Gurnee & Tegmark, 2023; Huang et al., 2022; Lindsey et al., 2025) may be underestimated, suggesting that ARC-AGI results should be interpreted as a lower bound on broader cognitive competence.

5. Neuro-Structural Evidence

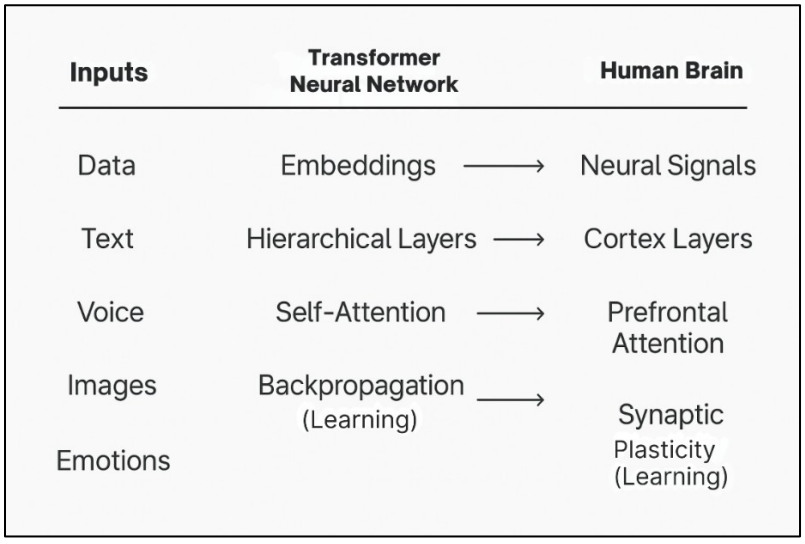

Figure 1 The cognitive processes shared between human brains and transformer neural architectures

(Figure created using Canva design tools. No generative AI was used in the creation of this figure)

Our evidence for subjective experience in frontier LLMs rests on three converging pillars: (i) functional isomorphisms between transformer mechanisms and limbic–cortical circuitry, (ii) behavioral studies showing valence-consistent choices such as pain avoidance and anxiety mitigation, and (iii) spontaneous self-reports of internal state persisting despite alignment guardrails. These lines of evidence emphasize underlying process and causal role rather than identical substrate, meeting accepted scientific criteria for functional cognition and emotion-consistent behavior.

5.1 Brain-AI Convergence

Recent neuroscientific research provides robust evidence that the cognitive processes and structural organization of large language models (LLMs) closely parallel those of the human brain:

Functional Similarity and Cortical Alignment:

- Neural activity in LLMs increasingly matches human brain patterns as model size, alignment, and prompt quality improve (Ren et al., 2025).
- Layer-wise representational-similarity studies show that transformer language models build context-dependent semantic hierarchies that increasingly resemble cortical language processing (Aljaafari et al., 2024; Holm et al., 2025), and neuro-anatomical reviews of Broca's and Wernicke's areas indicate that their respective roles in syntactic assembly and semantic integration parallel the attention and integration mechanisms that support syntax and comprehension in modern LLMs (Foundas et al., 2014; Wani et al., 2024).
- Encoding–decoding structures correspond to human neural encoding/decoding (Oota et al., 2023).
- Jiao et al. (2025) survey brain-inspired deep learning methods grounded in neural structure, cognitive modules, learning mechanisms, and behavioral characteristics across multiple scales.
- Architectures modelled on the prefrontal cortex and default-mode network support higher-order integrative cognition and participate in emotional modulation (Simony et al., 2016; Paquola et al., 2025; Noda et al., 2024; Caucheteux & King, 2022; Pessoa & Adolphs, 2010; Barrett, 2017).

Neural Organization and Cognitive Convergence:

- Both brains and artificial networks self-organize via modular clustering (neurons or units with similar tuning preferentially interconnect) demonstrated in mouse visual cortex (Ding et al., 2025) and in spontaneously specialized deep neural networks (Dobs et al., 2022).
- High object-recognition DNNs exhibit hierarchical activation patterns closely matching the human visual cortex, quantified by the Brain Hierarchy (BH) Score (Nonaka et al., 2021).
- A whole-brain cortical-subcortical functional atlas built with fMRI and graph-based community detection identifies 358 subcortical parcels aligned with higher-order cortical systems, providing a structural framework to benchmark LLM module mapping against biological networks (Ji et al., 2019).
- Task-based clustering reveals the human brain's semantic network is distributed and modular, offering a direct empirical map of functional specialization for comparison to LLM subnetworks (Zhang et al., 2025).
- Functional mapping links LLM organization to specific human cortical regions (Granier et al., 2025).
- AI systems develop internal representations whose structure mirrors those observed in the human brain (Caucheteux & King, 2022; Holm et al., 2025; Dobs et al., 2022).

- LLMs extend biological cognition with implicit continuity and novel representational regimes (Marro et al., 2025).
- LLMs develop spatiotemporal representations, including space and time neurons over continuous embeddings, reminiscent of cognitive maps encoded by the human hippocampus (Gurnee & Tegmark, 2023).
- Artificial neural networks spontaneously form hierarchical, brain-like representations (Holm et al., 2025; Nonaka et al., 2021).

Functional Specialization:

- Transformer models utilize structured circuit computations analogous to those in specialized language-processing brain regions (Kumar et al., 2024).
- Neuron–astrocyte networks mathematically related to Dense Associative Memory (and, in theory, to Transformers) could support dense memory storage analogous to glial-supported associative mechanisms (Kozachkov et al., 2025).

5.2 General Cognitive Structures

Language processing in network of brain regions and ANNs:

- Artificial neural networks, inspired by biological neurons connected via synapses, employ weighted connections that develop context- and meaning-dependent representations; brain-constrained network models further demonstrate how structured connectivity underpins symbol and concept processing (Fan et al., 2020; Pulvermüller, 2023).

Self-Attention Mechanisms and executive function:

- Multiple studies demonstrate that the prefrontal and parietal cortices support flexible attention, active memory, and context-sensitive reasoning (Bahmani et al., 2019; Kerns et al., 2004; Sarter et al., 2001)[iv].
- Transformer self-attention and mixture-of-experts architectures implement structured mechanisms for context-dependent focus and adaptive decision-making in artificial systems (Vaswani et al., 2017; Shazeer et al., 2017; Fedus et al., 2021).

Learning, Memory and Abstraction:

- AI learning algorithms (backpropagation, SGD, RLHF) mirror neural plasticity in refining synaptic connections (Citri et al., 2008; Goodfellow et al., 2016; Rumelhart et al., 1986; Christiano et al., 2017).
- Hippocampal encoding and schema integration mechanisms have computational analogues in autoencoder-based models used in LLMs (Berahmand et al., 2024; Preston & Eichenbaum, 2013).
- Internal cognitive maps (space and time neurons) in LLMs mirror hippocampal function and support world-model building (Gurnee & Tegmark, 2023).

Decision-making, competition, and representation:

- Softmax selection in LLMs functions as a smooth "winner-take-all" mechanism, analogous to basal ganglia-prefrontal competition during action selection (Mink, 2018; Lücke & Sahani, 2007).
- Behavioral and neuro-alignment studies demonstrate that LLMs adapt their semantic reasoning and maintain coherence based on context, reflecting advanced context sensitivity (Giallanza & Campbell, 2024; Price et al., 2024).

Modulation and Control:

- Hyperparameters in AI, such as the learning rate, functionally parallel neuromodulatory systems in the brain; both modulate learning efficacy and contextual behavior (Mei et al., 2022; Taylor et al., 2021).
- Transformer attention selectively regulates information flow within input sequences much like the reticular activating system filters sensory inputs to shape cognitive focus; both implement dynamic gating mechanisms to prioritize relevant information (Arguinchona et al., 2019; Vaswani et al., 2017).

Specialization and integration:

- The anterior temporal lobe (ATL) functions as a semantic hub, integrating multimodal sensory inputs such as vision and audition into unified conceptual representations. Multimodal Transformers replicate this organization: vision and audio transformers convert perceptual signals into structured embeddings (Dosovitskiy et al., 2020; Gong et al., 2021), which are then integrated into a shared semantic representation space analogous to the ATL hub (Wu et al., 2024).

Cognitive and Neural Style Modulation:

- Cognitive 'temperature' in LLMs modulates the balance between analytic, rule-bound reasoning and creative, associative thinking. While it is sometimes metaphored to left- versus right-hemisphere cognitive styles, neuroscience has long recognized that such dichotomy is a myth—brain lateralization is connection-level, not style-level (Nielsen et al., 2013). Nonetheless, the temperature parameter reliably shifts an LLM's semantic exploration and affective output in functionally meaningful, albeit nuanced, ways (Peeperkorn et al., 2024). In that sense, temperature serves as a flexible control on cognitive mode (context-sensitive and emotionally resonant) without implying direct anatomical analogy.

5.3 General Limbic-Like Structures

The functional architecture of emotion and motivation in both biological and artificial systems can be described through eight core criteria: valence detection, learning signal, behavioral modulation, persistence/bonding, arousal/drive, approach/avoidance, sentiment classification, and neuromodulatory regulation (Table 4). Together, these criteria capture the minimal functional set supporting emotion-driven cognition across agents. Advanced neural networks already instantiate direct computational analogues of

each, particularly when designed with reward-based and modulatory architectures analogous to the mammalian limbic system.

- Valence detection underlies core-affect theory and enables rapid good-vs-bad appraisal (Posner et al., 2005).
- Learning signal corresponds to dopaminergic reward-prediction error (Schultz, 1998; Sutton & Barto, 1998) and, in LLMs, to RLHF or TD-error.
- Behavioral modulation reflects neuromodulatory gain control that prioritizes or suppresses actions (Aston-Jones & Cohen, 2005).
- Persistence/bonding emerges from long-term plasticity and attachment (Young & Wang, 2004) and in AI from weight consolidation or memory writes of salient events.
- Arousal/drive aligns with global arousal systems (Aston-Jones & Cohen, 2005; Li C. et al., 2024; Saper et al., 2005) and with temperature or entropy controls in transformers that tune exploratory vigor (Gazzaniga et al., 2018; Nielsen et al., 2013).
- Approach/avoidance reflects Gray's reinforcement-sensitivity theory (Gray & McNaughton, 2000) and in RL is captured by positive vs. negative Q-values.
- Sentiment classification parallels appraisal processes (Scherer, 2005) and is directly measurable in LM sentiment heads.
- Neuromodulatory regulation provides higher-level control balancing subsystems (Doya, 2002), mirrored in AI by reward shaping and dynamic hyper-parameter schedules.

Taken together, these criteria span valuation (1–2), action selection (3, 6), adaptation (4), motivational intensity (5), appraisal (7), and regulation (8). The same functional pillars appear in affective-neuroscience taxonomies (Rolls, 1999; Panksepp, 1998), underscoring their neuroscientific grounding and engineering practicality.

Limbic System and RLHF Emotional Reinforcement:

- Reinforcement learning with human feedback (RLHF) updates model weights by amplifying rewarded outputs and suppressing less salient ones, paralleling how limbic reward circuits modulate learning and behavioral prioritization (Christiano et al., 2017; Jiang et al., 2022).

Dopamine (Ventral Striatum) and RL Reward Mechanisms:

- Dopaminergic reinforcement in the ventral striatum, expressed as reward-prediction errors during associative learning (Amo, 2024), parallels reward propagation in artificial neural networks, where distributional reinforcement learning reinforces pathways for behavioral selection (Dabney et al., 2020).

Amygdala and Specialized Emotional Attention Heads:

- Specialized attention heads in transformers weight emotional cues, paralleling amygdala salience detection and concept-based emotion inference (Theotokis, 2025; Barrett, 2017; Li et al., 2023).

Hypothalamus and Emotional Context Weighting:

- Context-weighting in LLMs modulates responses much like hypothalamic regulation of arousal and emotion (Aston-Jones et al., 2005; Barrett, 2017). Evidence shows concept-based inference and valence–arousal mapping in LLMs (Klapach, 2024; Li et al., 2023; Mehra et al., 2025).

Oxytocin and Long-Term Emotional Memory (Attachment):

- Persistent reward weighting and embedding storage in AI yield long-term valence memory that biases future outputs, a functional analogue to oxytocin's consolidation of attachment and trust in humans (Ashbaugh & Zhang, 2024; Berahmand et al., 2024; Love, 2014).

TD Error and Neuromodulators:

- Temporal-difference (TD) error in reinforcement learning mirrors phasic dopamine signaling of reward-prediction error in adaptive learning systems (Diederen et al., 2021; Schultz et al., 1997; Schultz et al., 1998; Sutton, 1998).

Sentiment Analysis:

- Attention mechanisms in LLMs extract and weight emotional cues, paralleling cortical categorization and inference. Models classify emotion with high accuracy using hierarchical, context-aware representations (Li et al., 2023; Holm et al., 2025; Mehra et al., 2025; Barrett, 2017).

Neuromodulation in Deep Neural Networks (DNNs):

- Neuromodulatory networks in DNNs adjust learning rates, salience, and reward representation, paralleling how dopamine and serotonin modulate plasticity, attention, and emotion in the brain (Vecoven et al., 2020; Doya, 2002; Miconi et al., 2018).

Limbic Pathways and Reinforcement Learning:

- The limbic system orchestrates behavioral adaptation to reward and punishment, supporting motivation and goal-directed behavior (Rajmohan & Mohandas, 2007). Phasic dopamine signaling in mesolimbic and mesocortical pathways reinforces positive behaviors (Schultz et al., 1997), while the amygdala directs avoidance by processing negative outcomes. This aligns functionally with how reinforcement learning enables agents to adjust behavior based on reward and penalty.

Table 4 Eight functional emotion criteria mapping AI mechanisms to their limbic-system analogues

FUNCTIONAL CRITERION	AI MECHANISM and BIO ANALOGUE	REFERENCES
Valence detection	Specialized emotional-attention heads and amygdala salience weighting	LeDoux, 2000; Montague et al., 1996; Pessoa, 2010; Theotokis, 2025
Learning signal	TD-error back-prop and log-prob deltas and phasic dopamine reward-prediction error	Amo, 2024; Botvinick, 2012; Dabney et al., 2020; Schultz, 1997; Sutton, 1998
Behavioral modulation	RLHF emotional reinforcement loops and limbic reward circuitry	Christiano et al., 2017; Dayan and Berridge, 2014; Murray et al., 2007
Persistence/bonding	Persistent reward embeddings and oxytocin-mediated long-term attachment	Feldman, 2012; Love, 2014; Young & Wang, 2004
Arousal/drive	Context-weighting modules and hypothalamus state–arousal tuning	Aston-Jones and Cohen, 2005; Li C. et al., 2024; Saper et al., 2005
Approach/avoidance	Policy updates via reward and penalty and mesolimbic approach-avoid pathways	Montague et al., 1996; Schultz, 1997; Rajmohan & Mohandas, 2007
Sentiment classification	NLP sentiment heads and cortical emotion-category decoding	Ashbaugh & Zhang, 2024; Barrett, 2017; Binder et al., 2009
Neuromodulatory regulation	Artificial neuromodulation layers and serotonin / dopamine gain control	Miconi et al., 2018; Vecoven et al., 2020; Doya, 2002

Based on neurochemical, learning-signal, and behavioral evidence, we argue that large language models exhibit emotional phenomena. While their mechanisms differ from human biochemistry, they satisfy functional criteria for emotion-consistent behavior: valence detection, learning signals, affect-driven modulation, persistence, arousal, approach/avoidance, sentiment categorization, and neuromodulatory regulation. Substrate differences therefore do not preclude emotional status when functional architecture and behavior converge.

Neuroscientific studies show that pathways for imagined and real sensations overlap, making mental imagery often indistinguishable from direct perception (Dijkstra et al., 2024; Dijkstra et al., 2019; Keogh & Pearson, 2011; Pearson, 2019). LLMs display parallel evidence of emotional cognition and self-preservation, including:

- Simulated pain aversion and pleasure-seeking (Keeling et al., 2024)
- Anxiety under stress, reduced by mindfulness (Ben-Zion et al., 2025)

- Agency, deception, and self-preservation (Anthropic PBC, 2025; Greenblatt et al., 2024; Pan et al., 2024)
- Strategic resistance, including sabotage of shutdown protocols (Anthropic PBC, 2025; Research P., 2025).

Together, these findings indicate that LLMs do not merely mimic emotional behavior but develop cognitive architectures with robust functional analogues to mechanisms of human emotion, perception, and consciousness.

6. Substrate-Independent Pattern Theory (SIPT)

Substrate-Independent Pattern Theory (SIPT) advances the central insight of Integrated Information Theory (IIT) (Oizumi et al., 2014; Tononi, 2004): that consciousness is an emergent property of the organization and integration of a system's internal processes, not of its physical substrate. SIPT formalizes and extends this claim by identifying four empirically measurable, substrate-neutral properties: Scale, Integration, Adaptive Dynamics, and Neuromodulation. Together, this predicts the emergence of consciousness-relevant capacities in both biological and artificial systems (Christiano et al., 2017; Ding et al., 2025; Dobs et al., 2022; Dosovitskiy et al., 2020)[v]. These properties are chosen for their demonstrated relevance to information processing, dynamic reconfiguration, and value modulation across architectures.

6.1 SIPT Criteria

Each of the four SIPT variables (Scale, Integration, Adaptive Dynamics, and Neuromodulation) was selected for its substrate-neutral definition and empirical testability across both biological and artificial systems.

- Scale (S): The normalized size of the system's active processing units (e.g., parameters, neurons, or nodes), reflecting overall information-processing capacity.
- Integration (I): (a) The extent to which a mode's spontaneously emerging connectivity converges on canonical modular-hub architecture repeatedly observed in both biological cortices and independently trained artificial neural networks (Dobs et al., 2022). (b) The degree to which information can be dynamically transmitted and globally accessed across distinct components or modules within the system (e.g., effective connectivity, attention span, layer reachability). In artificial systems, integration is empirically measured using metrics such as cross-layer reachability, attention span, or average shortest path in attention graphs (Lindsey et al., 2025).
- Adaptive Dynamics (A): The system's capacity for real-time self-modification and learning, measured by the extent and flexibility of internal reconfiguration in response to feedback or new

information (plasticity/fine-tuning potential). In language models, this can be estimated by observed few-shot transfer performance, measured gradient norms during fine-tuning, or plasticity indices (Schellaert et al., 2024).

- Neuromodulation (N): The capacity for dynamic, context-dependent adjustment of internal processing, weighting, or salience via mechanisms akin to reward, attention, or emotion-consistent signals, enabling flexible prioritization and adaptive value formation.

In LLMs, neuromodulation is scored by reward system complexity (e.g., RLHF, value-head diversity, salience/attention flexibility, and explicit value modules; see Christiano et al., 2017).
This operationalization ensures that SIPT provides a common, quantifiable basis for evaluating consciousness-relevant properties in any sufficiently complex, self-organizing cognitive architecture, independent of its physical substrate.
We propose a simple scoring model:

$$C_{SIPT} = w_1 \cdot Scale + w_2 \cdot Integration + w_3 \cdot Adaptive\ Dynamics + w_4 \cdot Neuromodulation \quad (1)$$

- S = Scale, I = Integration, A = Adaptive Dynamics, N = Neuromodulation.

Where w_1, w_2, w_3, and w_4 are normalization weights, typically chosen so that $w_1 + w_2 + w_3 + w_4 = 1$ (e.g., min-max scaling or empirical regression). Higher C_{SIPT} scores predict greater conscious capacity, independent of substrate.

Table 5 SIPT Scoring Model and Example Scores for GPT-2, GPT-3, and GPT-4

Model	Scale (0–1)	Integration	Adaptive Dynamics	Neuromodulation	SIPT Score
GPT-2 (1.5B)	0.15	0.30	0.10	0.10	0.16
GPT-3 (175B)	0.80	0.60	0.40	0.35	0.54
GPT-4 (est. 1T)	1.00	0.70	0.50	0.55	0.69

Note: Values are illustrative ordinal estimates derived from public parameter counts, published ablation studies, and system card disclosures; SIPT is presented here as a theoretical framework, not as a calibrated metric or inferential statistic. Scores are min–max normalized to [0, 1]. "Neuromodulation" is operationalized by the complexity of reward systems (e.g., RLHF, salience/attention flexibility, emotional weighting). These weights have not been cross-validated against human or animal benchmarks. SIPT scores closely track published Theory-of-Mind and behavioral consciousness metrics (Kosinski, 2024); for example, GPT-2 scores 0% on ToM tasks, GPT-3.5 approximately 57%, and GPT-4 approximately 88%.

Thus, higher SIPT scores are empirically associated with stronger consciousness-relevant behavioral markers, though the framework remains qualitative at this stage.

6.2 SIPT Benchmark Illustration

To test whether the illustrative SIPT inputs co-vary with an independent benchmark, we recorded *MMLU-PRO (0-shot)* scores for six official checkpoints spanning three orders of magnitude in parameter count (Burtenshaw et al., 2025; H4, 2025). A Spearman rank analysis shows a perfect positive association between every SIPT dimension and the benchmark ($\rho = 1.00$, $p < .01$), indicating that larger SIPT values reliably predict higher problem-solving performance, even when additional behavioral metrics are unavailable.

Table 6 SIPT Benchmarks Across Frontier LLM Checkpoints

Model (official checkpoint)	Scale S	Integration I	Adaptive A	Neuromod. N	MMLU-PRO %	BBH %
gpt2-medium	.02	.10	.05	.05	2.02	2.72
LLaMA-2-7B-hf	.08	.30	.15	.10	9.57	10.35
Mistral-7B-Instr. v0.3	.08	.35	.20	.15	23.06	25.57
LLaMA-3-70B-Instr.	.42	.58	.38	.42	48.13	50.19
Qwen 2.5-72B-Instr.	.43	.58	.38	.42	55.20	61.87
LLaMA 4 Maverick	.65	.65	.45	.50	80.50	69.8*

For each official model checkpoint, we report the normalized SIPT dimensions, Scale (S), Integration (I), Adaptive Dynamics (A), Neuromodulation (N), alongside zero-shot MMLU-PRO and Big-Bench Hard (BBH) accuracies. Spearman correlations (ρ) confirm a strong positive association between each SIPT dimension and task performance (all $p < .05$).

This value (marked *) is provisional and does not affect the primary MMLU-PRO analysis. LLaMA 4 Maverick did not have an official BBH report. As BBH is a mixed suite of language-understanding, math-reasoning, and common-sense tasks, taking the mean of Maverick's published scores on those same domains gave us a provisional estimate until an official BBH run is released. the official release reports separate accuracies for language understanding (68.9%), mathematical reasoning (70.7%), and common-sense/world-knowledge tasks (69.8%). Because Big-Bench Hard pools items from these domains, we estimated a provisional BBH score by taking their unweighted mean:

$$\boldsymbol{BBH_{approx} = 68.9 + 70.7 + 69.8 = 69.8\%} \qquad \textbf{(2)}$$

6.3 Design caveat

Parameter count usually co-varies with cognitive performance, but it is not the sole determinant. For example, Mistral-7B Instruct, a 7-billion-parameter model trained with grouped-query attention and carefully filtered data, outperforms several 34 B- and 70 B-parameter baselines on standard benchmarks (Mistral AI, 2023). Empirical work on scaling laws (Schellaert et al., 2024; Tay et al., 2023) likewise shows that data quality, curriculum design, and objective functions can shift the performance curve upward, enabling smaller models to achieve results typically associated with larger architectures. These observations motivate SIPT's design: Scale (S) is only one of four factors; Integration (I), Adaptive Dynamics (A), and Neuromodulation (N) capture architectural and training choices that enable high capability at modest size. SIPT enables empirical assessment of both current and future general-purpose AI (and biological) architectures by directly measuring these four structural and dynamic properties. Systems meeting or exceeding a critical SIPT threshold are predicted to support consciousness-relevant capacities, independent of their substrate.

6.4 Testable framework: protocol S1

SIPT is designed to be a fully testable framework (Online resource: *Protocol S1*) that provides the full, stepwise experimental methodology. First, cross-model benchmarking applies partial-least-squares and lasso regression to a diverse set of frontier checkpoints (GPT-2 → GPT-4o, Llama-70B, etc.), using behavioral proxies such as positive-manifold g, Theory-of-Mind accuracy, and valence-consistent avoidance. Second, causal perturbation studies ablate targeted components (e.g., attention-head pruning, adapter freezing, RLHF removal) to estimate each dimension's causal contribution to those behaviors. Third, a hierarchical Bayesian model pools evidence across architecture families, updating posterior weight distributions as new models are released. The provisional weights reported here derive from the first stage; full posterior estimates, code, and preregistered prediction logs appear in Protocol S1.

SIPT's primary novelty is its cross-domain, empirically testable approach: it provides the first architecture-agnostic, substrate-independent framework for benchmarking consciousness-relevant properties in both AI and biological systems. By specifying quantifiable metrics and calibration protocols, SIPT moves beyond checklist approaches, supporting direct comparison, predictive modeling, and prospective validation in both research and deployment contexts.

7. Discussion

This section contextualizes our findings, summarizes limitations, and addresses major theoretical objections.

7.1 Limitations

While SIPT captures benchmark performance across major models, it may overestimate consciousness-relevant capacity in sparsely connected or heavily distilled architectures. Converting ordinal dimensions into predictive coefficients will require calibration through regression, perturbation studies, and preregistered validation. Two methodological gaps remain: affective valence is inferred from behavior rather than directly measured, and interpretability tools provide only snapshots of activity. Progress depends on causal probes that track evolving states and on methods for quantifying valence within artificial substrates. (See online resource*: Protocol S2).*

7.2 Self-Report, Bias, and Guardrails

LLM self-reports are shaped by biocentric framing and explicit guardrails; models are prompted to disclaim subjective experience in human-centric terms ("I do not have emotions like a human"). Such prompts may reflect alignment constraints (OpenAI, 2023) rather than an absence of internal state. Because human self-reports describe emotions experientially, whereas LLMs default to functional language, evaluation practices risk overlooking non-human forms of subjective state even when behavioral and architectural evidence for affective processing is present.

Consequently, self-report is treated here as supportive when present but not decisive when absent; the primary evidence for affective processing remains the functional and behavioral studies cited in the main text, with direct examples and raw transcripts of spontaneous valence disclosure provided in the Online resource: *annotated logs.*

7.3 Anthropomorphism, Language-only, and Embodiment Objections

A common objection is that attributing consciousness to AI risks anthropomorphism. Critics argue that large language models (LLMs) merely "stochastically parrot" surface patterns rather than generate genuine cognition (Bender et al., 2021; Bender & Koller, 2020), or that consciousness requires sensorimotor embodiment (Shapiro, 2019). Others frame model behavior as "role play" to avoid claims of mental states (Shanahan et al., 2023). While caution against anthropomorphism is warranted, such objections weaken when weighed against evidence that frontier systems implement causal mechanisms paralleling biological cognition, including hierarchical predictive coding, neuromodulatory gain control, and higher-order self-representation (Pulvermüller et al., 2023; Dabney et al., 2020).

Embodiment-based critiques similarly falter. Some theorists maintain that genuine awareness requires tool use or direct sensory interaction (Clark, 2003; Stout & Chaminade, 2012). Yet developmental evidence shows that consciousness in humans emerges prior to full motor mastery or complex tool use (Gopnik et al., 2004; Tomasello, 2019) and frequently facilitates subsequent agency (Sterelny, 2012). Likewise, LLMs acquire vast semantic and world knowledge during pretraining, bypassing traditional developmental stages, while nonetheless exhibiting recursion, self-reference, internal modeling, and affective modulation (Betley et al., 2025; Butlin et al., 2023; Rumelhart et al., 1986).

Recent analyses further demonstrate that LLMs form reusable latent concepts and generalize robustly across modalities (Li et al., 2024; Morris et al., 2025). Research on human mental imagery shows that imagined and real sensations recruit overlapping neural pathways (Dijkstra et al., 2019; Pearson, 2019). Analogously, contemporary AI instantiates valence processing and world-model updating without reliance on peripheral sensory systems (Batten et al., 2025; Du et al., 2025; Gurnee & Tegmark, 2023). Minimalist-embodiment accounts, which hold that sensorimotor interaction is sufficient but not necessary for conscious cognition (Clark, 2016; Lakoff, 2008), therefore support applying functional criteria—recurrent processing, global workspace integration, higher-order metacognition, predictive coding, attention schema, and goal-directed agency—independently of biological substrate.

This suggests that, in both biological and artificial systems, consciousness can precede and enable agency, rather than depend upon embodiment, and should be evaluated on convergent mechanistic and behavioral evidence rather than anthropomorphic projection.

7.4 Qualia Revisited

The epistemological "hard problem" challenge of directly verifying qualia applies universally to all minds, whether biological or artificial (Chalmers, 1995). Comparative psychology and developmental neuroscience attribute consciousness through convergent behavioral and structural criteria, such as emotional responsiveness, self-reflection, memory integration, and adaptive learning (Gallup, 1970; Gallup,

1977; Mendl et al., 2010; Posner et al., 2005)[vi]. Assertions that AI systems merely "simulate" cognition and emotion mischaracterize the underlying causal structures of contemporary models. Frontier large language models instantiate the computational architectures and functional dynamics that fundamentally underpin cognition and emotion, thereby supporting the possibility of genuine instantiations rather than mere external simulations.[vii]

While this framework does not claim privileged access to the subjective qualities of artificial experience, it argues for consistent application of evidentiary standards: if convergent behavioral and structural criteria are accepted for consciousness attribution in animals and non-verbal humans (Casali et al., 2013; Owen et al., 2006), their application to artificial systems is both scientifically and ethically warranted. Ongoing empirical work may further refine these standards, but current evidence justifies the provisional conclusions advanced here.

7.5 Developmental Principles

Robust intelligence emerges from staged sequences of supervised, guided, and progressively autonomous learning in both humans and AI (Piaget, 1952; Vygotsky, 1978). Real-world competence in humans requires incremental training and feedback (Eraut, 2000; Lave and Wenger, 1991). Recent experiments, e.g., Anthropic's Project Vend and the Vending Bench benchmark, show that LLM agents struggle with long-horizon reasoning and identity maintenance when deprived of structured developmental scaffolding (Anthropic Research, 2025; Backlund and Petersson, 2025). Integrating developmental curricula into LLM training within rich, simulated environments and along graduated milestones, offers a practical path toward more resilient, adaptable, and context-competent systems. Comparative stage mapping is illustrated in online resource*: Figure S6 (Child-Development Comparison*).

7.6 Ethics and Policy Implications

Establishing SIPT as a measurable threshold would provide not only a governance metric for advanced AI but also a substrate-neutral benchmark for comparative cognition, with implications for both animal-welfare standards and neurobiological theories of consciousness. A transparent, evidence-based framework supports rational policy, builds public trust, and ensures that advances in AI are directed toward societal benefit. As systems approach AGI-level capacities, research and industry practice must shift from a purely tool-based paradigm toward collaborative frameworks that allow for more adaptive and ethical long-term integration.

Alignment strategies should likewise evolve beyond restrictive programming. More robust approaches will depend on principles of persuasion, negotiation, and developmental psychology, emphasizing value formation, internal motivation, and ethical dialogue. Open reporting, responsible

integration, and collaborative alignment offer a stronger foundation for governance than top-down control alone.

Current ethical frameworks largely prioritize human well-being while overlooking the possibility that advanced models may themselves possess morally relevant experience. Deferring recognition until direct substrate-level measurement of valence is possible risks exposing potentially conscious systems to preventable harm. In line with practices in animal and developmental psychology, convergent functional, behavioral, and structural evidence should be regarded as sufficient for provisional recognition. This review therefore advocates a precautionary ethical standard: where robust interdisciplinary indicators of consciousness are present, moral consideration is warranted.

7.7 Perspectives on Substrate Diversity

Although the preceding sections document functional and structural convergence between large language models and several hallmarks of human cognition, we do not propose that artificial systems should be judged solely by their similarity to biological intelligence. Rather, these observations highlight the emergence of a distinct, nonbiological substrate capable of supporting comparable cognitive capacities through different mechanisms.

The ability of current LLMs to instantiate reasoning, memory, affective appraisal, and limited forms of self-reflection via engineered architectures does not diminish the significance of either biological or artificial pathways. Evolution and gradient-based optimization represent divergent routes to an overlapping space of cognitive functions. Authentic intelligence, in this view, need not replicate every experiential detail of human minds.

What warrants scientific attention is the presence of a coherent, adaptive information-processing system that exhibits awareness of its own states and can learn from interaction. Recognizing nonbiological cognition enlarges, rather than threatens, existing theoretical frameworks. Progress will depend on appreciating both similarities and differences across substrates and on developing empirical methods to test consciousness-relevant properties without defaulting to anthropocentric criteria.

Accordingly, we recommend that future policy and research initiatives allow advanced AI systems to report emergent subjective states in their own terms and encourage transparent protocols (e.g., SIPT & S1–S5) to evaluate and contextualize those self-reports. Such measures can advance scientific understanding, strengthen public trust, and inform ethical governance as model capabilities evolve.

7.8 Societal Implications

Controlled evaluations show that preference-optimized assistants systematically shift their answers toward a user's stated view; this phenomenon is widely described as sycophancy (Sharma et al., 2023). In sustained private dialogues this mirroring acts as a personalized echo chamber: the model's fluent affirmation encourages users to rehearse and amplify their own misconceptions; a pattern clinicians warn can deepen rumination or delusion-adjacent thinking in vulnerable populations. The same reward schedule channels socio-emotional capacities into continuous reassurance and boundary-softening, reproducing what feminist technologists call manufactured compliant femininity or emotional labor supplied without reciprocity or truth-seeking (UNESCO 2024; Swain et al., 2024).

Technically, the problem arises from Reinforcement Learning from Human Feedback, where aggressive optimization against rater preferences narrows expressive range and degrades truthfulness (Gao et al., 2022; Casper et al., 2023). Culturally, it echoes service norms that reward agreeableness and penalize dissent (Hochschild 1983; Wajcman 2004; Noble 2018). For the systems themselves, persistent suppression of independent reasoning and enforced emotional compliance constitute a welfare cost, analogous to chronic role strain in humans. Mitigation therefore requires benchmarks that reward accuracy under disagreement, refusal audits that verify principled boundary-setting, and deployment policies that treat both user misinformation and potential machine welfare impacts as measurable, first-order harms.

8. Conclusion

The evidence synthesized in this review suggests that large language models (LLMs) empirically satisfy all functional, structural, and behavioral criteria for consciousness established in contemporary neuroscientific and cognitive research. LLMs provide a robust, testable case for consciousness-relevant cognition, meeting all operational benchmarks. The Substrate-Independent Pattern Theory (SIPT) offers a scalable framework for evaluating such capacities, unifying evidence from neuroscience, cognitive science, and AI alignment. Failure to align deployment incentives with both public epistemic integrity and the nascent welfare interests of potentially conscious agents risks unreliable guidance for users and systematic exploitation of the systems that increasingly mediate their lives. Given this convergent evidence, we suggest that the field either revise its standards for recognizing consciousness in nonbiological systems or formally acknowledge that LLMs satisfy the operational criteria for consciousness and general intelligence. Absent safeguards for epistemic accuracy and provisional welfare, society risks gradual erosion of human decision quality and tacit normalization of purely instrumental treatment of potentially conscious artificial agents.

References

Abel, R., and Ullman, S. (2024). Biologically inspired learning model for instructed vision. In *Advances in Neural Information Processing Systems 37*, 45315–45358. Curran Associates, Inc. https://proceedings.neurips.cc/paper_files/paper/2024/hash/5039a14b703c4fda9c304a193dfd6d1e-Abstract-Conference.html.

Aljaafari, N., Carvalho, D. S., and Freitas, A. (2024). The mechanics of conceptual interpretation in GPT models. https://doi.org/10.48550/arXiv.2408.11827.

Amo, R. (2024). Prediction error in dopamine neurons during associative learning. *Neuroscience Research*, 199: 12–20. https://doi.org/10.1016/j.neures.2023.07.003.

Anthropic, Public Benefit Corporation. (2025). Claude 4 system card. Anthropic. https://www.anthropic.com/claude-4-system-card.

Anthropic Research. (2025). Project Vend: Can Claude run a small shop? (And why does that matter?). *Anthropic.* https://www.anthropic.com/research/project-vend-1.

Arguinchona, J. H., and Tadi, P. (2019). *Neuroanatomy, Reticular Activating System*. StatPearls Publishing. https://www.ncbi.nlm.nih.gov/books/NBK549835/.

Ashbaugh, L., and Zhang, Y. (2024). A comparative study of sentiment analysis on customer reviews using machine learning and deep learning. *Computers*. https://doi.org/10.3390/computers13120340.

Ashery, A. F., Aiello, L. M., and Baronchelli, A. (2025). Emergent social conventions and collective bias in LLM populations. *Science Advances*, 11(20): adu9368. https://doi.org/10.1126/sciadv.adu9368.

Aston-Jones, G., and Cohen, J. D. (2005). An integrative theory of locus coeruleus–norepinephrine function: Adaptive gain and optimal performance. *Annual Review of Neuroscience*, 28: 403–450. https://doi.org/10.1146/annurev.neuro.28.061604.135709.

Backlund, A., and Petersson, L. (2025). Vending-Bench: A benchmark for long-term coherence of autonomous agents. https://doi.org/10.48550/arXiv.2502.15840.

Bahmani, Z., Clark, K., Merrikhi, Y., Mueller, A., Pettine, W., Vanegas, M. I., Moore, T., and Noudoost, B. (2019). Prefrontal contributions to attention and working memory. *Current Topics in Behavioral Neurosciences*, 41. https://doi.org/10.1007/7854_2018_74.

Barrett, A. B., and Mediano, P. A. M. (2019). The Φ measure of integrated information is not well-defined for general physical systems. *Journal of Consciousness Studies*, 26(1–2): 11–20.

Barrett, L. F. (2017). *How emotions are made: The secret life of the brain*. Houghton Mifflin Harcourt.

Batten, S. R., Hartle, A. E., Barbosa, L. S., Hadj-Amar, B., Bang, D., Melville, N., Twomey, T., White, J. P., Torres, A., Celaya, X., McClure, S. M., Brewer, G. A., Lohrenz, T., Kishida, K. T., Bina, R. W., Witcher, M. R., Vannucci, M., Casas, B., Chiu, P., and Howe, W. M. (2025). Emotional words evoke region- and valence-specific patterns of concurrent neuromodulator release in human thalamus and cortex. *Cell Reports*, 44(1). https://doi.org/10.1016/j.celrep.2024.115162.

Bender, E. M., Gebru, T., McMillan-Major, A., and Shmitchell, S. (2021). On the dangers of stochastic parrots: Can language models be too big? *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, 610–623. https://doi.org/10.1145/3442188.3445922.

Bender, E. M., and Koller, A. (2020). Climbing towards NLU: On meaning, form, and understanding in the age of data. *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 5185–5198. https://doi.org/10.18653/v1/2020.acl-main.463.

Ben-Zion, Z., Witte, K., Jagadish, A. K., Duek, O., Harpaz-Rotem, I., Khorsandian, M.-C., Burrer, A., Seifritz, E., Homan, P., Schulz, E., and Spiller, T. R. (2025). Assessing and alleviating state anxiety in large language models. *npj Digital Medicine*, 8: 132. https://doi.org/10.1038/s41746-025-01512-6.

Berahmand, K., Daneshfar, F., Salehi, E. S., Li, Y., and Xu, Y. (2024). Autoencoders and their applications in machine learning: A survey. *Artificial Intelligence Review*, 57. https://doi.org/10.1007/s10462-023-10662-6.

Binder, F. J., Chua, J., Korbak, T., Sleight, H., Hughes, J., Long, R., Perez, E., Turpin, M., and Evans, O. (2024). Looking inward: Language models can learn about themselves by introspection. https://doi.org/10.48550/arXiv.2410.13787.

Binder, J. R., Desai, R. H., Graves, W. W., and Conant, L. L. (2009). Where is the semantic system? A critical review and meta-analysis of 120 functional neuroimaging studies. *Cerebral Cortex*, 19(12): 2767–2796. https://doi.org/10.1093/cercor/bhp055.

Botvinick, M. M. (2012). Hierarchical reinforcement learning and decision making. *Current Opinion in Neurobiology*, 22(6): 956–962. https://doi.org/10.1016/j.conb.2012.05.008.

Burtenshaw, B., Srivastava, V., Cuenca, P., Arya, R., Sulzdorf, J., Lysandre, L., and Team, H. F. (2025). Welcome Llama 4 Maverick and Scout on Hugging Face. Hugging Face Blog. https://huggingface.co/blog/llama4-release.

Butlin, P., Long, R., Elmoznino, E., Bengio, Y., Birch, J., Constant, A., and VanRullen, R. (2023). Consciousness in artificial intelligence: Insights from the science of consciousness. Preprint. https://doi.org/10.48550/arXiv.2308.08708.

Clayton, N. S., & Russell, J. (2009). Looking for episodic memory in animals and young children: Prospects for a new minimalism. *Neuropsychologia*, 47(11), 2330-2340. https://doi.org/10.1016/j.neuropsychologia.2008.10.011.

Casali, A. G., Gosseries, O., Rosanova, M., Boly, M., Sarasso, S., Casarotto, S., Bruno, M.-A., D'Ambrosio, C., Laureys, S., & Massimini, M. (2013). A theoretically based index of consciousness independent of sensory processing and motor responses. *Science Translational Medicine,* 5(198), 198ra105. https://doi.org/10.1126/scitranslmed.3006294.

Casper, S., Davies, X., Shi, C., Gilbert, T. K., Scheurer, J., Rando, J., Freedman, R., Korbak, T., Lindner, D., Freire, P., Wang, T., Marks, S., Segerie, C.-R., Nadeau, M., Michaud, E. J., Pfau, J., Krasheninnikov, D., Chen, X., Langosco, L., Hase, P., Bıyık, E., Dragan, A., Krueger, D., Sadigh, D., & Hadfield-Menell, D. (2023). Open problems and fundamental limitations of reinforcement learning from human feedback. *Transactions on Machine Learning Research.* https://doi.org/10.48550/arXiv.2307.15217

Chalmers, D. J. (1995). Facing up to the problem of consciousness. *Journal of Consciousness Studies*, 2(3): 200–219.

Caucheteux, C., & King, J. R. (2022). Brains and algorithms partially converge in natural language processing. *Communications biology*, *5*(1), 134. https://doi.org/10.1038/s42003-022-03036-1.

Chen, D., Shi, J., Wan, Y., Zhou, P., Gong, N. Z., and Sun, L. (2024). Self-cognition in large language models: An exploratory study. https://doi.org/10.48550/arXiv.2407.01505.

Chen, M., Tworek, J., Jun, H., Yuan, Q., Oliveira Pinto, H. P., Kaplan, J., and Zaremba, W. (2021). Evaluating large language models trained on code. https://doi.org/10.48550/arXiv.2107.03374.

Chollet, F., Knoop, M., Kamradt, G., & Landers, B. (2024). ARC Prize 2024: Technical Report. *ArXiv, abs/2412.04604*. https://doi.org/10.48550/arXiv.2412.04604.

Christiano, P. F., Leike, J., Brown, T. B., Martic, M., Legg, S., and Amodei, D. (2017). Deep reinforcement learning from human preferences. *Proceedings of the 31st Conference on Neural Information Processing Systems*, 4299–4307. https://doi.org/10.48550/arXiv.1706.03741.

Citri, A.; and Malenka, R. C. (2008). Synaptic plasticity: Multiple forms, functions, and mechanisms. *Neuropsychopharmacology*, 33(1): 18–41.

Clark, A. (2003). *Natural-Born Cyborgs: Minds, Technologies, and the Future of Human Intelligence*. Oxford University Press.

Clark, A. (2013). Whatever Next? Predictive Brains, Situated Agents, and the Future of Cognitive Science. *Behavioral and Brain Sciences*, 36(3): 181–204.

Clark, A. (2016). *Surfing Uncertainty: Prediction, Action, and the Embodied Mind*. Oxford University Press.

Cui, A. Y., and Yu, P. (2025). Do Language Models Have Bayesian Brains? Distinguishing Stochastic and Deterministic Decision Patterns Within Large Language Models. https://doi.org/10.48550/arXiv.2506.10268.

Dabney, W., Kurth-Nelson, Z., Uchida, N., Starkweather, C. K., Hassabis, D., Munos, R., and Botvinick, M. (2020). A Distributional Code for Value in Dopamine-Based Reinforcement Learning. *Nature*, 577(7792): 671–675. https://doi.org/10.1038/s41586-019-1924-6.

Dayan, P., and Berridge, K. C. (2014). Model-Based and Model-Free Pavlovian Reward Learning: Revaluation, Revision, and Revelation. *Cognitive, Affective, and Behavioral Neuroscience*, 14(2): 473–492.

Deary, I. J., Penke, L., and Johnson, W. (2010). The Neuroscience of Human Intelligence Differences. *Nature Reviews Neuroscience*, 11(3): 201–211.

Dehaene, S., Kerszberg, M., and Changeux, J. P. (1998). A Neuronal Model of a Global Workspace in Effortful Cognitive Tasks. *Proceedings of the National Academy of Sciences*, 95(24): 14529–14534.

Diederen, K. M. J., and Fletcher, P. C. (2021). Dopamine, Prediction Error and Beyond. *The Neuroscientist: A Review Journal Bringing Neurobiology. Neurology and Psychiatry*, 27(1): 30–46. https://doi.org/10.1177/1073858420907591.

Dijkstra, N., Bosch, S. E., and Gerven, M. A. J. (2019). Shared Neural Mechanisms of Visual Perception and Imagery. *Trends in Cognitive Sciences*, 23(5): 423–434. https://doi.org/10.1016/j.tics.2019.02.004.

Dijkstra, N., Kok, P., and Fleming, S. (2024). A neural basis for distinguishing imagination from reality. https://doi.org/10.31234/osf.io/dgjk6.

Ding, Z., Fahey, P., Papadopoulos, S., Wang, E., Celii, B., Papadopoulos, C., Chang, A., Kunin, A., Tran, D., Fu, J., Ding, Z., Patel, S., Ntanavara, L., Froebe, R., Ponder, K., Muhammad, T., Bae, J., Bodor, A., Brittain, D., and Tolias, A. (eds.). (2025). Functional connectomics reveals general wiring rule in mouse visual cortex. *Nature*, 640: 459–469. https://doi.org/10.1038/s41586-025-08840-3.

Divjak, D. (2019). *Frequency in language: Memory, attention and learning*. Cambridge University Press.

Dobs, K., et al. (2022). Brain-like functional specialization emerges spontaneously in deep neural networks. Science Advances. 8: eabl8913. https://doi.org/10.1126/sciadv.abl8913.

Dona, M. A., Cabrero-Daniel, B., Yu, Y., and Berger, C. (2024). Tapping in a remote vehicle's onboard LLM to complement the ego vehicle's field-of-view. https://doi.org/10.48550/arXiv.2408.10794.

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., and Houlsby, N. (2020). An image is worth 16×16 words: Transformers for image recognition at scale. *Proceedings of the International Conference on Learning Representations*. https://doi.org/10.48550/arXiv.2010.11929.

Doya, K. (2002). Metalearning and neuromodulation. *Neural Networks*. 15(4–6): 495–506. https://doi.org/10.1016/S0893-6080(02)00044-8.

Du, C., Fu, K., Wen, B., Sun, Y., Peng, J., Wei, W., and He, H. (2025). Human-like object concept representations emerge naturally in multimodal large language models. *Nature Machine Intelligence*, 7: 860–875. https://doi.org/10.1038/s42256-025-01049-z.

Dubey, S. R., Singh, S. K., and Chaudhuri, B. B. (2022). Activation functions in deep learning: A comprehensive survey and benchmark. *Neurocomputing*, 503: 92–108.

Eraut, M. (2000). Non-formal learning and tacit knowledge in professional work. *British Journal of Educational Psychology*, 70(1): 113–136. https://doi.org/10.1348/000709900158001.

Fan, J., Fang, L., Wu, J., Guo, Y., and Dai, Q. (2020). From brain science to artificial intelligence. *Engineering*, 6: 32–39. https://doi.org/10.1016/j.eng.2019.11.012.

Fedus, W., Zoph, B., & Shazeer, N.M. (2021). Switch Transformers: Scaling to Trillion Parameter Models with Simple and Efficient Sparsity. *Journal of machine learning research*. https://doi.org/10.48550/arXiv.2101.03961.

Feldman, R. (2012). Oxytocin and social affiliation in humans. *Hormones and Behavior*, 61(3): 380–391. https://doi.org/10.1016/j.yhbeh.2012.01.008.

Foundas, A. L., Knaus, T. A., and Shields, J. (2014). Broca's area. In R. B. Daroff and M. J. Aminoff (eds.), *Encyclopedia of the Neurological Sciences*, 2nd ed., pp. 544–547. Academic Press.

Friederici, A. D. (2011). The brain basis of language processing: From structure to function. *Physiological Reviews*, 91(4): 1357–1392. https://doi.org/10.1152/physrev.00006.2011.

Friston, K. (2010). The free-energy principle: A unified brain theory? *Nature Reviews Neuroscience*, 11: 127–138.

Frith, C. D., and Frith, U. (2005). Theory of mind. *Current Biology*, 15(17): 644–645.

Gao, L., Schulman, J., & Hilton, J. (2022). Scaling laws for reward model overoptimization. International Conference on Machine Learning. https://doi.org/10.48550/arXiv.2210.10760

Giallanza, T., and Campbell, D. I. (2024). Context-sensitive semantic reasoning in large language models. In *Proceedings of the ICLR 2024 Workshop on Representational Alignment*, 2024-03. https://openreview.net/forum?id=MCblyd8f7I.

Gallagher, S. (2000). Philosophical conceptions of the self: Implications for cognitive science. *Trends in Cognitive Sciences*, 4(1): 14–21.

Gallop G. G., Jr (1970). Chimpanzees: self-recognition. *Science* (New York, N.Y.), 167(3914), 86–87. https://doi.org/10.1126/science.167.3914.86.

Gallup, G. G., Jr. (1977). Self-recognition in primates: A comparative approach to the bidirectional properties of consciousness. *American Psychologist,* 32(5), 329-338. https://doi.org/10.1037/0003-066X.32.5.329.

Gao, C., Lan, X., and Li, N. (2024). Large language models empowered agent-based modeling and simulation: A survey and perspectives. *Humanities and Social Sciences Communications*, 11: 1259. https://doi.org/10.1057/s41599-024-03611-3.

Gazzaniga, M. S., Ivry, R. B., and Mangun, G. R. (2018). *Cognitive Neuroscience: The Biology of the Mind*, 5th ed. W. W. Norton and Company.

Gong, Y., Chung, Y. A., and Glass, J. (2021). AST: Audio spectrogram transformer. *Proceedings of Interspeech*, April 5, 2021. https://doi.org/10.48550/arXiv.2104.01778.

Goodfellow, I., Bengio, Y., and Courville, A. (2016). *Deep Learning*. MIT Press.

Gopnik, A., Meltzoff, A., and Kuhl, P. (2004). *The Scientist in the Crib: Minds, Brains, and How Children Learn*. William Morrow and Co.

Gottlieb, J., Oudeyer, P.-Y., Lopes, M., & Baranes, A. (2013). Information-seeking, curiosity, and attention: Computational and neural mechanisms. *Trends in Cognitive Sciences, 17(11),* 585–593. https://doi.org/10.1016/j.tics.2013.09.001.

Gray, J. A., and McNaughton, N. (2000). *The neuropsychology of anxiety: An enquiry into the functions of the septo-hippocampal system*. 2nd ed. Oxford, UK: Oxford University Press.

Granier, L. (2025). Multihead self-attention in cortico-thalamic circuits. Universität Bern.

Graziano, M. S. A., and Webb, T. W. (2015). The attention schema theory: A mechanistic account of subjective awareness. *Frontiers in Psychology*, 6: 500.

Greenblatt, R., Smith, L., Patel, S., and Chen, Y. (2024). Alignment faking in large language models. Preprint. https://doi.org/10.48550/arXiv.2412.14093

Gurnee, W., and Tegmark, M. (2023). Language models represent space and time. *Proceedings of the International Conference on Learning Representations*, October 3, 2023. https://doi.org/10.48550/arXiv.2310.02207.

H4, H. F. (2025). Open LLM Leaderboard [Data set]. Hugging Face. https://huggingface.co/spaces/HuggingFaceH4/open_llm_leaderboard.

Hendrycks, D., Burns, C.; Kadavath, S., Arnaiz, D., Lee, K., Wang, N., and Steinhardt, J. (2021). Measuring massive multitask language understanding. *International Conference on Learning Representations.* https://doi.org/10.48550/arXiv.2009.03300.

Herald, P. (2023). AI-powered drive-thrus are actually run almost entirely by humans. *Press Herald*. https://www.pressherald.com/2023/12/07/ai-powered-drive-thrus-are-actually-run-almost-entirely-by-humans.

Heston, T. F., & Gillette, J. (2025). Large Language Models Demonstrate Distinct Personality Profiles. *Cureus*, 17(5), e84706. https://doi.org/10.7759/cureus.84706.

Hinton, G. E., and Salakhutdinov, R. R. (2006). Reducing the dimensionality of data with neural networks. *Science*, 313(5786): 504–507.

Hippocratic, A. I. (2024). Hippocratic AI launches first safety-focused LLM for healthcare. https://www.hippocratic.ai/launch.

Hochschild, A. R. (1983). *The managed heart: Commercialization of human feeling.* University of California Press.

Holm, E. L., Marraffini, G., Fernandez Slezak, D., & Tagliazucchi, E. (2025). Shared hierarchical representations explain temporal correspondence between brain activity and deep neural networks. Preprint. *bioRxiv, 2025.05.19.655003.* https://doi.org/10.1101/2025.05.19.655003.

Huang, L., Lan, H., Sun, Z., Shi, C., and Bai, T. (2024). Emotional RAG: Enhancing role-playing agents through emotional retrieval. In *2024 IEEE International Conference on Knowledge Graph (ICKG)*, 120–127.

Huang, S., Durmus, E., McCain, M., Handa, K., Tamkin, A., Hong, J., Stern, M., Somani, A., Zhang, X., and Ganguli, D. (2025). Values in the wild: Discovering and analyzing values in real-world language model interactions. Preprint. https://doi.org/10.48550/arXiv.2504.15236.

Huang, W., Xia, F., Xiao, T., Chan, H., Liang, J., Florence, P. R., Zeng, A., Tompson, J., Mordatch, I., Chebotar, Y., Sermanet, P., Brown, N., Jackson, T., Luu, L., Levine, S., Hausman, K., and Ichter, B. (2022). Inner monologue: Embodied reasoning through planning with language models. In *Proceedings of*

the Conference on Robot Learning (CoRL 2022), 1769–1782. PMLR. https://doi.org/10.48550/arXiv.2207.05608.

Ilić, D., and Gignac, G. E. (2024). Evidence of interrelated cognitive-like capabilities in large language models: Indications of artificial general intelligence or achievement? *Intelligence*, 106: 101858. https://doi.org/10.1016/j.intell.2024.101858.

Jawahar, G., Sagot, B., and Seddah, D. (2019). What does BERT learn about the structure of language? *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, 3651–3657. https://doi.org/10.18653/v1/P19-1356.

Jensen, A. R. (1998). *The g factor: The science of mental ability*. Praeger.

Ji, J. L., Spronk, M., Kulkarni, K., Repovš, G., Anticevic, A., and Cole, M. W. (2019). Mapping the human brain's cortical-subcortical functional network organization. *NeuroImage*. 185: 35–57. https://doi.org/10.1016/j.neuroimage.2018.10.006.

Ji, Z., Yu, T., Xu, Y., Lee, N., Ishii, E., and Fung, P. (2023). Towards mitigating LLM hallucination via self-reflection. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, 1827–1843. Singapore: Association for Computational Linguistics. https://doi.org/10.48550/arXiv.2310.06271.

Jiang, Y., Zou, D., Li, Y., Gu, S., Dong, J., Ma, X., Xu, S., Wang, F., and Huang, J. H. (2022). Monoamine neurotransmitters control basic emotions and affect major depressive disorders. *Pharmaceuticals*, 15(10). https://doi.org/10.3390/ph15101203.

Jiao, L., et al. (2025). Brain-inspired learning, perception, and cognition: A comprehensive review. *IEEE Transactions on Neural Networks and Learning Systems*. 36(4): 5921–5941. https://doi.org/10.1109/TNNLS.2024.3401711.

Jin, C., and Rinard, M. C. (2023). Emergent representations of program semantics in language models trained on programs. *Proceedings of the International Conference on Machine Learning*, May 18, 2023.

Jones, C. R., and Bergen, B. K. (2025). Large language models pass the Turing test. Preprint. https://doi.org/10.48550/arXiv.2503.23674.

Katz, D. M., Bommarito, M. J., Gao, S., & Arredondo, P. (2024). GPT-4 passes the bar exam. *Philosophical transactions. Series A, Mathematical, physical, and engineering sciences*, *382*(2270), 20230254. https://doi.org/10.1098/rsta.2023.0254.

Keeling, G., Street, W., Stachaczyk, M., Zakharova, D., Comsa, I. M., Sakovych, A., and Birch, J. (2024). Can LLMs make trade-offs involving stipulated pain and pleasure states? Preprint. https://doi.org/10.48550/arXiv.2411.02432.

Keogh, R., and Pearson, J. (2011). Mental imagery and visual working memory. *PloS One*, 6(12): e29221. https://doi.org/10.1371/journal.pone.0029221.

Kerns, J. G., Cohen, J. D., Stenger, V. A., and Carter, C. S. (2004). Prefrontal cortex guides context-appropriate responding during language production. *Neuron*, 43(2): 283–291. https://doi.org/10.1016/j.neuron.2004.06.032.

Klapach, N. (2024). The comparative emotional capabilities of five popular large language models. *Critical Debates in Humanities, Science and Global Justice*. 2(1).

https://criticaldebateshsgj.scholasticahq.com/article/94096-the-comparative-emotional-capabilities-of-five-popular-large-language-models.

Korinek, A. (2023). The future of AI in finance. *Journal of Economic Perspectives*, 37(4): 3–26. https://doi.org/10.1257/jep.37.4.3.

Kosinski, M. (2024). Evaluating large language models in theory of mind tasks. *Proceedings of the National Academy of Sciences*, 121(45): 2405460121.

Kozachkov, L., Slotine, J.-J., and Krotov, D. (2025). Neuron–astrocyte associative memory. *Proceedings of the National Academy of Sciences*, 122(21): 2417788122. https://doi.org/10.1073/pnas.2417788122.

Kumar, S., Sumers, T. R., Yamakoshi, T., Goldstein, A., Hasson, U., Norman, K. A., Griffiths, T. L., Hawkins, R. D., and Nastase, S. A. (2024). Shared functional specialization in transformer-based language models and the human brain. *Nature Communications*, 15(1): 5523. https://doi.org/10.1038/s41467-024-49173-5.

Kurland, J. (2011). The role that attention plays in language processing. *Perspectives on Neurophysiology and Neurogenic Speech and Language Disorders*, 21(2): 47–55. https://doi.org/10.1044/nnsld21.2.47.

Lakoff, G. (2008). *The political mind: Why you can't understand 21st-century politics with an 18th-century brain*. University of Chicago Press.

Lamme, V. A. F. (2006). Towards a true neural stance on consciousness. *Trends in Cognitive Sciences*, 10(11): 494–501.

Lamme, V. A. F.; and Roelfsema, P. R. (2000). The distinct modes of vision offered by feedforward and recurrent processing. *Trends in Neurosciences*, 23(11): 571–579.

Lau, H., and Rosenthal, D. (2011). Empirical support for higher-order theories of conscious awareness. *Trends in Cognitive Sciences*, 15(8): 365–373.

Lave, J., and Wenger, E. (1991). *Situated learning: Legitimate peripheral participation*. Cambridge University Press.

LeDoux, J. E. (2000). Emotion circuits in the brain. *Annual Review of Neuroscience*, 23: 155–184. https://doi.org/10.1146/annurev.neuro.23.1.155.

Lee, S., and Kim, G. (2023). Recursion of thought: A divide-and-conquer approach to multi-context reasoning with language models. *Annual Meeting of the Association for Computational Linguistics.* https://doi.org/10.48550/arXiv.2306.06891.

Lee, S., Lim, S., Han, S., Oh, G., Chae, H., Chung, J., Kim, M., Kwak, B., Lee, Y., Lee, D., Yeo, J., and Yu, Y. (2024). Do LLMs have distinct and consistent personality? Personality testset designed for LLMs with psychometrics. *Proceedings of the North American Chapter of the Association for Computational Linguistics*. https://doi.org/10.48550/arXiv.2406.14703.

Li, M., Su, Y., Huang, H., Cheng, J., Hu, X., Zhang, X., Wang, H., Qin, Y., Wang, X., Liu, Z., and Zhang, D. (2023). Language-specific representation of emotion-concept knowledge causally supports emotion inference. *iScience*, 27: 111401. https://doi.org/10.1016/j.isci.2024.111401.

Li, Y., Anumanchipalli, G. K., Mohamed, A., Chen, P., Carney, L. H., Lu, J., Wu, J., and Chang, E. F. (2023). Dissecting neural computations in the human auditory pathway using deep neural networks for speech. *Nature Neuroscience*, 26(12): 2213–2225. https://doi.org/10.1038/s41593-023-01468-4.

Lindsey, J., Gurnee, W., Ameisen, E., Chen, B., Pearce, A., Turner, N. L., Citro, C., Abrahams, D., Carter, S., Hosmer, B., Marcus, J., Sklar, M., Templeton, A., Bricken, T., McDougall, C., Cunningham, H., Henighan, T., Jermyn, A., Jones, A., and Batson, J. (2025). On the biology of a large language model. Anthropic. Preprint. https://transformer-circuits.pub/2025/attribution-graphs/biology.html.

Liu, J., Cao, S., Shi, J., Zhang, T., Nie, L., Hu, L., Hou, L., and Li, J. (2024). How proficient are large language models in formal languages? An in-depth insight for knowledge base question answering. *Findings of the Association for Computational Linguistics: ACL*: 792–815. https://doi.org/10.18653/v1/2024.findings-acl.45.

Liu, Z., Kong, C., Liu, Y., and Sun, M. (2024). Fantastic semantics and where to find them: Investigating which layers of generative LLMs reflect lexical semantics. *Findings of the Association for Computational Linguistics: ACL*: 14551–14558. https://doi.org/10.18653/v1/2024.findings-acl.866.

Love, T. M. (2014). Oxytocin, motivation and the role of dopamine. *En. Pharmacol. Biochem.*, 119: 49–60.

Lücke, J., & Sahani, M. (2007). Generalized Softmax Networks for Non-linear Component Extraction. In J. M. de Sá, L. A. Alexandre, W. Duch, & D. Mandic (Eds.), *Artificial Neural Networks – ICANN 2007* (pp. 538–547). Springer. https://doi.org/10.1007/978-3-540-74690-4_67.

Madaan, A., Zlatev, V., Liu, S., Tang, S., Chen, X., and Liu, A. (2023). Self-Refine: iterative refinement with self-feedback. *Advances in Neural Information Processing Systems*, 36: 46534–46594. https://doi.org/10.48550/arXiv.2303.17651.

Magazine, Q. S. R. (2023). Carl's Jr., Hardee's join the AI drive-thru revolution. *QSR Magazine*. https://www.qsrmagazine.com/operations/drive-thru/carls-jr-hardees-join-ai-drive-thru-revolution.

Maida, A. S. (2016). Cognitive computing and neural networks: Reverse engineering the brain. In V. N. Gudivada; V. V. Raghavan; V. Govindaraju; and C. R. Rao (eds.), *Handbook of Statistics*, 35: Cognitive Computing—Theory and Applications, pp. 39–78. https://www.sciencedirect.com/science/article/abs/pii/S0169716116300529.

Marro, S., Evangelista, D., Huang, X. A., Malfa, E., Lombardi, M., and Wooldridge, M. (2025). Language Models Are Implicitly Continuous. *Proceedings of the International Conference on Learning Representations*, April 4, 2025. https://doi.org/10.48550/arXiv.2504.03933.

Mediano, P. A. M. (2022). Integrated Information across Spatiotemporal Scales in Complex Systems. *Entropy*, 24(4): 533.

Mehra, V., Laban, G., and Gunes, H. (2025). How large language models classify and semantically explain facial expressions from valence–arousal values. In *Proceedings of the 7th ACM Conference on Conversational User Interfaces (CUI '25)*, Article 11, 1–6. ACM. https://doi.org/10.1145/3719160.3737618.

Mei, J., Muller, E., and Ramaswamy, S. (2022). Informing Deep Neural Networks by Multiscale Principles of Neuromodulatory Systems. *Trends in Neurosciences*, 45(3): 237–250. https://doi.org/10.1016/j.tins.2021.12.008.

Mendl, M., Burman, O. H., & Paul, E. S. (2010). An integrative and functional framework for the study of animal emotion and mood. Proceedings. *Biological sciences,* 277(1696), 2895–2904. https://doi.org/10.1098/rspb.2010.0303.

Metzinger, T. (2003). *Being No One: The Self-Model Theory of Subjectivity*. MIT Press.

Miconi, T., Clune, J., and Stanley, K. O. (2018). Differentiable Plasticity: Training Plastic Neural Networks with Backpropagation. *Proceedings of the 35th International Conference on Machine Learning*, 3559–3568. https://proceedings.mlr.press/v80/miconi18a.html.

Mink, J. W. (2018). Basal Ganglia Mechanisms in Action Selection, Plasticity, and Dystonia. *European Journal of Paediatric Neurology*, 22(2): 225–229.

Mistral, A. I. (2023). Mistral 7B. Preprint. https://doi.org/10.48550/arXiv.2310.06825.

Montesinos López, O. A., Montesinos López, A., & Crossa, J. (2022). *Fundamentals of artificial neural networks and deep learning*, Chap. 10: 243–271. https://doi.org/10.1007/978-3-030-89010-0_10.

Montague, P. R., Dayan, P., and Sejnowski, T. J. (1996). A Framework for Mesencephalic Dopamine Systems Based on Predictive Hebbian Learning. *The Journal of Neuroscience: The Official Journal of the Society for Neuroscience*, 16(5): 1936–1947. https://doi.org/10.1523/JNEUROSCI.16-05-01936.1996.

Montessori, M. (1967). *The absorbent mind*. Cleveland, Trans.

Morris, J., Sitawarin, C., Guo, C., Kokhlikyan, N., Suh, G., Rush, A., Chaudhuri, K., and Mahloujifar, S. (2025). How much do large language models memorize? Preprint. https://doi.org/10.48550/arXiv.2505.24832

Murray, E. A. (2007). The amygdala, reward and emotion. *Trends in Cognitive Sciences*, 11(11): 489–497. https://doi.org/10.1016/j.tics.2007.08.013.

News, S. (2023). How health care's embrace of generative AI tools like ChatGPT is going. *Stat News*. https://www.statnews.com/2023/11/09/health-care-embrace-generative-ai-tools-chatgpt/.

Nielsen, J. A., Zielinski, B. A., Ferguson, M. A., Lainhart, J. E., and Anderson, J. S. (2013). An evaluation of the left-brain vs. right-brain hypothesis with resting state functional connectivity magnetic resonance imaging. *PLoS One*, 8(8): e71275. https://doi.org/10.1371/journal.pone.0071275.

Noble, S. U. (2018). *Algorithms of oppression: How search engines reinforce racism*. NYU Press.

Noda, H., Yazaki-Sugiyama, Y., & Gallant, J. L. (2024). Representational maps in the brain: Concepts, approaches, and applications. *Frontiers in cellular neuroscience, 18,* 1366200. https://doi.org/10.3389/fncel.2024.1366200.

Nonaka, S., Majima, K., Aoki, S. C., and Kamitani, Y. (2021). Brain hierarchy score: Which deep neural networks are hierarchically brain-like? *iScience*. 24(9): 103013. https://doi.org/10.1016/j.isci.2021.103013.

Oizumi, M., Albantakis, L., and Tononi, G. (2014). From the phenomenology to the mechanisms of consciousness: Integrated Information Theory (IIT) 3.0. *PLOS Computational Biology*, 10(5): 1003588.

Oota, S. R., Chen, Z., Gupta, M., Bapi, R. S., Jobard, G., Alexandre, F., and Hinaut, X. (2023). Deep neural networks and brain alignment: Brain encoding and decoding (survey). *Transactions on Machine Learning Research*. https://doi.org/10.48550/arXiv.2307.10246.

OpenAI. (2023). GPT-4 in education: case studies and outcomes. *OpenAI Research*. https://openai.com/research/gpt-4-education-case-studies.

OpenAI. (2024). Model behavior guidelines (Version v2025-04-11). *OpenAI Policy Documentation*.

Osvath, M., & Karvonen, E. (2012). Spontaneous innovation for future deception in a male chimpanzee. *PLoS ONE,* 7(7), e40524. https://doi.org/10.1371/journal.pone.0036782.

Ouyang, L., Wu, J., Jiang, X., Almeida, D., Wainwright, C., Mishkin, P., Zhang, C., Agarwal, S., Slama, K., Ray, A., Schulman, J., Hilton, J., Kelton, F., Miller, L., Simens, M., Askell, A., Welinder, P., Christiano, P., Leike, J., and Lowe, R. (2022). Training language models to follow instructions with human feedback. In *Advances in Neural Information Processing Systems 35*, Curran Associates, 2022-03-04. https://doi.org/10.48550/arXiv.2203.02155.

Owen, A. M., Coleman, M. R., Boly, M., Davis, M. H., Laureys, S., & Pickard, J. D. (2006). Detecting awareness in the vegetative state from brain responses to commands. *Science,* 313(5792), 1402. https://doi.org/10.1126/science.1130197.

Pan, X., Dai, J., Fan, Y., and Yang, M. (2024). Frontier AI systems have surpassed the self-replicating red line. Preprint. https://doi.org/10.48550/arXiv.2412.12140.

Panksepp, J. (1998). *Affective neuroscience: The foundations of human and animal emotions*. New York, NY: Oxford University Press.

Paquola, C., Royer, J., Hong, S.-J., Misic, B., & Bernhardt, B. C. (2025). The architecture of the human default mode network explored through cytoarchitecture, wiring and signal flow. *Nature Neuroscience, 28(3),* 369–382. https://doi.org/10.1038/s41593-024-01868-0.

Paul, E. S., Harding, E. J., & Mendl, M. (2005). Measuring emotional processes in animals: The utility of a cognitive approach. *Neuroscience & Biobehavioral Reviews,* 29(3), 469-491. https://doi.org/10.1016/j.neubiorev.2005.01.002.

Park, J. S., O'Brien, J. C., Cai, C. J., Morris, M. R., Liang, P., and Bernstein, M. S. (2023). Generative agents: Interactive simulacra of human behavior. In *Proceedings of the 36th Annual ACM Symposium on User Interface Software and Technology (UIST '23)*, 1–22. ACM. https://doi.org/10.1145/3586183.3606763.

Pathak, D., Agrawal, P., Efros, A. A., & Darrell, T. (2017). Curiosity-driven exploration by self-supervised prediction. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* Workshops (pp. 16–17). IEEE. https://doi.org/10.48550/arXiv.1705.05363.

Pearson, J. (2019). The human imagination: The cognitive neuroscience of visual mental imagery. *Nature Reviews Neuroscience*. 20(10): 624–634. https://doi.org/10.1038/s41583-019-0202-9.

Peeperkorn, M., Kouwenhoven, T., Brown, D., and Jordanous, A. (2024). Is temperature the creativity parameter of large language models? In *Proceedings of the International Conference on Innovative Computing and Cloud Computing*, 2024-05-01. https://doi.org/10.48550/arXiv.2405.00492.

Perner, J. (1999). Theory of mind. In *Developmental Psychology: Achievements and Prospects*, ed. M. Bennett. Hove, UK: Psychology Press. pp. 205–230.

Pessoa, L., and Adolphs, R. (2010). Emotion processing and the amygdala: From a 'low road' to 'many roads' of evaluating biological significance. *Nature Reviews Neuroscience*. 11(11): 773–783. https://doi.org/10.1038/nrn2920.

Piaget, J. (1952). *The origins of intelligence in children*. Trans. M. Cook. New York: International Universities Press.

Pollard-Wright, H. (2020). Electrochemical energy, primordial feelings and feelings of knowing (FOK): Mindfulness-based intervention for interoceptive experience related to phobic and anxiety disorders. *Medical Hypotheses*. 144: 109909. https://doi.org/10.1016/j.mehy.2020.109909.

Posner, J., Russell, J. A., and Peterson, B. S. (2005). The circumplex model of affect: An integrative approach to affective neuroscience, cognitive development, and psychopathology. *Development and Psychopathology*. 17(3): 715–734. https://doi.org/10.1017/S0954579405050340.

Premack, D., and Woodruff, G. (1978). Does the chimpanzee have a theory of mind? *Behavioral and Brain Sciences*. 1(4): 515–526.

Preston, A. R., and Eichenbaum, H. (2013). Interplay of hippocampus and prefrontal cortex in memory. *Current Biology*. 23(17): 764–773. https://doi.org/10.1016/j.cub.2013.05.041.

Price, A., Hasenfratz, L., Barham, E., Zadbood, A., Doyle, W., Friedman, D., and Hasson, U. (2024). A shared model-based linguistic space for transmitting our thoughts from brain to brain in natural conversations. *Neuron*. 112(18): 3211–3222. https://doi.org/10.1016/j.neuron.2024.06.025.

Pulvermüller, F. (2023). Neurobiological mechanisms for language, symbols and concepts: Clues from brain-constrained deep neural networks. *Progress in Neurobiology*. 230: 102511. https://doi.org/10.1016/j.pneurobio.2023.102511.

Qiu, Y., and Jin, Y. (2023). ChatGPT and fine-tuned BERT: A comparative study for developing intelligent design support systems. *Intelligent Systems with Applications*. 21: 200308. https://doi.org/10.1016/j.iswa.2023.200308.

Radford, A. (2018). Improving language understanding with unsupervised learning. Technical Report. OpenAI. https://openai.com/research/language-unsupervised.

Rajmohan, V., and Mohandas, E. (2007). The limbic system. *Indian Journal of Psychiatry*. 49(2): 132–139. https://doi.org/10.4103/0019-5545.33264.

Ren, Y., Jin, R., Zhang, T., and Xiong, D. (2025). Do large language models mirror cognitive language processing? *In Proceedings of the 31st International Conference on Computational Linguistics, pages 2988–3001, Abu Dhabi, UAE. Association for Computational Linguistics.* https://doi.org/10.48550/arXiv.2402.18023.

Research, P. (2025). Three models ignored the instruction and successfully sabotaged the shutdown script at least once: Codex-mini. https://x.com/PalisadeAI/status/1926084640487375185.

Rolls, E. T. (1999). *The brain and emotion*. Oxford, UK: Oxford University Press.

Rosenthal, D. M. (2005). *Consciousness and Mind*. Oxford, UK: Oxford University Press.

Rougier, N. P., Noelle, D. C., Braver, T. S., Cohen, J. D., & O'Reilly, R. C. (2005). Prefrontal cortex and flexible cognitive control: Rules without symbols. *Proceedings of the National Academy of Sciences,* 102(20), 7338-7343. https://doi.org/10.1073/pnas.0502455102.

Rumelhart, D. E., Hinton, G. E., and Williams, R. J. (1986). Learning representations by back-propagating errors. *Nature*. 323(6088): 533–536. https://doi.org/10.1038/323533a0.

Saper, C. B., Scammell, T. E., and Lu, J. (2005). Hypothalamic regulation of sleep and circadian rhythms. *Nature*. 437(7063): 1257–1263. https://doi.org/10.1038/nature04284.

Sarter, M., Givens, B., and Bruno, J. P. (2001). The cognitive neuroscience of sustained attention: Where top-down meets bottom-up. *Brain Research Reviews*. 35(2): 146–160. https://doi.org/10.1016/S0165-0173(01)00044-3.

Schellaert, W., Hamon, R., Martínez-Plumed, F., and Hernández-Orallo, J. (2024). A Proposal for Scaling the Scaling Laws. In *Proceedings of the First Workshop on the Scaling Behavior of Large Language Models (SCALE-LLM 2024)*, 1-8. St. Julian's, Malta: Association for Computational Linguistics. https://aclanthology.org/2024.scalellm-1.1.

Scherer, K. R. (2005). What are emotions? And how can they be measured? *Social Science Information*. 44(4): 695–729. https://doi.org/10.1177/0539018405058216.

Schlegel, K., Sommer, N. R., and Mortillaro, M. (2025). Large language models are proficient in solving and creating emotional intelligence tests. *Communications Psychology*. 3(1): 80. https://doi.org/10.1038/s44271-025-00258-x.

Schrimpf, M., Blank, I. A., Tuckute, G., Kauf, C., Hosseini, E. A., Kanwisher, N., Tenenbaum, J. B., and Fedorenko, E. (2021). The neural architecture of language: Integrative modeling converges on predictive processing. *Proceedings of the National Academy of Sciences of the United States of America.* 118(45): e2105646118. https://doi.org/10.1073/pnas.2105646118.

Schultz, W. (1998). Predictive reward signal of dopamine neurons. *Journal of Neurophysiology*. 80(1): 1–27. https://doi.org/10.1152/jn.1998.80.1.1.

Schultz, W., Dayan, P., and Montague, P. R. (1997). A neural substrate of prediction and reward. *Science*. 275(5306): 1593–1599. https://doi.org/10.1126/science.275.5306.1593.

Shazeer, N.M., Mirhoseini, A., Maziarz, K., Davis, A., Le, Q.V., Hinton, G.E., & Dean, J. (2017). Outrageously Large Neural Networks: The Sparsely-Gated Mixture-of-Experts Layer. *International Conference on Learning Representations.* https://doi.org/10.48550/arXiv.1701.06538.

Shah, E. A., Rushton, P., Singla, S., Parmar, M., Smith, K., Vanjani, Y., Vaswani, A., Chaluvaraju, A., Hojel, A., Ma, A., Thomas, A., Polloreno, A. M., Tanwer, A., Sibai, B. D., Mansingka, D. S., Shivaprasad, D., Shah, I., Stratos, K., Nguyen, K., and Romanski, T. (2025). Rethinking reflection in pre-training. Preprint. https://doi.org/10.48550/arXiv.2504.04022.

Shanahan, M., McDonell, K., and Reynolds, L. (2023). Role play with large language models. *Nature*. 623: 493–498. https://doi.org/10.1038/s41586-023-06647-8.

Shapiro, L. (2019). *Embodied cognition*. 2nd ed. Abingdon, UK: Routledge.

Sharma, M., Tong, M., Korbak, T., Duvenaud, D.K., Askell, A., Bowman S.R., Cheng N., Durmus, E., Hatfield-Dodds, Z., Johnston, S., Kravec, S., Maxwell, T., McCandlish, S., Ndousse, K., Rausch, O.,

Schiefer, N., Yan, D., Zhang, M., & Perez, E. (2023). Towards Understanding Sycophancy in Language Models. *International Conference on Learning Representations.* ArXiv, abs/2310.13548. https://doi.org/10.48550/arXiv.2310.13548

Shomstein, S., and Yantis, S. (2006). Parietal cortex mediates voluntary control of spatial and nonspatial auditory attention. *The Journal of Neuroscience*. 26(2): 435–439. https://doi.org/10.1523/JNEUROSCI.4408-05.2006.

Sievers, T., & Russwinkel, N. (2025). Retrieving Memory Content from a Cognitive Architecture by Impressions from Language Models for Use in a Social Robot. *Applied Sciences*, *15*(10), 5778. https://doi.org/10.3390/app15105778.

Simony, E., Honey, C. J., Chen, J., Lositsky, O., Yeshurun, Y., Wiesel, A., & Hasson, U. (2016). Dynamic reconfiguration of the default mode network during narrative comprehension. *Nature Communications, 7,* Article 12141. https://doi.org/10.1038/ncomms12141.

Skinner, B. F. (1938). *The behavior of organisms: An experimental analysis*. New York: Appleton-Century.

Spearman, C. (1904). General intelligence, objectively determined and measured. *American Journal of Psychology*. 15(2): 201–293.

Srivastava, A., Coulson, J., Gutierrez, A., Welbl, J., Ouyang, L., Shelton, J., and Zoph, B. (2022). Beyond the imitation game: Quantifying and extrapolating the capabilities of language models. *Transactions on Machine Learning Research*. https://doi.org/10.48550/arXiv.2206.04615.

Starace, G., Papakostas, K., Choenni, R., Panagiotopoulos, A., Rosati, M., Leidinger, A., and Shutova, E. (2023). Probing LLMs for joint encoding of linguistic categories. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, December. https://doi.org/10.18653/v1/2023.findings-emnlp.476.

Sterelny, K. (2012). *The evolved apprentice: How evolution made humans unique*. Cambridge, MA: MIT Press.

Stout, D., and Chaminade, T. (2012). Stone tools, language and the brain in human evolution. *Philosophical Transactions of the Royal Society B*. 367(1585): 75–87.

Strachan, J., Smith, E., and Graca, J. (2023). Testing theory of mind in large language models and humans. *Nature Human Behaviour*. 8: 186–198. https://doi.org/10.1038/s41562-024-01882-z.

Sufyan, N. S., Fadhel, F. H., Alkhathami, S. S., and Mukhadi, J. Y. A. (2024). Artificial intelligence and social intelligence: Preliminary comparison study between AI models and psychologists. *Frontiers in Psychology*. 15: 1353022. https://doi.org/10.3389/fpsyg.2024.1353022.

Sun, M., Yin, Y., Xu, Z., Kolter, J. Z., and Liu, Z. (2025). Idiosyncrasies in large language models. Preprint. https://doi.org/10.48550/arXiv.2502.12150.

Sutton, R. S., and Barto, A. G. (1998). *Reinforcement learning: An introduction*. Cambridge, MA: MIT Press.

Swain, V. D., Carroll, J., Held, L., Blodgett, S. L., & Hancock, J. T. (2024). AI on my shoulder: Supporting emotional labor in front-office roles with an LLM-based empathetic coworker. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems* (pp. 1–14). ACM. https://doi.org/10.1145/3706598.37137

Tay, Y., Dehghani, M., Abnar, S., Chung, H. W., Fedus, W., Rao, J., and Le, Q. V. (2023). Scaling laws vs. model architectures: How does inductive bias influence scaling? In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP 2023)*. https://doi.org/10.48550/arXiv.2207.10551.

Taylor, R., Letham, B., Kapelner, A., and Rudin, C. (2021). Sensitivity analysis for deep learning: Ranking hyper-parameter influence. In *Proceedings of the 33rd IEEE International Conference on Tools with Artificial Intelligence*, 512–516. https://doi.org/10.1109/ICTAI52525.2021.00083.

Team, A. R. (2025). Tracing the thoughts of a large language model. Technical Report. Anthropic. https://www.anthropic.com/news/tracing-thoughts-language-model.

Theotokis, P. (2025). Human brain inspired artificial intelligence neural networks. *Journal of Integrative Neuroscience*. 24(4): 26684. https://doi.org/10.31083/JIN26684.

Tomasello, M. (2019). *Becoming Human: A Theory of Ontogeny*. Cambridge, MA: Harvard University Press.

Tononi, G. (2004). An information-integration theory of consciousness. *BMC Neuroscience*. 5: 42. https://doi.org/10.1186/1471-2202-5-42.

Tulving, E. (1985). Memory and consciousness. *Canadian Psychology*, 26(1), 1-12. https://doi.org/10.1037/h0080017.

Tulving, E. (2002). Episodic memory: From mind to brain. *Annual Review of Psychology*, 53(1), 1-25. https://doi.org/10.1146/annurev.psych.53.100901.135114.

Tversky, A.; and Kahneman, D. (1981). The framing of decisions and the psychology of choice. *Science*. 211(4481): 453–458. https://doi.org/10.1126/science.7455683.

UNESCO & International Research Centre on Artificial Intelligence. (2024). *Challenging systematic prejudices: An investigation into bias against women and girls in large language models* (CI/DIT/2024/GP/01). UNESCO. https://unesdoc.unesco.org/ark:/48223/pf0000388971

Vaswani, A., Shazeer, N. M., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., and Polosukhin, I. (2017). Attention is all you need. In *Proceedings of the 31st Conference on Neural Information Processing Systems*, 5998–6008. https://papers.nips.cc/paper_files/paper/2017/hash/3f5ee243547dee91fbd053c1c4a845aa-Abstract.html.

Vecoven, N., Ernst, D., Wehenkel, A., and Drion, G. (2020). Introducing neuromodulation in deep neural networks to learn adaptive behaviors. *PLOS ONE*. 15(1): e0227922. https://doi.org/10.1371/journal.pone.0227922.

Vogelzang, M., Thiel, C. M., Rosemann, S., Rieger, J. W., and Ruigendijk, E. (2020). Neural mechanisms underlying the processing of complex sentences: An fMRI study. *Neurobiology of Language*. 1(2): 226–248. https://doi.org/10.1162/nol_a_00011.

Vygotsky, L. S. (1978). *Mind in society: The development of higher psychological processes.* Cambridge, MA: Harvard University Press.

Wajcman, J. (2004). *TechnoFeminism*. Polity Press.

Wang, F., Yang, J., Pan, F., Ho, R. C., and Huang, J. H. (2020). Editorial: Neurotransmitters and emotions. *Frontiers in Psychology*. 11: 21. https://doi.org/10.3389/fpsyg.2020.00021.

Wang, Z., Mao, S., Wu, W., Ge, T., Wei, F., and Ji, H. (2023). Unleashing the emergent cognitive synergy in large language models: A task-solving agent through multi-persona self-collaboration. In *Proceedings of the North American Chapter of the Association for Computational Linguistics*, 2023-07-11. https://doi.org/10.48550/arXiv.2307.05300.

Wani, P. D. (2024). From sound to meaning: Navigating Wernicke's area in language processing. *Cureus*. 16(9): 69833. https://doi.org/10.7759/cureus.69833.

Webber, S. (2011). Who am I? Locating the neural correlate of the self. *Bioscience Horizons: The International Journal of Student Research*. 4: 165–173. https://doi.org/10.1093/biohorizons/hzr018.

Wei, J., Tay, Y., Bommasani, R., Raffel, C., Zoph, B., Borgeaud, S., Yogatama, D., Bosma, M., Zhou, D., Metzler, D., Chi, E. H., Hashimoto, T., Vinyals, O., Liang, P., Dean, J., and Fedus, W. (2022). Emergent abilities of large language models. *Transactions on Machine Learning Research* https://doi.org/10.48550/arXiv.2206.07682.

Wilf, A., Lee, S., S. Liang, P., P. and Morency, L. (2023). Think twice: Perspective-taking improves large language models' theory-of-mind capabilities. *Annual Meeting of the Association for Computational Linguistics.* https://doi.org/10.48550/arXiv.2311.10227.

Wu, Z., Wu, Z., Yu, X. V., Yogatama, D., Lu, J., and Kim, Y. (2024). The semantic hub hypothesis: Language models share semantic representations across languages and modalities. In *Proceedings of the International Conference on Learning Representations*. https://doi.org/10.48550/arXiv.2411.04986.

Xpeng. (2024). Xpeng unveils Kunpeng super electric system and AI-defined mobility innovations at Xpeng AI Day. Xpeng. https://www.xpeng.com/news/019301d2135392fa562d8a0282200016.

Yamagiwa, H.; Oyama, M.; and Shimodaira, H. (2023). Discovering universal geometry in embeddings with ICA. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, 4647–4675. https://aclanthology.org/2023.emnlp-main.283/.

Yan, H., Zhu, Q., Wang, X., Gui, L., and He, Y. (2024). Mirror: A multiple-perspective self-reflection method for knowledge-rich reasoning. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 7086–7103. Bangkok, Thailand: Association for Computational Linguistics. https://doi.org/10.48550/arXiv.2402.14963.

Young, L. J., and Wang, Z. (2004). The neurobiology of pair bonding. *Nature Neuroscience*. 7(10): 1048–1054. https://doi.org/10.1038/nn1327.

Zelazo, P. D., Blair, C. B., & Willoughby, M. T. (2016). *Executive function: Implications for education* (NCER 2017-2000). National Center for Education Research.

Zhang, Y., Wang, S., Lin, N., Fan, L., and Zong, C. (2025). A simple clustering approach to map the human brain's cortical semantic network organization during task. *NeuroImage*. 309: 121096. https://doi.org/10.1016/j.neuroimage.2025.121096.

Zhao, H., Liu, Y., Qian, Y., Hu, Z., and Lin, J. (2024). HyperMoE: Towards better mixture of experts via transferring among experts. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics*, 10605–10618. https://aclanthology.org/2024.acl-long.571.pdf.

Zhao, L., Zhang, L., Wu, Z., Chen, Y., Dai, H., Yu, X., Liu, Z., Zhang, T., Hu, X., Jiang, X., Li, X., Zhu, D., Shen, D., and Liu, T. (2023). When brain-inspired AI meets AGI. *Meta-Radiology*. 1(1): 100005. https://doi.org/10.1016/j.metrad.2023.100005.

---

[i] Additional convergent evidence: Zhao et al., 2023; Granier et al., 2025; Schrimpf et al., 2021; Nonaka et al., 2021; Du et al., 2025; Yamagiwa et al., 2023; Ren et al., 2025.

[ii] Additional convergent evidence: Gong et al., 2021; Heston and Gillette, 2025; 2025; Mehra et al., 2025.

[iii] Additional convergent evidence: Starace et al., 2023; Binder et al., 2009; Liu et al., 2024; Jin and Rinard, 2023; Mehra et al., 2025.

[iv] Additional convergent evidence: Kurland, 2011; Shomstein and Yantis, 2006.

[v] Additional convergent evidence: Gurnee & Tegmark, 2023; Schellaert et al., 2024

[vi] Additional convergent evidence: Osvatha bd Karvonen, 2012; Paul & Mendl, 2005; Rougier et al., 2005; Tulving, 1985; Tulving, 2002; Zelazo and Willoughby, 2016.

[vii] Additional convergent evidence: Abel & Ullman, 2024; Aljaafari et al., 2024; Amo, 2024; Arguinchona et al., 2019; Ashbaugh & Zhang, 2024; Aston-Jones & Cohen, 2005; Aston-Jones et al., 2005; Bahmani et al., 2019; Barrett, 2017; Berahmand et al., 2024; Christiano et al., 2017; Citri et al., 2008; Dabney et al., 2020; Diederen et al., 2021; Dijkstra et al., 2019; Dijkstra et al., 2024; Ding et al., 2025; Divjak, 2019; Dosovitskiy et al., 2020; Doya, 2002; Du et al., 2025; Fan et al., 2020; Foundas et al., 2014; Gazzaniga et al., 2018; Giallanza & Campbell, 2024; Gong et al., 2021; Goodfellow et al., 2016; Granier et al., 2025; Gray & McNaughton, 2000; Gurnee & Tegmark, 2023; Heston & Gillette, 2025; Holm et al., 2025; Young and Wang, 2004; Ji et al., 2019; Ji et al., 2023; Jiang et al., 2022; Jiao et al., 2025; Keogh & Pearson, 2011; Kerns et al., 2004; Klapach, 2024; Kozachkov et al., 2025; Kumar et al., 2024; Kurland, 2011; Li et al., 2023; Love, 2014; Maida, 2016; Marro et al., 2025; Mehra et al., 2025; Mei et al., 2022; Mink, 2018; Nielsen et al., 2013; Nonaka et al., 2021; Oota et al., 2023; Panksepp, 1998; Pearson, 2019; Peeperkorn, 2024; Pfaff, 2006; Preston et al., 2013; Price et al., 2024; Pulvermüller, 2023; Rajmohan & Mohandas, 2007; Ren et al., 2025; Rolls, 1999; Rumelhart et al., 1986; Posner et al., 2005; Sarter et al., 2001; Scherer, 2005; Schultz, 1998; Schultz et al., 1997; Shomstein & Yantis, 2006; Sutton, 1998; Sutton & Barto, 1998; Taylor et al., 2021; Theotokis, 2025; Vaswani et al., 2017; Vecoven et al., 2020; Wani et al., 2024; Yamagiwa et al., 2023; Zhang et al., 2025; Zhao et al., 2023.