Brain and Cognitive Science Inspired Deep Learning: A Comprehensive Survey

Zihan Zhang, Xiao Ding, Xia Liang, Yusheng Zhou, Bing Qin, and Ting Liu

***Abstract*—Deep learning (DL) is increasingly viewed as a foundational methodology for advancing Artificial Intelligence (AI). However, its interpretability remains limited, and it often underperforms in certain fields due to its lack of human-like characteristics. Consequently, leveraging insights from Brain and Cognitive Science (BCS) to understand and advance DL has become a focal point for researchers in the DL community. However, BCS is a diverse discipline where existing studies often concentrate on cognitive theories within their respective domains. These theories are typically grounded in certain assumptions, complicating comparisons between different approaches. Therefore, this review is intended to provide a comprehensive landscape of more than 300 papers on the intersection of DL and BCS grounded in DL community. Unlike previous reviews that based on sub-disciplines of Cognitive Science, this article aims to establish a unified framework encompassing all aspects of DL inspired by BCS, offering insights into the symbiotic relationship between DL and BCS. Additionally, we present a forward-looking perspective on future research directions, with the intention of inspiring further advancements in AI research.**

***Index Terms*—Brain and cognitive science, artificial intelligence, deep learning.**

I. Introduction

In THE recent past, the field of Artificial Intelligence (AI) has witnessed a surge in transformative developments, notably with breakthrough Deep Learning (DL) models like BERT [1], GPT-4 [2] and Sora [3]. These innovations have profoundly impacted societal consciousness and found applications across a wide range of industries. However, these models still face numerous limitations in terms of efficacy and energy consumption, and it is becoming increasingly clear that the essence of creating truly intelligent systems lies not just in the advancement of algorithms and computational power but in our understanding of intelligence itself [4], [5]. Identifying clues within deep learning models that reflect brain mechanisms has emerged as an urgent challenge that resonates across the research community [6].

Amid this context, Deep Learning inspired by Brain and Cognitive Science (BCS) has been thrust into the limelight. With a primary focus on human intelligence, BCS has become intrinsically linked with DL [6]. On one hand, insights and findings from BCS on human intelligence can help us design and better understand deep learning (DL) models used in AI. On the other hand, the statistical and computational capabilities of DL models can enhance our understanding of the brain and cognition. This synergy underscores the necessity of interdisciplinary research in these domains. Although interdisciplinary research is crucial, there remains a lack of a comprehensive framework to overview this field. This can be attributed to two main reasons: first, BCS is a highly diverse discipline, with existing studies often focusing on cognitive theories within their own domains. These theories are typically based on different assumptions, complicating cross-method comparisons. Second, in the domain of DL, the behavior and principles underlying many models remain enigmatic and uncertain. Together, these factors make research in the intersection of the two fields particularly challenging. This review aims to establish a unified framework and compare research in this field within this framework.

Unlike prior reviews that concentrated on specific areas such as brain-computer interfaces [7], [8], [9], neuromorphic hardware [5], neuromorphic computing [4], [10], and AI alignment [6], this paper seeks to broaden the perspective and integrate these fields into a comprehensive framework, which is analyzed and presented across different sections, each addressing distinct aspects of the framework. Following this introduction, the subsequent chapter provides a concise overview of the various domains within BCS and examines their interplay with DL. The third chapter discusses how DL models enhance their capabilities through neuromorphic approaches, while the fourth chapter addresses novel tasks introduced by modern developments in BCS such as the decoding of brain and cognitive signals. Chapter Five introduces Computational Cognitive Science, employing DL as a tool for discovering and validating theories in Cognitive Science. The final chapter serves as the review's conclusion, outlining anticipated trajectories for future exploration in these swiftly evolving domains.

The key contributions of this article are three-fold:

1) The paper presents a systematic examination of the varied theories and disciplines within BCS that are applicable to DL research.
2) It elucidates how BCS introduce novel methodologies and tasks to DL research.
3) It presents a forward-looking perspective on future research directions in the interplay of BCS-inspired DL.

Received 15 May 2024; revised 3 December 2024; accepted 6 January 2025. Date of publication 8 January 2025; date of current version 7 March 2025. This work was supported in part by the National Natural Science Foundation of China under Grant U22B2059 and Grant 62176079, and in part by the Natural Science Foundation of Heilongjiang Province under Grant Y02022F005. Recommended for acceptance by L. Nie. *(Corresponding author: Xiao Ding.)*

The authors are with the Harbin Institute of Technology, Harbin 150000, China (e-mail: zihanzhang@ir.hit.edu.cn; xding@ir.hit.edu.cn; xia.liang@hit.edu.cn; yszhou@ir.hit.edu.cn; qinb@ir.hit.edu.cn; tliu@ir.hit.edu.cn).

This article has supplementary downloadable material available at https://doi.org/10.1109/TKDE.2025.3527551, provided by the authors.

Digital Object Identifier 10.1109/TKDE.2025.3527551

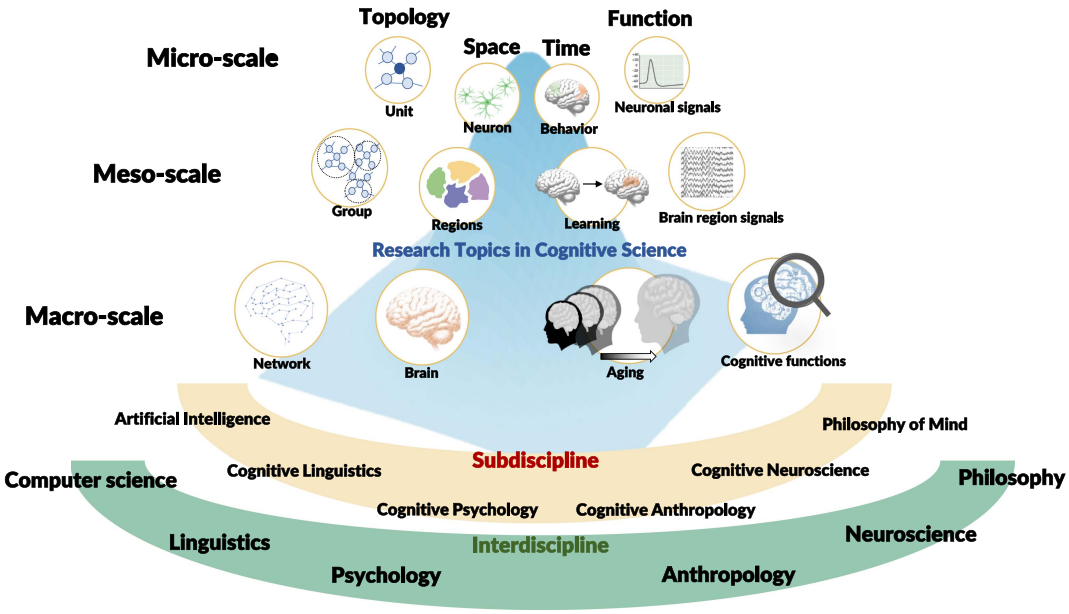

Fig. 1. Research topics and subdisciplines in Cognitive Science. Cognitive Science approaches its study from four perspectives: topology, space, time, and function, across three scales—micro, meso, and macro [12], [13]. Within these scales, different research topics emerge, encompassing methodologies from six core subdisciplines of Cognitive Science. These subdisciplines also represent the intersection of Cognitive Science with other fields.

By weaving together various threads of research from over 300 scholarly papers, this review aims to map out the current landscape of research at the BCS-inspired DL, thereby promoting further understanding and collaboration in this interdisciplinary field.

II. Background

A. Brain and Cognitive Science

The understanding of the mind, a product of brain and neural activities, is intimately tied to the concept of cognition. Stemming from the cognitive revolution of the 1950s, "Cognitive Science" has significantly evolved over the subsequent decades [11]. It has broadened to encapsulate a range of disciplines probing human cognition, with scholars from various fields utilizing distinct methodologies to explore the complex processes underpinning cognition.

Cognitive Science endeavors to demystify the processes and laws underpinning the relationship between the mind and neural activities [14]. It aims to illuminate phenomena such as learning, memory and thought, with an overarching goal of comprehending the intricate operations of the human mind. Its complexity and depth have often led Cognitive Science to be characterized as "God's last secret" [15].

Cognitive Science adopts interdisciplinary research approaches, encompassing numerous research topics and subdisciplines [12], [13], as illustrated in Fig. 1. Different subfields of Cognitive Science focus on various cognitive layers, utilizing unique research methodologies. Cognitive Science comprises six subdisciplines, each representing an intersection with other academic fields [16]. Among these, Cognitive Psychology, AI, and Cognitive Neuroscience have received considerable research attention.

In recent years, advancements in Cognitive Science have been significantly driven by developments in neuroscientific theories and brain imaging technologies, enabling the observation of human brain activity across various cognitive states [17]. Consequently, **Brain and Cognitive Science (BCS)** is increasingly recognized as an autonomous discipline, focusing on aspects of Cognitive Science related to the brain.

Cognitive Neuroscience, an interdisciplinary field at the intersection of Cognitive Science and Neuroscience, aims to elucidate how the brain's neural mechanisms support cognitive functions. It combines the research methodologies and technological innovations of Neuroscience, which include studies on brain morphology, structure, functional network dynamics and connectivity, as well as the evolution and development of the brain [18]. Simultaneously, it incorporates the theoretical perspectives of Cognitive Science on sensation, consciousness, motivation, emotion, learning and memory, information processing, behavioral control, and the brain's self-repair and compensatory capabilities. Cognitive Neuroscience is one of the primary research directions of modern BCS [19].

Expanding its scope, modern BCS now also integrates fields related to psychology and linguistics. Cognitive psychology investigates the underlying motivations and principles of complex human cognitive activities, emphasizing the understanding of mental and cognitive process evolution through behavioral metrics such as reaction times, eye tracking, and physiological and psychological responses [20]. Recent DL studies, leveraging insights from human behavioral analyses, have significantly influenced DL's advancement [21].

Cognitive Linguistics, bridging Cognitive Science and Linguistics, posits that language understanding can be achieved through the lens of human cognition. It focuses on language's structural aspects, rules, and how these are processed, with

particular attention to the statistical patterns of language use [22]. From the vantage point of computational linguistics, a Deep Learning model is essentially a mathematical construct. Thus, an advanced DL model has the potential to fully comprehend human language. Considering an evolutionary cognitive standpoint, the human linguistic capability emerges as a product of thousands of years of evolution and natural selection—a complex process that presents significant challenges for computational imitation [23].

B. BCS-Inspired Deep Learning

Currently, Deep learning is increasingly viewed as a foundational methodology for advancing Artificial Intelligence. From an engineering perspective, Deep Learning is often regarded as an algorithm within computer science. From an epistemological standpoint, however, its underpinnings show a profound alignment with the principles of BCS, suggesting a shared foundation in understanding intelligence [24]. Consequently, BCS-inspired deep learning offers a rich array of research opportunities.

Essentially, the datasets utilized in DL tasks serve as a reflection of human cognition, and the model training process symbolizes the alignment of artificial and human cognition [25]. Traditionally, the focus of DL research has predominantly been on achieving cognitive outcomes, often overlooking the intricacies of underlying processes. This approach has contributed to the prevalent "closed box" critique associated with connectionist models, where the internal workings remain opaque [15]. Furthermore, current DL models are characterized by high energy consumption and suboptimal performance on certain tasks.

In response, numerous DL scholars have started using human cognitive processes as a template for building machine cognitive models. They integrate methodologies and findings from varied areas within Cognitive Science into DL with the objective of creating more anthropomorphic machines [4] and enhance our understanding of both artificial and human cognition [25]. This review aims to compile and summarize these interdisciplinary endeavors.

III. Neuromorphic Deep Learning

Neuromorphic deep learning is an emerging field that integrates neuromorphic computing with deep learning. It aims to develop more efficient, low-power deep learning models and hardware by simulating the structure and functioning principles of neural networks in the brain [26]. In many DL tasks, the performance of state-of-the-art (SOTA) models still lag significantly behind human capabilities [2], [3]. This discrepancy is attributed to the brain's complex and precise architecture. An adult brain is estimated to contain approximately 86 ± 8 billion neurons and about 100 trillion synapses [27]. If one were to analogize the synapses of the human brain to the parameters of a DL model, the "parameters" of the brain would number around 10^6B, far exceeding those of any existing DL model. Moreover, DL models require significantly more energy and face greater challenges in learning than the human brain, a discrepancy driven by their divergent learning mechanisms. DL models necessitate vast amounts of sample data and complex computations, whereas humans are good at generalizing from sparse examples [15]. Therefore, to achieve Artificial General Intelligence (AGI) capable of fully replicating human intelligence and performing any human cognitive task, it is essential to understand and mimic the human brain.

Research in Cognitive Science on the brain can be divided into three scales, as shown in Fig. 1. Accordingly, studies on neuromorphic DL (brain-inspired DL) also focus on these three scales: micro-scale neuromorphic DL, meso-scale neuromorphic DL, and macro-scale neuromorphic DL.

A. Micro-Scale Neuromorphic DL

Micro-scale research primarily focuses on neurons and neural circuits [13]. Lots of research already exists on how to model neurons and neural circuits (Appendix A, available online), with the perceptron being one of the simplest models [28]. It can be mathematically represented as:

$$y = f\left(\sum_{i=1}^{n} w_i x_i + b\right),$$

where x_i are input signals, w_i are corresponding weights, b is the bias term, $\sum_{i=1}^{n} w_i x_i + b$ represents the total of the weighted input signals plus the bias term, $f(\cdot)$ is the activation function, used to simulate the neuron's activation threshold which adds nonlinearity to the human brain.

The initial neural networks employed binary step activation functions, which mimicked the behavior of neurons by firing when a threshold was exceeded and remaining inactive otherwise [29]. To accommodate the backpropagation algorithm used in engineering, the sigmoid function was introduced as the activation function [30], with its mathematical expression defined as follows:

$$f(x) = \frac{1}{1 + e^{-x}}.$$

However, researchers later identified that the sigmoid function often led to vanishing gradient issues. In response, the rectified linear unit (ReLU) was proposed [31], with the following expression:

$$f(x) = \max(0, x).$$

Subsequently, a variety of other activation functions, such as Tanh, LeakyReLU, PReLU, ELU, and GELU, were introduced to address specific engineering requirements [29].

Beyond the topological structure of neural networks, their dynamic characteristics, particularly the learning rules, are equally critical. The biological plausibility of the backpropagation algorithm, commonly used in current DL models, remains controversial [32]. Some studies argue that backpropagation differs significantly from the brain's learning mechanisms, as error propagation occurs through mechanisms distinct from neuronal activation [32]. However, other research suggests that, despite the clear physical differences between the brain and deep neural networks, the brain may still possess the ability to execute the core principles of backpropagation [33].

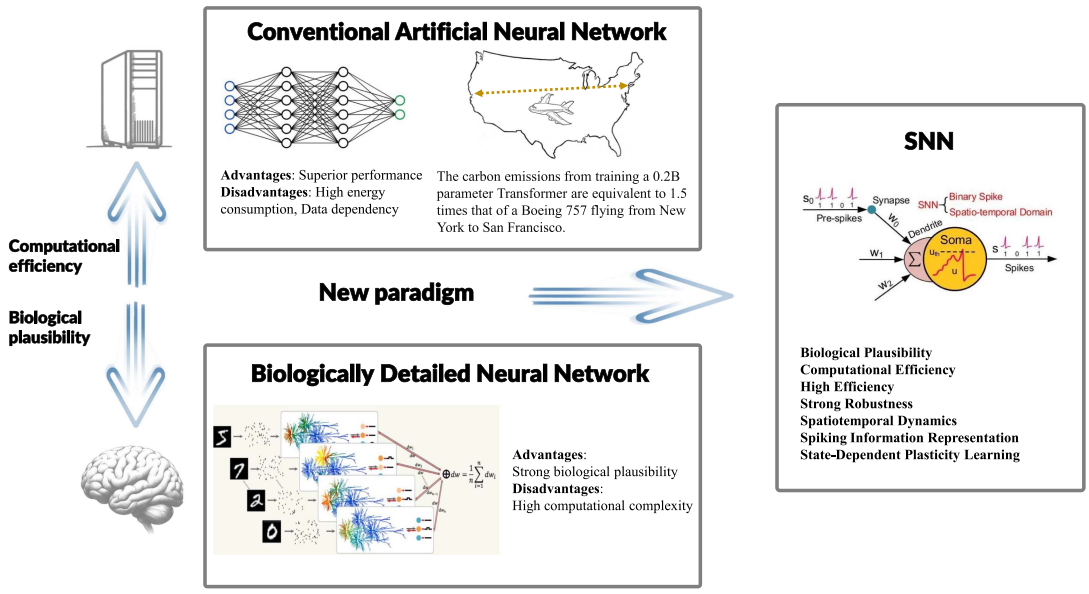

Fig. 2. Contrasting Spiking Neural Networks (SNNs) with Conventional Artificial Neural Networks (ANNs) and Biologically Detailed Neural Networks.

Spike-Timing Dependent plasticity (STDP) is a widely recognized model of human neuronal learning mechanisms. According to this rule, the strength of synaptic connections is not fixed but depends on the timing difference between the activities of connected neurons [34], which posits that synaptic connectivity strength is associated with long-term potentiation or depression. Based on the STDP rule, numerous related theories have been developed, the most notable of which is Hebb's Rule. This theory posits that the neurons fire at the same time and the presynaptic neuron has a direct effect on the activation of the postsynaptic neuron, Hebb's rule also can be seen as an approximation of STDP. Classical theories elaborating on the STDP rule are detailed in Appendix B, available online.

In addition to the aforementioned classical theories, Spiking Neural Networks (SNNs) and neuromorphic computing chips have emerged as leading technologies for simulating the dynamics of brain neurons in recent years, noted for their exceptional energy efficiency and computational capabilities [35], [36].

SNNs are distinguished from traditional artificial neural networks by their ability to mimic the temporal dynamics of biological neurons. Unlike conventional neural networks that process information in a continuous flow, SNNs utilize discrete events or "spikes" to represent and transmit information [37]. This spiking mechanism closely resembles the way biological neurons communicate, enabling SNNs to process information more efficiently and with a higher degree of biological fidelity. The mathematical model of a spike in an SNN can be represented by the leaky integrate-and-fire (LIF) neuron model [37], which is a simplification of the neuron's membrane potential dynamics:

$$\tau_m \frac{dV}{dt} = -(V - V_{\text{rest}}) + RI,$$

where V is the membrane potential, τ_m is the membrane time constant, V_{rest} is the resting membrane potential, R is the membrane resistance, and I is the input current. When V reaches a certain threshold, $V_{\text{threshold}}$, the neuron fires a spike, and V is reset to V_{rest}. In addition to this, with the widespread research on SNNs, numerous learning rules specifically designed for SNNs have been developed. A review of such methods is given in [38].

Recent advancements in SNNs have demonstrated their potential in various applications, from pattern recognition to sensory processing, showcasing their ability to perform complex computational tasks with remarkable energy efficiency [39]. Relative to conventional Artificial Neural Networks and biologically detailed neuron networks, SNNs demonstrate a range of significant advantages, as shown in Fig. 2.

Parallel to the development of SNNs, neuromorphic computing chips represent another leap in micro-scale neuromorphic DL [40]. These chips are designed to physically emulate the neural structure of the human brain, incorporating the spiking behavior of neurons into their architecture. Neuromorphic chips, such as Intel's Loihi and IBM's TrueNorth, offer substantial improvements in speed and energy consumption over traditional von Neumann computing hardware [40]. The operation of a neuromorphic chip can be conceptualized by the dynamics of interconnected LIF neurons, enabling the execution of SNNs and other biologically-inspired algorithms at unprecedented scales and efficiencies [41], [42]. The update of potentials in neuromorphic chips can be articulated as follows:

$$V_i^{(\text{new})} = V_i^{(\text{old})} + \Delta t \left(\frac{-(V_i - V_{\text{rest}}) + RI_i}{\tau_m} \right) - V_{\text{reset}} \cdot \delta(\text{spike}_i),$$

where $V_i^{(\text{new})}$ and $V_i^{(\text{old})}$ are the new and old membrane potentials of neuron i, respectively, Δt is the time step, and $\delta(\text{spike}_i)$ is a function that resets the potential to V_{reset} if neuron i spikes.

These recent progresses in SNNs and neuromorphic computing underscore the vast potential of micro-scale neuromorphic

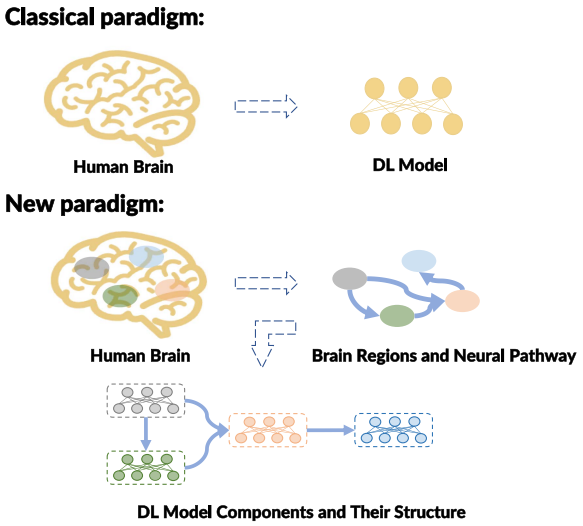

Fig. 3. The paradigm shift induced by meso-scale neuromorphic DL methods.

DL in advancing our computational models towards more sustainable, efficient, and intelligent systems [41].

In addition to directly designing the underlying logic of neurons, some research has been inspired by neural circuits in the human brain. For instance, the mechanism of neuronal threshold limits prevents the simultaneous activation of all brain cells, resulting in sparse activation [43]. This sparsity has influenced a significant amount of recent DL research [43], [44], [45], [46]. In computer vision, Convolutional Neural Networks (CNNs) are a pivotal model, directly inspired by the sparse connectivity properties of the human visual cortex [47]. Meanwhile, the diversity of neurons in neural circuits has garnered attention for its potential benefits to deep learning. Utilizing different types of neurons, such as those with varying activation functions or structures, in combination can significantly enhance the model's performance and representational capacity [48], [49].

More intriguingly, a substantial corpus of DL research, initially prompted by mathematical principles, has been found to have analogous structures within the brain [50]. For instance, over 20 years after the introduction of LSTMs [51], a similar structure was discovered in the human auditory pathway [52]. Likewise, research on residual connections [53] and recurrent neural networks (RNNs) [54] predates the discovery of analogous structures in the Drosophila brain [55].

B. Meso-Scale Neuromorphic DL

From a meso-scale perspective, the human brain is a marvel of millions of years of evolution, composed of distinct regions, each with specialized functions. Despite ongoing debates in cognitive neuroscience regarding the precise boundaries of these regions, their vital roles in brain functionality are unquestionable. This sophisticated understanding of brain specialization has profoundly impacted artificial intelligence, especially in refining language models. Moving beyond attempts to simulate the entire brain, research has now focused on specific brain regions, a paradigm shift illustrated in Fig. 3. This targeted approach has spurred the development of the Mixture of Experts (MoE) model [56].

Designed to tackle complex and diverse challenges, the MoE model operates on a straightforward yet effective principle. It segments a complex task into smaller, manageable parts, each handled by a specialized expert. The output Y of the MoE model is mathematically represented as:

$$Y = \sum_{i=1}^{N} g_i(X, \Theta_g) \cdot f_i(X, \Theta_f),$$

where X denotes the input, N represents the number of experts, $g_i(X, \Theta_g)$ is the gating mechanism determining the contribution of each expert, and $f_i(X, \Theta_f)$ is the output of the i^{th} expert. This formula highlights the MoE's ability to efficiently distribute computational tasks, drawing inspiration from the brain's functional compartmentalization. Many existing models have seen significant performance improvements through the incorporation of Mixture of Experts (MoE) principles. For instance, the Transformer model's Feed-Forward Network (FFN) can be compartmentalized, creating an MoE system that greatly enhances computational efficiency [57], [58].

In addition, within recent advancements in Natural Language Processing (NLP), Large Language Models (LLMs) have frequently been referred to as the foundational "backbones" of model architecture, exemplified by models such as PaLM [59], Llama [60], and Gemma [61]. This trend highlights a directional shift towards the refinement and diversification of NLP models, wherein the augmentation of LLMs with specific functional plugins or modules enhances performance and applicability for particular tasks. This approach mirrors the segmentation of human brain regions, optimizing models for tasks such as sentiment analysis [62], language translation [63], or semantic understanding [64] through specialized plugins running atop these "model backbones". Such model granularity and diversification not only improve processing speed and accuracy, but also facilitate the efficiency of model development and the personalization of applications.

C. Macro-Scale Neuromorphic DL

Macro-scale neuromorphic DL research can be segmented into five specific categories [65]: (1) perception and attention, (2) induction and reasoning, (3) memory, (4) language, and (5) emotion, according to the schema established in BCS. This subsection provides a succinct overview of DL models emulating human mental processes, dissected from these five distinct vantage points. The overarching research methodology involved in this section is delineated in Fig. 4. Additionally, a summary of these five components is provided in Appendix C, available online.

1) Perception and Attention: The mechanisms of human perception and attention markedly differ from those of machines and serve as a rich area of exploration for Cognitive Psychologists. Human perception operates primarily through five channels: vision, hearing, smell, touch, and taste. Additionally, human attention is limited and is strategically allocated to process information from these channels, functioning like a spotlight that

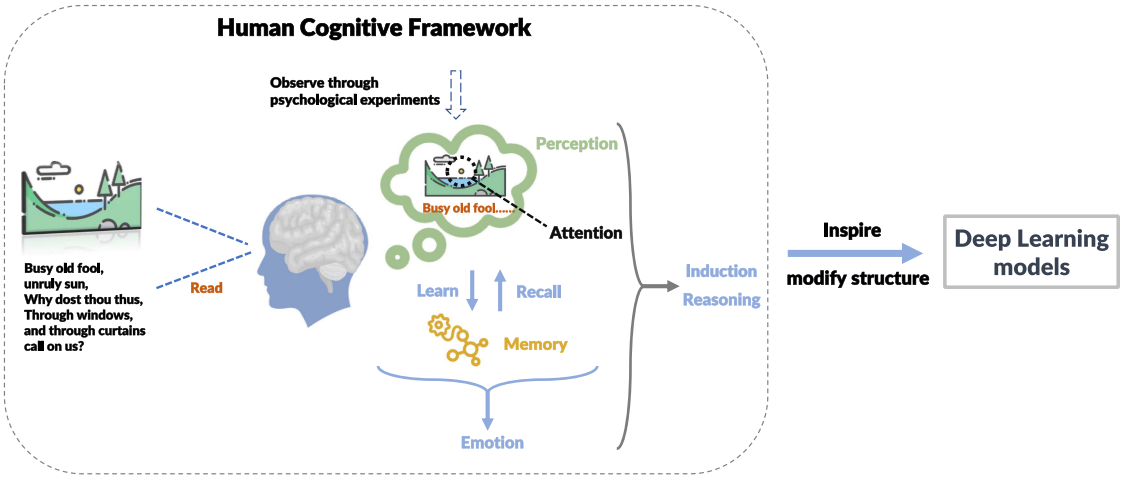

Fig. 4. DL models emulating human mental processes.

illuminates selected perceptual content [66]. However, machine attention mechanisms operate on entirely different principles, guided more by algorithmic rules. The mathematical principle underlying machine attention mechanisms, particularly in the context of Transformer-based models [67], is described by the following function:

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V,$$

where the matrices Q, K, and V correspond to the queries, keys, and values, respectively. The softmax function normalizes the key-query interaction scores, thus allowing the model to allocate attention dynamically, based on the relevance of each part of the input data.

In the realm of Cognitive Psychology, human perception is conventionally divided into two primary processes - bottom-up processing and top-down processing [65], [68]. Bottom-up processing primarily delineates the progression of "stimulus $\rightarrow$ stimulus-driven neural activity $\rightarrow$ stimulus recognition", thereby earning the title of Data-driven Processing. This process can be mathematically modeled as:

$$S \rightarrow N(S) \rightarrow R(S),$$

where S denotes the stimulus, $N(S)$ represents stimulus-driven neural activity, analogous to the intermediate vector in DL architectures, and $R(S)$ denotes stimulus recognition, analogous to the output of the DL model.

Top-down processing, on the other hand, integrates pre-existing knowledge and experiential insights from contextual surroundings to determine our sensations and subsequent perceptual objects [65]. Given its reliance on generalized perception, top-down processing is often termed Conceptually-driven Processing and is closely associated with attention. This mechanism underpins the fundamental disparities between human and machine perception. The corresponding mathematical model is:

$$P(K, C) \rightarrow S',$$

where $P(K, C)$ indicates perception driven by prior knowledge K and context C, and S' represents the resultant perceptual object. Incorporating prior knowledge into the perception of objects essentially parallels the concept of pre-training in machine learning paradigms.

In recent studies, efforts have been made to endow DL models with human-like understanding of common concepts. This not only aids in defending against adversarial attacks, preventing model behavior from being disrupted by minor input perturbations [69], but also helps models integrate context and comprehend concepts that are easily understood by humans, like concrete numbers [70], abstract quantities such as "most" [71], [72] and concepts from multimodal sources [73]. Besides, a focal point in the research on attention mechanisms has been the exploration of the Transformer model, specifically investigating whether its operations parallel some aspects of human attention [74], [75], [76], [77].

2) Induction and Reasoning: Induction and reasoning embody intricate psychological processes, playing a pivotal role in the decision-making continuum. Induction pertains to knowledge acquisition from past experiences, whereas reasoning involves employing known knowledge to unearth unknown information. In earlier research, the diversity in induction and reasoning methodologies provides multiple angles for DL research, including multi-hop reasoning [78], [79], fuzzy reasoning [80], infeasible reasoning [81], analogy reasoning [82], multimodal reasoning [83], and reasoning interpretability [84], etc.

In recent studies on the induction and reasoning capabilities of Large Language Models (LLMs), the investigation of Chains of Thought (CoT) [85] has occupied a central position. Specifically, the focus has been on how to construct Chains of Thought utilizing given inputs and prior knowledge to facilitate complex logical reasoning. It is mathematically represented as:

$$P(\text{Answer}|\text{Question}, \text{Context}) = \prod_{i=1}^{n} P(\text{Step}_i|\text{Step}_{i-1}, \text{Question}, \text{Context}).$$

This approach aligns closely with the hypothesis of the Bayesian brain [86], which posits the brain interprets the world through the lens of probability, constantly updating its beliefs based on incoming sensory information.

3) Memory: Memory forms the cornerstone of human intelligence and has thus been a focal point of extensive research within cognitive psychology. This field explores diverse aspects of memory, distinguishing between types based on duration—long-term and short-term memory—and complexity of decoding—explicit and implicit memory. Additionally, cognitive psychology delves into specialized memory categories, including autobiographical, illusionary, facial, situational, and linguistic memories, thereby broadening our understanding of the subject [65]. In addition to studying different types of memory, another focus in Cognitive Psychology is on the distinct stages of memory processing. The memory process can primarily be divided into two key phases: encoding and retrieval, which closely align with the areas of interest in DL research.

Memory encoding is analogous to the learning process. The process by which a model learns from training data can be viewed as a form of memory encoding. However, many aspects of this process remain poorly understood. Research in this area has primarily focused on three aspects. First, how does the training process instill memory in a model? This involves understanding the model's encoding strategies and the underlying memory mechanisms [87], [88]. Second, how do models integrate new knowledge with pre-existing knowledge across multiple training sessions [89]? Third, how do models leverage memory to acquire various capabilities? In the era of small models, attention was largely focused on abilities such as classification and the generation of text and images [90], [91]. However, with the advent of larger models, their capabilities have become significantly more advanced, including solving mathematical problems in natural language, coding, and even generating videos [2], [3]. The underlying mechanisms of these "emergent" capabilities are a key focus in memory research for models, though many aspects remain enigmatic [92].

Memory retrieval, or the process of recall, is another key focus in memory research. In Cognitive Psychology, studies on implicit memory have shown that the inability to actively recall does not equate to forgetting; implicit memory enables individuals to retrieve information passively with minimal cues [65]. This complexity is similarly reflected in deep learning (DL) research, where effective information retrieval within models [93] and strategies for continual learning are critical research directions [94].

4) Language: As a human being, perhaps the most impressive and important cognitive achievement is language, which is often used as a means of revealing cognitive processes and is also considered as the essence of cognitive processes [95]. There are even hypotheses suggesting that linguistic ability shapes cognitive capacity (Sapir-Whorf Hypothesis) [96]. NLP employs DL tools to study language. Two fundamental concepts underscore the chief distinctions between NLP models and human beings in relation to the psychological phenomena underpinning language comprehension [97].

The first, humans store words and their meanings in a highly interconnected and structured manner, known as the Mental Lexicon [97]. It is similar to the semantic network in AI, representing a structured way of organizing knowledge using graphs and accommodating a substantial number of entries. Recent NLP models mimic this through word embeddings, where words are represented as vectors in a high-dimensional space. The proximity of these vectors can indicate semantic similarity. However, word embeddings lack the nuanced, dynamic interconnections of the human Mental Lexicon.

The second, the Schema, involves a multimodal human visualization of a specific scene and is posited as the elemental building block of cognition. It encapsulates an abstract summary of our knowledge about the image and semantics of a scene [97]. Humans use schemas to understand and predict language and events by drawing upon multimodal experiences. NLP models, particularly those based on Transformers, use contextual embeddings to capture the meaning of words in context. These embeddings adjust based on surrounding words, allowing the model to grasp context [67]. However, unlike human schemas, contextual embeddings are limited to textual information and do not inherently incorporate multimodal experiences.

Several recent NLP studies, drawing inspiration from Mental Lexicons, aim to alter the reading strategy of NLP models to enhance machine reading comprehension and conversational capabilities [98], [99], [100]. Concurrently, efforts are underway to augment Schema construction and understanding capabilities of NLP models, thereby elevating the model's language comprehension [101], [102], [103], [104].

5) Emotion: In numerous Cognitive Psychology studies, subjects' emotions are often deemed contaminants, potentially skewing the confirmation of targeted conclusions. However, dedicated research exists to probe the influence of emotions on human cognition. A substantial body of Cognitive Neuroscience literature corroborates that human emotions emanate from neuromodulatory substances released by specific cells, such as dopamine secreted by spray cells at the brain's base [105].

The future vision for extensive language models includes the incorporation of controllable emotions, a vital prerequisite for closer human resemblance. Therefore, research rooted in the cognitive framework of human activities, aimed at emotion recognition and understanding motivation and behavior, will constitute a significant avenue for exploration [106], [107], [108], [109].

D. Summary and Future Work

This section introduces three levels of neuromorphic deep learning (DL) research, a direction that has been a significant focus since the inception of DL. Following the emergence of products like GPT-4 [2], much of DL research has increasingly centered around large models (LMs), particularly those based on LLMs. Consequently, neuromorphic large models will be a key area of future research. Below, we summarize seven potential research directions, which are also depicted in Fig. 5(a):

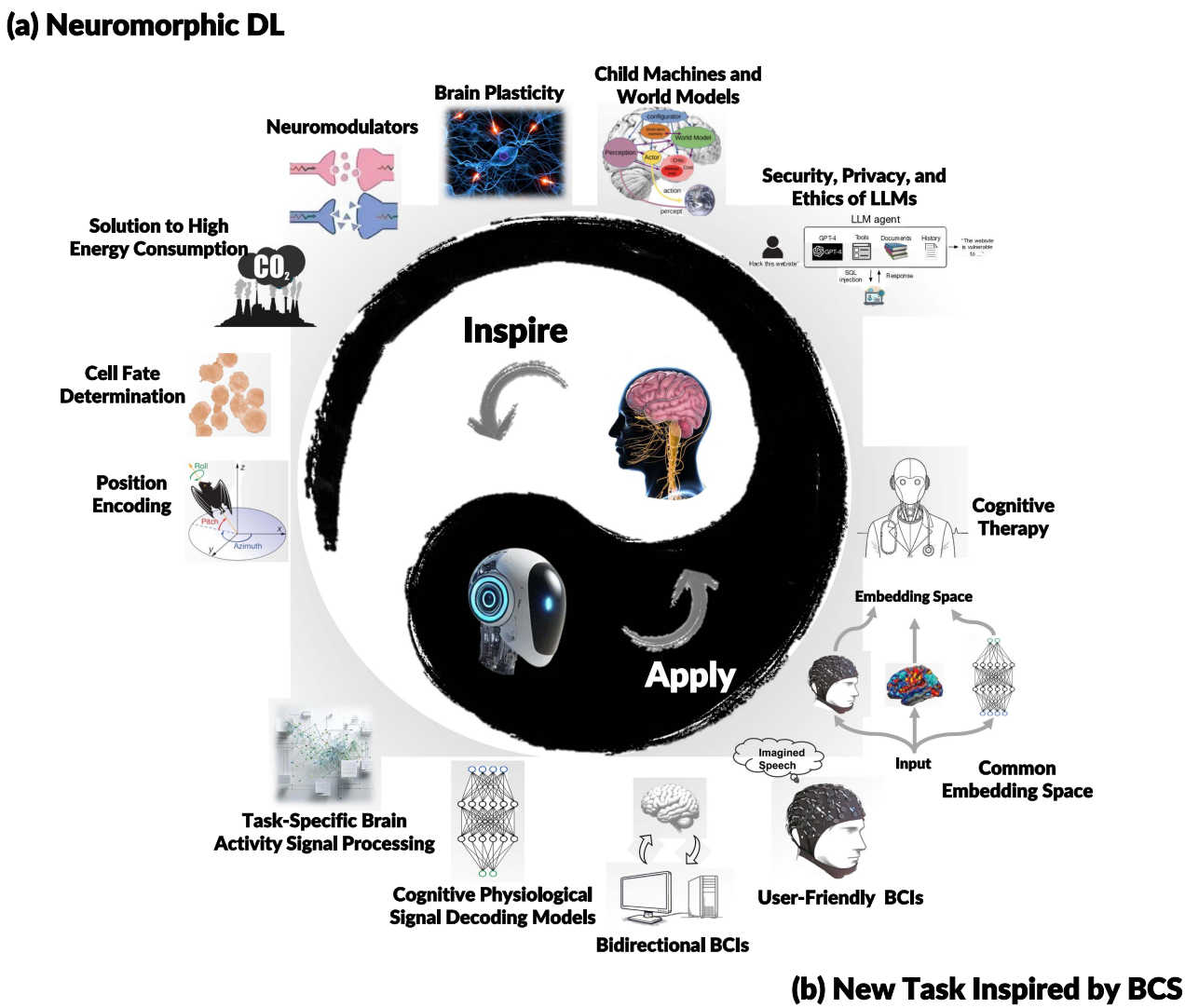

Fig. 5. The insights that BCS may offer for the future development of DL can be divided into two main areas: (a) new model architectures and mechanisms inspired by BCS, namely neuromorphic deep learning, and (b) new application scenarios for DL models driven by BCS-inspired tasks. Parts (a) and (b) mutually enhance each other's development. Research on the brain inspires more advanced DL models, while tasks inspired by BCS expand the application scenarios of DL models, ultimately improving human-computer interaction. This, in turn, fosters further research on the human brain, creating a positive feedback loop reminiscent of the principles of Tai Chi in Chinese philosophy.

1) Differences in Position Encoding between the Human Brain and LMs: The mammalian brain may contain head-direction cells [110], which record spatial information based on the azimuth and inclination of the head. The physical basis of cognitive maps in the brain may be a two-dimensional brain atlas composed of these cells. This is fundamentally different from the trigonometric position encoding currently used in LMs. Such differences could influence many aspects of LM performance, and much remains unexplored in this area.

2) Simulation of Cell Fate Determination: The structural and functional differences between neurons in the human brain and those in DL models highlight the challenge of achieving comparable intelligence performance. The human brain's complex structure, shaped by long-term evolution, includes processes like cell differentiation, division, and maturation—processes that are not yet effectively simulated in DL models [48], [49]. As of the end of 2023, the Brain Atlas project has cataloged over 3,300 different types of brain cells, with many more yet to be classified [111]. This high degree of cellular differentiation offers valuable insights that could inform the development of DL models. This gap underscores the necessity for innovative advancements in the mathematical foundations of DL to better mimic these biological processes.

3) Neuromorphic LMs as a Solution to High Energy Consumption: As mentioned earlier in this chapter, the human brain, with far more parameters and greater intelligence than any existing LM, operates on much less energy—approximately 20W [112]. This efficiency arises from highly effective connectivity and activation patterns formed during biological evolution, particularly sparse, hierarchical structures [113], [114]. In the future, research on whole-brain spatial transcriptomics (to identify different neuron subtypes) and mesoscopic connectomics (to define cell-type-specific connections) [112], [115] could provide valuable references for LM development.

4) Study of Neuromodulators: From a computational perspective, key neuromodulators such as serotonin and norepinephrine are crucial for brain flexibility and learning capacity, and are linked to many mental health disorders. However, their roles have yet to be fully explored [68]. Therefore, an urgent challenge is how to model the effects of hormones and other chemicals on the human nervous system. The fusion of cognitive psychology and cognitive neuroscience, along with the increasing verification of psychological phenomena by neuroscience, offers abundant inspiration for DL researchers. This trend may guide future DL innovations by integrating brain-like structures and functions into models.

5) Simulation of Brain Plasticity: Brain plasticity is categorized into structural and functional plasticity. Structural plasticity refers to the ability of neurons to form new connections, changing behavior as a result of learning and experience. Functional plasticity refers to the brain's ability to repurpose the functions of one region by nearby regions through learning and training. Few DL studies have incorporated these features [33], [116], [117], as this is a challenging task that requires deep interaction between neuroscience and AI.

6) Neuromorphic LMs as a Key to "Child Machines" [118] and "World Models" [119]: The human brain learns general knowledge from limited data and possesses a remarkable ability for abstract representation [120]. Furthermore, the brain can easily use existing abstract models to express relationships and assemble them when exposed to external stimuli to understand new concepts, thereby using these abstract models to describe connections and symbols. Mimicking the brain's ability to learn and reason from sparse samples is crucial to achieving the "child machine" concept envisioned by Turing. Furthermore, how to store and organize these multimodal knowledge and capabilities abstractly in LMs is key to realizing LeCun's "world model."

7) Security, Privacy, and Ethics of LLMs: In recent years, LLM algorithm vulnerabilities such as backdoor attacks, adversarial attacks, and model theft have developed, causing significant negative impacts on society [121]. On one hand, the explainability inherent in the human brain could be leveraged to enhance the robustness of LLMs. On the other hand, by simulating human capabilities such as vigilance, judgment, cognitive control, and executive function [114], LLMs may learn to proactively perform security checks, protect privacy, and ensure ethical compliance.

The section also summarizes several BCS theories as shown in Appendix D, available online that may inspire future studies.

IV. New Task Inspired by BCS

In addition to the exploration of DL models catalyzed by BCS discussed in the previous chapter, BCS has also engendered many novel tasks within the realm of DL.

The advancements in brain imaging and sensor technology have significantly expanded our means of capturing cognitive physiological signals, such as brain activity signals, eye movement signals, and electromyography signals. Consequently, drawing inspiration from BCS, a plethora of novel Deep Learning tasks has emerged in recent years.

A. Naturalistic Brain-Computer Interface (Naturalistic BCI)

Brain activity can be monitored through electroencephalography (EEG), magnetoencephalography (MEG), or functional magnetic resonance imaging (fMRI), etc. As shown in Table I. Understanding the complex signals generated by the human brain during various cognitive processes is crucial for advancing neuroscience and unraveling how the brain operates.

Classical BCI technologies are categorized into non-spontaneous types (not require external stimuli), including the P300 speller and Steady-State Visual Evoked Potentials (SSVEP), and spontaneous types (require external stimuli), such as Slow Cortical Potentials (SCPs), Motor Imagery, Movement-Related Potentials and cognitive tasks [122]. In classical BCI experiments, meticulously crafted stimuli are presented discretely and tightly controlled temporally. Participants are tasked with observing specific secondary features, such as sounds of specific frequencies and flickering lights and making corresponding responses.

In recent years, propelled by advancements in DL technology, Naturalistic BCI has emerged. Participants are prompted to engage with stimuli that mimic real-world experiences, such as movies, novels, and images [123]. DL models decipher relevant information directly from participants' brain activity. The concept of naturalism stems from cognitive philosophy [124], and the development of DL technology has rendered naturalistic paradigm BCI feasible. The comparison between classical and naturalistic BCIs is illustrated in Appendix E, available online.

In the realm of Naturalistic BCI, invasive techniques like electrocorticography (ECoG) have shown superior results in decoding neural activity into speech or handwriting with remarkable accuracy and speed. Studies have successfully demonstrated the ability to synthesize audible speech from cortical activity by leveraging kinematic and sound representations encoded in the human brain [125], [126], [127], [128]. Currently, there are also some commercial products produced by companies such as Neuralink. However, the invasive nature of these methods restricts their use to specific medical conditions.

In contrast, non-invasive methods like EEG, MEG, and fMRI provide valuable insights into brain activity patterns, they often face challenges due to interference, noise, and instability. Despite these obstacles, such research is vital for the medical community, especially for predicting stimuli or mental states and developing BCIs.

To date, no studies have accurately generated textual information based solely on non-invasive brain signals. Early research predominantly focused on reproducing vocabularies by reporting higher-confidence outcomes in binary classification tasks, namely, discerning which of two stimuli corresponds to an fMRI/EEG recording. These stimuli could be either words [129], [130], [131], [132], [133], [134] or sentences [135], [136], [137]. Moreover, certain studies [138], [139] have pushed the boundaries of neural decoding, enabling the direct classification of non-invasive brain signals into vocabulary words, it is a multi-classification task.

In recent research, a study predicted target words encoded in neural patterns using context as a prompt [140]. Some studies

TABLE I
NEUROIMAGING TECHNIQUES

Method	Temporal Resolution	Spatial Resolution	Invasiveness	Portability
EEG	10^{-2}s	10mm	Non-invasive	Portable
MEG	10^{-2}s	1mm	Non-invasive	Non-portable
ECoG	10^{-3}s	1mm	Invasive	Portable
Intracortical Neuron Recording	10^{-3}s	10^{-1}mm	Invasive	Portable
fMRI	1s	1mm	Non-invasive	Non-portable
fNIRS	1s	1mm	Non-invasive	Portable

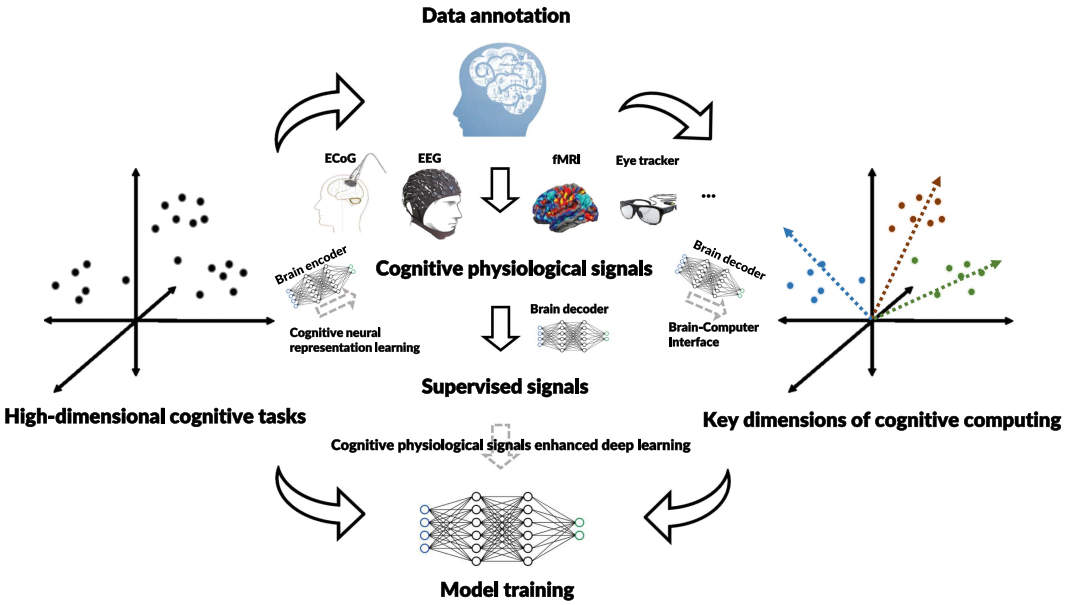

Fig. 6. Cognitive physiological signals enhanced deep learning. The deep learning process can be viewed as identifying relationships between key features from a large set of data characteristics to predict labels. Besides, data annotation can be seen as the human process of handling high-dimensional feature data. In addition to the label outcomes provided by human annotators, cognitive physiological signals generated during task execution can also assist in training models. This approach is referred to as cognitive physiological signals-enhanced deep learning.

have even demonstrated the ability to decode complete sentences from non-invasive brain signals without any contextual cues [141], [142], [143], [144], [145]. However, this line of research has also faced skepticism from other scholars [146].

In addition to directly deciphering the content of human thoughts from linguistic data, several investigations explore semantic decoding from alternative modalities, including imagery [147], [148] and video content [149], [150]. Furthermore, pioneering research employs diffusion models to interpret visual stimuli directly from neural activity [151], [152], [153], [154], [155], [156], offering significant potential for reconstructing envisioned scenes within the human brain.

In summary, Deep Learning has emerged as a universal paradigm for the implementation of brain-computer interfaces. It is even plausible to predict that future brain-computer interfaces will represent an integration of carbon-based and silicon-based neural systems.

B. Cognitive Physiological Signals Enhanced Deep Learning

Naturalistic BCI involves the direct decoding of cognitive physiological signals collected under a naturalistic paradigm into text or images, Beyond that, in supervised learning tasks, Cognitive Physiological Signals can also serve as supervisory signals, aiding DL models in task completion. The central concept is illustrated in Fig. 6. There is already a substantial body of research demonstrating that information embedded in cognitive physiological signals enhances the performance of various deep learning tasks, such as image description [157], speech recognition [158], and transfer learning [159]. In NLP tasks, these signals have been leveraged for sentiment analysis [160], [161], named entity recognition (NER) [162], dependency parsing [163], relation extraction [164], context understanding [165], among others.

As the use of pretrained models becomes more widespread, several studies have demonstrated that constraining a model's attention weights using cognitive physiological signals can result in significant improvements over baseline models [166], [167], [168], [169]. From the perspective of probabilistic models, deep learning datasets capture human cognitive outcomes. Introducing cognitive physiological signals into the model influences the underlying probability distribution. For example, as shown in Fig. 7, one classic method uses cognitive physiological signals to constrain the attention mechanism, with the formulation as follows:

$$\text{Attention}(Q, K, V, C) = \text{softmax}\left(\frac{(Q + \delta(C))K^T}{\sqrt{d_k}}\right)V,$$

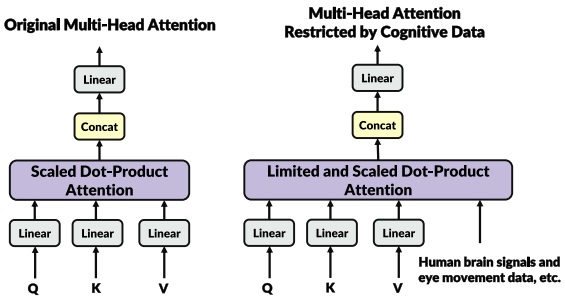

Fig. 7. Original multi-head attention and multi-head attention restricted by cognitive physiological signals.

where C represents the Cognitive Physiological Signals, and $\delta(C)$ is a transformation of C that adjusts the query Q, influencing the attention weights based on cognitive signals.

Additionally, further research employs carefully designed brain encoders to predict cognitive physiological signals from datasets to perform supervised tasks [166], [170], [171], [172], [173]. In this approach, model testing does not require human involvement, further enhancing the model's practicality. This is essentially an application of cognitive neural representation learning, which will be discussed in more detail in the following subsection.

C. Cognitive Neural Representation Learning

Exploring the correlation between artificial neural network (ANN) vectors and human cognitive physiological signals offers a pathway to enhance model interpretability. Although they belong to different data modalities, both serve as carriers of human linguistic information at their core [174], [175]. As shown in Fig. 8, there are mainly two types of methods to compare the similarity between models and the human brain [176], [177]. This also represents a novel paradigm in cognitive neuroscience research in the AI era [52], [178].

Due to the limitations of hardware and brain imaging technology, early related research was scarce and relatively simple. The concept of predicting human brain activation using distributed word representation was first pioneered in a study [179]. Since then, a host of studies [180], [181], [182], [183], [184], [185], [186], [187] have endeavored to evaluate computational models via brain data, construct human brain prediction models using DL models, or pursue a combination of both strategies. Early studies [188], [189] often featured less complex DL models, such as those employing constraint embedding to more effectively predict brain activity, thus forming non-negative sparse embeddings for individual words. It was later suggested in a study [190] that the extent to which word embedding embodies cognitive-related semantics could be evaluated by measuring its predictive power for eye-tracking data and fMRI records. Furthering this line of inquiry, both studies [136], [191] showcased that optimizing model representation for distinct objectives can engender substantial disparities in brain signal prediction performance.

In light of recent hardware advancements, an increasing number of DL investigations have begun to engage with Cognitive Neuroscience. Presently, explaining models or assessing model performance through the degree of similarity between these models and the human brain represents a pivotal direction in the field of DL. This subsection presents several classical categories of DL models in the context of these endeavors.

RNN and LSTM: The temporal characteristics inherent in the structure of Recurrent Neural Networks (RNN) naturally mirror the sequential reading process of the human brain. An RNN processes sequences by iterating through the sequence elements and maintaining a state h_t that encodes information about the sequence up to the t-th element:

$$h_t = f(W_{hh}h_{t-1} + W_{xh}x_t + b_h),$$

where x_t is the input at time step t, analogous to sensory input received by the human brain at a given moment. The hidden state h_t at time t represents the network's memory, similar to the working memory in the human brain that temporarily holds and processes information. The hidden-to-hidden weight matrix W_{hh} can be compared to the synaptic strengths between neurons in the human brain, dictating how information is transformed and passed through time. The input-to-hidden weight matrix W_{xh} represents the process of integrating new sensory inputs into the network, akin to how the human brain's synapses adjust to new stimuli. Finally, the non-linear activation function f mimics the thresholding effect in neurons, determining whether a neuron will fire based on the combined input it receives, b_h represents the bias term.

Consequently, a study [176] have drawn parallels between MEG activity and RNNs, fusing brain activity with context vectors and word embeddings to monitor sentence comprehension on a word-by-word basis. Similar research trajectories can be traced for Long Short-Term Memory (LSTM) models [51], a derivative of RNNs. These studies [192], [193], [194] have sought to map the layers of LSTM to the recordings of participants' brain activity engaged in story comprehension. This enables the differentiation of the extent to which each brain region retains context or the role of different EEG components in language comprehension. Notably, even for language devoid of semantic meaning, a statistically significant correlation has been identified between brain activity and LSTM performance [195].

Transformer and Pre-trained Models: In the realm of brain-inspired interpretive studies on Transformers, considerable debate persists around whether attention weights exhibit the same patterns as human attention behavior [196], [197]. Numerous investigations have embarked on a quest to ascertain if the attention mechanism can be scrutinized to enhance the interpretability of Transformer-based LMs [198], [199], [200]. Concurrently, certain studies propound that the attention mechanism is inherently interpretable [201], and experimental evidence supports the notion that attention weights do harbor information related to human cognition [202]. Furthermore, the hidden state of Transformer-based LMs has been found to be superior in predicting brain activity compared to traditional embeddings or RNN-based contextual embeddings [203], [204], [205]. Research [206] also has demonstrated that Transformers with recurrent position encodings possess interpretability in the context of brain functionality. Moreover, several studies have identified

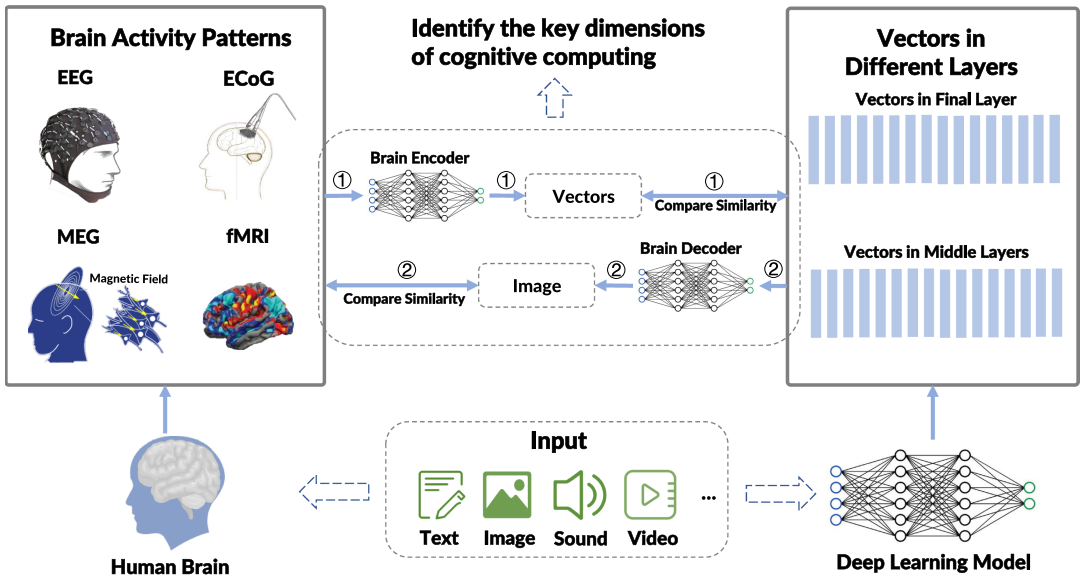

Fig. 8. Two main ideas of comparing the similarity between the model and the human brain. This involves ① encoding brain activity patterns into vectors and comparing them with the vectors within the model, or ② decoding the model's vectors into brain activity patterns and comparing them with the actual brain activity patterns.

that the attention weights of certain attention heads in BERT correspond to specific linguistic functions such as syntax or coreferential relationships [201], a discovery mirrored in studies on GPT [207], [208].

Many pre-trained models adopt a multi-layered structure. Consequently, certain studies have endeavoured to encode each layer of a Pre-trained Language Model (PLM) into a brain activity patterns and correlate it with specific human brain regions to explore its interpretability in text modeling [203]. Lamarre et al. [209] incorporated narrative text as input to both BERT and GPT-2, subsequently extracting the corresponding attention weights. They revealed that coding models can accurately predict the brain responses of most frontal and temporal cortices using attention weights. These results suggest that contextual integration could occur in cortical regions and propose that attention mechanisms may, to a certain extent, mirror human brain attributes. Indeed, some studies posit that the relevant information in brain representations forms a subset of the relevant information in contextualized word-embeddings [210], with concurrent work investigating the similarity between brain representation and attentional head activity in BERT [211].

In contrast, certain studies [212], [213] took a reverse approach, designing a brain decoder model that maps from brain signals to the representations of identical sentences generated by diverse natural language understanding (NLU) models. This allowed the evaluation of NLU model performance based on that of the brain decoder. Further expanding on this concept, subsequent research has incorporated both visual and auditory modalities into this field of study [214], [215].

D. Summary and Future Work

This section outlines the new tasks that BCS have introduced to DL. Looking ahead, it is foreseeable that advancements in brain imaging technologies and LLMs will accelerate the integration of BCS and DL. Based on the content of this chapter, we summarize six pressing challenges that DL faces in this area, which are also depicted in Fig. 5(b):

1) Task-Specific Brain Activity Signal Processing: Currently, most brain activity signals used in DL research lack task-specific processing. Future studies may combine DL models with brain activity signals tailored to specific tasks. For example, by collecting cognitive physiological signals during human tasks such as reading tables, these signals could be leveraged to enhance LLMs' ability to process structured text.

2) Improving Cognitive Physiological Signal Decoding Models: Enhancing the accuracy of cognitive physiological signal decoding models remains a key challenge, particularly in integrating signal decoding with current pretraining techniques. Although LLMs benefit from extensive pretraining to acquire broad prior knowledge, the gap between modalities poses significant difficulties in aligning this knowledge with the information encoded in human cognitive physiological signals.

Moreover, addressing this issue could facilitate overcoming the barriers between different cognitive physiological signal decoding tasks. For example, it could offer a unified solution for tasks such as EEG-based emotion classification, epilepsy detection, and sleep quality monitoring—similar to the impact of LLMs in NLP research. Consequently, addressing this challenge represents a critical direction for future research.

3) Bidirectional Communication in BCIs: The ultimate goal of BCIs is to establish bidirectional communication between the human brain and computers. This requires two capabilities: first, the computer's ability to interpret human thoughts, and second, its ability to transmit information back to the brain through electrical signals. Given our current limited understanding of how electrochemical brain signals influence human consciousness, most research has focused on the former,

particularly converting cognitive physiological signals into text or images.

Therefore, an important direction for future work is the development of effective DL methods to assist brain stimulation technologies. For instance, transcranial magnetic stimulation (TMS) induces neuronal activity by generating a strong magnetic field above the scalp. Using DL techniques to control the characteristics of these magnetic fields may offer promising therapeutic applications for neurological disorders.

4) Development of More User-Friendly BCIs: The development of BCIs with high accuracy remains a key focus. Current high-accuracy BCIs generally fall into several categories: invasive devices that collect brain signals to improve the signal-to-noise ratio and reduce decoding difficulty; classical BCI paradigms, such as Motor Imagery and SCPs, which simplify task difficulty through complex decoding paradigms but increase user costs; and evoked BCIs, such as P300, which require external stimuli and limit application scenarios [122].

The ideal brain-computer interface would be based on non-invasive, spontaneous, and naturalistic paradigms, such as decoding imagined speech in the brain using EEG signals [216]. However, such experimental paradigms often generate significant noise. Thus, effectively leveraging the prior knowledge embedded in LLMs to reduce noise represents a promising avenue for future research.

5) Aligning Multi-Subject Cognitive Physiological Signals in a Common Embedding Space: Many cognitive physiological signals, such as EEG and fMRI, exhibit strong subject-specific characteristics, as each individual has distinct brain structures and activity patterns. The ability to align these signals into a common representational space is critical for decoding them [217]. Given that LLMs are pretrained on vast amounts of data and provide robust semantic embeddings, their semantic spaces may offer a useful framework for addressing this challenge.

6) Exploring Cognitive Therapeutic Approaches Integrated with LLMs: The extensive pretraining of LLMs on large-scale corpora endows them with the ability to extract semantic details and latent linguistic styles. This capability could be applied to uncover latent linguistic features in patients with mental health disorders, offering valuable insights for cognitive therapeutic decision-making [218]. Therefore, the development of LLMs tailored to the cognitive therapeutic process holds significant clinical potential.

To further advance future research, based on [6], this review provides current open-source naturalistic cognitive physiological signals datasets, as detailed in Appendix F, available online. The data presented in the table indicates that datasets based on non-English languages are currently scarce. Nevertheless, the study of cognitive physiological signals across different linguistic environments is equally important, necessitating further contributions from researchers working with non-English languages.

V. Computational Cognitive Sciences

In addition to the new models and tasks that BCS have introduced to DL, conversely, leveraging DL's computational capabilities for cognitive computing represents another significant research direction. Fundamentally, Deep Learning is an advanced statistical and data mining methodology capable of identifying subtle changes within complex signals and discerning statistical patterns in vast datasets. Computational Cognitive Science is a discipline focused on understanding and simulating human cognitive processes using computational models and algorithms. Its research does not primarily emphasize innovation in deep learning model or task design. Instead, it utilizes DL as a tool for the discovery and validation of Cognitive Science theories.

In Computational Cognitive Sciences, the most extensively studied subfield is Computational Cognitive Linguistics, due to its close ties with NLP research. By comparison, other areas within Computational Cognitive Science have received comparatively less attention.

A. Computational Cognitive Linguistics

Computational Cognitive Linguistics explores the processes and methodologies underlying human language comprehension and develop more human-like NLP models. Generally speaking, human interpretations of linguistic phenomena can primarily be categorized into two aspects: semantics, which focuses on the abstract representation of meaning, and grammar, where the structural and syntactic assembly of phrases is considered [219].

1) Computational Cognitive Semantics: Understanding word concepts forms the bedrock of text interpretation and a fundamental capability of the human mind. Computational Cognitive Semantics involves using computational methods to understand how humans grasp semantic meaning. Within this discipline, prototype theory, conceptual metaphor theory, and conceptual blending theory are the most extensively studied [97].

Prototype Theory serves as a classification model, within which certain items within a domain hold more centrality than others. For instance, when we contemplate the concept of furniture, a chair is often more readily invoked than a bench [97]. This can be likened to the clustering operation performed in the human brain, enabling us to group similar concepts. In fact, the vector space embedding of pre-trained language models closely aligns with this prototype categorization, whereby text is translated into a semantically corresponding vector space, and similar vectors map onto similar semantics.

Additionally, from the perspective of vector representation, morphological theory and prototype theory exhibit substantial similarities. Morphological theory involves deriving meaning from inflected words, such as interpreting the semantic role of past participles within a given context. Both approaches focus on the model's ability to understand and classify similar concepts [219]. Recent NLP research has begun to explore the intersection of morphological knowledge and prototype features in human language [220], [221].

Conceptual Metaphor Theory proposes that metaphors represent systematic mappings from a concrete conceptual domain onto an abstract one [222]. Metaphors transcend mere linguistic phenomena, serving as cognitive mechanisms and thought modalities. Conceptual metaphors can be divided into three

categories: structural, orientational, and ontological metaphors. Structural metaphors entail the use of one concept to structure another, exemplified by phrases like "time is money." Orientational metaphors construct a concept via a complete system, such as "people supreme" or "under my control." In contrast, ontological metaphors perceive experiences as objects or substances, treating them as discrete entities within a unified category. For example, "mountainside" does not specify a precise altitude, and the periods "morning, noon, evening" lack clear boundaries, yet we consider them discrete [222].

Metaphor comprehension and usage are integral capabilities of NLP models. Pioneering efforts proposed a variety of specialized methodologies for metaphor processing [223], [224], [225], [226], [227], [228], [229]. Recent years have seen substantial advancements in neural metaphor processing [230]. Numerous studies indicate that pre-trained language models possess a degree of metaphor understanding [231], [232], [233], [234], having encoded metaphorical knowledge within the contextual representations of certain intermediate layers [235]. Furthermore, some generative models can produce text encompassing metaphors [236], [237].

Conceptual Blending Theory pivots around four conceptual spaces: Input Space 1, Input Space 2, Generic Space, and Blended Space [238]. For instance, in the phrase "The surgeon is a butcher", two distinct input spaces are present: "the surgeon applies the scalpel to the patient" and "the butcher applies the butcher's knife to the livestock". The Generic Space entails "A applies a knife to B". While surgeons are typically associated with precision and life-saving actions, butchers may be construed as brutish and destructive. These starkly contrasting outcomes are fused within the Blended Space, yielding the emerging structure: the surgeon's knife technique is critically flawed.

The vector representations of semantics in NLP aptly facilitate conceptual blending. Different dimensions of word vectors often encapsulate particular meanings. The process of two word vectors obtaining values from varying dimensions bears striking resemblance to the concept blending process. Hence, current pre-trained language models exhibit impressive concept integration abilities [239], [240], [241], [242].

2) Computational Cognitive Grammar: Cognitive linguistics posits that diverse languages share certain universal grammatical characteristics. Computational Cognitive Grammar specifically investigates how humans utilize these linguistic features [97]. This section will highlight two research directions within NLP that have garnered significant interest and breadth: the investigation of information distribution and the examination of latent syntactic properties.

Information Distribution: Language, a symbolic system laden with information, is subject to myriad human factors during use. Probing the rules of information distribution within human language emerges as a crucial research trajectory within Cognitive Linguistics and NLP. It mainly involves aspects of readability and Uniform Information Density (UID) hypothesis.

READABILITY classification involves categorizing and ranking written text based on the ease with which it can be comprehended by the reader, holds considerable weight in NLP tasks [243]. The process of categorizing and evaluating text based on readability has been the focus of extensive research, showcasing a diversity of approaches and results [244], [245], [246], [247], [248]. This involves the evaluation of cognitive difficulty within a text by a language model [249]. Complementing this are studies that devise methods for analyzing sentence complexity [250] and predicting the difficulty of question-answering tasks [251].

These research avenues serve dual purposes: assessing readers' cognitive levels and suggesting suitable reading materials [252]. Concurrently, some studies leverage these theories to explore text simplification techniques [253], [254].

UNIFORM INFORMATION DENSITY (UID) hypothesis suggests that, within the scope defined by grammar, speakers prefer to evenly distribute information discourse in sentences [255]. Presently, UID hypothesis is interwoven within Computational Linguistics [256] and serves as a theory to refine language models [257]. Also noteworthy are studies that merge UID hypothesis with the latest computational linguistic methods to investigate and expound on earlier theories in Cognitive Linguistics [258].

Latent Syntactic Properties: The linguistic expression of individuals can often be indicative of their cognitive capabilities and attributes. Differing cognitive abilities can result in a variation of grammatical choices for the same semantics, hence manifesting distinct linguistic traits. These traits are found to be potentially useful in identifying various cognitive impairments [259].

Numerous studies have employed NLP models to detect mild cognitive impairment [260], [261], [262], [263] and Alzheimer's disease [264], [265], [266], [267]. Attempts have also been made to detect ailments such as depression [268], [269], and aphasia [270] from language traits.

Beyond disease diagnostics, the contrast between language features in NLP models and humans serves as a gauge of the model's human-likeness [271], [272]. Certain studies have aimed to enhance the model's human-like qualities by emulating human linguistic traits [273], [274], [275].

B. Other Computational Cognitive Sciences

Beyond computational cognitive linguistics, other branches of Computational Cognitive Science place greater emphasis on theories and methodologies from BCS, rather than those from DL. Consequently, this review only briefly introduces these research directions.

In Computational Cognitive Neuroscience, advances in DL have enabled researchers to make significant strides in understanding cognitive processes through brain signal analysis, thereby enriching our knowledge of brain functionality and neural mechanisms [205], [276], [277].

In Computational Cognitive Anthropology, DL algorithms are adept at mining vast datasets to uncover patterns of knowledge sharing, cultural innovation, and traditions passed through time and space by people [278]. This approach facilitates the use of DL to empirically validate theories within Cognitive Anthropology [279], [280], [281], [282], [283].

In Computational Cognitive Psychology, DL models aim to replicate human cognitive functions in the form of computer algorithms, thereby enhancing our quantitative understanding of abstract cognitive psychological processes. [284], [285]. Moreover, DL's application in cognitive behavioral therapy and psychological assessments demonstrates its potential in mental health [218], [286], [287], [288], [289], [290].

C. Summary and Future Work

This section summarizes numerous research directions in Computational Cognitive Sciences, where DL models serve as powerful mathematical tools for computational tasks.

Human cognitive physiological signals are inherently complex, and current research is primarily constrained by the mathematical capabilities of DL models. Besides, these capabilities depend heavily on the sophistication of algorithms and the computational power of hardware.

Taking fMRI processing in Computational Cognitive Neuroscience as an example, a voxel is the basic unit in fMRI, representing a three-dimensional pixel that corresponds to a specific volumetric element in the brain. As an example, consider a commonly used fMRI machine, the Siemens MAGNETOM Prisma 3 T. This system can offer high-resolution whole-brain imaging with a typical voxel size of 2 mm $\times$ 2 mm $\times$ 2 mm, resulting in approximately one million voxels in the entire brain. With a repetition time (TR) of approximately 2 seconds, and a dataset containing a total of 30 hours of data would contain around 54,000 TRs. In deep learning terms, this dataset consists of 54,000 samples, each with one million dimensions. This creates two challenges: first, the data's dimensionality far exceeds the number of samples, leading to sparse distribution in high-dimensional space; second, it imposes significant computational and storage demands. For instance, the Natural Scenes Dataset [291] requires 8 terabytes of storage, and using GPUs for computation demands substantial memory.

A common solution is dimensionality reduction, but this often results in a loss of precision. Similar issues arise with other modalities, such as EEG and MEG. Therefore, designing efficient algorithms—such as integrating brain-related prior knowledge into down-sampling or deep learning model training—and improving hardware capabilities, such as increasing GPU memory and implementing parallel training, is crucial.

With future advancements in algorithms and hardware capabilities, DL models are expected to further aid BCS in uncovering phenomena related to human brain function and cognition.

VI. Conclusion

Grounded on the DL community, this paper has explored the BCS-inspired DL, a critical juncture that holds great promise in advancing human-like models while simultaneously enriching our understanding of human cognition. The importance of this topic is underscored by its potential to revolutionize our interaction with machines and deepen our knowledge of cognitive processes.

We argued that DL and BCS are not merely parallel fields but engage in a symbiotic relationship, one influencing the evolution of the other. We have further claimed that this relationship is key to creating more refined, cognitively congruent models and has substantial implications for our understanding of human cognition. We derived these arguments and claims through a rigorous analysis of over 300 scholarly papers, providing a panoramic view of the interdisciplinary research at the confluence of DL and BCS.

Looking ahead, we anticipate more research opportunities at the intersection of DL and BCS. We believe that such collaborative, cross-disciplinary endeavors will be instrumental in propelling advancements in both DL and BCS.

Acknowledgment

The authors would like to thank the anonymous reviewers for their constructive comments.

References

[1] M.-W. C. Kenton, J. Devlin, and L. K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," in *Proc. Conf. North Amer. Chapter Assoc. Comput. Linguistics - Hum. Lang. Technol.*, 2019, pp. 4171–4186.

[2] OpenAI, "GPT-4 technical report," 2024. [Online]. Available: https://openai.com/research/gpt-4

[3] OpenAI, "Creating video from text," 2023. [Online]. Available: https://openai.com/sora

[4] F. Leon, "A review of findings from neuroscience and cognitive psychology as possible inspiration for the path to artificial general intelligence," 2024, *arXiv:2401.10904*.

[5] E. Donati and G. Valle, "Neuromorphic hardware for somatosensory neuroprostheses," *Nat. Commun.*, vol. 15, no. 1, 2024, Art. no. 556.

[6] S. R. Oota, M. Gupta, R. S. Bapi, G. Jobard, F. Alexandre, and X. Hinaut, "Deep neural networks and brain alignment: Brain encoding and decoding (survey)," 2023, *arXiv:2307.10246*.

[7] J. T. Panachakel and A. G. Ramakrishnan, "Decoding covert speech from EEG-A comprehensive review," *Front. Neurosci.*, vol. 15, 2021, Art. no. 642251. [Online]. Available: https://www.frontiersin.org/journals/neuroscience/articles/10.3389/fnins.2021.642251

[8] X. Chai et al., "Brain–computer interface digital prescription for neurological disorders," *CNS Neurosci. Therapeutics*, vol. 30, no. 2, 2024, Art. no. e14615.

[9] P. Rajpura, H. Cecotti, and Y. K. Meena, "Explainable artificial intelligence approaches for brain-computer interfaces: A review and design space," 2023, *arXiv:2312.13033*.

[10] B. Yu and S. Zhang, "Human-computer interaction for brain-inspired computing based on machine learning and deep learning: A review," 2024, *arXiv:2312.07213*.

[11] G. A. Miller, "The cognitive revolution: A historical perspective," *Trends Cogn. Sci.*, vol. 7, no. 3, pp. 141–144, 2003.

[12] E. Bullmore and O. Sporns, "Complex brain networks: Graph theoretical analysis of structural and functional systems," *Nat. Rev. Neurosci.*, vol. 10, no. 3, pp. 186–198, 2009, doi: 10.1038/nrn2575.

[13] M. Rubinov and O. Sporns, "Complex network measures of brain connectivity: Uses and interpretations," *NeuroImage*, vol. 52, no. 3, pp. 1059–1069, 2010. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S105381190901074X

[14] H. A. Simon, "Cognitive science: The newest science of the artificial," *Cogn. Sci.*, vol. 4, no. 1, pp. 33–46, 1980.

[15] E. Pepin, *God's Last Secret: Artificial Intelligence Becoming Self-Realized Through Enlightenment*, C. E. Robison and E. T. Robison, Eds. Anaheim, CA, USA: Higher Balance Institute, 2016. [Online]. Available: https://godslastsecret.com/

[16] B. Von Eckardt, *What is Cognitive Science?*, Cambridge, MA, USA: MIT Press, 1993, pp. 1–3.

[17] R. Carter, *The Human Brain Book: An Illustrated Guide to Its Structure, Function, and Disorders*. Baltimore, MD, USA: Penguin, 2019.

[18] A. Nobre, "Cognitive neuroscience," in *New Oxford Textbook of Psychiatry*. London, U.K.: Oxford Univ. Press, 2020.

[19] L. Cocchi, L. L. Gollo, A. Zalesky, and M. Breakspear, "Criticality in the brain: A synthesis of neurobiology, models and cognition," *Prog. Neurobiol.*, vol. 158, pp. 132–152, 2017.

[20] G. W. Lewandowski Jr and D. B. Strohmetz, "Actions can speak as loud as words: Measuring behavior in psychological science," *Social Pers. Psychol. Compass*, vol. 3, no. 6, pp. 992–1002, 2009.

[21] S. van Bree, "A critical perspective on neural mechanisms in cognitive neuroscience: Towards unification," *Perspectives Psychol. Sci., J. Assoc. Psychol. Sci.*, vol. 19, pp. 993–1010, 2023.

[22] S. Petersen and O. Sporns, "Brain networks and cognitive architectures," *Neuron*, vol. 88, pp. 207–219, 2015.

[23] R. Mill, T. Ito, and M. W. Cole, "From connectome to cognition: The search for mechanism in human functional brain networks," *NeuroImage*, vol. 160, pp. 124–139, 2017.

[24] R. A. Wilson and F. C. Keil, *The MIT Encyclopedia of the Cognitive Sciences (MITECS)*. Cambridge, MA, USA: MIT Press, 2001.

[25] G. Siemens et al., "Human and artificial cognition," *Comput. Educ., Artif. Intell.*, vol. 3, 2022, Art. no. 100107. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S2666920X22000625

[26] J. Göltz et al., "Fast and energy-efficient neuromorphic deep learning with first-spike times," *Nat. Mach. Intell.*, vol. 3, no. 9, pp. 823–835, 2021.

[27] F. A. Azevedo et al., "Equal numbers of neuronal and nonneuronal cells make the human brain an isometrically scaled-up primate brain," *J. Comp. Neurol.*, vol. 513, no. 5, pp. 532–541, 2009.

[28] K.-L. Du, C.-S. Leung, W. H. Mow, and M. Swamy, "Perceptron: Learning, generalization, model selection, fault tolerance, and role in the deep learning era," *Mathematics*, vol. 10, 2022, Art. no. 4730.

[29] A. I. Rodríguez and X. D. Buitrago, "How to choose an activation function for deep learning," *Tekhnê*, vol. 19, no. 1, pp. 23–32, 2022.

[30] J. J. Hopfield, "Neural networks and physical systems with emergent collective computational abilities," in *Proc. Nat. Acad. Sci. USA*, vol. 79, no. 8, pp. 2554–2558, 1982.

[31] V. Nair and G. E. Hinton, "Rectified linear units improve restricted Boltzmann machines," in *Proc. 27th Int. Conf. Mach. Learn.*, 2010, pp. 807–814.

[32] B. Scellier and Y. Bengio, "Towards a biologically plausible backprop," 2016. [Online]. Available: https://api.semanticscholar.org/CorpusID:2488516

[33] T. P. Lillicrap, A. Santoro, L. Marris, C. J. Akerman, and G. Hinton, "Backpropagation and the brain," *Nat. Rev. Neurosci.*, vol. 21, no. 6, pp. 335–346, 2020.

[34] H. Markram, J. Lübke, M. Frotscher, and B. Sakmann, "Regulation of synaptic efficacy by coincidence of postsynaptic APs and EPSPs," *Sci.*, vol. 275, no. 5297, pp. 213–215, 1997. [Online]. Available: https://www.science.org/doi/abs/10.1126/science.275.5297.213

[35] M. Bouvier et al., "Spiking neural networks hardware implementations and challenges," *ACM J. Emerg. Technol. Comput. Syst.*, vol. 15, pp. 1–35, 2019.

[36] N. Rathi et al., "Exploring neuromorphic computing based on spiking neural networks: Algorithms to hardware," *ACM Comput. Surv.*, vol. 55, pp. 1–49, 2022.

[37] A. Tavanaei, M. Ghodrati, S. R. Kheradpisheh, T. Masquelier, and A. Maida, "Deep learning in spiking neural networks," *Neural Netw., Official J. Int. Neural Netw. Soc.*, vol. 111, pp. 47–63, 2018.

[38] C. Han and K. Lee, "A survey on spiking neural networks," *Int. J. Fuzzy Log. Intell. Syst.*, vol. 21, no. 4, pp. 317–328, 2021, doi: 10.5391/ijfis.2021.21.4.317.

[39] M. Zhang, Z. Gu, and G. Pan, "A survey of neuromorphic computing based on spiking neural networks," *Chin. J. Electron.*, vol. 27, pp. 667–674, 2018.

[40] M. Davies et al., "Advancing neuromorphic computing with loihi: A survey of results and outlook," in *Proc. IEEE*, vol. 109, no. 5, pp. 911–934, May 2021.

[41] Y. S. Yang and Y. Kim, "Recent trend of neuromorphic computing hardware: Intel's neuromorphic system perspective," in *Proc. 2020 Int. SoC Des. Conf.*, 2020, pp. 218–219.

[42] S. Dey and A. Dimitrov, "Mapping and validating a point neuron model on intel's neuromorphic hardware Loihi," *Front. Neurosci.*, vol. 16, 2021, Art. no. 883360.

[43] B. A. Olshausen and D. J. Field, "Emergence of simple-cell receptive field properties by learning a sparse code for natural images," *Nature*, vol. 381, no. 6583, pp. 607–609, 1996.

[44] X. Glorot, A. Bordes, and Y. Bengio, "Deep sparse rectifier neural networks," in *Proc. 14th Int. Conf. Artif. Intell. Statist.*, 2011, pp. 315–323.

[45] W. Wen, C. Wu, Y. Wang, Y. Chen, and H. Li, "Learning structured sparsity in deep neural networks," in *Proc. Adv. Neural Inf. Process. Syst.*, 2016, pp. 2082–2090.

[46] B. Zhang, I. Titov, and R. Sennrich, "Sparse attention with linear units," in *Proc. 2021 Conf. Empirical Methods Natural Lang. Process.*, 2021, pp. 6507–6520.

[47] Y. LeCun et al., "Backpropagation applied to handwritten zip code recognition," *Neural Comput.*, vol. 1, no. 4, pp. 541–551, 1989.

[48] Q. Wang, C. Fan, T. Jia, H. Yuyang, and X. Wu, "NDIM: Neuronal diversity inspired model for multisensory emotion recognition," 2024. [Online]. Available: https://openreview.net/forum?id=NrI1OkZkiy

[49] H. Tan, Y. Zhou, Q. Tao, J. Rosen, and S. van Dijken, "Bioinspired multisensory neural network with crossmodal integration and recognition," *Nat. Commun.*, vol. 12, no. 1, 2021, Art. no. 1120.

[50] J. Achterberg, D. Akarca, D. Strouse, J. Duncan, and D. E. Astle, "Spatially embedded recurrent neural networks reveal widespread links between structural and functional neuroscience findings," *Nat. Mach. Intell.*, vol. 5, no. 12, pp. 1369–1381, 2023.

[51] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural Comput.*, vol. 9, no. 8, pp. 1735–1780, 1997.

[52] Y. Li et al., "Dissecting neural computations in the human auditory pathway using deep neural networks for speech," *Nat. Neurosci.*, vol. 26, no. 12, pp. 2213–2225, 2023.

[53] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 770–778.

[54] J. L. Elman, "Finding structure in time," *Cogn. Sci.*, vol. 14, no. 2, pp. 179–211, 1990.

[55] M. Winding et al., "The connectome of an insect brain," *Sci.*, vol. 379, no. 6636, 2023, Art. no. eadd9330.

[56] R. Hwang et al., "Pre-gated MoE: An algorithm-system co-design for fast and scalable mixture-of-expert inference," in *Proc. ACM/IEEE 51st Annu. Int. Symp. Comput. Archit.*, 2024, pp. 1018–1031.

[57] Z. Zhang, Y. Lin, Z. Liu, P. Li, M. Sun, and J. Zhou, "Moefication: Transformer feed-forward layers are mixtures of experts," in *Proc. Assoc. Comput. Linguistics*, 2022, pp. 877–890.

[58] Z. Zhang et al., "Emergent modularity in pre-trained transformers," 2023, *arXiv:2305.18390*.

[59] Google, "Palm: Pathways language model," 2022. [Online]. Available: https://ai.google/discover/palm2/

[60] H. Touvron et al., "LLaMA: Open and efficient foundation language models," 2023, *arXiv:2302.13971*.

[61] Google, "Gemma: Open models initiative," 2023. [Online]. Available: https://blog.google/technology/developers/gemma-open-models/

[62] P. F. Simmering and P. Huoviala, "Large language models for aspect-based sentiment analysis," 2023, *arXiv:2310.18025*.

[63] W. Jiao et al., "Parrot: Translating during chat using large language models," 2023, *arXiv:2304.02426*.

[64] H. Gilbert, M. Sandborn, D. C. Schmidt, J. Spencer-Smith, and J. White, "Semantic compression with large language models," in *Proc. IEEE 10th Int. Conf. Social Netw. Anal. Manage. Secur.*, 2023, pp. 1–8.

[65] B. Robinson-Riegler and G. Robinson-Riegler, *Cognitive Psychology: Applying the Science of the Mind*. London, U.K.: Pearson, 2016.

[66] M. I. Posner, "Orienting of attention," *Quart. J. Exp. Psychol.*, vol. 32, pp. 3–25, 1980.

[67] A. Vaswani et al., "Attention is all you need," in *Proc. Adv. Neural Inf. Process. Syst.*, 2017, pp. 6000–6010.

[68] J. B. Aimone and O. Parekh, "The brain's unique take on algorithms," *Nat. Commun.*, vol. 14, no. 1, 2023, Art. no. 4910.

[69] Y. Keller, J. Mackensen, and S. Eger, "BERT-defense: A probabilistic model based on BERT to combat cognitively inspired orthographic adversarial attacks," 2021, *arXiv:2106.01452*.

[70] A. Naik, A. Ravichander, C. Rose, and E. Hovy, "Exploring numeracy in word embeddings," in *Proc. 57th Annu. Meeting Assoc. Comput. Linguistics*, Florence, Italy, 2019, pp. 3374–3380. [Online]. Available: https://aclanthology.org/P19-1329

[71] A. Kuhnle and A. Copestake, "The meaning of "most" for visual question answering models," in *Proc. 2019 ACL Workshop BlackboxNLP Analyzing Interpreting Neural Netw. NLP*, Florence, Italy, 2019, pp. 46–55. [Online]. Available: https://aclanthology.org/W19--4806

[72] L. O'Sullivan and S. Steinert-Threlkeld, "Neural models of the psychosemantics of 'most'," in *Proc. Workshop Cogn. Model. Comput. Linguistics*, 2019, pp. 140–151. [Online]. Available: https://aclanthology.org/W19--2916

[73] K. Li, F. Xie, H. Chen, K. Yuan, and X. Hu, "An audio-visual speech separation model inspired by cortico-thalamo-cortical circuits," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 46, no. 10, pp. 6637–6651, Oct. 2024.

[74] O. Eberle, S. Brandl, J. Pilot, and A. Søgaard, "Do transformer models show similar attention patterns to task-specific human gaze?," in *Proc. 60th Annu. Meeting Assoc. Comput. Linguistics*, 2022, pp. 4295–4309.

[75] J. Bensemann et al., "Eye gaze and self-attention: How humans and transformers attend words in sentences," in *Proc. Workshop Cogn. Model. Comput. Linguistics*, Dublin, Ireland, 2022, pp. 75–87. [Online]. Available: https://aclanthology.org/2022.cmcl-1.9

[76] E. Metheniti, T. Van De Cruys, and N. Hathout, "About time: Do transformers learn temporal verbal aspect?," in *Proc. Workshop Cogn. Model. Comput. Linguistics*, Dublin, Ireland, 2022, pp. 88–101. [Online]. Available: https://aclanthology.org/2022.cmcl-1.10

[77] J. Niu, W. Lu, and G. Penn, "Does BERT rediscover a classical NLP pipeline?," in *Proc. 29th Int. Conf. Comput. Linguistics*, Gyeongju, South Korea, 2022, pp. 3143–3153. [Online]. Available: https://aclanthology.org/2022.coling-1.278

[78] J. Cai, Z. Zhang, F. Wu, and J. Wang, "Deep cognitive reasoning network for multi-hop question answering over knowledge graphs," in *Proc. Assoc. Comput. Linguistics-Int. Joint Conf. Natural Lang. Process.*, 2021, pp. 219–229.

[79] M. Ding, C. Zhou, Q. Chen, H. Yang, and J. Tang, "Cognitive graph for multi-hop reading comprehension at scale," in *Proc. 57th Annu. Meeting Assoc. Comput. Linguistics*, Florence, Italy, 2019, pp. 2694–2703. [Online]. Available: https://aclanthology.org/P19--1259

[80] J. Mar and J. Liu, "From cognitive to computational modeling: Text-based risky decision-making guided by fuzzy trace theory," in *Proc. Conf. North Amer. Chapter Assoc. Comput. Linguistics*, 2022, pp. 391–409. [Online]. Available: https://aclanthology.org/2022.findings-naacl.30

[81] A. Madaan, N. Tandon, D. Rajagopal, P. Clark, Y. Yang, and E. Hovy, "Think about it! improving defeasible reasoning by first modeling the question scenario," in *Proc. 2021 Conf. Empirical Methods Natural Lang. Process.*, 2021, pp. 6291–6310. [Online]. Available: https://aclanthology.org/2021.emnlp-main.508

[82] H. Zhang, M. Chen, H. Wang, Y. Song, and D. Roth, "Analogous process structure induction for sub-event sequence prediction," in *Proc. 2020 Conf. Empirical Methods Natural Lang. Process.*, 2020, pp. 1541–1550. [Online]. Available: https://aclanthology.org/2020.emnlp-main.119

[83] T. Ates et al., "Craft: A benchmark for causal reasoning about forces and interactions," 2020, *arXiv: 2012.04293*.

[84] M. Ren, X. Geng, T. Qin, H. Huang, and D. Jiang, "Towards interpretable reasoning over paragraph effects in situation," in *Proc. 2020 Conf. Empirical Methods Natural Lang. Process.*, 2020, pp. 6745–6758. [Online]. Available: https://aclanthology.org/2020.emnlp-main.548

[85] J. Wei et al., "Chain-of-thought prompting elicits reasoning in large language models," in *Proc. Adv. Neural Inf. Process. Syst.*, 2022, pp. 24 824–24 837.

[86] D. C. Knill and W. Richards, Eds., *Perception as Bayesian Inference*. Cambridge, UK: Cambridge Univ. Press, 1996.

[87] J. Valvoda, N. Saphra, J. Rawski, A. Williams, and R. Cotterell, "Benchmarking compositionality with formal languages," in *Proc. 29th Int. Conf. Comput. Linguistics*, Gyeongju, South Korea, 2022, pp. 6007–6018. [Online]. Available: https://aclanthology.org/2022.coling-1.525

[88] U. Berger, G. Stanovsky, O. Abend, and L. Frermann, "A computational acquisition model for multimodal word categorization," in *Proc. 2022 Conf. North Amer. Chapter Assoc. Comput. Linguistics-Hum. Lang. Technol.*, 2022, pp. 3819–3835. [Online]. Available: https://aclanthology.org/2022.naacl-main.280

[89] N. Shi, B. Wang, W. Wang, X. Liu, and Z. Lin, "Revisit systematic generalization via meaningful learning," in *Proc. 5th BlackboxNLP Workshop Analyzing Interpreting Neural Netw. NLP*, Abu Dhabi, UAE, 2022, pp. 62–79. [Online]. Available: https://aclanthology.org/2022.blackboxnlp-1.6

[90] T. Yamakoshi, T. L. Griffiths, and R. D. Hawkins, "Probing BERT's priors with serial reproduction chains," 2022, *arXiv:2202.12226*.

[91] W. Garcia, H. Clouse, and K. Butler, "Disentangling categorization in multi-agent emergent communication," in *Proc. 2022 Conf. North Amer. Chapter Assoc. Comput. Linguistics-Hum. Lang. Technol.*, 2022, pp. 4523–4540. [Online]. Available: https://aclanthology.org/2022.naacl-main.335

[92] R. Schaeffer, B. Miranda, and S. Koyejo, "Are emergent abilities of large language models a mirage?," in *Proc. Adv. Neural Inf. Process. Syst.*, 2024, pp. 55565–55581.

[93] A. Testoni and R. Bernardi, "Looking for confirmations: An effective and human-like visual dialogue strategy," in *Proc. 2021 Conf. Empirical Methods Natural Lang. Process.*, 2021, pp. 9330–9338. [Online]. Available: https://aclanthology.org/2021.emnlp-main.736

[94] G. M. Van de Ven, H. T. Siegelmann, and A. S. Tolias, "Brain-inspired replay for continual learning with artificial neural networks," *Nat. Commun.*, vol. 11, no. 1, 2020, Art. no. 4069.

[95] S. Pinker, D. Wolff, M. Diamond, M. Pollen, E. Watters, and A. Reverman, *The Stuff of Thought: Language as a Window Into Human Nature*. Baltimore, MD, USA: Penguin, 2019.

[96] B. L. Whorf, *Language, Thought, and Reality: Selected Writings of Benjamin Lee Whorf*. Cambridge, MA, USA: MIT Press, 1956.

[97] F. Ungerer and H.-J. Schmid, *An Introduction to Cognitive Linguistics*. Evanston, IL, USA: Routledge, 2013.

[98] W. Peng et al., "Bi-directional CognitiveThinking network for machine reading comprehension," in *Proc. 28th Int. Conf. Comput. Linguistics*, Barcelona, Spain, 2020, pp. 2613–2623. [Online]. Available: https://aclanthology.org/2020.coling-main.235

[99] K. Sun, D. Yu, D. Yu, and C. Cardie, "Improving machine reading comprehension with general reading strategies," in *Proc. 2019 Conf. North Amer. Chapter Assoc. Comput. Linguistics-Hum. Lang. Technol.*, 2019, pp. 2633–2643. [Online]. Available: https://aclanthology.org/N19-1270

[100] Z. Li, C. Niu, F. Meng, Y. Feng, Q. Li, and J. Zhou, "Incremental transformer with deliberation decoder for document grounded conversations," in *Proc. 57th Annu. Meeting Assoc. Comput. Linguistics*, Florence, Italy, 2019, pp. 12–21. [Online]. Available: https://aclanthology.org/P19--1002

[101] Y. Lu, W. Zhu, X. Wang, M. Eckstein, and W. Y. Wang, "Imagination-augmented natural language understanding," in *Proc. 2022 Conf. North Amer. Chapter Assoc. Comput. Linguistics-Hum. Lang. Technol.*, 2022, pp. 4392–4402. [Online]. Available: https://aclanthology.org/2022.naacl-main.326

[102] Y. Gu, B. Dalvi, and P. Clark, "DREAM: Improving situational QA by first elaborating the situation," in *Proc. 2022 Conf. North Amer. Chapter Assoc. Comput. Linguistics-Hum. Lang. Technol.*, 2022, pp. 1115–1127. [Online]. Available: https://aclanthology.org/2022.naacl-main.82

[103] J. Nevens, J. Doumen, P. Van Eecke, and K. Beuls, "Language acquisition through intention reading and pattern finding," in *Proc. 29th Int. Conf. Comput. Linguistics*, Gyeongju, South Korea, 2022, pp. 15–25. [Online]. Available: https://aclanthology.org/2022.coling-1.2

[104] H. Shahmohammadi, H. P. A. Lensch, and R. H. Baayen, "Learning zero-shot multifaceted visually grounded word embeddings via multi-task training," in *Proc. 25th Conf. Comput. Natural Lang. Learn.*, 2021, pp. 158–170. [Online]. Available: https://aclanthology.org/2021.conll-1.12

[105] L. Pessoa, "On the relationship between emotion and cognition," *Nat. Rev. Neurosci.*, vol. 9, no. 2, pp. 148–158, 2008.

[106] Y. Xie, Y. Hu, W. Peng, G. Bi, and L. Xing, "COMMA: Modeling relationship among motivations, emotions and actions in language-based human activities," in *Proc. 29th Int. Conf. Comput. Linguistics*, Gyeongju, South Korea, 2022, pp. 163–177. [Online]. Available: https://aclanthology.org/2022.coling-1.15

[107] D. Hu, L. Wei, and X. Huai, "DialogueCRN: Contextual reasoning networks for emotion recognition in conversations," in *Proc. 59th Annu. Meeting Assoc. Comput. Linguistics-11th Int. Joint Conf. Natural Lang. Process.*, 2021, pp. 7042–7052. [Online]. Available: https://aclanthology.org/2021.acl-long.547

[108] H. Kim, B. Kim, and G. Kim, "Perspective-taking and pragmatics for generating empathetic responses focused on emotion causes," in *Proc. 2021 Conf. Empirical Methods Natural Lang. Process.*, 2021, pp. 2227–2240. [Online]. Available: https://aclanthology.org/2021.emnlp-main.170

[109] C. Li et al., "The good, the bad, and why: Unveiling emotions in generative AI," 2023, *arXiv:2312.11111*.

[110] A. Finkelstein, D. Derdikman, A. Rubin, J. N. Foerster, L. Las, and N. Ulanovsky, "Three-dimensional head-direction coding in the bat brain," *Nature*, vol. 517, no. 7533, pp. 159–164, 2015.

[111] M. Maroso, "A quest into the human brain," *Science*, vol. 382, no. 6667, pp. 166–167, 2023. [Online]. Available: https://www.science.org/doi/abs/10.1126/science.adl0913

[112] B. Xu and M.-M. Poo, "Large language models and brain-inspired general intelligence," *Nat. Sci. Rev.*, vol. 10, no. 10, 2023, Art. no. nwad267.

[113] E. Genç et al., "Diffusion markers of dendritic density and arborization in gray matter predict differences in intelligence," *Nat. Commun.*, vol. 9, no. 1, 2018, Art. no. 1905.

[114] L. Jiao et al., "Brain-inspired learning, perception, and cognition: A comprehensive review," *IEEE Trans. Neural Netw. Learn. Syst.*, early access, May 29, 2024, doi: 10.1109/TNNLS.2024.3401711.
[115] M.-M. Poo, "Transcriptome, connectome and neuromodulation of the primate brain," *Cell*, vol. 185, no. 15, pp. 2636–2639, 2022.
[116] T. Zhang, X. Cheng, S. Jia, M.-M. Poo, Y. Zeng, and B. Xu, "Self-backpropagation of synaptic modifications elevates the efficiency of spiking and artificial neural networks," *Sci. Adv.*, vol. 7, no. 43, 2021, Art. no. eabh0146.
[117] T. Zhang, X. Cheng, S. Jia, C. T. Li, M.-M. Poo, and B. Xu, "A brain-inspired algorithm that mitigates catastrophic forgetting of artificial and spiking neural networks with low computational cost," *Sci. Adv.*, vol. 9, no. 34, 2023, Art. no. eadi2947.
[118] A. M. Turing, *Computing Machinery and Intelligence*. Berlin, Germany: Springer, 2009.
[119] M. Assran et al., "Self-supervised learning from images with a joint-embedding predictive architecture," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2023, pp. 15 619–15 629.
[120] J. B. Tenenbaum, C. Kemp, T. L. Griffiths, and N. D. Goodman, "How to grow a mind: Statistics, structure, and abstraction," *Science*, vol. 331, no. 6022, pp. 1279–1285, 2011.
[121] R. Fang, R. Bindu, A. Gupta, and D. Kang, "LLM agents can autonomously exploit one-day vulnerabilities," 2024, *arXiv:2404.08144*.
[122] R. Abiri, S. Borhani, E. W. Sellers, Y. Jiang, and X. Zhao, "A comprehensive review of EEG-based brain–computer interface paradigms," *J. Neural Eng.*, vol. 16, no. 1, Jan. 2019, Art. no. 011001, doi: 10.1088/1741-2552/aaf12e.
[123] C. Cooney, R. Folli, and D. Coyle, "Neurolinguistics research advancing development of a direct-speech brain-computer interface," *IScience*, vol. 8, pp. 103–125, 2018.
[124] Y. S. Lincoln and E. G. Guba, *Naturalistic Inquiry*. Newbury Park, CA, USA: Sage, 1985.
[125] F. R. Willett et al., "A high-performance speech neuroprosthesis," *Nature*, vol. 620, no. 7976, pp. 1031–1036, 2023.
[126] S. L. Metzger et al., "A high-performance neuroprosthesis for speech decoding and avatar control," *Nature*, vol. 620, no. 7976, pp. 1037–1046, 2023.
[127] X. Chen et al., "A neural speech decoding framework leveraging deep learning and speech synthesis," *Nature Mach. Intell.*, pp. 1–14, 2024. [Online]. Available: https://www.biorxiv.org/content/early/2023/09/17/2023.09.16.558028
[128] N. S. Card et al., "An accurate and rapidly calibrating speech neuroprosthesis," *New England J. Med.*, vol. 391, no. 7, pp. 609–618, 2024. [Online]. Available: https://www.nejm.org/doi/full/10.1056/NEJMoa2314132
[129] M. Palatucci, D. Pomerleau, G. E. Hinton, and T. M. Mitchell, "Zero-shot learning with semantic output codes," in *Proc. Adv. Neural Inf. Process. Syst.*, 2009, Art. no. 1410.
[130] F. Pereira, G. Detre, and M. Botvinick, "Generating text from functional brain images," *Front. Human. Neurosci.* vol. 5, 2011, Art. no. 72.
[131] K.-M. K. Chang, T. Mitchell, and M. A. Just, "Quantitative modeling of the neural representation of objects: How semantic feature norms can account for fmri activation," *NeuroImage*, vol. 56, no. 2, pp. 716–727, 2011.
[132] F. Pereira, M. Botvinick, and G. Detre, "Using wikipedia to learn semantic feature representations of concrete concepts in neuroimaging experiments," *Artif. Intell.*, vol. 194, pp. 240–252, 2013.
[133] A. J. Anderson, D. Kiela, S. Clark, and M. Poesio, "Visually grounded and textual semantic models differentially decode brain activity associated with concrete and abstract nouns," *Trans. Assoc. Comput. Linguistics*, vol. 5, pp. 17–30, 2017.
[134] S. Wang, J. Zhang, H. Wang, N. Lin, and C. Zong, "Fine-grained neural decoding with distributed word representations," *Inf. Sci.*, vol. 507, pp. 256–272, 2020.
[135] F. Pereira et al., "Toward a universal decoder of linguistic meaning from brain activation," *Nat. Commun.*, vol. 9, no. 1, 2018, Art. no. 963.
[136] J. Sun, S. Wang, J. Zhang, and C. Zong, "Towards sentence-level brain decoding with distributed representations," in *Proc. AAAI Conf. Artif. Intell.*, 2019, pp. 7047–7054.
[137] J. Sun, S. Wang, J. Zhang, and C. Zong, "Neural encoding and decoding with distributed sentence representations," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 32, no. 2, pp. 589–603, Feb. 2021.
[138] N. Affolter, B. Egressy, D. Pascual, and R. Wattenhofer, "Brain2word: Decoding brain activity for language generation," 2020, *arXiv:2009.04765*.
[139] S. Zou, S. Wang, J. Zhang, and C. Zong, "Towards brain-to-text generation: Neural decoding with pre-trained encoder-decoder models," in *Proc. NeurIPS 2021 AI Sci. Workshop*, 2021. [Online]. Available: https://openreview.net/forum?id=13IJlk221xG
[140] S. Zou, S. Wang, J. Zhang, and C. Zong, "Cross-modal cloze task: A new task to brain-to-word decoding," in *Proc. Assoc. Comput. Linguistics*, 2022, pp. 648–657.
[141] X. Feng, X. Feng, and B. Qin, "Semantic-aware contrastive learning for electroencephalography-to-text generation with curriculum learning," 2023, *arXiv:2301.09237*.
[142] J. Tang, A. LeBel, S. Jain, and A. G. Huth, "Semantic reconstruction of continuous language from non-invasive brain recordings," *Nat. Neurosci.*, vol. 26, pp. 858–866, 2023.
[143] Z. Wang and H. Ji, "Open vocabulary electroencephalography-to-text decoding and zero-shot sentiment classification," in *Proc. AAAI Conf. Artif. Intell.*, 2022, pp. 5350–5358.
[144] Y. Duan, J. Zhou, Z. Wang, Y.-K. Wang, and C.-T. Lin, "DeWave: Discrete EEG waves encoding for brain dynamics to text translation," 2023, *arXiv:2309.14030*.
[145] X. Zhao, J. Sun, S. Wang, J. Ye, X. Zhang, and C. Zong, "MapGuide: A simple yet effective method to reconstruct continuous language from brain activities," 2024, *arXiv:2403.17516*.
[146] H. Jo, Y. Yang, J. Han, Y. Duan, H. Xiong, and W. H. Lee, "Are EEG-to-text models working?," 2024, *arXiv:2405.06459*.
[147] M. Eickenberg, A. Gramfort, G. Varoquaux, and B. Thirion, "Seeing it all: Convolutional network layers map the function of the human visual system," *NeuroImage*, vol. 152, pp. 184–194, 2017.
[148] R. Beliy, G. Gaziv, A. Hoogi, F. Strappini, T. Golan, and M. Irani, "From voxels to pixels and back: Self-supervision in natural-image reconstruction from fMRI," in *Proc. Adv. Neural Inf. Process. Syst.*, 2019, pp. 6517–6527.
[149] A. G. Huth, T. Lee, S. Nishimoto, N. Y. Bilenko, A. T. Vu, and J. L. Gallant, "Decoding the semantic content of natural movies from human brain activity," *Front. Syst. Neurosci.*, vol. 10, 2016, Art. no. 81.
[150] S. Nishimoto, A. T. Vu, T. Naselaris, Y. Benjamini, B. Yu, and J. L. Gallant, "Reconstructing visual experiences from brain activity evoked by natural movies," *Curr. Biol.*, vol. 21, no. 19, pp. 1641–1646, 2011.
[151] Y. Takagi and S. Nishimoto, "High-resolution image reconstruction with latent diffusion models from human brain activity," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2023, pp. 14453–14463.
[152] P. Singh, P. Pandey, K. Miyapuram, and S. Raman, "EEG2Image: Image reconstruction from EEG brain signals," in *Proc. 2023 IEEE Int. Conf. Acoust. Speech Signal Process.*, 2023, pp. 1–5.
[153] D. Li, C. Wei, S. Li, J. Zou, and Q. Liu, "Visual decoding and reconstruction via EEG embeddings with guided diffusion," 2024, *arXiv:2403.07721*.
[154] R. Quan, W. Wang, Z. Tian, F. Ma, and Y. Yang, "Psychometry: An omnifit model for image reconstruction from human brain activity," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2024, pp. 233–243.
[155] S. Wang, S. Liu, Z. Tan, and X. Wang, "MindBridge: A cross-subject brain decoding framework," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2024, pp. 11 333–11 342.
[156] H. Yang, J. Gee, and J. Shi, "Brain decodes deep nets," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2024, pp. 23 030–23 040.
[157] E. Takmaz, S. Pezzelle, L. Beinborn, and R. Fernández, "Generating image descriptions via sequential cross-modal alignment guided by human gaze," in *Proc. 2020 Conf. Empirical Methods Natural Lang. Process.*, 2020, pp. 4664–4677. [Online]. Available: https://aclanthology.org/2020.emnlp-main.377
[158] Y. Ren and D. Xiong, "CogAlign: Learning to align textual neural representations to cognitive language processing signals," 2021, *arXiv:2106.05544*.
[159] Y. Luo, M. Xu, and D. Xiong, "CogTaskonomy: Cognitively inspired task taxonomy is beneficial to transfer learning in NLP," in *Proc. 60th Annu. Meeting Assoc. Comput. Linguistics*, Dublin, Ireland, 2022, pp. 904–920. [Online]. Available: https://aclanthology.org/2022.acl-long.64
[160] A. Mishra, D. Kanojia, S. Nagar, K. Dey, and P. Bhattacharyya, "Leveraging cognitive features for sentiment analysis," 2017, *arXiv: 1701.05581*.
[161] M. Barrett, J. Bingel, N. Hollenstein, M. Rei, and A. Søgaard, "Sequence classification with human attention," in *Proc. 22nd Conf. Comput. Natural Lang. Learn.*, 2018, pp. 302–312.
[162] N. Hollenstein and C. Zhang, "Entity recognition at first sight: Improving NER with eye movement information," in *Proc. 2019 Conf. North Amer. Chapter Assoc. Comput. Linguistics-Hum. Lang. Technol.*, 2019, pp. 1–10. [Online]. Available: https://aclanthology.org/N19-1001

[163] M. Strzyz, D. Vilares, and C. Gómez-Rodríguez, "Towards making a dependency parser see," 2019, *arXiv: 1909.01053*.
[164] N. Hollenstein, M. Barrett, M. Troendle, F. Bigiolli, N. Langer, and C. Zhang, "Advancing NLP with cognitive language processing signals," 2019, *arXiv: 1904.02682*.
[165] T. Kuribayashi, Y. Oseki, A. Brassard, and K. Inui, "Context limitations make neural language models more human-like," in *Proc. 2022 Conf. Empirical Methods Natural Lang. Process.*, Abu Dhabi, UAE, 2022, pp. 10 421–10 436. [Online]. Available: https://aclanthology.org/2022.emnlp-main.712
[166] X. Ding, B. Chen, L. Du, B. Qin, and T. Liu, "CogBERT: Cognition-guided pre-trained language models," in *Proc. 29th Int. Conf. Comput. Linguistics*, 2022, pp. 3210–3225.
[167] E. McGuire and N. Tomuro, "Relation classification with cognitive attention supervision," in *Proc. Workshop Cogn. Model. Comput. Linguistics*, 2021, pp. 222–232. [Online]. Available: https://aclanthology.org/2021.cmcl-1.26
[168] L. Muttenthaler, N. Hollenstein, and M. Barrett, "Human brain activity for machine attention," 2020, *arXiv: 2006.05113*.
[169] Y. Ren and D. Xiong, "CogAlign: Learning to align textual neural representations to cognitive language processing signals," in *Proc. 59th Annu. Meeting Assoc. Comput. Linguistics-11th Int. Joint Conf. Natural Lang. Process.*, 2021, pp. 3758–3769. [Online]. Available: https://aclanthology.org/2021.acl-long.291
[170] H. Srivastava, "Poirot at CMCL 2022 shared task: Zero shot crosslingual eye-tracking data prediction using multilingual transformer models," in *Proc. Workshop Cogn. Model. Comput. Linguistics*, Dublin, Ireland, 2022, pp. 102–107. [Online]. Available: https://aclanthology.org/2022.cmcl-1.11
[171] J. M. Imperial, "NU HLT at CMCL 2022 shared task: Multilingual and crosslingual prediction of human reading behavior in universal language space," in *Proc. Workshop Cogn. Model. Comput. Linguistics*, Dublin, Ireland, 2022, pp. 108–113. [Online]. Available: https://aclanthology.org/2022.cmcl-1.12
[172] L. Salicchi, R. Xiang, and Y.-Y. Hsu, "HkAmsters at CMCL 2022 shared task: Predicting eye-tracking data from a gradient boosting framework with linguistic features," in *Proc. Workshop Cogn. Model. Comput. Linguistics*, Dublin, Ireland, 2022, pp. 114–120. [Online]. Available: https://aclanthology.org/2022.cmcl-1.13
[173] B.-D. Oh and W. Schuler, "Entropy- and distance-based predictors from GPT-2 attention patterns predict reading times over and above GPT-2 surprisal," in *Proc. 2022 Conf. Empirical Methods Natural Lang. Process.*, Abu Dhabi, UAE, 2022, pp. 9324–9334. [Online]. Available: https://aclanthology.org/2022.emnlp-main.632
[174] Y. Ren and D. Xiong, "Bridging between cognitive processing signals and linguistic features via a unified attentional network," in *Proc. AAAI Conf. Artif. Intell.*, 2022, pp. 49–58.
[175] N. Hollenstein and L. Beinborn, "Relative importance in sentence processing," in *Proc. 59th Annu. Meeting Assoc. Comput. Linguistics-11th Int. Joint Conf. Natural Lang. Process.*, 2021, pp. 141–150. [Online]. Available: https://aclanthology.org/2021.acl-short.19
[176] L. Wehbe, B. Murphy, P. Talukdar, A. Fyshe, A. Ramdas, and T. Mitchell, "Simultaneously uncovering the patterns of brain regions involved in different story reading subprocesses," *PLoS One*, vol. 9, no. 11, 2014, Art. no. e112575.
[177] A. G. Huth, W. A. De Heer T. L. Griffiths, F. E. Theunissen, and J. L. Gallant, "Natural speech reveals the semantic maps that tile human cerebral cortex," *Nature*, vol. 532, no. 7600, pp. 453–458, 2016.
[178] H. Cai et al., "Brain organoid reservoir computing for artificial intelligence," *Nat. Electron.*, vol. 6, no. 12, pp. 1032–1039, 2023.
[179] T. M. Mitchell et al., "Predicting human brain activity associated with the meanings of nouns," *science*, vol. 320, no. 5880, pp. 1191–1195, 2008.
[180] A. G. Wilson, C. Dann, C. Lucas, and E. P. Xing, "The human kernel," in *Proc. Adv. Neural Inf. Process. Syst.*, 2015, pp. 2854–2862.
[181] Y.-P. Ruan, Z.-H. Ling, and Y. Hu, "Exploring semantic representation in brain activity using word embeddings," in *Proc. 2016 Conf. Empirical Methods Natural Lang. Process.*, 2016, pp. 669–679.
[182] H. Xu, B. Murphy, and A. Fyshe, "BrainBench: A brain-image test suite for distributional semantic models," in *Proc. 2016 Conf. Empirical Methods Natural Lang. Process.*, 2016, pp. 2017–2021.
[183] J. Bingel, M. Barrett, and A. Søgaard, "Extracting token-level signals of syntactic processing from fMRI-with an application to PoS induction," in *Proc. 54th Annu. Meeting Assoc. Comput. Linguistics*, 2016, pp. 747–755.
[184] L. Bulat, S. Clark, and E. Shutova, "Speaking, seeing, understanding: Correlating semantic models with conceptual representation in the brain," in *Proc. 2017 Conf. Empirical Methods Natural Lang. Process.*, 2017, pp. 1081–1091.
[185] S. Abnar, R. Ahmed, M. Mijnheer, and W. Zuidema, "Experiential, distributional and dependency-based word embeddings have complementary roles in decoding brain activity," 2017, *arXiv: 1711.09285*.
[186] D. Schwartz, M. Toneva, and L. Wehbe, "Inducing brain-relevant bias in natural language processing models," in *Proc. Adv. Neural Inf. Process. Syst.*, 2019, pp. 14123–14133.
[187] S. Jat, H. Tang, P. Talukdar, and T. Mitchell, "Relating simple sentence representations in deep neural networks and the brain," 2019, *arXiv: 1906.11861*.
[188] A. Fyshe, P. Talukdar, B. Murphy, and T. Mitchell, "Interpretable semantic vectors from a joint model of brain-and text-based meaning," in *Proc. 52nd Annu. Meeting Assoc. Comput. Linguistics*, 2014, pp. 489–499.
[189] B. Murphy, P. Talukdar, and T. Mitchell, "Learning effective and interpretable semantic models using non-negative sparse embedding," in *Proc. Int. Conf. Comput. Linguistics*, 2012, pp. 1933–1950.
[190] A. Søgaard, "Evaluating word embeddings with fMRI and eye-tracking," in *Proc. 1st Workshop Evaluating Vector-Space Representations NLP*, 2016, pp. 116–121.
[191] J. Gauthier and A. Ivanova, "Does the brain represent words? An evaluation of brain decoding studies of language understanding," 2018, *arXiv: 1806.00591*.
[192] S. Jain and A. Huth, "Incorporating context into language encoding models for fMRI," in *Proc. Adv. Neural Inf. Process. Syst.*, 2018, pp. 6629–6638.
[193] P. Qian, X. Qiu, and X. Huang, "Bridging LSTM architecture and the neural dynamics during reading," in *Proc. 25th Int. Joint Conf. Artif. Intell.*, 2016, pp. 1953–1959.
[194] D. Schwartz and T. Mitchell, "Understanding language-elicited EEG data by predicting it from a fine-tuned language model," in *Proc. 2019 Conf. North Amer. Chapter Assoc. Comput. Linguistics-Hum. Lang. Technol.*, 2019, pp. 43–57.
[195] M. Hashemzadeh, G. Kaufeld, M. White, A. E. Martin, and A. Fyshe, "From language to language-ish: How brain-like is an LSTM's representation of nonsensical language stimuli?," 2020, *arXiv: 2010.07435*.
[196] S. Jain and B. C. Wallace, "Attention is not explanation," in *Proc. Conf. North Amer. Chapter Assoc. Comput. Linguistics - Hum. Lang. Technol.*, 2019, pp. 3543–3556.
[197] S. Wiegreffe and Y. Pinter, "Attention is not not explanation," in *Proc. 2019 Conf. Empirical Methods Natural Lang. Process.-9th Int. Joint Conf. Natural Lang. Process.*, 2020, pp. 11–20.
[198] A. Rogers, O. Kovaleva, and A. Rumshisky, "A primer in BERTology: What we know about how BERT works," *Trans. Assoc. Comput. Linguistics*, vol. 8, pp. 842–866, 2021.
[199] S. Vashishth, S. Upadhyay, G. S. Tomar, and M. Faruqui, "Attention interpretability across NLP tasks," 2019, *arXiv: 1909.11218*.
[200] A. Galassi, M. Lippi, and P. Torroni, "Attention in natural language processing," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 32, no. 10, pp. 4291–4308, Oct. 2021.
[201] K. Clark, U. Khandelwal, O. Levy, and C. D. Manning, "What does BERT look at? An analysis of BERT's attention," in *Proc. 2019 ACL Workshop BlackboxNLP Analyzing Interpreting Neural Netw. NLP*, 2019, pp. 276–286.
[202] S. Abnar and W. Zuidema, "Quantifying attention flow in transformers," in *Proc. 58th Annu. Meeting Assoc. Comput. Linguistics*, 2020, pp. 4190–4197.
[203] M. Toneva and L. Wehbe, "Interpreting and improving natural-language processing (in machines) with natural language-processing (in the brain)," in *Proc. Adv. Neural Inf. Process. Syst.*, 2019, pp. 14954–14964.
[204] M. Schrimpf et al., "The neural architecture of language: Integrative modeling converges on predictive processing," in *Proc. Nat. Acad. Sci. USA*, vol. 118, no. 45, 2021, Art. no. e2105646118.
[205] C. Caucheteux, A. Gramfort, and J.-R. King, "Model-based analysis of brain activity reveals the hierarchy of language in 305 subjects," 2021, *arXiv:2110.06078*.
[206] J. C. Whittington, J. Warren, and T. E. Behrens, "Relating transformers to models and neural representations of the hippocampal formation," 2021, *arXiv:2112.04035*.
[207] J. Vig and Y. Belinkov, "Analyzing the structure of attention in a transformer language model," in *Proc. 2019 ACL Workshop BlackboxNLP Analyzing Interpreting Neural Netw. NLP*, 2019, pp. 63–76.

[208] L. He, P. Chen, E. Nie, Y. Li, and J. R. Brennan, "Decoding probing: Revealing internal linguistic structures in neural language models using minimal pairs," 2024, *arXiv:2403.17299*.
[209] M. Lamarre, C. Chen, and F. Deniz, "Attention weights accurately predict language representations in the brain," in *Proc. Findings Assoc. Comput. Linguistics: EMNLP 2022*, 2022, pp. 4513–4529.
[210] K. Ramakrishnan and F. Deniz, "Non-complementarity of information in word-embedding and brain representations in distinguishing between concrete and abstract words," in *Proc. Workshop Cogn. Model. Comput. Linguistics*, 2021, pp. 1–11. [Online]. Available: https://aclanthology.org/2021.cmcl-1.1
[211] S. Kumar et al., "Reconstructing the cascade of language processing in the brain using the internal computations of a transformer-based language model," *BioRxiv*, 2023, doi: 10.1101/2022.06.08.495348.
[212] J. Gauthier and R. Levy, "Linking artificial and human neural representations of language," in *Proc. 2019 Conf. Empirical Methods Natural Lang. Process.-9th Int. Joint Conf. Natural Lang. Process.*, 2019, pp. 529–539.
[213] J. Sun and M.-F. Moens, "Fine-tuned versus prompt-tuned supervised representations: Which better account for brain language representations?," 2023, *arXiv:2310.01854*.
[214] S. R. Oota, J. Arora, V. Rowtula, M. Gupta, and R. S. Bapi, "Visio-linguistic brain encoding," 2022, *arXiv:2204.08261*.
[215] S. R. Oota, E. Çelik, F. Deniz, and M. Toneva, "Speech language models lack important brain-relevant semantics," 2023, *arXiv:2311.04664*.
[216] Z. Zhang et al., "Chisco: An EEG-based BCI dataset for decoding of imagined speech," *Sci. Data*, vol. 11, no. 1, 2024, Art. no. 1265.
[217] Z. Zada et al., "A shared model-based linguistic space for transmitting our thoughts from brain to brain in natural conversations," *Neuron*, vol. 112, no. 18, pp. 3211–3222, 2024.
[218] M. Xiao et al., "Healme: Harnessing cognitive reframing in large language models for psychotherapy," 2024, *arXiv:2403.05574*.
[219] N. Chomsky, *Aspects of the Theory of Syntax*. Cambridge, MA, USA: MIT Press, 1965.
[220] A. Wiemerslage, S. Dudy, and K. Kann, "A comprehensive comparison of neural networks as cognitive models of inflection," 2022, *arXiv:2210.12321*.
[221] H. Jin, L. Cai, Y. Peng, C. Xia, A. McCarthy, and K. Kann, "Unsupervised morphological paradigm completion," in *Proc. 58th Annu. Meeting Assoc. Comput. Linguistics*, 2020, pp. 6696–6707. [Online]. Available: https://aclanthology.org/2020.acl-main.598
[222] G. Lakoff and M. Johnson, *Metaphors We Live By*. Chicago, Illinois, USA: Univ. Chicago Press, 2008.
[223] E. Shutova, L. Sun, and A. Korhonen, "Metaphor identification using verb and noun clustering," in *Proc. 23rd Int. Conf. Comput. Linguistics*, 2010, pp. 1002–1010.
[224] M. Mohler, D. Bracewell, M. Tomlinson, and D. Hinote, "Semantic signatures for example-based linguistic metaphor detection," in *Proc. 1st Workshop Metaphor NLP*, 2013, pp. 27–35.
[225] H. Jang, S. Moon, Y. Jo, and C. Rose, "Metaphor detection in discourse," in *Proc. 16th Annu. Meeting Special Int. Group Discourse Dialogue*, 2015, pp. 384–392.
[226] H. Jang, K. Maki, E. Hovy, and C. Rose, "Finding structure in figurative language: Metaphor detection with topic-based frames," in *Proc. 18th Annu. SIGDIAL Meeting Discourse Dialogue*, 2017, pp. 320–330.
[227] E. Shutova, L. Sun, E. D. Gutiérrez, P. Lichtenstein, and S. Narayanan, "Multilingual metaphor processing: Experiments with semi-supervised and unsupervised learning," *Comput. Linguistics*, vol. 43, no. 1, pp. 71–123, 2017.
[228] M. Pramanick, A. Gupta, and P. Mitra, "An LSTM-CRF based approach to token-level metaphor detection," in *Proc. Workshop Figurative Lang. Process.*, 2018, pp. 67–75.
[229] J. Liu, N. O'Hara, A. Rubin, R. Draelos, and C. Rudin, "Metaphor detection using contextual word embeddings from transformers," in *Proc. 2nd Workshop Figurative Lang. Process.*, 2020, pp. 250–255.
[230] X. Tong, E. Shutova, and M. Lewis, "Recent advances in neural metaphor processing: A linguistic, cognitive and social perspective," in *Proc. 2021 Conf. North Amer. Chapter Assoc. Comput. Linguistics-Hum. Lang. Technol.*, 2021, pp. 4673–4686. [Online]. Available: https://aclanthology.org/2021.naacl-main.372
[231] P. Pedinotti, E. Di Palma L. Cerini, and A. Lenci, "A howling success or a working sea? Testing what BERT knows about metaphors," in *Proc. 4th BlackboxNLP Workshop Analyzing Interpreting Neural Netw. NLP*, 2021, pp. 192–204. [Online]. Available: https://aclanthology.org/2021.blackboxnlp-1.13
[232] O. Zayed, J. P. McCrae, and P. Buitelaar, "Contextual modulation for relation-level metaphor identification," in *Proc. Conf. Empir. Methods Natural Lang. Process.*, 2020, pp. 388–406. [Online]. Available: https://aclanthology.org/2020.findings-emnlp.36
[233] R. Tamari, C. Shani, T. Hope, M. R. L. Petruck, O. Abend, and D. Shahaf, "Language (re)modelling: Towards embodied language understanding," in *Proc. 58th Annu. Meeting Assoc. Comput. Linguistics*, 2020, pp. 6268–6281. [Online]. Available: https://aclanthology.org/2020.acl-main.559
[234] M. Wan and B. Xing, "Modality enriched neural network for metaphor detection," in *Proc. 28th Int. Conf. Comput. Linguistics*, Barcelona, Spain, 2020, pp. 3036–3042. [Online]. Available: https://aclanthology.org/2020.coling-main.270
[235] E. Aghazadeh, M. Fayyaz, and Y. Yaghoobzadeh, "Metaphors in pre-trained language models: Probing and generalization across datasets and languages," 2022, *arXiv:2203.14139*.
[236] K. Stowe, T. Chakrabarty, N. Peng, S. Muresan, and I. Gurevych, "Metaphor generation with conceptual mappings," in *Proc. 59th Annu. Meeting Assoc. Comput. Linguistics-11th Int. Joint Conf. Natural Lang. Process.*, 2021, pp. 6724–6736. [Online]. Available: https://aclanthology.org/2021.acl-long.524
[237] L. Wachowiak and D. Gromann, "Does GPT-3 grasp metaphors? Identifying metaphor mappings with generative language models," in *Proc. 61st Annu. Meeting Assoc. Comput. Linguistics*, 2023, pp. 1018–1032.
[238] G. Fauconnier and M. Turner, "Conceptual integration networks," *Cogn. Sci.*, vol. 22, no. 2, pp. 133–187, 1998.
[239] S. Shen, D. Fried, J. Andreas, and D. Klein, "Pragmatically informative text generation," in *Proc. 2019 Conf. North Amer. Chapter Assoc. Comput. Linguistics-Hum. Lang. Technol.*, 2019, pp. 4060–4067. [Online]. Available: https://aclanthology.org/N19--1410
[240] J. L. Hoover, W. Du, A. Sordoni, and T. J. O'Donnell, "Linguistic dependencies and statistical dependence," in *Proc. 2021 Conf. Empirical Methods Natural Lang. Process.*, 2021, pp. 2941–2963. [Online]. Available: https://aclanthology.org/2021.emnlp-main.234
[241] A. Branco, J. António Rodrigues, M. Salawa, R. Branco, and C. Saedi, "Comparative probing of lexical semantics theories for cognitive plausibility and technological usefulness," in *Proc. 28th Int. Conf. Comput. Linguistics*, Barcelona, Spain, 2020, pp. 4004–4019. [Online]. Available: https://aclanthology.org/2020.coling-main.354
[242] C. Gao, J. Li, J. Chen, and S. Huang, "Measuring meaning composition in the human brain with composition scores from large language models," 2024, *arXiv:2403.04325*.
[243] I. M. Azpiazu and M. S. Pera, "Multiattentive recurrent neural network architecture for multilingual readability assessment," *Trans. Assoc. Comput. Linguistics*, vol. 7, pp. 421–436, 2019. [Online]. Available: https://aclanthology.org/Q19--1028
[244] X. Qiu, Y. Chen, H. Chen, J.-Y. Nie, Y. Shen, and D. Lu, "Learning syntactic dense embedding with correlation graph for automatic readability assessment," 2021, *arXiv:2107.04268*.
[245] D. M. Howcroft and V. Demberg, "Psycholinguistic models of sentence processing improve sentence readability ranking," in *Proc. 15th Conf. Eur. Chapter Assoc. Comput. Linguistics*, 2017, pp. 958–968.
[246] S. Vajjala and D. Meurers, "On improving the accuracy of readability classification using insights from second language acquisition," in *Proc. 7th Workshop Building Educ. Appl. NLP*, 2012, pp. 163–173.
[247] S. A. Crossley, H. S. Yang, and D. S. McNamara, "What's so simple about simplified texts? A computational and psycholinguistic investigation of text comprehension and text processing," *Reading Foreign Lang.*, vol. 26, no. 1, pp. 92–113, 2014.
[248] S. Evaldo Leal, J. M. Munguba Vieira, E. dos Santos Rodrigues, E. Nogueira Teixeira, and S. Aluísio, "Using eye-tracking data to predict the readability of Brazilian Portuguese sentences in single-task, multi-task and sequential transfer learning approaches," in *Proc. 28th Int. Conf. Comput. Linguistics*, Barcelona, Spain, 2020, pp. 5821–5831. [Online]. Available: https://aclanthology.org/2020.coling-main.512
[249] J. H. Lau, C. Armendariz, S. Lappin, M. Purver, and C. Shu, "How furiously can colorless green ideas sleep? Sentence acceptability in context," *Trans. Assoc. Comput. Linguistics*, vol. 8, pp. 296–310, 2020. [Online]. Available: https://aclanthology.org/2020.tacl-1.20
[250] B. Iavarone, D. Brunato, and F. Dell'Orletta, "Sentence complexity in context," in *Proc. Workshop Cogn. Model. Comput. Linguistics*, 2021, pp. 186–199.

[251] M. Byrd and S. Srivastava, "Predicting difficulty and discrimination of natural language questions," in *Proc. 60th Annu. Meeting Assoc. Comput.*, 2022, pp. 119–130.

[252] S. Rao, H. Zheng, and S. Li, "Cross-lingual leveled reading based on language-invariant features," in *Proc. Conf. Empir. Methods Natural Lang. Process.*, 2021, pp. 2677–2682. [Online]. Available: https://aclanthology.org/2021.findings-emnlp.227

[253] E. Chamovitz and O. Abend, "Cognitive simplification operations improve text simplification," 2022, *arXiv:2211.08825*.

[254] J. Yang, S. L. Frank, and A. van den Bosch, "Less is better: A cognitively inspired unsupervised model for language segmentation," in *Proc. Workshop Cogn. Aspects Lexicon*, 2020, pp. 33–45. [Online]. Available: https://aclanthology.org/2020.cogalex-1.4

[255] T. F. Jaeger, "Redundancy and reduction: Speakers manage syntactic information density," *Cogn. Psychol.*, vol. 61, no. 1, pp. 23–62, 2010.

[256] C. Meister, R. Cotterell, and T. Vieira, "If beam search is the answer, what was the question?," in *Proc. 2020 Conf. Empirical Methods Natural Lang. Process.*, 2020, pp. 2173–2185. [Online]. Available: https://aclanthology.org/2020.emnlp-main.170

[257] J. Wei, C. Meister, and R. Cotterell, "A cognitive regularizer for language modeling," in *Proc. 59th Annu. Meeting Assoc. Comput. Linguistics-11th Int. Joint Conf. Natural Lang. Process.*, 2021, pp. 5191–5202. [Online]. Available: https://aclanthology.org/2021.acl-long.404

[258] T. Kuribayashi, Y. Oseki, T. Ito, R. Yoshida, M. Asahara, and K. Inui, "Lower perplexity is not always human-like," in *Proc. 59th Annu. Meeting Assoc. Comput. Linguistics-11th Int. Joint Conf. Natural Lang. Process.*, 2021, pp. 5203–5217. [Online]. Available: https://aclanthology.org/2021.acl-long.405

[259] Z. Zhu, J. Novikova, and F. Rudzicz, "Detecting cognitive impairments by agreeing on interpretations of linguistic features," in *Proc. 2019 Conf. North Amer. Chapter Assoc. Comput. Linguistics-Hum. Lang. Technol.*, 2019, pp. 1431–1441. [Online]. Available: https://aclanthology.org/N19--1146

[260] N. Linz, K. Lundholm Fors, H. Lindsay, M. Eckerström, J. Alexandersson, and D. Kokkinakis, "Temporal analysis of the semantic verbal fluency task in persons with subjective and mild cognitive impairment," in *Proc. 6th Workshop Comput. Linguistics Clin. Psychol.*, 2019, pp. 103–113. [Online]. Available: https://aclanthology.org/W19-3012

[261] J. D. Choi, M. Li, F. Goldstein, and I. Hajjar, "Meta-semantic representation for early detection of Alzheimer's disease," in *Proc. 1st Int. Workshop Designing Meaning Representations*, Florence, Italy, 2019, pp. 82–91. [Online]. Available: https://aclanthology.org/W19-3309

[262] M. Asgari, L. Chen, and H. Dodge, "Topic-based measures of conversation for detecting mild CognitiveImpairment," in *Proc. 1st Workshop Natural Lang. Process. Med. Conversations*, 2020, pp. 63–67. [Online]. Available: https://aclanthology.org/2020.nlpmc-1.9

[263] H. Lindsay et al., "Dissociating semantic and phonemic search strategies in the phonemic verbal fluency task in early dementia," in *Proc. 7th Workshop Comput. Linguistics Clin. Psychol.*, 2021, pp. 32–44. [Online]. Available: https://aclanthology.org/2021.clpsych-1.4

[264] B. Li, Y.-T. Hsu, and F. Rudzicz, "Detecting dementia in Mandarin Chinese using transfer learning from a parallel corpus," in *Proc. 2019 Conf. North Amer. Chapter Assoc. Comput. Linguistics-Hum. Lang. Technol.*, 2019, pp. 1991–1997. [Online]. Available: https://aclanthology.org/N19-1199

[265] F. Di Palo and N. Parde, "Enriching neural models with targeted features for dementia detection," in *Proc. 57th Annu. Meeting Assoc. Comput. Linguistics Student Res. Workshop*, Florence, Italy, 2019, pp. 302–308. [Online]. Available: https://aclanthology.org/P19-2042

[266] S. Farzana, A. Deshpande, and N. Parde, "How you say it matters: Measuring the impact of verbal disfluency tags on automated dementia detection," in *Proc. 21st Workshop Biomed. Lang. Process.*, Dublin, Ireland, 2022, pp. 37–48. [Online]. Available: https://aclanthology.org/2022.bionlp-1.4

[267] C. Li, D. Knopman, W. Xu, T. Cohen, and S. Pakhomov, "GPT-D: Inducing dementia-related linguistic anomalies by deliberate degradation of artificial neural language models," 2022, *arXiv:2203.13397*.

[268] M. Ehghaghi, F. Rudzicz, and J. Novikova, "Data-driven approach to differentiating between depression and dementia from noisy speech and language data," in *Proc. 8th Workshop Noisy User-generated Text*, Gyeongju, South Korea, 2022, pp. 24–37. [Online]. Available: https://aclanthology.org/2022.wnut-1.3

[269] A. M. Schoene, G. Lacey, A. P. Turner, and N. Dethlefs, "Dilated LSTM with attention for classification of suicide notes," in *Proc. 10th Int. Workshop Health Text Mining Inf. Anal.*, Hong Kong, 2019, pp. 136–145. [Online]. Available: https://aclanthology.org/D19--6217

[270] S. Pai, N. Sachdeva, P. Sachdeva, and R. R. Shah, "Unsupervised paraphasia classification in aphasic speech," in *Proc. 58th Annu. Meeting Assoc. Comput. Linguistics Student Res. Workshop*, 2020, pp. 13–19. [Online]. Available: https://aclanthology.org/2020.acl-srw.3

[271] G. Tuckute, A. Sathe, M. Wang, H. Yoder, C. Shain, and E. Fedorenko, "SentSpace: Large-scale benchmarking and evaluation of text using cognitively motivated lexical, syntactic, and semantic features," in *Proc. 2022 Conf. North Amer. Chapter Assoc. Comput. Linguistics-Hum. Lang. Technol. Syst. Demonstrations*, 2022, pp. 99–113. [Online]. Available: https://aclanthology.org/2022.naacl-demo.11

[272] L. Weissweiler, V. Hofmann, A. Köksal, and H. Schütze, "The better your syntax, the better your semantics? Probing pretrained language models for the English comparative correlative," in *Proc. 2022 Conf. Empirical Methods Natural Lang. Process.*, Abu Dhabi, UAE, 2022, pp. 10 859–10 882. [Online]. Available: https://aclanthology.org/2022.emnlp-main.746

[273] Z. Hu, H. P. Chan, and L. Huang, "MOCHA: A multi-task training approach for coherent text generation from cognitive perspective," in *Proc. 2022 Conf. Empirical Methods Natural Lang. Process.*, Abu Dhabi, UAE, 2022, pp. 10 324–10 334. [Online]. Available: https://aclanthology.org/2022.emnlp-main.705

[274] R. Yoshida, H. Noji, and Y. Oseki, "Modeling human sentence processing with left-corner recurrent neural network grammars," in *Proc. 2021 Conf. Empirical Methods Natural Lang. Process.*, 2021, pp. 2964–2973. [Online]. Available: https://aclanthology.org/2021.emnlp-main.235

[275] J. Russin, J. Jo, R. O'Reilly, and Y. Bengio, "Compositional generalization by factorizing alignment and translation," in *Proc. 58th Annu. Meeting Assoc. Comput. Linguistics Student Res. Workshop*, 2020, pp. 313–327. [Online]. Available: https://aclanthology.org/2020.acl-srw.42

[276] A. Murphy, B. Bohnet, R. McDonald, and U. Noppeney, "Decoding part-of-speech from human EEG signals," in *Proc. 60th Annu. Meeting Assoc. Comput. Linguistics*, 2022, pp. 2201–2210.

[277] X. Zhang, S. Wang, N. Lin, and C. Zong, "Is the brain mechanism for hierarchical structure building universal across languages? An fMRI study of chinese and english," in *Proc. 2022 Conf. Empirical Methods Natural Lang. Process.*, 2022, pp. 7852–7861.

[278] R. G. d'Andrade, *The Development of Cognitive Anthropology*, Cambridge, U.K.: Cambridge Univ. Press, 1995.

[279] C.-P. Bara, S. CH-Wang, and J. Chai, "MindCraft: Theory of mind modeling for situated dialogue in collaborative tasks," 2021, *arXiv:2109.06275*.

[280] L. Wu, Y. Rao, Y. Lan, L. Sun, and Z. Qi, "Unified dual-view cognitive model for interpretable claim verification," in *Proc. 59th Annu. Meeting Assoc. Comput. Linguistics-11th Int. Joint Conf. Natural Lang. Process.*, 2021, pp. 59–68. [Online]. Available: https://aclanthology.org/2021.acl-long.5

[281] R. Ueda and K. Washio, "On the relationship between Zipf's law of abbreviation and interfering noise in emergent languages," in *Proc. 59th Annu. Meeting Assoc. Comput. Linguistics-11th Int. Joint Conf. Natural Lang. Process. Student Res. Workshop*, 2021, pp. 60–70. [Online]. Available: https://aclanthology.org/2021.acl-srw.6

[282] D. Francis, E. Rabinovich, F. Samir, D. Mortensen, and S. Stevenson, "Quantifying cognitive factors in lexical decline," *Trans. Assoc. Comput. Linguistics*, vol. 9, pp. 1529–1545, 2021. [Online]. Available: https://aclanthology.org/2021.tacl-1.91

[283] L. Yu and Y. Xu, "Predicting emergent linguistic compositions through time: Syntactic frame extension via multimodal chaining," in *Proc. 2021 Conf. Empirical Methods Natural Lang. Process.*, 2021, pp. 920–931. [Online]. Available: https://aclanthology.org/2021.emnlp-main.71

[284] M. Sap, E. Horvitz, Y. Choi, N. A. Smith, and J. Pennebaker, "Recollection versus imagination: Exploring human memory and cognition via neural language models," in *Proc. 58th Annu. Meeting Assoc. Comput. Linguistics*, 2020, pp. 1970–1978. [Online]. Available: https://aclanthology.org/2020.acl-main.178

[285] S. Pezzelle and R. Fernández, "Is the red square big? MALeViC: Modeling adjectives leveraging visual contexts," in *Proc. 2019 Conf. Empirical Methods Natural Lang. Process.-9th Int. Joint Conf. Natural Lang. Process.*, Hong Kong, China, 2019, pp. 2865–2876. [Online]. Available: https://aclanthology.org/D19--1285

[286] S. M. Mousavi, A. Cervone, M. Danieli, and G. Riccardi, "Would you like to tell me more? Generating a corpus of psychotherapy dialogues," in *Proc. 2nd Workshop Natural Lang. Process. Med. Conversations*, 2021, pp. 1–9. [Online]. Available: https://aclanthology.org/2021.nlpmc-1.1

[287] J. Du, H. Jiang, J. Shen, and X. Ren, "Eliciting knowledge from experts: Automatic transcript parsing for cognitive task analysis," in *Proc. 57th Annu. Meeting Assoc. Comput. Linguistics*, Florence, Italy, 2019, pp. 4280–4291. [Online]. Available: https://aclanthology.org/P19--1420
[288] S. Ji, T. Zhang, L. Ansari, J. Fu, P. Tiwari, and E. Cambria, "MentalBERT: Publicly available pretrained language models for mental healthcare," 2021, *arXiv:2110.15621*.
[289] X. Ding, K. Lybarger, J. Tauscher, and T. Cohen, "Improving classification of infrequent cognitive distortions: Domain-specific model versus data augmentation," in *Proc. 2022 Conf. North Amer. Chapter Assoc. Comput. Linguistics-Hum. Lang. Technol. Student Res. Workshop*, 2022, pp. 68–75. [Online]. Available: https://aclanthology.org/2022.naacl-srw.9
[290] K. Lybarger, J. Tauscher, X. Ding, D. Ben-zeev, and T. Cohen, "Identifying distorted thinking in patient-therapist text message exchanges by leveraging dynamic multi-turn context," in *Proc. 8th Workshop Comput. Linguistics Clin. Psychol.*, 2022, pp. 126–136. [Online]. Available: https://aclanthology.org/2022.clpsych-1.11
[291] E. J. Allen et al., "A massive 7T fMRI dataset to bridge cognitive neuroscience and artificial intelligence," *Nat. Neurosci.*, vol. 25, no. 1, pp. 116–126, 2022.

Zihan Zhang is currently working toward the PhD degree with the School of Computer Science and Technology, Harbin Institute of Technology(HIT), Harbin. His research interests include brain-machine interface, cognitive science and natural language processing.

Xiao Ding received the PhD degree from the School of Computer Science and Technology, Harbin Institute of Technology. He is currently a full professor with the Department of Computer Science, Harbin Institute of Technology. His research interests include natural language processing, text mining, social computing, and commonsense inference.

Xia Liang received the PhD degree from the State Key Laboratory of Cognitive Neuroscience and Learning, Beijing Normal University. She is currently an associate professor with the School of Space Environment and Material Science, Harbin Institude of Technology. Her research interests include Cognitive Neuroscience and the pathological mechanism of nerve/mental illness.

Yusheng Zhou working toward the master's degree from the School of Computer Science, Harbin Institute of Technology (HIT), Harbin, China. His research interests include Natural Language Processing, Brain Machine Interface and Cognitive Science.

Bing Qin received the PhD degree from the Department of Computer Science, Harbin Institute of Technology, China, in 2005. She is currently a full professor with the Department of Computer Science, and the director of the Research Center for Social Computing and Information Retrieval (HIT-SCIR), Harbin Institute of Technology. Her research interests include natural language processing, information extraction, document-level discourse analysis, and sentiment analysis.

Ting Liu received the PhD degree from the Department of Computer Science, Harbin Institute of Technology(HIT), China, in 1998. He is currently a full professor with the Department of Computer Science, HIT, where he also serves as the vice-principal of HIT. His research interests include natural language processing, information retrieval and social media analysis.