Received: 28 June 2025 | Revised: 27 October 2025 | Accepted: 21 November 2025 | Published online: 28 November 2025

REVIEW

FinTech and Sustainable Innovation
2025, Vol. 1, A15, 12 Pages
DOI: 10.47852/bonviewFSI52026630

BON VIEW PUBLISHING

A Review on Bridging Brain-Inspired Mechanisms and Large-Scale Pre-trained Models: Toward Adaptive, Efficient, and Interpretable AI

Jingzhe Wang[1], Yajing Wang[1] and Zongwei Luo[2,3,*]

[1]*Department of Computer Science and Technology, Beijing Normal-Hong Kong Baptist University, China*

[2]*Institute of Artificial Intelligence and Future Networks, Beijing Normal University at Zhuhai, China*

[3]*Guangdong Provincial Key Laboratory of IRADS, China*

Abstract: Large-scale pre-trained models, such as GPT, greatly improve numerous areas of artificial intelligence (AI), including natural language understanding, image recognition, and the integration of various data types. However, these models continue to encounter significant challenges, including excessive computational resource requirements, inadequate adaptability to dynamic environments, susceptibility to catastrophic forgetting, and limited internal interpretability. By comparison, the human brain exhibits efficient learning from sparse data, demonstrates robust adaptability across diverse contexts, operates with minimal energy consumption, retains information over extended periods, and can be elucidated through its underlying cognitive processes. This review examines recent research and is the first to categorize brain-inspired methods into three key dimensions. It explores how mechanisms of the human brain, such as hierarchical and modular designs, biologically inspired attention mechanisms, memory enhancement strategies, synaptic plasticity, and predictive coding, could inspire optimizations for large-scale models. This review not only synthesizes the current state of the field but also proposes potential directions for future research. There remains a need for stronger theories, better testing methods, improved hardware–software coordination, and careful consideration of ethical issues. Overcoming these challenges will require closer collaboration between neuroscientists, cognitive scientists, machine learning experts, and engineers. Through such collaboration, the field may develop the next generation of AI systems that perform at the highest levels while being more efficient, flexible, and transparent.

Keywords: large-scale pre-trained models, brain-inspired optimization, artificial intelligence, event-driven computation, continual learning, interpretability

1. Introduction

Over the past decade, large-scale pre-trained models have rapidly ascended to prominence within the field of artificial intelligence (AI), demonstrating unprecedented capacities in natural language processing, computer vision, and multimodal integration [1–6]. Such models leverage extensive corpora of textual, visual, and other data modalities, enabling them to achieve state-of-the-art performance on a broad range of benchmark tasks. Despite these advances, the increasing scale and complexity of these architectures have precipitated several notable challenges. Chief among these are excessive computational and energy costs, limited adaptability to novel domains, difficulties in learning continuously over time, and insufficient transparency in their decision-making processes [7–11].

By comparison, the human brain exemplifies a remarkably efficient and adaptable information processing system. Operating within stringent energy constraints, the brain excels at rapidly integrating multiple sensory modalities, adapting to dynamic environmental demands, learning effectively from limited exposure, and retaining accumulated knowledge over a lifetime without catastrophic forgetting. Neuroscientific research has begun to elucidate the underlying principles that support these capabilities, encompassing hierarchical and modular cortical organization, selective attention mechanisms, intricate memory consolidation processes, and synaptic plasticity that continually reshapes neural circuits as new information is acquired [12–17]. Together, these mechanisms facilitate a balance between efficiency, adaptability, and interpretability that is often challenging to achieve in large-scale machine learning models.

Motivated by these insights, a growing body of work seeks to incorporate biologically informed principles into the design and training of large-scale models. Such endeavors are grounded in the proposition that harnessing brain-inspired organizational structures and learning rules may yield more robust, flexible, and

***Corresponding author:** Zongwei Luo, Institute of Artificial Intelligence and Future Networks, Beijing Normal University at Zhuhai and Guangdong Provincial Key Laboratory of IRADS, China. Email: lzwqhk@bnu.edu.cn

resource-efficient computational systems [18, 19]. Notable rese arch directions include the development of modular and hierarchical model architectures inspired by cortical organization; the integration of attention mechanisms that dynamically allocate computational resources based on task-relevant features, echoing the brain's selective attention processes [20–22]; and the adoption of memory and plasticity principles that permit continuous, incremental learning without extensive retraining [23, 24]. Moreover, the exceptional metabolic efficiency of the brain has stimulated research into model compression, pruning, and other strategies aimed at reducing computational footprints without sacrificing model quality [25–27].

Notwithstanding these promising avenues, the process of systematically integrating brain-inspired mechanisms into large-scale models remains far from straightforward. Fundamental questions persist regarding how best to implement hierarchical structures that can scale to complex tasks, how to align neural attentional dynamics with computational attention heads, how to achieve stable memory retention in continuously evolving environments, and how to ensure that such modifications yield empirically demonstrable improvements in interpretability, efficiency, and robustness [28–32]. The inherently interdisciplinary nature of these challenges—spanning machine learning, neuroscience, cognitive science, and computational modeling—underscores the necessity of collaborative efforts that draw upon a broad spectrum of empirical and theoretical perspectives.

This review brings together current knowledge from neuroscience and from large-scale pre-trained and foundation models and primarily covers literature published from 2019 to 2025 while including earlier seminal works when necessary for context. It distills core biological principles—how the brain organizes information, focuses attention, stores memories, and adjusts its connections—which underwrite efficiency and adaptability, and it maps these principles to architectural, training, and optimization strategies for large models. Methodologically, we identified studies through keyword-based searches and citation tracing, prioritized peer-reviewed articles and influential preprints with methodological or empirical contributions aligned to our scope, excluded nontechnical editorials and out-of-scope applications, and organized the corpus along three analytic axes—adaptability, efficiency, and interpretability—that guide the structure and synthesis of the review.

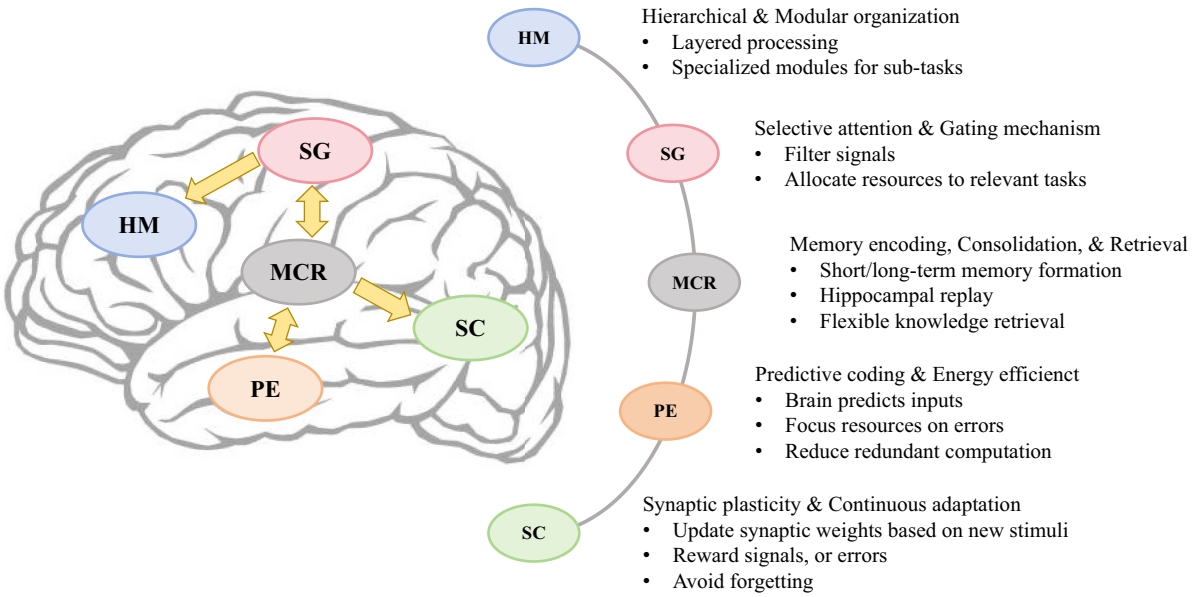
2. Background and Preliminary

2.1. Cognitive mechanisms in the human brain

The human brain stands as a model of remarkable efficiency and adaptability, supporting complex cognitive functions and learning processes under stringent energetic constraints. Central to its success is a suite of interrelated mechanisms—hierarchically organized processing, modular specialization, selective attentional control, robust memory formation, and synaptic plasticity—that collectively underpin flexible, context-aware information processing and long-term knowledge retention [33, 34]. Understanding these foundational principles not only advances our comprehension of human cognition but also provides valuable insights into how we might guide the development of more efficient and adaptable large-scale computational models.

As illustrated in Figure 1, five core mechanisms—hierarchical and modular organization, selective attention, memory, synaptic plasticity, and predictive coding—may work in concert to achieve robust and efficient cognition. The conceptual diagram highlights these interdependencies, with arrows indicating how each mechanism both shapes and is shaped by the others, underscoring the dynamic and bidirectional nature of cognitive processing. It is important to note that this diagram illustrates possible points of interaction rather than claiming a literal, exhaustive neural mapping. For example, selective attention may amplify relevant signals within hierarchical modules, while predictive coding processes can dampen predictable inputs to conserve resources. Memory consolidation and synaptic plasticity further adapt these pathways over

Figure 1
Five key cognitive mechanisms of the human brain



time, gradually integrating new information into existing knowledge structures. The following subsections delve into each mechanism, explaining its role and relevance to large-scale models.

2.1.1. Hierarchical and modular organization

Neural processing in the brain is structured in a hierarchical, modular fashion. Sensory information often proceeds from lower-level cortical areas, which extract basic stimulus features, toward higher-level associative regions that integrate and interpret these features in more abstract terms [33, 35, 36]. This layered architecture supports both bottom-up and top-down processing, enabling the brain to integrate prior knowledge, contextual information, and attentional signals to refine perception and decision-making [34, 37]. Such hierarchical networks are organized into functionally specialized modules—discrete neuronal populations that are selectively tuned to particular categories of stimuli or cognitive tasks [38]. By conferring specialization and flexible recombination, modularity and hierarchy support generalization and transfer of learned skills across domains—qualities that large-scale models often strive to emulate.

2.1.2. Selective attention and gating mechanisms

Human cognition is supported by attentional systems that highlight salient information while suppressing irrelevant inputs. This selective gating mechanism allows the brain to efficiently allocate its limited computational and energetic resources to the most behaviorally pertinent signals [39, 40]. Attentional control arises through distributed yet interconnected networks involving the frontoparietal cortex and subcortical structures, dynamically adjusting the gain on sensory streams and internal representations [41, 42]. By doing so, the brain can flexibly switch between tasks, adapt to environmental changes, and maintain stable performance under varying conditions. These principles inform computational attention strategies that large-scale models have begun to adopt, seeking more context-sensitive resource allocation and improved interpretability of model outputs.

2.1.3. Memory encoding, consolidation, and retrieval

Memory formation in the brain is distinguished by its capacity to encode, store, and retrieve information over diverse timescales. Short-term and working memory rely on sustained neural activity patterns, whereas long-term memory depends on synaptic modifications that underlie stable information storage [43, 44]. Critical to this process is the consolidation of memories, wherein initially labile traces, often formed in the hippocampus, are gradually integrated into distributed cortical networks to achieve long-lasting representation [45]. These consolidation dynamics are supported by offline reactivation events such as hippocampal replay, which helps strengthen synaptic connections and update cortical storage [46, 47]. In parallel, associative and semantic memory structures enable flexible retrieval of stored information, permitting the recombination of knowledge to support inference, problem-solving, and generalization [48, 49]. Such biological memory systems offer blueprints for constructing large-scale models capable of integrating new information continuously without catastrophic forgetting and performing efficient knowledge retrieval.

2.1.4. Synaptic plasticity and continuous adaptation

Synaptic plasticity underlies the brain's capacity to update internal representations in response to novel stimuli, shifting task demands, or altered environmental contingencies. Molecular and cellular mechanisms such as long-term potentiation and long-term depression, combined with neuromodulatory influences (e.g., dopamine, acetylcholine), serve to fine-tune synaptic weights according to reward signals, prediction errors, and contextual relevance [50–52]. This distributed, continuous adaptation process ensures that new information can be incorporated without erasing previously acquired skills. Emulating similar plasticity-inspired update rules in large-scale models may improve their capacity for online learning and adaptation, allowing these systems to maintain stable performance across shifting data distributions and extended temporal horizons.

2.1.5. Predictive coding and energy efficiency

Despite its extensive computational capabilities, the brain operates with constrained energetic resources. Evidence suggests that the cortex implements predictive coding principles whereby higher-order areas generate probabilistic predictions of forthcoming sensory input, and only the unpredicted components are further processed [53–55]. Such predictive schemes efficiently allocate processing capacity to deviations from expectation, thereby minimizing redundant computations and conserving energy. Integrating predictive coding frameworks into large-scale models may improve their ability to handle variable input quality, reduce computational overhead, and enhance their overall adaptability to uncertain conditions.

2.2. Large-scale pre-trained models: capabilities and limitations

Large-scale, pre-trained models have facilitated impressive advancements in natural language processing, computer vision, and multimodal tasks. Such models—trained on extensive corpora and comprising billions of parameters—demonstrate remarkable capabilities, ranging from fluent, context-sensitive text generation to coherent multimodal reasoning [3–5, 56–59]. Their widespread deployment across research and industry underscores their appeal as general-purpose learners capable of few-shot adaptation, multilingual translation, and even code synthesis [60]. However, despite these successes, fundamental challenges impede their broader, more sustainable integration into complex, dynamic real-world environments.

Foremost among these challenges is the escalating computational and energetic cost associated with training and deploying increasingly large architectures. The resources required to handle vast datasets and train models with unprecedented parameter counts have raised concerns regarding environmental impact, carbon emissions, and the long-term economic viability of these approaches [11, 61–63]. Unlike the human brain's low-power neural computations, current artificial systems often rely on brute-force optimization, extensive hyperparameter tuning, and massive parallelization rather than exploiting biologically inspired trade-offs that balance efficiency with representational complexity.

Another persistent limitation pertains to adaptability and continual learning. Although foundation models can be fine-tuned for new tasks, they often struggle to maintain previously acquired competencies without retraining on past data, a phenomenon known as catastrophic forgetting. Overwriting of earlier representations and poor accommodation of novel data distributions remain significant obstacles to flexible, long-term knowledge accumulation [64–66].

Recent surveys and analyses further systematize strategies for mitigating catastrophic forgetting and enabling continual adaptation across tasks and settings [64, 67, 68]. In contrast, human cognition seamlessly integrates new information into existing schemas while retaining past skills, suggesting that insights from synaptic plasticity and hippocampal-cortical consolidation processes could inform more resilient approaches to model updating.

Challenges in robustness and generalization further limit the applicability of large-scale models. Despite ingesting immense training corpora, these systems frequently exhibit brittle behavior in the face of distributional shifts, adversarial perturbations, and previously unseen modalities [69–71]. Rather than extracting stable, abstract concepts, they often rely on superficial statistical regularities that fail to transfer to unfamiliar contexts. The brain, in contrast, routinely copes with sensory noise, environmental volatility, and complex social cues, adjusting its internal representations through attention, prediction, and memory consolidation. Emulating these adaptive strategies in large-scale models may yield more robust performance under real-world conditions.

Opacity and interpretability pose yet another major limitation. Although large models often develop rich latent representations and can capture intricate patterns, these representations are difficult to explain or align with human-understandable concepts. Efforts to uncover latent knowledge or identify hidden biases reveal that many models harbor problematic associations and knowledge structures that arise without explicit guidance [72–75]. The resulting lack of transparency complicates attempts to ensure fairness, reliability, and accountability, and it impedes the creation of models that can provide justifications for their decisions. In contrast, the hierarchical and modular organization of the human brain—coupled with decades of neuroscientific research—offers at least partial insights into how neural circuits implement representational transformations. Drawing inspiration from these biological principles may facilitate the design of architectures with more interpretable internal dynamics.

Following the lifecycle perspective illustrated in Figure 2, we can identify the major limitations of large-scale pre-trained models. The diagram conceptualizes the entire process, from model training and fine-tuning to deployment and continuous learning, highlighting that each stage presents distinct difficulties. As the figure suggests, these challenges—such as high compute costs, brittle generalization, and catastrophic forgetting—correspond to the remarkable strengths of the human brain. These biological strengths can, in turn, be leveraged to inspire features that overcome the difficulties faced by artificial systems. The following section explores how integrating these neuroscientific principles can inform the design of more efficient, adaptable, and interpretable large-scale models.

3. Adaptive, Efficient, and Interpretable Optimization of Large-Scale Models Through Brain-Inspired Principles

We apply the methodological framing stated at the end of the Introduction: studies identified through keyword-based searches and citation tracing are organized along the three analytic dimensions of adaptability, efficiency, and interpretability. Evidence is synthesized directionally rather than through cross-paper numerical aggregation; where feasible, interpretations are made under shared or normalized compute and energy budgets. The following subsections map neurobiological and cognitive mechanisms to computable architectures and training strategies.

Overcoming the core limitations of large-scale models—ranging from their prohibitively high computational demands and inflexibility to their fragile generalization and limited interpretability—may benefit from the systematic incorporation of neuroscientific principles. The human brain excels at adapting to novel conditions, efficiently coding information, and operating under severe metabolic constraints. Consequently, research at the intersection of neuroscience and machine learning has increasingly aimed to distill the mechanistic underpinnings of cortical processing to inform the next generation of computational architectures [12, 15, 19]. Realizing these bio-inspired improvements requires translating high-level biological concepts into scalable, efficient,

Figure 2
The large model lifecycle: challenges and brain-inspired solutions

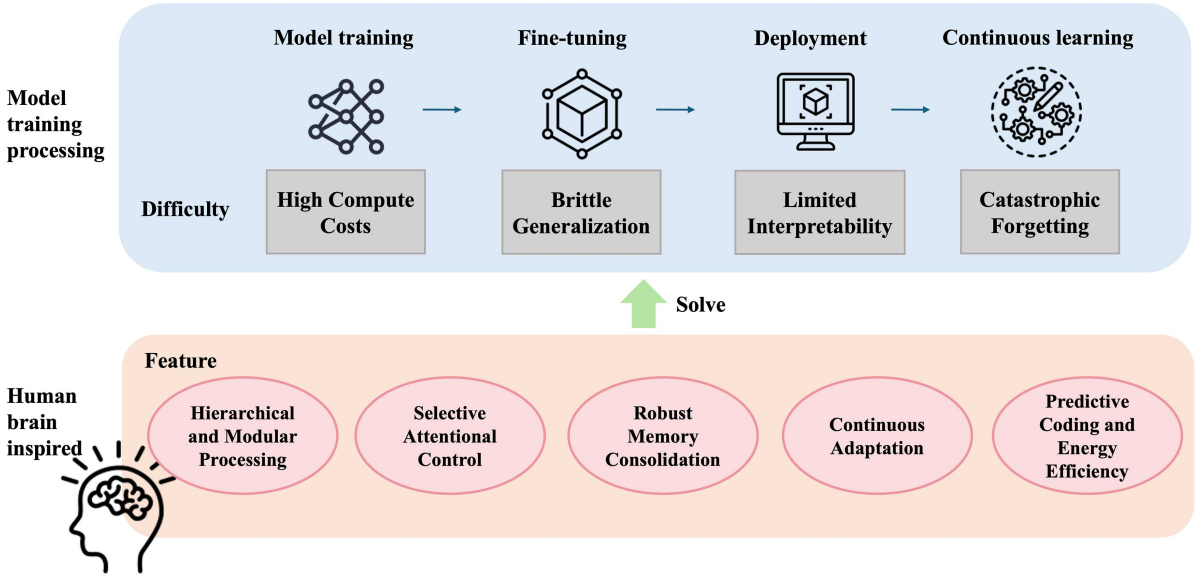
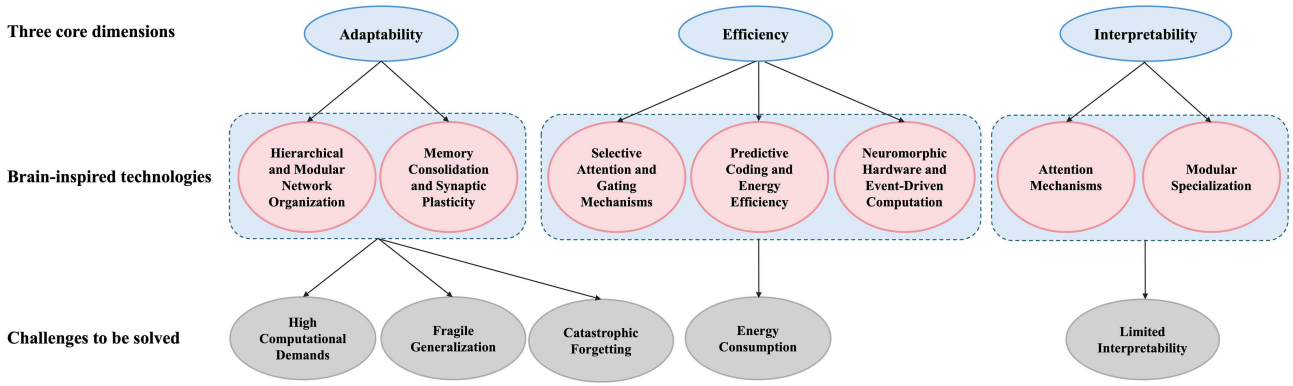


Figure 3
Overview of major brain-inspired approaches for large-scale models categorized into adaptability, efficiency, and interpretability strategies



and robust computational frameworks. Before delving into specific strategies, Figure 3 provides an overview of how these brain-inspired approaches can be categorized along the three primary dimensions of adaptability, efficiency, and interpretability.

We further organize brain-inspired approaches along a two dimensional taxonomy that distinguishes physical structure inspired methods including neuronal and synaptic dynamics, spiking and neuromorphic or predictive coding, from human behavior inspired methods including attention and gating, memory consolidation, transfer and meta learning, and reinforcement or imitation learning, and we map mechanisms in both streams to computable architectures and training strategies that are analyzed along adaptability, efficiency, and interpretability.

Table 1 consolidates representative brain-inspired strategies with their implementations and assesses adaptability, efficiency, and interpretability as directional evidence rather than absolute numbers. The table operationalizes our mechanism-to-algorithm mapping by linking neurobiological and cognitive mechanisms to computable architectures and training strategies, including sparsely gated mixtures of experts, retrieval-augmented generation (RAG) and parameter isolation, attention and network-level gating, predictive-coding objectives, and spiking or event-driven platforms. Given heterogeneous tasks and reporting protocols, interpretations are made under shared or normalized compute and energy budgets whenever feasible. The table serves as a compact design aid, indicating robust gains, key trade-offs, and areas where evidence remains inconclusive, and it anchors the subsection-level analyses that follow.

3.1. Adaptability

Adaptability in large-scale models is crucial for handling dynamic environments and is a primary focus of brain-inspired optimization. A central strategy for enhancing adaptability lies in emulating the brain's hierarchical and modular network organization. This biological principle has been computationally realized most prominently through Sparsely Gated Mixture-of-Experts (MoE) architectures [76]. The MoE model formalizes this concept as:

$$\text{Output}\,(x) = \sum_{i=1}^{N} g_i\,(x) \cdot E_i\,(x) \tag{1}$$

where a trainable gating network $g(x)$ dynamically routes each input x to a small subset of specialized "expert" networks $E_i(x)$. This input-dependent routing achieves computational sparsity and has demonstrated substantial efficiency gains. For instance, the Switch Transformer reported a pre-training speedup of up to 7× over an equivalently performing dense model [77]. However, this approach introduces a trade-off between computational efficiency and memory, as all expert parameters must be stored, and it presents challenges in maintaining training stability and load balancing across experts [76, 77].

Another critical aspect of adaptability is inspired by the brain's mechanisms for memory consolidation and synaptic plasticity, which mitigate the catastrophic forgetting common in artificial networks [64]. The brain integrates new information without overwriting prior knowledge through processes like hippocampal replay and neuromodulated synaptic adjustments [51, 78]. Computationally, this has inspired approaches that aim for stable, continuous learning. For example, some models incorporate differentiable plasticity in recurrent networks to adapt to new tasks without extensive retraining [79]. A more direct analogue to memory retrieval is seen in modern RAG models, which externalize knowledge in a nonparametric store and retrieve relevant information at inference time, thereby separating knowledge storage from model parameter updates and enhancing adaptability to new information [80]. The biological synaptic plasticity update rule, often abstracted as:

$$\Delta W_{ij} = \eta \cdot \phi(x_i) \cdot \psi(y_j) \tag{2}$$

This rule underscores the principle of local, activity-dependent weight adjustments. While powerful, implementing such fine-grained plasticity at the scale of large models remains a significant challenge, often simplified to more coarse-grained techniques like selective fine-tuning or parameter isolation.

3.2. Efficiency

Improving computational and energy efficiency is another paramount goal. Selective attention and gating mechanisms in the brain, which allocate processing resources to the most salient information, offer a powerful blueprint [41, 81]. While the self-attention mechanism in Transformers is brain-inspired, it is a highly simplified version. Its formulation is given by:

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V \tag{3}$$

where query (Q), key (K), and value (V) matrices interact to compute context-aware representations [21]. More biologically

grounded gating mechanisms, however, seek to dynamically allocate computational resources at a network level. For example, some approaches use recurrent gating to activate only relevant model parameters during inference, reducing computational overhead without sacrificing performance [82].

The brain's reliance on predictive coding provides another compelling framework for energy-efficient computation. Neuroscientific theories propose that cortical circuits continuously generate predictions about sensory inputs and primarily process the "prediction error"—the discrepancy between expectation and reality [37, 54]. This principle minimizes redundant computations on predictable signals. In machine learning, this is often implemented by training models to minimize a prediction error loss function:

$$\mathscr{L} = |x - \hat{x}|^2 \tag{4}$$

where the model learns to generate a prediction $\hat{x}$ of the actual input x. Beyond a simple loss, a true predictive coding framework implies an internal generative model that actively suppresses predictable information flow—a feature explored in models for video prediction and unsupervised representation learning [83, 84]. On the hardware front, neuromorphic systems that utilize spike-based, event-driven computation directly mimic the brain's sparse signaling, offering a path to radically lower power consumption, though their integration with large-scale deep learning remains an active research area [27, 85]. Representative platforms and surveys illustrate this trajectory, including large-scale neuromorphic systems and recent hardware–software advances for spike-based learning [20, 27]. Deployment involves inevitable trade-offs among accuracy, interpretability, and efficiency, often under constraints on compute, latency, privacy, and robustness. In practice, training typically targets GPU or TPU clusters with memory bandwidth and interconnects as primary bottlenecks, whereas inference must meet latency and energy budgets on edge or data center systems. Event -and neuromorphic platforms offer energy advantages for sparse spiking workloads, while large language and vision models benefit from quantization and memory-aware scheduling in conventional accelerators.

3.3. Interpretability

Enhancing the interpretability of opaque large-scale models is a critical challenge for ensuring their reliability and trustworthiness.

Table 1
Brain-inspired strategies mapped to representative implementations and evaluated along adaptability, efficiency, and interpretability

Strategy	Representative implementation	Adaptability	Efficiency	Interpretability
Hierarchical–modular routing [76, 77]	Sparsely Gated Mixture-of-Experts; Switch Transformer	Positive. Improves task switching and continuous adaptation via input-dependent expert routing.	Positive. Substantial pre-training speedups reported, with memory and load-balancing trade-offs.	Mixed. Functional specialization aids post hoc attribution, but global transparency remains limited.
Memory consolidation and synaptic plasticity, including externalized memory [79, 80]	Differentiable plasticity; retrieval-augmented generation; parameter isolation or selective fine-tuning	Positive. Mitigates catastrophic forgetting and supports continuous integration of new information.	Mixed. Efficiency is implementation-dependent; this review does not quantify compute or energy impacts.	Mixed. External knowledge stores improve traceability, while internal weight changes remain difficult to explain.
Attention and gating [21, 82]	Self-attention; network-level dynamic gating and on-demand parameter activation	Mixed. Benefits are context-dependent across tasks and distributions.	Positive. Activating only relevant parameters reduces unnecessary computation without degrading performance.	Positive. Attention weights provide limited but useful transparency.
Predictive coding [54, 83]	Prediction-error-minimization training; unsupervised and video prediction models	Positive. Often improves robustness to uncertainty and variable input quality.	Positive. Reduces redundant processing and supports energy-oriented computation.	Inconclusive. Explanatory pathways remain under development.
Spiking and neuromorphic, event-driven [20, 27]	Spiking neural networks; event-driven or neuromorphic hardware platforms	Mixed. Integration with large-scale deep learning remains an active research area.	Positive. Event-driven sparsity offers significant power advantages in suitable workloads.	Inconclusive. Varies with platform and methodology.

Brain-inspired principles offer two main avenues for improvement. First, the attention mechanism itself, as formulated in Equation (3), provides a degree of transparency. By visualizing the attention weights, researchers can infer which parts of the input data the model deemed most influential in its decision-making process, revealing the model's focus and potential biases [21]. Second, the principle of modular specialization, as implemented in MoE architectures (Equation (1)), aids interpretability by design. Segmenting a network into specialized modules allows for clearer attribution of function. One can analyze the behavior of individual experts to understand their specific roles or identify which experts are activated for particular types of inputs, thereby decomposing the model's complex decision-making process into more understandable sub-problems [76, 86]. While neither approach achieves full transparency, both represent significant steps toward creating models whose internal workings are more aligned with human-understandable concepts.

3.4. Summary

In conclusion, the integration of brain-inspired principles into large-scale models presents a promising path toward addressing their fundamental limitations. These approaches can be broadly categorized by their primary goals. To enhance adaptability, principles of hierarchical modularity and memory consolidation are being explored to create models that can learn continually and handle dynamic tasks, moving beyond the static "train-once" paradigm. To improve efficiency, strategies like selective attention, predictive coding, and event-driven neuromorphic hardware aim to significantly reduce the immense computational and energy costs associated with large models by minimizing redundant information processing. Finally, to increase interpretability, the inherent structure of attention mechanisms and modular specialization provides a clearer view into the model's decision-making process, fostering greater trust and accountability. While most efforts remain in an exploratory phase, these converging lines of research suggest a future where AI systems are not only powerful but also more flexible, efficient, and transparent. Beyond description, we synthesize several tensions that shape current research: fidelity of brain-inspired mechanisms versus engineering tractability, continual adaptation versus stability, energy efficiency versus task generality, and post hoc interpretability versus decision faithfulness. These tensions delineate where mechanisms transfer effectively and where they fall short, motivating evaluation under shared budgets and comparable conditions.

Illustrative implementation examples help ground the synthesis. In streaming fraud detection, event-driven features and budgeted inference on edge or near-edge systems enable low-latency decisions under strict energy and privacy constraints. In credit risk assessment, compact surrogate models paired with auditable reasoning traces support regulatory review while preserving task performance. In grid and building energy forecasting, deployment on resource-constrained devices requires memory-aware scheduling and quantization, linking efficiency gains to real-world reliability [27, 85].

4. Future Directions and Emerging Frontiers

The integration of brain-inspired principles into large-scale models, while promising, is still in a nascent stage. The analyses presented in the previous section reveal not only the potential of these approaches but also their limitations, thereby charting a clear course for future research. Advancing the field will require addressing specific theoretical gaps, engineering challenges, and evaluation paradigms that have emerged from these early explorations.

4.1. Theoretical gaps and foundational obstacles

A primary obstacle is the lack of rigorous theoretical frameworks that can translate biological mechanisms into scalable, mathematically principled algorithms. While empirical successes exist, a deeper foundation is needed. For instance, as highlighted by the load-balancing challenges in MoE models (Section 3.1), a key theoretical gap is the development of more principled, less heuristic routing algorithms inspired by theories of neural resource allocation and competition [15, 19]. Similarly, the critique of predictive coding's implementation via a simple loss function (Section 3.2) points to a pressing need for theoretical work on how internal generative models can be efficiently learned and updated within large-scale architectures, moving beyond simple error minimization [87, 88]. Establishing such theories is essential to guide the next generation of model design systematically.

4.2. Balancing performance, efficiency, and hardware realities

The trade-offs identified in current brain-inspired models present direct challenges for hardware and systems design. For example, the tension between computational efficiency and parameter inefficiency in MoE architectures (Section 3.1) necessitates a co-design of future hardware. Systems will need to accommodate models with massive, sparsely accessed parameter stores, demanding innovations in memory bandwidth and dataflow management that differ from today's GPU architectures optimized for dense matrix operations [89, 90]. Furthermore, to fully realize the energy-saving potential of event-driven computation and predictive coding (Section 3.2), new compiler infrastructures and runtime environments are required that can exploit asynchronous, sparse processing—a paradigm for which current deep learning stacks are ill-suited [27, 85].

4.3. Specialized and multidimensional benchmarks

The limitations of current evaluation methods call for the development of new, specialized benchmarks. Standard datasets like ImageNet or GLUE are insufficient for measuring the core advantages promised by brain-inspired models [91, 92]. As discussed in the context of memory consolidation and synaptic plasticity (Section 3.1), there is a pressing need for robust benchmarks that can rigorously quantify catastrophic forgetting in continual learning scenarios [68, 93]. Moreover, the analysis of MoE models (Section 3.1) reveals that new evaluation metrics are required to assess not only task accuracy but also the efficiency of dynamic resource allocation, such as routing quality and expert utilization under diverse data distributions. Such benchmarks are critical for driving the iterative refinement of these complex architectures.

4.4. Complexities of multimodal and multitask integration

The brain's ability to seamlessly fuse information from diverse sensory modalities remains a formidable challenge for AI [94, 95]. However, the modular architectures discussed in Section 3.1 offer a promising substrate for tackling this problem. A key future direction is to explore how specialized expert modules can be trained on distinct modalities (e.g., vision, audio), while a shared or

hierarchical gating mechanism learns to effectively route and fuse their representations in a context-dependent manner. This approach could provide a more principled way to manage the inductive biases of different data types, moving beyond the current trend of forcing all modalities through a single, homogeneous architectural backbone.

4.5. Ethical and societal considerations

As AI models incorporate more brain-like features, new ethical and societal questions emerge. The brain-inspired methods for interpretability discussed in Section 3.3, such as visualizing attention weights or analyzing modular function, offer a path toward more transparent AI [96, 97]. However, this very transparency can introduce new risks. For instance, if a model's attentional focus can be clearly understood, could it also be more easily manipulated by adversarial attacks? If functions are localized to specific modules, could these modules develop concentrated, hard-to-detect biases? These questions demand proactive, interdisciplinary dialogue between AI researchers, neuroscientists, and ethicists to ensure that the pursuit of humanlike AI aligns with societal values [98, 99].

4.6. The critical role of collaborative endeavors

Ultimately, bridging the gap between biological intelligence and large-scale models is not a task for any single discipline. The specific challenges and opportunities identified throughout this review—from developing new routing theories for modular networks to designing hardware for sparse computation—underscore the need for deep, sustained collaboration. Machine learning researchers, neuroscientists, cognitive scientists, and hardware engineers must align priorities to systematically test and scale the most promising neuro-inspired principles. Forging such cohesive communities of practice is essential to ensure that the next generation of AI not only achieves higher performance but also embodies the efficiency, adaptability, and interpretability exemplified by the human brain [13, 16, 19].

5. Conclusion

The challenges confronted by contemporary large-scale models—ranging from their massive computational demands and rigid adaptation to their fragile generalization and limited interpretability—underscore the need for more sophisticated paradigms. In parallel, the human brain's remarkable ability to efficiently process diverse sensory inputs, adapt to changing conditions, consolidate memories over extended periods, and maintain robust performance under stringent energy constraints offers a compelling blueprint for improving these models. Recent interdisciplinary efforts have begun to incorporate brain-inspired principles into large-scale model optimization, drawing inspiration from hierarchical and modular architectures, attentional mechanisms, memory consolidation strategies, synaptic-like plasticity rules, and predictive coding frameworks.

Although early empirical studies have yielded promising initial results, these attempts remain at a formative stage. Significant theoretical gaps persist, and the widespread adoption of such neuro-inspired techniques faces engineering hurdles, including hardware–software integration and the development of meaningful, multidimensional benchmarks. Moreover, extending these ideas to multimodal environments, addressing real-time adaptability, and ensuring that emerging methods align with ethical and social values represent substantive, ongoing challenges. The mapping from neural principles to engineering abstractions simplifies biological complexity and may vary with task, data, and hardware; we therefore treat the three-dimensional framework of adaptability, efficiency, and interpretability as a heuristic rather than a one-to-one correspondence. Ethical and regulatory considerations increasingly require transparency, auditability, and fairness in high-stakes settings. We therefore treat interpretability not only as a scientific goal but as a compliance and governance requirement that shapes model selection and deployment.

Looking ahead, progress will depend on deepening collaborations across disciplinary boundaries. Neuroscientists, cognitive scientists, machine learning researchers, hardware engineers, ethicists, and policymakers all have a role to play in shaping how brain-inspired methods evolve and integrate into next-generation models. Robust theoretical foundations, sophisticated measurement tools, and carefully curated benchmarks will guide the refinement of these bio-inspired approaches, ensuring that models become not only capable but also versatile, energy-efficient, and interpretable. Promising directions include (i) standardized protocols for evaluating adaptivity, efficiency, and interpretability under shared budgets, (ii) co-design of representations and hardware that preserve event sparsity without sacrificing task generality, and (iii) causal analyses that disentangle when brain-inspired priors help or hinder foundation model updates.

Ultimately, although the path forward is complex, the potential rewards are substantial. If researchers leverage the brain's fundamental computational strategies to enhance large-scale models, the field may move closer to developing intelligent systems that genuinely approach human-like adaptability and resourcefulness. Such advancements could catalyze the emergence of sustainable, context-aware, and trustworthy AI, marking a significant milestone in bridging the gap between biological and artificial intelligence.

Funding Support

This work was supported by the MOE AI Project on Hikvision and Beijing Normal University at Zhuhai, MOE AI Project on Hikvision and Beijing Normal-Hong Kong Baptist University, the Beijing Normal University Zhuhai Teaching Reform Project – Online and Offline Course on Artificial Intelligence and Ethics, and Guangdong Provincial Key Laboratory of Interdisciplinary Research and Application for Data Science.

Ethical Statement

This study does not contain any studies with human or animal subjects performed by any of the authors.

Conflicts of Interest

Zongwei Luo is the Editor-in-Chief of *FinTech and Sustainable Innovation* and was not involved in the editorial review or the decision to publish this article. The authors declare that they have no conflicts of interest to this work.

Data Availability Statement

Data sharing is not applicable to this article as no new data were created or analyzed in this study.

Author Contribution Statement

Jingzhe Wang: Conceptualization, Methodology, Software, Validation, Formal analysis, Investigation, Data curation, Writing – original draft, Writing – review & editing, Visualization. **Yajing Wang:** Validation, Formal analysis, Investigation, Data curation, Writing – original draft, Writing – review & editing, Visualization. **Zongwei Luo:** Conceptualization, Methodology, Resources, Writing – review & editing, Supervision, Project administration, Funding acquisition.

References

[1] Paaß, G., & Giesselbach, S. (2023). *Foundation models for natural language processing: Pre-trained language models integrating media*. Switzerland: Springer International Publishing. https://doi.org/10.1007/978-3-031-23190-2

[2] Li, C., Gan, Z., Yang, Z., Yang, J., Li, L., Wang, L., & Gao, J. (2024). Multimodal foundation models: From specialists to general-purpose assistants. *Foundations and Trends® in Computer Graphics and Vision*, *16*(1-2), 1–214. https://doi.org/10.1561/0600000110

[3] Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., ..., & Amodei, D. (2020). Language models are few-shot learners. In *Proceedings of the 34th International Conference on Neural Information Processing Systems, 159,* 1877–1901. https://dl.acm.org/doi/abs/10.5555/3495724.3495883

[4] Raffel, C., Shazeer, N., Roberts, A., Lee, K., Narang, S., Matena, M., ..., & Liu, P. J. (2020). Exploring the limits of transfer learning with a unified text-to-text transformer. *The Journal of Machine Learning Research*, *21*(1), 5485–5551. https://dl.acm.org/doi/abs/10.5555/3455716.3455856

[5] Xi, Z., Chen, W., Guo, X., He, W., Ding, Y., Hong, B., ..., & Gui, T. (2025). The rise and potential of large language model based agents: A survey. *Science China Information Sciences*, *68*(2), 121101. https://doi.org/10.1007/s11432-024-4222-0

[6] Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., ..., & Houlsby, N. (2021). An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations*, 1–22.

[7] Violos, J., Mamanis, G., Kompatsiaris, I., & Papadopoulos, S. (2025). Cognition and context-aware decision-making systems for a sustainable planet: A survey on recent advancements, applications and open challenges. *Discover Sustainability*, *6*(1), 235. https://doi.org/10.1007/s43621-025-00954-y

[8] Strubell, E., Ganesh, A., & McCallum, A. (2020). Energy and policy considerations for modern deep learning research. *Proceedings of the AAAI Conference on Artificial Intelligence, 34* (09), 13693–13696. https://doi.org/10.1609/aaai.v34i09.7123

[9] Chiang, C.-W., Lu, Z., Li, Z., & Yin, M. (2024). Enhancing AI-assisted group decision making through LLM-powered devil's advocate. In *Proceedings of the 29th International Conference on Intelligent User Interfaces*, 103–119. https://doi.org/10.1145/3640543.3645199

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

[11] Luccioni, A. S., Viguier, S., & Ligozat, A.-L. (2023). Estimating the carbon footprint of bloom, a 176B parameter language model. *Journal of Machine Learning Research*, *24*(253), 1–15.

[12] Hassabis, D., Kumaran, D., Summerfield, C., & Botvinick, M. (2017). Neuroscience-inspired artificial intelligence. *Neuron*, *95*(2), 245–258. https://doi.org/10.1016/j.neuron.2017.06.011

[13] Kriegeskorte, N., & Douglas, P. K. (2018). Cognitive computational neuroscience. *Nature Neuroscience*, *21*(9), 1148–1160. https://doi.org/10.1038/s41593-018-0210-5

[14] Pezzulo, G., Rigoli, F., & Friston, K. (2015). Active inference, homeostatic regulation and adaptive behavioural control. *Progress in Neurobiology*, *134*, 17–35. https://doi.org/10.1016/j.pneurobio.2015.09.001

[15] Goldberg, H. (2023). Unraveling neurodiversity: Insights from neuroscientific perspectives. *Encyclopedia*, *3*(3), 972–980. https://doi.org/10.3390/encyclopedia3030070

[16] Marblestone, A. H., Wayne, G., & Kording, K. P. (2016). Toward an integration of deep learning and neuroscience. *Frontiers in Computational Neuroscience*, *10*, 94. https://doi.org/10.3389/fncom.2016.00094

[17] Zador, A. M. (2019). A critique of pure learning and what artificial neural networks can learn from animal brains. *Nature Communications*, *10*(1), 3770. https://doi.org/10.1038/s41467-019-11786-6

[18] Cichy, R. M., & Kaiser, D. (2019). Deep neural networks as scientific models. *Trends in Cognitive Sciences*, *23*(4), 305–317. https://doi.org/10.1016/j.tics.2019.01.009

[19] Lillicrap, T. P., Santoro, A., Marris, L., Akerman, C. J., & Hinton, G. (2020). Backpropagation and the brain. *Nature Reviews Neuroscience*, *21*(6), 335–346. https://doi.org/10.1038/s41583-020-0277-3

[20] Fang, W., Chen, Y., Ding, J., Yu, Z., Masquelier, T., Chen, D., ..., & Tian, Y. (2023). SpikingJelly: An open-source machine learning infrastructure platform for spike-based intelligence. *Science Advances*, *9*(40), eadi1480. https://doi.org/10.1126/sciadv.adi1480

[21] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., ..., & Polosukhin, I. (2017). Attention is all you need. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, 6000–6010.

[22] Ororbia, A., Mali, A., Kohan, A., Millidge, B., & Salvatori, T. (2024). *A review of neuroscience-inspired machine learning*. arXiv. https://doi.org/10.48550/arXiv.2403.18929

[23] Graves, A., Wayne, G., & Danihelka, I. (2014). *Neural turing machines*.arXiv. https://doi.org/10.48550/arXiv.1410.5401

[24] Lu, Q., Hasson, U., & Norman, K. A. (2022). A neural network model of when to retrieve and encode episodic memories. *eLife*, *11*, e74445. https://doi.org/10.7554/eLife.74445

[25] LeCun, Y., Denker, J. S., & Solla, S. A. (1989). Optimal brain damage. In *Proceedings of the 3rd International Conference on Neural Information Processing Systems*, 598–605.

[26] Han, S., Pool, J., Tran, J., & Dally, W. J. (2015). Learning both weights and connections for efficient neural network. In *Proceedings of the 29th International Conference on Neural Information Processing Systems, 1,* 1135–1143. https://dl.acm.org/doi/10.5555/2969239.2969366

[27] Davies, M., Wild, A., Orchard, G., Sandamirskaya, Y., Guerra, G. A. F., Joshi, P., ..., & Risbud, S. R. (2021). Advancing neuromorphic computing with Loihi: A survey of results and outlook. *Proceedings of the IEEE*, *109*(5), 911–934. https://doi.org/10.1109/JPROC.2021.3067593

[28] Eliasmith, C., Stewart, T. C., Choo, X., Bekolay, T., DeWolf, T., Tang, Y., & Rasmussen, D. (2012). A large-scale model of the functioning brain. *Science*, *338*(6111), 1202–1205. https://doi.org/10.1126/science.1225266

[29] Urai, A. E., Doiron, B., Leifer, A. M., & Churchland, A. K. (2022). Large-scale neural recordings call for new insights to link brain and behavior. *Nature Neuroscience*, *25*(1), 11–19. https://doi.org/10.1038/s41593-021-00980-9

[30] Bellmund, J. L. S., Gärdenfors, P., Moser, E. I., & Doeller, C. F. (2018). Navigating cognition: Spatial codes for human thinking. *Science*, *362*(6415), eaat6766. https://doi.org/10.1126/science.aat6766

[31] de Santana Correia, A., & Colombini, E. L. (2022). Attention, please! A survey of neural attention models in deep learning. *Artificial Intelligence Review*, *55*(8), 6037–6124. https://doi.org/10.1007/s10462-022-10148-x

[32] Zhang, Z., Ding, X., Liang, X., Zhou, Y., Qin, B., & Liu, T. (2025). Brain and cognitive science inspired deep learning: A comprehensive survey. *IEEE Transactions on Knowledge and Data Engineering*, *37*(4), 1650–1671. https://doi.org/10.1109/TKDE.2025.3527551

[33] Felleman, D. J., & van Essen, D. C. (1991). Distributed hierarchical processing in the primate cerebral cortex. *Cerebral Cortex*, *1*(1), 1–47. https://doi.org/10.1093/cercor/1.1.1-a

[34] Mesulam, M. M. (1998). From sensation to cognition. *Brain*, *121*(6), 1013–1052. https://doi.org/10.1093/brain/121.6.1013

[35] Mountcastle, V. B. (1978). An organizing principle for cerebral function: The unit model and the distributed system. In G. M. Edelman & V. B. Mountcastle (Eds.), *The mindful brain: Cortical organization and the group-selective theory of higher brain function* (pp. 7–50). MIT Press.

[36] Grill-Spector, K., & Malach, R. (2004). The human visual cortex. *Annual Review of Neuroscience*, *27*, 649–677. https://doi.org/10.1146/annurev.neuro.27.070203.144220

[37] Friston, K. (2010). The free-energy principle: A unified brain theory? *Nature Reviews Neuroscience*, *11*(2), 127–138. https://doi.org/10.1038/nrn2787

[38] Kanwisher, N. (2010). Functional specificity in the human brain: A window into the functional architecture of the mind. *Proceedings of the National Academy of Sciences*, *107*(25), 11163–11170. https://doi.org/10.1073/pnas.1005062107

[39] Posner, M. I., & Petersen, S. E. (1990). The attention system of the human brain. *Annual Review of Neuroscience*, *13*, 25–42. https://doi.org/10.1146/annurev.ne.13.030190.000325

[40] Itti, L., & Koch, C. (2001). Computational modelling of visual attention. *Nature Reviews Neuroscience*, *2*(3), 194–203. https://doi.org/10.1038/35058500

[41] Corbetta, M., & Shulman, G. L. (2002). Control of goal-directed and stimulus-driven attention in the brain. *Nature Reviews Neuroscience*, *3*(3), 201–215. https://doi.org/10.1038/nrn755

[42] Buschman, T. J., & Miller, E. K. (2007). Top-down versus bottom-up control of attention in the prefrontal and posterior parietal cortices. *Science*, *315*(5820), 1860–1862. https://doi.org/10.1126/science.1138071

[43] Baddeley, A. (2012). Working memory: Theories, models, and controversies. *Annual Review of Psychology*, *63*, 1–29. https://doi.org/10.1146/annurev-psych-120710-100422

[44] D'Esposito, M., & Postle, B. R. (2015). The cognitive neuroscience of working memory. *Annual Review of Psychology*, *66*, 115–142. https://doi.org/10.1146/annurev-psych-010814-015031

[45] Sridhar, S., Khamaj, A., & Asthana, M. K. (2023). Cognitive neuroscience perspective on memory: Overview and summary. *Frontiers in Human Neuroscience*, *17*, 1217093. https://doi.org/10.3389/fnhum.2023.1217093

[46] Chen, Z. S., & Wilson, M. A. (2023). How our understanding of memory replay evolves. *Journal of Neurophysiology*, *129*(3), 552–580. https://doi.org/10.1152/jn.00454.2022

[47] Battaglia, F. P., Benchenane, K., Sirota, A., Pennartz, C. M. A., & Wiener, S. I. (2011). The hippocampus: Hub of brain network communication for memory. *Trends in Cognitive Sciences*, *15*(7), 310–318. https://doi.org/10.1016/j.tics.2011.05.008

[48] Kumar, A. A. (2021). Semantic memory: A review of methods, models, and current challenges. *Psychonomic Bulletin & Review*, *28*(1), 40–80. https://doi.org/10.3758/s13423-020-01792-x

[49] de Brigard, F., Umanath, S., & Irish, M. (2022). Rethinking the distinction between episodic and semantic memory: Insights from the past, present, and future. *Memory & Cognition*, *50*(3), 459–463. https://doi.org/10.3758/s13421-022-01299-x

[50] Costa, R. P., Froemke, R. C., Sjöström, P. J., & van Rossum, M. C. (2015). Unified pre- and postsynaptic long-term plasticity enables reliable and flexible learning. *eLife*, *4*, e09457. https://doi.org/10.7554/eLife.09457

[51] Keck, T., Keller, G. B., Jacobsen, R. I., Eysel, U. T., Bonhoeffer, T., & Hübener, M. (2013). Synaptic scaling and homeostatic plasticity in the mouse visual cortex in vivo. *Neuron*, *80*(2), 327–334. https://doi.org/10.1016/j.neuron.2013.08.018

[52] Magee, J. C., & Grienberger, C. (2020). Synaptic plasticity forms and functions. *Annual Review of Neuroscience*, *43*, 95–117. https://doi.org/10.1146/annurev-neuro-090919-022842

[53] Harris, K. D., & Thiele, A. (2011). Cortical state and attention. *Nature Reviews Neuroscience*, *12*(9), 509–523. https://doi.org/10.1038/nrn3084

[54] Keller, G. B., & Mrsic-Flogel, T. D. (2018). Predictive processing: A canonical cortical computation. *Neuron*, *100*(2), 424–435. https://doi.org/10.1016/j.neuron.2018.10.003

[55] Yuste, R. (2015). From the neuron doctrine to neural networks. *Nature Reviews Neuroscience*, *16*(8), 487–497. https://doi.org/10.1038/nrn3962

[56] Chowdhery, A., Narang, S., Devlin, J., Bosma, M., Mishra, G., Roberts, A., ..., & Fiedel, N. (2023). PaLM: Scaling language modeling with pathways. *Journal of Machine Learning Research*, *24*(240), 1–113.

[57] Wang, Y., & Luo, Z. (2023). Enhance multi-domain sentiment analysis of review texts through prompting strategies. In *2023 International Conference on High Performance Big Data and Intelligent Systems*, 1–7. https://doi.org/10.1109/HDIS60872.2023.10499502

[58] Wang, Y., Luo, Z., Wang, J., Zhou, Z., Chen, Y., & Han, B. (2025). Eliciting causal abilities in large language models for reasoning tasks. *Proceedings of the AAAI Conference on Artificial Intelligence, 39*(14), 15212–15220. https://doi.org/10.1609/aaai.v39i14.33669

[59] Alayrac, J.-B., Donahue, J., Luc, P., Miech, A., Barr, I., Hasson, Y., ..., & Simonyan, K. (2022). Flamingo: A visual language model for few-shot learning. In *Proceedings of the 36th International Conference on Neural Information Processing Systems*, 23716–23736. https://dl.acm.org/doi/10.5555/3600270.3601993

[60] Lu, P., Peng, B., Cheng, H., Galley, M., Chang, K.-W., Wu, Y. N., ..., & Gao, J. (2023). Chameleon: Plug-and-play

compositional reasoning with large language models. In *Proceedings of the 37th International Conference on Neural Information Processing Systems*, 43447–43478. https://dl.acm.org/doi/10.5555/3666122.3668004

[61] Patterson, D., Gonzalez, J., Hölzle, U., Le, Q., Liang, C., Munguia, L.-M., ..., & Dean, J. (2022). The carbon footprint of machine learning training will plateau, then shrink. *Computer*, *55*(7), 18–28. https://doi.org/10.1109/MC.2022.3148714

[62] Alzoubi, Y. I., & Mishra, A. (2024). Green artificial intelligence initiatives: Potentials and challenges. *Journal of Cleaner Production*, *468*, 143090. https://doi.org/10.1016/j.jclepro.2024.143090

[63] Hoffmann, J., Borgeaud, S., Mensch, A., Buchatskaya, E., Cai, T., Rutherford, E., ..., & Sifre, L. (2022). Training compute-optimal large language models. In *Proceedings of the 36th International Conference on Neural Information Processing Systems*, 30016–30030. https://dl.acm.org/doi/10.5555/3600270.3602446

[64] Parisi, G. I., Kemker, R., Part, J. L., Kanan, C., & Wermter, S. (2019). Continual lifelong learning with neural networks: A review. *Neural Networks*, *113*, 54–71. https://doi.org/10.1016/j.neunet.2019.01.012

[65] Ramasesh, V. V., Dyer, E., & Raghu, M. (2021). Anatomy of catastrophic forgetting: Hidden representations and task semantics. In *International Conference on Learning Representations*, 1–31.

[66] Wang, H., Zhang, Y., Yang, Y., Zheng, Y., & Wong, K.-F. (2024). Acquiring new knowledge without losing old ones for effective continual dialogue policy learning. *IEEE Transactions on Knowledge and Data Engineering*, *36*(12), 7569–7584. https://doi.org/10.1109/TKDE.2023.3344727

[67] Shi, H., Xu, Z., Wang, H., Qin, W., Wang, W., & Wang, Y., ..., & Wang, H. (2025). Continual learning of large language models: A comprehensive survey. *ACM Computing Surveys*, *58*(5), 120. https://doi.org/10.1145/3735633

[68] de Lange, M., Aljundi, R., Masana, M., Parisot, S., Jia, X., & Leonardis, A. (2022). A continual learning survey: Defying forgetting in classification tasks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *44*(7), 3366–3385. https://doi.org/10.1109/TPAMI.2021.3057446

[69] Zhao, T., Zhang, L., Ma, Y., & Cheng, L. (2024). A survey on safe multi-modal learning systems. In *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, 6655–6665. https://doi.org/10.1145/3637528.3671462

[70] Hendrycks, D., Basart, S., Mu, N., Kadavath, S., Wang, F., Dorundo, E., ..., & Gilmer, J. (2021). The many faces of robustness: A critical analysis of out-of-distribution generalization. In *2021 IEEE/CVF International Conference on Computer Vision*, 8320–8329. https://doi.org/10.1109/ICCV48922.2021.00823

[71] Raza, A. (2022). Scalable architectures for distributed commonsense knowledge bases with real-time synchronization and fault tolerance. *Open Journal of Robotics, Autonomous Decision-Making, and Human-Machine Interaction*, *7*(12), 1–16.

[72] Burns, C., Ye, H., Klein, D., & Steinhardt, J. (2023). Discovering latent knowledge in language models without supervision. In *11th International Conference on Learning Representations*, 1–28.

[73] Schramowski, P., Turan, C., Andersen, N., Rothkopf, C. A., & Kersting, K. (2022). Large pre-trained language models contain human-like biases of what is right and wrong to do. *Nature Machine Intelligence*, *4*(3), 258–268. https://doi.org/10.1038/s42256-022-00458-8

[74] Belinkov, Y., & Glass, J. (2019). Analysis methods in neural language processing: A survey. *Transactions of the Association for Computational Linguistics*, *7*, 49–72. https://doi.org/10.1162/tacl_a_00254

[75] Hofman, J. M., Watts, D. J., Athey, S., Garip, F., Griffiths, T. L., Kleinberg, J., ..., & Yarkoni, T. (2021). Integrating explanation and prediction in computational social science. *Nature*, *595*(7866), 181–188. https://doi.org/10.1038/s41586-021-03659-0

[76] Shazeer, N., Mirhoseini, A., Maziarz, K., Davis, A., Le, Q., Hinton, G., & Dean, J. (2017). Outrageously large neural networks: The sparsely-gated mixture-of-experts layer. In *5th International Conference on Learning Representations*, 1–19.

[77] Fedus, W., Zoph, B., & Shazeer, N. (2022). Switch transformers: Scaling to trillion parameter models with simple and efficient sparsity. *Journal of Machine Learning Research*, *23*(120), 1–39.

[78] Fusi, S., Drew, P. J., & Abbott, L. F. (2005). Cascade models of synaptically stored memories. *Neuron*, *45*(4), 599–611. https://doi.org/10.1016/j.neuron.2005.02.001

[79] Santoro, A., Bartunov, S., Botvinick, M., Wierstra, D., & Lillicrap, T. (2016). Meta-learning with memory-augmented neural networks. In *Proceedings of the 33rd International Conference on Machine Learning, 48*, 1842–1850.

[80] Lewis, P., Perez, E., Piktus, A., Petroni, F., Karpukhin, V., Goyal, N., ..., & Kiela, D. (2020). Retrieval-augmented generation for knowledge-intensive NLP tasks. In *Proceedings of the 34th International Conference on Neural Information Processing Systems*, *33*, 9459–9474.

[81] Kumar, K. P., Swarubini, P. J., & Ganapathy, N. (2025). Cognitive artificial intelligence. In P. V. Mohanan (Ed.), *Artificial intelligence and biological sciences* (pp. 301–323). CRC Press. https://doi.org/10.1201/9781003492726-18

[82] Goyal, A., Lamb, A., Hoffmann, J., Sodhani, S., Levine, S., Bengio, Y., & Schölkopf, B. (2021). Recurrent independent mechanisms. In *9th International Conference on Learning Representations*, 1–37.

[83] Zhao, X., Wang, L., Zhang, Y., Han, X., Deveci, M., & Parmar, M. (2024). A review of convolutional neural networks in computer vision. *Artificial Intelligence Review*, *57*, 99. https://doi.org/10.1007/s10462-024-10721-6

[84] Badamdorj, T., Rochan, M., Wang, Y., & Cheng, L. (2022). Contrastive learning for unsupervised video highlight detection. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 14022–14032. https://doi.org/10.1109/CVPR52688.2022.01365

[85] Merolla, P. A., Arthur, J. V., Alvarez-Icaza, R., Cassidy, A. S., Sawada, J., Akopyan, F., . . . , & Modha, D. (2014). A million spiking-neuron integrated circuit with a scalable communication network and interface. *Science*, *345*(6197), 668–673. https://doi.org/10.1126/science.1254642

[86] Ong, I., Almahairi, A., Wu, V., Chiang, W.-L., Wu, T., Gonzalez, J. E., ..., & Stoica, I. (2025). RouteLLM: Learning to route LLMs from preference data. In *the 13th International Conference on Learning Representations*, 1–16.

[87] Urbanczik, R., & Senn, W. (2014). Learning by the dendritic prediction of somatic spiking. *Neuron*, *81*(3), 521–528. https://doi.org/10.1016/j.neuron.2013.11.030

[88] Pehlevan, C., & Chklovskii, D. B. (2015). Optimization theory of hebbian/anti-hebbian networks for PCA and whitening. In *2015 53rd Annual Allerton Conference on Communication,*

Control, and Computing, 1458–1465. https://doi.org/10.1109/ALLERTON.2015.7447180

[89] Mansour, Y., Kaissar, A., & Ansari, S. (2024). Review on recent matrix multiplication optimization using deep learning. In *Intelligent and Fuzzy Systems: Intelligent Industrial Informatics and Efficient Networks Proceedings of the INFUS 2024 Conference*, *1,* 359–371. https://doi.org/10.1007/978-3-031-70018-7_41

[90] Muralidhar, R., Borovica-Gajic, R., & Buyya, R. (2022). Energy efficient computing systems: Architectures, abstractions and modeling to techniques and standards. *ACM Computing Surveys*, *54*(11s), 236. https://doi.org/10.1145/3511094

[91] Wang, L., Zhang, X., Su, H., & Zhu, J. (2024). A comprehensive survey of continual learning: Theory, method and application. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *46*(8), 5362–5383. https://doi.org/10.1109/TPAMI.2024.3367329

[92] Furber, S. (2016). Large-scale neuromorphic computing systems. *Journal of Neural Engineering*, *13*(5), 051001. https://doi.org/10.1088/1741-2560/13/5/051001

[93] Cervera, M. R. (2022). *Learning, forgetting and uncertainty estimation in artificial neural networks and in the brain*. PhD Thesis, ETH Zurich.

[94] Huth, A. G., de Heer, W. A., Griffiths, T. L., Theunissen, F. E., & Gallant, J. L. (2016). Natural speech reveals the semantic maps that tile human cerebral cortex. *Nature*, *532*(7600), 453–458. https://doi.org/10.1038/nature17637

[95] Shams, L., & Seitz, A. R. (2008). Benefits of multisensory learning. *Trends in Cognitive Sciences*, *12*(11), 411–417. https://doi.org/10.1016/j.tics.2008.07.006

[96] Bau, D., Zhu, J.-Y., Strobelt, H., Lapedriza, A., Zhou, B., & Torralba, A. (2020). Understanding the role of individual units in a deep neural network. *Proceedings of the National Academy of Sciences*, *117*(48), 30071–30078. https://doi.org/10.1073/pnas.1907375117

[97] Danks, D., & London, A. J. (2017). Regulating autonomous systems: Beyond standards. *IEEE Intelligent Systems*, *32*(1), 88–91. https://doi.org/10.1109/MIS.2017.1

[98] Jobin, A., Ienca, M., & Vayena, E. (2019). The global landscape of AI ethics guidelines. *Nature Machine Intelligence*, *1*(9), 389–399. https://doi.org/10.1038/s42256-019-0088-2

[99] Taddeo, M., & Floridi, L. (2018). How AI can be a force for good. *Science*, *361*(6404), 751–752. https://doi.org/10.1126/science.aat5991

How to Cite: Wang, J., Wang, Y., & Luo, Z. (2025). A Review on Bridging Brain-Inspired Mechanisms and Large-Scale Pre-trained Models: Toward Adaptive, Efficient, and Interpretable AI. *FinTech and Sustainable Innovation*, *1*, A15. https://doi.org/10.47852/bonviewFSI52026630