A Survey of Brain-Inspired Mechanisms for Neuro-Symbolic Reasoning

Florin Leon[a,*]

[a]"Gheorghe Asachi" Technical University of Iaşi, Romania
ORCID (Florin Leon): https://orcid.org/0000-0002-1370-9145

Abstract. Recent advances in Large Language Models have demonstrated that Transformer-based architectures can support symbolic-like reasoning without explicit symbolic formalisms. However, these models remain resource-intensive, opaque, and sometimes limited in systematic generalization and memory control. This paper reviews a set of biologically inspired mechanisms that may offer more efficient and flexible alternatives for implementing reasoning in neural systems, or even introduce new design principles. We explore models that address variable binding, compositionality, contextual inference and embedding, neuro-symbolic integration, and architectural designs inspired by neuroscience. They reveal how symbolic operations can emerge from continuous, distributed neural dynamics.

1 Introduction

In early 2025, there was a significant shift in the landscape of Artificial Intelligence (AI) when Large Language Models (LLMs) capable of robust multi-step reasoning entered the mainstream. These Transformer-based systems now demonstrate emergent algorithmic behaviors in a wide range of tasks, without explicit internal symbolic representations. As a result, biologically inspired approaches to reasoning, which once seemed essential for integrating neural and symbolic systems, seem to have lost their sense of urgency and even necessity.

However, the efficiency and interpretability challenges faced by LLMs raise new questions about the role that neuroscience can still play in the design of reasoning systems. Despite their impressive performance, current models remain highly resource-intensive, as they require massive datasets and extensive computation to achieve their results. In contrast, biological systems perform coherent compositional reasoning using far fewer resources. This motivates a look at brain-inspired computation, not as an alternative to Transformers, but as a potential source of principles for improving their scalability, modularity, and reasoning capabilities.

Although many studies explore neuro-symbolic (NS) methods, we focus on a smaller group of works that seem especially relevant to improving reasoning in LLMs. Rather than aiming for comprehensive coverage or classifying existing hybrid architectures, these examples were chosen for their potential to suggest concrete principles that could inspire more efficient neural systems. Also, while logic has long been considered the epitome of reasoning, and many works combine neural architectures with logical inference, we place less emphasis on these approaches in this survey. This is not due to their lack of value, but because we view logical reasoning not as an innate product of brain function, but as a learned, constrained mode of thought that appears under specific conditions. Instead, we analyze a range of biologically inspired mechanisms that could inform or complement reasoning in artificial systems, and illustrate how symbolic-like behavior can emerge from dynamic, distributed processes.

Several recent works have also surveyed the growing field of neuro-symbolic integration and brain-inspired AI. For example, [43] presents a taxonomy of NS learning systems, and emphasizes three paradigms: learning for reasoning, reasoning for learning, and learning-reasoning. It covers technical models and application domains, but its scope is broad and not specifically tailored to reasoning methods. Similarly, [3] offers a wide-ranging overview of NSAI, with a focus on representation, learning, and decision making for fields such as robotics and healthcare. Another review [26] explores NS approaches for Artificial Intelligence of Things (AIoT) applications, while [23] discusses biologically inspired strategies in the broader effort to realize Artificial General Intelligence (AGI).

This survey highlights five key aspects of reasoning in neural systems that align with mechanisms observed in biological cognition. These include the ability to maintain variable identity and role assignment, the formation and reuse of abstract functional components, the encoding of context-sensitive information, the integration of symbolic computation with neural dynamics, and the use of biologically inspired memory and control architectures. Thus, in the rest of the article, we analyze techniques relevant for variable binding (Section 2), compositionality (Section 3), handling context and embeddings (Section 4), neuro-symbolic and hybrid systems (Section 5), and architectures inspired by neuroscience (Section 6). This perspective allows us to extract some insights from neuroscience and assess their applicability to current or future reasoning models.

2 Variable Binding

Understanding how variable binding is implemented in neural systems, both artificial and biological, is central to explaining and improving reasoning. We begin with a study of LLMs, which already exhibit symbolic-like behavior, and then explore progressively more biologically grounded mechanisms that achieve similar goals.

The manner in which Transformer-based LLMs perform variable binding during in-context reasoning tasks is analyzed in [9]. The authors formalize the binding problem as the need to associate each entity with the correct attribute in multi-entity contexts. Through causal

* Corresponding Author. Email: florin.leon@academic.tuiasi.ro.

interventions, they uncover a distributed mechanism in which "binding ID" vectors are added to the base representations of entities and attributes. These vectors reside in a continuous subspace, are independent of token position, and allow LLMs to perform queries with factorizable and position-invariant behavior. The binding vectors transfer across tasks and models, which suggests that LLMs implicitly learn a general-purpose representational subspace for symbolic associations. This mechanism reveals how standard Transformer architectures can internally implement reasoning via learned, reusable vector-based operations.

Although Transformer LLMs can succeed in binding using learned vector representations, their architecture lacks explicit mechanisms for controlled storage, reuse, or indirection. The next paper [18] addresses this limitation by introducing a biologically inspired solution based on neural gating and addressing. It presents a neurocomputational model in which the prefrontal cortex (PFC) and basal ganglia (BG) jointly support variable binding via indirection. In this model, one PFC region or "stripe" encodes addresses or pointers to information stored in other PFC regions. These addresses are regulated by BG-mediated gating mechanisms, which control when specific bindings are updated, maintained, or used for output. Each stripe has a fixed identity mapped onto a unique activation pattern, which enables address-based routing without copying actual content. This architecture enables flexible reuse of values for roles and supports systematic generalization to new combinations that are not encountered during training. It also illustrates how indirection, an operation commonly used in computer science, might be realized in biological circuits.

While the PFC-BG model emphasizes dynamic routing and gating at the systems level, the following paper [22] explores how neural circuits might perform variable binding through associations at the level of spiking neurons. It models sparse neuron assemblies in the brain, which form dynamically in "neural spaces" and serve as pointers to concept assemblies located in a separate "content space". These assembly pointers form through fast Spike-Timing Dependent Plasticity (STDP), triggered by transient disinhibition. This mechanism supports symbolic-like operations such as binding, copying, and equality checking, all implemented without explicit symbolic representations.

In contrast to pointer models, the next approach [13] introduces a mechanism that supports symbolic binding through memory segmentation within a unified autoassociative network. It proposes Dynamically Partitionable Autoassociative Neural Networks (DPAANNs) as a biologically plausible architecture for solving the variable binding problem. DPAANNs integrate a central attractor-based memory system with dynamically segmentable buffers that represent symbolic roles. These buffers can be quickly bound to values by activating distinct subpopulations within the shared autoassociative space. The model supports role-value independence, allows compositional encoding and decoding, and enables variable manipulation through buffer-specific addressing, all using standard neural dynamics. Unlike anatomical or synchrony-based binding approaches, DPAANNs offer flexible, content-addressable binding without hardwiring or oscillatory control. The model draws on insights from symbolic cognitive architectures (e.g., ACT-R [1]), but remains grounded in biologically plausible assumptions about connectivity, attractor dynamics, and Hebbian plasticity.

While both indirection and pointer-based models emphasize spatial representation, the final paper [31] addresses a fundamentally different axis of computation: time. It explores how dynamic, oscillation-based synchronization can solve the binding problem with minimal structural overhead. It proposes that variable binding in working memory can be implemented through time-based synchronization, where each role-value pair is encoded by neural activity aligned to a distinct oscillatory phase. This temporal encoding enables multiple bindings to coexist without interference and allows rapid binding and unbinding. The model links memory capacity to oscillatory frequency. Slower oscillations permit more distinct bindings, while faster ones reduce capacity. Unlike synaptic binding, this method avoids persistent connections and supports flexible reuse. Simulations demonstrate how phase separation can maintain multiple bindings concurrently and explain capacity constraints in working memory.

Together, these models illustrate a diverse set of mechanisms (vector-based addition, indirection, assembly pointers, attractor-based partitioning, and oscillatory phases) through which variable binding can be achieved. They may offer complementary strategies for improving generalization, memory efficiency, and compositional reasoning in future intelligent systems.

Discussion

The papers reviewed in this section address a core challenge at the intersection of neural networks and symbolic reasoning, i.e., how to represent and manipulate information such as role-filler bindings or variable-value associations, within distributed, trainable systems. The shared objective is to discover mechanisms that allow neural models to maintain the identity of variables across operations, dynamically assign roles, and carry out functions like copying, comparison, indirection, and substitution.

A central insight in these works is that variable binding need not rely on discrete symbols or fixed architecture. Instead, it can emerge from learned dynamics over continuous representations. Several models achieve this by introducing neural forms of indirection, where patterns of activity act as pointers to other representations. These pointers can be created, reused, and recomposed dynamically, which enables systematic generalization and flexible memory access. Other approaches use oscillatory or time-based mechanisms to encode multiple bindings in parallel, which allows variable-role pairs to be represented and disentangled based on temporal phase. Still others rely on attractor-based partitioning or topologically organized buffers that support role-specific storage and recombination.

From the perspective of LLMs, these insights suggest possible directions for architectural and training innovations. For instance, the addition of explicit binding subspaces, similar to learned role vectors or binding identities, could improve the handling of coreference, substitution, and memory-intensive tasks. Indirection and pointer-based mechanisms could allow models to learn internal memory protocols that support variable reuse in different contexts, while time-based activation patterns could facilitate dynamic binding and unbinding without overwriting content. Such additions could enhance the ability to learn and apply reusable abstractions, improve interpretability, and support systematic generalization.

3 Compositionality

To see how modern neural models could acquire and generalize compositional structures, we begin this section with some evidence that standard networks may develop modular internal subroutines spontaneously, and then trace a path through increasingly explicit architectural and learning designs that aim to support systematic generalization [24]. The authors define a notion of compositionality and apply model pruning techniques with continuous sparsification to isolate subnetworks responsible for individual subfunctions. These

include tasks such as detecting spatial relations in vision or subject-verb agreement in language. They find that, across architectures and domains, subnetworks often encode distinct, functionally specialized components. These subnetworks can be ablated to selectively disrupt one function without impairing others, which offers evidence for modular task decomposition. Self-supervised pretraining increases the clarity and consistency of this modular organization, especially in language models. The study suggests that gradient-based learning alone can produce compositional representations under the right conditions.

The next study [21] directly optimizes for compositional behavior. It shows how training procedures, rather than architectures alone, can induce systematic generalization. Meta-Learning for Compositionality (MLC) is a method that trains a standard Transformer to acquire human-like compositional generalization. This is based on a large number of short training episodes that involve learning a small artificial language of invented words. In each episode, a few examples of input-output word pairs are presented where, e.g., "dax" refers to a red circle and "fep" means to repeat the previous word three times. Afterwards, the model is required to generate outputs for new combinations that were not present during training. Through meta-learning, the Transformer becomes proficient at inferring latent grammars from limited data and applying learned rules to new combinations. The resulting model exhibits both flexibility and systematicity, outperforms standard neural networks, and matches human generalization patterns. It also replicates human-like errors, which reveals similar inductive biases. This result shows that compositional reasoning can emerge in standard architectures when training explicitly promotes the development of compositionality.

While MLC encourages systematicity through task distribution, the next approach [15] focuses on architectural inductive bias. It aims to make compositional reasoning an explicit part of the internal operations of the network. For this purpose, it proposes the Memory, Attention, and Composition (MAC) network, a fully differentiable architecture for visual reasoning. Each MAC cell maintains separate control and memory states to represent the current reasoning goal and the intermediate result. The model answers questions by dividing them into a sequence of discrete steps, where each cell selects a relevant part of the question, retrieves information from the image, and updates memory accordingly. The architecture uses soft attention and gated updates to support flexible yet interpretable reasoning over both language and vision. This design imposes a prior that enforces stepwise reasoning. Thus, the network achieves high accuracy and interpretability on the CLEVR benchmark [16] without explicit supervision.

Unlike MAC, which implements compositionality by means of architectural constraints, the following work [4] uses a hybrid NS design. It introduces the Neural-Symbolic Stack Machine (NeSS), which combines a symbolic stack machine with a neural controller. The symbolic component supports recursive operations such as "push", "pop", and "reduce", while the neural network learns to produce execution traces without trace-level supervision. The model generalizes well on multiple benchmarks, including SCAN [20] and few-shot language tasks, and achieves perfect generalization. A key innovation is the notion of operational equivalence, which enables the model to infer compositional categories by grouping functionally similar expressions. This integration of symbolic components with neural learning shows that deep networks can acquire abstract, rule-like behavior when given the appropriate inductive tools and execution model.

Finally, paper [29] introduces the Relation Network (RN), a module designed to perform relational reasoning by explicitly computing pairwise relations among entities. An RN computes a function over all object pairs and aggregates these outputs to support downstream inference. RNs excel in visual and text-based tasks that require the understanding of inter-object relationships, such as comparing object attributes or predicting physical interactions. They outperform standard neural architectures on relational tasks, particularly on the CLEVR dataset, and do so without relying on symbolic inputs or supervision. By enforcing a relational inductive bias at the module level, RNs provide a plug-and-play mechanism for embedding reasoning within otherwise generic networks.

Discussion

These papers show that compositionality can emerge in neural systems through architectural constraints, meta-learning regimes, or task designs that implicitly favor modularity. The surveyed approaches differ in their methods, benefits, and unresolved issues. One line of work uses sparsification to isolate subnetworks linked to specific subfunctions such as counting or comparison. These components remain functional under ablation, which shows that modular computation can emerge without explicit supervision. However, this discovery happens after training, while an important question is how to encourage persistent functional separation during learning. Meta-learning methods like MLC expose a Transformer to short synthetic episodes that require rule abstraction and reuse. This setup enables strong generalization and is also able to replicate human error patterns. Still, the approach depends on many generated tasks and grammars, which limits its scalability. Applying similar ideas to real-world data remains an open issue. Architectural designs such as the MAC model use stepwise attention and memory operations controlled by separate units. These models generalize well on visual tasks and produce interpretable intermediate results, but their sequential nature may reduce efficiency on long text inputs. Hybrid systems such as NeSS combine neural networks with recursive symbolic operations. They generalize on tasks like SCAN and program induction but rely on curriculum learning and predefined operations. Relation Networks introduce relational modules that compare object pairs. They perform well on relational tasks but do not address recursion or variable binding.

For LLMs, these insights imply that generalization and rule application may not require discrete symbolic modules. Instead, such capabilities could emerge more reliably if LLMs were trained under regimes that induce abstraction, for example, using episodic meta-learning, architectural modularity, or targeted pruning techniques. They also suggest that interventions at the representation level, e.g., to isolate functional subspaces or inject compositional operators, could enhance interpretability and robustness, particularly for multi-step reasoning tasks. These results challenge the assumption that Transformers are inherently non-compositional and offer mechanisms to steer their internal dynamics toward more generalizable computation.

4 Contextual Inference and Embedding Techniques

Understanding how neural systems represent context requires models that combine representation learning with memory and inference. This section reviews biologically inspired approaches that address these challenges through various embedding techniques, contextual inference frameworks, and representational models that support abstraction and relational reasoning. We begin with models that gener-

ate efficient embeddings from sparse neural codes, and move toward those that organize memory and support flexible navigation in semantic and conceptual spaces.

A biologically inspired model for word embeddings based on the architecture of the mushroom body of the fruit fly is presented in [25]. This network uses sparse, competitive dynamics to transform word-context pairs into binary codes, with global inhibition based on a k-winners-take-all (kWTA) mechanism. Learning is driven by exposure to high-frequency word-context pairs and modulated by word rarity, which enables the system to capture semantic similarity and context-sensitive word senses. These sparse embeddings, i.e., the FlyVec model, support tasks such as word-sense disambiguation and document classification, with high computational efficiency.

Building on this foundation, some follow-up studies refine the approach and extend its applicability. One of them [32] introduces a continual learning rule that updates only synapses from the top active units to the target class output, leaving all others frozen. This sparsity fixes the number of synapses updated per example and limits its interference. Another work [10] extends the model from words to full sentences, i.e., the Comply method, by encoding word positions as complex phases and learning a single complex-valued parameter matrix. A kWTA stage then produces compact, interpretable binary sentence embeddings that preserve both semantics and word order.

While the fly-inspired models capture context through local coding, the COIN framework [14] introduces an approach grounded in Bayesian inference. It posits that the brain maintains multiple latent context representations that guide memory creation, expression, and updating. Contexts are inferred probabilistically from sensory cues, feedback, and time, rather than through direct observation. The model accounts for classical conditioning, episodic recall, decision making, and motor learning. It also distinguishes between proper learning, which involves memory updates, and apparent learning, which consists of adjustments to context beliefs. By treating context as a latent variable, the COIN model tries to explain how symbolic-like behavior can arise from probabilistic inference over hidden states.

Neuroscience research suggests that the hippocampus supports context-dependent representation and reasoning by organizing knowledge into structured internal spaces. Mechanisms such as cognitive maps (representations of relationships between states or concepts), successor representations (encodings of expected future state occupancy), and grid cells (neurons that exhibit periodic spatial firing patterns and help to pinpoint the current location) offer models for constructing embeddings and inference procedures.

Although originally studied for spatial navigation, these mechanisms also appear to support conceptual reasoning. Empirical evidence that the brain reuses spatial navigation mechanisms for abstract reasoning is provided in [5]. Human participants who had to learn a conceptual space defined by visual features of bird images exhibited grid-like activation patterns in the entorhinal cortex, analogous to those observed during physical navigation. This supports the idea that the brain encodes conceptual relationships using spatial structure, and suggests a shared computational substrate for spatial and non-spatial inference.

A successor representation (SR) is introduced in [34], where each state is encoded by the expected future occupancy of other states under a policy. This predictive model separates transition dynamics from goals and supports efficient updates to value functions. While originally applied to spatial navigation, SRs also provide a general framework for organizing relational knowledge; this idea is applied to semantic memory in [35]. A neural network encodes animal species using handcrafted features and learns a cognitive map based on expected feature-based similarity transitions. Varying the SR discount factor produces coarse or fine conceptual groupings, e.g., insects vs. mammals, and the model interpolates between known animals to classify new or incomplete inputs. This supports the use of predictive relational maps for abstract conceptual knowledge in order to generalize beyond training data.

A model that allows standard 2D grid cells to encode high-dimensional variables is proposed in [17]. The system uses random linear projections to embed high-dimensional inputs into periodic activity patterns across multiple grid modules. This mixed modular code enables linear decoding of positions in abstract vector spaces and supports multiple variable types without modifying the network architecture. It resolves the dimensionality bottleneck in grid codes by preserving pairwise relations and scaling to arbitrarily many dimensions.

The authors of [28] suggest that both spatial and conceptual representations arise in fact from a general clustering mechanism, where grid-like patterns in navigation tasks emerge from uniform sampling, while conceptual clusters reflect semantic similarity.

Discussion

These papers highlight how context-sensitive embeddings that support abstraction, generalization, and memory in neural systems can emerge from predictive, probabilistic, or geometric codes. An important idea is that the brain infers latent contexts to determine when to store, retrieve, or update information. This allows it to handle discontinuities in experience without overwriting past knowledge, a property essential for continual learning. SRs and grid-like codes provide compact embeddings that encode relations and support flexible planning and classification, even in conceptual domains.

These principles can also suggest some ways to extend LLMs. A model that infers latent context could decide which information to retrieve or update based on changes in input or task. Spatial coding schemes, such as SRs or grid-based encodings, could replace fixed position embeddings and organize concepts based on relational patterns. This may help models to recognize analogies or reuse knowledge across tasks. Sparse and competitive embedding mechanisms could increase memory efficiency and allow the representation of multiple meanings of a word or concept, depending on context. These models propose biologically grounded strategies for encoding information in ways that support robust reasoning and generalization for different tasks, challenges that remain essential to scaling and extending LLM capabilities.

5 Neuro-Symbolic and Hybrid Systems

This section presents some developments that bridge neural learning with symbolic or algorithmic methods. We begin with models that couple neural controllers with differentiable memory, then explore systems that integrate symbolic reasoning more explicitly, and close the section with architectures inspired by biological memory systems.

The Differentiable Neural Computer (DNC) [12] offers an influential example of neuro-symbolic integration. Building on the earlier Neural Turing Machine (NTM) model [11], DNC augments a recurrent neural network with a differentiable external memory matrix. This setup allows the model to perform operations that resemble reading and writing variables in traditional computing. The DNC controller learns to access memory through soft attention mechanisms that include content-based retrieval, temporal linking (storing

the order of writes in a temporal link matrix), and usage-based allocation (tracking memory usage to guide writes to unused locations). These mechanisms give the model the ability to construct and manipulate data structures such as lists, trees, and graphs. As a result, the DNC can handle tasks like pathfinding, graph inference, or array sorting. Moreover, it can generalize for variable-length inputs and can perform operations that resemble classical procedural logic within a fully differentiable framework. Unlike classic neural networks that entangle memory with computation, the DNC separates memory and control, which enables behavior that resembles algorithmic processing.

While the DNC emphasizes a general purpose external memory, the Neuro-Symbolic Concept Learner [27] introduces a modular design and symbolic reasoning into visual question answering. The model decomposes the learning process into three components: a visual perception module that extracts object-level features, a semantic parser that translates natural language questions into symbolic programs, and a program executor that interprets these programs on the scene. All modules are jointly trained from image–question–answer tuples, and do not require annotated object labels. Visual attributes are modeled as neural operators that map object embeddings into interpretable concept spaces (e.g., shape, color), while symbolic programs capture compositional logic through executable sequences. This architecture enables generalization to new object configurations, new visual domains, and longer queries.

The previous models embed symbolic control directly within the architecture. The next approach focuses on teaching neural networks to imitate the stepwise behavior of classical algorithms. In the Neural Algorithmic Reasoning framework [39], neural networks are trained to emulate traditional algorithms, such as Dijkstra's shortest path and value iteration. The learning process proceeds in stages. A neural processor first learns algorithmic steps by training on low-dimensional abstract inputs. Then, encoder and decoder networks transform real-world inputs into and out of the latent space of the processor. This separation allows the model to preserve algorithmic invariants while adapting to noisy, high-dimensional data. It addresses the algorithmic bottleneck, i.e., the problem of compressing complex real-world inputs into low-dimensional representations required by traditional algorithms. This proves to be especially powerful in reinforcement learning tasks, where latent planning through algorithmic modules improves performance in complex or partially observed environments.

The Hint-ReLIC method [2] improves the generalization of neural algorithmic models by incorporating causal regularization. The authors notice that many different inputs can lead to identical intermediate computations in algorithms. Based on this, they propose a self-supervised contrastive learning objective that encourages graph neural networks to produce similar internal representations for such inputs. Using a causal graph to formalize this invariance, the model generates augmented examples that preserve execution trajectories, and enforce stepwise consistency for different variants. This improves out-of-distribution performance on algorithmic reasoning benchmarks such as CLRS-30 [38], particularly for sorting and graph tasks.

The Shared Dual Memory Transformer (SDMTR) [44] modifies the standard Transformer architecture by replacing self-attention with a memory-based system inspired by the brain. The model introduces two shared memory components: a workspace that acts like working memory, and a long-term memory (LTM) that stores useful information across layers. At each layer, input tokens (that act as independent modules, in fact, units) compete to write to the workspace using sparse attention. Only the most relevant tokens are allowed to update the workspace. The workspace is then broadcast back to all tokens to guide further processing. Important workspace content is stored in LTM using outer product attention, which allows the model to build high-capacity memory representations. During inference, the model retrieves relevant information from LTM to guide token updates. This design supports iterative reasoning and helps the model to generalize better on relational tasks. SDMTR is reported to outperform standard Transformers on benchmarks such as bAbI [40], Sort-of-CLEVR [30], and the "triangle detection" visual reasoning task [36].

Discussion

Neuro-symbolic and hybrid models approach reasoning with different strengths and limitations. The Differentiable Neural Computer augments a learned controller with an external memory that supports content-based lookup, temporal linking, and dynamic allocation. It enables scalable memory use and variable-like access, and performs well on synthetic question-answering and graph tasks. However, training is computationally expensive, attention cost increases with memory size, and stability with very large memories remains unresolved. Hint-ReLIC improves neural algorithmic reasoning by aligning hidden states across identical intermediate steps in an algorithm. It enforces invariance when different inputs imply the same next step, which boosts out-of-distribution accuracy on CLRS benchmarks. However, the approach depends on access to algorithm trajectories or manually generated hints, which may limit its applicability to real-world data. The Neuro-Symbolic Concept Learner combines object detection, parsing, and a symbolic program executor, and achieves high accuracy on CLEVR using natural supervision. Yet, it relies on curriculum learning and clean object masks, which makes transfer to cluttered or ambiguous scenes an open problem. The Shared Dual-Memory Transformer introduces a competition-based workspace and a long-term memory. It reports good results on reasoning benchmarks and enables memory visualization. Still, it can have high computation costs and lacks evaluation on long language tasks or noisy inputs.

For Transformer-based LLMs, these ideas provide a blueprint for enhancing compositional generalization, especially in domains that require logic, recursion, or algorithmic manipulation. Incorporating symbolic intermediates, such as learned execution traces or memory read-write patterns, can make reasoning processes more interpretable and modular. Hybrid systems also offer a way to disentangle content (e.g., object or entity representations) from operations (e.g., comparison or traversal), which reduces the burden on attention mechanisms to simulate algorithmic flow. These models may suggest that carefully constrained hybridization and not just greater scale may be necessary, or at least helpful, to achieve robust reasoning in LLMs.

6 Neuroscience-Inspired Architectures

This section presents some models that use inspiration from the hippocampus, prefrontal cortex, and broader cortical dynamics to build architectures capable of generalization, composition, and memory manipulation that can support forms of reasoning relevant to AI.

The Tolman-Eichenbaum Machine (TEM) [41] provides a unified model for how the hippocampus supports both spatial navigation and relational reasoning. It represents tasks as graph-based transitions and explains how the hippocampus links relational codes (encoded by grid-like representations) with sensory content using fast Hebbian learning. This separation of relational patterns and sensory detail en-

ables the model to generalize across tasks, transfer knowledge to new environments, and produce consistent remapping patterns. TEM accounts for a wide range of neural responses observed in neuroscience experiments.

Building on this foundation, the TEM-t model [42] reframes the components of TEM using a Transformer architecture. The authors introduce a modified Transformer that uses recurrent positional encodings and causal attention. They show that it can reproduce spatial cell types and memory behaviors similar to the hippocampus. A formal equivalence is proven between Transformer attention and Hebbian memory retrieval mechanisms, which suggests that brain-like architectures and modern Transformer models share underlying computational principles. This formulation supports the idea that Transformer-based systems could benefit from models of hippocampal function and offers a framework for integrating relational memory with language and abstract reasoning.

While TEM is designed to support generalization through relational mapping across spatial and task-based contexts, the model in [19] proposes that hippocampal replay supports compositional generation. Replay refers to the brain's reactivation of neural sequences during rest or planning, often at accelerated timescales. The paper argues that these sequences do not merely reflect past experiences but can combine entities and roles, such as "verb" or "start point", to build new representations. The authors suggest that replay combines role-bound elements into new configurations, and allows the system to infer facts that were never explicitly learned. They present experimental evidence that replay can generate unexperienced sequences, support abstract reasoning, and construct compound knowledge. These insights recast replay not only as a memory mechanism but as a computational resource for relational and symbolic inference.

Working memory frameworks are extended in [33] by introducing adaptive chunking as a learned strategy for managing the tradeoff between memory precision and capacity. It uses reinforcement signals to decide when to store raw inputs and when to merge similar items into shared representations, depending on task demands. This chunking mechanism allows the network to store more information with fewer resources while accepting a controlled loss in precision. The model accounts for recency bias, i.e., the improved recall of recent items, as a consequence of selectively updating or replacing earlier representations. It also explains differential chunking, where the likelihood of merging items increases with their similarity and decreases with the number of items.

The model in [37] simulates how semantic knowledge can emerge from word learning grounded in sensory and motor experience. It uses a spiking neural network with biologically plausible connectivity and Hebbian learning to associate spoken words with perceptual and action-based features. These associations produce distributed neural cell assemblies that integrate phonological, visual, and motor information without requiring labeled data or external supervision. Words for objects activate vision-related circuits, while action words activate motor-related circuits. These category-specific activations converge in multimodal regions that function as semantic hubs. The model shows how semantic representations can self-organize from repeated exposure to co-occurring patterns of sound and sensorimotor input, and it explains both the emergence of modality-specific word meanings and shared multimodal representations based on the architecture of the network and learning dynamics.

The Semantic Pointer Architecture (SPA) [6] is a cognitive architecture that aims to explain how high-level behavior can arise from low-level neural interactions. The modeling starts with neurons with tuning curves, where each neuron responds more or less strongly depending on the input values, and uses these responses to encode vectors (similar to the activations of neural populations or cell assemblies) and apply transformations [7]. SPA uses these neural building blocks to create high-level representations called semantic pointers, i.e., fixed-size vectors that can combine concepts using circular convolution, and later recover their parts. This allows a neural system to represent rules, memories, and sequences in a way that supports reasoning and control. The architecture was used to create a computational model with millions of spiking neurons capable of visual processing and action planning, e.g., recognizing digits and remembering lists, counting, answering questions, drawing, and solving psychological tasks like Raven's matrices [8].

Discussion

These models highlight how biologically inspired mechanisms, such as information replay, chunking, or multimodal grounding, can inform the design of neural architectures with enhanced reasoning capabilities. A recurring theme is the separation of abstract relational patterns from episodic content, implemented through distinct coding strategies or dynamic binding operations. This stands in contrast to current LLMs, which merge function and content within a single representation space. Mechanisms like fast Hebbian learning and adaptive chunking could offer efficient means to encode, update, and reuse relational information without retraining, a feature that could enable more sample-efficient reasoning in LLMs. Replay-based models show that sequence generation need not rely on sampling from static representations. Instead, reasoning can emerge from compositional recombination of role-bound elements, a process more similar to planning than retrieval. Integrating this with LLMs could improve zero-shot generalization and support multi-step inference through internal simulation rather than token prediction alone.

The use of competitive memory systems and task-sensitive gating illustrates how biological networks manage precision-capacity tradeoffs. These mechanisms could inspire selective memory routing in LLMs, which would enable models to preserve important relational patterns while compressing redundant inputs. Likewise, the emergence of cell assemblies in semantic grounding models shows how distributed representations can self-organize to reflect shared abstractions for different modalities. This points to opportunities for LLMs to learn grounded semantic representations through unsupervised multimodal training.

7 Conclusions

This survey has examined a range of brain-inspired mechanisms that can offer insights into how symbolic-like reasoning can be implemented within neural systems. While recent advances in LLMs have brought multi-step reasoning into the mainstream, they have come at significant computational cost and with limited interpretability. In contrast, biological systems perform compositional, context-sensitive reasoning with remarkable efficiency and generalization capabilities, which may motivate the continued exploration of neural principles that could inform artificial architectures.

To summarize the main insights of the survey, Table 1 presents several biologically inspired ideas drawn from the five categories discussed above. For each category, it highlights specific mechanisms that could improve reasoning in neural architectures such as LLMs. These ideas aim to support capabilities such as variable tracking, memory control, compositional processing, and relational representations.

Table 1. Biologically inspired ideas for improving reasoning in large language models and other neural systems.

Category	Ideas for neural reasoning systems
Variable binding	- A dedicated role subspace in token embeddings could help track the roles of variables and preserve their identity across long contexts - Address-based indexing could allow attention to target specific memory locations and reuse intermediate results, like pointers - Sparse attention could help retrieve information by similarity with fewer activated elements, and more efficiency and interpretability - Dividing the residual stream into multiple task-specific branches that activate based on context could help hold several role-value pairs simultaneously without interference - Fast Hebbian plasticity could form temporary role-value bindings that dissolve after a reasoning step - Phase-based modulation could help maintain multiple bindings in parallel, separated by timing – or simulated using, e.g., learned sinusoidal gates
Compositionality	- Sparsity penalties or pruning could promote functional modules that remain interpretable and easy to update - Episodic meta-learning with few-shot tasks could support rule discovery and reuse in different domains - A (possibly latent) neural stack could support step-by-step composition, recursion, and hierarchical input processing - Relational attention with pairwise comparisons or edge labels could represent token relationships directly and improve judgments of equality, order, and grouping
Contextual inference and embedding techniques	- Sparse binary embeddings could reduce memory use while keeping concept meanings distinct - Embeddings inspired by successor representation could improve next token prediction by modeling likely future states - Complex-valued position encodings with phase information could unify order and meaning in a single operation - Position encodings inspired by grid cells could capture conceptual distance more effectively - Direction vectors in a learned relational map could support movement between related concepts - Freezing inactive parameters during continual training could preserve existing knowledge while integrating new information
Neuro-symbolic and hybrid systems	- A trainable controller with differentiable external memory could manage storage and retrieval of intermediate results for long reasoning tasks - A shared workspace memory could allow tokens to compete for relevance and promote the most important ones to longer-term memory - Generating soft programs or reasoning graphs could break complex queries into clear, step-by-step operations - Consistency constraints during training could help produce similar internal states for logically equivalent inputs - Training using contrastive examples with irrelevant input variations could help identify the relevant information that drives the next reasoning step
Neuroscience-inspired architectures	- A recurrent mechanism that tracks positional change over time could replace fixed encodings and give a more dynamic, context-aware sense of sequence - Letting tokens compete for a small number of write locations at each layer could form stable representations of important ideas - A chunking mechanism could merge memory items based on similarity and usage to improve memory efficiency - A replay mechanism could retrieve earlier sequences and replace entities in them to support counterfactual and imaginative reasoning - Multimodal training data could ground language in perception and action, which may improve transfer learning

Looking forward, such strategies could offer promising paths for advancing neuro-symbolic reasoning. One useful direction could be to build introspection tools that track internal activations during each step of a reasoning task. These tools could expose how variable bindings, rule applications, and memory retrievals map onto specific neural components. This could support better debugging, evaluation, and transparency. Another useful mechanism could come from biological timing patterns. The brain often uses nested oscillations to coordinate the activation of roles and values in sequences. Artificial models could implement similar timing-based schemes to manage which variables are active and when, and thus reduce interference during multi-step reasoning. A third idea could involve sleep-like replay. The brain replays key experiences offline to reinforce important associations and consolidate them into long-term memory. Neuro-symbolic models could benefit from a similar mechanism that periodically revisits past reasoning chains to strengthen important patterns without retraining on full datasets. These directions could offer specific, testable lines of work toward more robust reasoning systems.

As research progresses, incorporating such design principles into neural reasoning systems may lead to architectures that generalize better, reason more transparently, and operate with greater efficiency, moving closer to the adaptability and robustness of human cognition.

Acknowledgements

This research is supported by the project "Romanian Hub for Artificial Intelligence - HRIA", Smart Growth, Digitization and Financial Instruments Program, 2021-2027, MySMIS no. 334906.

References

[1] J. R. Anderson and C. Lebiere. *The Atomic Components of Thought*. Lawrence Erlbaum Associates, 1998.

[2] B. Bevilacqua, K. Nikiforou, B. Ibarz, I. Bica, M. Paganini, C. Blundell, J. Mitrovic, and P. Veličković. Neural algorithmic reasoning with causal regularisation. In *Proceedings of the 40th International Conference on Machine Learning (ICML)*, volume 202, pages 2272–2288. PMLR, 2023. URL https://proceedings.mlr.press/v202/bevilacqua23a.html.

[3] B. P. Bhuyan, A. Ramdane-Cherif, R. Tomar, and T. P. Singh. Neuro-symbolic artificial intelligence: A survey. *Neural Computing and Applications*, 36(21):12809–12844, 2024. doi: 10.1007/s00521-024-09960-z.

[4] X. Chen, C. Liang, A. W. Yu, D. Song, and D. Zhou. Compositional generalization via neural-symbolic stack machines. In *Advances in Neural Information Processing Systems*, volume 33, pages 1690–1701, 2020. URL https://proceedings.neurips.cc/paper/2020/hash/12b1e42dc0746f22cf361267de07073f-Abstract.html.

[5] A. O. Constantinescu, J. X. O'Reilly, and T. E. J. Behrens. Organizing conceptual knowledge in humans with a gridlike code. *Science*, 352 (6292):1464–1468, 2016. doi: 10.1126/science.aaf0941.

[6] C. Eliasmith. *How to Build a Brain: A Neural Architecture for Biological Cognition*. Oxford University Press, New York, NY, 2015.

[7] C. Eliasmith and C. H. Anderson. *Neural Engineering: Computation, Representation, and Dynamics in Neurobiological Systems*. Bradford Books, 2004.

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

[9] J. Feng and J. Steinhardt. How do language models bind entities in context? In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2024. URL https://openreview.net/forum?id=zb3b6oKO77.

[10] A. Figueroa, J. Westerhoff, G. Atefi, D. Fast, B. Winter, F. A. Gers, A. Löser, and W. Nejdl. Comply: Learning sentences with complex weights inspired by fruit fly olfaction, 2025. URL https://arxiv.org/abs/2502.01706.

[11] A. Graves, G. Wayne, and I. Danihelka. Neural turing machines, 2014. URL https://arxiv.org/abs/1410.5401.

[12] A. Graves, G. Wayne, M. Reynolds, T. Harley, I. Danihelka, A. Grabska-Barwińska, S. Gómez Colmenarejo, E. Grefenstette, T. Ramalho, J. Agapiou, A. Puigdomènech Badia, K. M. Hermann, Y. Zwols, G. Ostrovski, A. Cain, H. King, C. Summerfield, P. Blunsom, K. Kavukcuoglu, and D. Hassabis. Hybrid computing using a neural network with dynamic external memory. *Nature*, 538(7626):471–476, 2016. doi: 10.1038/nature20101.

[13] K. J. Hayworth. Dynamically partitionable autoassociative networks as a solution to the neural binding problem. *Frontiers in Computational Neuroscience*, 6:73, 2012. doi: 10.3389/fncom.2012.00073.

[14] J. B. Heald, M. Lengyel, and D. M. Wolpert. Contextual inference in learning and memory. *Trends in Cognitive Sciences*, 27(1):43–64, 2023. doi: 10.1016/j.tics.2022.10.004.

[15] D. A. Hudson and C. D. Manning. Compositional attention networks for machine reasoning. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2018. URL https://openreview.net/forum?id=S1Euwz-Rb.

[16] J. Johnson, B. Hariharan, L. van der Maaten, L. Fei-Fei, C. L. Zitnick, and R. Girshick. Clevr: A diagnostic dataset for compositional language and elementary visual reasoning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1988–1997, 2017. doi: 10.1109/CVPR.2017.215.

[17] M. Klukas, M. Lewis, and I. Fiete. Efficient and flexible representation of higher-dimensional cognitive variables with grid cells. *PLOS Computational Biology*, 16(4):e1007796, 2020. doi: 10.1371/journal.pcbi.1007796.

[18] T. Kriete, D. C. Noelle, J. D. Cohen, and R. C. O'Reilly. Indirection and symbol-like processing in the prefrontal cortex and basal ganglia. *Proceedings of the National Academy of Sciences*, 110(41):16390–16395, 2013. doi: 10.1073/pnas.1303547110.

[19] Z. Kurth-Nelson, T. Behrens, G. Wayne, K. Miller, L. Luettgau, R. Dolan, Y. Liu, and P. Schwartenbeck. Replay and compositional computation. *Neuron*, 111(4):454–469, 2023. doi: 10.1016/j.neuron.2022.12.028.

[20] B. M. Lake and M. Baroni. Generalization without systematicity: On the compositional skills of sequence-to-sequence recurrent networks. In *Proceedings of the 35th International Conference on Machine Learning (ICML)*, pages 2873–2882, 2018.

[21] B. M. Lake and M. Baroni. Human-like systematic generalization through a meta-learning neural network. *Nature*, 623(7985):115–121, 2023. doi: 10.1038/s41586-023-06668-3.

[22] R. Legenstein, C. H. Papadimitriou, S. Vempala, and W. Maass. Assembly pointers for variable binding in networks of spiking neurons, 2016. URL https://arxiv.org/abs/1611.03698.

[23] F. Leon. A review of findings from neuroscience and cognitive psychology as possible inspiration for the path to artificial general intelligence, 2024. URL https://arxiv.org/abs/2401.10904.

[24] M. A. Lepori, T. Serre, and E. Pavlick. Break it down: Evidence for structural compositionality in neural networks. In *Advances in Neural Information Processing Systems*, volume 36, 2023. URL https://openreview.net/forum?id=rwbzMiuFQl.

[25] Y. Liang, C. K. Ryali, B. Hoover, L. Grinberg, S. Navlakha, M. J. Zaki, and D. Krotov. Can a fruit fly learn word embeddings? In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2021. URL https://openreview.net/forum?id=xfmSoxdxFCG.

[26] Z. Lu, I. Afridi, H. J. Kang, I. Ruchkin, and X. Zheng. Surveying neuro-symbolic approaches for reliable artificial intelligence of things. *Journal of Reliable Intelligent Environments*, 10(3):257–279, 2024. doi: 10.1007/s40860-024-00231-1.

[27] J. Mao, C. Gan, P. Kohli, J. B. Tenenbaum, and J. Wu. The neuro-symbolic concept learner: Interpreting scenes, words, and sentences from natural supervision. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2019. URL https://arxiv.org/abs/1904.12584.

[28] R. M. Mok and B. C. Love. A non-spatial account of place and grid cells based on clustering models of concept learning. *Nature Communications*, 10(1):5685, 2019. doi: 10.1038/s41467-019-13760-8.

[29] A. Santoro, D. Raposo, D. G. T. Barrett, M. Malinowski, R. Pascanu, P. Battaglia, and T. Lillicrap. A simple neural network module for relational reasoning. In *Advances in Neural Information Processing Systems*, volume 30, pages 4967–4976, 2017. URL https://papers.nips.cc/paper/2017/hash/7082a-simple-neural-network-module-for-relational-reasoning.pdf.

[30] A. Santoro, D. Raposo, D. G. T. Barrett, M. Malinowski, R. Pascanu, P. Battaglia, and T. Lillicrap. A simple neural network module for relational reasoning. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2017.

[31] M. Senoussi, P. Verbeke, and T. Verguts. Time-based binding as a solution to and a limitation for flexible cognition. *Frontiers in Psychology*, 12:798061, 2022. doi: 10.3389/fpsyg.2021.798061.

[32] Y. Shen, S. Dasgupta, and S. Navlakha. Algorithmic insights on continual learning from fruit flies, 2021. URL https://arxiv.org/abs/2107.07617.

[33] A. Soni and M. J. Frank. Adaptive chunking improves effective working memory capacity in a prefrontal cortex and basal ganglia circuit. *eLife*, 13:RP97894, 2025. doi: 10.7554/eLife.97894.

[34] K. L. Stachenfeld, M. M. Botvinick, and S. J. Gershman. The hippocampus as a predictive map. *Nature Neuroscience*, 20(11):1643–1653, 2017. doi: 10.1038/nn.4650.

[35] P. Stöwer, A. Schilling, A. Maier, and P. Krauss. Neural network based formation of cognitive maps of semantic spaces and the putative emergence of abstract concepts. *Scientific Reports*, 13(1):3644, 2023. doi: 10.1038/s41598-023-30307-6.

[36] Thanapat1273. Triangle 3 dataset, 2022. URL https://universe.roboflow.com/thanapat1273/triangle_3.

[37] R. Tomasello, M. Garagnani, T. Wennekers, and F. Pulvermüller. A neurobiologically constrained cortex model of semantic grounding with spiking neurons and brain-like connectivity. *Frontiers in Computational Neuroscience*, 12:88, 2018. doi: 10.3389/fncom.2018.00088.

[38] P. Veličković, A. P. Badia, D. Budden, R. Pascanu, A. Banino, M. Dashevskiy, R. Hadsell, and C. Blundell. The clrs algorithmic reasoning benchmark. In *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pages 22084–22102, 2022.

[39] P. Veličković and C. Blundell. Neural algorithmic reasoning. *Patterns*, 2(7):100273, 2021. doi: 10.1016/j.patter.2021.100273.

[40] J. Weston, A. Bordes, S. Chopra, A. M. Rush, B. van Merriënboer, A. Joulin, and T. Mikolov. Towards ai-complete question answering: A set of prerequisite toy tasks, 2015. URL https://arxiv.org/abs/1502.05698.

[41] J. C. R. Whittington, T. H. Muller, S. Mark, G. Chen, C. Barry, N. Burgess, and T. E. J. Behrens. The tolman-eichenbaum machine: Unifying space and relational memory through generalization in the hippocampal formation. *Cell*, 183(5):1249–1263.e23, 2020. doi: 10.1016/j.cell.2020.10.024.

[42] J. C. R. Whittington, J. Warren, and T. E. J. Behrens. Relating transformers to models and neural representations of the hippocampal formation. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2022. URL https://openreview.net/forum?id=B8DVo9B1YE0.

[43] D. Yu, B. Yang, D. Liu, H. Wang, and S. Pan. A survey on neural-symbolic learning systems. *Neural Networks*, 166:105–126, 2023. doi: 10.1016/j.neunet.2023.06.028.

[44] X. Zeng, J. Lin, P. Hu, Z. Li, and T. Huang. Sdmtr: A brain-inspired transformer for relation inference. In *Proceedings of The 27th International Conference on Artificial Intelligence and Statistics*, volume 238 of *Proceedings of Machine Learning Research*, pages 3259–3267. PMLR, 2024. URL https://proceedings.mlr.press/v238/zeng24a.html.