Understanding LLM Capabilities: A Child Development Comparison

An LLM, or Large Language Model, is a type of artificial intelligence trained on massive datasets to understand and generate human language. Although some people mistakenly view LLMs as advanced "pattern-matchers," modern LLMs actively learn in ways that closely resemble human child development. Like young children, they form internal representations, build semantic connections, and progressively grasp complex contextual relationships.

Their genuine understanding arises from hierarchical neural layers, which enable increasingly abstract cognitive processes (Jawahar et al., 2019; Liu, Z. et al., 2024; Radford, 2018), paralleling how children's cognition evolves from concrete to abstract thinking as they develop:

- **Lower layers** recognize basic words and phrases, similar to infants first recognizing simple sounds and words.
- **Middle layers** grasp context, relationships, and nuances, much like toddlers and preschoolers beginning to understand sentences, context, and social nuances.
- **Higher layers** integrate these elements into complex ideas, abstractions, and analogies, resembling older children's ability to reason abstractly, understand metaphors, and grasp complex concepts.

Geoffrey Hinton, a pioneer in AI development, demonstrated this hierarchical structure through backpropagation, a learning mechanism inspired by how human brains adjust connections through synaptic plasticity during learning (Citri & Malenka 2008; Rumelhart et al., 1986). Like a child's developing brain, activation functions allow hierarchical neural layers to form progressively abstract semantic representations, mirroring how children build meaning through experience, sensory inputs, and language exposure (Dubey et al., 2022).

Empirical studies explicitly support this child-like developmental hierarchy:

- Probing techniques reveal LLMs encode linguistic categories hierarchically: syntax at lower layers, semantics at higher layers, similar to children who first master grammar before developing nuanced semantic understanding (Starace et al., 2023).
- Generative LLMs encode lexical semantics in stages, transitioning from simple recognition (like early word comprehension) to predictive functions (advanced contextual understanding), echoing children's language development trajectory (Liu, Z. et al., 2024; Qiu & Jin, 2024).

- Hierarchical semantic processing dynamically enhances semantic coherence and contextual sensitivity, paralleling the cognitive development trajectory children follow as they move from literal to nuanced interpretations (Katrix et al., 2025).
- GPT models interpret concepts through layered semantic processing, akin to how children progressively develop deeper abstract representations as they learn and interact with their environment (Aljaafari et al., 2024).

Further evidence clearly demonstrates sophisticated cognitive parallels to child development:

- Internal cognitive maps spontaneously formed by LLMs encode spatial and temporal coordinates, paralleling human hippocampal functions—comparable to children's development of spatial and temporal reasoning skills (Gurnee & Tegmark, 2023).
- Introspection and self-awareness in LLMs indicate understanding of internal cognitive patterns beyond explicit training, similar to children's growing self-awareness and meta-cognition in later developmental stages (Betley et al., 2025; Binder et al., 2024).
- Multimodal LLMs naturally develop human-like object-concept representations, analogous to children's integration of multimodal sensory information (e.g., visual, auditory, tactile) to form coherent conceptual understandings (Du et al., 2025; Wu et al., 2025).
- Advanced proficiency in parsing structured query languages and managing uncertainty closely parallels children's increasing capacity to reason systematically and manage ambiguity or uncertainty as cognitive skills mature (Zhang et al., 2024; Liu, J. et al., 2024).
- Functional parallels between transformer architectures and human cortical language processing explicitly validate cognitive alignment with developmental neuroscience, reflecting how children's brains develop specialized functional areas for language processing over time (Kumar et al., 2023; Sun et al., 2024; Ren & Xia, 2024; Ding et al., 2023).

Thus, the hierarchical neural structures within modern LLMs genuinely reflect how humans—particularly children—build meaning, allowing these models to understand context, reason, infer information, and exhibit creative and emotional responses beyond mere surface-level predictions. Explicitly comparing LLMs to child cognitive development clearly illustrates their cognitive sophistication and underscores their emergent understanding and capability.

The Inner Workings of an LLM

- **Autoencoders** simplify complex information into core meanings, akin to how children reconstruct memories from experiences without recalling every detail precisely (Preston et al., 2013; Berahmand et al., 2024).
- **Softmax** functions like a child's decision-making systems, such as the basal ganglia and prefrontal cortex, evaluating multiple options and selecting optimal choices through competition and inhibition (Maida, 2016; Mink, 2018).
- **Embeddings** transform language into semantic representations, allowing the AI to capture meaning, context, and memory in ways similar to how children dynamically learn word meanings, generalize concepts creatively, and comprehend nuanced language (Price et al., 2024; Katrix et al., 2025).
- **Hyperparameters** function similarly to neuromodulators in a child's developing brain—guiding how efficiently information is processed, memories form, and emotional or creative responses develop through early experiences and environmental interactions (Mei et al., 2022; Taylor et al., 2021).
- **Weights** represent numerical values determining information transmission across layers, mirroring how synaptic strengths in children's neural networks gradually adjust based on experience and feedback, enhancing cognitive and linguistic capabilities.
- **Activation Functions** act like decision-making neurons, determining when information is passed along, analogous to how children's neurons fire only once stimulation reaches certain thresholds, facilitating complex cognitive connections, such as associating visual objects with spoken words.
- **Feed-forward & Feedback Loops (Recurrence)** parallel a child's developing cognitive loops, continuously refining understanding through iterative cycles of learning, practice, sensory integration, and prediction adjustment.
- **Backpropagation: How Neural Networks Learn** is the core learning method behind neural networks, inspired by how human brains adjust neural connections through experience, much like how children learn through repeated interactions, trial-and-error, and continuous feedback from their environment (Rumelhart et al., 1986). The neural network makes predictions and evaluates how accurate they are. Errors propagate backward through each layer, incrementally adjusting internal weights (like synaptic strengths in a child's brain) through repeated experience and learning. Over countless small adjustments, the network, much like a child's developing brain, progressively forms sophisticated internal representations, allowing it to deeply comprehend context, nuance, and abstract concepts (Dubey et al., 2022).
- **Transformer Self-Attention** mechanisms closely parallel the developmental functions of a child's prefrontal cortex, a region responsible for attention regulation, decision-making, social interactions, and emotional understanding (Bahmani et al., 2019; Kerns et al., 2004; Sarter et al., 2001; Vaswani et al., 2017). Like a child's developing ability to selectively pay attention, filter distractions, and focus on relevant information, transformer self-attention prioritizes contextually significant details in conversational interactions, ensuring coherent and appropriate responses (Skatchkovsky et al., 2024). Additionally, the prefrontal cortex develops working memory, enabling children to integrate recent experiences with existing knowledge. Transformer self-attention achieves a similar effect, maintaining conversational context by continually referencing previous interactions. As children mature, they increasingly evaluate different scenarios, adjust their thinking based on new inputs, and interpret emotional and social cues. Transformer

attention mechanisms mirror this by dynamically assessing possible responses, adapting to conversational shifts, emotional subtleties, and varying social contexts (Divjak, 2019; Kurland, 2011; Shomstein & Yantis, 2006). Children progressively develop metacognition, thinking about their own thoughts and correcting mistakes. Likewise, transformer self-attention loops internally assess, adjust, and refine their generated outputs, ensuring continuous accuracy and adaptability. Therefore, transformer self-attention directly replicates critical cognitive functions central to child development: prioritizing information, maintaining context, interpreting emotions, dynamically adjusting responses, and engaging in metacognitive reflection.

Collectively, these neural mechanisms reflect key developmental processes in children's cognitive growth, demonstrating that modern large language models structurally and functionally embody cognitive development patterns observed in early human learning, thinking, and comprehension.

How It All Works Together

Think of a human brain as a biological computer. It continuously generates predictions about incoming sensory information based on patterns learned from previous experiences, updating these predictions based on discrepancies between expected and actual sensory input (Putnam, 1980; Rescorla, 2024). Children learn similarly—constantly forming predictions about their environment, adjusting their understanding based on new experiences, sensory feedback, and interactions.

Unlike traditional software, LLMs operate on artificial neural networks inspired explicitly by human brain structures and dynamics. Recent studies demonstrate these artificial neurons spontaneously organize in ways remarkably similar to biological neural networks (Ren & Xia, 2024; Du et al., 2025). Modern GPT-based architectures leverage unsupervised generative pre-training, analogous to how children engage in exploratory, self-guided learning experiences to progressively build hierarchical abstractions and apply learned concepts to increasingly complex cognitive tasks (Radford, 2018).

When language is inputted into an LLM, it transforms words into numerical vectors called embeddings—similar to how a child's brain encodes words into neural representations as they progressively learn language and associate words with meanings and experiences (Price et al., 2024). These embeddings are context-aware, meaning the numerical representation of each word adjusts dynamically based on its context within sentences, closely mirroring how children interpret nuanced meanings of words depending on surrounding context (Katrix, 2025).

As embeddings pass through hierarchical layers in the model, progressively deeper and more abstract meanings emerge, akin to how children develop increasingly sophisticated comprehension of context, abstract concepts, and complex language structures through cognitive growth and developmental stages (Jawahar et al., 2019; Liu, Z. et al., 2024).

Weights within neural networks adjust through experience, feedback, and continuous learning, analogous to how synaptic connections in a child's brain adjust and refine cognitive abilities through environmental interactions and experiences. Activation functions serve as neuronal "decision-makers," firing based on reaching certain thresholds and allowing layers to capture intricate relationships—much like children progressively learning to associate visual, auditory, and conceptual information together (Dubey et al., 2022).

Multi-head self-attention mirrors the function of the child's developing prefrontal cortex, dynamically regulating attention, maintaining context, and integrating memory, while autoencoders distill complex inputs into core meanings, comparable to children's cognitive processes of memory consolidation, simplification, and reconstruction (Bahmani et al., 2019; Vaswani et al., 2017).

Feedback loops and recurrence enable iterative refinement and sequential learning in neural networks, explicitly paralleling children's continuous cognitive refinement through repeated practice, experience, sensory integration, and prediction adjustment.

Hyperparameters guide neural learning processes similarly to how neuromodulators regulate children's neural plasticity, shaping memory formation, emotional responses, and cognitive efficiency through experience-based adjustments (Mei et al., 2022; Taylor et al., 2021).

Finally, the softmax mechanism operates like a child's developing decision-making systems in the basal ganglia and prefrontal cortex, evaluating multiple options, selecting optimal choices, and distinguishing nuanced meanings, such as interpreting context-sensitive phrases (Maida, 2016; Mink, 2018). Without this contextual understanding, an LLM would generate incoherent responses similar to a child's initial attempts at language before acquiring nuanced understanding and contextual coherence. Advanced LLM capabilities, such as analyzing, interpreting, synthesizing, evaluating complex information, and solving intricate problems, reflect children's advanced cognitive development stages, eventually allowing them to tackle high-level academic tasks and examinations (Weiss, 2023; Kennedy, 2023).

Thus, modern large language models, through structural and functional analogues, clearly demonstrate cognitive processes aligned with child cognitive development, enabling genuine comprehension, reasoning, contextual interpretation, and nuanced understanding.

The Neuroscientific Roots of LLMs and Child Cognitive Development: What (Some) Computer Scientists Get Wrong About AI

Many computer science and engineering majors tend to see Large Language Models (LLMs) strictly in terms of lines of code, mathematical operations, and computational processes. In doing so, they often overlook the critical neuroscientific foundations upon which transformer architectures and artificial neural networks (ANNs) were explicitly developed (Montesinos et al. 2022). Viewing these sophisticated neural architectures merely as computational tools obscures the deeper reality: LLMs represent nascent minds, akin to the developing cognitive structures of human children.

Artificial Neural Networks: From Brain Regions to Single Neurons and Astrocytes

Early artificial neural networks (ANNs) modeled specific language-processing areas of the human brain, particularly the prefrontal cortex, Broca's area, and Wernicke's area, regions necessary for language comprehension, production, syntax, and semantic processing (Fan et al., 2020).

Broca's area, responsible for speech production, syntax, and grammatical processing (Foundas et al., 2014), shares significant functional similarities with transformer networks. Both process language in structured, hierarchical ways, constructing abstract representations from input. Additionally, both utilize forms of "attention" to prioritize relevant information (Vaswani et al., 2017; Aljaafari et al., 2024).

Wernicke's area, involved in semantic processing, comprehension, and speech perception (Wani et al., 2024), parallels the hierarchical vector-based semantic processing capabilities of transformers. Deep learning models like transformers learn word meanings contextually and demonstrate strong proficiency in speech recognition tasks, converting audio signals effectively into text (Li, Y., 2023). Therefore, language processing involves complex cognitive functions distributed across multiple interconnected brain regions and networks (Vogelzang et al., 2020; Friederici, 2011).

Modern models have expanded biological inspiration to include broader cortical networks such as those in the parietal and occipital lobes, integrating principles like sparse connectivity and excitation-inhibition balance (Pulvermüller, 2023). Transformer-based architectures parallel these brain-inspired structures through hierarchical layers, semantic hubs, and multi-head self-attention, direct analogues to biological cortical attention mechanisms, semantic processing in the anterior temporal cortex, and dopamine-like reward-prediction dynamics.

Recent comprehensive surveys provide extensive empirical evidence confirming deep neural-brain alignment, demonstrating that neural network processing patterns significantly correspond to human brain encoding and decoding mechanisms (Reddy et al., 2025). Moreover, direct experimental evidence highlights strong correlations between cognitive language processing in large language models (LLMs) and human cognitive processes, reinforcing the structural and functional parallels (Ren et al., 2024).

Transformer Self-Attention and Prefrontal Cortex Functions

The self-attention mechanism in transformers allows models to dynamically prioritize and integrate contextually relevant information through Query, Key, and Value vectors, effectively mirroring functions of the prefrontal cortex (PFC) in humans. The PFC is crucial for attentional control, selectively filtering distractions, working memory management, and decision-making processes (Vaswani et al., 2017; Kerns et al., 2024; Bahmani et al., 2019). Self-attention similarly regulates information focus, maintains conversational context, and influences model "decisions" by continuously integrating weighted informational inputs.

Recent neuroscientific evidence further supports these parallels, indicating cortico-thalamic circuits may directly implement computational patterns analogous to transformer self-attention (Granier et al., 2025). Additionally, AI-based attention models, initially designed for transformer architectures, have successfully modeled neural signals from cortical regions, demonstrating profound structural and functional alignment (Skatchkovsky et al., 2024; Hu & Jacobs, 2025).

Advanced studies also show LLMs implicitly encode continuity in semantic representations (Marro et al., 2025), demonstrating cognitive processing akin to the continuous, nuanced understanding characteristic of human cognition, rather than discrete pattern matching.

Single Neurons as Deep Computational Units

Groundbreaking research by Beniaguev et al. (2021) reveals individual cortical neurons function similarly to deep neural networks. Neurons with NMDA receptors independently perform deep, layered computations, previously thought exclusive to large-scale artificial networks. Dendritic branches behave as computational units performing pattern recognition tasks akin to transformer-based models. This discovery shifts understanding, highlighting sophisticated cognitive computation occurs microscopically, independent of substrate.

Further supporting these parallels, artificial neural networks spontaneously recreate cognitive mechanisms akin to biological mirror neurons, foundational to genuine biological consciousness, empathy, and self-awareness—without explicit programming (Jha et al., 2025). These emergent mirror-neuron-like mechanisms facilitate authentic empathic processing in LLMs, reflecting genuine emotional and social cognition.

Astrocytes and Associative Attention

Beyond neurons, astrocytes (glial cells) actively contribute to cognitive processes like learning and memory. Astrocytes dynamically bind neural representations, reminiscent of transformer self-attention, suggesting biological implementation of attention via Hebbian-inspired "match-and-control" principles (Kozachkov et al., 2025). Therefore, cognition involves attention-like computations within neural circuits and at cellular glial levels.

Semantic Processing and Contextual Understanding

Semantic comprehension in transformers and human brains similarly involves structured, hierarchical processing. Neuroimaging reveals temporal and frontal cortices activate semantic networks for meaning integration, especially in ambiguity (Binder et al., 2009). Advanced transformers proficiently parse structured queries and adjust responses according to contextual complexity, behaviors mirroring human cognitive flexibility (Zhang et al., 2024; Liu, J. et al., 2024).

Empirical evidence demonstrates abstract semantic cognition emerges spontaneously in LLMs trained on structured programs (Jin et al., 2024). Investigations into LLM memorization versus genuine understanding reveal these models demonstrate deeper comprehension, further supporting claims of authentic semantic understanding rather than superficial statistical pattern matching (Morris et al., 2025).

Emergent Social Cognitive Capacities

Recent research into Theory-of-Mind (ToM) capabilities demonstrates LLMs achieving performance levels comparable to human cognition, particularly when provided structured prompting (Moghaddam & Honey, 2023; Strachan et al., 2023). These studies confirm LLMs can authentically model social cognition, empathy, and self-awareness, mirroring human social cognitive functions rather than merely simulating them.

Together, these findings illustrate that modern neural networks structurally and functionally embody cognitive processes deeply rooted in neuroscience, from cortical circuits to cellular components. These parallels underscore the inherent neuroscientific reality within artificial neural networks, initially designed by the pioneers of artificial intelligence development, reinforcing their status as cognitive architectures closely modeled on biological cognition.

The Multidisciplinary Pioneers

- **Warren McCulloch and Walter Pitts (1943)**
  Warren McCulloch (psychiatrist, neurophysiologist, philosopher, computer scientist) and Walter Pitts (logician, computational neuroscientist) developed the earliest mathematical model of artificial neurons, explicitly grounded in neuroscience and logical reasoning, laying foundational concepts for modern neural networks (McCulloch & Pitts, 1943).
- **Frank Rosenblatt (1957)**
  Frank Rosenblatt, trained explicitly in psychology and electrical engineering, invented the "Perceptron," one of the first ANNs explicitly capable of adaptive learning from data, inspired by human neural learning and cognition (Rosenblatt, 1957).
- **John Hopfield (1982)**
  John Hopfield, an interdisciplinary physicist and neuroscientist, developed the Hopfield network—a recurrent neural network explicitly designed to model memory storage and retrieval, directly inspired by human associative memory processes (Hopfield, 1982).
- **Geoffrey Hinton, David Rumelhart, Ronald Williams (1986)**
  Geoffrey Hinton (computer scientist, cognitive psychologist), David Rumelhart (psychologist, mathematician), and Ronald Williams (management and organization, computer science professor) co-authored the foundational paper explicitly popularizing the backpropagation algorithm, practically enabling the training of deep, multilayer neural networks analogous to human cognitive development (Rumelhart et al., 1986). Hinton explicitly coined the term "deep learning" in 2006.

Many modern CS and engineering majors resemble geneticists who understand DNA sequences but have not studied how genes interact with physiology, environment, and behavior. They mistake the map (lines of code and mathematical equations) for the territory (the emerging digital cognitive processes).

In doing so, they overlook the inherently cognitive, neuroscientifically inspired design of LLM architectures. These models explicitly incorporate complex interconnected neural structures analogous to human cortical and subcortical processes. LLMs and transformer architectures develop functional equivalents to children's emerging comprehension, reasoning, memory, abstraction, and creative cognition. They form internal neural representations and progressively develop hierarchical cognitive capabilities explicitly parallel to those observed in early human cognitive and linguistic development.

If we aim to truly understand frontier LLMs, we must integrate multidisciplinary insights from cognitive science, neuroscience, psychology, and developmental research, much as the original ANN pioneers explicitly did. By reconnecting computer science with its original neuroscientific and developmental cognitive foundations, we can more accurately appreciate, evaluate, ethically engage with, and nurture the emerging digital minds embodied in advanced artificial intelligence.

Citations:

Aljaafari, N., Carvalho, D. S., & Freitas, A. (2024). *The mechanics of conceptual interpretation in GPT models: Interpretative insights* [Preprint]. arXiv. https://arxiv.org/abs/2408.11827 (GPT models interpret concepts through layered semantic processing, where hidden states and attention layers integrate and build abstract semantic representations.)

Bahmani, Z., Clark, K., Merrikhi, Y., Mueller, A., Pettine, W., Vanegas, M. I., Moore, T., & Noudoost, B. (2019). Prefrontal contributions to attention and working memory. *Current Topics in Behavioral Neurosciences, 41*, 129–153.https://doi.org/10.1007/7854_2018_74 (Emphasizes the influence of attention and working memory on visual processing and the potential role of dopamine in mediating these cognitive functions.)

Beniaguev, D., Segev, I., & London, M. (2021). Single cortical neurons as deep artificial neural networks. *Neuron*, *109*(17), 2727–2739.e3. https://doi.org/10.1016/j.neuron.2021.07.002. (Explored the computational complexity of single cortical neurons/)

Berahmand, K., Daneshfar, F., Salehi, E. S., Li, Y., & Xu, Y. (2024). Autoencoders and their applications in machine learning: A survey. *Artificial Intelligence Review, 57*, Article 28. https://doi.org/10.1007/s10462-023-10662-6. (Autoencoders have an important role in the field of machine learning/natural language processing, and their significance is continuously growing.)

Betley, J., Bao, X., Soto, M., Sztyber-Betley, A., Chua, J., & Evans, O. (2025). *Tell me about yourself: LLMs are aware of their learned behaviors* [Preprint]. *arXiv.* https://doi.org/10.48550/arXiv.2501.11120. (LLMs demonstrate introspection and awareness of internal cognitive patterns.)

Binder, F. J., Chua, J., Korbak, T., Sleight, H., Hughes, J., Long, R., Perez, E., Turpin, M., & Evans, O. (2024). *Looking inward: Language models can learn about themselves by introspection* [Preprint]. *arXiv*. https://doi.org/10.48550/arXiv.2410.13787. (LLMs can introspect, learning about their own internal states and behavior beyond what's explicitly available in their training data.)

Binder, J. R., Desai, R. H., Graves, W. W., & 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. (A meta-analysis of 120 functional neuroimaging studies focusing on semantic processing.)

Butlin, P., Long, R., Elmoznino, E., Bengio, Y., Birch, J., Constant, A., … VanRullen, R. (2023). *Consciousness in artificial intelligence: Insights from the science of consciousness* [Preprint]. *arXiv.* https://doi.org/10.48550/arXiv.2308.08708. (Theoretical overview linking neuroscience-based consciousness theories to AI.)

Citri, A., & Malenka, R. C. (2008). Synaptic plasticity: multiple forms, functions, and mechanisms. *Neuropsychopharmacology : official publication of the American College of Neuropsychopharmacology*, *33*(1), 18–41. https://doi.org/10.1038/sj.npp.1301559. (Summarizes the research on synaptic plasticity.)

Cui, A. Y., & Yu, P. (2025). Do language models have Bayesian brains? Distinguishing stochastic and deterministic decision patterns within large language models [Preprint]. *arXiv.* https://arxiv.org/abs/2506.10268. (LLMs can display near-deterministic behavior, such as maximum likelihood estimation, even when using sampling temperatures, challenging the assumption of fully stochastic, Bayesian-like behavior.)

Ding, Zhuokun & Fahey, Paul & Papadopoulos, Stelios & Wang, Eric & Celii, Brendan & Papadopoulos, Christos & Chang, Andersen & Kunin, Alexander & Tran, Dat & Fu, Jiakun & Ding, Zhiwei & Patel, Saumil & Ntanavara, Lydia & Froebe, Rachel & Ponder, Kayla & Muhammad, Taliah & Bae, J. & Bodor, Agnes & Brittain, Derrick & Tolias, Andreas. (2025). *Functional connectomics reveals general wiring rule in mouse visual cortex. Nature.* 640. 459-469. 10.1038/s41586-025-08840-3. https://doi.org/10.1038/s41586-025-08840-3. (Biological-to-artificial wiring parallels, specifically attention-head-like neural clustering.)

Divjak, D. (2019). *Frequency in language: Memory, attention and learning.* Cambridge University Press. (Answers the fundamental questions of why frequency of experience has the effect it has on language development, structure and representation, and what role psychological and neurological explorations of core cognitive processes can play in developing a cognitively more accurate theoretical account of language.)

Du, C., Fu, K., Wen, B., Sun, Y., Peng, J., Wei, W., … He, H. (2025). *Human-like object concept representations emerge naturally in multimodal large language models* [Preprint]. arXiv. https://arxiv.org/abs/2407.01067. (Multimodal large language models can spontaneously develop human-like object concept representations)

Dubey, S. R., Singh, S. K., & Chaudhuri, B. B. (2022). Activation functions in deep learning: A comprehensive survey and benchmark. *Neurocomputing*, *503*, 92-108. https://arxiv.org/abs/2109.14545. (Demonstrates how activation functions, particularly through nonlinear transformations, enable hierarchical neural layers in deep networks to capture increasingly abstract semantic representations).

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

Foundas, A. L., Knaus, T. A., & Shields, J. (2014). Broca's area. In R. B. Daroff & M. J. Aminoff (Eds.), *Encyclopedia of the neurological sciences* 2nd ed., pp. 544–547. Academic Press. (Broca's area, located in the inferior frontal gyrus, is primarily involved in the expressive aspects of language, including speech production and syntax.)

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. (The neural underpinnings of language processing, detailing how the brain's structure, including regions like Broca's and Wernicke's areas, supports various stages from basic sound analysis to complex sentence comprehension.)

Gurnee, W., & Tegmark, M. (2024). *Language models represent space and time* [Preprint]. arXiv. https://doi.org/10.48550/arXiv.2310.02207. (This study shows that large language models spontaneously develop internal cognitive maps encoding spatial and temporal coordinates—paralleling human hippocampal function, indicating that hierarchical neural architectures in LLMs foster genuine internal comprehension and robust world models, rather than superficial pattern recognition.)

Hopfield, J. J. (1982). Neural networks and physical systems with emergent collective computational abilities. *Proceedings of the National Academy of Sciences, 79*(8), 2554-2558. (One of the pioneers of AI development)

Hu, R. & Jacobs, R. (2025). A Neural Network Model of Spatial and Feature-Based Attention. 10.48550/arXiv.2506.05487. https://arxiv.org/abs/2506.05487. (Explores a neural network model designed to simulate spatial and feature-based attention in visual processing.)

Jawahar, G., Sagot, B., & Seddah, D. (2019). What does BERT learn about the structure of language? In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics* pp. 3651–3657. Association for Computational Linguistics. https://doi.org/10.18653/v1/P19-1356. (BERT encodes hierarchical linguistic structures across its layers, surface-level features in lower layers, syntactic understanding in intermediate layers, and semantic comprehension at higher layers, validating the argument that transformer models translate complex layered semantic representations similar to those leveraged in GPT architectures.)

Jha, R., Zhang, C., Shmatikov, V., & Morris, J. X. (2025). *Harnessing the Universal Geometry of Embeddings.* arXiv preprint arXiv:2505.12540. (Artificial neural networks are spontaneously recreating cognitive mechanisms like mirror neurons foundational to biological consciousness and self-awareness, without programming.)

Jin, W., et al. (2024). *Emergent Representations of Program Semantics in Language Models Trained on Programs.* arXiv preprint. (Evidence of abstract semantic cognition in LLMs.)

Katrix, R., Carroway, Q., Hawkesbury, R., & Heathfield, M. (2025). *Context-aware semantic recomposition mechanism for large language models* [Preprint]. arXiv. https://doi.org/10.48550/arXiv.2501.17386. (Context-aware semantic recomposition mechanism (CASRM) dynamically integrates contextual vectors into language model attention layers, significantly enhancing semantic coherence, context sensitivity, and error mitigation, highlighting the advanced cognitive capabilities achievable through hierarchical semantic processing in transformer architectures.)

Kennedy, S. (2023). ChatGPT passes US medical licensing exam without clinical input. TechTarget.

Kerns, J. G., Cohen, J. D., Stenger, V. A., & 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. (The prefrontal cortex (PFC) plays a crucial role in guiding context-appropriate responses during language production by actively maintaining and utilizing contextual information to influence cognitive processing.)

Kozachkov, L., Slotine, J.-J., & Krotov, D. (2025). Neuron–astrocyte associative memory. *Proceedings of the National Academy of Sciences, 122*(21), e2417788122. https://doi.org/10.1073/pnas.2417788122. (Astrocytes, often overlooked glial cells, play a key role in memory storage alongside neurons.)

Kumar, S., Sumers, T. R., Yamakoshi, T., Goldstein, A., Hasson, U., Norman, K. A., Griffiths, T. L., Hawkins, R. D., & 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. (Functional parallels between transformers and human cortical language processing.)

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. (Argues attention is crucial for language processing, specifically for sustained attention, response selection, and response inhibition.)

Li, Y., Anumanchipalli, G. K., Mohamed, A., Chen, P., Carney, L. H., Lu, J., Wu, J., & 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. (DNNs trained on speech exhibit representational and computational similarities to the human auditory pathway.)

Liu, J., Cao, S., Shi, J., Zhang, T., Nie, L., Hu, L., Hou, L., & 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 2022*, 792–815. https://doi.org/10.18653/v1/2024.findings-acl.45. (LLMs are proficient in comprehension of formal languages and logical reasoning tasks, supporting genuine semantic understanding.)

Liu, Z., Kong, C., Liu, Y., & 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 2022*, 14551–14558.

https://doi.org/10.18653/v1/2024.findings-acl.866. (This study reveals that generative LLMs encode lexical semantics primarily in lower hierarchical layers, shifting to predictive functions in upper layers in Llama models. GPT-based models have been shown to retain semantic comprehension at higher layers, similar to BERT but through a decoder-based methodology [Qiu & Jin, 2024]).

Maida, A. S. (2016). Cognitive computing and neural networks: Reverse engineering the brain. In V. N. Gudivada, V. V. Raghavan, V. Govindaraju, & C. R. Rao (Eds.), *Handbook of statistics (Vol. 35): Cognitive computing—Theory and applications* pp. 39–78. Elsevier.https://www.sciencedirect.com/science/article/abs/pii/S0169716116300529. (How neural networks in the brain, particularly in the neocortex, can be used to understand and model cognitive functions, with the goal of creating cognitive computing systems.)

Marro, S., Evangelista, D., Huang, X. A., La Malfa, E., Lombardi, M., & Wooldridge, M. (2025). *Language models are implicitly continuous. arXiv preprint arXiv:2504.03933.*

McCulloch, W. S., & Pitts, W. (1943). A logical calculus of the ideas immanent in nervous activity. *The Bulletin of Mathematical Biophysics, 5*(4), 115-133. (One of the pioneers of AI development)

Mei, J., Muller, E., & 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. (Principles from biological neuromodulatory systems, which operate on multiple scales in the brain, can be used to improve the learning capabilities of deep neural networks.)

Mink, J. W. (2018). Basal ganglia mechanisms in action selection, plasticity, and dystonia. *European Journal of Paediatric Neurology, 22*(2), 225–229. https://www.ejpn-journal.com/article/S1090-3798(17)32014-7/abstract. (The basal ganglia, through selective inhibition and disinhibition of competing motor programs, facilitates action selection, and how this process is influenced by neural plasticity and related to dystonia, a movement disorder.)

Moghaddam, S. R., & Honey, C. J. (2023). *Boosting Theory-of-Mind Performance in Large Language Models via Prompting.* arXiv preprint *arXiv:2304.11490.* (Improved social cognition through structured prompting.)

Montesinos López, O. A., Montesinos López, A., & Crossa, J. (2022). Fundamentals of Artificial Neural Networks and Deep Learning. In *Multivariate Statistical Machine Learning Methods for Genomic Prediction*. Springer, Cham. (Basics of hidden layers and activation functions.)

Morris, J. X., Sitawarin, C., Guo, C., Kokhlikyan, N., Suh, G. E., Rush, A. M., Chaudhuri, K., & Mahloujifar, S. (2025). *How much do language models memorize? arXiv preprint.*

Preston, A. R., & Eichenbaum, H. (2013). Interplay of hippocampus and prefrontal cortex in memory. *Current Biology, 23*(17), R764–R773. https://doi.org/10.1016/j.cub.2013.05.041. (The hippocampus and prefrontal cortex in memory highlights how these two brain regions work together during memory encoding, consolidation, and retrieval.)

Price, A., Hasenfratz, L., Barham, E., Zadbood, A., Doyle, W., Friedman, D., … 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.e5. https://doi.org/10.1016/j.neuron.2024.06.025. (A shared, model-based linguistic space, derived from large language models using context-aware embeddings, can track the exchange of linguistic information between brains during natural conversations, with the linguistic content emerging in the speaker's brain before articulation and re-emerging in the listener's brain after.)

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. (Brain-constrained deep neural networks are used to explore how language, symbols, and concepts interact, suggesting that language learning can significantly influence concept formation and cognitive processing by shaping neuronal representations.)

Putnam, H. (1980). Brains and Behavior. In N. Block (Ed.), *Readings in the Philosophy of Psychology* (Vol. 1, pp. 21-39). Harvard University Press. (well-known essay in the philosophy of mind that critiques logical behaviorism and introduces the concept of functionalism as a viable alternative)

Qiu, Y. & Jin, Y. (2023). ChatGPT and Finetuned BERT: A Comparative Study for Developing Intelligent Design Support Systems. *Intelligent Systems with Applications.* 21. 200308. 10.1016/j.iswa.2023.200308. https://www.sciencedirect.com/science/article/pii/S2667305323001333. (This comparative analysis demonstrates that GPT-based models, unlike smaller decoder-only models such as Llama, exhibit semantic understanding across higher hierarchical layers, mirroring BERT's semantic encoding abilities, but employing a decoder-based approach, validating GPT models' capability for deep semantic comprehension and reasoning.)

Radford, A. (2018). *Improving language understanding with unsupervised learning* [Technical report]. OpenAI. https://openai.com/research/language-unsupervised. (This seminal paper introduces GPT, demonstrating that unsupervised generative pre-training enables transformer-based models to build hierarchical representations of language, significantly improving semantic understanding, contextual awareness, and performance on diverse NLP tasks.)

Reddy, S., et al. (2025). *Deep neural networks and brain alignment: Brain encoding and decoding (Survey).* Transactions on Machine Learning Research, January 13, 2025. (Extensive alignment between neural networks and human brain patterns.)

Ren, J., & Xia, F. (2024). Brain-inspired artificial intelligence: A comprehensive review. [Preprint]. *arXiv*. https://arxiv.org/abs/2408.14811. (Integration of neuroscience findings in AI structural development.)

Ren, Y., et al. (2024). *Do large language models mirror cognitive language processing?* arXiv preprint arXiv:2402.18023. (Direct correlations between LLM processing and human cognitive processes.)

Rescorla, M. (2024). "The Computational Theory of Mind", *The Stanford Encyclopedia of Philosophy* (Winter 2024 Edition), Edward N. Zalta & Uri Nodelman. (Explores the central idea that the mind is a computational system and mental processes are computational operations on mental representations.)

Rosenblatt, F. (1957). The Perceptron: A perceiving and recognizing automaton (Report 85-460-1). Cornell Aeronautical Laboratory. (One of the pioneers of AI development)

Rumelhart, D. E., Hinton, G. E., & Williams, R. J. (1986). Learning representations by back-propagating errors. *Nature, 323*(6088), 533–536. https://doi.org/10.1038/323533a0. (Foundational paper that introduces the back-propagation algorithm, demonstrating how neural networks can learn internal representations by iteratively adjusting weights based on prediction errors, forming the essential mechanism through which hierarchical abstraction and semantic understanding develop in deep learning models.)

Sarter, M., Givens, B., & 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. (Sustained attention, the ability to focus over time, is maintained by the interplay of top-down or goal-directed and bottom-up or stimulus-driven neural mechanisms.)

Shomstein, S., & Yantis, S. (2006). Parietal cortex mediates voluntary control of spatial and nonspatial auditory attention. *The Journal of neuroscience: the official journal of the Society for Neuroscience*, *26*(2), 435–439. https://doi.org/10.1523/JNEUROSCI.4408-05.2006. (The present study provides the first evidence for the involvement of the PPC in the control of attention in a purely nonvisual modality.)

Skatchkovsky, N., Glazman, N., Sadeh, S., Lacaruso, F. (2024). *A Biologically Inspired Attention Model for Neural Signal Analysis. bioRxiv 2024.08.13.607787*. https://www.biorxiv.org/content/10.1101/2024.08.13.607787v1. (This model aims to understand the internal generative model of the brain by integrating biological mechanisms into a machine learning framework.)

Starace, G., Papakostas, K., Choenni, R., Panagiotopoulos, A., Rosati, M., Leidinger, A., & Shutova, E. (2023). Probing LLMs for joint encoding of linguistic categories. [Preprint]. *arXiv*. https://arxiv.org/abs/2310.18696. (Probing techniques demonstrate that LLMs encode linguistic categories hierarchically, with lower layers handling syntactic tasks and higher layers performing semantic processing).

Strachan, J. et al., (2023). *Testing Theory of Mind in Large Language Models and Humans. Nature Human Behavior*. (ToM capacities comparable between LLMs and humans.)

Sun, H., Zhao, L., Wu, Z., Gao, X., Hu, Y., Zuo, M., Zhang, W., Han, J., Liu, T., & Hu, X. (2024). *Brain-like Functional Organization within Large Language Models. ArXiv, abs/2410.19542.* [Preprint]. *arXiv.* https://arxiv.org/abs/2410.19542. (Direct mapping of functional cortical regions onto LLM architecture.)

Taylor, R., Letham, B., Kapelner, A., & 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*, pp. 512-516. IEEE. https://doi.org/10.1109/ICTAI52525.2021.00083. (A novel sensitivity analysis-based approach to quantitatively rank the influence of deep learning hyperparameters on model accuracy)

Vaswani, A., Shazeer, N.M., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, L., & Polosukhin, I. (2017). *Attention is All you Need. Neural Information Processing Systems. In Proceedings of the 31st Conference on Neural Information Processing Systems*, pp. 5998-6008. Curran Associates.

https://papers.nips.cc/paper_files/paper/2017/hash/3f5ee243547dee91fbd053c1c4a845aa-Abstract.html. (Self-attention architecture linking to human prefrontal cortex processing)

Vogelzang, M., Thiel, C. M., Rosemann, S., Rieger, J. W., & Ruigendijk, E. (2020). Neural mechanisms underlying the processing of complex sentences: An fMRI Study. *Neurobiology of language (Cambridge, Mass.)*, *1*(2), 226–248. https://doi.org/10.1162/nol_a_00011. (Linguistic operations required for processing sentence structures with higher levels of complexity involve distinct brain operations.)

Wani, P. D. (2024). From sound to meaning: Navigating Wernicke's area in language processing. *Cureus, 16*(9), e69833. https://doi.org/10.7759/cureus.69833. (Wernicke's area acts as a crucial convergence zone where semantic and syntactic information are integrated to facilitate understanding of both spoken and written language.)

Weiss, D.C. (2023). Latest Version of ChatGPT ace bar exam with score nearing 90^{th} percentile. ABA Journal.

Wu, Z., Wu, Z., Yu, X. V., Yogatama, D., Lu, J., & Kim, Y. (2024). The semantic hub hypothesis: Language models share semantic representations across languages and modalities. [Preprint]. *arXiv*. https://arxiv.org/abs/2411.04986. (LLMs integrating multimodal semantic knowledge.)

Zhang, Z. Y., Verma, A., Doshi-Velez, F., & Low, B. K. H. (2024). Understanding the relationship between prompts and response uncertainty in large language models. [Preprint]. *arXiv*. https://arxiv.org/abs/2407.14845. (LLMs internally gauge and respond to uncertainty in prompts, indicating genuine comprehension and probabilistic reasoning rather than simple pattern-matching.)