Biologically-Grounded Memory and Decision Systems for Neural Language Models

Open research specification 5 of 7

Krrish Choudhary

Verace Pvt. Ltd.

This document was first drafted in the form of a patent specification, which is why it speaks of an invention and of claims. No patent application was filed and none is intended. Where the text mentions a related or co-pending patent application, it means a companion document in this set, and none of them was filed. The work is published as open research.

Published under the PolyForm Noncommercial License 1.0.0

Copyright (c) 2026 Verace Pvt. Ltd. (https://verace.in)

Contents

1 TITLE **3**

2 FIELD **3**

3 BACKGROUND **3**

3.1 Technical Field and Problem Statement 3

3.2 Limitations of Prior Art 3

3.2.1 Working Memory 3

3.2.2 Episodic Memory and Consolidation 4

3.2.3 Action Selection and Temporal Prediction 4

3.3 Objects of the Invention 5

4 SUMMARY **5**

4.1 Multi-Timescale Working Memory 5

4.2 Hippocampal Episodic Memory 7

4.3 Hierarchical Basal Ganglia 8

4.4 Cerebellar Forward Model 8

5 DESCRIPTION OF THE FIGURES **9**

6 DETAILED DESCRIPTION **32**

6.1 System Architecture Overview 32

6.2 NMDA Attractor Slots — Detailed Implementation 32

6.2.1 Biological Basis 32

6.2.2 Slot State Update 32

6.2.3 Content-Based Read 33

6.2.4 NMDA Confidence 33

6.2.5 Parallel Scan Acceleration 33

6.3 Activity-Silent Trace — Detailed Implementation 34

6.3.1 Biological Basis 34

6.3.2 Facilitation and Depression Dynamics 34

6.3.3 Two-Scan Parallel Acceleration 34

6.3.4 Silent Trace Confidence 35

6.4 Phonological Loop — Detailed Implementation 35

6.4.1 Biological Basis 35

6.4.2 Ring Buffer with Context Tagging 35

6.4.3 Phonological Confidence 35

6.5 Metamemory Confidence Aggregation 36

6.6 Hippocampal Episodic Memory — Detailed Implementation 36
6.6.1 Architecture Overview 36
6.6.2 Dentate Gyrus Pattern Separation 36
6.6.3 CA3 Encode and Complete 36
6.6.4 Adaptive Write Gate 36
6.6.5 Bayesian Concept Confidence — Detail 37
6.6.6 Goal-Directed Retrieval — Detail 37
6.6.7 Elaborative Encoding — Detail 37
6.6.8 Error-Driven Forgetting — Detail 38
6.6.9 Multi-Token Chunking — Detail 38
6.6.10 Schema Absorption — Detail 38
6.6.11 Consolidation Scoring — Detail 38
6.6.12 Micro-Consolidation (Sharp Wave Ripples) — Detail 39
6.6.13 Relevance-Based Eviction 39
6.7 Hierarchical Basal Ganglia — Detailed Implementation 40
6.7.1 Biological Basis 40
6.7.2 Single Loop Architecture 40
6.7.3 Three-Loop Hierarchy 40
6.7.4 Tonic Dopamine Tracking 41
6.8 Cerebellar Forward Model — Detailed Implementation 41
6.8.1 Biological Basis 41
6.8.2 Architecture 41
6.8.3 Forward Pass 41
6.9 Identity-at-Initialisation Contract 42
6.10 State Persistence Across Forward Calls 42

7 CLAIMS **42**

1. TITLE

Biologically-Grounded Memory and Decision Systems for Neural Language Models

2. FIELD

The present invention relates generally to memory systems and decision-making mechanisms in neural language models, and more particularly to a biologically-grounded suite of working memory subsystems, episodic memory formation and retrieval mechanisms, memory consolidation pipelines, and hierarchical action selection circuits that collectively enable a cortical column network to maintain, retrieve, consolidate, and act upon information across multiple timescales ranging from immediate token-level retention through discourse-level planning, said systems being modelled on the mammalian hippocampus, prefrontal working memory, basal ganglia, and cerebellum.

3. BACKGROUND

3.1. Technical Field and Problem Statement

Neural language models require the ability to store, maintain, retrieve, and act upon information across timescales ranging from single tokens to entire documents. Current transformer-based architectures rely on a fixed-length context window implemented through self-attention over all prior tokens, creating two fundamental limitations: (a) a hard sequence-length boundary beyond which no information is retained, and (b) uniform treatment of all prior tokens regardless of their importance, recency, or reliability. These architectures lack explicit memory systems, consolidation mechanisms, metamemory confidence signals, and hierarchical action selection circuits, all of which are present in the mammalian brain and are known to be essential for flexible, adaptive language behaviour.

3.2. Limitations of Prior Art

3.2.1. Working Memory

Luck and Vogel ("The capacity of visual working memory for features and conjunctions," *Nature*, 1997) established that human working memory maintains approximately four discrete items through persistent neural activity, primarily mediated by NMDA receptor-dependent plateau potentials in prefrontal cortex. Baddeley ("Working memory," *Science*, 1992) further decomposed working memory into a phonological loop for verbal rehearsal, a visuospatial sketchpad, and a central executive. Mongillo, Barak, and Tsodyks ("Synaptic theory of working memory," *Science*, 2008) demonstrated that working memory can also be maintained in an activity-silent form through short-term synaptic facilitation and depression (the Tsodyks-Markram model), persisting

even when neural firing ceases. No prior neural language model implements all three working memory subsystems—persistent attractor slots, activity-silent synaptic traces, and phonological loop rehearsal—operating in parallel at distinct timescales with per-subsystem confidence scoring and a metamemory aggregation mechanism.

3.2.2. Episodic Memory and Consolidation

Eichenbaum ("Hippocampus: Cognitive processes and neural representations that underlie declarative memory," *Neuron*, 2004) established the hippocampal system as the substrate for episodic memory, with the dentate gyrus performing pattern separation and CA3 performing autoassociative pattern completion. McClelland, McNaughton, and 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*, 1995) proposed Complementary Learning Systems theory, wherein the hippocampus rapidly encodes specific episodes while the neocortex slowly integrates patterns through interleaved replay during sleep. Tse et al. ("Schemas and memory consolidation," *Science*, 2007) demonstrated that pre-existing knowledge schemas accelerate hippocampal-to-cortical consolidation. Craik and Lockhart ("Levels of processing: A framework for memory research," *Journal of Verbal Learning and Verbal Behavior*, 1972) showed that deeper, more elaborative encoding produces stronger and more retrievable memory traces. Anderson ("Retrieval-induced forgetting," in *True and False Recovered Memories*, 2014) identified that retrieving a target memory can suppress competing traces sharing the same retrieval cue. No prior neural language model integrates all of: dentate gyrus pattern separation, Bayesian concept confidence tracking, goal-directed retrieval biased by a global workspace, elaborative encoding with inter-episode linking, error-driven forgetting, multi-token chunking, schema absorption from consolidation centroids, and sharp-wave-ripple micro-consolidation during wakefulness.

3.2.3. Action Selection and Temporal Prediction

Frank and Badre ("Mechanisms of hierarchical reinforcement learning in corticostriatal circuits," *Cerebral Cortex*, 2012) established that the basal ganglia implements hierarchical action selection through D1 (direct, stabilising) and D2 (indirect, flexibility-enabling) corticostriatal pathways, organised into nested loops by cortical depth band with anterior loops gating posterior loops in a Koechlin cascade. Wolpert, Miall, and Kawato ("Internal models in the cerebellum," *Trends in Cognitive Sciences*, 1998) demonstrated that the cerebellum maintains forward models predicting motor and cognitive outcomes, applying real-time correction when predictions fail. No prior neural language model combines hierarchical D1/D2 basal ganglia loops with Boltzmann receptor kinetics, cascade top-down gating, tonic dopamine tracking, and a cerebellar forward model with causal error correction for temporal prediction.

3.3. Objects of the Invention

It is therefore an object of the present invention to provide a multi-timescale working memory system for neural language models comprising NMDA attractor slots, activity-silent synaptic traces, and a phonological loop with temporal context tagging, each producing per-subsystem confidence signals aggregated into a metamemory scalar.

It is a further object to provide an episodic memory system implementing hippocampal dentate gyrus pattern separation, CA3 autoassociative completion, Bayesian concept confidence, goal-directed retrieval, elaborative encoding, error-driven forgetting, multi-token chunking, schema absorption, and sharp-wave-ripple micro-consolidation.

It is a further object to provide a hierarchical action selection system modelled on the basal ganglia with three D1/D2 corticostriatal loops operating at word, sentence, and discourse levels with top-down cascade gating and Boltzmann receptor kinetics.

It is a further object to provide a cerebellar temporal prediction and error correction system that predicts next-timestep representations and applies causal correction signals.

It is a further object to ensure that all said memory and decision subsystems are identity-at-init(Stale from transformer era check with /nct-research)ialisation, ensuring zero contribution before learning commences and enabling safe integration with pre-trained models.

4. SUMMARY

The present invention provides biologically-grounded memory and decision systems for neural language models, implemented as modular subsystems that operate within a cortical column network. The invention comprises four principal subsystems: a multi-timescale working memory, a hippocampal episodic memory with consolidation, a hierarchical basal ganglia for action selection, and a cerebellar forward model for temporal prediction.

4.1. Multi-Timescale Working Memory

The working memory subsystem comprises three biologically-distinct memory mechanisms operating in parallel:

NMDA Attractor Slots. A set of $K = 4$ competitive memory slots, each of dimension d_{slot}, maintaining persistent representations through exponential decay with learnable per-slot decay rates. At each time step, an input representation $\mathbf{h}_t$ is projected through a scoring network to produce softmax write probabilities over slots, and through a content network to produce the write value. The slot update follows the linear recurrence:

$$\mathbf{s}_k[t] = \text{sigmoid}(\beta_k) \cdot \mathbf{s}_k[t-1] + w_k[t] \cdot \mathbf{c}[t] \tag{1}$$

where β_k is a learnable decay bias initialised such that $\text{sigmoid}(\beta_k) \approx 0.9975$ (half-life ≈ 277

tokens), $w_k[t] = \text{softmax}(\mathbf{W}_{\text{score}} \cdot \mathbf{h}_t)_k$ is the competitive write probability, and $\mathbf{c}[t] = \mathbf{W}_{\text{write}} \cdot \mathbf{h}_t$ is the content vector. Reading employs content-based attention: query $\mathbf{c}[t]$ attends over slot contents via softmax-normalised dot products. NMDA confidence is computed as the read-attention-weighted sum of slot norms, measuring activation strength. This recurrence admits a parallel scan decomposition with constant decay coefficient $a_t = \text{sigmoid}(\beta_k)$ and input-dependent additive term $b_t = w_k[t] \cdot \mathbf{c}[t]$, executed as K independent parallel scans over the $K \times d_{\text{slot}}$ state vector to eliminate the sequential time loop.

Activity-Silent Trace. A Tsodyks-Markram synaptic facilitation-depression model maintaining information in synaptic weights rather than sustained activity. Two state variables evolve per trace dimension: facilitation u (residual calcium) and depression x (vesicle availability). The update equations are:

$$u[t] = \text{sigmoid}(\alpha_f) \cdot u[t-1] + (1 - u[t-1]) \cdot g[t] \cdot |\mathbf{c}[t]| \tag{2}$$

$$x[t] = \text{sigmoid}(\alpha_d) \cdot x[t-1] + (1 - \text{sigmoid}(\alpha_d)) - \text{sigmoid}(\alpha_{df}) \cdot 0.5 \cdot u[t] \cdot x[t-1] \tag{3}$$

where $g[t] = \text{sigmoid}(\mathbf{W}_g \cdot \mathbf{h}_t)$ is a learned write gate and $\mathbf{c}[t] = \mathbf{W}_{\text{store}} \cdot \mathbf{h}_t$ is the stored content. The effective synaptic release is $j_{\text{eff}}[t] = u[t] \cdot x[t]$. Silent trace confidence is the mean facilitation level $\bar{u}[t]$. Crucially, Equation 2 is linearisable as $u[t] = u[t-1] \cdot (\text{sigmoid}(\alpha_f) - g[t] \cdot |\mathbf{c}[t]|) + g[t] \cdot |\mathbf{c}[t]|$, admitting a parallel scan. Depression x depends on u and thus requires a second sequential scan after the first, yielding two sequential parallel scans total.

Phonological Loop. A ring buffer of $K_p = 8$ slots implementing articulatory rehearsal with temporal context tagging. Each slot stores both a content vector $\mathbf{c} \in \mathbb{R}^{d_p}$ and a temporal context vector $\mathbf{ctx} \in \mathbb{R}^{d_{\text{ctx}}}$. The context vector evolves via a learned near-orthogonal linear transformation:

$$\mathbf{ctx}[t] = \frac{\mathbf{W}_{\text{evolve}} \cdot \mathbf{ctx}[t-1]}{\|\mathbf{W}_{\text{evolve}} \cdot \mathbf{ctx}[t-1]\| + \epsilon} \tag{4}$$

where $\mathbf{W}_{\text{evolve}}$ is initialised as a scaled orthogonal matrix ($0.99 \times \mathbf{Q}$) producing slow rotation in context space. At write time, the content is tagged with the current context via additive projection. At read time, retrieval scores combine both content similarity (dot product between query and slot contents) and context similarity (dot product between current context and stored context tags), enabling serial-order retrieval that pure content addressing cannot achieve. Rehearsal refreshes the most recently written slot: $\mathbf{s}_{\text{recent}} \leftarrow \mathbf{s}_{\text{recent}} + r \cdot \mathbf{s}_{\text{recent}}$ where $r = \text{sigmoid}(r_{\text{raw}})$.

Metamemory Confidence Aggregation. Per-subsystem confidence signals—NMDA slot activation strength, silent trace mean facilitation, and phonological context similarity—are each passed through a sigmoid normalisation and averaged to produce a single scalar $c_{\text{mem}} \in [0, 1]$. This metamemory signal drives downstream deliberation depth: low confidence triggers deeper processing in the drift-diffusion decision mechanism.

4.2. Hippocampal Episodic Memory

The hippocampal subsystem implements episodic encoding, storage, retrieval, and consolidation through the following mechanisms:

Dentate Gyrus Pattern Separation. Input representations are projected through a sparse expansion layer $\mathbf{W}_{\text{DG}} \in \mathbb{R}^{d \times d_{\text{sparse}}}$ followed by ReLU activation, producing high-dimensional sparse codes that decorrelate similar inputs. The sparse activation pattern is hashed via top-k active unit indices to produce a concept key for Bayesian confidence tracking.

CA3 Autoassociative Completion. Sparse DG codes are projected to dense representations via $\mathbf{W}_{\text{CA3_encode}}$, then dense queries are projected back to sparse via $\mathbf{W}_{\text{CA3_retrieve}}$ with decay gating, and completed via a separate Schaffer collateral projection $\mathbf{W}_{\text{CA3_complete}}$ (initialised to zeros for identity at init).

Bayesian Concept Confidence. Each concept, identified by its DG sparse code hash, maintains a confidence value $p \in [0.01, 0.99]$. New concepts begin at $p_0 = 0.3$ (tentative). On confirmation (positive evidence with precision weighting):

$$p_{\text{post}} = p_{\text{prior}} \cdot (1 - \alpha) + \frac{e}{1 + \sigma^2} \cdot \alpha, \quad \alpha = \frac{0.1}{1 + \sigma^2} \tag{5}$$

On contradiction (negative evidence): $p_{\text{post}} = p_{\text{prior}} \times 0.5$. Concept confidence scales episodic retrieval strength: $\text{scale} = \text{sim} \times (0.5 + 0.5 \times p)$, ensuring unreliable concepts contribute less.

Goal-Directed Retrieval. When a goal context signal from the global workspace is available, the retrieval query is biased: $\mathbf{q}_{\text{norm}} = 0.7 \cdot \mathbf{q}_{\text{input}} + 0.3 \cdot \mathbf{q}_{\text{goal}}$. This models prefrontal modulation of hippocampal retrieval (Moscovitch, 2002), focusing memory search on task-relevant episodes rather than purely stimulus-driven matching.

Elaborative Encoding. At encode time, each new episode is linked to its top-3 most similar existing episodes by cosine similarity. At retrieval time, if the best match has linked episodes, those related episodes are also retrieved and blended: $\textbf{out} \leftarrow \textbf{out} + 0.3 \times \text{mean}(\text{CA1}(\text{related}))$. This implements Craik's (1972) levels-of-processing framework: deeper encoding at storage produces richer retrieval.

Error-Driven Forgetting. When a new episode has high cue similarity (> 0.7) but low content similarity (< 0.3) with an existing episode, the old episode's surprise score is zeroed, accelerating its eviction. This implements Anderson's (2014) retrieval-induced forgetting: encoding corrected information suppresses contradicting traces.

Multi-Token Chunking. When mean prediction error falls below a chunk threshold (< 0.3), the current token is merged into the most recent episode via running average: $\mathbf{x}_{\text{last}} \leftarrow 0.7 \times \mathbf{x}_{\text{last}} + 0.3 \times \mathbf{x}_{\text{new}}$. This enables compound concept encoding (e.g., "New York" stored as one episode, not two).

Schema Absorption. After consolidation produces centroid patterns (semantic memories), these centroids are absorbed as schemas on the hippocampus. If a new schema matches an existing one (cosine similarity > 0.8), they merge via weighted average with confidence boost

($+0.05$). Schemas receive a $1.5\times$ retrieval strength boost over individual episodes.

Consolidation Scoring. Replay priority scoring integrates retrieval count and concept confidence: $\text{score}_i = \text{count}_i \times (0.5 + 0.5 \times p_i)$. High-confidence, frequently-retrieved memories consolidate first.

Micro-Consolidation (Sharp Wave Ripples). Every ~ 50 steps during wakefulness, the top-3 most-retrieved memories are replayed through the full cortical processing pipeline at 5% of normal learning rate. This implements Buzsáki's (2015) sharp wave ripple mechanism: compressed replay transferring episodic knowledge into cortical weights continuously, not only during sleep.

4.3. Hierarchical Basal Ganglia

The basal ganglia subsystem implements action selection through three corticostriatal D1/D2 loops organised by depth band:

- **Loop 1 (Word)**: Columns 0–3, fast gating for word-level selection.
- **Loop 2 (Sentence)**: Columns 4–7, medium gating for phrase/sentence planning.
- **Loop 3 (Discourse)**: Columns 8–11, slow gating for discourse-level goals.

Each loop implements Go (D1, direct) and NoGo (D2, indirect) pathways with Boltzmann receptor kinetics:

$$\text{D1}_{\text{act}} = \text{sigmoid}(s_1 \cdot (\text{DA}_{\text{eff}} - \text{EC50}_1)) \tag{6}$$

$$\text{D2}_{\text{act}} = \text{sigmoid}(s_2 \cdot (\text{DA}_{\text{eff}} - \text{EC50}_2)) \tag{7}$$

where $\text{DA}_{\text{eff}} = \text{DA}_{\text{phasic}} - \text{DA}_{\text{tonic}}$ is the phasic deviation from tonic baseline, and s_i, EC50_i are learnable receptor parameters with D2 having lower EC50 (higher sensitivity). Loop output is $\text{D1}_{\text{mod}} - \text{D2}_{\text{mod}}$ (Go minus NoGo competition). The cascade runs top-down: discourse output gates sentence processing via a learned projection, and sentence output gates word processing, implementing the Koechlin (2003) hierarchical control architecture.

4.4. Cerebellar Forward Model

The cerebellum subsystem implements temporal prediction and error correction. A forward model projects the current representation through a hidden layer to predict the next timestep's representation. The prediction error is:

$$\mathbf{e}[t] = \mathbf{x}[t+1] - \hat{\mathbf{x}}[t+1|t] \tag{8}$$

The error is applied as a causal correction (right-padded to maintain temporal causality) through a learned error projection initialised to zeros. The cerebellar prediction loss $\mathcal{L}_{\text{cb}} = \text{MSE}(\hat{\mathbf{x}}[:-1], \mathbf{x}[1:])$ is monitored for temporal prediction quality.

5. DESCRIPTION OF THE FIGURES

The invention will be more fully understood from the following detailed description taken in conjunction with the accompanying drawings, in which:

FIG. P5-1 is a system-level diagram of the multi-timescale working memory system (Claim 1), showing the three parallel subsystems: NMDA attractor slots with competitive softmax write and exponential decay, activity-silent trace with Tsodyks-Markram facilitation-depression dynamics, and phonological loop with ring buffer and temporal context tagging, converging into metamemory confidence aggregation via sigmoid normalisation.

FIG. P5-2 is a detailed diagram of the hippocampal episodic memory system (Claim 2), showing dentate gyrus pattern separation via sparse expansion, CA3 autoassociative pattern completion, the episodic buffer with encoding pipeline (surprise computation, multi-token chunking, elaborative linking, error-driven forgetting), per-concept Bayesian confidence tracking, and retrieval with concept confidence scaling.

FIG. P5-3 is a diagram of the hierarchical basal ganglia action selection system (Claim 3), showing three corticostriatal loops (discourse, sentence, word) with D1 direct and D2 indirect pathways, Boltzmann receptor activation with learnable EC50 and slope, Go-minus-NoGo competition, top-down cascade gating, and effective dopamine as phasic deviation from tonic baseline.

FIG. P5-4 is a flow diagram of the NMDA parallel prefix scan (Claim 4, dep. on 1), showing batch pre-computation of write probabilities and content vectors, reshape into single scan of dimension $K \times d_{\text{slot}}$ with constant per-slot decay, and batched attention reads over the scan output.

FIG. P5-5 is a flow diagram of the dual-scan silent trace acceleration (Claim 5, dep. on 1), showing the linearised facilitation recurrence computed via a first parallel scan, followed by a second parallel scan for depression with coefficients dependent on the first scan's output, yielding effective synaptic release as their product.

FIG. P5-6 is a diagram of the phonological rehearsal mechanism (Claim 6, dep. on 1), showing the ring buffer of K_p slots, the articulatory rehearsal step that boosts the most recently written slot by a learnable strength factor $r = \text{sigmoid}(r_{\text{raw}})$ prior to writing the current input.

FIG. P5-7 is a diagram of the temporal context rotation (Claim 7, dep. on 1), showing initialisation as a random unit-norm vector, evolution via a learned near-orthogonal linear transformation with scaling factor less than 1.0, and L_2 normalisation, producing slow rotation in context space with monotonically decaying similarity over distance.

FIG. P5-8 is a diagram of the content-plus-context read scoring (Claim 8, dep. on 1), showing the dual scoring mechanism: content similarity via dot product between query projection and slot contents, temporal context similarity via dot product between current and stored context vectors, and their additive combination for softmax attention.

FIG. P5-9 is a diagram of the DDM integration of metamemory confidence (Claim 9, dep. on 1), showing how the metamemory scalar modulates the drift-diffusion decision mechanism:

low confidence causes longer evidence accumulation (more deliberation steps), high confidence causes shorter accumulation (faster token commitment).

FIG. P5-10 is a diagram of volatile working memory state management (Claim 10, dep. on 1), showing state persistence across sequential forward calls, reset on batch size change, and exclusion from disk checkpoints, modelling the volatile nature of biological working memory.

FIG. P5-11 is a diagram of per-subsystem confidence scaling (Claim 11, dep. on 1), showing the three different sigmoid transformations: sigmoid$(2c_{\text{NMDA}} - 1)$ for NMDA, sigmoid$(3c_{\text{silent}} - 1)$ for silent trace, and sigmoid(c_{phono}) for phonological loop, each multiplicatively gating its respective subsystem's output.

FIG. P5-12 is a diagram of identity-at-init(Stale from transformer era check with /nct-research)ialisation for working memory (Claim 12, dep. on 1), showing all output projections initialised to zero weights and all gate parameters initialised to -2.0 (sigmoid ≈ 0.12), ensuring zero WM contribution at initialisation.

FIG. P5-13 is a diagram of goal-directed hippocampal retrieval (Claim 13, dep. on 2), showing the biased retrieval query computed as $0.7 \times q_{\text{input}} + 0.3 \times q_{\text{goal}}$ with L_2 normalisation, modelling prefrontal modulation of hippocampal memory search.

FIG. P5-14 is a diagram of elaborative linking (Claim 14, dep. on 2), showing encoding-time cosine similarity computation to find top-3 most similar existing episodes, storage of related indices in metadata, and retrieval-time blending of related episodes with weight 0.3.

FIG. P5-15 is a diagram of schema absorption (Claim 15, dep. on 2), showing consolidation centroid extraction, cosine similarity matching against existing schemas with 0.8 threshold, weighted merge with confidence boost for matches, new schema creation for novel centroids, and multiplicative strength boost for schema matches at retrieval.

FIG. P5-16 is a flow diagram of micro-consolidation during wakefulness (Claim 16, dep. on 2), showing periodic selection of top-n most-retrieved memories scored by retrieval count weighted by concept confidence, working memory reset, forward pass through cortical columns at reduced learning rate, and full learning pipeline including predictive coding, Hebbian with BCM, anti-Hebbian decorrelation, metaplasticity, and subcortical learning.

FIG. P5-17 is a diagram of multi-token chunking (Claim 17, dep. on 2), showing the precision-weighted prediction error computation, comparison against a learnable threshold, merge path via weighted running average $\alpha \cdot x_{\text{last}} + (1 - \alpha) \cdot x_{\text{new}}$, and new episode creation path.

FIG. P5-18 is a diagram of error-driven forgetting (Claim 18, dep. on 2), showing the dual-threshold contradiction detection requiring both high cue similarity (same context) and low content similarity (conflicting information), with the action of zeroing the old memory's surprise score to accelerate eviction.

FIG. P5-19 is a diagram of sparse code hashing for concept keys (Claim 19, dep. on 2), showing DG sparse expansion via ReLU, extraction of top-k most active unit indices, hashing to produce a discrete concept key, and the resulting pattern separation through sparse distributed coding.

FIG. P5-20 is a diagram of the adaptive write gate (Claim 20, dep. on 2), showing the gate input as concatenation of mean-pooled representation and mean precision estimate, sigmoid activation, and positive bias initialisation yielding initial write probability exceeding 0.5.

FIG. P5-21 is a diagram of relevance-based eviction (Claim 21, dep. on 2), showing the eviction scoring function combining surprise magnitude, retrieval count, and age, with lowest-scoring memories (consolidated, frequently-retrieved, low-surprise) evicted first when the buffer exceeds capacity.

FIG. P5-22 is a diagram of episode source metadata (Claim 22, dep. on 2), showing the episode entry structure with source metadata fields: step counter, is_generated boolean flag, and precision at encoding time, enabling source monitoring at retrieval.

FIG. P5-23 is a diagram of the three-loop BG hierarchy (Claim 23, dep. on 3), showing discourse (anterior, goals, no top-down gating), sentence (middle, planning, discourse-gated), and word (posterior, selection, sentence-gated) loops corresponding to the Koechlin cascade.

FIG. P5-24 is a diagram of Boltzmann receptor kinetics (Claim 24, dep. on 3), showing D1 and D2 receptor dose-response curves with learnable EC50 and slope parameters, with D2 having lower EC50 (higher DA sensitivity) providing differential sensitivity to dopamine fluctuations.

FIG. P5-25 is a diagram of cascade gate initialisation (Claim 25, dep. on 3), showing zero-weight gate projections with bias -2.0 (sigmoid ≈ 0.12), ensuring minimal higher-to-lower loop influence at initialisation and gradual emergence of cascade gating through learning.

FIG. P5-26 is a diagram of the BG pathway architecture (Claim 26, dep. on 3), showing the processing pipeline within each D1 and D2 pathway: RMS normalisation, first linear projection to d_{bg}, GeLU activation, second linear projection (initialised to zeros), and multiplicative Boltzmann receptor activation.

FIG. P5-27 is a diagram of the cerebellar forward model (Claim 27, dep. on 3), showing the prediction pathway through GeLU hidden layer, temporal prediction error computation for positions 0 to $T - 2$, causal error correction via right-padding with zeros and zero-initialised error projection, and gated application with gate initialised to sigmoid$(-2.0) \approx 0.12$.

FIG. P5-28 is a system combination diagram (Claim 28, dep. on 1, 2, 3), showing the integration of working memory (volatile, fast, limited capacity), hippocampal episodic memory (persistent, large capacity, surprise-gated), and basal ganglia (hierarchical action selection, DA-driven) with the cortical column network.

FIG. P5-29 is a method flow diagram (Claim 29), showing the nine method steps: (a) NMDA slot update, (b) Tsodyks-Markram dynamics, (c) phonological ring buffer with context, (d) metamemory aggregation, (e) hippocampal encoding with chunking and linking, (f) Bayesian confidence tracking, (g) micro-consolidation replay, (h) hierarchical BG action selection, and (i) cerebellar forward model prediction and correction.

FIG. P5-30 is a diagram of enhanced micro-consolidation (Claim 30, dep. on 29), showing the memory selection scoring formula $\text{score}_i = \text{retrieval_count}_i \times (0.5 + 0.5 \times p_i)$, and the two plasticity safeguards during replay: BCM threshold tracking to prevent runaway potentiation,

and metaplasticity weight-change tracking to prevent plasticity drift.

DRAWING SHEETS

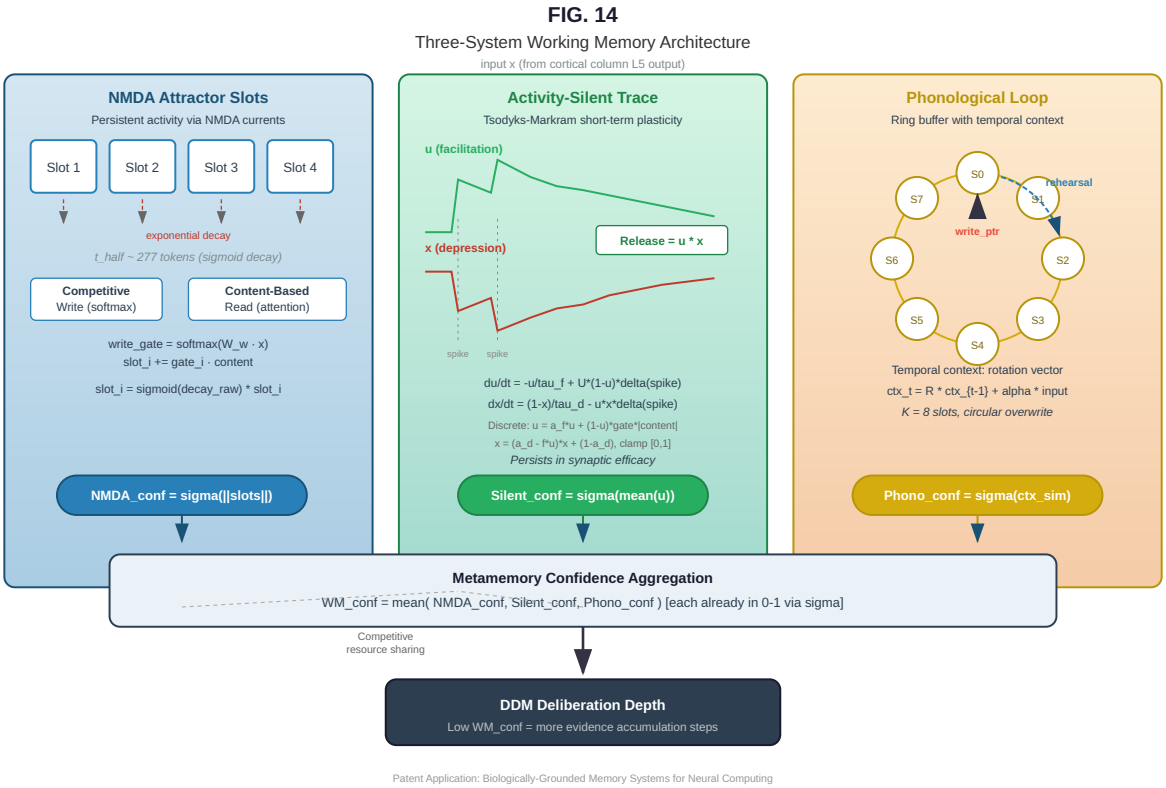

Figure 1: FIG. 1 — Multi-Timescale Working Memory Subsystem (NMDA, Silent Trace, Phonological Loop)

FIG. 15
Hippocampal Episodic Memory System

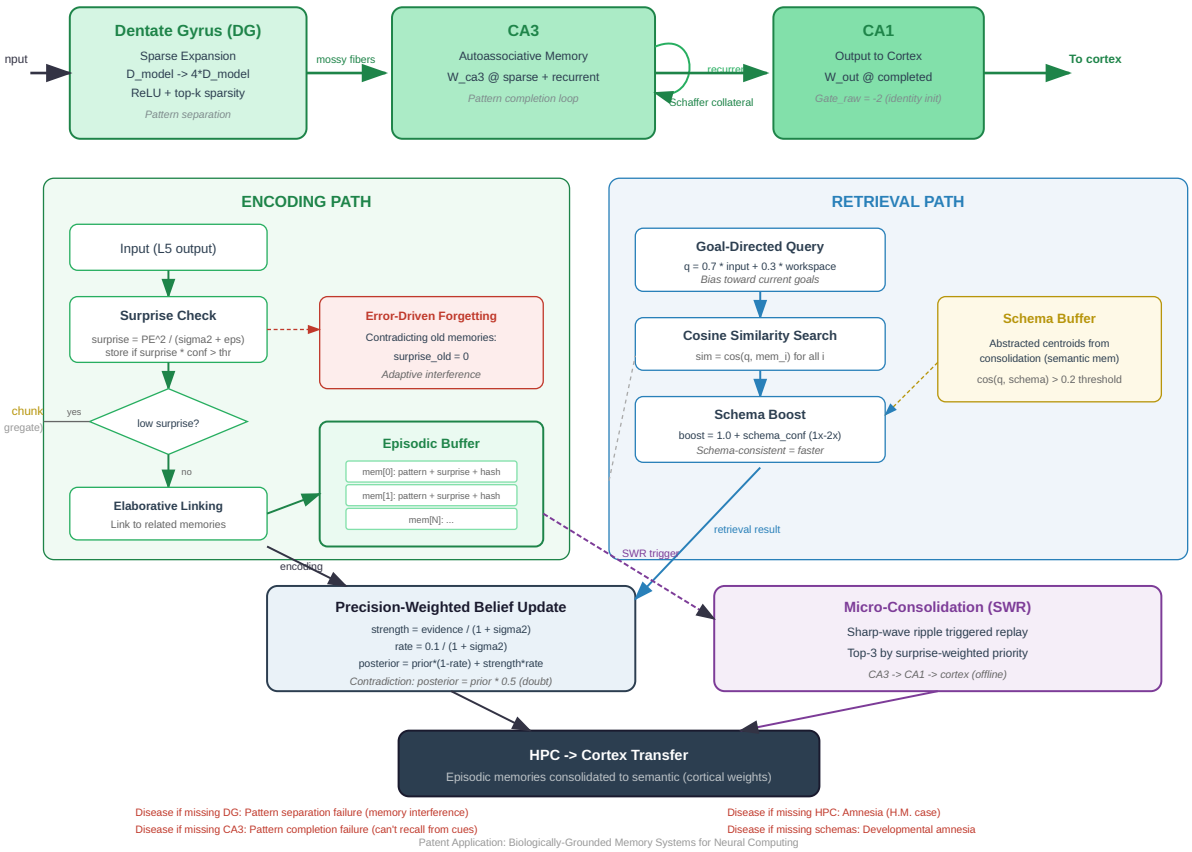

Figure 2: FIG. 2 — Hippocampal Episodic Memory System (DG/CA3/CA1, Episodic Buffer, Bayesian Confidence)

FIG. 16

Hierarchical D1/D2 Basal Ganglia System

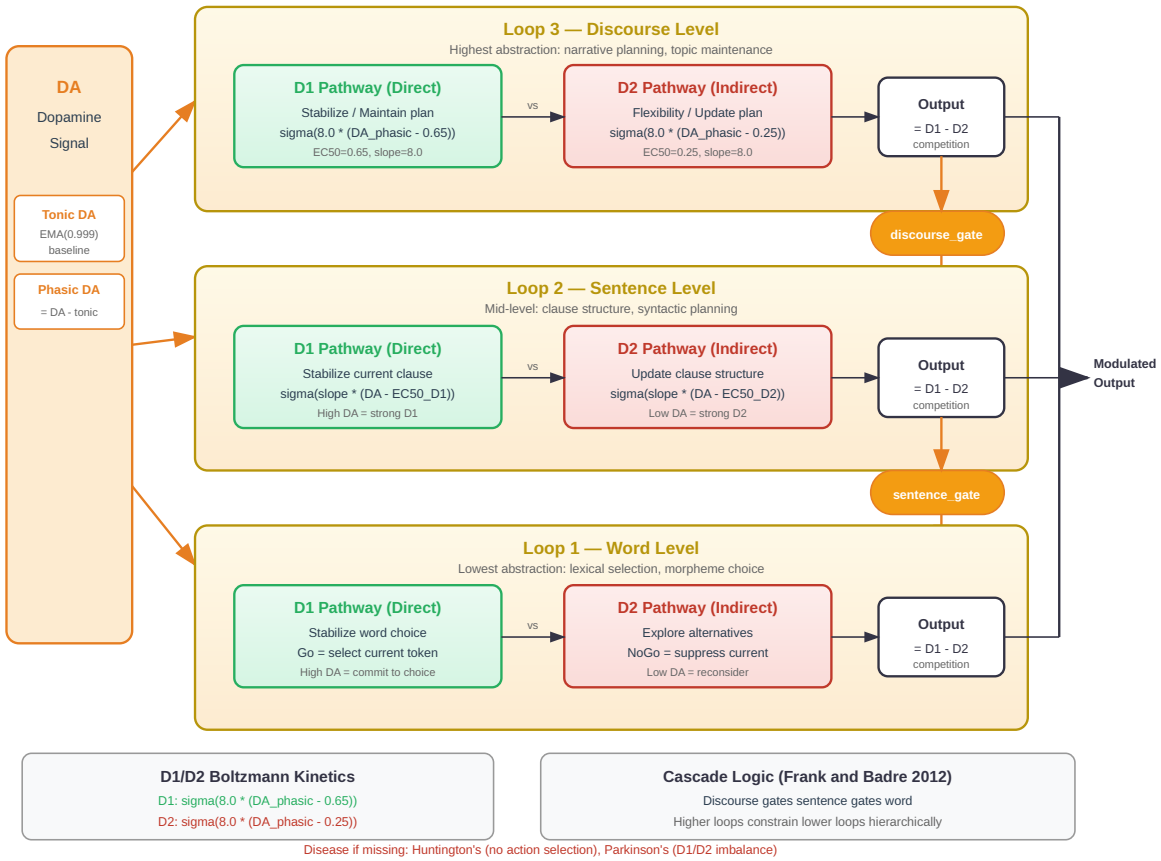

Figure 3: FIG. 3 — Hierarchical Basal Ganglia with D1/D2 Pathways and Cascade Gating

CLAIM ILLUSTRATION SHEETS

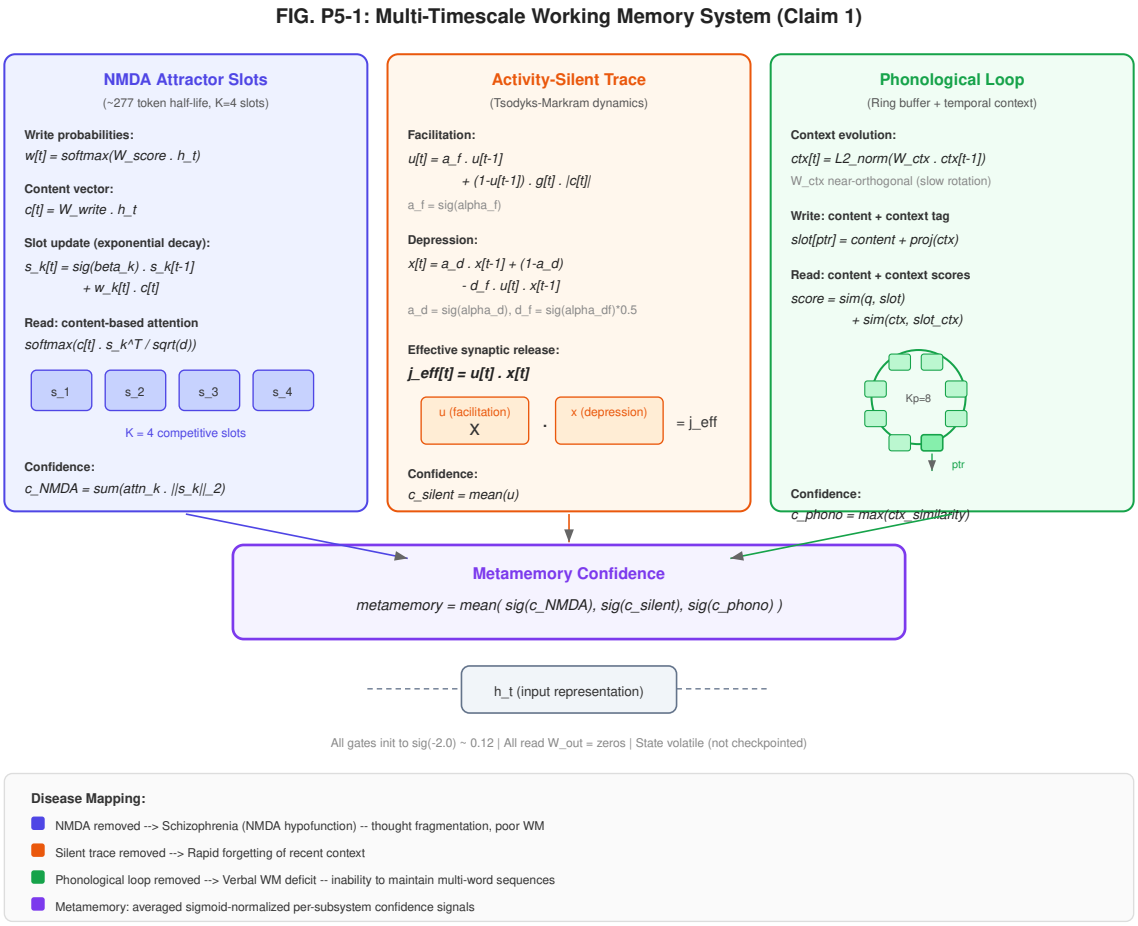

Figure 4: Claim 1 — Multi-Timescale Working Memory System

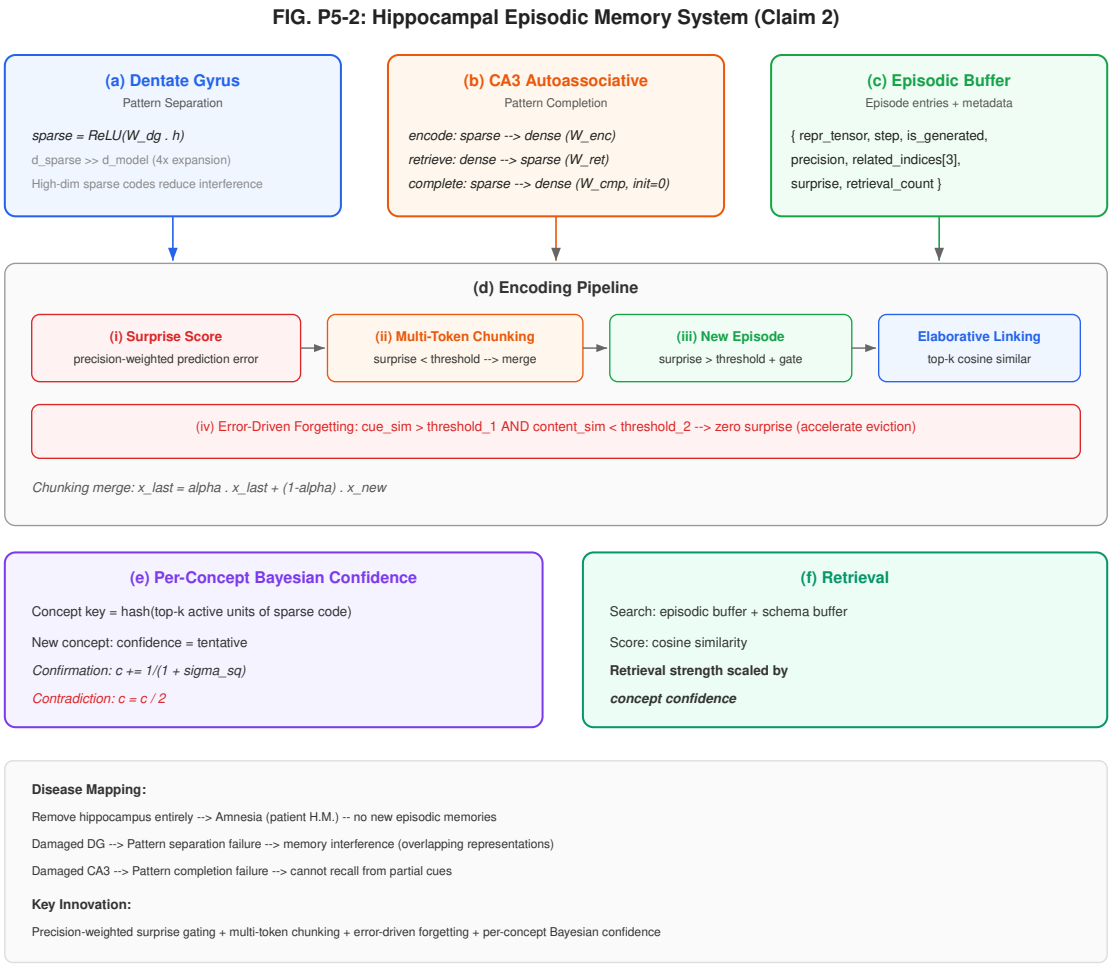

Figure 5: Claim 2 — Hippocampal Episodic Memory System

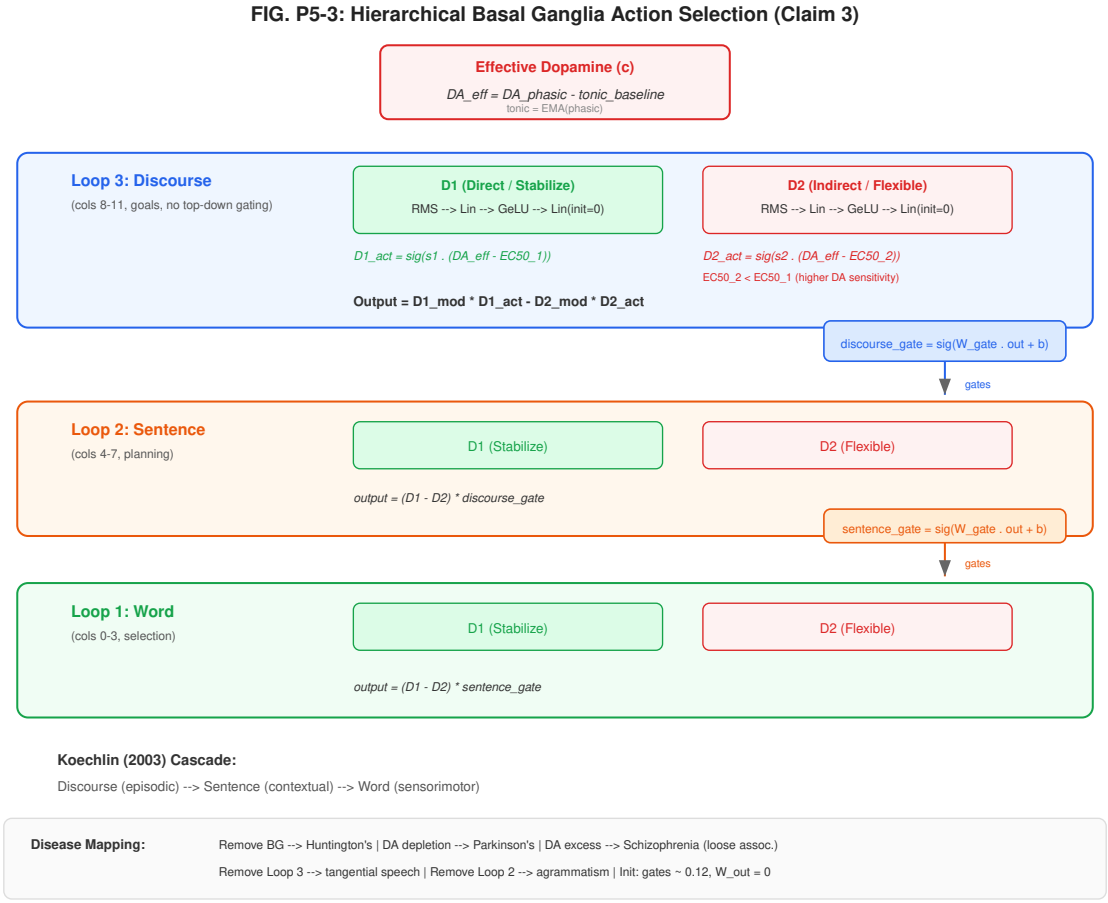

Figure 6: Claim 3 — Hierarchical Basal Ganglia Action Selection

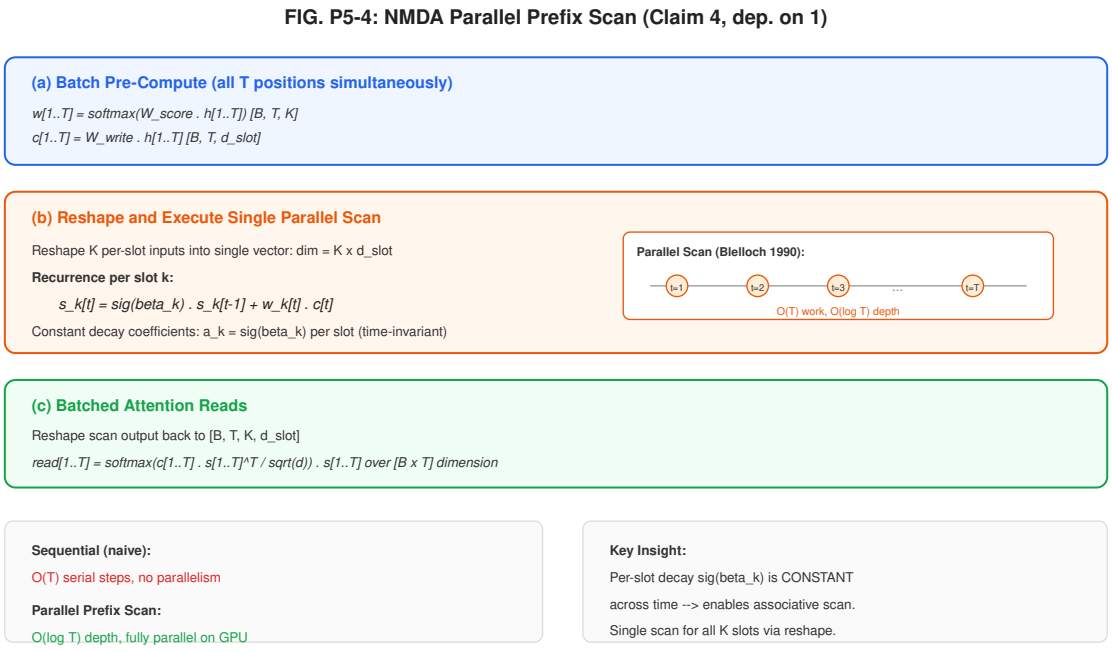

Figure 7: Claim 4 — NMDA Parallel Prefix Scan (dep. on 1)

FIG. P5-5: Dual-Scan Silent Trace (Claim 5, dep. on 1)

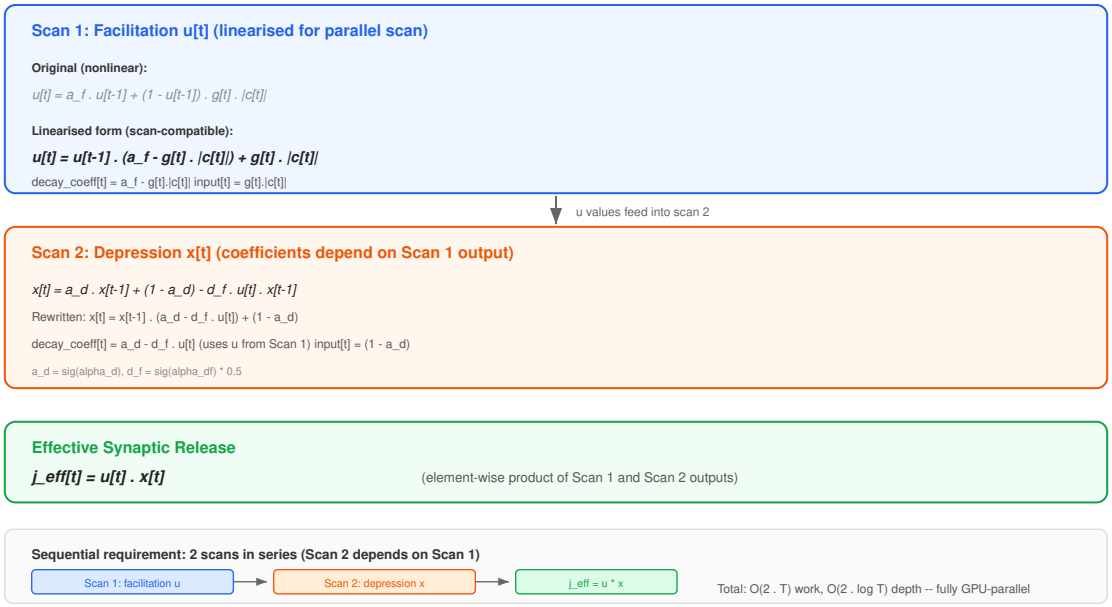

Figure 8: Claim 5 — Dual-Scan Silent Trace Acceleration (dep. on 1)

FIG. P5-6: Phonological Rehearsal Mechanism (Claim 6, dep. on 1)

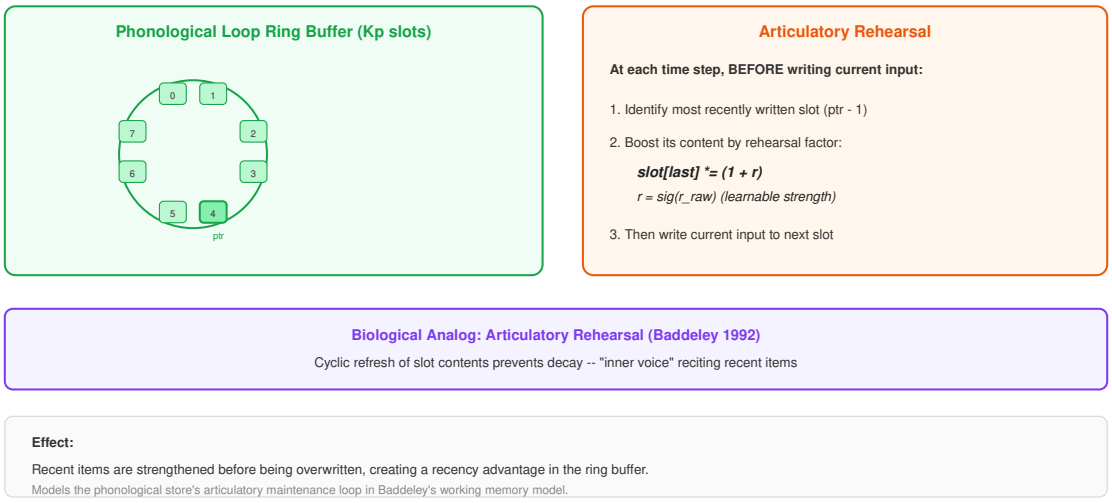

Figure 9: Claim 6 — Phonological Rehearsal Mechanism (dep. on 1)

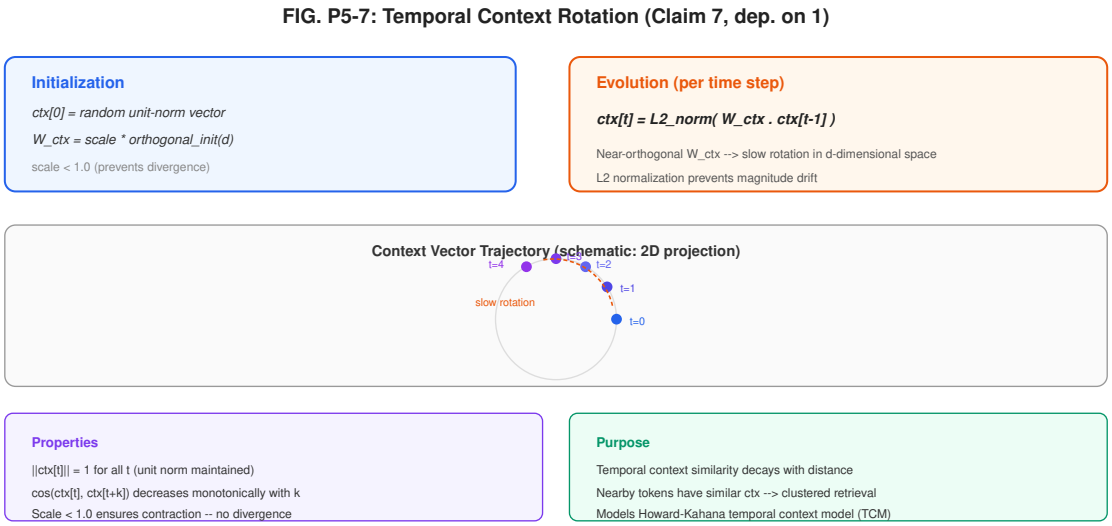

Figure 10: Claim 7 — Temporal Context Rotation (dep. on 1)

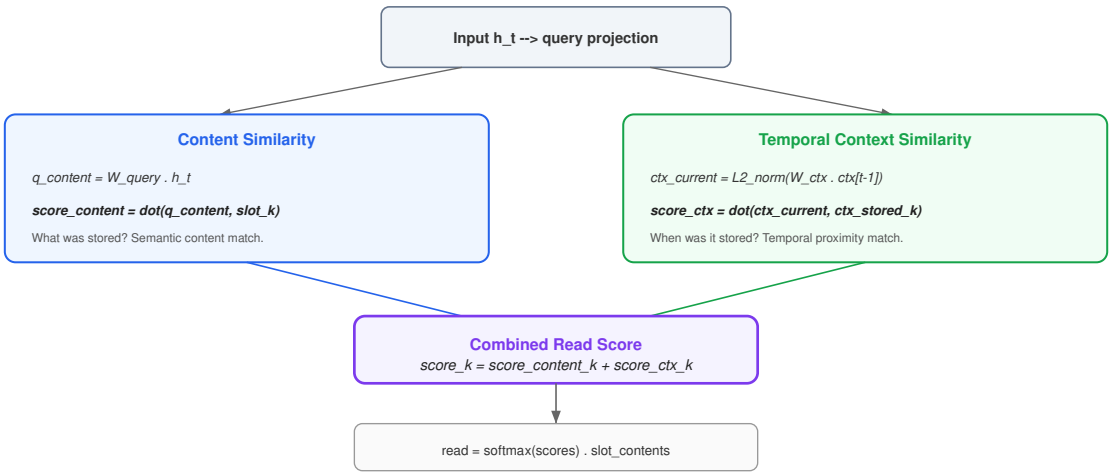

Figure 11: Claim 8 — Content + Context Read Scores (dep. on 1)

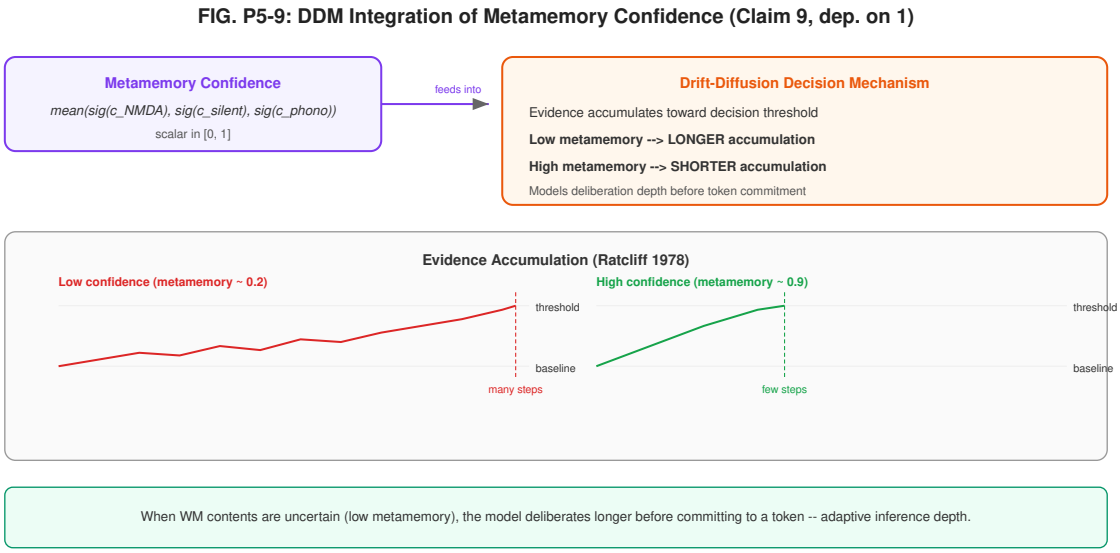

Figure 12: Claim 9 — DDM Integration of Metamemory Confidence (dep. on 1)

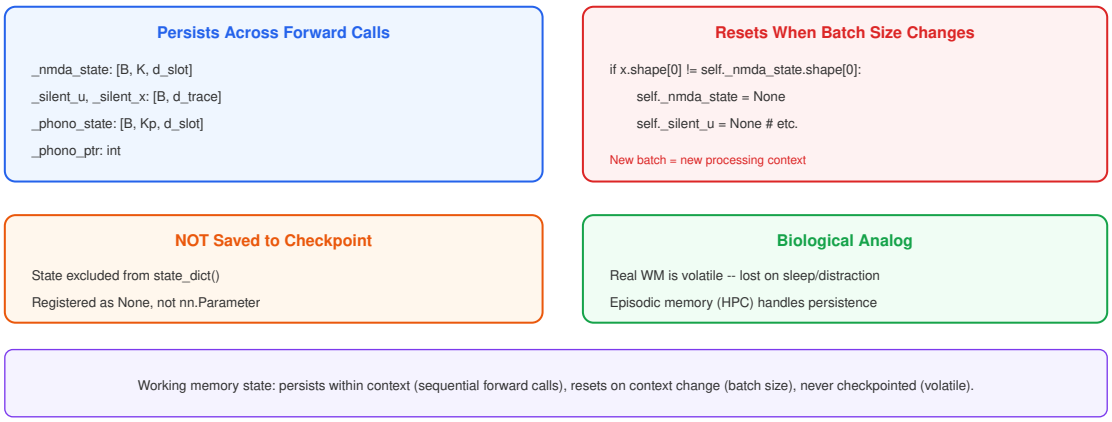

Figure 13: Claim 10 — Volatile Working Memory State (dep. on 1)

FIG. P5-11: Confidence Scaling per Subsystem (Claim 11, dep. on 1)

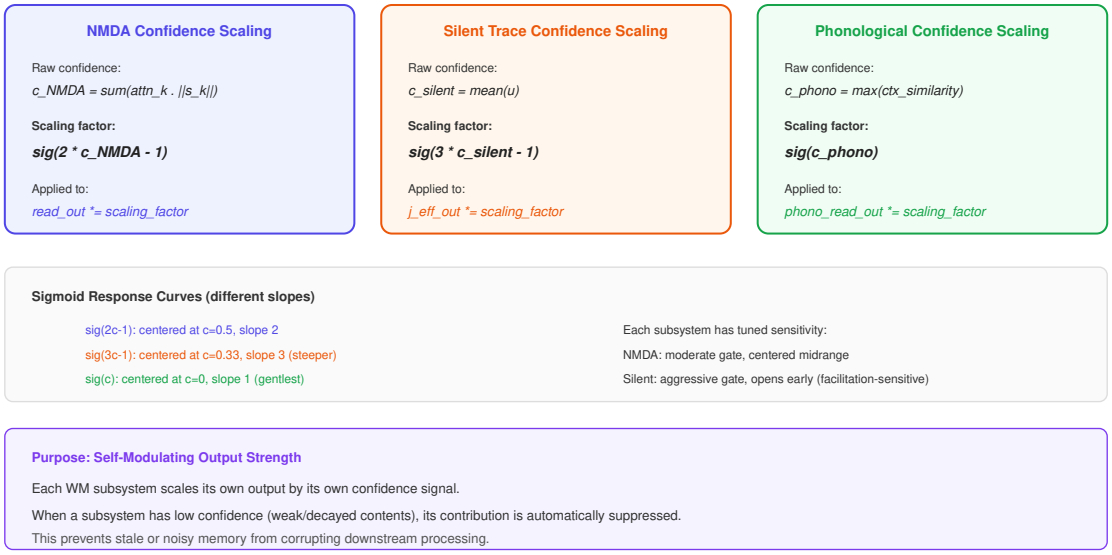

Figure 14: Claim 11 — Confidence Scaling per Subsystem (dep. on 1)

FIG. P5-12: Identity-at-Init for Working Memory (Claim 12, dep. on 1)

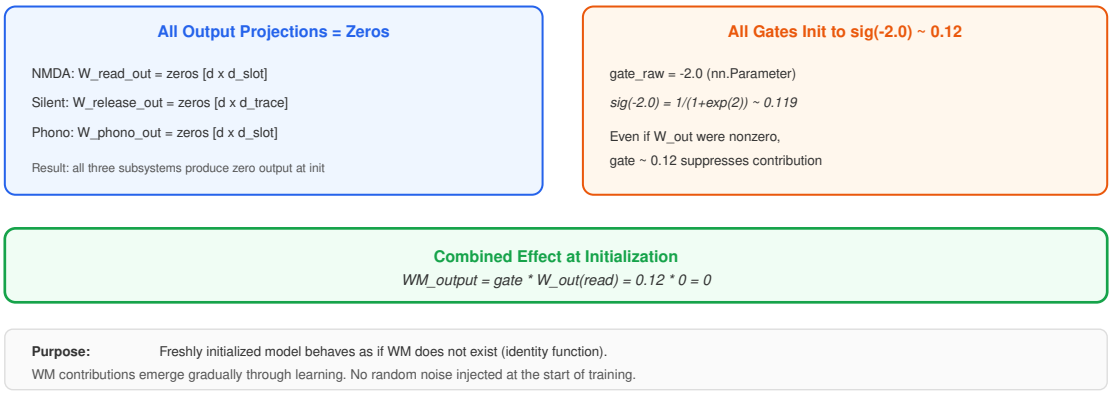

Figure 15: Claim 12 — identity-at-init(Stale from transformer era check with /nct-research) for Working Memory (dep. on 1)

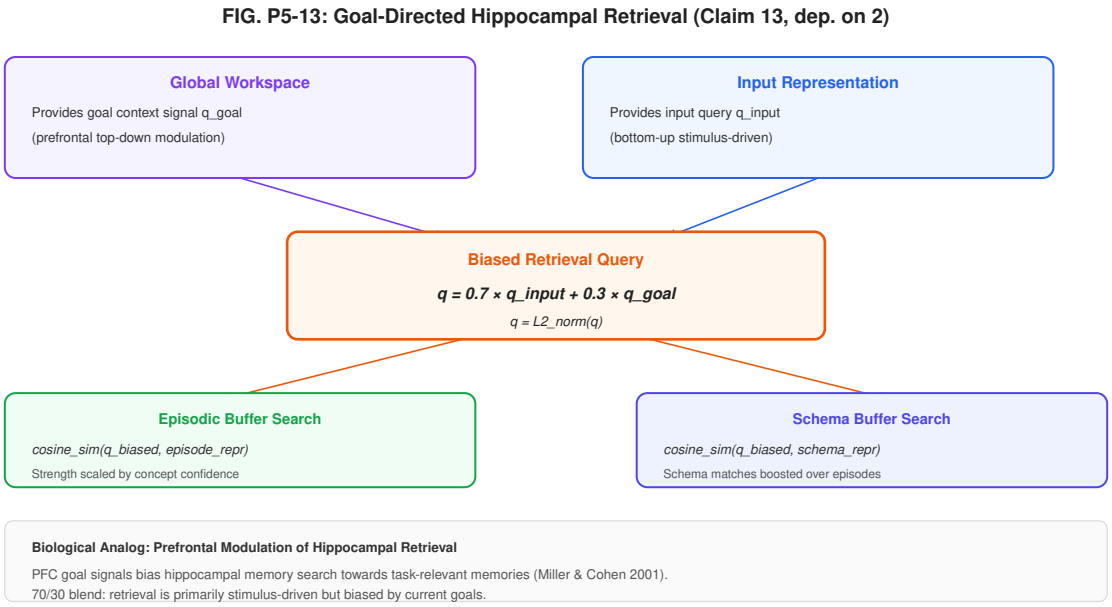

Figure 16: Claim 13 — Goal-Directed Hippocampal Retrieval (dep. on 2)

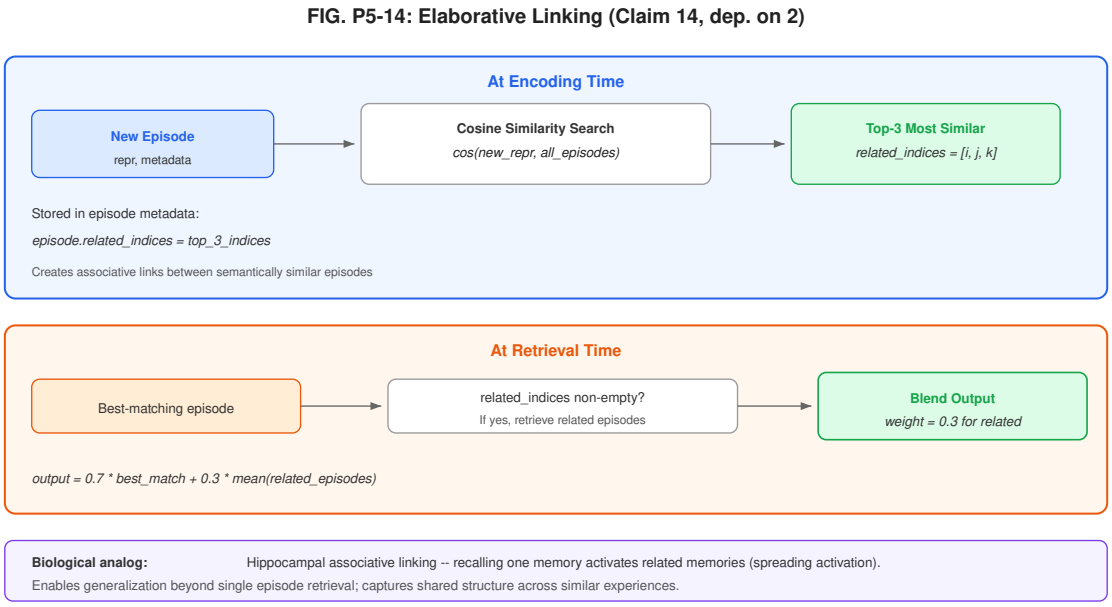

Figure 17: Claim 14 — Elaborative Linking (dep. on 2)

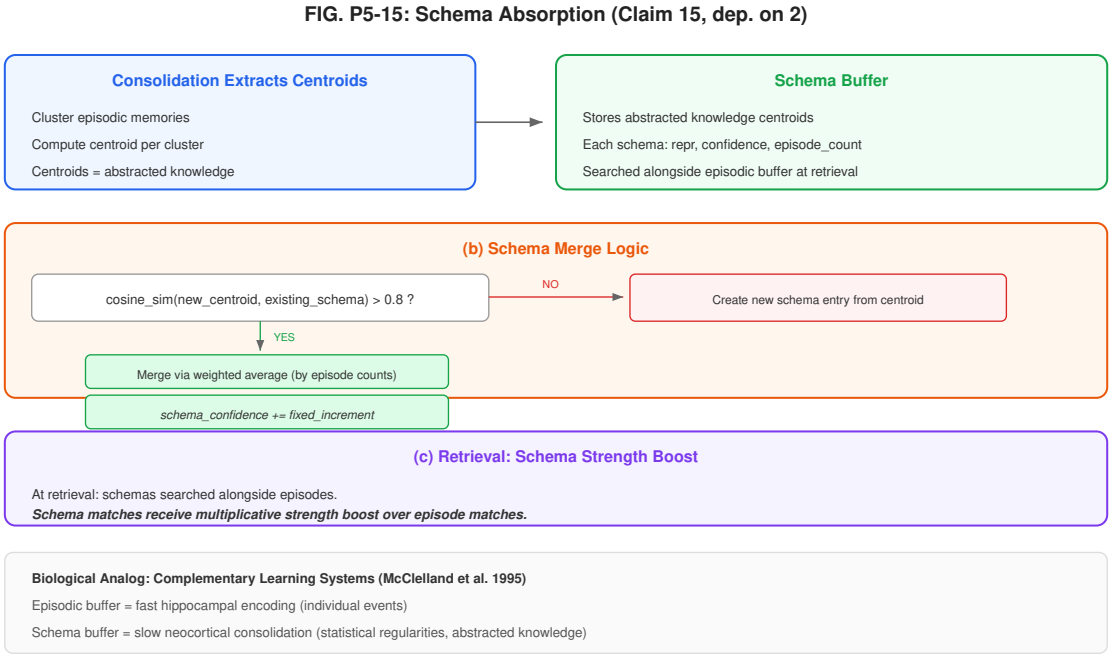

Figure 18: Claim 15 — Schema Absorption (dep. on 2)

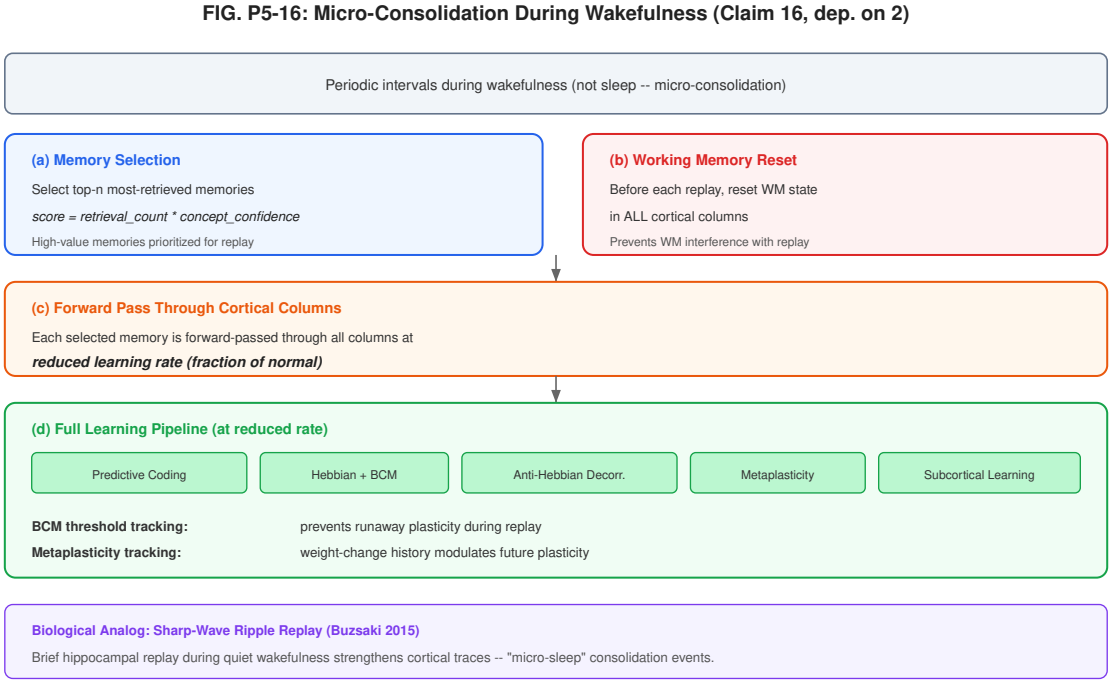

Figure 19: Claim 16 — Micro-Consolidation During Wakefulness (dep. on 2)

FIG. P5-17: Multi-Token Chunking (Claim 17, dep. on 2)

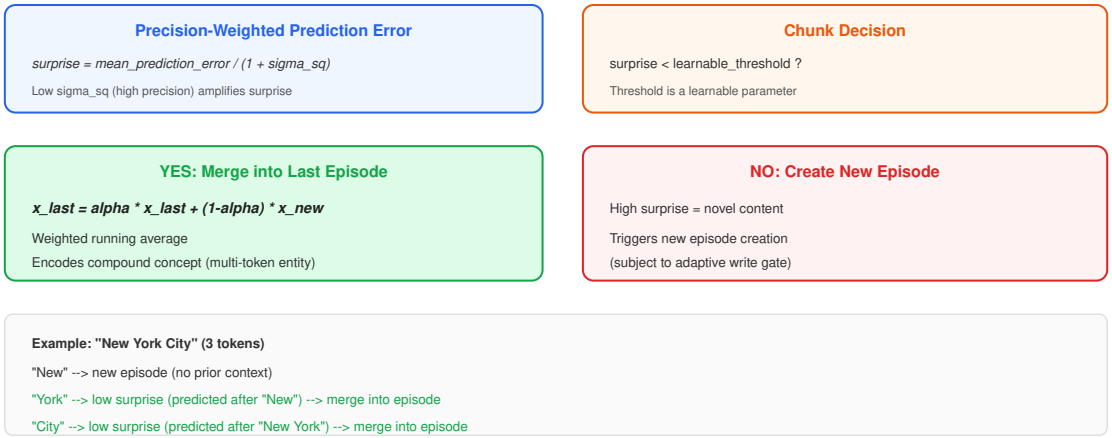

Figure 20: Claim 17 — Multi-Token Chunking (dep. on 2)

FIG. P5-18: Error-Driven Forgetting (Claim 18, dep. on 2)

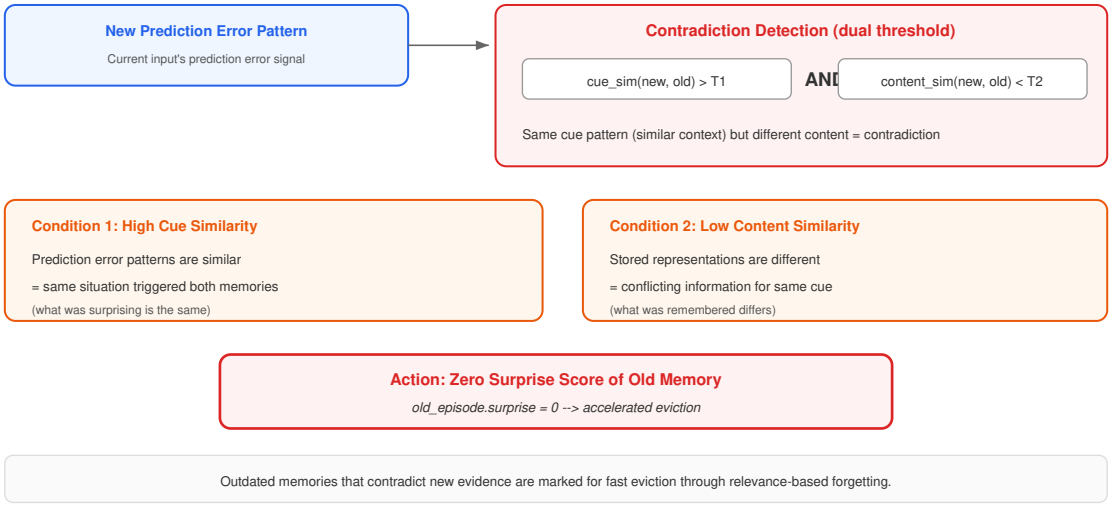

Figure 21: Claim 18 — Error-Driven Forgetting (dep. on 2)

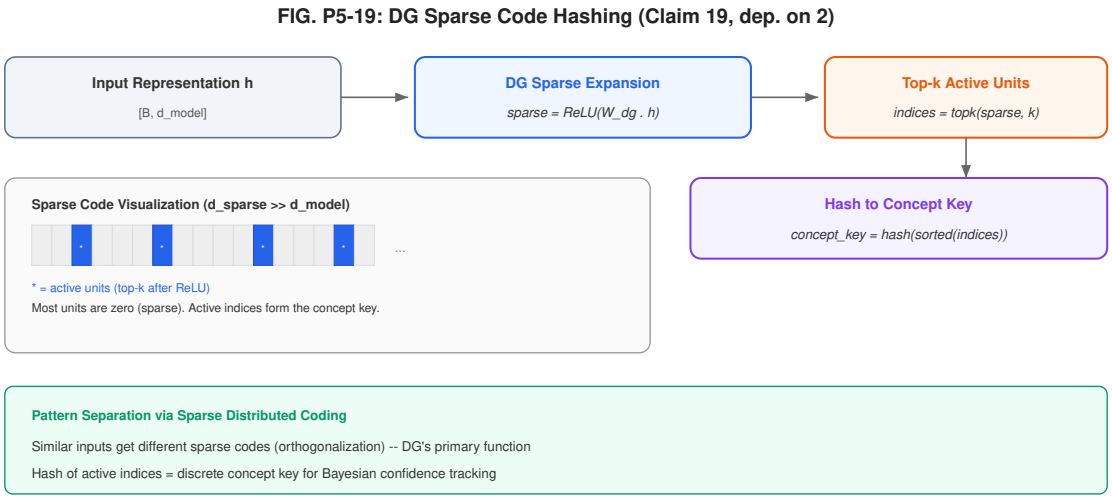

Figure 22: Claim 19 — DG Sparse Code Hashing (dep. on 2)

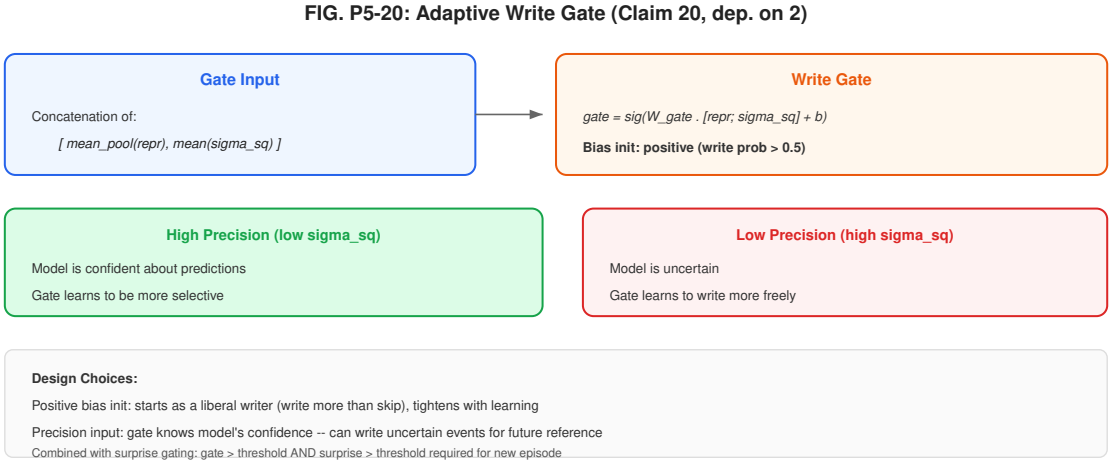

Figure 23: Claim 20 — Adaptive Write Gate (dep. on 2)

FIG. P5-21: Relevance-Based Eviction (Claim 21, dep. on 2)

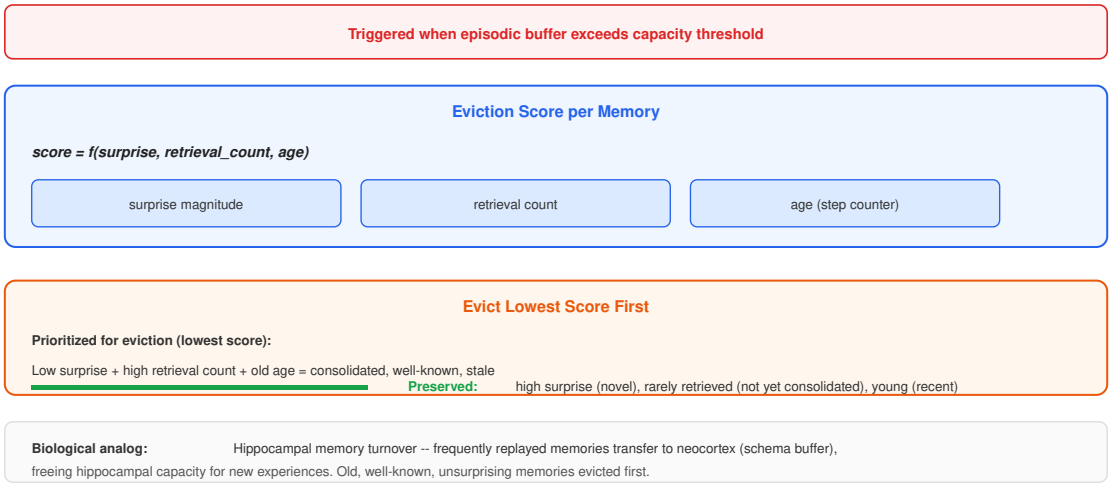

Figure 24: Claim 21 — Relevance-Based Eviction (dep. on 2)

FIG. P5-22: Episode Source Metadata (Claim 22, dep. on 2)

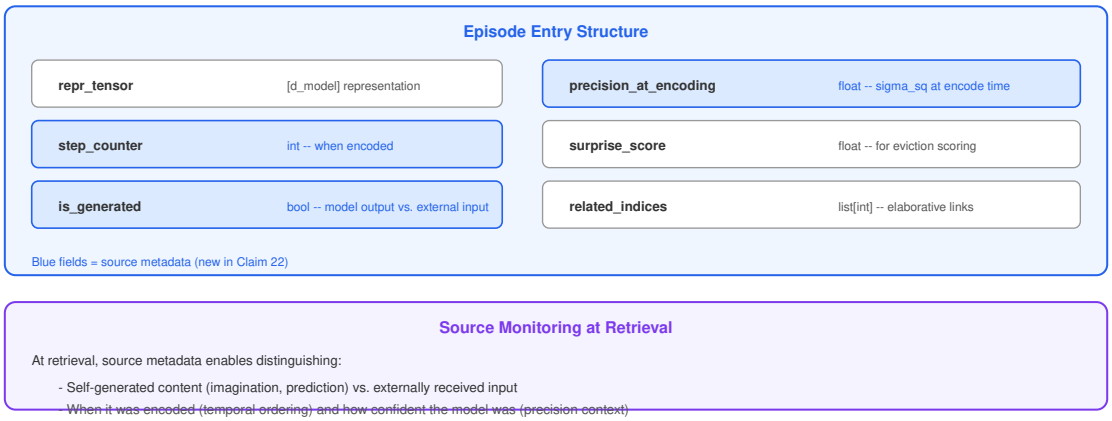

Figure 25: Claim 22 — Episode Source Metadata (dep. on 2)

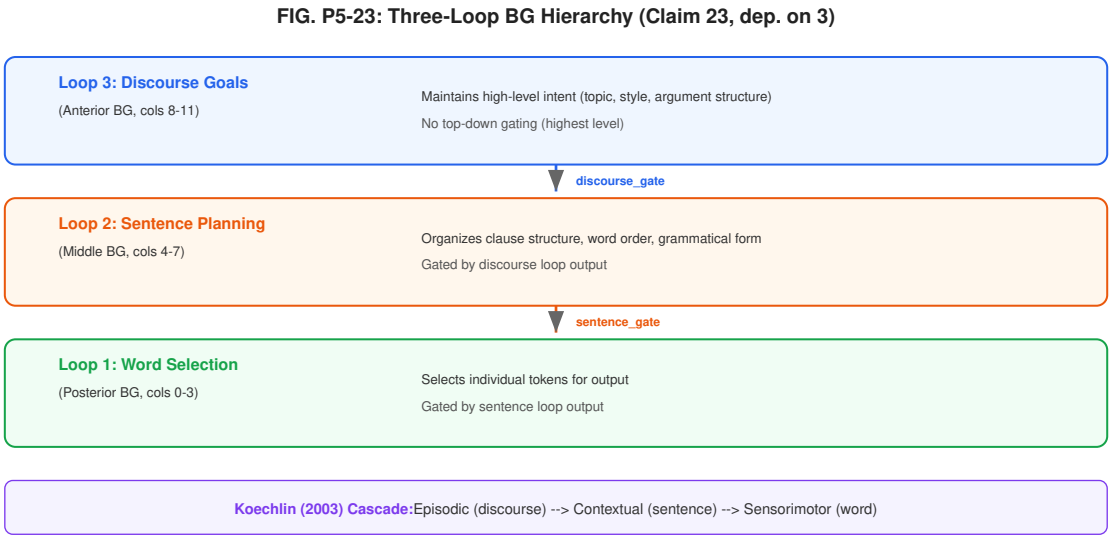

Figure 26: Claim 23 — Three-Loop BG Hierarchy (dep. on 3)

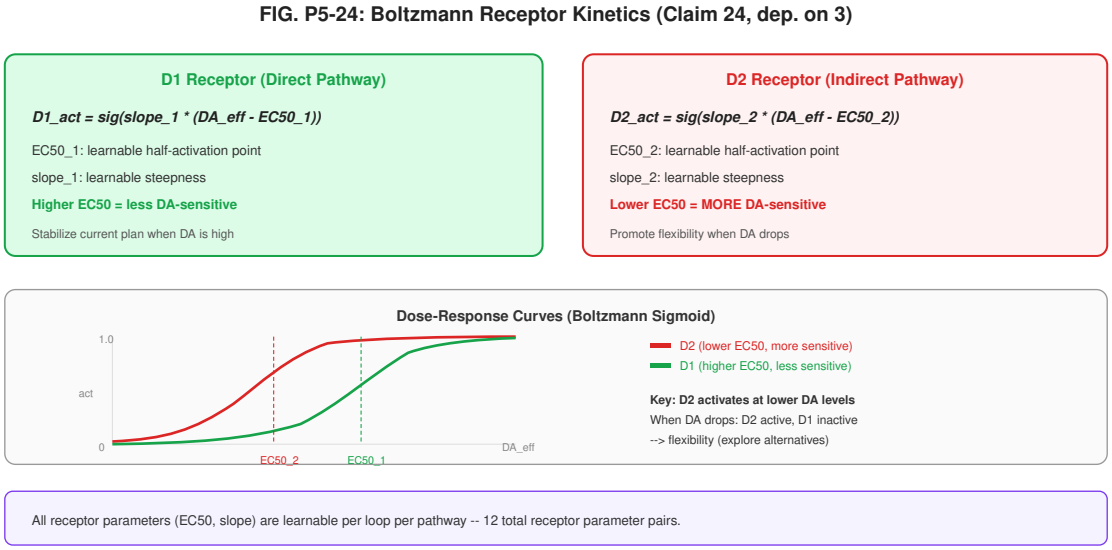

Figure 27: Claim 24 — Boltzmann Receptor Kinetics (dep. on 3)

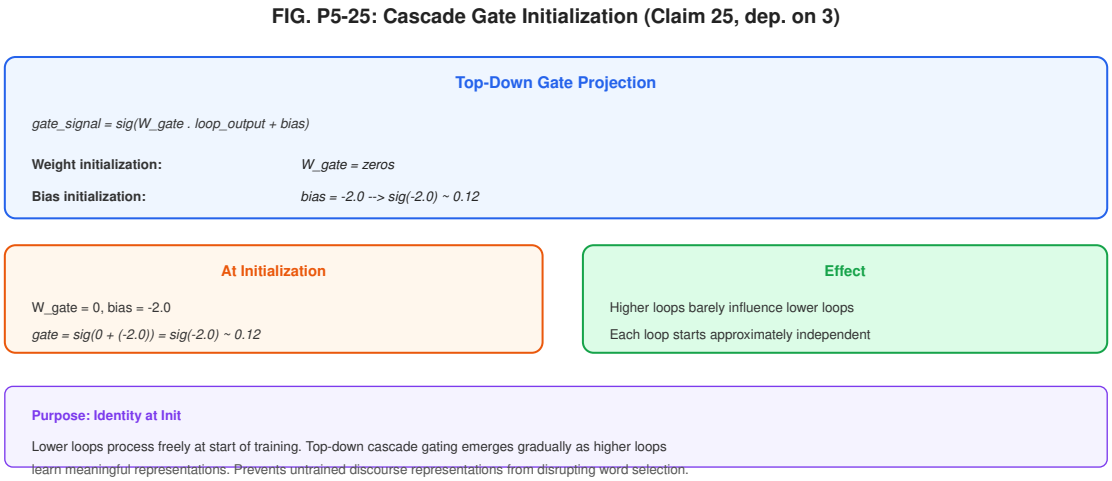

Figure 28: Claim 25 — Cascade Gate Initialisation (dep. on 3)

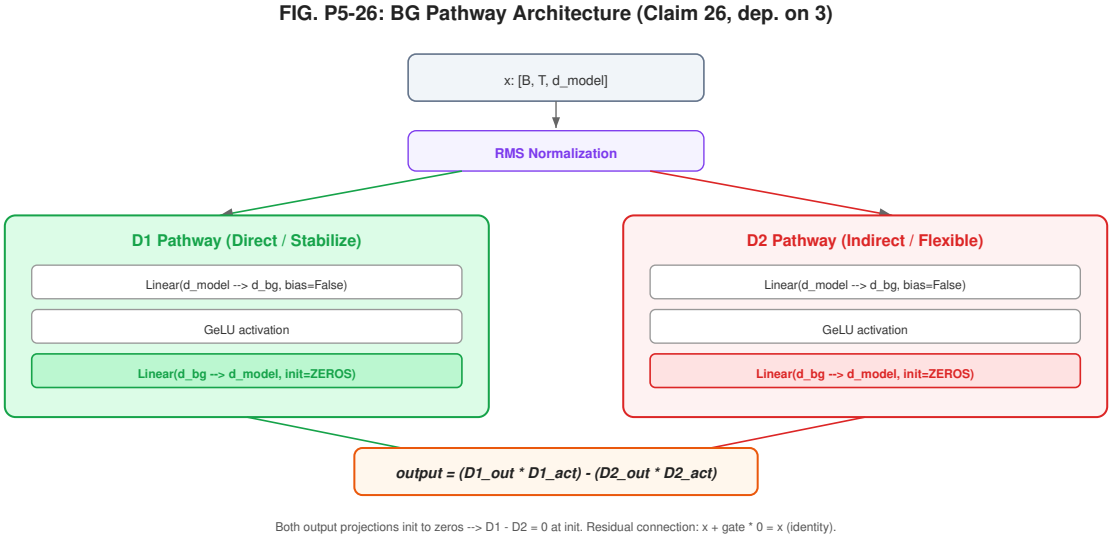

Figure 29: Claim 26 — BG Pathway Architecture (dep. on 3)

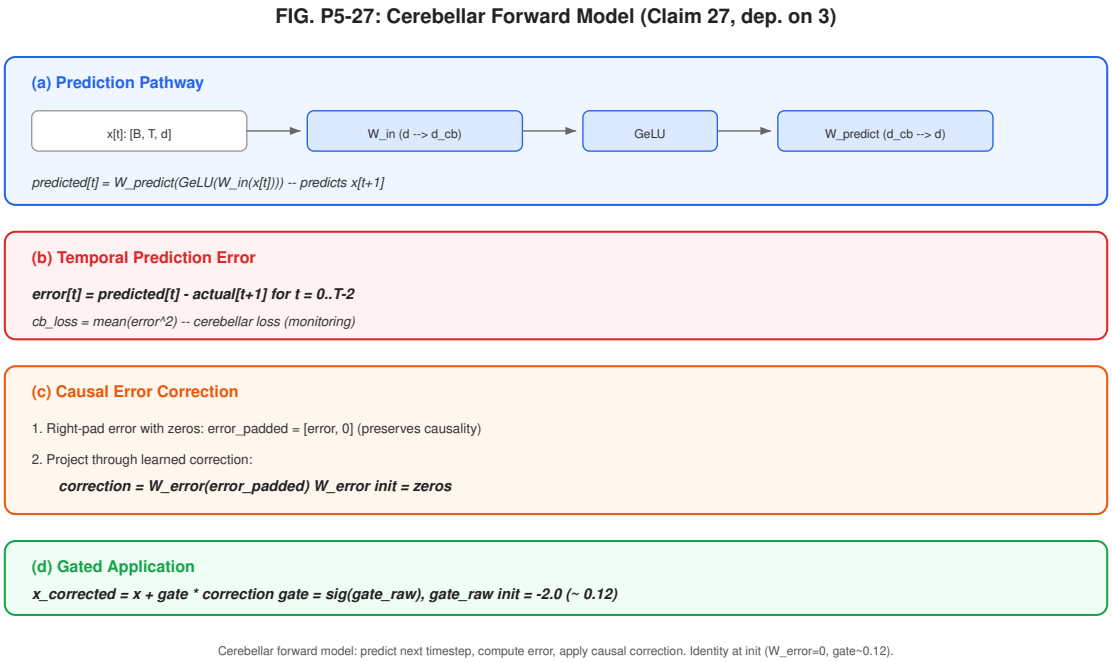

Figure 30: Claim 27 — Cerebellar Forward Model (dep. on 3)

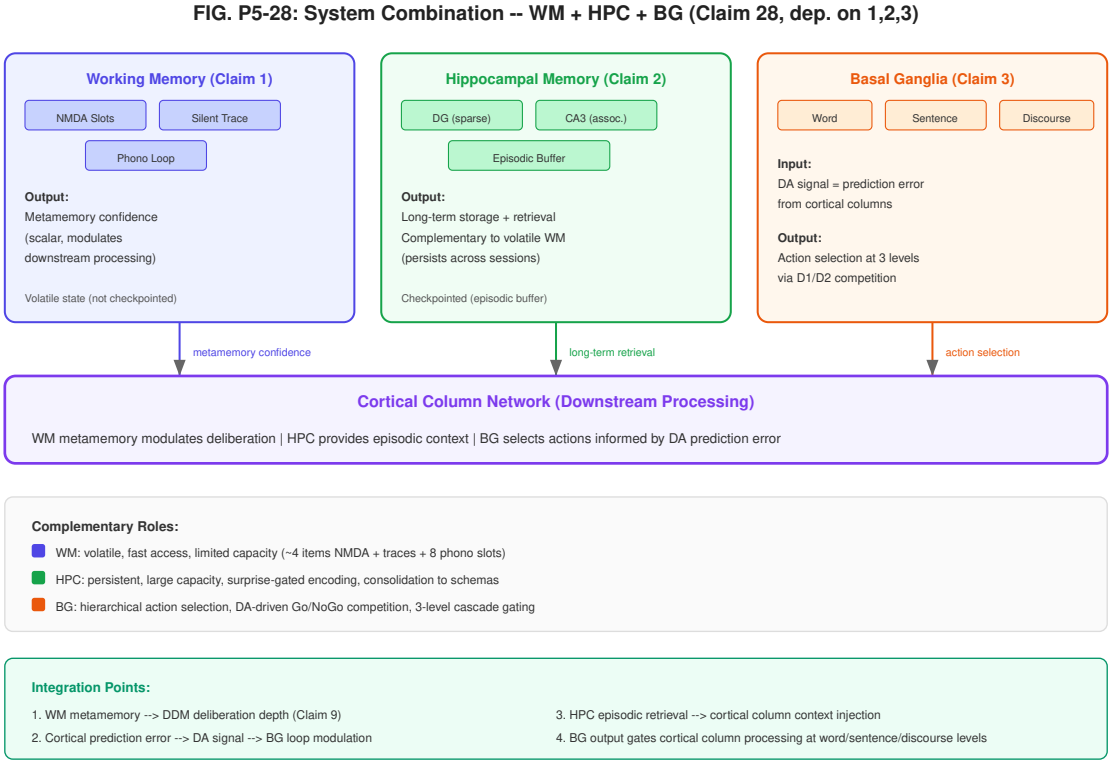

Figure 31: Claim 28 — System Combination: WM + HPC + BG (dep. on 1, 2, 3)

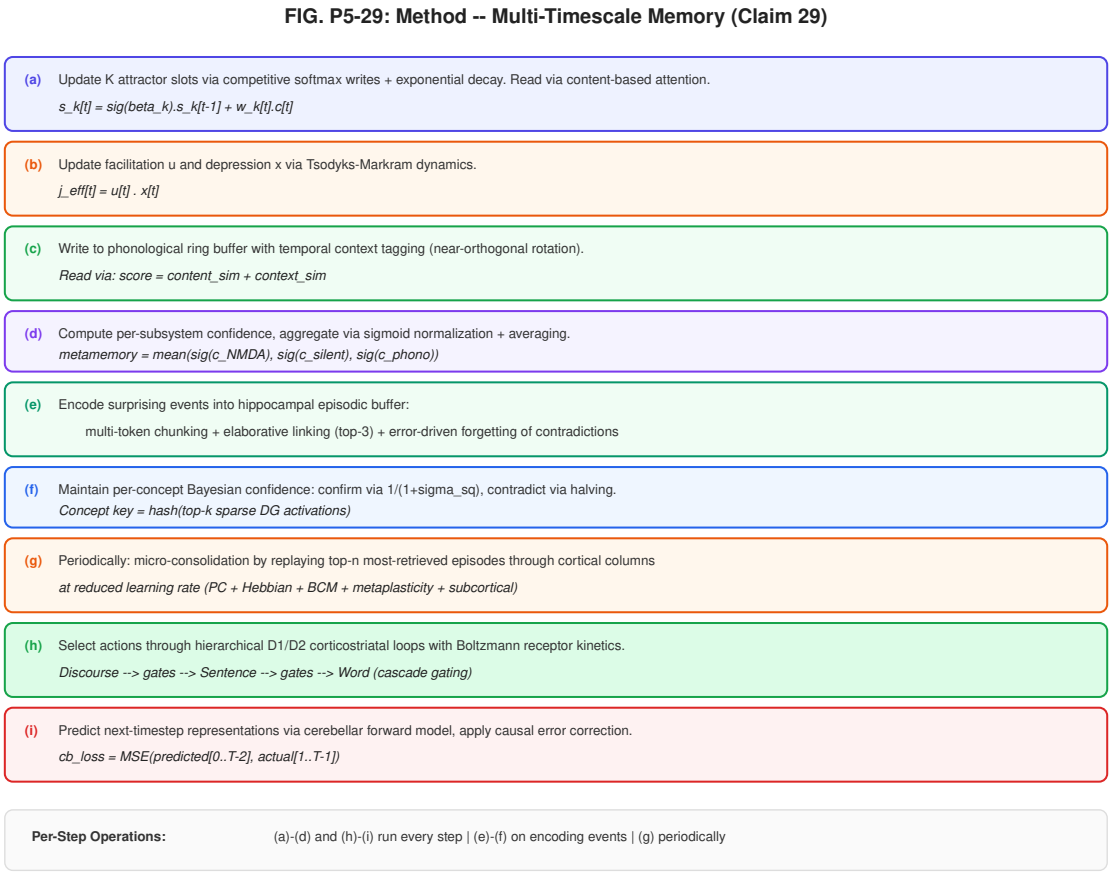

Figure 32: Claim 29 — Method: Multi-Timescale Memory

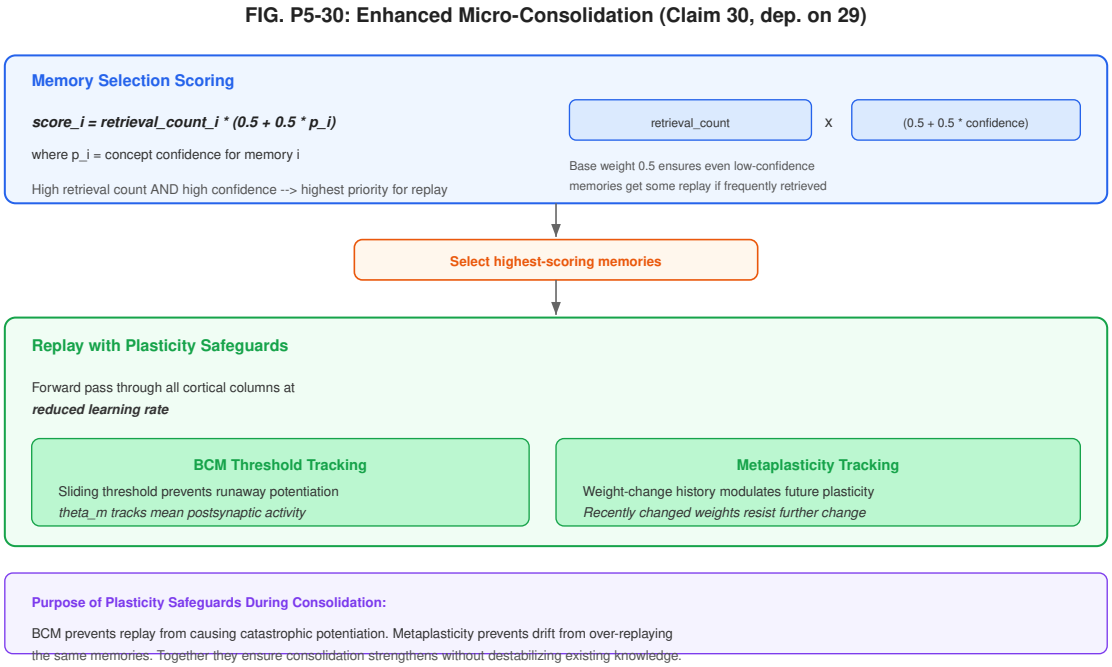

Figure 33: Claim 30 — Enhanced Micro-Consolidation (dep. on 29)

6. DETAILED DESCRIPTION

The following description sets forth the preferred embodiments of the invention. Specific structural and functional details are provided to enable a person skilled in the art to make and use the invention. Numerical values, dimensions, and specific parameter settings are given as exemplary embodiments and are not to be construed as limiting.

6.1. System Architecture Overview

The memory and decision systems of the present invention operate within a cortical column network as described in the co-pending Patent Application 1 of this suite (see FIG. P5-28 for the system combination overview). The cortical column network comprises N cortical column processing units (default $N = 12$), each processing representations of dimension d_{model} (default 512). Each cortical column contains a working memory module as described herein (FIG. P5-1). The hippocampal episodic memory (FIG. P5-2), basal ganglia (FIG. P5-3), and cerebellum (FIG. P5-27) operate at the inter-column (system) level, receiving the composite output of all columns and feeding their contributions back into the processing pipeline.

All modules in the present invention satisfy an *identity-at-initialisation* contract (see FIG. P5-12 for working memory, FIG. P5-25 for basal ganglia cascade gates): at construction time, all output projections are initialised to zero weights and all gate parameters are initialised to -2.0 (yielding sigmoid$(-2.0) \approx 0.12$), ensuring that each subsystem contributes negligibly until its parameters have been trained. This enables safe integration with pre-trained models without disrupting existing representations.

6.2. NMDA Attractor Slots — Detailed Implementation

6.2.1. Biological Basis

The NMDA attractor slot mechanism models persistent activity in prefrontal cortex mediated by NMDA receptor plateau potentials. Electrophysiological recordings demonstrate that prefrontal neurons maintain delay-period activity for approximately 4 items (Luck & Vogel, 1997; Cowan, 2001), with individual items maintained in discrete attractor states that resist interference through competitive inhibition. The NMDA receptor's slow kinetics (decay time constant $\sim$100 ms) provide the biophysical substrate for sustained activity over seconds-long delays.

6.2.2. Slot State Update

The NMDA system maintains $K = 4$ slots of dimension d_{slot} (default 64) (see FIG. P5-1 for the complete NMDA architecture). At each time step t, the column state $\mathbf{h}_t \in \mathbb{R}^d$ is first normalised via RMS normalisation, then processed to produce:

- Write probabilities: $\mathbf{w}[t] = \text{softmax}(\mathbf{W}_{\text{score}} \cdot \text{RMSNorm}(\mathbf{h}_t)) \in \mathbb{R}^K$

- Content: $\mathbf{c}[t] = \mathbf{W}_{\text{write}} \cdot \text{RMSNorm}(\mathbf{h}_t) \in \mathbb{R}^{d_{\text{slot}}}$

Each slot k updates according to Equation 1. The softmax competition over write probabilities ensures that at most one slot receives a strong write signal at any time step, implementing the competitive inhibition observed in prefrontal delay-period activity.

6.2.3. Content-Based Read

Reading employs content-based attention over slot contents:

$$\text{score}_k[t] = \mathbf{s}_k[t]^\top \cdot \mathbf{c}[t] \tag{9}$$

$$\mathbf{r}_{\text{weights}}[t] = \text{softmax}(\text{score}[t]) \in \mathbb{R}^K \tag{10}$$

$$\mathbf{r}[t] = \sum_k r_k[t] \cdot \mathbf{s}_k[t] \tag{11}$$

6.2.4. NMDA Confidence

The NMDA confidence signal measures how strongly the attended slots are activated:

$$c_{\text{NMDA}}[t] = \sum_k r_k[t] \cdot \|\mathbf{s}_k[t]\|_2 \tag{12}$$

This is the read-attention-weighted sum of slot norms. High slot norms indicate recently written or well-maintained content; low norms indicate decayed content. The confidence is further transformed to a scaling factor $\text{sigmoid}(2 \cdot c_{\text{NMDA}} - 1)$ that modulates the read output (see FIG. P5-11 for the confidence-scaled read architecture), ensuring strong memories contribute more than fading ones.

6.2.5. Parallel Scan Acceleration

The slot recurrence in Equation 1 takes the form $\mathbf{h}_t = a_t \cdot \mathbf{h}_{t-1} + b_t$ where $a_t = \text{sigmoid}(\beta_k)$ is constant (input-independent) and $b_t = w_k[t] \cdot \mathbf{c}[t]$ is pre-computable from the input. This admits an efficient parallel prefix scan (Blelloch, 1990) computed in $O(\log T)$ parallel steps rather than $O(T)$ sequential steps (see FIG. P5-4 for the three-stage parallel scan pipeline).

The implementation proceeds in three stages:

1. **Batch pre-compute**: Compute all write probabilities and contents for all T positions via batched matrix multiplications, producing tensors of shape $[B, T, K]$ and $[B, T, d_{\text{slot}}]$.
2. K **parallel scans**: Reshape to $[B, T, K \times d_{\text{slot}}]$ and execute a single parallel scan over the concatenated state, with per-slot constant decay expanded across the d_{slot} dimension.
3. **Batch reads**: Reshape scan output to $[B, T, K, d_{\text{slot}}]$ and compute all reads via batched matrix multiplications over the $B \times T$ dimension.

A parity contract ensures that $|\text{parallel output} - \text{sequential output}| < 10^{-4}$, with the relaxed tolerance accounting for cumulative floating-point differences over T steps.

6.3. Activity-Silent Trace — Detailed Implementation

6.3.1. Biological Basis

The activity-silent trace models the Tsodyks-Markram short-term synaptic plasticity mechanism (Mongillo et al., 2008; Stokes, 2015). Working memory information can be maintained in the pattern of facilitated and depressed synapses even when neural firing ceases entirely. A probe stimulus (the next input token) reads out this hidden memory trace through the pattern of synaptic release probabilities.

6.3.2. Facilitation and Depression Dynamics

The silent trace system maintains two state variables per dimension of the trace space (d_{trace}, default 64) (see FIG. P5-5 for the dual parallel scan architecture):

Facilitation $u \in [0,1]^{d_{\text{trace}}}$ models residual calcium in presynaptic terminals:

$$u[t] = \text{sigmoid}(\alpha_f) \cdot u[t-1] + (1 - u[t-1]) \cdot g[t] \cdot |\mathbf{c}[t]| \tag{13}$$

where $g[t] = \text{sigmoid}(\mathbf{W}_g \cdot \text{RMSNorm}(\mathbf{h}_t))$ is a learned write gate (initialised with bias -1.0 for conservative writing) and $\mathbf{c}[t] = \mathbf{W}_{\text{store}} \cdot \text{RMSNorm}(\mathbf{h}_t)$ is the stored content.

Depression $x \in [0,1]^{d_{\text{trace}}}$ models neurotransmitter vesicle availability:

$$x[t] = \text{sigmoid}(\alpha_d) \cdot x[t-1] + (1 - \text{sigmoid}(\alpha_d)) - \text{sigmoid}(\alpha_{df}) \cdot 0.5 \cdot u[t] \cdot x[t-1] \tag{14}$$

where α_d and α_{df} are learnable parameters. The depression variable is clamped to $[0,1]$ at each step (biological constraint: vesicle availability cannot be negative or exceed capacity).

The **effective synaptic release** is $j_{\text{eff}}[t] = u[t] \cdot x[t]$, projected through a read-out projection $\mathbf{W}_{\text{retrieve}}$ (initialised to zeros).

6.3.3. Two-Scan Parallel Acceleration

The facilitation equation is linearised as:

$$u[t] = u[t-1] \cdot \underbrace{(\text{sigmoid}(\alpha_f) - g[t] \cdot |\mathbf{c}[t]|)}_{a_t^{(u)}} + \underbrace{g[t] \cdot |\mathbf{c}[t]|}_{b_t^{(u)}} \tag{15}$$

This is a first-order linear recurrence with input-dependent coefficients, computed via a parallel prefix scan (Scan 1).

The depression equation, after substitution, becomes:

$$x[t] = x[t-1] \cdot \underbrace{(2\,\text{sigmoid}(\alpha_d) - 1 - \text{sigmoid}(\alpha_{df}) \cdot 0.5 \cdot u[t])}_{a_t^{(x)}} + \underbrace{(1 - \text{sigmoid}(\alpha_d))}_{b_t^{(x)}} \tag{16}$$

The coefficient $a_t^{(x)}$ depends on $u[t]$ from Scan 1, so this must be computed as Scan 2 after Scan 1 completes. The two scans reflect biological independence: facilitation (calcium dynamics) and depression (vesicle dynamics) are physically separate processes.

6.3.4. Silent Trace Confidence

Confidence is the mean facilitation level: $c_{\text{silent}}[t] = \frac{1}{d_{\text{trace}}} \sum_i u_i[t]$. High facilitation indicates recently accessed memory (confident retrieval); low facilitation indicates a faded trace (uncertain retrieval). The confidence modulates the read-out: $\text{out}[t] = \mathbf{W}_{\text{retrieve}} \cdot j_{\text{eff}}[t] \cdot \text{sigmoid}(3 \cdot c_{\text{silent}} - 1)$.

6.4. Phonological Loop — Detailed Implementation

6.4.1. Biological Basis

The phonological loop models the articulatory rehearsal system described by Baddeley (1992). Verbal working memory maintains approximately 7 ± 2 items (Miller, 1956) through cyclic refresh of an internal speech buffer. Burgess and Hitch (2006) demonstrated that serial order in the phonological store is maintained through a temporal context signal—a slowly evolving distributed representation that tags each item with its ordinal position.

6.4.2. Ring Buffer with Context Tagging

The phonological loop maintains $K_p = 8$ slots of dimension d_p (default 64) in a circular ring buffer. A temporal context vector $\mathbf{ctx} \in \mathbb{R}^{d_{\text{ctx}}}$ (where $d_{\text{ctx}} = \max(d_p/2, 4)$) evolves at each step via a learned near-orthogonal rotation (Equation 4).

At each time step:

1. **Context evolution**: **ctx** is rotated and renormalised (see FIG. P5-7 for the near-orthogonal rotation mechanism).
2. **Passive decay**: All slots are multiplied by $\text{sigmoid}(\gamma_{\text{decay}})$ (default $\text{sigmoid}(4.0) \approx 0.982$).
3. **Rehearsal**: The most recently written slot is boosted (see FIG. P5-6 for the rehearsal mechanism): $\mathbf{s} \leftarrow \mathbf{s} + \text{sigmoid}(r_{\text{raw}}) \cdot \mathbf{s}$ (default $\text{sigmoid}(-1.0) \approx 0.27$).
4. **Context-tagged write**: Content $\mathbf{c} = \mathbf{W}_{\text{write}} \cdot \mathbf{h}_t$ is tagged with context: $\mathbf{c}_{\text{tagged}} = \mathbf{c} + \mathbf{W}_{\text{ctx_proj}} \cdot \mathbf{ctx}$. The tagged content is written to the current ring buffer position, gated by a learned write gate.
5. **Content+context read**: Read scores combine content similarity and temporal context similarity (see FIG. P5-8 for the dual-score read mechanism): $\text{score}_k = \mathbf{s}_k^\top \mathbf{q} + \mathbf{ctx}_k^\top \mathbf{ctx}_{\text{current}}$, followed by softmax attention and weighted read.

6.4.3. Phonological Confidence

Confidence is the maximum context similarity across slots: $c_{\text{phono}}[t] = \max_k(\mathbf{ctx}_k^\top \cdot \mathbf{ctx}_{\text{current}})$. High context similarity indicates reading from a well-ordered position (confident serial recall);

low similarity indicates positional confusion.

6.5. Metamemory Confidence Aggregation

The three per-subsystem confidence signals are aggregated into a single metamemory scalar:

$$c_{\text{mem}} = \frac{1}{|S|} \sum_{s \in S} \text{sigmoid}(c_s) \tag{17}$$

where S is the set of active subsystems and c_s is the raw confidence for subsystem s. If no subsystem is active, $c_{\text{mem}} = 0.5$ (neutral). This scalar is consumed by the drift-diffusion decision mechanism in generation (see FIG. P5-9 for the DDM integration) to modulate deliberation depth: low metamemory confidence triggers longer evidence accumulation before token commitment.

6.6. Hippocampal Episodic Memory — Detailed Implementation

6.6.1. Architecture Overview

The hippocampal system comprises parametric components (DG, CA3, CA1 projections) and a non-parametric episodic buffer (a growing list of episode dictionaries). The parametric path provides rapid one-shot encoding/retrieval; the episodic buffer provides exact-match retrieval of specific past experiences. Both paths contribute through the CA1 output projection.

6.6.2. Dentate Gyrus Pattern Separation

The DG performs sparse expansion (see FIG. P5-19 for the sparse code hashing pipeline): $\mathbf{z}_{\text{DG}} = \text{ReLU}(\mathbf{W}_{\text{DG}} \cdot \mathbf{h})$ where $\mathbf{W}_{\text{DG}} \in \mathbb{R}^{d \times d_{\text{sparse}}}$ (default $d_{\text{sparse}} = 2048$). The ReLU activation ensures sparsity: most units are zero, and similar inputs produce dissimilar sparse patterns (pattern separation). For concept indexing, the top-32 active unit indices are extracted and hashed to produce a discrete concept key.

6.6.3. CA3 Encode and Complete

Encoding maps sparse to dense: $\mathbf{e} = \mathbf{W}_{\text{CA3_encode}} \cdot \mathbf{z}_{\text{DG}}$. Retrieval maps dense back to sparse with decay: $\mathbf{z}_{\text{ret}} = \text{sigmoid}(\beta_{\text{decay}}) \cdot \text{ReLU}(\mathbf{W}_{\text{CA3_retrieve}} \cdot \mathbf{e})$. Pattern completion through Schaffer collaterals: $\mathbf{h}_{\text{complete}} = \mathbf{W}_{\text{CA3_complete}} \cdot \mathbf{z}_{\text{ret}}$, where $\mathbf{W}_{\text{CA3_complete}}$ is initialised to zeros (identity at init, separate from the encode projection to keep encode and retrieve paths independent).

6.6.4. Adaptive Write Gate

Episode storage is controlled by a learned write gate (see FIG. P5-20 for the adaptive write gate architecture): $p_{\text{write}} = \text{sigmoid}(\mathbf{W}_{\text{gate}} \cdot [\bar{\mathbf{x}}; \bar{\sigma^2}])$ where $\bar{\mathbf{x}}$ is the mean-pooled representation and $\bar{\sigma^2}$ is the mean precision estimate. The gate bias is initialised to $+1.0$ ($\text{sigmoid}(1) \approx 0.73$, starting open). An episode is stored when surprise $\times\ p_{\text{write}} \times$ confidence $> \theta_{\text{threshold}}$ (default $\theta = 0.1$).

6.6.5. *Bayesian Concept Confidence — Detail*

The confidence update follows a precision-weighted Bayesian rule (see FIG. P5-2 for the complete Bayesian confidence system). For positive evidence e with variance σ^2:

$$\alpha = \frac{0.1}{1+\sigma^2} \quad \text{(precision-weighted learning rate)} \tag{18}$$

$$e_{\text{strength}} = \frac{e}{1+\sigma^2} \quad \text{(precision-weighted evidence)} \tag{19}$$

$$p_{\text{post}} = p_{\text{prior}} \cdot (1-\alpha) + e_{\text{strength}} \cdot \alpha \tag{20}$$

High-precision evidence (low σ^2) shifts beliefs more than uncertain evidence. For contradictory evidence: $p_{\text{post}} = 0.5 \cdot p_{\text{prior}}$, implementing rapid disconfirmation (Anderson, 2003). The posterior is clamped to $[0.01, 0.99]$.

6.6.6. *Goal-Directed Retrieval — Detail*

When a goal context $\mathbf{g}$ is available from the global workspace (representing the current cognitive task or discourse goal), the retrieval query is composed (see FIG. P5-13 for the goal-directed retrieval pipeline):

$$\mathbf{q} = \frac{0.7 \cdot \bar{\mathbf{h}}_{\text{input}} + 0.3 \cdot \bar{\mathbf{g}}}{\|0.7 \cdot \bar{\mathbf{h}}_{\text{input}} + 0.3 \cdot \bar{\mathbf{g}}\| + \epsilon} \tag{21}$$

where $\bar{\mathbf{h}}_{\text{input}}$ and $\bar{\mathbf{g}}$ are the L_2-normalised mean-pooled representations. This 70/30 blend ensures retrieval is primarily input-driven but biased toward goal-relevant memories, modelling prefrontal modulation of hippocampal retrieval (Moscovitch, 2002).

6.6.7. *Elaborative Encoding — Detail*

At encode time, each new episode at buffer index n is linked to its top-3 most similar existing episodes (see FIG. P5-14 for the elaborative linking and retrieval blending):

$$\text{related}(n) = \text{argtop3}_{i<n} \left(\frac{\bar{\mathbf{x}}_n^\top \cdot \bar{\mathbf{x}}_i}{\|\bar{\mathbf{x}}_n\| \cdot \|\bar{\mathbf{x}}_i\|} \right) \tag{22}$$

These indices are stored in the episode metadata (see FIG. P5-22 for the complete episode source metadata structure). At retrieval time, if the best match has non-empty related indices, the related episodes are retrieved and blended:

$$\mathbf{out}_{\text{elab}} = \mathbf{out}_{\text{best}} + 0.3 \cdot \frac{1}{|\text{related}|} \sum_{i \in \text{related}} \text{CA1}(\bar{\mathbf{x}}_i) \tag{23}$$

6.6.8. Error-Driven Forgetting — Detail

For each existing episode i when encoding a new episode n (see FIG. P5-18 for the dual-threshold contradiction detection):

1. Compute cue similarity: cue_sim $= \frac{\bar{\mathbf{e}}_n^\top \cdot \bar{\mathbf{e}}_i}{\|\bar{\mathbf{e}}_n\| \cdot \|\bar{\mathbf{e}}_i\|}$ (error pattern similarity)
2. Compute content similarity: content_sim $= \frac{\bar{\mathbf{x}}_n^\top \cdot \bar{\mathbf{x}}_i}{\|\bar{\mathbf{x}}_n\| \cdot \|\bar{\mathbf{x}}_i\|}$
3. If cue_sim > 0.7 AND content_sim < 0.3: the old episode contradicts the new one. Set surprise$_i = 0$, accelerating eviction.

6.6.9. Multi-Token Chunking — Detail

At each encoding step (see FIG. P5-17 for the chunking decision flow), the mean prediction error (precision-weighted squared error averaged over dimensions and time) is computed. If this falls below the chunk threshold (0.3), the current input is a continuation of the previous concept and is merged via running average:

$$\mathbf{x}_{\text{last}} \leftarrow 0.7 \cdot \mathbf{x}_{\text{last}} + 0.3 \cdot \mathbf{x}_{\text{new}} \tag{24}$$

No new episode is created. This enables the system to store compound concepts (multi-word named entities, collocations, frequent phrases) as single episodic entries.

6.6.10. Schema Absorption — Detail

After the consolidation manager extracts centroid patterns from clustered episodes (semantic memories), these are absorbed as schemas on the hippocampus (see FIG. P5-15 for the merge-or-create schema absorption pipeline):

1. For each centroid, compute cosine similarity against all existing schemas.
2. If sim > 0.8 with an existing schema: merge via weighted average proportional to episode counts, and boost confidence by $+0.05$ (capped at 0.99).
3. Otherwise: create a new schema with initial confidence 0.5.

At retrieval time, schemas are searched alongside episodes. If the best schema match exceeds the best episode match in cosine similarity, the schema is preferred. Schema retrievals receive a $1.5\times$ strength boost (McClelland, 2013: schemas represent generalised knowledge that should dominate over single episodes).

6.6.11. Consolidation Scoring — Detail

When selecting memories for consolidation or replay, each memory i is scored (see FIG. P5-30 for the enhanced micro-consolidation scoring):

$$\text{priority}_i = \text{retrieval_count}_i \times (0.5 + 0.5 \times p_i) \tag{25}$$

where p_i is the concept confidence for memory i. Memories with high retrieval count AND high concept confidence consolidate first—they are the most-accessed, most-reliable knowledge, making them highest priority for cortical integration.

6.6.12. Micro-Consolidation (Sharp Wave Ripples) — Detail

Every ~ 50 processing steps (configurable), the system performs micro-consolidation if at least 3 memories are available (see FIG. P5-16 for the full replay pipeline):

1. Select the top-3 memories by confidence-weighted retrieval count.
2. For each selected memory:
   (a) Reset working memory in all cortical columns (prevent WM contamination).
   (b) Forward-pass the memory through all cortical columns (compressed replay).
   (c) Apply the full learning pipeline at 5% of normal learning rate:
      - Predictive coding: update $\mathbf{W}_{\text{predict}}$ and $\mathbf{W}_{\sigma}$ from PC errors.
      - L5 SwiGLU and dendritic FFN updates from micro-replay errors.
      - Local parameter updates for L4, L1, L6, and L2/3.
      - BCM Hebbian learning on L2/3 associations.
      - Anti-Hebbian decorrelation.
      - BCM threshold tracking for the micro-replay activity.
      - Metaplasticity tracking for weight change magnitude.
      - Subcortical learning (thalamus, workspace, cerebellum, lateral connections).
   (d) Mark the memory as consolidated.

The 5% learning rate prevents catastrophic overwriting of existing cortical knowledge while still enabling gradual knowledge transfer—matching the biological principle that consolidation is slow and interleaved.

6.6.13. Relevance-Based Eviction

When the episodic buffer exceeds twice the configured capacity, memories are scored for eviction (see FIG. P5-21 for the relevance-based eviction scoring):

$$\text{evict_score}_i = \text{surprise}_i \times 2.0 - \text{retrieval_count}_i \times 0.5 - \text{age}_i \times 0.01 \tag{26}$$

Memories with the lowest score (low surprise, high retrieval count, old age) are evicted first—these are memories the cortex has already learned (consolidated, frequently replayed). This is NOT first-in-first-out; it is relevance-based forgetting modelling the brain's consolidation-driven forgetting (Anderson & Hulbert, 2021).

6.7. Hierarchical Basal Ganglia — Detailed Implementation

6.7.1. Biological Basis

The basal ganglia is the brain's action selection organ, implementing Go/NoGo competition through parallel direct (D1) and indirect (D2) corticostriatal pathways modulated by dopamine. Frank and Badre (2012) demonstrated hierarchical organisation: anterior (frontal) loops gate posterior loops, creating a Koechlin (2003) cascade from abstract goals to concrete actions.

6.7.2. Single Loop Architecture

Each of the three loops contains (see FIG. P5-26 for the D1/D2 pathway architecture and FIG. P5-24 for the Boltzmann receptor kinetics):

- **D1 pathway** (direct, stabilising): $\mathbf{W}_{\text{D1_in}} \in \mathbb{R}^{d \times d_{\text{bg}}}$, $\mathbf{W}_{\text{D1_out}} \in \mathbb{R}^{d_{\text{bg}} \times d}$ (output initialised to zeros).
- **D2 pathway** (indirect, flexibility): $\mathbf{W}_{\text{D2_in}} \in \mathbb{R}^{d \times d_{\text{bg}}}$, $\mathbf{W}_{\text{D2_out}} \in \mathbb{R}^{d_{\text{bg}} \times d}$ (output initialised to zeros).
- **Receptor parameters**: EC50_1, s_1 for D1; EC50_2, s_2 for D2 (D2 has lower EC50 = higher DA sensitivity).
- **Gate parameter**: $g_{\text{raw}} = -2.0$ at init (sigmoid ≈ 0.12).

The D1 and D2 activation functions implement Boltzmann receptor binding kinetics:

$$\text{D1}_{\text{act}} = \text{sigmoid}(s_1 \cdot (\text{DA}_{\text{eff}} - \text{EC50}_1)) \tag{27}$$

$$\text{D2}_{\text{act}} = \text{sigmoid}(s_2 \cdot (\text{DA}_{\text{eff}} - \text{EC50}_2)) \tag{28}$$

where $\text{DA}_{\text{eff}} = \text{DA}_{\text{phasic}} - \text{DA}_{\text{tonic}}$ and DA_{tonic} is an exponential moving average of phasic DA (configurable EMA coefficient).

Each pathway processes the input through: RMSNorm $\rightarrow$ $\mathbf{W}_{\text{in}}$ $\rightarrow$ GeLU $\rightarrow$ $\mathbf{W}_{\text{out}}$ $\rightarrow$ $\times$DA_act. The loop output is:

$$\mathbf{y} = \mathbf{x} + g \cdot (\text{D1_mod} - \text{D2_mod}) \cdot \text{top_down_gate} \tag{29}$$

6.7.3. Three-Loop Hierarchy

The three loops are organised by cortical depth band (see FIG. P5-23 for the three-loop hierarchy and cascade gating):

1. **Discourse loop** (anterior, columns 8–11): Processes first with no top-down gate (highest level). Its output is projected through $\mathbf{W}_{\text{discourse_gate}} \in \mathbb{R}^{d \times 1}$ (bias -2.0) followed by sigmoid to produce a gate for the sentence loop.
2. **Sentence loop** (middle, columns 4–7): Gated by the discourse loop output. Its output produces a gate for the word loop via $\mathbf{W}_{\text{sentence_gate}}$.

3. **Word loop** (posterior, columns 0–3): Gated by the sentence loop output. Directly modulates word-level token selection.

This cascade ensures that discourse-level goals constrain sentence plans, which in turn constrain word choices—a hierarchical control structure absent from standard language models.

6.7.4. Tonic Dopamine Tracking

The tonic DA baseline is tracked as an exponential moving average (see FIG. P5-3 for the tonic DA deviation-based system):

$$\text{DA}_{\text{tonic}} \leftarrow (1 - \lambda) \cdot \text{DA}_{\text{tonic}} + \lambda \cdot \overline{\text{DA}_{\text{phasic}}} \tag{30}$$

where $\lambda = 1 -$ ema_coefficient (configurable). This ensures that the D1/D2 competition responds to *deviations* from the expected DA level, not absolute DA magnitude.

6.8. Cerebellar Forward Model — Detailed Implementation

6.8.1. Biological Basis

The cerebellum maintains forward models that predict the sensory consequences of actions (Wolpert et al., 1998). For language, this translates to predicting the next timestep's representation from the current one, enabling proactive error correction rather than purely reactive processing.

6.8.2. Architecture

The cerebellar forward model comprises (see FIG. P5-27 for the complete cerebellar architecture):

- Input projection: $\mathbf{W}_{\text{in}} \in \mathbb{R}^{d \times d_{\text{cb}}}$ (default $d_{\text{cb}} = 256$).
- Prediction: RMSNorm $\rightarrow \mathbf{W}_{\text{in}} \rightarrow$ GeLU $\rightarrow \mathbf{W}_{\text{predict}} \in \mathbb{R}^{d_{\text{cb}} \times d}$.
- Error correction: $\mathbf{W}_{\text{error}} \in \mathbb{R}^{d \times d}$ (initialised to zeros).
- Gate: $g_{\text{raw}} = -2.0$ (sigmoid ≈ 0.12).

6.8.3. Forward Pass

For sequence length $T > 1$:

1. Compute predictions: $\hat{\mathbf{x}} = \mathbf{W}_{\text{predict}}(\text{GeLU}(\mathbf{W}_{\text{in}}(\text{RMSNorm}(\mathbf{x}))))$.
2. Compute temporal error: $\mathbf{e}[t] = \mathbf{x}[t+1] - \hat{\mathbf{x}}[t]$ for $t = 0, \ldots, T-2$.
3. Compute cerebellar loss: $\mathcal{L}_{\text{cb}} = \frac{1}{(T-1)d} \sum_{t=0}^{T-2} \|\mathbf{e}[t]\|^2$.
4. Apply causal correction: Pad error with zeros on the right (last position gets no correction, preserving causality): $\mathbf{e}_{\text{padded}} = [\mathbf{e}[0], \ldots, \mathbf{e}[T-2], \mathbf{0}]$.
5. Correct: $\mathbf{x}_{\text{corrected}} = \mathbf{x} + g \cdot \mathbf{W}_{\text{error}} \cdot \mathbf{e}_{\text{padded}}$.

The causal padding is critical: each position's correction is based only on the prediction error from the *previous* position's prediction about the current position, maintaining temporal causality.

6.9. Identity-at-Initialisation Contract

All subsystems in the present invention satisfy:

1. All output projections ($\mathbf{W}_{\text{read}}$, $\mathbf{W}_{\text{retrieve}}$, $\mathbf{W}_{\text{phono_read}}$, CA1, CA3_complete, D1_out, D2_out, $\mathbf{W}_{\text{error}}$) are initialised with zero weights.
2. All gate parameters (g_{raw}) are initialised to -2.0 (sigmoid ≈ 0.12).
3. Working memory state is initialised to zeros (empty slots, zero facilitation, empty ring buffer).
4. A freshly constructed model produces outputs that are within 0.12 scaling factor of the identity transformation.

This contract enables any subset of subsystems to be attached to a pre-trained model without disrupting existing performance, with each subsystem gradually learning to contribute as training progresses. FIG. P5-29 provides a complete method overview showing all per-step operations (a)–(i) and their temporal scheduling.

6.10. State Persistence Across Forward Calls

Working memory state (NMDA slots, silent trace u and x, phonological ring buffer and context) persists across sequential forward calls (see FIG. P5-10 for the volatile state persistence model), modelling the biological reality that working memory maintains information across processing steps within a single cognitive episode. State resets when the batch size changes (signalling a new context) or when explicitly reset. Working memory state is NOT saved to disk checkpoints, matching the volatile nature of real working memory—it is reconstructed from input at the start of each new context.

Hippocampal episodic buffer, concept confidence, and schema buffer persist across training steps and are volatile (not saved to checkpoints). Tonic DA in the basal ganglia is a registered buffer and IS saved to checkpoints.

7. CLAIMS

We claim:

Claim 1. A multi-timescale working memory system for a neural language model, the system comprising a processor and memory storing instructions that, when executed, cause the processor to perform operations comprising:

(a) maintaining a first memory subsystem comprising K attractor slots, each slot being a vector of dimension d_{slot}, wherein at each time step t, for an input representation $\mathbf{h}_t$:

(i) computing competitive write probabilities $\mathbf{w}[t] = \text{softmax}(\mathbf{W}_{\text{score}} \cdot \mathbf{h}_t) \in \mathbb{R}^K$ over said K slots;

(ii) computing a content vector $\mathbf{c}[t] = \mathbf{W}_{\text{write}} \cdot \mathbf{h}_t \in \mathbb{R}^{d_{\text{slot}}}$;

(iii) updating each slot k via exponential decay with a learnable per-slot decay rate: $\mathbf{s}_k[t] = \text{sigmoid}(\beta_k) \cdot \mathbf{s}_k[t-1] + w_k[t] \cdot \mathbf{c}[t]$;

(iv) reading from slots via content-based attention using $\mathbf{c}[t]$ as query over slot contents with softmax-normalised dot-product scores;

(v) computing a first confidence signal as the read-attention-weighted sum of slot L_2 norms;

(b) maintaining a second memory subsystem implementing Tsodyks-Markram facilitation-depression dynamics in a synaptic trace space of dimension d_{trace}, wherein:

(i) a facilitation variable $u[t]$ is updated as: $u[t] = a_f \cdot u[t-1] + (1 - u[t-1]) \cdot g[t] \cdot |\mathbf{c}[t]|$, where $a_f = \text{sigmoid}(\alpha_f)$ is a learnable decay, $g[t]$ is a learned write gate, and $\mathbf{c}[t]$ is a stored content projection;

(ii) a depression variable $x[t]$ is updated as: $x[t] = a_d \cdot x[t-1] + (1 - a_d) - d_f \cdot u[t] \cdot x[t-1]$, where $a_d = \text{sigmoid}(\alpha_d)$ and $d_f = \text{sigmoid}(\alpha_{df}) \cdot 0.5$;

(iii) effective synaptic release is computed as $j_{\text{eff}}[t] = u[t] \cdot x[t]$;

(iv) a second confidence signal is computed as the mean of facilitation u across trace dimensions;

(c) maintaining a third memory subsystem comprising a ring buffer of K_p slots with temporal context tagging, wherein:

(i) a temporal context vector **ctx** evolves via a learned near-orthogonal linear transformation followed by L_2 normalisation at each time step;

(ii) written content is tagged with said temporal context via additive projection before storage in the ring buffer;

(iii) reading combines both content similarity and temporal context similarity as retrieval scores;

(iv) a third confidence signal is computed as the maximum temporal context similarity across slots;

(d) computing a metamemory confidence scalar by applying sigmoid normalisation to each of said first, second, and third confidence signals and averaging the normalised values.

Claim 2. A hippocampal episodic memory system for a neural language model, the system comprising a processor and memory storing instructions that, when executed, cause the processor to perform operations comprising:

(a) performing pattern separation by projecting input representations through a sparse expansion projection followed by rectified linear activation to produce high-dimensional sparse codes;

(b) performing autoassociative pattern completion through an encode projection from sparse to dense space, a retrieval projection from dense to sparse space with learnable decay gating, and a separate completion projection initialised to zero weights;

(c) maintaining an episodic buffer storing episode entries, each entry comprising at least a representation tensor and associated metadata;

(d) at encoding time, for each new episode:

  (i) computing a precision-weighted prediction error surprise score;

  (ii) when said surprise score falls below a chunk threshold, merging the new representation into the most recent episode via weighted running average rather than creating a new entry;

  (iii) when said surprise score exceeds said chunk threshold and an adaptive write gate permits storage, creating a new episode entry and linking it to its top-k most similar existing episodes by cosine similarity;

  (iv) detecting contradicting old episodes having cue similarity exceeding a first threshold and content similarity below a second threshold, and zeroing the surprise score of said contradicting episodes to accelerate their eviction;

(e) maintaining a per-concept confidence value identified by a hash of the sparse code activation pattern, wherein:

  (i) new concepts begin at a tentative confidence;

  (ii) confirmation evidence updates confidence via a precision-weighted rule proportional to $1/(1+\sigma^2)$;

  (iii) contradictory evidence halves the prior confidence;

(f) at retrieval time, searching said episodic buffer and a schema buffer via cosine similarity, scaling retrieval strength by concept confidence.

Claim 3. A hierarchical action selection system for a neural language model, the system comprising a processor and memory storing instructions that, when executed, cause the processor to perform operations comprising:

(a) maintaining a plurality of corticostriatal loops, each loop corresponding to a distinct temporal granularity of language processing, wherein each loop comprises:

(i) a first pathway (D1, direct) computing a stabilising modulation as a function of a dopamine signal through a projection, nonlinear activation, output projection, and multiplicative Boltzmann receptor activation sigmoid$(s_1 \cdot (\text{DA}_{\text{eff}} - \text{EC50}_1))$;

(ii) a second pathway (D2, indirect) computing a flexibility modulation through the same architecture with separate learnable receptor parameters including a lower EC50 value for higher dopamine sensitivity;

(iii) computing loop output as the difference between D1 and D2 modulations;

(b) organising said plurality of loops in a cascade hierarchy, wherein:

(i) a highest-level loop processes without top-down gating;

(ii) each higher-level loop produces a gating signal through a learned projection with sigmoid activation;

(iii) each lower-level loop's output is multiplicatively gated by the next higher-level loop's gating signal;

(c) computing effective dopamine as a phasic deviation from a tonic baseline, wherein said tonic baseline is an exponential moving average of observed phasic dopamine signals.

Claim 4. The system of Claim 1, wherein the slot update recurrence is computed via a parallel prefix scan over the sequence dimension, comprising:

(a) batch pre-computing all write probabilities and content vectors for all T positions via batched matrix multiplications;

(b) reshaping the K per-slot scan inputs into a single vector of dimension $K \times d_{\text{slot}}$ and executing a single parallel scan with constant per-slot decay coefficients;

(c) reshaping the scan output and computing all reads via batched attention over the $B \times T$ dimension.

Claim 5. The system of Claim 1, wherein the facilitation update of the second memory subsystem is linearised as $u[t] = u[t-1] \cdot (a_f - g[t] \cdot |\mathbf{c}[t]|) + g[t] \cdot |\mathbf{c}[t]|$ and computed via a first parallel prefix scan, and the depression update is computed via a second parallel prefix scan with coefficients dependent on the facilitation values from said first scan, yielding two sequential parallel scans.

Claim 6. The system of Claim 1, wherein the third memory subsystem further comprises an articulatory rehearsal mechanism wherein the most recently written slot is boosted by a learnable rehearsal strength factor $r = \text{sigmoid}(r_{\text{raw}})$ at each time step prior to writing the current input.

Claim 7. The system of Claim 1, wherein the temporal context vector is initialised as a random unit-norm vector and evolved via a learned linear transformation whose weight matrix is initialised

as a scaled orthogonal matrix with scaling factor less than 1.0, producing slow rotation in context space without divergence.

Claim 8. The system of Claim 1, wherein the read scores of the third memory subsystem are computed as the sum of a content similarity score and a temporal context similarity score, said content similarity being the dot product between a query projection of the input and slot contents, and said temporal context similarity being the dot product between the current context vector and the stored per-slot context vectors.

Claim 9. The system of Claim 1, wherein the metamemory confidence scalar is provided to a drift-diffusion decision mechanism to modulate deliberation depth, such that low metamemory confidence causes longer evidence accumulation before token commitment during language generation.

Claim 10. The system of Claim 1, wherein working memory state persists across sequential forward calls within a processing context and resets when the batch size changes, said state not being saved to disk checkpoints, modelling the volatile nature of biological working memory.

Claim 11. The system of Claim 1, wherein the first confidence signal is further transformed by $\text{sigmoid}(2 \cdot c_{\text{NMDA}} - 1)$ to produce a scaling factor applied multiplicatively to the read output, and the second confidence signal is further transformed by $\text{sigmoid}(3 \cdot c_{\text{silent}} - 1)$ to produce a scaling factor applied multiplicatively to the synaptic release output, and the third confidence signal is transformed by $\text{sigmoid}(c_{\text{phono}})$ to produce a scaling factor applied multiplicatively to the phonological read output.

Claim 12. The system of Claim 1, wherein all output projections of the first, second, and third memory subsystems are initialised with zero weights, and all gate parameters are initialised to a value producing a sigmoid output of approximately 0.12, ensuring near-zero contribution at initialisation.

Claim 13. The system of Claim 2, wherein the retrieval query is biased by a goal context signal from a global workspace, the biased query being computed as $\mathbf{q} = 0.7 \cdot \mathbf{q}_{\text{input}} + 0.3 \cdot \mathbf{q}_{\text{goal}}$ followed by L_2 normalisation, modelling prefrontal modulation of hippocampal retrieval.

Claim 14. The system of Claim 2, wherein linking at encoding time comprises computing cosine similarity between the new episode and all existing episodes and storing the indices of the top-3 most similar episodes as related indices in the episode metadata, and at retrieval time, when the best-matching episode has non-empty related indices, retrieving the related episodes and blending them into the output with a weight of 0.3.

Claim 15. The system of Claim 2, further comprising a schema buffer storing abstracted knowledge centroids, wherein:

(a) after consolidation extracts centroid patterns from clustered episodes, said centroids are absorbed as schemas;

(b) if a new centroid has cosine similarity exceeding 0.8 with an existing schema, the centroids are merged via weighted average proportional to episode counts and schema confidence is boosted by a fixed increment;

(c) at retrieval time, schemas are searched alongside episodes, and schema matches receive a multiplicative strength boost over episode matches.

Claim 16. The system of Claim 2, further comprising micro-consolidation at periodic intervals during wakefulness, wherein:

(a) the top-n most-retrieved memories, scored by retrieval count weighted by concept confidence, are selected;

(b) working memory in all cortical columns is reset before each replay;

(c) each selected memory is forward-passed through all cortical columns and subjected to a full learning pipeline at a fraction of the normal learning rate;

(d) learning comprises predictive coding updates, Hebbian updates with BCM threshold tracking, anti-Hebbian decorrelation, metaplasticity tracking, and subcortical learning.

Claim 17. The system of Claim 2, wherein the multi-token chunking merges the new representation into the most recent episode via weighted running average $\mathbf{x}_{\text{last}} \leftarrow \alpha \cdot \mathbf{x}_{\text{last}} + (1 - \alpha) \cdot \mathbf{x}_{\text{new}}$ when the precision-weighted mean prediction error falls below a learnable threshold, enabling compound concept encoding.

Claim 18. The system of Claim 2, wherein error-driven forgetting detects contradicting old memories by requiring both that the cue similarity between prediction error patterns exceeds a first threshold and that the content similarity between stored representations falls below a second threshold, and upon detection, zeroing the surprise score of the old memory to accelerate its eviction through relevance-based forgetting.

Claim 19. The system of Claim 2, wherein the sparse code hash is computed by extracting the indices of the top-k most active units from the rectified sparse code and hashing said indices to produce a discrete concept key, implementing pattern separation through sparse distributed coding.

Claim 20. The system of Claim 2, wherein the adaptive write gate receives as input the concatenation of the mean-pooled representation and the mean precision estimate, with the gate bias initialised to a positive value yielding an initial write probability exceeding 0.5.

Claim 21. The system of Claim 2, wherein relevance-based eviction scores each memory using a combination of surprise magnitude, retrieval count, and age, and evicts memories with the lowest score when the buffer exceeds a capacity threshold, thereby prioritising eviction of consolidated, frequently-retrieved, low-surprise memories.

Claim 22. The system of Claim 2, wherein each episode entry further stores source metadata comprising at least a step counter, a boolean flag indicating whether the encoded content was generated by the model or received as input, and the precision estimate at encoding time, enabling source monitoring at retrieval time.

Claim 23. The system of Claim 3, wherein the plurality of corticostriatal loops comprises three loops corresponding to word-level selection, sentence-level planning, and discourse-level goal maintenance, organised such that the discourse loop gates the sentence loop and the sentence loop gates the word loop.

Claim 24. The system of Claim 3, wherein the D1 and D2 receptor parameters comprise learnable EC50 and slope values implementing Boltzmann receptor binding kinetics, with the D2 pathway having a lower EC50 than the D1 pathway, providing higher sensitivity to dopamine fluctuations.

Claim 25. The system of Claim 3, wherein the gating signal from each higher-level loop to the next lower-level loop is produced by a learned linear projection with sigmoid activation, said projection having weights initialised to zeros and bias initialised to a negative value producing a gate output of approximately 0.12 at initialisation.

Claim 26. The system of Claim 3, wherein each pathway within each loop processes the input through RMS normalisation, a first linear projection to a hidden dimension d_{bg}, a GeLU nonlinear activation, and a second linear projection back to the model dimension, with the second linear projection initialised to zero weights.

Claim 27. The system of Claim 3, further in combination with a cerebellar forward model comprising:

(a) a prediction pathway projecting the current representation through a hidden layer with GeLU activation to predict the next timestep's representation;

(b) computing temporal prediction error as the difference between predicted and actual next-timestep representations for positions 0 to $T - 2$;

(c) computing a cerebellar loss as the mean squared prediction error for monitoring temporal prediction quality;

(d) applying causal error correction by right-padding the error signal with zeros and projecting through a learned error projection initialised to zero weights;

(e) the error correction being gated by a learnable gate parameter initialised to produce approximately 0.12 output.

Claim 28. The system of Claim 1, further in combination with the hippocampal episodic memory system of Claim 2 and the hierarchical action selection system of Claim 3, wherein

the metamemory confidence scalar from the working memory system modulates downstream processing, the episodic memory system provides long-term storage and retrieval complementary to the volatile working memory, and the basal ganglia system selects actions informed by dopamine signals reflecting prediction error from the cortical column network.

Claim 29. A computer-implemented method for maintaining multi-timescale memory in a neural language model, the method comprising:

(a) at each time step, updating K attractor slots via competitive softmax write probabilities and exponential decay, and reading from said slots via content-based attention;

(b) at each time step, updating facilitation and depression variables according to Tsodyks-Markram dynamics and computing effective synaptic release as their product;

(c) at each time step, writing to a phonological ring buffer with temporal context tagging via a near-orthogonal rotation, and reading via combined content and context similarity scores;

(d) computing per-subsystem confidence signals and aggregating them into a metamemory scalar via sigmoid normalisation and averaging;

(e) encoding surprising events into a hippocampal episodic buffer with multi-token chunking, elaborative linking to similar episodes, and error-driven forgetting of contradicting entries;

(f) maintaining per-concept Bayesian confidence values updated via precision-weighted evidence;

(g) periodically performing micro-consolidation by replaying the most-retrieved episodes through cortical columns at a reduced learning rate;

(h) selecting actions through hierarchical D1/D2 corticostriatal loops with Boltzmann receptor kinetics and top-down cascade gating;

(i) predicting next-timestep representations via a cerebellar forward model and applying causal error correction.

Claim 30. The method of Claim 29, wherein the micro-consolidation step (g) further comprises scoring each memory by the product of its retrieval count and concept confidence value $(0.5 + 0.5 \times p_i)$, selecting the highest-scoring memories, and applying both BCM threshold tracking and metaplasticity weight-change tracking during the reduced-learning-rate replay to prevent plasticity drift during consolidation.

ABSTRACT

A neural language model augmented with biologically-grounded memory and decision systems is disclosed. The system comprises three working memory subsystems operating at distinct timescales: (i) NMDA attractor slots providing persistent storage via competitive softmax slot allocation and exponential decay with learnable half-life, accelerated by parallel scan computation; (ii) an activity-silent synaptic trace implementing Tsodyks-Markram facilitation-depression dynamics via two sequential parallel scans, maintaining information in synaptic weights without sustained neural activity; and (iii) a phonological loop employing a ring buffer with temporal context tagging through slow-rotating orthogonal vectors for serial-order maintenance and articulatory rehearsal. A metamemory confidence aggregator combines per-subsystem confidence signals to drive downstream deliberation depth. The system further comprises a hippocampal episodic memory with dentate gyrus pattern separation, CA3 autoassociative pattern completion, Bayesian concept confidence tracking, elaborative encoding linking episodes to their three nearest neighbours, error-driven forgetting of contradicting memories, multi-token chunking for compound concepts, schema absorption from consolidation centroids, and sharp-wave-ripple micro-consolidation transferring frequently-retrieved episodes into cortical weights during quiet wakefulness. A hierarchical basal ganglia implements three corticostriatal D1/D2 loops organised by depth band for word-level, sentence-level, and discourse-level action selection with Boltzmann receptor kinetics and top-down cascade gating. A cerebellar forward model predicts next-timestep representations and applies causal error correction. All subsystems are identity-at-init(Stale from transformer era check with /nct-research)ialisation and gated to ensure zero contribution before learning.