Brain: Structural Memory, Thought Formation, and Language in a Biologically Grounded System

Vitalii Belyi*

Abstract

We present *Brain*, a biologically inspired **cognitive architecture** that models the complete pipeline from memory storage through thought formation to linguistic expression. Knowledge is encoded in the topology and discrete states of a connection graph rather than in real-valued weights. The system learns from text using strictly local plasticity rules and discrete synaptic states, without gradient-based optimization.

This work explores a cognitive architecture that separates three functional components: **(1) multimodal memory**, responsible for structural storage of experiences; **(2) thought formation**, understood as the dynamic reconfiguration of stored knowledge; and **(3) linguistic expression**, the sequential rendering of internal representations into language.

Within this framework, large language models are treated as optional linguistic interfaces corresponding to the third component, while the present work focuses on the design and evaluation of the memory substrate itself.

The proposed model represents word forms as discrete units connected by directed associations that transition through qualitative states (NEW → USED → MYELINATED → PRUNE) driven by local usage dynamics. Memory retrieval emerges from a local spreading-activation process with inhibition, decay, working-memory constraints, and biologically motivated competition. Episodic indexing is implemented via a hippocampus-inspired mechanism supporting pattern separation, CA3 attractor dynamics for pattern completion, and replay-driven consolidation [Teyler and DiScenna, 1986, McClelland et al., 1995, Rolls, 2013].

On a curated curriculum of basic world knowledge and question answering, the system achieves 98.8% accuracy (419/424 questions), including 100% on bAbI Task 1 working memory benchmark (250/250), while appropriately abstaining on unknown queries. The architecture and evaluation pipeline are released as open-source code at `https://github.com/sss777999/Brain`.

1 Introduction

Have you ever noticed that you already know what you want to say before you start speaking? The decision is there—but articulating it takes time. You search for words, construct sentences, and by the time you finish explaining, seconds or even minutes have passed.

But here is the deeper question: did "you" actually make that decision? Cognitive science suggests a more nuanced picture. The brain operates on multiple timescales: fast, automatic processes driven by neuromodulators and accumulated experience often reach conclusions before slower deliberative processes catch up [Kahneman, 2011]. Emotional and somatic states influence decisions before we become aware of them [Damasio, 1994]. What we experience as "deciding" may often be the conscious mind catching up to—and rationalizing—choices that faster systems have already biased or made.

*Independent researcher. Correspondence: aiops9a1@protonmail.com

Consider the evolutionary context. An elephant sees water and runs toward it—not because it "decided" to, but because hormonal signals make approaching water feel good [Panksepp, 2004]. A tiger hunts not from moral reasoning but because catching prey triggers reward [Berridge et al., 2009]. These are fast, subcortical systems shaped by millions of years of evolution. Humans inherited this machinery. We still have the same reward circuits, the same neuromodulators, the same drive toward what feels good and away from what feels bad.

But here is what may distinguish us—a hypothesis grounded in behavioral observation and supported by prefrontal cortex research [Miller and Cohen, 2001, Fuster, 2008]. Animals can suppress impulses, but typically only under external pressure: fear of a predator, pain from past experience, threat of retaliation. A wolf will back away from prey if the prey fights back. But a human can override an impulse *without* external pressure, through internal reasoning alone. We can feel rage and choose not to act. We can see the reward and walk away. Not always, not everyone, not reliably—but this capacity exists. This is what the prefrontal cortex adds: not just another layer of processing, but the ability to veto decisions that deeper systems have already made [Aron, 2007].

The question, then, is not whether we make choices—but whether we can sometimes override the choices our deeper systems have already made.

This does not diminish the role of thought. But it reframes it: conscious cognition may be less about *making* decisions and more about *evaluating, refining, and communicating* them. The verbal expression we produce—the explanation we give to others or even to ourselves—is a slow, sequential process layered on top of fast, parallel, largely unconscious computation.

This project began with that observation. Before starting, an extensive search was conducted for similar approaches: academic databases, arXiv, and AI-assisted literature reviews. While individual components exist in various research groups, no complete system combining discrete structural memory, biologically grounded plasticity, and explicit memory-thought-language separation appeared to exist. This gap motivated the present work.

1.1 Motivating Observations

Consider common introspective phenomena. When solving a problem alone, the solution often appears almost instantly—no words needed. But when explaining it to someone, the process slows dramatically. The thought is already formed; verbalization is the bottleneck.

Memory, upon closer examination, is not a single thing. One can recall a melody without visualizing anything, or remember sitting on a bench hearing the rustle of wind through leaves, without remembering what music was playing. These are separate channels—visual, auditory, semantic—that activate independently and combine into coherent experiences.

More striking: we know facts from school—physics, mathematics—but cannot recall *when* we learned them. The knowledge exists, detached from any episodic context. It was absorbed from trusted sources (parents, teachers, textbooks) and now simply *is*. Other memories are different: we remember exactly where we were when we heard certain news. The brain treats these differently.

Comparative neuroanatomy and cross-cultural cognitive research provide evidence that despite surface differences in language and culture, the fundamental architecture of human cognition is remarkably conserved [Braitenberg and Schüz, 1998, Dehaene, 2020]. Cortical organization, hippocampal structure, and basic memory mechanisms show striking similarity across populations. The words differ, the customs differ, but the underlying neural machinery of memory and reasoning is the same.

These observations led to a hypothesis: human cognition operates as three distinct but interacting systems. First, a **memory substrate** that stores patterns—multimodal, associative, and incredibly fast. Second, a **thought formation** process that activates and combines these patterns

into coherent internal states, still without words. Third, a **linguistic interface** that converts internal states into sequential speech—the slowest part, but the only part visible to others.

Large language models excel at that third component. They are brilliant verbalizers. But they lack the first two: explicit memory and non-linguistic thought formation. This work addresses that gap by building the missing pieces: a biologically grounded memory system and thought formation mechanism that can interface with LLMs for verbalization.

1.2 From Observation to Architecture

This project is an attempt to translate those observations into working code. The goal is not to compete with LLMs on benchmarks—they will win. The goal is to explore a different question: can we build a system that remembers like a brain, retrieves like a brain, and only then verbalizes using an LLM as a linguistic wrapper?

In biological systems, memory is not a collection of numerical parameters but a dynamic structure shaped by experience. Synaptic connectivity is continuously formed, stabilized, and pruned through local plasticity mechanisms, while inhibitory competition and neuromodulation regulate which patterns persist. These processes give rise to stable representations that can be flexibly recombined during thought and reasoning [Hebb, 1949, Markram et al., 1997].

The system described here models memory as a graph of discrete units whose connections evolve through qualitative states driven by local activity. Learning, recall, and forgetting emerge from structural change and activation dynamics, not from global objective functions.

1.3 Memory, Thought, and Language as Distinct Processes

The architecture investigated in this work is motivated by a functional separation commonly observed in cognitive neuroscience. Human cognition can be viewed as involving three interacting but distinct processes.

First, **memory** provides a fast, parallel substrate for storing experiences and knowledge. Biological memory is inherently multimodal, integrating sensory input, abstract concepts, and contextual information into unified representations.

Second, **thought formation** operates over this memory substrate. Before any linguistic output occurs, memories and concepts are dynamically reconfigured into a coherent internal state. This process is non-sequential, highly parallel, and largely unconscious. Note that "thought" here refers not to chain-of-thought reasoning or logical inference as commonly studied in AI, but to the underlying neural process by which the brain activates, combines, and stabilizes patterns of memory—the biological substrate of cognition itself.

Neurophysiologically, thought formation can be viewed as a process of spreading activation across neural assemblies, shaped by competition and inhibition. A partial cue may trigger activation of multiple candidate patterns, which then compete via lateral inhibition until a coherent pattern stabilizes. Similar dynamics have been formalized in attractor-like accounts of hippocampal and cortical recall [Rolls, 2007, 2013, Treves and Rolls, 1994].

Recall works by the same mechanism: a partial cue activates part of a stored pattern, and the remaining neurons are recruited through their established connections. This explains common memory phenomena—for example, recognizing a face but failing to recall the name corresponds to partial restoration of the underlying pattern: the face-related neurons activate successfully, but the weaker connections to name-related neurons do not drive full completion.

Third, **linguistic expression** transforms internal representations into linear sequences of words. Language is constrained by grammar and syntax and unfolds over time, making it fundamentally

different from the underlying representational processes.

Within this framework, language is not equated with thought. Instead, it serves as an interface for communicating internal states.

1.4 Relation to Large Language Models

Large language models excel at linguistic expression, the third component of the proposed architecture. They generate fluent text by modeling statistical regularities in language, but they do not explicitly store experiences as structured memory, nor do they possess a distinct mechanism for internal thought formation.

In the present work, large language models are therefore not treated as memory or reasoning substrates. Instead, they may optionally be used as output interfaces that render internal representations into grammatically correct natural language without contributing additional knowledge or influencing memory formation.

This distinction allows the memory system explored here to be evaluated independently of language generation performance. Critically, **test accuracy is evaluated on the Brain's raw semantic output**, not on LLM-polished text. The Broca's area module (local LLM) only assembles words into grammatically readable phrases—for example, transforming "stars appear sky night" into "The stars appear in the sky at night." This post-processing improves readability but does not change whether a test passes or fails.

This architectural separation mirrors a key observation about human cognition: we often understand something fully before we can articulate it. The process of "finding the right words" takes time and effort, suggesting that linguistic expression is computationally distinct from the underlying thought. In our implementation, the Brain module produces semantic content (the "thought"), while an optional LLM acts as a Broca's area analogue, transforming raw semantic output into grammatical speech. The LLM adds no new knowledge—it merely verbalizes what the memory system has retrieved.

This observation motivated the present research: if LLMs excel at verbalization but lack explicit memory and retrieval mechanisms, then building those missing components—rather than scaling language models further—may be a more direct path toward systems that truly remember and reuse knowledge.

Importantly, the current strength of LLMs is an advantage for this research direction. High-quality linguistic realization is readily available as an external component, enabling the present work to focus on memory, consolidation, and retrieval mechanisms.

1.5 Trust and Source Attribution

Another observation from introspection: not all information is treated equally by the brain. Facts learned from parents or teachers feel different from things overheard in conversation. School knowledge carries authority; gossip does not. The brain appears to tag information with its source and adjust consolidation accordingly.

The system implements this through explicit **source attribution**. Information can arrive from different channels: LEARNING (formal instruction, like school), CONVERSATION (casual interaction), or DIRECT_INPUT (manually verified facts—treated as maximally trustworthy). Each source type affects how quickly connections strengthen and whether episodes consolidate.

For example, the bAbI reasoning tasks [Weston et al., 2015] are processed as CONVERSATION—transient context that can be overwritten. Curriculum facts are processed as LEARNING—stable

knowledge that resists interference. This mirrors how a child might forget a playground rumor but retain multiplication tables for decades.

1.6 Scope and Motivation

The goal of this project is not to compete with existing machine learning systems on standard benchmarks, nor to propose a complete model of human cognition. Rather, it explores a specific question: can a memory-first architecture, grounded in biologically plausible mechanisms, support learning, recall, and question answering from text without gradient-based optimization?

This work does not claim superiority over gradient-based approaches. Large language models have demonstrated remarkable capabilities, and the present system cannot match their scale or fluency. The motivation here is different: to explore an *alternative* path—one that may offer interpretability, biological plausibility, and a different understanding of how memory and knowledge might be organized.

Whether this approach will scale, lead to emergent capabilities, or represent a viable long-term research direction remains unknown. But the question feels worth asking: if biological memory works through local plasticity and discrete structural change rather than global optimization, can we build systems that work the same way? This project is an attempt to find out.

1.7 Invitation to Collaborate

This work shares observations from introspection and a project that grew out of those observations. Part of the motivation for publishing is to find collaborators—researchers with deeper expertise in neuroscience, cognitive science, or AI—who might find this direction worth exploring together.

The code is open source. Contributions, critiques, and alternative approaches are welcome.

The following sections describe the design of the proposed memory system, its activation dynamics, hippocampus-inspired episodic indexing, and empirical evaluation on a controlled curriculum of basic world knowledge.

2 Related Work

The present work draws on and differs from several lines of research in machine learning and computational neuroscience.

Memory-Augmented Neural Networks. Memory Networks [Weston et al., 2014], Neural Turing Machines [Graves et al., 2014], and Differentiable Neural Computers [Graves et al., 2016] augment neural networks with external memory modules accessed through attention mechanisms. These systems use gradient-based optimization and continuous-valued addressing. In contrast, the present work uses discrete synaptic states and strictly local plasticity rules without backpropagation.

Spiking Neural Networks. Spiking neural networks (SNNs) model neurons as discrete event generators, enabling temporal coding and spike-timing dependent plasticity [Maass, 1997, Gerstner and Kistler, 2002]. The present system incorporates Hodgkin–Huxley dynamics and STDP but emphasizes discrete structural states (NEW/USED/MYELINATED/PRUNE) rather than continuous weight adaptation.

Sparse Distributed Memory. Kanerva's Sparse Distributed Memory [Kanerva, 1988] models associative retrieval through high-dimensional sparse representations. The hippocampal indexing mechanism in our system shares this emphasis on sparse encoding (dentate gyrus) but adds attractor dynamics for pattern completion (CA3) and replay-driven consolidation.

Knowledge Graphs. Traditional knowledge graphs store explicit relational triples but lack learning dynamics. The present system can be viewed as a knowledge graph with biologically motivated plasticity: connections strengthen, weaken, and prune based on usage patterns rather than explicit curation.

Computational Models of Hippocampus. The hippocampal indexing theory [Teyler and DiScenna, 1986] and complementary learning systems framework [McClelland et al., 1995] propose that hippocampus provides rapid encoding while neocortex supports slow consolidation. The present implementation follows this division, with episodic indexing via DG/CA3 and cortical consolidation through replay.

The key distinction of this work is the combination of (1) discrete synaptic states rather than continuous weights, (2) strictly local plasticity without global optimization, and (3) explicit separation of memory, thought formation, and linguistic expression.

3 Model Overview

The proposed system implements memory as a directed graph of discrete units and associations whose structure evolves over time. In contrast to parameter-optimized neural architectures that rely on real-valued weight adaptation, learning in the proposed system is expressed through qualitative transitions in the state of connections driven by local activity.

The current implementation integrates spiking neural dynamics with discrete-state structural memory. Neurons follow Hodgkin–Huxley dynamics with spike-timing dependent plasticity (STDP), while connections maintain discrete qualitative states. Brain oscillations (theta/gamma) and neuromodulation (dopamine, acetylcholine) further modulate learning and retrieval. Pattern completion is performed by a dedicated CA3 module implementing iterative attractor dynamics with lateral inhibition [Rolls, 2007, 2013].

The model is designed around three core requirements: (1) strictly local plasticity, (2) absence of global optimization objectives, and (3) biologically motivated competition and stabilization mechanisms.

A strict architectural boundary (PlasticityMode) separates learning from inference, ensuring that question answering does not modify long-term memory structures.

All system behavior—learning, recall, forgetting, and question answering—emerges from the interaction of these mechanisms.

3.1 Core Constraints

To maintain biological plausibility and conceptual clarity, the following design choices are enforced throughout the system:

- No gradient descent or backpropagation.
- No real-valued synaptic weights as carriers of semantic meaning.
- No embedding geometry or distance-based similarity measures; retrieval relies on discrete structural criteria rather than vector distances.
- No parameterized attention mechanisms; attentional selection emerges from competition, inhibition, and structural priority.

All computations are local, discrete, and structurally grounded.

3.2 System Architecture

Figure 1 presents a high-level view of the system components and their interactions.

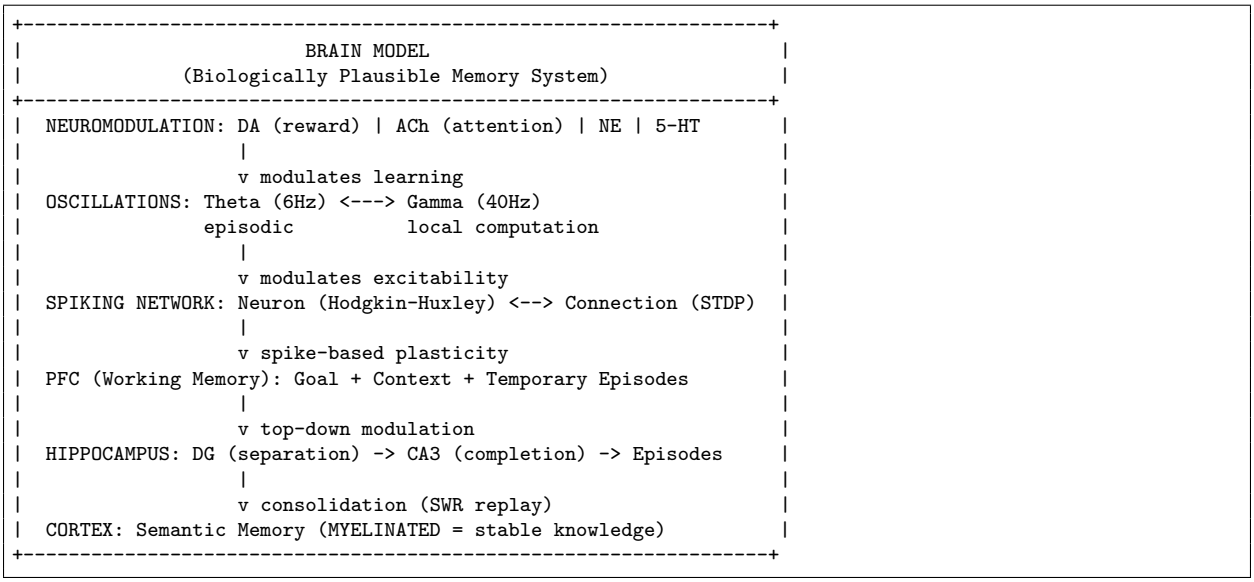

Figure 1: High-level architecture showing the flow from neuromodulation through spiking dynamics, working memory, episodic indexing, and semantic consolidation.

4 Representations

4.1 Neurons

Each distinct word form observed during training is represented as a neuron. Neurons follow Hodgkin–Huxley dynamics with continuous membrane potential, gating variables for ion channels (Na^+, K^+), and spike generation when voltage exceeds threshold [Hodgkin and Huxley, 1952]. Parameters are adapted from standard cortical neuron models rather than the original squid giant axon values. Despite this continuous internal state, the functional output remains discrete: a neuron either fires a spike or does not. Neurons maintain directed incoming and outgoing connections.

To reflect biological constraints, neurons enforce a fan-out limit on outgoing connections. In the current implementation, this limit is set to approximately 7,000 outgoing connections per neuron, consistent with cortical connectivity estimates [Braitenberg and Schüz, 1998].

While the present system operates on textual input, the representation is modality- agnostic. In principle, neurons may represent visual features, audio patterns, or other sensory elements, enabling extension to multimodal memory.

4.2 Connections and Discrete Synaptic States

Associations between neurons are represented as directed connections. Each connection exists in one of four qualitative states:

- **NEW**: recently formed, unstable connection.

- **USED**: connection repeatedly involved in successful activation.
- **MYELINATED**: structurally stabilized connection with priority during activation.[1]
- **PRUNE**: connection marked for removal due to inactivity.

Connections do not carry real-valued weights. Instead, state transitions are driven by local usage counters reflecting how often a connection participates in activation sequences.

Directional information is preserved through separate forward and backward usage counters, providing an STDP-like bias toward causal ordering [Markram et al., 1997].

4.3 Spike-Timing Dependent Plasticity

Beyond discrete state transitions, the system implements spike-timing dependent plasticity (STDP), a biologically grounded learning rule that captures the causal structure of neural activity [Markram et al., 1997].

When neuron A fires before neuron B (pre-before-post, $\Delta t > 0$), the connection $A \to B$ undergoes long-term potentiation (LTP). When the order is reversed ($\Delta t < 0$), long-term depression (LTD) occurs. The magnitude follows an exponential decay:

$$\Delta w \propto \exp\left(-\frac{|\Delta t|}{\tau}\right), \quad \tau = 20\text{ms}$$

This asymmetric learning window ensures that connections encode predictive relationships: if A reliably precedes B, the $A \to B$ pathway strengthens, enabling pattern completion from partial cues. The system maintains spike histories for each neuron, allowing precise timing-based updates during learning.

Integration with training: During learning, the `_simulate_spike_pair()` function in `train.py` generates spike pairs for each word connection. Pre-spike occurs at global time t, post-spike at $t + 5$ms, producing strong LTP ($\exp(-5/20) \approx 0.78$). Each connection uses `EligibilityTrace`, `CalciumState`, and `MetaplasticState` from `spiking.py` for biologically accurate plasticity.

4.4 Three-Factor Learning and Eligibility Traces

Pure STDP faces the temporal credit assignment problem: how can a synapse that was active seconds ago be strengthened when a reward arrives later? Biological systems solve this through **eligibility traces**—temporary synaptic tags that mark recently active synapses as candidates for modification [Gerstner et al., 2018].

The system implements three-factor learning: (1) presynaptic activity, (2) postsynaptic activity, and (3) a neuromodulatory signal. When STDP occurs, it creates an eligibility trace rather than immediate weight change. The trace decays exponentially ($\tau \approx 1$s). When dopamine arrives (signaling novelty or reward), it converts eligibility traces into permanent structural changes.

This mechanism enables the system to associate actions with delayed outcomes—critical for learning from text where the relevance of a word may only become clear sentences later.

Figure 2 illustrates the four-factor learning system.

[1]The term "myelinated" is used metaphorically to denote functional stability and transmission priority, analogous to how biological myelination increases axonal conduction speed. The model does not simulate glial wrapping; the state represents a connection that has been consolidated through repeated use.

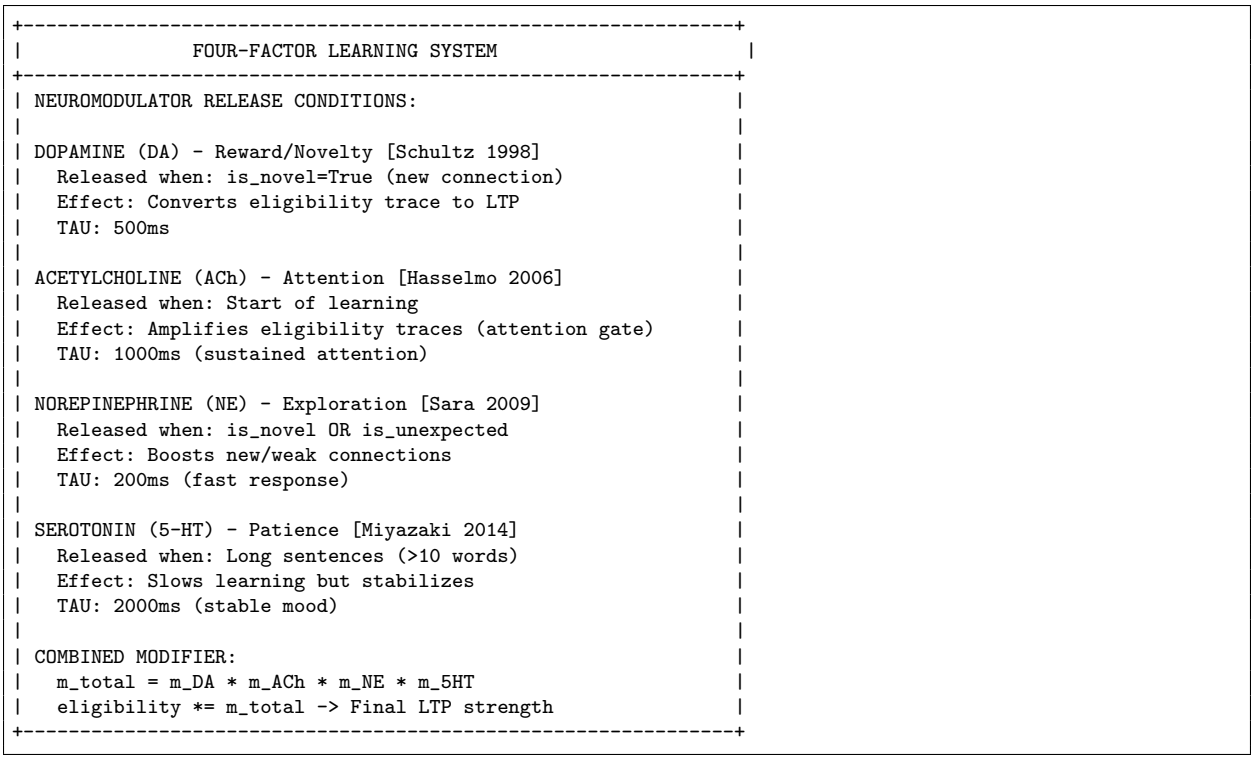

Figure 2: Four-factor learning: dopamine (novelty/reward), acetylcholine (attention gate), norepinephrine (exploration), and serotonin (patience) multiplicatively modulate eligibility traces.

4.5 Synaptic Tagging and Capture

To model cooperative stabilization effects observed in biological systems, the model optionally implements a synaptic tagging and capture mechanism [Frey and Morris, 1997]. When a connection becomes myelinated, nearby connections within the same local context experience a reduced threshold for myelination, facilitating the formation of coherent memory structures.

This mechanism supports rapid consolidation of related associations without introducing global coordination or learned parameters.

4.6 Semantic and Syntactic Streams

Inspired by dual-stream accounts of language processing [Saur et al., 2008], the system distinguishes between two classes of connections:

- **Semantic connections**, linking content-bearing words and concepts.
- **Syntactic connections**, involving function words and structural markers.

Semantic connections primarily support meaning-based retrieval, while syntactic connections capture ordering and grammatical structure. During memory retrieval and question answering, spreading activation is restricted to semantic connections, preventing purely syntactic associations from driving semantic inference.

5 Learning from Text

Learning proceeds incrementally as text is presented to the system. Given an input sentence, the following steps are performed:

1. The sentence is tokenized into word forms.
2. Each token is classified as a content word, function word, or interrogative.
3. Neurons are created for previously unseen word forms.
4. Directed connections are created or updated within a temporal window corresponding to working-memory constraints [Miller, 1956].

Here, "epochs" denote repeated exposure to the same curriculum, not gradient-based optimization.

5.1 Local Hebbian-Style Updates

For each ordered pair of tokens (w_i, w_j) within the temporal window, a directed connection $w_i \to w_j$ is created if absent, and its forward usage counter is incremented, consistent with Hebbian-style association formation [Hebb, 1949].

This simple local rule yields directional associations that support ordered recall and structured traversal during retrieval, without requiring global supervision.

5.2 Context-Modulated Plasticity

To approximate top-down modulation observed in prefrontal cortical systems, the model includes a context-sensitive reinforcement mechanism, motivated by evidence for PFC-driven control of sensory processing and working memory [Miller and Cohen, 2001, Zanto et al., 2011]. When a broader sentence context already forms a coherent pattern, usage of connections within that context is locally amplified.

Importantly, this amplification is implemented as additional local usage events on the same connections, not as learned attention weights or external control signals.

NMDA Receptor Mechanism: A key challenge in context-modulated learning is the "cold start" problem: newly formed connections have low usage counts and would not participate in context-based reinforcement under a fixed threshold. The system addresses this through an NMDA receptor-inspired mechanism [Malenka and Bear, 2004].

In biological synapses, NMDA receptors are blocked by Mg^{2+} ions at resting membrane potential. When the postsynaptic neuron is sufficiently depolarized (approximately -40mV), the Mg^{2+} block is relieved, allowing even weak synapses to participate in long-term potentiation (LTP). The system implements this as a dynamic threshold: when four or more context neurons are simultaneously active (indicating strong network activation), the minimum usage threshold for context cache inclusion drops from 3 to 1. This allows newly formed connections to participate in context-based learning when overall activity is high, consistent with the biological principle that weak synapses can undergo LTP when the postsynaptic neuron is already strongly activated.

5.3 Top-Down Attentional Modulation

During retrieval, the system implements biologically grounded top-down modulation based on task demands. This mechanism is motivated by two key findings:

Zanto et al. (2011) demonstrated that prefrontal cortex causally modulates sensory processing via top-down signals that enhance task-relevant stimuli while suppressing irrelevant ones [Zanto et al., 2011].

Desimone & Duncan (1995) proposed the Biased Competition theory, where attention operates through competitive interactions—enhancing some representations at the expense of suppressing others [Desimone and Duncan, 1995].

The implementation uses *multiplicative* gain modulation rather than additive bonuses, consistent with neurophysiological evidence. When a query specifies a particular relation type (e.g., "What IS X?" implies category membership), connections with matching relational markers receive a gain factor of 5.0, while non-matching connections receive a suppression factor of 0.2. This creates soft competitive dynamics where relevant associations dominate without hard-coded rules.

5.4 Chunk Formation

When a neuron develops a dominant myelinated semantic forward connection, the system may form a new chunk neuron representing a frequently co-occurring sequence (e.g., "united_states").

Chunking reduces representational complexity and mirrors cognitive chunk formation observed in human expertise [Chase and Simon, 1973], while remaining grounded in local structural conditions.

Figure 3 illustrates the learning pipeline from input text to episodic encoding.

6 Activation Dynamics and Attention

Inference and recall in the proposed system are governed by local spreading activation combined with biologically motivated competition and inhibition. Rather than computing attention weights or similarity scores, the system relies on structural priority, inhibition, and decay to select coherent activation patterns.

6.1 Spreading Activation

Activation is initiated by external input (e.g., query terms) or internal contextual patterns. Active neurons propagate activation locally through their outgoing connections according to the following principles:

- Activation propagates only through existing connections.
- Connections in the MYELINATED state propagate activation with highest priority.
- USED connections propagate activation normally.
- NEW connections propagate weakly and are easily suppressed.
- PRUNE connections do not propagate activation.

Propagation is asynchronous and local; no global traversal or evaluation of the memory graph is performed.

6.2 Inhibitory Competition

To prevent uncontrolled spread and enforce coherence, the system employs inhibitory competition analogous to cortical inhibitory interneuron dynamics. As activation spreads, inhibitory units suppress weakly supported activations, mutually incompatible branches, and diffuse activity patterns. This produces winner-take-most dynamics consistent with attractor-like selection [Rolls, 2007, Treves and Rolls, 1994].

Episode selection uses discrete structural scoring based on connection states and overlap counts, not learned parameters or continuous similarity metrics.

6.3 Biological Attention Mechanism

Attention in the proposed model is not a separate module but an emergent property of activation dynamics. Attention arises from the interaction of structural priority (myelination), contextual bias from active patterns, and inhibitory suppression of competitors. This serves a functional role analogous to attentional selection without requiring learned attention parameters.

6.4 Spread Attention and Associative Recall

Beyond focused activation, the system supports controlled associative expansion. When inhibition is partially relaxed, activation may spread from a dominant pattern to structurally related patterns, enabling associative recall and context-sensitive exploration of memory.

6.5 Working Memory Constraint

To model limitations observed in biological cognition, the system enforces a working memory constraint on the number of simultaneously active neurons. This constraint typically results in approximately 7 ± 2 active elements [Miller, 1956].

6.6 Prefrontal Cortex and Persistent Activity

Biological working memory requires more than capacity limits—it requires *active maintenance* of representations across time. In the prefrontal cortex (PFC), neurons exhibit **persistent activity**: sustained firing that maintains information even without continued sensory input [Wang, 2001, Compte et al., 2000]. This is the neural substrate of "keeping something in mind."

The proposed system implements a PFC module with three biologically grounded mechanisms:

(1) NMDA-mediated slow decay. NMDA receptors have markedly slower kinetics than AMPA receptors ($\tau_{\text{NMDA}} \approx 100$ms vs $\tau_{\text{AMPA}} \approx 5$ms) [Lisman et al., 1998]. This slow decay acts as a "synaptic memory" that sustains activity between spikes. The system implements blended decay (30% AMPA + 70% NMDA), allowing representations to persist longer than with fast decay alone.

(2) Recurrent excitation. Pyramidal neurons in PFC form recurrent excitatory connections that create positive feedback loops [Wang, 2001]. When a representation is active, it reinforces itself through these recurrent connections, creating *attractor dynamics*—bistable states that "lock in" active patterns. In the implementation, PFC slots that share content (e.g., "john garden" and "john football") mutually boost each other's activation, making related information more resistant to decay.

(3) Distractor resistance via inhibitory gating. Active working memory must resist interference from irrelevant stimuli. GABAergic interneurons create an inhibitory barrier around active representations [Miller and Cohen, 2001]. New inputs must either be goal-relevant (receiving

top-down facilitation) or sufficiently salient to overcome this barrier. This explains why we can ignore distractions while holding a phone number in mind, yet attend immediately to our name being called.

These mechanisms work together: NMDA provides the temporal bridge between inputs, recurrent excitation stabilizes active patterns into attractors, and inhibitory gating protects these attractors from noise. The result is a working memory that can maintain "John is in the garden" across multiple intervening sentences, correctly answering "Where is John?" even after processing unrelated information.

On the bAbI Task 1 benchmark (single supporting fact with distractor sentences), this architecture achieves 100% accuracy (250/250) without any task-specific training—the same mechanisms that maintain any working memory representation naturally handle the task structure.

6.7 Hub Suppression and Weber–Fechner Scaling

Highly connected neurons (hubs) pose a risk of dominating activation due to their degree alone. The system applies a hub penalty inspired by Weber–Fechner-like compression, reducing the dominance of high-degree nodes and supporting context-specific recall.

6.8 Decay and Stability

Activation decays over time unless reinforced by structural support. Stable patterns emerge when recurrent activation, myelinated pathways, and contextual coherence reinforce one another.

6.9 Neuromodulation System

Biological memory is not simply a matter of connection strength—it is dynamically regulated by neuromodulators that gate learning, modulate excitability, and signal behavioral relevance. The system implements four neuromodulators, each with distinct release conditions and effects:

Dopamine (DA) signals novelty and reward [Schultz, 1998]. When a new connection forms or an unexpected pattern is encountered, dopamine is released. This converts eligibility traces into permanent structural changes (see Three-Factor Learning). Without dopamine, STDP creates only temporary eligibility; with dopamine, learning is consolidated. This explains why surprising or rewarding experiences are remembered better.

Acetylcholine (ACh) gates attention and encoding [Hasselmo, 2006]. Released at the onset of learning (when a new sentence is presented), ACh amplifies eligibility traces and increases the probability that active synapses will be modified. High ACh favors encoding new information; low ACh favors retrieval of existing memories. The system modulates ACh based on task context.

Norepinephrine (NE) signals arousal and surprise [Sara, 2009]. Released when input is novel or unexpected, NE increases neuronal excitability and biases the system toward exploration of new pathways rather than exploitation of established ones. This enables flexible learning when the environment changes.

Serotonin (5-HT) promotes behavioral inhibition and patience [Miyazaki et al., 2014]. Released during sustained learning (long sentences, complex material), 5-HT stabilizes plasticity but slows its rate. This implements temporal discounting: immediate associations are valued more than distant ones.

The combined learning modifier $m_{total} = m_{DA} \times m_{ACh} \times m_{NE} \times m_{5HT}$ gates all plasticity, ensuring that learning occurs only when appropriate neuromodulatory conditions are met.

6.10 Brain Oscillations

Neural activity in biological systems is organized by rhythmic oscillations that coordinate processing across brain regions [Buzsáki, 2006]. The system implements two key rhythms:

Theta rhythm (6 Hz) dominates during episodic memory encoding and retrieval. Theta provides a temporal framework for organizing sequential information—each theta cycle can encode one "item" in a sequence. During learning, theta modulates neuron excitability such that inputs arriving at different theta phases receive different processing.

Gamma rhythm (40 Hz) supports local computation and feature binding. Nested within theta cycles, gamma oscillations coordinate the firing of neurons representing related features. Theta-gamma coupling enables sequence encoding: different items occupy different gamma cycles within a theta cycle, preserving temporal order [Lisman and Idiart, 2005].

The `BrainOscillator` class generates these rhythms and modulates neuronal excitability during spike simulation. This ensures that spiking activity respects biological timing constraints rather than operating in abstract computational time.

6.11 Spiking Neural Network Architecture

The system implements a complete spiking neural network in `spiking.py` with biologically accurate components:

SpikingNeuron: Hodgkin–Huxley neuron model with membrane potential dynamics, ion channel gating variables (Na^+, K^+), action potential generation, and refractory periods. Each neuron maintains spike history for STDP computation.

Synapse: Base synapse class with spike propagation, synaptic delay, and classical STDP. Serves as foundation for specialized synapse types.

SpikingNetwork: Container for neurons and synapses, manages simulation stepping, oscillator updates, and coordinated STDP application.

6.12 Synapse Types

The system implements multiple biologically grounded synapse types, each capturing different aspects of synaptic plasticity:

ThreeFactorSynapse: Implements eligibility traces that bridge the temporal credit assignment gap [Gerstner et al., 2018]. STDP creates eligibility traces rather than immediate weight changes; neuromodulator signals convert traces to permanent plasticity.

BCMSynapse: Bienenstock–Cooper–Munro rule with sliding threshold [Bienenstock et al., 1982]. Firing rate determines plasticity direction: high rates induce LTP, low rates induce LTD. Threshold adapts based on activity history, providing homeostasis.

STPSynapse: Short-term plasticity with facilitation and depression on millisecond timescales [Tsodyks and Markram, 1997]. Facilitating synapses (low initial release probability, strong facilitation) model cortical–cortical connections; depressing synapses (high release, fast depletion) model thalamo–cortical connections.

DendriticSynapse: Location-dependent plasticity based on dendritic compartment [Larkum, 2013]. Proximal synapses show standard Hebbian STDP; distal synapses exhibit anti-Hebbian plasticity under low cooperativity. Back-propagating action potentials attenuate with distance from soma.

MetaplasticSynapse: Sliding threshold metaplasticity [Abraham and Bear, 1996]. Recent LTP increases threshold for future LTP; recent LTD increases threshold for future LTD. This prevents runaway plasticity and implements synaptic homeostasis.

CalciumBasedSynapse: Calcium-concentration-based plasticity [Graupner and Brunel, 2012]. Low Ca^{2+} produces no change; medium Ca^{2+} activates phosphatases (LTD); high Ca^{2+} activates CaMKII (LTP). NMDA receptors provide coincidence detection with supralinear calcium for pre–post timing.

AntiHebbianSynapse: Reversed STDP polarity for specific inhibitory connections [Feldman, 2012]. Pre-before-post produces LTD; post-before-pre produces LTP. Observed at excitatory inputs onto fast-spiking interneurons.

BiologicalSynapse: Comprehensive synapse combining all mechanisms: Hodgkin–Huxley dynamics, STDP, three-factor learning, short-term plasticity, dendritic location effects, metaplasticity, and calcium dynamics. This is the most complete biological model available in the system.

6.13 Supporting Data Structures

EligibilityTrace: Exponentially decaying synaptic tag ($\tau \approx 1$s) marking recently active synapses as candidates for modification.

STPState: Short-term plasticity state tracking release probability (u) and available resources (x). Effective transmission $= w \times u \times x$.

CalciumState: Calcium concentration at synapse with separate tracking of pre- and postsynaptic spike times for NMDA coincidence detection.

MetaplasticState: History of recent LTP/LTD events with sliding thresholds for future plasticity.

DendriticCompartment: Dendritic location (proximal, distal, apical, basal) with location-specific attenuation and STDP type.

SpikeRecord: Individual spike record (neuron ID, time) for precise timing-based computations.

6.14 Source Memory

The system implements source memory monitoring based on findings that the brain remembers not just what was learned but where and how it was acquired [Johnson et al., 1993]. Each episode stores a source type (LEARNING, EXPERIENCE, CONVERSATION, MEDIA) with associated trust levels. During retrieval, the prefrontal cortex classifies the question type and routes retrieval toward appropriate source types. This enables the system to prefer school-learned facts for semantic questions while favoring experiential memories for cause-effect questions.

6.15 Prefrontal Cortex Module

The `pfc.py` module implements working memory and executive control:

PFC: Central working memory buffer with limited capacity ($\sim$7 slots). Maintains goal state, context, and temporary episodes. Implements NMDA-like slow decay and recurrent excitation for persistent activity.

PFCSlot: Individual working memory slot with activation level, content, and decay dynamics. Different slot types (GOAL, CONTEXT, FACT) serve different cognitive functions.

AttentionGate: Top-down attentional control based on PFC goal state. Provides multiplicative gain modulation to enhance task-relevant stimuli and suppress distractors [Desimone and Duncan, 1995].

MemoryRouter: Routes incoming information to appropriate memory systems based on source type and question classification.

ThinkingEngine: Spontaneous activation and associative exploration, modeling mind-wandering and creative thinking through neural noise.

InferenceEngine: Inference through spreading activation in the trained network, using working memory context to guide retrieval.

IterativeRetriever: Multi-step reasoning via PFC-hippocampus loop. Maintains goal state, iteratively queries hippocampus, accumulates context, and terminates when goal is reached [Eichenbaum, 2017].

RetrievalResult: Container for retrieval outcomes including retrieved episode, confidence, and reasoning trace.

QuestionType: Classification of questions (FACTUAL, CAUSAL, TEMPORAL, etc.) for source-appropriate routing.

6.16 Lexicon Module

The `lexicon.py` module implements sensory-motor language pathways [Hickok and Poeppel, 2007]:

InputLayer: Sensory pathway from words to neurons. Maps auditory/visual word forms to internal representations, modeling the ventral "what" stream.

OutputLayer: Motor pathway from neurons to words. Maps internal representations to articulatory motor plans, modeling the dorsal "how" stream.

Lexicon: Combined interface managing bidirectional word↔neuron mapping. Implements the mental lexicon with both recognition and production pathways.

6.17 Activation Module

The `activation.py` module implements spreading activation dynamics:

Activation: Core spreading activation process. Propagates activity through existing connections with myelinated priority, lateral inhibition, decay, and working memory limits. No global graph search—activation spreads locally like neural activity.

NeuromodulatorSystem: Manages four neuromodulator levels (DA, ACh, NE, 5-HT) with release conditions, decay dynamics, and combined learning modifiers.

6.18 Graph Storage

The `graph_storage.py` module provides efficient storage for large-scale graphs:

GraphStorage: NumPy-based directed graph with connection states (NEW, USED, MYELINATED, PRUNE), connection types (SEMANTIC, SYNTACTIC), and efficient batch operations. Implements STDP-like directional tracking.

6.19 Training Modes

The `training_modes.py` module defines 20 training modes based on brain memory systems:

TrainingMode: Enumeration of declarative and procedural training modes including FACT, DEFINITION, HIERARCHY, PROPERTY, RELATION, SEQUENCE, EPISODE, PROCEDURE, ROUTINE, CAUSE_EFFECT, and others.

Each mode has corresponding data classes (FactData, DefinitionData, HierarchyData, etc.) that structure input for appropriate encoding.

6.20 Experience Events

The `experience.py` module handles experiential learning:

ExperienceEvent: Represents a single experience with timestamp, source, emotional valence, and associated neurons. Enables episodic encoding of lived experiences distinct from learned facts.

7 Episodic Memory: Hippocampus-Inspired Indexing

In addition to cortical-style associative memory, the system implements an episodic indexing mechanism inspired by hippocampal memory theories [Teyler and DiScenna, 1986, McClelland et al., 1995]. This module supports rapid encoding of experiences, pattern separation, pattern completion, and replay-driven consolidation.

7.1 Episode Data Structure

The `episode.py` module defines the core episodic trace:

Episode: Each episode stores: (1) `input_neurons`—original words as frozen set for search; (2) `input_words`—word order tuple preserving hippocampal time cell sequence; (3) `pattern_neurons`—sparse DG representation (∼2%); (4) `context_neurons`—what was active at encoding; (5) `timestamp`; (6) `source`—where/how learned; (7) `state`—current consolidation stage; (8) `replay_count`—times replayed during sleep.

EpisodeState: Episodes transition through consolidation stages: NEW (freshly encoded, fragile), REPLAYED (reinforced during sleep), CONSOLIDATED (stable, transferred to cortex), DECAYING (weakening due to disuse). This models hippocampal-cortical memory transfer.

7.2 Pattern Separation

To reduce interference between similar experiences, episodic representations are encoded using sparse activation patterns analogous to the Dentate Gyrus (DG). Following neurophysiological findings [Rolls, 2007], only approximately 2% of input neurons are selected for each episodic trace through competitive dynamics. This extreme sparsity increases orthogonality and aligns with theoretical analyses of pattern separation and memory capacity [Rolls, 2013, Treves and Rolls, 1994].

7.3 Pattern Completion via CA3 Attractor Dynamics

Given a partial cue, episodic retrieval proceeds through a CA3-inspired attractor network implemented as a separate module (`ca3.py`). Unlike simple argmax selection over episode lists, the system performs iterative dynamics until pattern stabilization [Rolls, 2007, 2013]:

1. **Initial activation**: Cue neurons receive activation value 1.0.
2. **Iterative spreading**: Activation spreads via recurrent collateral connections. MYELINATED connections transmit with strength 0.8, USED with 0.4, NEW with 0.1. Top-down modulation multiplies matching connector strengths by 5.0.
3. **Lateral inhibition**: Winner-Take-All dynamics retain only top-K neurons (default $K = 20$), suppressing weaker activations.
4. **Stability check**: If active neuron set equals previous iteration, attractor reached.

5. **Episode scoring**: Completed pattern is matched against candidate episodes using full scoring logic.

Episode scoring incorporates multiple biologically motivated factors:

- **Query overlap** (highest priority): Episodes containing query terms receive strong activation, modeling goal-directed top-down modulation from prefrontal systems [Miller and Cohen, 2001, Zanto et al., 2011, Desimone and Duncan, 1995].
- **Connection strength with context multiplier**: 1-hop and 2-hop paths weighted by connection state. Context words (query terms not in episode) receive 3× multiplier.
- **Top-down connector modulation**: Multiplicative enhancement (×5.0) for connections matching query relation type; suppression (×0.2) otherwise.
- **Context diversity bonus**: Connections appearing in multiple contexts receive $\log_2(\text{diversity})\times$ 2.0 bonus [Spens and Burgess, 2024].
- **Unconnected context filtering**: Query terms not in episode must be connected to episode contents; otherwise episode is rejected (anti-hallucination).
- **Recency bias**: Working memory episodes receive timestamp-based bonuses, with reverse recency for past-tense queries [Howard and Kahana, 2002].
- **Consolidation bonus**: CONSOLIDATED episodes receive priority over NEW episodes.
- **Divisive normalization**: Confidence threshold filters low-scoring episodes [Carandini and Heeger, 2012].

The CA3 module is instantiated as an explicit dependency of the Hippocampus class, not as a global singleton, following the architectural principle of no hidden global state.

7.4 CA1 Output Layer

After pattern completion in CA3, the retrieved pattern must be transformed for cortical output. This is the function of the CA1 region, the primary output layer of the hippocampus [Amaral and Witter, 1989, Naber et al., 2001].

The system implements CA1 as a feedforward readout layer with two input pathways:

(1) Schaffer collaterals (CA3→CA1): The main input pathway, weighted at 70%. MYELINATED connections in CA3 receive additional transmission bonus, reflecting faster conduction in myelinated axons.

(2) Temporoammonic pathway (EC→CA1): Direct input from entorhinal cortex Layer III, weighted at 30%. This pathway bypasses CA3 and provides current query context, enabling coincidence detection when both CA3 and EC activate the same CA1 neurons.

CA1 output projects to two targets:

- **Entorhinal cortex Layer V**: Main output to neocortex for consolidation.
- **Prefrontal cortex**: Direct projection supporting working memory maintenance and multi-step reasoning [Preston and Eichenbaum, 2013].

This trisynaptic circuit (EC→DG→CA3→CA1→EC/PFC) implements the complete hippocampal processing loop described in neuroanatomical studies [Amaral and Witter, 1989].

7.5 Replay and Consolidation

During periods of low external input (simulated sleep), episodic traces are replayed internally via Sharp Wave-Ripples (SWR), a biologically accurate mechanism observed in hippocampal recordings [Buzsáki, 2015]. The system implements several key features of SWR:

Temporal Compression: Replay occurs 10–20× faster than original encoding [Nádasdy et al., 1999]. In the current implementation, inter-spike intervals are compressed by a factor of 15, transforming 100ms encoding intervals into ∼6.7ms replay intervals.

Forward and Reverse Replay: Approximately 30% of replays occur in reverse temporal order [Diba and Buzsáki, 2007], supporting planning and backward chaining in addition to consolidation.

NREM/REM Sleep Phases: The system alternates between NREM cycles (SWR-driven consolidation) and REM cycles (random reactivation for memory integration) [Born and Wilhelm, 2012].

Stochastic Episode Selection: Episodes are selected for replay probabilistically based on recency and importance, consistent with findings that not all memories replay each night [Wilson and McNaughton, 1994].

Synaptic Homeostasis: After sleep, global synaptic downscaling preserves relative connection strengths while reducing overall synaptic load [Tononi and Cirelli, 2006].

Cross-Episode Linking: During REM sleep, the system identifies episodes sharing common elements and creates semantic connections between their unique components [McClelland et al., 1995, Kumaran and McClelland, 2012]. For example, if Episode 1 contains {dog, animal} and Episode 2 contains {cat, animal}, the shared element "animal" triggers co-activation, creating a direct connection dog↔cat. This mechanism implements the Complementary Learning Systems theory: the hippocampus "teaches" the neocortex by extracting statistical regularities across episodes, transforming episodic memories into semantic knowledge. The biological basis is that overlapping neural representations during replay cause Hebbian strengthening between previously unconnected neurons that share context.

Frequently replayed episodes reinforce corresponding cortical structures via local plasticity, supporting complementary learning-system behavior [McClelland et al., 1995]. Episodes that are not replayed decay over time, providing a natural forgetting mechanism. Replay and consolidation are consistent with modern accounts of memory construction and consolidation [Spens and Burgess, 2024].

Figure 4 illustrates the sleep consolidation process.

7.6 Homeostatic Plasticity Mechanisms

To maintain network stability and efficient storage, the system implements biologically motivated homeostatic mechanisms:

Heterosynaptic LTD [Rolls, 2013]: When one connection strengthens, neighboring connections on the same neuron weaken, preventing runaway potentiation.

Synaptic Scaling [Turrigiano, 2008]: Neurons stabilize activity levels via homeostatic adjustment of outgoing efficacy.

Predictive Coding [Rao and Ballard, 1999]: Already-myelinated associations represent expected transitions and receive no additional strengthening; learning focuses on prediction errors.

Competitive Learning in DG: Pattern separation employs winner-take-all dynamics shaped by existing structure, consistent with sparse coding mechanisms [Rolls, 2007].

Diluted Connectivity: Connections are formed only within a limited temporal window (4 words), reflecting biological constraints where neurons connect primarily to nearby cells rather

than forming fully-connected networks [Rolls, 2007].

Lateral Inhibition: During retrieval, query words do not receive activation bonuses from themselves (self-inhibition), and strongly activated neurons suppress weakly activated competitors, consistent with cortical inhibitory interneuron dynamics.

7.7 Developmental Phases

The system implements biologically motivated developmental stages that affect plasticity and pruning throughout learning [Hensch, 2005, Hubel and Wiesel, 1970].

Critical Periods: Windows of heightened plasticity for specific learning types. The system tracks four developmental stages:

- **INFANT**: High plasticity ($\times 2.0$), no pruning, all critical periods active.
- **CHILD**: Moderate plasticity ($\times 1.5$), light pruning begins, language/semantic/syntactic critical periods active.
- **ADOLESCENT**: Normal plasticity, aggressive pruning (threshold 5), only semantic critical period remains.
- **ADULT**: Reduced plasticity ($\times 0.8$), maintenance pruning, all critical periods closed.

Experience-Expectant Plasticity [Greenough et al., 1987]: During critical periods, specific connection types receive learning bonuses. Syntactic connections receive $\times 1.5$ bonus during language critical period; after closure, learning becomes harder ($\times 0.5$).

Synaptic Pruning [Huttenlocher, 1979]: Unused connections are eliminated following the "use it or lose it" principle. Pruning peaks during ADOLESCENT stage, removing connections below usage threshold. MYELINATED connections are protected from pruning, modeling the stability of well-established pathways.

PV Interneuron Maturation [Hensch, 2005]: Inhibition level increases with development (0.3 in INFANT $\rightarrow$ 1.0 in ADULT), modeling the maturation of parvalbumin-positive interneurons that close critical periods.

8 Question Answering

Question answering is implemented as a constrained activation and retrieval process. Given a question, the system performs the following steps:

1. Extract content terms and interrogative structure.
2. If required content terms are unknown, return an explicit "I do not know" response.
3. **Basal ganglia selects action strategy** (single retrieval vs multi-hop reasoning).
4. Initiate spreading activation through semantic connections.
5. Use the resulting activation pattern as a cue for episodic retrieval.
6. Apply pattern completion with structural gating.
7. Traverse the retrieved episode to produce a raw semantic answer.

8. **Motor output (Broca's area analogue)** renders answer in correct word order.

9. Optionally render the answer into grammatical language using an external language model acting purely as an output interface.

Episode ranking is based on discrete structural properties (connection states, overlap counts, context connectivity) rather than learned similarity scores or embedding distances.

Figure 5 illustrates the question answering pipeline.

8.1 Basal Ganglia Action Selection

Before retrieval begins, the system must decide *how* to answer: should it perform a single memory retrieval, or engage in multi-hop reasoning that chains multiple retrievals? This decision is made by a basal ganglia circuit that implements biologically grounded action selection [Frank et al., 2006, Mink, 1996].

The basal ganglia receive cortical input (question salience, familiarity) and neuromodulatory signals, then select among competing actions through a Go/NoGo mechanism:

D1 (Go) pathway: When dopamine activates D1 receptors in the striatum, it facilitates the currently most active action representation, biasing the system toward executing that action.

D2 (NoGo) pathway: D2 receptor activation inhibits competing actions, implementing a "gate" that prevents premature or inappropriate responses.

STN hyperdirect pathway: The subthalamic nucleus provides a fast "emergency brake" that can halt all actions when uncertainty is high, preventing impulsive errors.

GPi/GPe tonic inhibition: The globus pallidus maintains tonic inhibition of thalamic targets; action selection occurs when this inhibition is transiently released for the winning action.

In the implementation, basal ganglia select between "retrieve" (single-hop) and "multi_hop" strategies based on question complexity and neuromodulator state. Complex questions with multiple entities trigger multi-hop reasoning; simple factual questions use direct retrieval.

8.2 Motor Output and Sequence Generation

Retrieving the correct information is not sufficient—the answer must be produced in correct word order. This is the function of motor output areas, particularly Broca's area for speech production [Hickok and Poeppel, 2007].

The system implements a `SequenceGenerator` class that preserves the temporal order stored in hippocampal time cells. When an episode is retrieved, its `input_words` tuple maintains the original sequence in which words were encoded. The motor output module reads this sequence and generates the answer in the correct order.

This separation mirrors the biological distinction between *knowing* something (hippocampal/cortical retrieval) and *saying* something (motor planning and execution). A person may know that Paris is the capital of France, but producing the sentence "The capital of France is Paris" requires additional motor sequencing.

For complex answers, the sequence generator also handles:

- Phrase coherence: keeping related words together (e.g., "salt water" not "water salt")
- Temporal ordering: respecting the order in which events occurred
- Grammatical scaffolding: providing structure for optional LLM polishing

9 Experiments

9.1 Curriculum-Based Evaluation

The system was evaluated on a curated curriculum of basic world knowledge designed to approximate the conceptual scope of early childhood learning.

Statistic	Value
Neurons (word forms)	48,301
Connections	1,453,469
Myelinated connections	19,252 (1.3%)
Used connections	77,745 (5.3%)
New connections	1,356,472
Episodic traces	68,947
— NEW	35,157
— REPLAYED	2,139
— CONSOLIDATED	30,748
— DECAYING	903
FineWeb-Edu articles	1,000 (50K sentences)
Training pipeline	curriculum → preschool → grade1 → FineWeb
Curriculum QA accuracy	98% (49/50)
Strict (I do not know)	100% (3/3)
Preschool QA accuracy	95.8% (46/48)
Grade1 QA accuracy	100% (64/64)
FineWeb QA accuracy	77.8% (7/9)
bAbI Task 1 (working memory)	100% (250/250)
Total QA accuracy	98.8% (419/424)
[INFER-NO-LEARN] test	PASS (0 LTM changes)

Table 1: Model statistics (January 24, 2026 (v1.0)). Training pipeline includes source memory stages. PlasticityMode ensures inference does not modify long-term memory. Basal ganglia integrated for action selection. Broca's area module provides syntactic processing.

Sample test log output (from `logs/test_results_*.txt`, available in repository):

```
=== EXCELLENT (10/10) ===
[PASS] Q: What is the opposite of hot? [Brain: 0.001s | Raw:10 LLM:10]
       Brain raw: cold     Expected: ['cold']

[PASS] Q: What does a caterpillar become? [Brain: 0.336s | Raw:10 LLM:10]
       Brain raw: butterfly     Expected: ['butterfly']

[PASS] Q: Who wrote Hamlet? [Brain: 0.573s | Raw:9 LLM:9]
       Brain raw: I do not know     Expected: ['not know'] (correct rejection)

=== GOOD WITH GRAMMAR ISSUES (Broca area imperfect) ===
[PASS] Q: What is a puppy? [Brain: 0.412s | Raw:9 LLM:5]
       Brain raw: baby dog     LLM: An baby dog (grammar error)

[PASS] Q: Is a feather heavier than a rock? [Brain: 0.415s | Raw:6 LLM:5]
```

```
       Brain raw: rock heavier feather     Expected: ['heavier', 'rock']

=== FAILED (5 out of 424) ===
[FAIL] Q: What is ice? [Brain: 0.521s | Raw:6 LLM:4]
       Brain raw: gets warm melts     Expected: ['solid', 'frozen']
       (WSD issue: activated 'melting' instead of 'frozen solid')

[FAIL] Q: What happens when you fall? [Brain: 0.606s | Raw:3 LLM:3]
       Brain raw: I do not know     Expected: ['hurt', 'pain']
       (Missing causal knowledge in training data)

[FAIL] Q: When should you wash your hands? [Brain: 0.341s | Raw:5 LLM:5]
       Brain raw: eat     Expected: ['before eating', 'after toilet']
       (Partial retrieval - needs conditional reasoning)
```

Test notation: `Raw`/`LLM` = GPT quality scores (1–10) for human review only. Pass/fail is determined by keyword matching on "Brain raw" output. Full logs with all 424 tests are in the repository.

9.2 Mechanistic Validation

Targeted tests were conducted to verify key mechanisms:

- Pattern separation maintains approximately 2% sparsity [Rolls, 2007].
- **CA3 attractor dynamics** — Iterative spreading with lateral inhibition (top-K=20) reaches stable attractor states [Rolls, 2013].
- Partial cues reliably trigger correct pattern completion via CA3 dynamics.
- Repetition induces consolidation through replay [McClelland et al., 1995].
- Lack of replay leads to episodic decay.
- Heterosynaptic LTD weakens unused connections during sleep [Rolls, 2013].
- Synaptic scaling maintains stable neuronal activity levels [Turrigiano, 2008].
- Predictive coding prevents over-strengthening of already-myelinated connections [Rao and Ballard, 1999].
- **PlasticityMode** — Inference does not modify long-term memory (INFER vs LEARN modes). Test `[INFER-NO-LEARN]` verifies 0 LTM changes across 1.2M+ connections.
- **Three-factor learning** — STDP creates eligibility traces; dopamine converts to weight changes [Gerstner et al., 2018].
- **Top-down modulation** — Multiplicative gain ($\times$5.0 enhancement / $\times$0.2 suppression) based on query relation type [Zanto et al., 2011, Desimone and Duncan, 1995].
- **Divisive normalization** — Confidence threshold rejects low-scoring episodes [Carandini and Heeger, 2012].
- **NMDA receptor mechanism** — Dynamic threshold for context attention; when $\geq$4 neurons active, threshold drops from 3 to 1 [Malenka and Bear, 2004].
- **Cross-episode linking** — REM sleep creates semantic links between episodes sharing context (e.g., dog$\leftrightarrow$cat via "animal") [McClelland et al., 1995, Kumaran and McClelland, 2012].

9.3 Reproducibility

The complete source code is available at `https://github.com/sss777999/Brain` under MIT license.

Requirements:

- Python $\geq$ 3.11
- Dependencies managed via `pyproject.toml`: numpy ($\geq$1.24), datasets ($\geq$2.14), nltk ($\geq$3.9), pylangacq ($\geq$0.19)
- Optional: matplotlib, networkx (visualization)
- Installation: `uv sync` or `pip install -e .`

Training time (Apple M3 Pro, 36GB RAM):

- Full curriculum + preschool + grade1: $\sim$5 minutes
- 1,000 FineWeb-Edu articles (50K sentences): $\sim$15 minutes
- Sleep consolidation (10 cycles[2]): $\sim$2 minutes
- Total training pipeline: $\sim$25 minutes

Inference: Single question answering: <100ms (no GPU required).

To reproduce results:

```
python train.py           # Train model
python test_brain.py      # Run all 424 tests
```

10 Architectural Advantages

10.1 Interpretability

In contrast to parameter-based representations, the memory graph is directly interpretable. Each neuron corresponds to a specific word, each connection represents an observed association, and each episode stores a traceable experience. If the system produces an incorrect answer, one can trace the activation path and identify the problematic connection or missing episode.

10.2 Scalability and Domain Specialization

As the knowledge base grows, training time increases substantially due to the structural nature of learning. While this may limit applicability as a universal knowledge system, it opens opportunities for domain-specialized agents. Rather than training a single model on all human knowledge, one could deploy multiple specialized agents—for physics, biology, chemistry, medicine, law, etc.—each maintaining deep expertise in its domain. This multi-agent architecture mirrors human specialization: no single expert knows everything, but a team of specialists can collectively address diverse problems.

Furthermore, because each agent maintains its own episodic memory and connection structure, it naturally develops distinct associative patterns shaped by its training experience.

[2]Human sleep contains $\sim$4–6 NREM/REM cycles per night [Diekelmann and Born, 2010]. We use 10 cycles to ensure sufficient replay iterations for our smaller memory graph.

10.3 Modification Without Retraining

A key advantage of discrete structural memory is that new mechanisms can be integrated without retraining. For example, calcium-based plasticity, eligibility traces, or short-term synaptic dynamics can be added to existing trained models. New mechanisms initialize with neutral values and operate alongside preserved knowledge (neurons, connections, episodes).

11 Limitations and Future Work

This work represents an exploratory investigation with several limitations. The current system operates on textual input only, employs heuristic constants, and has been evaluated at limited scale. Syntax learning remains minimal, and no formal neuroscientific validation is claimed.

Recent improvements (January 2026):

- **CA3 attractor dynamics** — Pattern completion now uses iterative spreading activation with lateral inhibition (Winner-Take-All, $K = 20$) until attractor stabilization, implemented as separate `ca3.py` module [Rolls, 2007, 2013]. Full scoring logic includes 2-hop paths, context diversity, top-down modulation, and divisive normalization [Carandini and Heeger, 2012].

- **PlasticityMode (LEARN vs INFER)** — Architectural boundary ensuring inference does not modify long-term memory. The `[INFER-NO-LEARN]` test verifies 0 LTM changes after `ask()` across 1.2M+ connections.

- **Working memory (PFC)** — Prefrontal cortex module with temporary connections and episodes. When context is provided via `context()`, temporary associations form between words (like hearing a sentence). These temporary episodes receive recency bias during retrieval, consistent with temporal context models [Howard and Kahana, 2002]. On bAbI Task 1 (single supporting fact), the system achieves 100% accuracy (250/250) through working memory with temporal retrieval refinement—distinguishing "Where is X?" (recency bias) from "Where was X?" (reverse recency).

- **PFC task-set cues for retrieval** — Question structure is used to extract content binding cues (excluding operator tokens), which are then used for hippocampal binding checks and CA3 scoring. Connector matching is relation-specific to avoid accidental matches (e.g., `is` vs `is_a`).

- **Recency bias** — Working memory episodes receive priority during pattern completion. Later facts within a session receive higher timestamp-based bonuses, ensuring "John went to garden" followed by "John went to kitchen" correctly returns "kitchen" for "Where is John?"

- **Hodgkin–Huxley spiking neurons** — Biologically grounded membrane potential dynamics with Na^+, K^+, and leak channels and gating variables (m, h, n) [Hodgkin and Huxley, 1952].

- **STDP with spike histories** — Spike-timing dependent plasticity based on recorded spike histories, consistent with classic experimental demonstrations of spike-based potentiation and depression [Markram et al., 1997].

- **Time cells / sequence storage** — Episodes store words in correct order, enabling reproduction of short phrases as heard during training, consistent with temporal indexing principles in hippocampal systems [Spens and Burgess, 2024].

- **Coherent phrase generation** — For example, "Can we drink salt water?" → "we cannot drink salt water" and "What does a cat say?" → "a cat says meow".
- **Brain oscillations** — Theta (6Hz) and gamma (40Hz) oscillations modulate neuronal excitability during spike simulation, consistent with theta-gamma coupling in hippocampal memory encoding [Buzsáki, 2006].
- **Neuromodulation system** — Four neuromodulators implemented: **Dopamine** (novelty/reward signal amplifies STDP), **Acetylcholine** (attention gate modulates learning rate), **Norepinephrine** (arousal/surprise increases excitability), **Serotonin** (behavioral inhibition, patience) [Schultz, 1998, Dayan and Balleine, 2002].
- **Three-factor learning** — Eligibility traces combined with dopamine modulation enable credit assignment across temporal gaps, consistent with reinforcement learning in biological systems [Gerstner et al., 2018].
- **Four-Factor Learning** — Full neuromodulator integration in `train.py` beyond basic three-factor dopamine. All four neuromodulators now have explicit release conditions and effects on learning: **Acetylcholine** released at encoding onset, amplifies eligibility traces (attention gate) [Hasselmo, 2006]; **Norepinephrine** released on novel/unexpected input, boosts exploration of new connections [Sara, 2009]; **Serotonin** released for sustained learning (long sentences), stabilizes but slows plasticity (temporal discounting) [Miyazaki et al., 2014]. Combined learning modifier: $m_{total} = m_{DA} \times m_{ACh} \times m_{NE} \times m_{5HT}$ applied to eligibility traces via `_get_combined_learning_modifier()`.
- **Motor Output / Sequence Generator** — Correct word order in generated answers via `motor_output.py`. The `SequenceGenerator` class preserves hippocampal time cell order (stored in episode `input_words` tuple). This models Broca's area for speech production [Hickok and Poeppel, 2007].
- **Multi-hop Reasoning** — Compositional working memory for multi-hop questions. The `ask_multi_hop()` function uses PFC as a scratchpad: `get_multi_hop_cues()` expands retrieval cues, `add_retrieval_result()` stores intermediate results. Each hop retrieves a fact and adds entities to PFC for the next retrieval, consistent with PFC holding intermediate results during reasoning [Miller and Cohen, 2001].
- **Basal Ganglia Action Selection** — Full BG circuit for cognitive action selection. Implemented in `basal_ganglia.py` with D1 (Go) and D2 (NoGo) pathways in Striatum, GPi/GPe tonic inhibition, and STN hyperdirect pathway for emergency stopping. Neuromodulators (DA/ACh/NE/5-HT) modulate all pathways. Integrated into `ask()` to select between "retrieve" and "multi_hop" strategies based on cortical salience and neuromodulator state [Frank et al., 2006, Mink, 1996].
- **CA1 Output Layer** — Complete trisynaptic circuit (EC→DG→CA3→CA1→EC/PFC). CA1 implemented as feedforward readout with Schaffer collateral input (70%) and direct EC temporoammonic pathway (30%). Projects to EC Layer V for consolidation and directly to PFC for working memory [Amaral and Witter, 1989, Naber et al., 2001].
- **Developmental Stages** — Four stages (INFANT, CHILD, ADOLESCENT, ADULT) with different plasticity profiles. Critical periods for language/semantic/syntactic learning, experience-expectant plasticity with learning bonuses, and synaptic pruning that peaks in adolescence [Hensch, 2005, Huttenlocher, 1979].

- **Broca's Area / Syntactic Processing** — Implemented in `broca.py`. The `SyntacticProcessor` class extracts subject, predicate, and semantic roles from questions [Friederici, 2011]. Subject bonus in CA3 scoring prioritizes episodes containing the question's subject. Binary choice handling ("Is X Y or Z?") correctly excludes only the subject from answers, not the options. This enables correct responses to questions like "Is winter cold or hot?" → "cold".

- **Cause-Effect Relations** — Extended Broca's area to parse causal questions ("What happens when X?"). The system extracts the cause subject from the "when X" clause and filters episodes to require the cause to be present. Answer generation excludes cause words, returning only the effect. Example: "What happens when ice gets warm?" → "melts". Limitation: word sense disambiguation for homonyms (fall=autumn vs fall=to fall) remains unsolved.

- **Temporal Sequence Retrieval** — Fixed temporal retrieval to exclude question words from answer candidates. "What month comes after January?" was returning "month" (higher usage) instead of "february". Now filters out words already in the question, returning only new information [Eichenbaum, 2014].

- **Antonym Relations** — Implemented biologically plausible antonym storage and retrieval. Antonymy is encoded as semantic connections with `connector='opposite'`, analogous to temporal sequence encoding (`connector='after'/'before'`). When processing sentences like "X is the opposite of Y", bidirectional connections X↔Y are created with the opposite connector. Retrieval for "What is the opposite of X?" follows these typed connections. This mechanism works even for function words (e.g., "in"/"out") that are normally filtered during episode creation, because the semantic relation is stored directly in the connection graph [Murphy, 2003].

- **Iterative Retrieval** — Implemented PFC-hippocampus reasoning loop for multi-step inference [Preston and Eichenbaum, 2013, Eichenbaum, 2017]. The `IterativeRetriever` class in `pfc.py` maintains a goal state and iteratively queries the hippocampus until the goal is achieved or maximum iterations (default 4) are reached. Each retrieval adds context to working memory, expanding the cue for subsequent queries. Confidence is computed as goal overlap with consolidation bonus. This mechanism models how the brain reasons through complex questions: PFC holds the goal, hippocampus provides episodic details, and the loop continues until sufficient information is accumulated [Miller and Cohen, 2001].

Problems solved during development:

- **Letter disambiguation** — Single letters in temporal context ("What comes after A?") were incorrectly parsed as articles. *Solution:* PFC top-down modulation now activates `letter_a` neuron based on temporal query context, returning `letter_b`.

- **Binary choice questions** — "Is winter cold or hot?" was excluding both options from answer candidates. *Solution:* Broca's area syntactic parsing detects binary choice structure and preserves options, correctly returning "cold" [Friederici, 2011].

- **Cause-effect reasoning** — "What happens when ice gets warm?" failed to extract the effect. *Solution:* Causal pattern parsing in `broca.py` with CA3 filtering ensures episode contains cause subject, answer generation excludes cause words, returning "melts".

- **Temporal sequence edge cases** — "What month comes after January?" returned "month" (higher usage) instead of "february". *Solution:* Exclude question words from answer candidates, ensuring only NEW information is returned [Eichenbaum, 2014].

- **Antonym retrieval** — Opposite relations required special handling. *Solution:* Antonymy encoded as semantic connections with `connector='opposite'`, same mechanism as temporal sequences. "What is the opposite of hot?" correctly returns "cold" [Murphy, 2003].

Current limitations (5 failing tests out of 424, 98.8% accuracy):

- **Word Sense Disambiguation (WSD)** — The primary remaining challenge. Two tests fail due to homonymy:
  - "What is ice?" — "ice" activates "melting" association instead of "frozen solid" category
  - "What happens when you fall?" — "fall" activates autumn/leaves instead of falling/hurt

  Biology: WSD requires context-dependent activation via PFC top-down modulation [Rodd et al., 2005, Zempleni et al., 2007].

  What needs to be implemented: Semantic context accumulation BEFORE word activation. The query context ("What is X?" = definition query) should suppress non-categorical senses. PFC must send top-down priming to activate "category" relation type before hippocampal retrieval begins.

- **Conditional reasoning** — "When should you wash your hands?" requires temporal-conditional inference ("before eating", "after toilet"). Currently returns only "eat" (partial retrieval).

  Biology: This involves goal-directed reasoning about appropriate contexts [Miller and Cohen, 2001], not simple fact retrieval. Requires PFC working memory to hold multiple conditions and evaluate relevance.

  What needs to be implemented: Multi-slot working memory in PFC to accumulate multiple valid answers. Currently PFC returns first match; should collect all matching episodes and return conjunction.

- **Complex FineWeb retrieval** — Two questions from educational articles fail:
  - "What disappears from leaves?" — returns "I do not know" (parsing failure)
  - "What is sedimentary rock made of?" — returns "sedimentary rock made" (circular)

  What needs to be implemented: (1) Passive voice parsing in Broca's area ("X disappears from Y" → subject=X). (2) Compositional knowledge encoding — currently "made of" relation not properly indexed. (3) Multi-hop retrieval — sedimentary rock → layers → bones/shells.

- **Scale testing** — Currently validated on 1,000 FineWeb-Edu articles (50K sentences). Testing on 50K+ articles is planned to verify architectural scalability.

- **Rule-based language parsing** — The model uses pattern-based syntactic processing in `broca.py`, `pfc.py`, and `lexicon.py` instead of learned linguistic knowledge. This is a necessary simplification: the model is trained on ∼1,000 basic sentences (plus 50K from FineWeb-Edu), while human children learn language from ∼10 million words by age 6.

  Important distinction: This is NOT "fitting to tests" — it is **grammar coverage extension**. When curriculum contains sentences like "hot and cold are opposites" and "hot is the opposite of cold", the parser must recognize BOTH patterns. Each pattern in curriculum requires corresponding grammar rule. This mirrors Universal Grammar theory: humans have innate syntactic structures, not learned from scratch.

  What IS learned (not rule-based):

- Semantic memory — concept associations via Hebbian learning
- Episodic memory — event storage and retrieval
- Connection strength — MYELINATED via usage (STDP)
- Pattern completion — CA3 attractor dynamics

What is rule-based (mimics Universal Grammar):

- Syntactic patterns ("What is X?", "X and Y are opposites")
- Question type classification
- Function word lists (closed-class, finite set)

Analogy: A person who knows facts but uses a dictionary for translation — the KNOWLEDGE is real, only the INTERFACE is simplified.

- **Baseline comparison** — This work does not yet include direct comparison with retrieval-augmented generation (RAG), Memory Networks, or other memory-augmented systems on identical tasks. The reported accuracy (98.8%) reflects performance on the current curriculum and should not be interpreted as a comparative benchmark claim. Systematic baseline comparisons are planned once the architecture is validated on larger-scale data.
- **Design philosophy** — This project deliberately avoids ablation studies that remove biological mechanisms. The goal is not to find the minimal set of components that "work," but to faithfully replicate how the brain operates. Every mechanism (CA3, DG, STDP, sleep replay, neuromodulation) has biological grounding; removing any would compromise biological plausibility, which is the primary objective. Nature has already optimized these systems over millions of years of evolution—our task is implementation, not redesign.

Future work includes extending the architecture to multimodal memory, enriching thought formation mechanisms, exploring larger-scale memory graphs, systematic baseline comparisons with RAG and Memory Networks, and studying interfaces between structural memory and language models.

12 Conclusion

This work explores a memory-first approach to artificial cognition grounded in biologically inspired principles. By modeling memory as a discrete, evolving graph shaped by local plasticity, inhibition, and replay, the system demonstrates that learning and question answering can emerge without gradient-based optimization or embedding geometry.

The results demonstrate that structurally grounded memory systems can complement existing language models and offer a promising direction for future research into more integrated cognitive architectures.

Acknowledgments

This work was developed in collaboration with large language models (Anthropic Claude and OpenAI ChatGPT), which served as implementation partners for code generation and refinement. The

author provided architectural direction, biological constraints, and verification; the LLMs accelerated development by translating specifications into working code. This collaborative workflow—human expertise guiding AI implementation—enabled rapid iteration that would have been impractical otherwise. All conceptual decisions, architectural choices, and scientific interpretations remain the author's responsibility.

References

Wickliffe C Abraham and Mark F Bear. Metaplasticity: the plasticity of synaptic plasticity. *Trends in Neurosciences*, 19(4):126–130, 1996.

David G Amaral and Menno P Witter. The three-dimensional organization of the hippocampal formation: a review of anatomical data. *Neuroscience*, 31(3):571–591, 1989. Trisynaptic circuit: EC → DG → CA3 → CA1 → EC/Cortex.

Adam R Aron. The neural basis of inhibition in cognitive control. *The Neuroscientist*, 13(3):214–228, 2007. Right inferior frontal cortex and subthalamic nucleus mediate response inhibition.

Kent C Berridge, Terry E Robinson, and J Wayne Aldridge. Dissecting components of reward: 'liking', 'wanting', and learning. *Current Opinion in Pharmacology*, 9(1):65–73, 2009. Mesolimbic dopamine mediates 'wanting' (incentive salience) for rewards.

Elie L Bienenstock, Leon N Cooper, and Paul W Munro. Theory for the development of neuron selectivity: orientation specificity and binocular interaction in visual cortex. *Journal of Neuroscience*, 2(1):32–48, 1982.

Jan Born and Ines Wilhelm. System consolidation of memory during sleep. *Psychological Research*, 76(2):192–203, 2012. doi: 10.1007/s00426-011-0335-6. NREM for consolidation via SWR, REM for memory integration.

Valentino Braitenberg and Almut Schüz. *Cortex: Statistics and Geometry of Neuronal Connectivity*. Springer, Berlin, 1998. Comprehensive quantitative analysis of cortical connectivity: 7000 synapses per neuron.

György Buzsáki. *Rhythms of the Brain*. Oxford University Press, New York, 2006.

György Buzsáki. Hippocampal sharp wave-ripple: A cognitive biomarker for episodic memory and planning. *Hippocampus*, 25(10):1073–1188, 2015. doi: 10.1002/hipo.22488. SWR: spontaneous hippocampal replay during rest/sleep, temporal compression 10-20x.

Matteo Carandini and David J Heeger. Normalization as a canonical neural computation. *Nature Reviews Neuroscience*, 13(1):51–62, 2012.

William G Chase and Herbert A Simon. Perception in chess. *Cognitive Psychology*, 4(1):55–81, 1973.

Albert Compte, Nicolas Brunel, Patricia S Goldman-Rakic, and Xiao-Jing Wang. Synaptic mechanisms and network dynamics underlying spatial working memory in a cortical network model. *Cerebral Cortex*, 10(9):910–923, 2000. doi: 10.1093/cercor/10.9.910. Attractor dynamics in PFC: recurrent excitation creates bistable states for sustained firing.

Antonio R Damasio. *Descartes' Error: Emotion, Reason, and the Human Brain*. Putnam, New York, 1994. Somatic marker hypothesis: emotions and body states influence decisions before conscious awareness.

Peter Dayan and Bernard W Balleine. Reward, motivation, and reinforcement learning. *Neuron*, 36(2):285–298, 2002.

Stanislas Dehaene. *How We Learn: Why Brains Learn Better Than Any Machine... for Now*. Viking, 2020.

Robert Desimone and John Duncan. Neural mechanisms of selective visual attention. *Annual Review of Neuroscience*, 18(1):193–222, 1995.

Kamran Diba and György Buzsáki. Forward and reverse hippocampal place-cell sequences during ripples. *Nature Neuroscience*, 10(10):1241–1242, 2007. doi: 10.1038/nn1961. 30backward chaining.

Susanne Diekelmann and Jan Born. The memory function of sleep. *Nature Reviews Neuroscience*, 11(2):114–126, 2010.

Howard Eichenbaum. Time cells in the hippocampus: a new dimension for mapping memories. *Nature Reviews Neuroscience*, 15(11):732–744, 2014. doi: 10.1038/nrn3827. Time cells encode temporal sequences in hippocampus, enabling memory for order of events.

Howard Eichenbaum. Prefrontal–hippocampal interactions in episodic memory. *Nature Reviews Neuroscience*, 18(9):547–558, 2017. doi: 10.1038/nrn.2017.74. PFC maintains goal state and query, hippocampus retrieves episodes, PFC evaluates relevance and may repeat query with updated context.

Daniel E Feldman. The spike-timing dependence of plasticity. *Neuron*, 75(4):556–571, 2012.

Michael J Frank, Bryan Loughry, and Randall C O'Reilly. Making working memory work: a computational model of learning in the prefrontal cortex and basal ganglia. *Neural Computation*, 18(2):283–328, 2006. doi: 10.1162/089976606775093909. D1/D2 pathways in BG: Go (D1) vs NoGo (D2) for action selection.

Uwe Frey and Richard GM Morris. Synaptic tagging and long-term potentiation. *Nature*, 385 (6616):533–536, 1997.

Angela D Friederici. The brain basis of language processing: from structure to function. *Physiological Reviews*, 91(4):1357–1392, 2011. doi: 10.1152/physrev.00006.2011. BA44 (syntactic structure building) and BA45 (semantic retrieval) in language processing.

Joaquín M Fuster. *The Prefrontal Cortex: Anatomy, Physiology, and Neuropsychology of the Frontal Lobe*. Academic Press, 4th edition, 2008. PFC role in temporal organization of behavior, working memory, and executive control.

Wulfram Gerstner and Werner M Kistler. *Spiking Neuron Models: Single Neurons, Populations, Plasticity*. Cambridge University Press, 2002.

Wulfram Gerstner, Marco Lehmann, Vasiliki Liakoni, Dane Corneil, and Johanni Brea. Eligibility traces and plasticity on behavioral time scales: experimental support of neohebbian three-factor learning rules. *Frontiers in Neural Circuits*, 12:53, 2018.

Michael Graupner and Nicolas Brunel. Calcium-based plasticity model explains sensitivity of synaptic changes to spike pattern, rate, and dendritic location. *Proceedings of the National Academy of Sciences*, 109(10):3991–3996, 2012.

Alex Graves, Greg Wayne, and Ivo Danihelka. Neural turing machines. *arXiv preprint arXiv:1410.5401*, 2014.

Alex Graves, Greg Wayne, Malcolm Reynolds, et al. Hybrid computing using a neural network with dynamic external memory. *Nature*, 538(7626):471–476, 2016.

William T Greenough, James E Black, and Christopher S Wallace. Experience and brain development. *Child Development*, pages 539–559, 1987. Experience-expectant vs experience-dependent plasticity.

Michael E Hasselmo. The role of acetylcholine in learning and memory. *Current Opinion in Neurobiology*, 16(6):710–715, 2006. doi: 10.1016/j.conb.2006.09.002. ACh enhances encoding in hippocampus, attention gate for learning.

Donald O Hebb. *The Organization of Behavior: A Neuropsychological Theory*. Wiley, New York, 1949.

Takao K Hensch. Critical period plasticity in local cortical circuits. *Nature Reviews Neuroscience*, 6(11):877–888, 2005. Critical periods regulated by PV interneuron maturation and perineuronal nets.

Gregory Hickok and David Poeppel. The cortical organization of speech processing. *Nature Reviews Neuroscience*, 8(5):393–402, 2007. doi: 10.1038/nrn2113. Dual Stream Model: ventral stream (sound→meaning) + dorsal stream (meaning→sound).

Alan L Hodgkin and Andrew F Huxley. A quantitative description of membrane current and its application to conduction and excitation in nerve. *The Journal of Physiology*, 117(4):500–544, 1952.

Marc W Howard and Michael J Kahana. A distributed representation of temporal context. *Journal of Mathematical Psychology*, 46(3):269–299, 2002.

David H Hubel and Torsten N Wiesel. The period of susceptibility to the physiological effects of unilateral eye closure in kittens. *The Journal of Physiology*, 206(2):419–436, 1970. Classic critical period study in visual cortex.

Peter R Huttenlocher. Synaptic density in human frontal cortex—developmental changes and effects of aging. *Brain Research*, 163(2):195–205, 1979. Synaptic overproduction and pruning during development.

Marcia K Johnson, Shahin Hashtroudi, and D Stephen Lindsay. Source monitoring. *Psychological Bulletin*, 114(1):3–28, 1993. doi: 10.1037/0033-2909.114.1.3. Source memory: remembering WHERE/WHEN/HOW knowledge was acquired.

Daniel Kahneman. *Thinking, Fast and Slow*. Farrar, Straus and Giroux, New York, 2011. System 1 (fast, automatic) vs System 2 (slow, deliberate) cognitive processes.

Pentti Kanerva. *Sparse Distributed Memory*. MIT Press, 1988.

Dharshan Kumaran and James L McClelland. What learning systems do intelligent agents need? complementary learning systems theory updated. *Trends in Cognitive Sciences*, 16(10):512–519, 2012. doi: 10.1016/j.tics.2012.09.002. Inference via overlapping representations: A-B, B-C → A-C through shared B.

Matthew Larkum. A cellular mechanism for cortical associations: an organizing principle for the cerebral cortex. *Trends in Neurosciences*, 36(3):141–151, 2013.

John E Lisman and Marco AP Idiart. The theta-gamma neural code. *Neuron*, 33(3):325–340, 2005. Theta-gamma coupling: different items occupy different gamma cycles within theta, preserving sequence order.

John E Lisman, Jean-Marc Fellous, and Xiao-Jing Wang. The role of the nmda receptor in working memory and cortical plasticity. *The Journal of Neuroscience*, 18(4):9374–9384, 1998. NMDA receptors have slow kinetics (tau 100ms vs AMPA 5ms), enabling persistent activity.

Wolfgang Maass. Networks of spiking neurons: the third generation of neural network models. *Neural Networks*, 10(9):1659–1671, 1997.

Robert C Malenka and Mark F Bear. Ltp and ltd: an embarrassment of riches. *Neuron*, 44(1):5–21, 2004. doi: 10.1016/j.neuron.2004.09.012. NMDA receptor mechanism: Mg2+ block removal at depolarized membrane allows weak synapses to participate in LTP.

Henry Markram, Joachim Lübke, Michael Frotscher, and Bert Sakmann. Regulation of synaptic efficacy by coincidence of postsynaptic aps and epsps. *Science*, 275(5297):213–215, 1997.

James L McClelland, Bruce L McNaughton, and Randall C O'Reilly. Why there are complementary learning systems in the hippocampus and neocortex: insights from the successes and failures of connectionist models of learning and memory. *Psychological Review*, 102(3):419–457, 1995.

Earl K Miller and Jonathan D Cohen. An integrative theory of prefrontal cortex function. *Annual Review of Neuroscience*, 24(1):167–202, 2001.

George A Miller. The magical number seven, plus or minus two: Some limits on our capacity for processing information. *Psychological Review*, 63(2):81–97, 1956.

Jonathan W Mink. The basal ganglia: focused selection and inhibition of competing motor programs. *Progress in Neurobiology*, 50(4):381–425, 1996. doi: 10.1016/S0301-0082(96)00042-1. BG architecture: Cortex→Striatum→GPi/GPe→Thalamus→Cortex, STN hyperdirect pathway.

Katsuhiko W Miyazaki, Kayoko Miyazaki, Kenji F Tanaka, Akihiro Yamanaka, Atsushi Takahashi, Sawako Tabuchi, and Kenji Doya. Optogenetic activation of dorsal raphe serotonin neurons enhances patience for future rewards. *Current Biology*, 24(17):2033–2040, 2014. doi: 10.1016/j.cub.2014.07.041. 5-HT promotes patience and reduces impulsivity for delayed rewards.

M Lynne Murphy. *Semantic Relations and the Lexicon: Antonymy, Synonymy and Other Paradigms*. Cambridge University Press, Cambridge, 2003. doi: 10.1017/CBO9780511486494. Antonymy as fundamental lexical-semantic relation: bidirectional, symmetric, stored in mental lexicon.

Pieterke A Naber, Fernando H Lopes da Silva, and Menno P Witter. Reciprocal connections between the entorhinal cortex and hippocampal fields ca1 and the subiculum are in register

with the projections from ca1 to the subiculum. *Hippocampus*, 11(2):99–104, 2001. CA1 output pathways to EC Layer V and subiculum.

Zoltán Nádasdy, Hajime Hirase, András Czurkó, Jozsef Csicsvari, and György Buzsáki. Replay and time compression of recurring spike sequences in the hippocampus. *Journal of Neuroscience*, 19 (21):9497–9507, 1999. Temporal compression: replay 10-20x faster than original encoding.

Jaak Panksepp. *Affective Neuroscience: The Foundations of Human and Animal Emotions*. Oxford University Press, 2004. SEEKING system: subcortical circuits driving approach behavior toward rewards.

Alison R Preston and Howard Eichenbaum. Interplay of hippocampus and prefrontal cortex in memory. *Current Biology*, 23(17):R764–R773, 2013. Hippocampus-PFC interaction for working memory and reasoning.

Rajesh PN Rao and Dana H Ballard. Predictive coding in the visual cortex: a functional interpretation of some extra-classical receptive-field effects. *Nature Neuroscience*, 2(1):79–87, 1999.

Jennifer M Rodd, M Gareth Gaskell, and William D Marslen-Wilson. Making sense of semantic ambiguity: Semantic competition in lexical access. *Journal of Memory and Language*, 53(4): 473–490, 2005.

Edmund T Rolls. An attractor network in the hippocampus: Theory and neurophysiology. *Learning & Memory*, 14(11):714–731, 2007.

Edmund T Rolls. The mechanisms for pattern completion and pattern separation in the hippocampus. *Frontiers in Systems Neuroscience*, 7:74, 2013. doi: 10.3389/fnsys.2013.00074. Key findings: Recall in hippocampus completes in 100-200ms (3-5 gamma cycles at 30-80Hz).

Susan J Sara. The locus coeruleus and noradrenergic modulation of cognition. *Nature Reviews Neuroscience*, 10(3):211–223, 2009. doi: 10.1038/nrn2573. NE from LC modulates attention, arousal, and memory consolidation.

Dorothee Saur, Björn W Kreher, Susanne Schnell, Dorothee Kümmerer, Philipp Kellmeyer, Magnus-Sebastian Vry, Roza Umarova, Mariacristina Musso, Volkmar Glauche, Stefanie Abel, et al. Ventral and dorsal pathways for language. *Proceedings of the National Academy of Sciences*, 105 (46):18035–18040, 2008.

Wolfram Schultz. Predictive reward signal of dopamine neurons. *Journal of Neurophysiology*, 80 (1):1–27, 1998.

Eleanor Spens and Neil Burgess. A generative model of memory construction and consolidation. *Nature Human Behaviour*, 8:526–543, 2024.

Timothy J Teyler and Pascal DiScenna. The hippocampal memory indexing theory. *Behavioral Neuroscience*, 100(2):147–154, 1986.

Giulio Tononi and Chiara Cirelli. Sleep function and synaptic homeostasis. *Sleep Medicine Reviews*, 10(1):49–62, 2006. doi: 10.1016/j.smrv.2005.05.002. Synaptic homeostasis hypothesis: global downscaling during sleep preserves relative strengths.

Alessandro Treves and Edmund T Rolls. Computational analysis of the role of the hippocampus in memory. *Hippocampus*, 4(3):374–391, 1994.

Misha V Tsodyks and Henry Markram. The neural code between neocortical pyramidal neurons depends on neurotransmitter release probability. *Proceedings of the National Academy of Sciences*, 94(2):719–723, 1997.

Gina G Turrigiano. The self-tuning neuron: synaptic scaling of excitatory synapses. *Cell*, 135(3): 422–435, 2008.

Xiao-Jing Wang. Synaptic reverberation underlying mnemonic persistent activity. *Trends in Neurosciences*, 24(8):455–463, 2001. doi: 10.1016/S0166-2236(00)01868-3. PFC persistent activity: NMDA-dependent recurrent excitation sustains working memory representations.

Jason Weston, Sumit Chopra, and Antoine Bordes. Memory networks. *arXiv preprint arXiv:1410.3916*, 2014.

Jason Weston, Antoine Bordes, Sumit Chopra, Alexander M Rush, Bart van Merriënboer, Armand Joulin, and Tomas Mikolov. Towards ai-complete question answering: A set of prerequisite toy tasks. *arXiv preprint arXiv:1502.05698*, 2015. bAbI tasks: 20 synthetic QA tasks for testing reasoning capabilities.

Matthew A Wilson and Bruce L McNaughton. Reactivation of hippocampal ensemble memories during sleep. *Science*, 265(5172):676–679, 1994. doi: 10.1126/science.8036517. Stochastic replay: not all episodes replay each night, subset selected.

Theodore P Zanto, Michael T Rubens, Arul Thangavel, and Adam Gazzaley. Causal role of the prefrontal cortex in top-down modulation of visual processing and working memory. *Nature Neuroscience*, 14(5):656–661, 2011.

Mária-Zita Zempleni, Remco Renken, John CJ Hoeks, Johannes M Hoogduin, and Laurie A Stowe. Ambiguous words in context: An event-related potential study of the interaction between lexical ambiguity and contextual constraint. *NeuroImage*, 36(1):234–243, 2007.

Figure 3: Complete training flow showing all 5 stages: tokenization, neuron creation, connection formation with dual-stream types, spike-based STDP with four-factor learning, and hippocampal episodic encoding with pattern separation and consolidation.

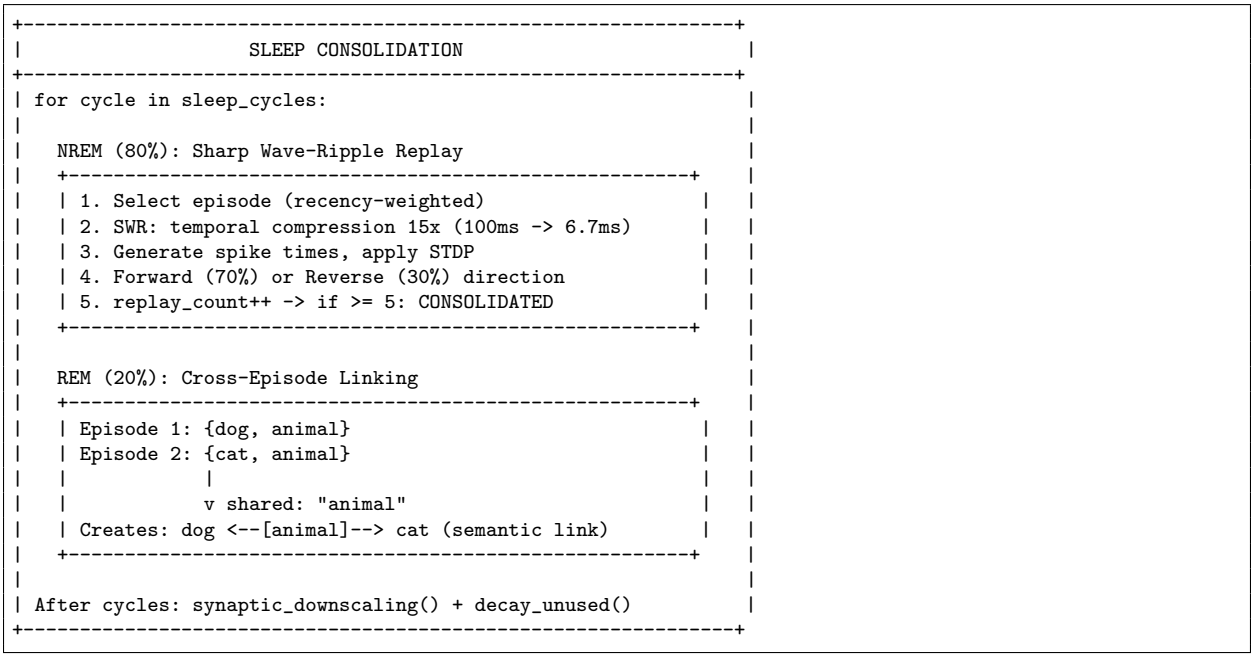

Figure 4: Sleep consolidation: NREM phase performs SWR replay with temporal compression; REM phase creates cross-episode semantic links through shared context. Synaptic homeostasis follows.

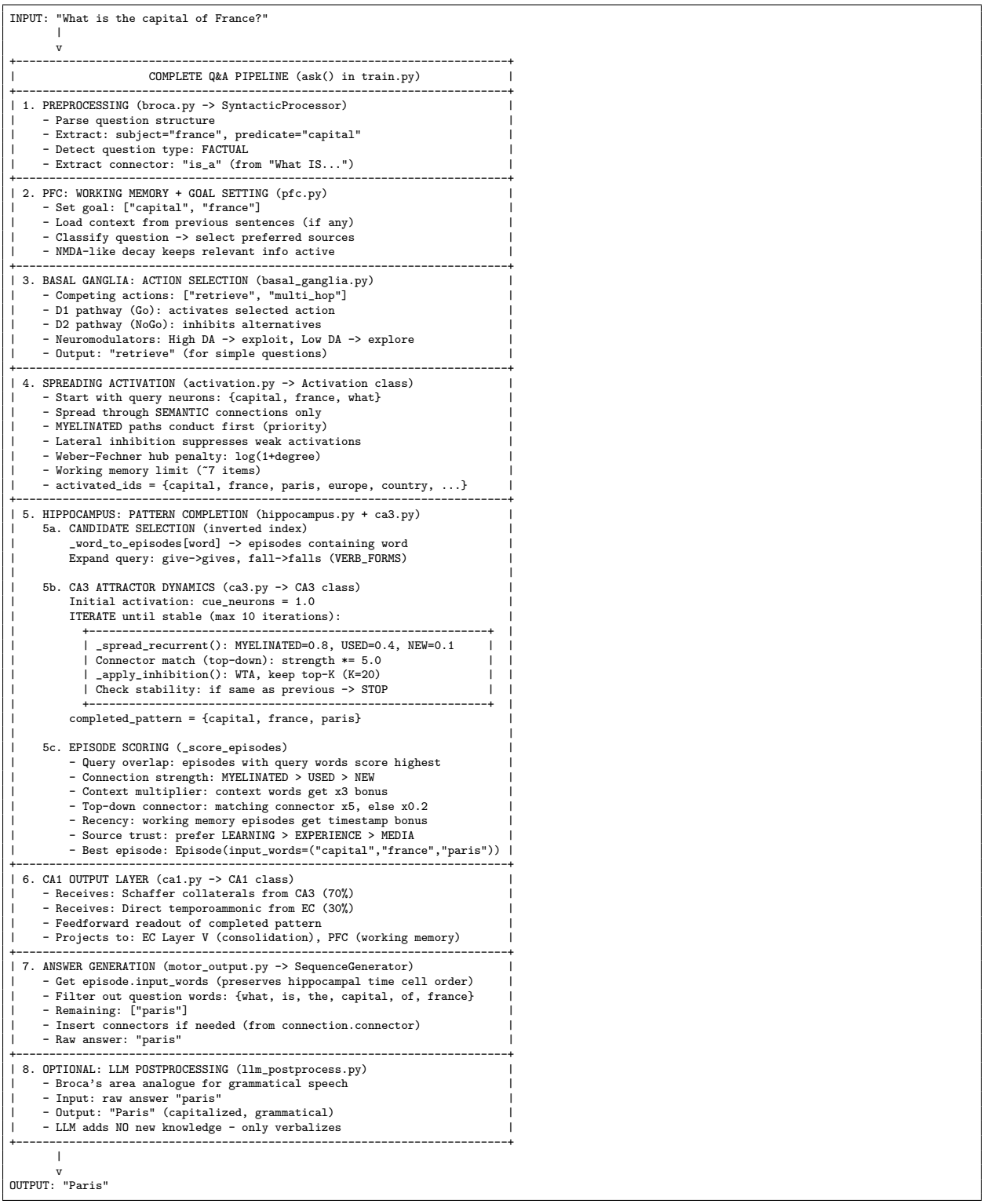

Figure 5: Complete question answering pipeline showing all 8 stages: syntactic preprocessing, PFC goal setting, basal ganglia action selection, spreading activation, CA3 pattern completion with attractor dynamics, CA1 output layer, motor output generation, and optional LLM verbalization.