A Cortical Column Network for Natural Language Processing

Open research specification 1 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 FIELD **3**

2 BACKGROUND **3**

2.1 Prior Art: Transformer Neural Networks 3

2.2 Limitations of the Prior Art 3

2.2.1 Quadratic Computational Complexity 4

2.2.2 Absence of Biological Grounding 4

2.2.3 Dependence on Global Backpropagation 4

2.2.4 No Intrinsic Working Memory 4

2.2.5 No Uncertainty Quantification 4

2.2.6 No Temporal Organisation 5

2.2.7 No Self-Monitoring 5

2.3 Need for Improvement 5

3 SUMMARY **5**

4 DESCRIPTION OF THE FIGURES **7**

5 DETAILED DESCRIPTION **28**

5.1 System Overview (FIG. 1, FIG. P1-1) 28

5.2 Cortical Column Processing Unit (FIG. 2, FIG. P1-2) 29

5.2.1 Layer 4: Thalamic Input Sublayer 29

5.2.2 Layer 1: Top-Down Modulation Sublayer 29

5.2.3 Interneuron Gating Circuits — Pre-Association Phase (FIG. P1-9) . . . 30

5.2.4 Layer 2/3: Windowed Hebbian Association (FIG. 3, FIG. P1-5 through P1-8) 31

5.2.5 Working Memory Subsystem 33

5.2.6 Oscillatory Dynamics Module (FIG. P1-15, P1-16) 34

5.2.7 Interneuron Gating Circuits — Post-Computation Phase (FIG. P1-10) . 35

5.2.8 Layer 5: Output and Prediction Sublayer (FIG. P1-21, P1-22) 35

5.2.9 Layer 6: Thalamic Feedback Sublayer 37

5.3 Thalamic Routing System (FIG. 4, FIG. P1-11, P1-12) 37

5.3.1 Thalamic Relay 37

5.3.2 Thalamic Reticular Nucleus (TRN) 38

5.3.3 Bimodality Loss 38

5.3.4 Routed Output 38

5.4 Lateral Column Connections (FIG. P1-13) 38

5.5 Global Workspace Broadcast (FIG. P1-14) 39

5.5.1 Write Projections . . . 39
5.5.2 Surprise-Weighted Competition . . . 39
5.5.3 Broadcast and Read . . . 39
5.6 BCM Homeostasis (FIG. P1-17) . . . 39
5.7 Metaplasticity (FIG. P1-18) . . . 40
5.8 Structural Plasticity (FIG. P1-19) . . . 40
5.8.1 Connection Mask . . . 40
5.8.2 Pruning . . . 40
5.8.3 Synaptogenesis . . . 41
5.9 Metacognitive Self-Assessment (FIG. P1-20) . . . 41
5.9.1 Conflict Detection . . . 41
5.9.2 Adaptive Computation . . . 41
5.9.3 Health Monitoring . . . 41
5.10 Identity-at-Initialisation Design Principle (FIG. P1-4) . . . 41
5.11 Forward Pass Through the Complete System (FIG. P1-2, P1-23) . . . 42
5.12 Disease-Mapping Validation Methodology (FIG. P1-27) . . . 43
5.13 Implementation Considerations . . . 44
5.13.1 Mixed Precision Training . . . 44
5.13.2 Custom CUDA Kernels . . . 45
5.13.3 Torch Compilation . . . 45
5.13.4 Scalability (FIG. P1-28) . . . 45

TITLE

A Cortical Column Network for Natural Language Processing

1. FIELD

The present invention relates generally to neural network architectures for natural language processing, and more particularly to a biologically-grounded cortical column network system that replaces transformer-based self-attention mechanisms with a hierarchical stack of cortical column processing units, each implementing a six-layer pipeline modelled on the laminar structure of the mammalian neocortex, incorporating local Hebbian learning, predictive coding, interneuron gating circuits, multi-timescale working memory, oscillatory dynamics, thalamic routing, lateral column connections, and global workspace broadcast for processing sequential language data.

2. BACKGROUND

2.1. Prior Art: Transformer Neural Networks

The dominant architecture for natural language processing (NLP) since 2017 has been the transformer, introduced by Vaswani et al. in "Attention Is All You Need" (Advances in Neural Information Processing Systems, 2017). The transformer architecture processes input sequences through a stack of identical blocks, each comprising a multi-head self-attention mechanism and a position-wise feedforward network. Self-attention computes, for each token in a sequence, a weighted combination of all other tokens using learned query (Q), key (K), and value (V) projections, with attention weights derived from a scaled dot-product followed by a softmax normalisation:

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^\top}{\sqrt{d_k}}\right)V \tag{1}$$

This mechanism has proven effective for capturing long-range dependencies in language, leading to large-scale models such as GPT, BERT, PaLM, LLaMA, and their successors.

2.2. Limitations of the Prior Art

Despite their commercial success, transformer architectures suffer from several fundamental limitations that the present invention addresses:

2.2.1. Quadratic Computational Complexity

The self-attention mechanism in transformers has $O(T^2)$ time and memory complexity with respect to the sequence length T, because every token must attend to every other token. This quadratic scaling limits practical deployment to context windows of thousands to hundreds of thousands of tokens, and imposes significant energy and hardware costs for long-document processing. Various linear attention approximations (e.g., Performer, Linformer, Longformer) have been proposed, but these typically sacrifice representational fidelity for efficiency and do not address the other limitations described below.

2.2.2. Absence of Biological Grounding

Transformers bear no structural correspondence to the neural circuits of the mammalian neocortex. The brain processes language through hierarchical cortical columns with six distinct laminar layers, local recurrent circuits, and inhibitory interneuron gating — none of which appear in transformer architectures. This lack of biological grounding means that transformers cannot benefit from the extensive neuroscience literature on cortical computation, and cannot exhibit brain-like properties such as graceful degradation, local learning, or self-monitoring.

2.2.3. Dependence on Global Backpropagation

Transformer training relies exclusively on global backpropagation of error gradients through the entire network. This requires storing all intermediate activations during the forward pass (consuming substantial memory), propagating gradients through potentially hundreds of layers (creating vanishing and exploding gradient problems), and performing a global synchronisation step after each batch. The brain, by contrast, learns primarily through local synaptic modification rules (Hebbian learning, predictive coding) that require no global error signal.

2.2.4. No Intrinsic Working Memory

Transformers have no persistent state between inference steps. All context is captured within the fixed-length attention window, and extending context requires either recomputing attention over the full history (at quadratic cost) or discarding old context. The brain maintains information across multiple timescales through distinct working memory subsystems (NMDA-receptor-mediated persistent activity, activity-silent synaptic traces, and articulatory rehearsal loops), enabling effectively unlimited temporal context.

2.2.5. No Uncertainty Quantification

Standard transformer outputs provide logit scores but no built-in measure of prediction uncertainty. Confidence estimation requires external calibration techniques or ensemble methods. The brain natively computes prediction precision (σ^2) at every cortical level through precision-weighted

predictive coding (PWPC), enabling context-dependent resource allocation: uncertain predictions trigger additional processing ("thinking harder"), while confident predictions proceed rapidly.

2.2.6. No Temporal Organisation

Transformers rely on position embeddings (sinusoidal or learned) to encode sequential order, but have no mechanism for the nested temporal organisation observed in cortical oscillations. The brain organises information processing through phase-amplitude coupling between oscillatory bands (gamma, theta, alpha, delta), binding items in working memory to distinct temporal phases and coordinating computation across cortical regions at different timescales.

2.2.7. No Self-Monitoring

Transformers have no metacognitive capacity to monitor their own processing quality. The brain employs anterior cingulate cortex (ACC) conflict monitoring, which detects inter-column disagreement and allocates additional computational resources (inference iterations) to difficult inputs. This adaptive computation is absent from uniform-depth transformer processing.

2.3. Need for Improvement

There exists a need in the art for a neural network architecture for natural language processing that: (a) achieves linear rather than quadratic computational complexity in sequence length; (b) is grounded in the laminar structure of the neocortex; (c) supports local learning without global backpropagation; (d) incorporates persistent multi-timescale working memory; (e) provides built-in uncertainty quantification; (f) implements oscillatory temporal organisation; and (g) includes metacognitive self-monitoring for adaptive computation. The present invention addresses all of these needs.

3. SUMMARY

The present invention provides a computer-implemented neural network system and method for natural language processing that replaces the transformer self-attention architecture with a biologically-grounded cortical column network (CCN).

In a first aspect, the invention provides a computer-implemented neural network system for natural language processing, comprising: a plurality of N cortical column processing units arranged in a hierarchical stack from a lowest column (column 0) to a highest column (column $N-1$); a thalamic routing system comprising a plurality of relay groups and a thalamic reticular nucleus, configured to gate information flow between cortical columns; a set of lateral column connections providing horizontal communication between adjacent columns; and a global workspace module implementing surprise-weighted broadcast competition across all columns.

Each cortical column processing unit comprises six functional layers arranged in a biologically-grounded pipeline:

(i) a Layer 4 thalamic input sublayer receiving feedforward input from the thalamic routing system;

(ii) a Layer 1 top-down modulation sublayer receiving predictions from a higher column;

(iii) an interneuron gating circuit comprising VIP disinhibitory neurons, SST dendritic inhibitory neurons applied before lateral association, and PV/Chandelier output veto neurons applied after oscillatory dynamics;

(iv) a Layer 2/3 windowed Hebbian association sublayer performing content-addressable recall within a local causal window of size w tokens with $O(T \times w)$ time complexity, using learned association weights, anti-Hebbian lateral decorrelation weights, distance-weighted context aggregation, and k-winner-take-all competitive coding;

(v) a working memory subsystem operating at three distinct timescales: NMDA attractor slots with persistent activity, activity-silent synaptic traces with facilitation and depression dynamics, and a phonological loop with decay, rehearsal, and temporal context tagging;

(vi) an oscillatory dynamics module implementing nested theta-gamma phase-amplitude coupling with learnable oscillatory periods;

(vii) a Layer 5 output and prediction sublayer computing a feedforward output using SwiGLU activation with dendritic gating (basal and apical), k-winner-take-all sparsity, a top-down prediction for the column below, and a per-token prediction precision estimate (σ^2) via softplus activation; and

(viii) a Layer 6 thalamic feedback sublayer generating a corticothalamic feedback signal for modulating the thalamic relay.

The system further comprises: a BCM homeostatic plasticity mechanism per column, maintaining a sliding threshold that prevents runaway Hebbian potentiation; a metaplasticity mechanism per column that adjusts learning rates based on recent plasticity history; a structural plasticity mechanism that prunes weak connections and grows new connections based on prediction error magnitude; and a metacognitive self-assessment module that computes inter-column conflict and dynamically allocates additional predictive coding inference iterations to difficult inputs.

In a second aspect, the invention provides a computer-implemented method for processing natural language input through the cortical column network system, comprising: embedding input tokens into a continuous representation; routing the embedded representation through the thalamic system to a first cortical column; processing the representation through the six-layer pipeline of each cortical column in hierarchical sequence; computing top-down prediction errors

between each column's prediction and the actual input of the column below; iteratively settling column representations through predictive coding inference iterations under adaptive convergence control; computing a global workspace broadcast based on surprise-weighted competition among columns; and producing output logits through a language model head.

In a third aspect, the invention provides a non-transitory computer-readable storage medium storing instructions that, when executed by one or more processors, cause the processors to implement the cortical column network system and method described above.

The architecture provides the following advantages over the prior art:

- Linear computational complexity $O(T \times w)$ in sequence length, replacing the $O(T^2)$ complexity of self-attention;
- Biological grounding in the six-layer laminar structure of the neocortex, enabling systematic architecture design guided by neuroscience;
- Support for local learning without global backpropagation, through predictive coding and Hebbian learning rules;
- Persistent multi-timescale working memory providing effectively unlimited temporal context;
- Built-in per-token uncertainty quantification through precision-weighted predictive coding;
- Oscillatory temporal organisation through nested phase-amplitude coupling;
- Metacognitive self-monitoring with adaptive computation allocation;
- identity-at-init(Stale from transformer era check with /nct-research)ialisation design ensuring stable training from random initialisation; and
- A principled disease-mapping validation methodology wherein removal of any architectural component produces a specific, predictable failure mode corresponding to a known neurological or psychiatric condition.

4. 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. 1 is a high-level architectural overview of the cortical column network system, showing the embedding layer, thalamic routing, hierarchical column stack, lateral connections, global workspace, subcortical systems, and language model head.

FIG. 2 is a detailed view of a single cortical column processing unit, illustrating the nine sublayers (L4, L1, SST/VIP, L2/3, WM, Oscillations, PV/Chandelier, L5, L6) and their internal data flow.

FIG. 3 is a diagram of the windowed Hebbian association mechanism in Layer 2/3, contrasting it with standard self-attention and showing the $O(T \times w)$ local causal window computation.

FIG. 4 is a diagram of the thalamic routing system, showing relay groups, TRN relevance gating, and L6 corticothalamic feedback.

FIG. P1-1 is a block diagram showing the overall architecture of the cortical column network system, including the embedding layer, thalamic routing system, hierarchical stack of N cortical columns, lateral connections, global workspace, and the language model head (Claim 1).

FIG. P1-2 is a flowchart showing the method for processing natural language input, illustrating the nine-step pipeline from embedding through output logit generation with predictive coding settling loops (Claim 2).

FIG. P1-3 is a diagram illustrating the non-transitory computer-readable storage medium claim, showing the six-layer pipeline specification stored on the medium (Claim 3).

FIG. P1-4 is a diagram illustrating the identity-at-init(Stale from transformer era check with /nct-research)ialisation principle, showing the gate parameter initialisation at $g = -2.0$ yielding sigmoid$(-2.0) \approx 0.12$, zero output weight matrices, and the residual bypass ensuring near-identity at init (Claim 4).

FIG. P1-5 is a diagram illustrating the Layer 2/3 windowed Hebbian association sublayer, showing the direct association path via $\mathbf{W}_{\text{assoc}}$ and the Hebbian update rule $\Delta W = \eta \cdot (\mathbf{X}^\top \mathbf{X})/N$ (Claim 5).

FIG. P1-6 is a diagram illustrating the windowed context aggregation mechanism, showing the distance-weighted combination within the causal window using softplus$((\mathbf{x}_t - \mathbf{x}_j) \cdot \mathbf{W}_{\text{dist}})$ and its $O(T \times w)$ complexity (Claim 6).

FIG. P1-7 is a diagram illustrating the anti-Hebbian lateral decorrelation matrix, showing the update rule $\Delta\mathbf{W}_{\text{lateral}} = \eta \cdot ((\mathbf{Y}^\top \mathbf{Y})/N - \mathbf{I})$ that drives head outputs toward orthogonality (Claim 7).

FIG. P1-8 is a diagram illustrating the k-winner-take-all competitive coding mechanism, showing $k = \lfloor d_h \times \rho \rfloor$ with sparsity ratio $\rho = 0.15$ and the resulting sparse population code (Claim 8).

FIG. P1-9 is a circuit diagram illustrating the VIP disinhibition and SST dendritic inhibition pathway, showing $\mathbf{s}_{\text{eff}} = \mathbf{s}_{\text{SST}} \cdot (1 - \mathbf{v}_{\text{VIP}})$ clamped to a maximum of 0.9 (Claim 9).

FIG. P1-10 is a diagram illustrating the PV/Chandelier output veto gate, showing the pass probability p_{pass} with bias initialised to 5.0, the hard binary Bernoulli gate in the forward pass, and the straight-through gradient estimator in the backward pass (Claim 10).

FIG. P1-11 is a diagram illustrating the thalamic reticular nucleus (TRN) relevance gate, showing $r = \text{sigmoid}((\mathbf{W}_q \cdot \mathbf{x}) \odot (\mathbf{W}_k \cdot \delta_{\text{relay}})/\sqrt{d_{\text{TRN}}})$ and the bimodality loss mean$(r \cdot (1-r))$ (Claim 11).

FIG. P1-12 is a diagram illustrating the Layer 6 thalamic feedback sublayer, showing the cortex-thalamus feedback loop where relay output is modulated as $\delta_{\text{relay}} + 0.1 \cdot \mathbf{f}_{\text{L6}}$ (Claim 12).

FIG. P1-13 is a diagram illustrating the lateral column connections, showing the low-rank send

projection $\mathbf{W}_{\text{send}} \in \mathbb{R}^{r \times d_{\text{model}}}$ and receive projection $\mathbf{W}_{\text{recv}} \in \mathbb{R}^{d_{\text{model}} \times r}$ with $r = \max(d_{\text{model}}/8, 32)$ between adjacent columns (Claim 13).

FIG. P1-14 is a block diagram illustrating the global workspace broadcast mechanism, showing the surprise computation via mean squared prediction error, the softmax competition with learnable temperature, and the per-column write and read projections (Claim 14).

FIG. P1-15 is a diagram illustrating the oscillatory gamma amplitude modulation equation $A_\gamma(t) = \max(0, 1 + \kappa_1 \cos(\phi_\theta) + \kappa_2 \cos(\phi_\delta) - \kappa_3(1 + \cos(\phi_\alpha))/2)$ with learnable periods via softplus (Claim 15).

FIG. P1-16 is a diagram illustrating the phase-content bilinear coupling and oscillatory modulation, showing $\mathbf{m}_{\text{osc}} = \mathbf{W}_{\text{out}} \cdot \tanh(\mathbf{W}_{\text{content}} \cdot \mathbf{h} + \mathbf{W}_{\text{phase}} \cdot \mathbf{p})$ and the gated application $\mathbf{h} + \text{sigmoid}(g_{\text{osc}}) \cdot A_\gamma \cdot \mathbf{m}_{\text{osc}}$ (Claim 16).

FIG. P1-17 is a diagram illustrating the BCM homeostatic plasticity mechanism, showing the sliding threshold θ_{BCM} updated via exponential moving average of mean squared postsynaptic activity, with potentiation and depression zones (Claim 17).

FIG. P1-18 is a diagram illustrating the metaplasticity module, showing the tracking of recent plasticity magnitude via EMA and the learning rate multiplier $1/(1 + \text{plasticity_ema})$ (Claim 18).

FIG. P1-19 is a diagram illustrating the structural plasticity mechanism in the Layer 2/3 Hebbian association sublayer, showing the binary connection mask, below-threshold pruning counter, and error-driven synaptogenesis lifecycle (Claim 19).

FIG. P1-20 is a diagram illustrating the metacognitive self-assessment module with zero learnable parameters, showing the inter-column conflict computed as $\text{Var}(\overline{\text{output}_1}, \ldots, \overline{\text{output}_N})$ and the mapping to additional inference iterations (Claim 20).

FIG. P1-21 is a diagram illustrating the Layer 5 dendritic gates, showing the apical dendritic gate modulated by Layer 1 top-down context and the basal dendritic gate modulated by Layer 4 thalamic input, both with bias initialised to 3.0 ($\text{sigmoid}(3.0) \approx 0.95$) (Claim 21).

FIG. P1-22 is a diagram illustrating the per-token prediction precision σ^2 computation, showing the softplus activation with depth-dependent initialisation and the developmental gradient from shallow to deep columns (Claim 22).

FIG. P1-23 is a flow diagram illustrating the predictive coding inference iterations performed without gradient computation, showing the convergence monitoring via relative prediction error change and early termination criteria (Claim 23).

FIG. P1-24 is a diagram illustrating the Layer 2/3 recurrent settling dynamics, showing the iterative association computation with re-anchoring to the pre-L2/3 input at each settling step (Claim 24).

FIG. P1-25 is a diagram illustrating the depth-dependent σ^2 temporal smoothing via exponential moving average, showing the time constant $\alpha_i = 1/(1 + s \cdot i)$ from instantaneous tracking at shallow columns to integration over approximately 16 tokens at deep columns (Claim 25).

FIG. P1-26 is a diagram illustrating the three Hebbian learning modes for the association matrix $\mathbf{W}_{\text{assoc}}$: basic Hebbian, BCM-modulated Hebbian, and error-modulated BCM Hebbian

with blend weight λ (Claim 26).

FIG. P1-27 is a table diagram illustrating the disease-mapping validation methodology, showing the correspondence between architectural component removal and specific neurological or psychiatric disease analogues (Claim 27).

FIG. P1-28 is a diagram illustrating the multi-scale parameter configurations, showing the proportional scaling of d_{model}, N (number of columns), and d_{ff} across the 106M, 300M, and 1B parameter presets while preserving all biological mechanisms (Claim 28).

DRAWING SHEETS

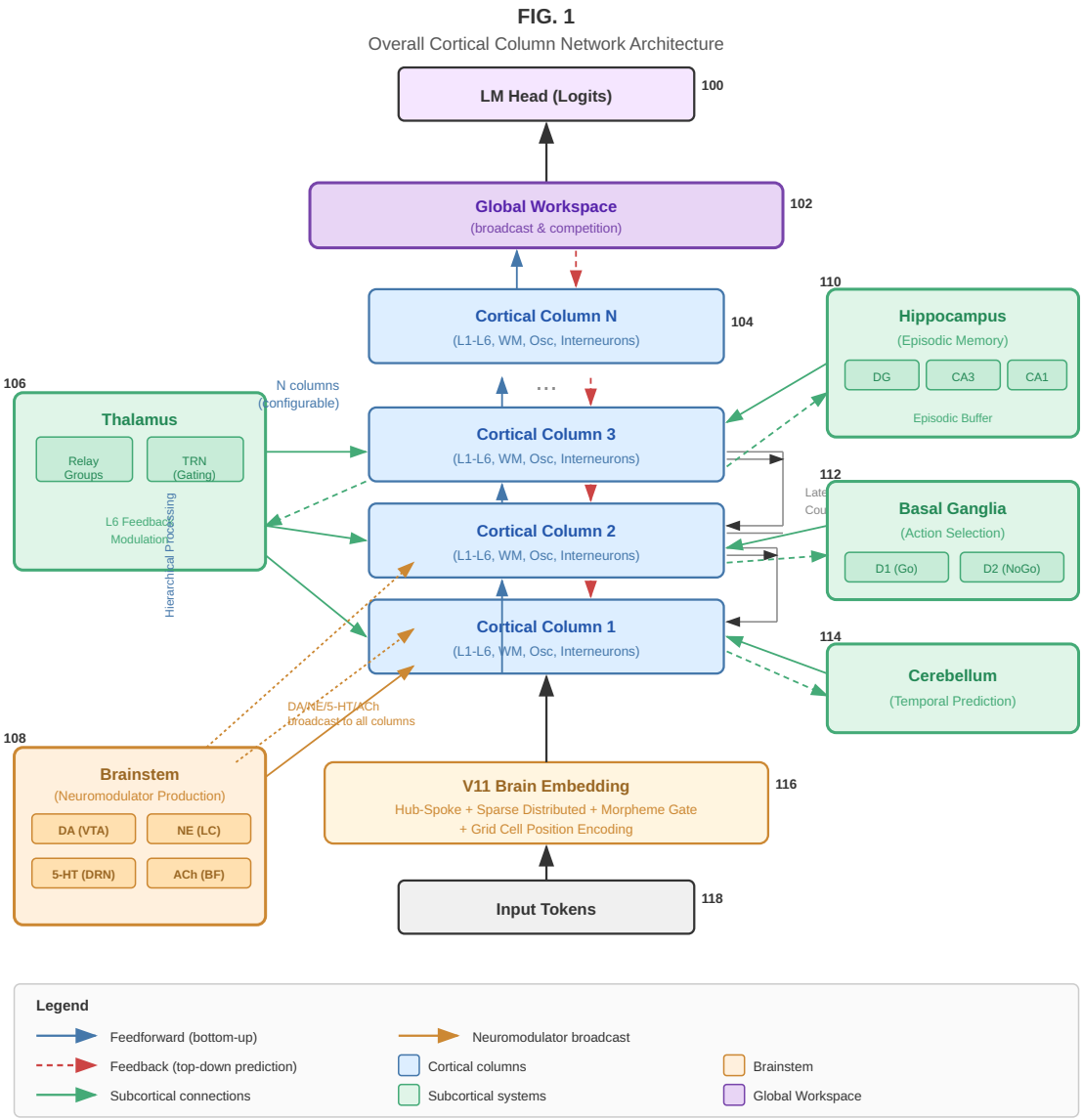


Figure 1: FIG. 1 — Overall Architecture of the Cortical Column Network System

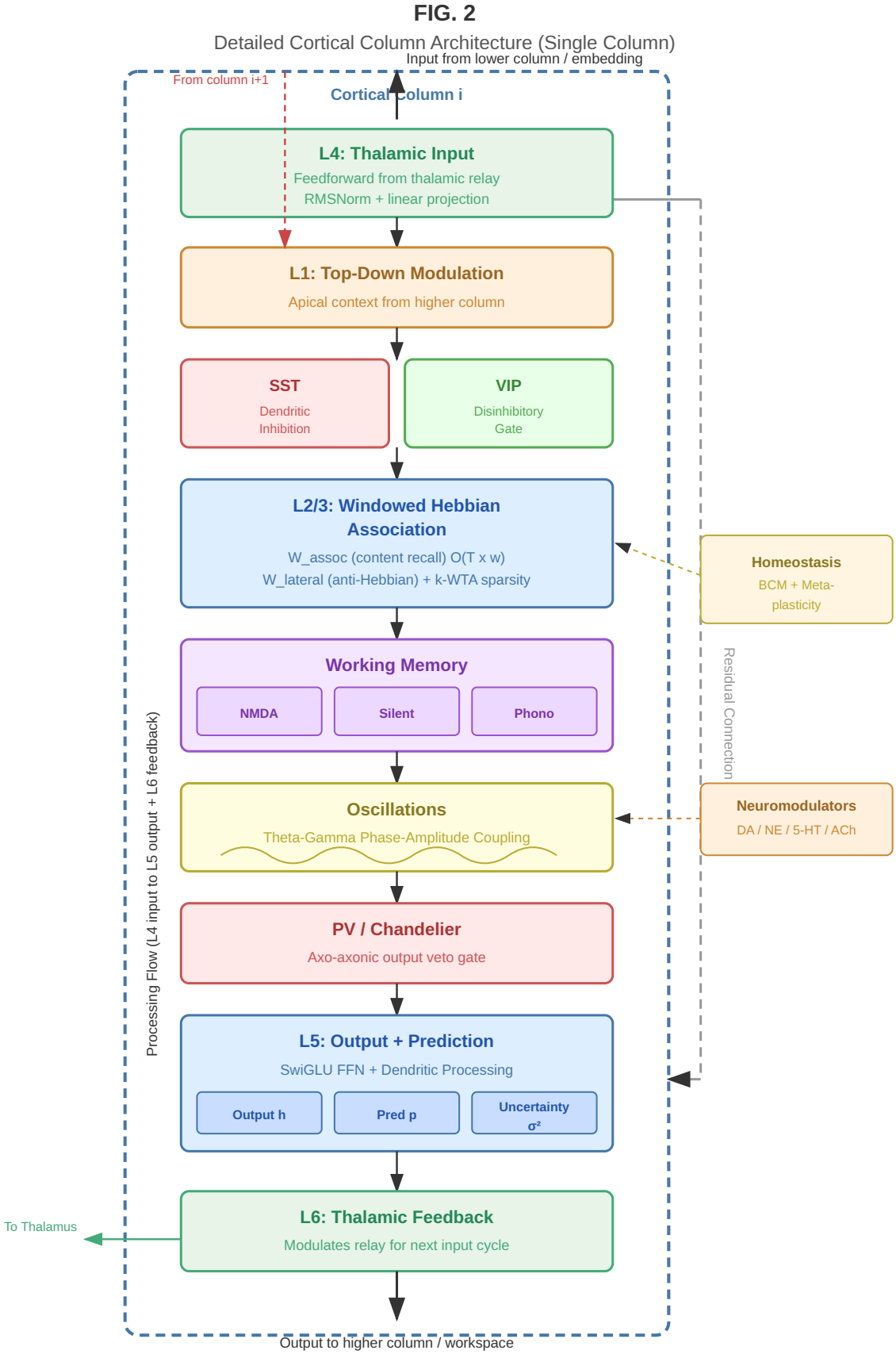


Figure 2: FIG. 2 — Cortical Column Processing Unit: Nine-Sublayer Pipeline

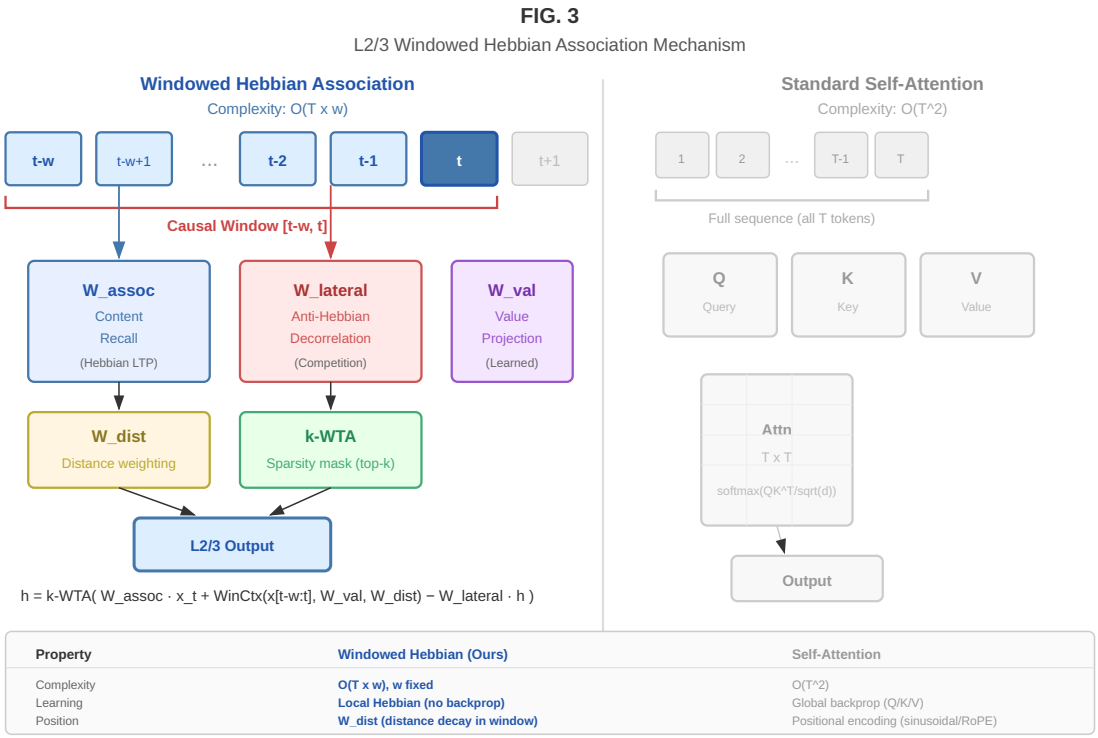


Figure 3: FIG. 3 — Windowed Hebbian Association vs. Standard Self-Attention

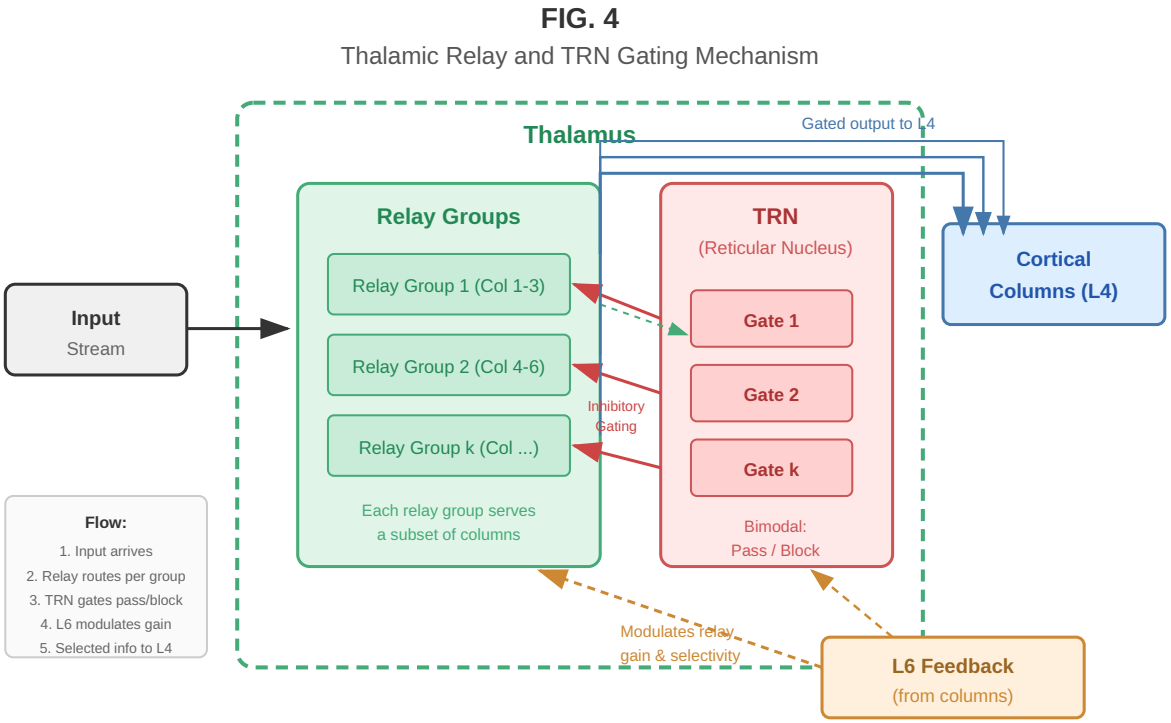


Figure 4: FIG. 4 — Thalamic Routing System: Relay Groups, TRN Gate, and L6 Feedback

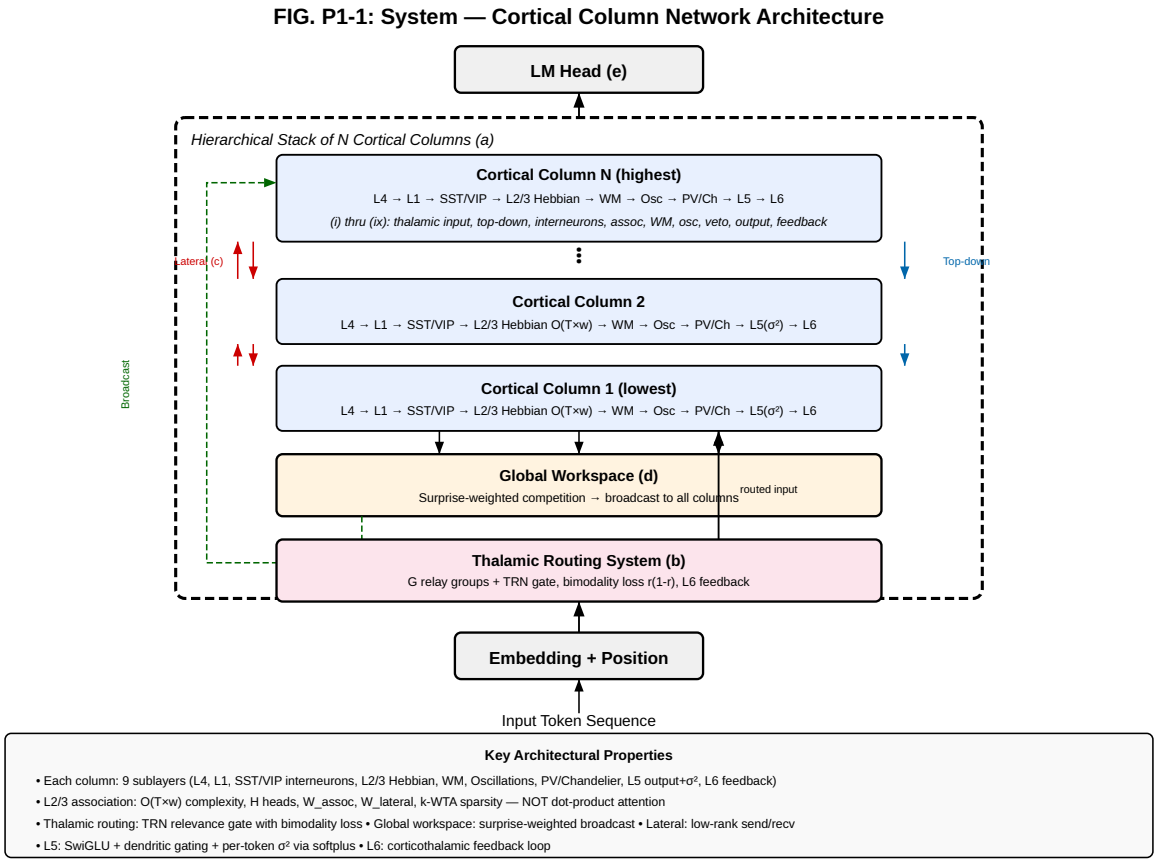


Figure 5: FIG. P1-1 — System: Cortical Column Network Architecture (Claim 1)

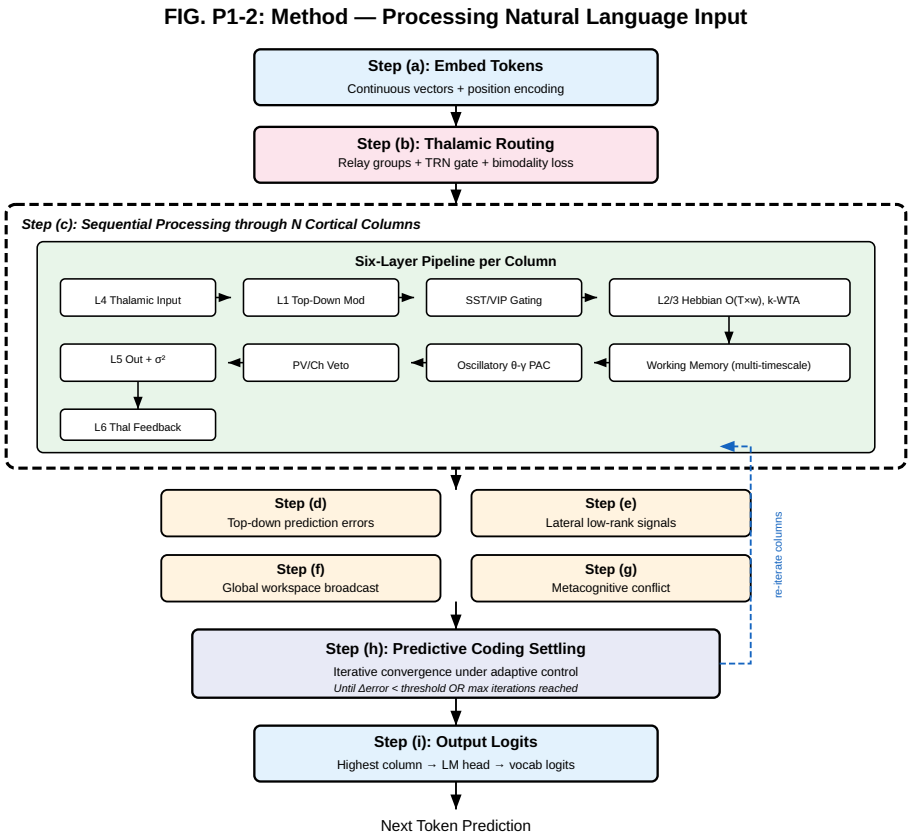


Figure 6: FIG. P1-2 — Method: Processing Natural Language Input (Claim 2)

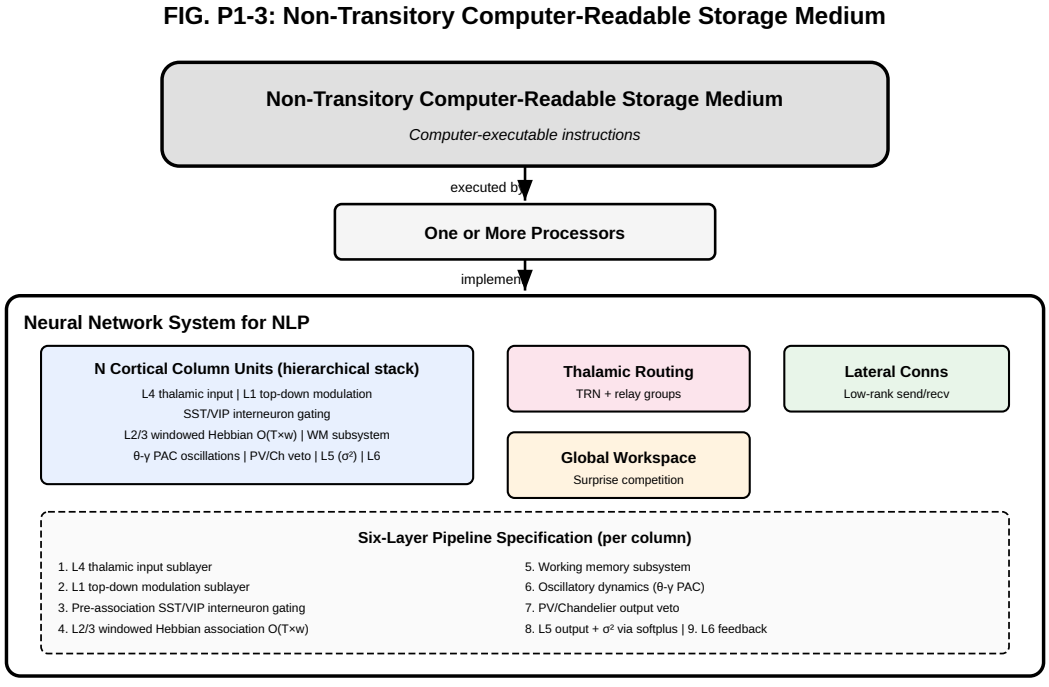


Figure 7: FIG. P1-3 — Non-Transitory Computer-Readable Storage Medium (Claim 3)

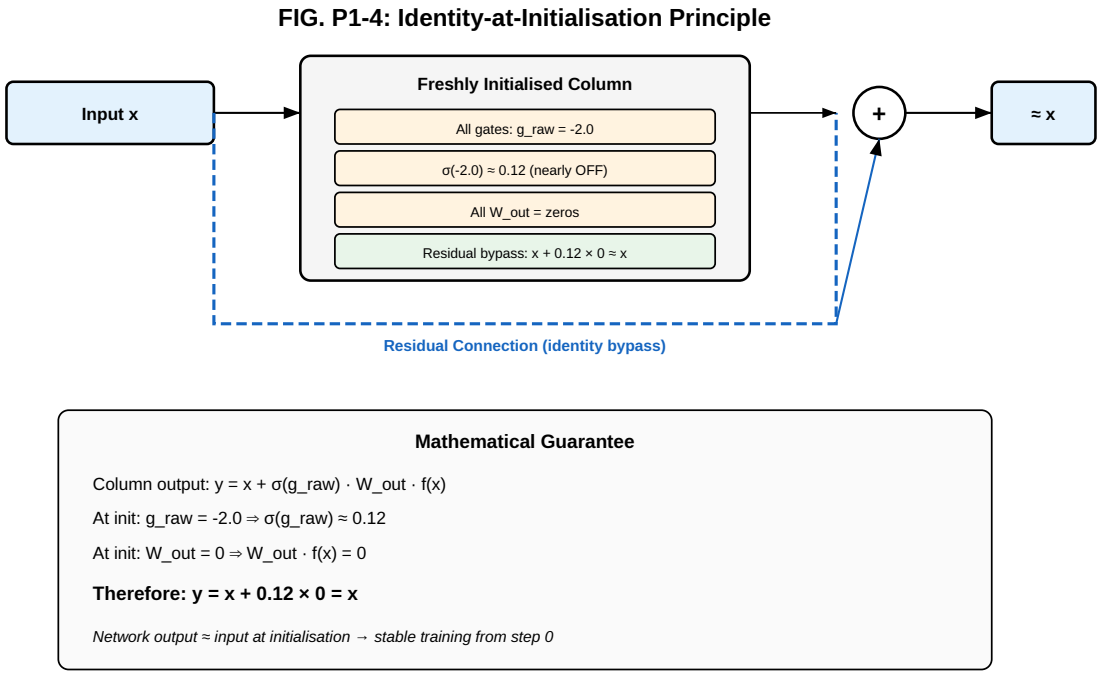


Figure 8: FIG. P1-4 — identity-at-init(Stale from transformer era check with /nct-research)ialisation Principle (Claim 4)

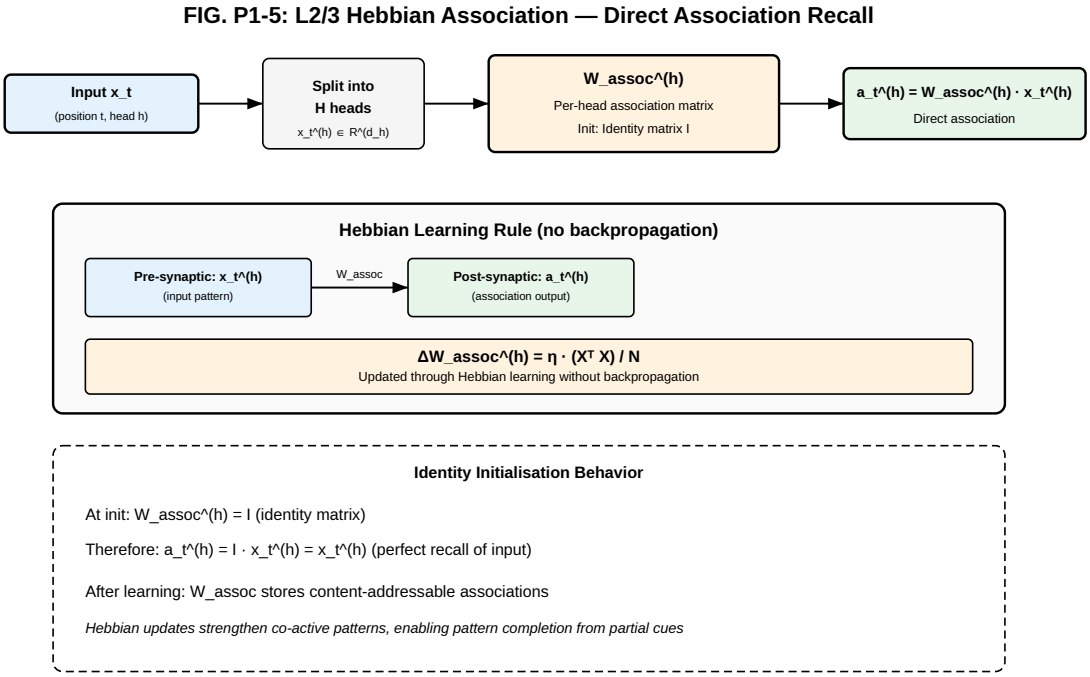


Figure 9: FIG. P1-5 — L2/3 Hebbian Association: Direct Association Recall (Claim 5)

FIG. P1-6: Windowed Context Aggregation with Distance Weighting

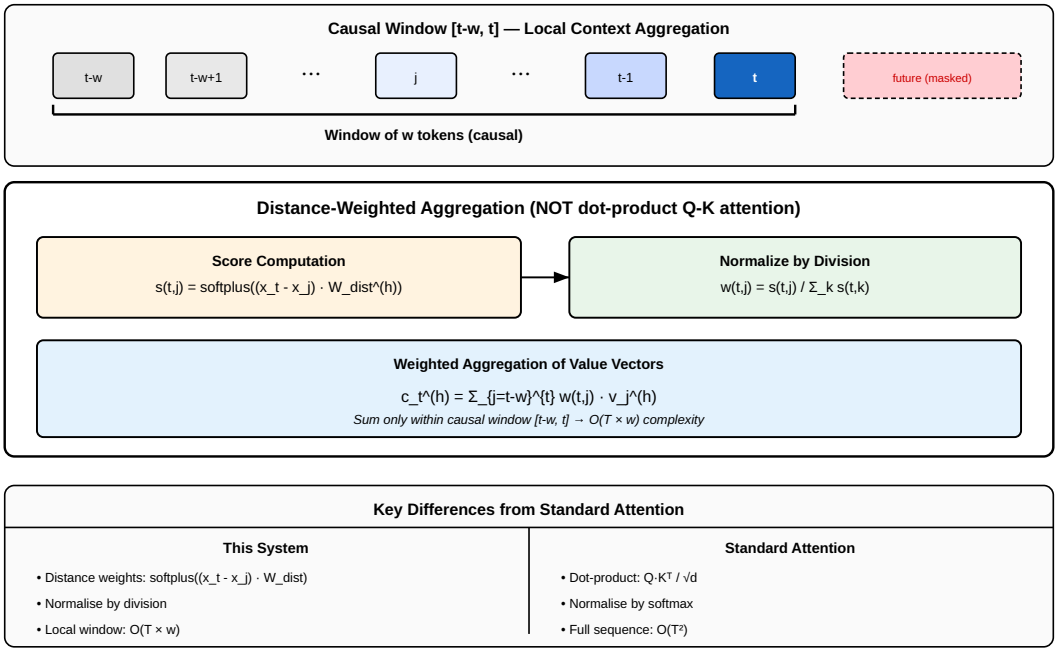


Figure 10: FIG. P1-6 — Windowed Context Aggregation with Distance Weighting (Claim 6)

FIG. P1-7: Anti-Hebbian Lateral Decorrelation

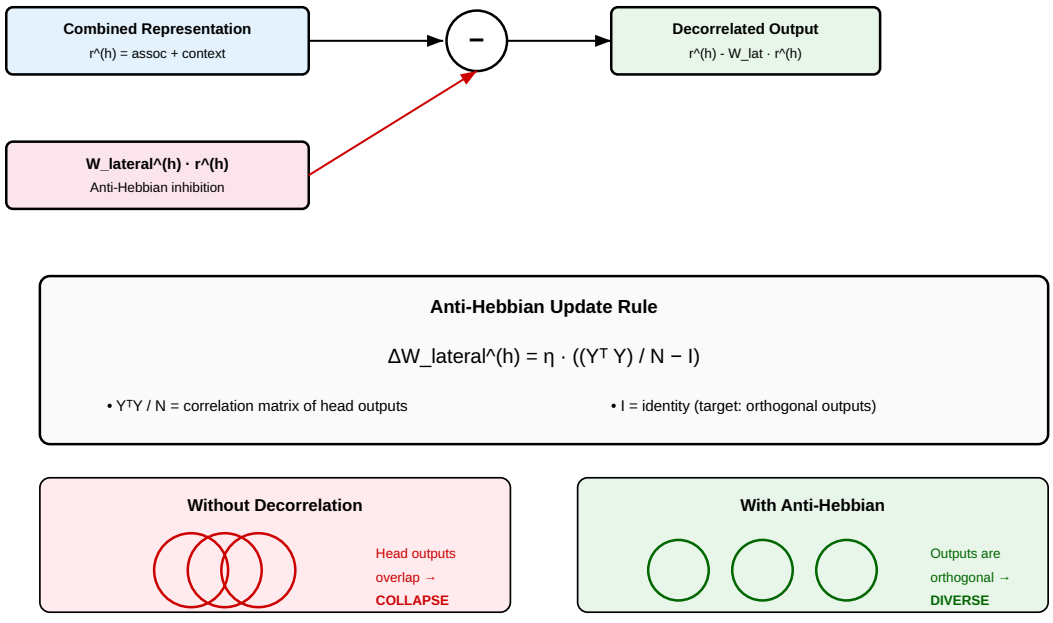


Figure 11: FIG. P1-7 — Anti-Hebbian Lateral Decorrelation (Claim 7)

FIG. P1-8: k-Winner-Take-All Competitive Coding

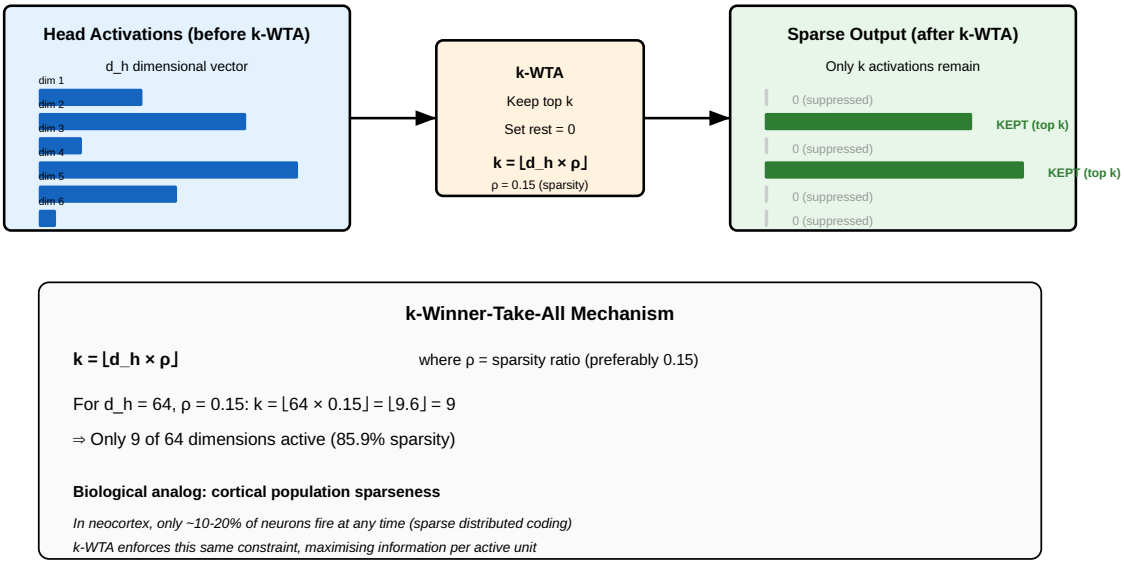


Figure 12: FIG. P1-8 — k-Winner-Take-All Competitive Coding (Claim 8)

FIG. P1-9: VIP Disinhibition and SST Dendritic Inhibition

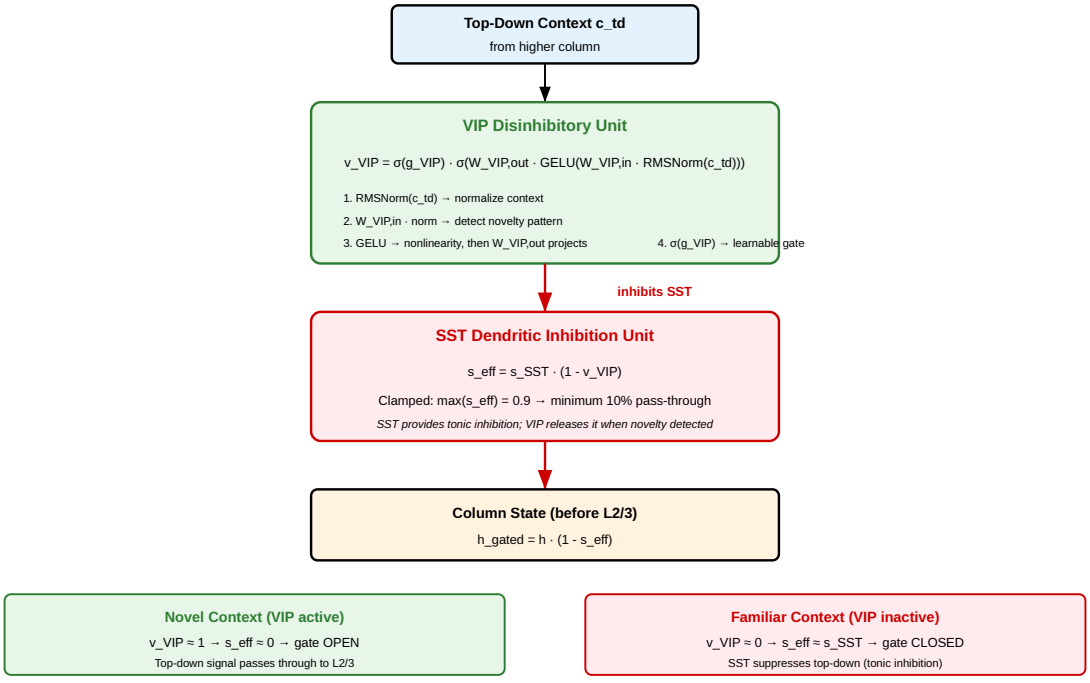


Figure 13: FIG. P1-9 — VIP Disinhibition and SST Dendritic Inhibition (Claim 9)

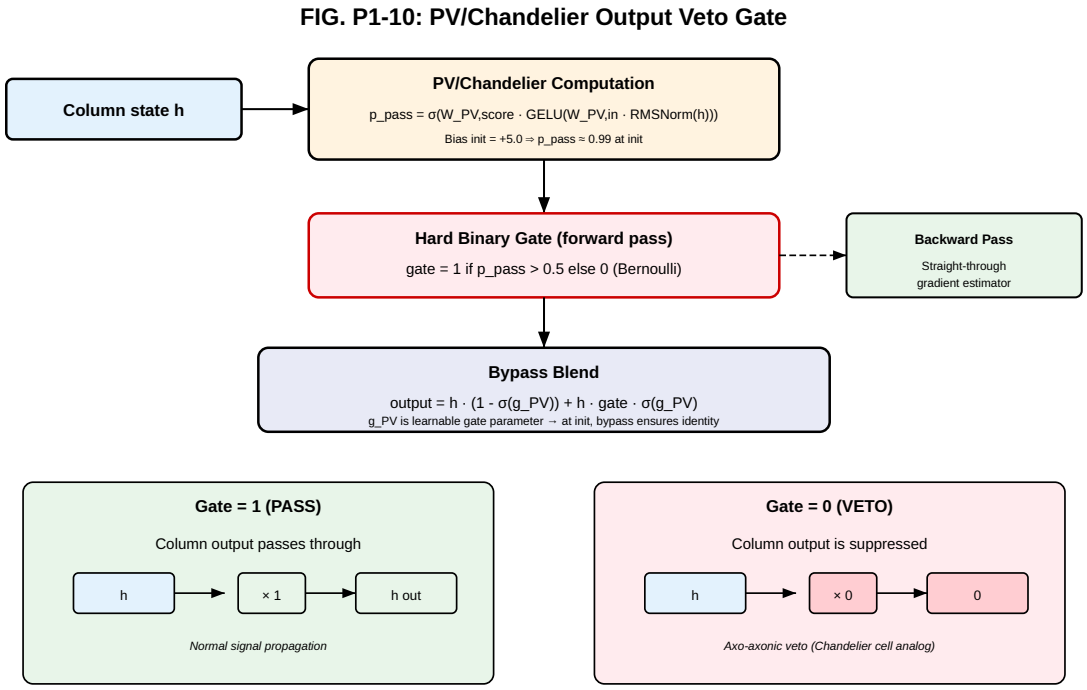


Figure 14: FIG. P1-10 — PV/Chandelier Output Veto Gate (Claim 10)

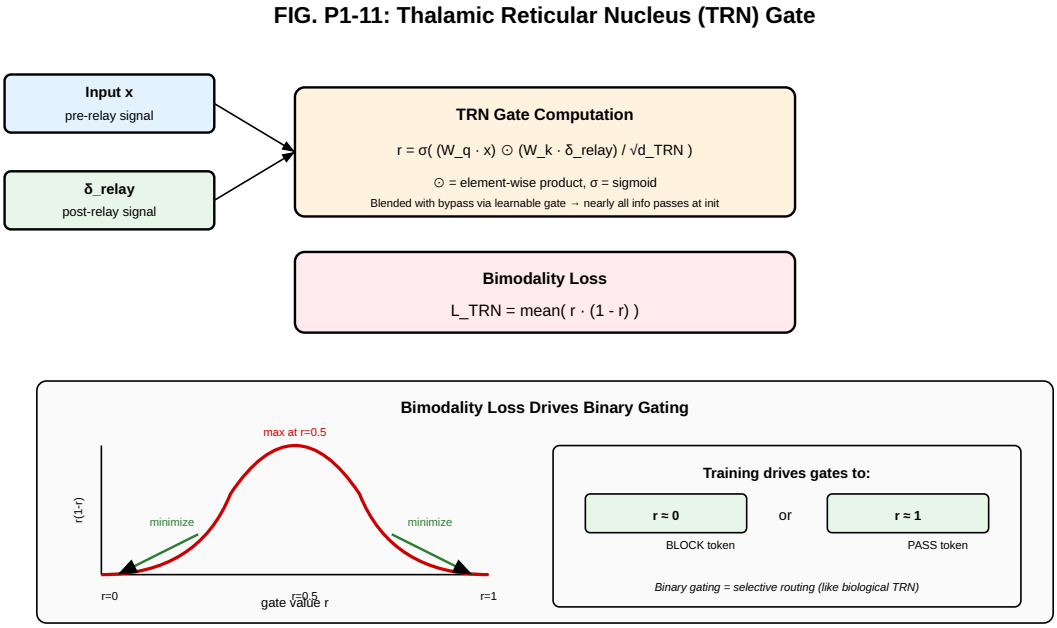


Figure 15: FIG. P1-11 — Thalamic Reticular Nucleus (TRN) Gate (Claim 11)

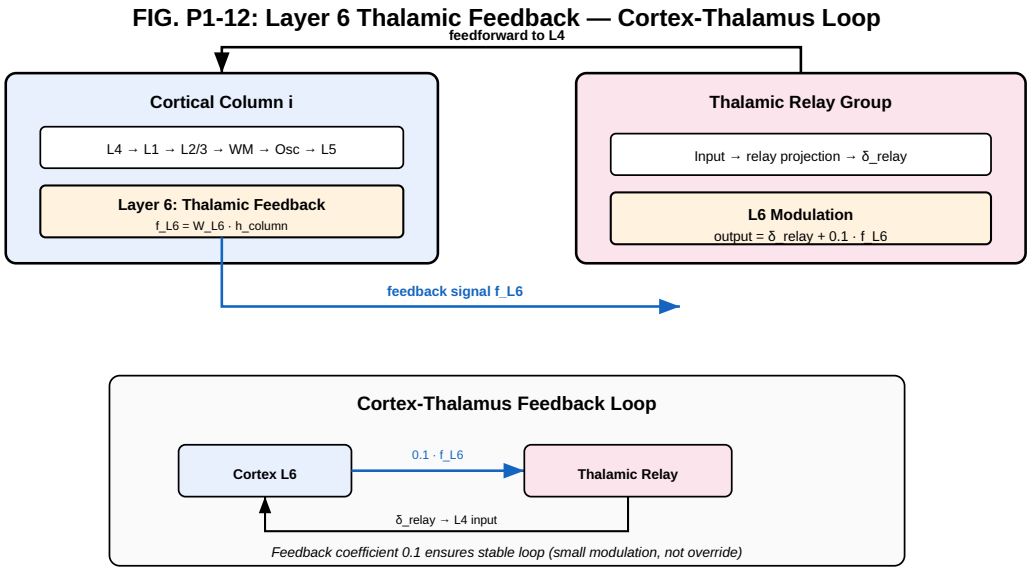


Figure 16: FIG. P1-12 — Layer 6 Thalamic Feedback: Cortex-Thalamus Loop (Claim 12)

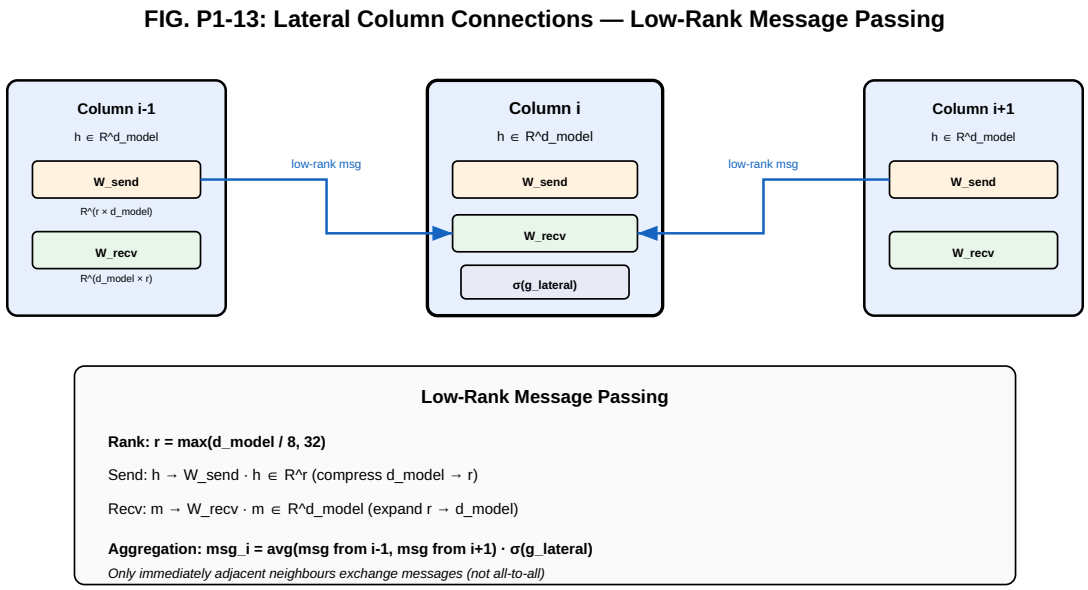


Figure 17: FIG. P1-13 — Lateral Column Connections: Low-Rank Message Passing (Claim 13)

FIG. P1-14: Global Workspace — Surprise-Weighted Broadcast

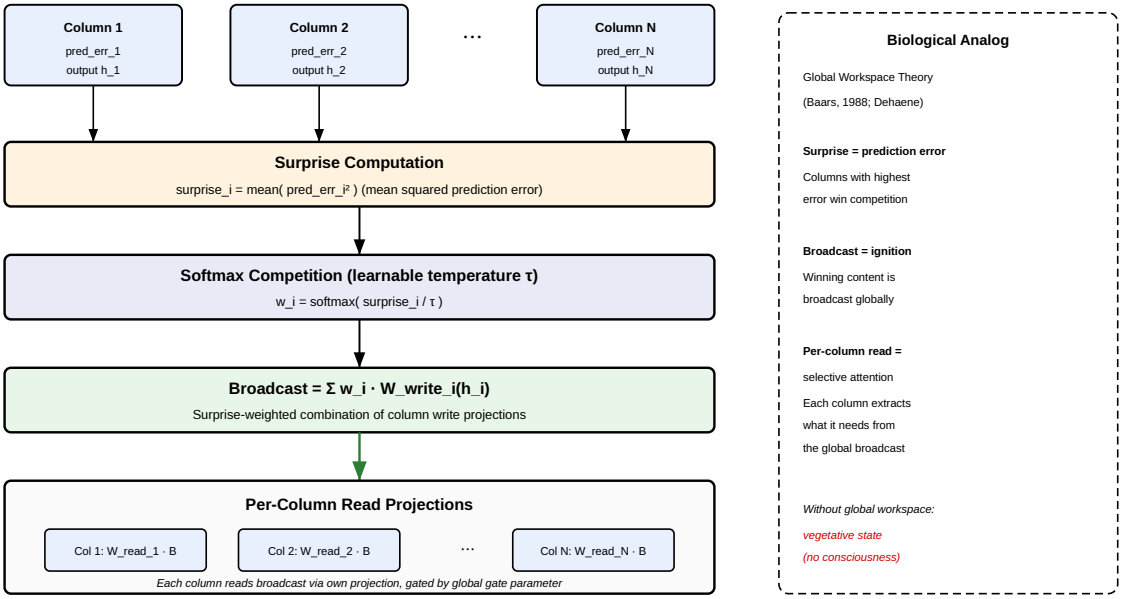


Figure 18: FIG. P1-14 — Global Workspace: Surprise-Weighted Broadcast (Claim 14)

FIG. P1-15: Oscillatory Gamma Amplitude Modulation

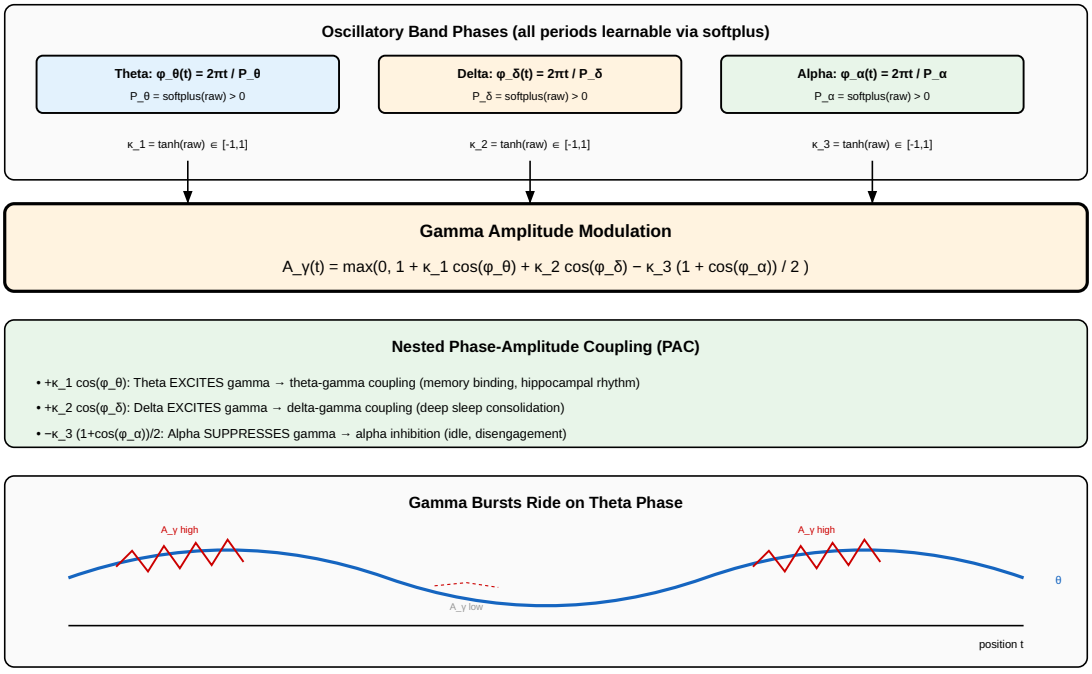


Figure 19: FIG. P1-15 — Oscillatory Gamma Amplitude Modulation (Claim 15)

FIG. P1-16: Phase-Content Bilinear Coupling + Oscillatory Modulation

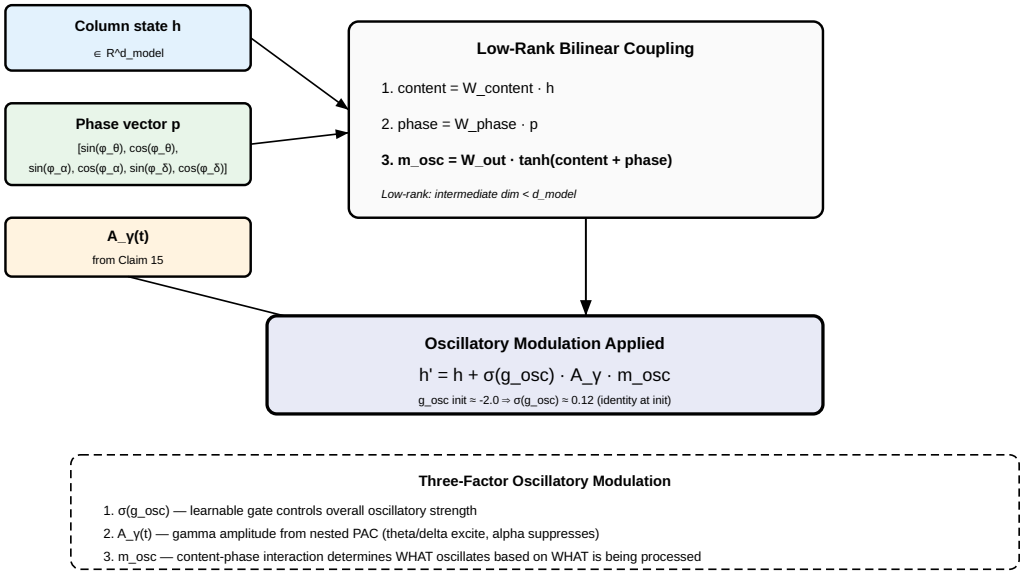


Figure 20: FIG. P1-16 — Phase-Content Bilinear Coupling and Oscillatory Modulation (Claim 16)

FIG. P1-17: BCM Homeostatic Plasticity

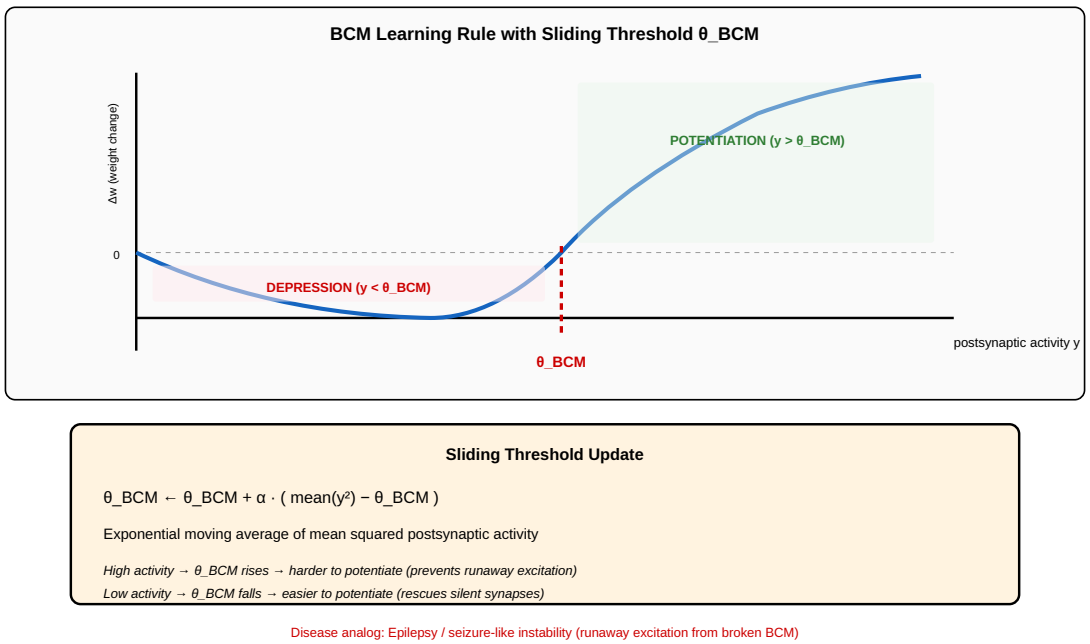
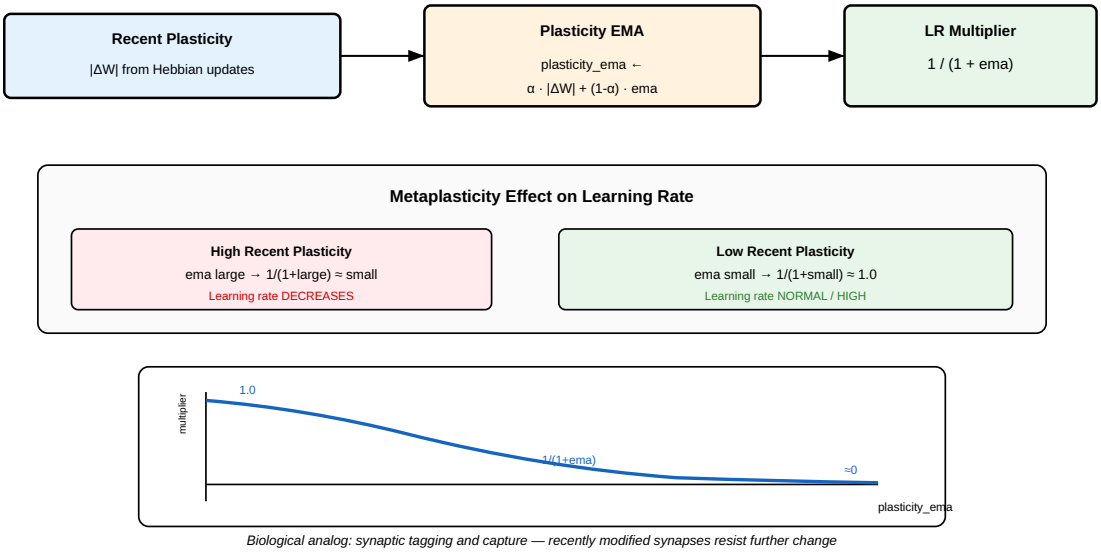


Figure 21: FIG. P1-17 — BCM Homeostatic Plasticity (Claim 17)

FIG. P1-18: Metaplasticity Module



Biological analog: synaptic tagging and capture — recently modified synapses resist further change

Figure 22: FIG. P1-18 — Metaplasticity Module (Claim 18)

FIG. P1-19: Structural Plasticity in L2/3 Hebbian Association

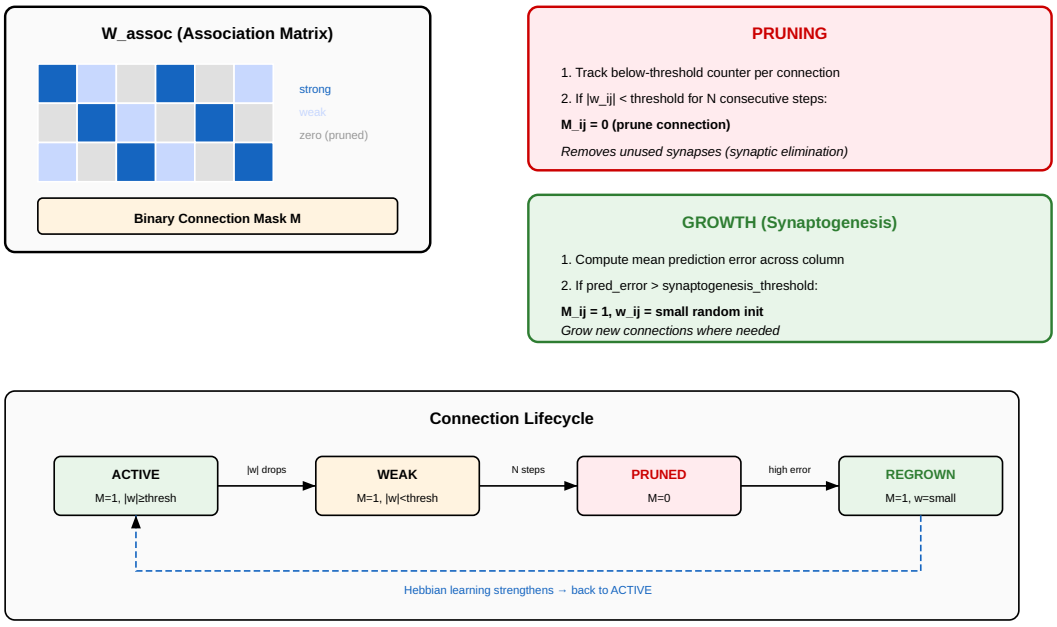


Figure 23: FIG. P1-19 — Structural Plasticity in L2/3 Hebbian Association (Claim 19)

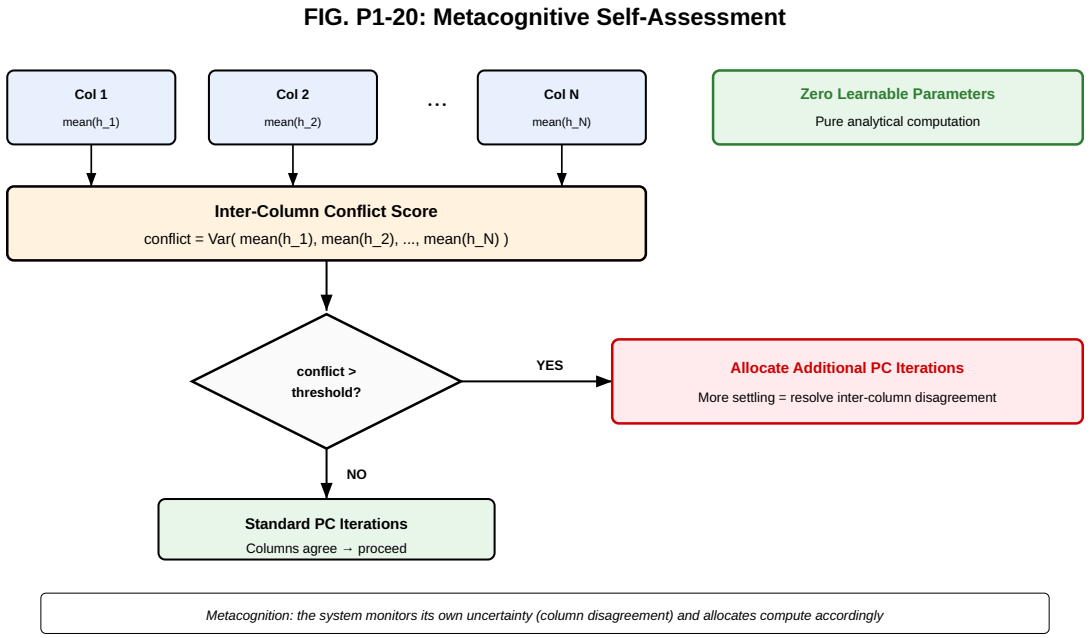


Figure 24: FIG. P1-20 — Metacognitive Self-Assessment (Claim 20)

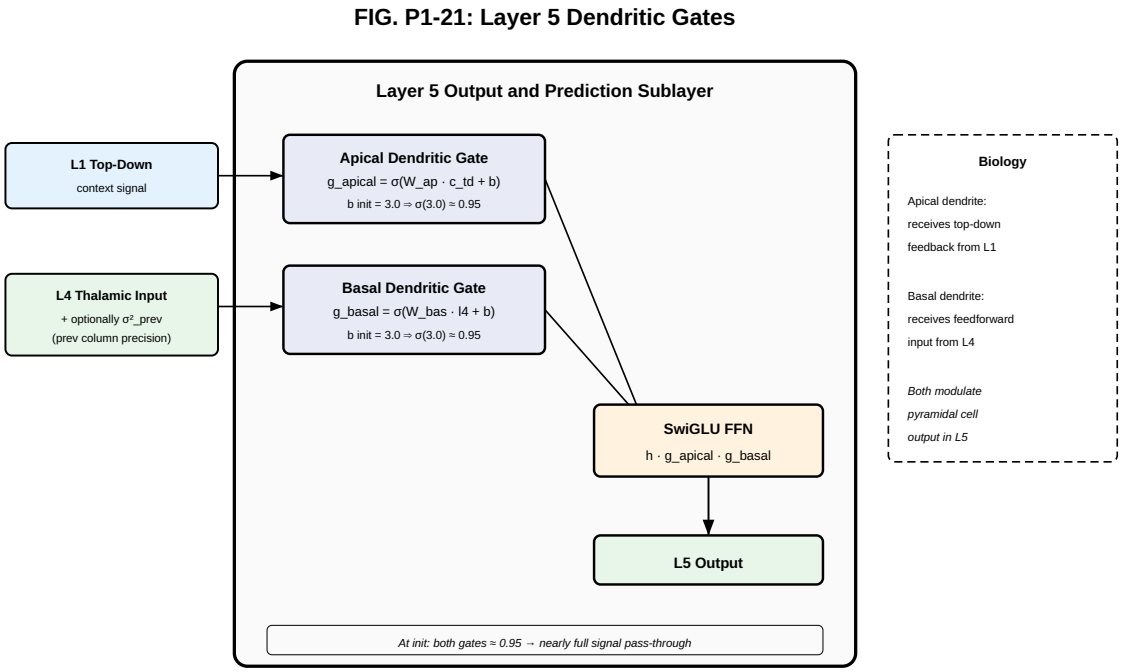


Figure 25: FIG. P1-21 — Layer 5 Dendritic Gates (Claim 21)

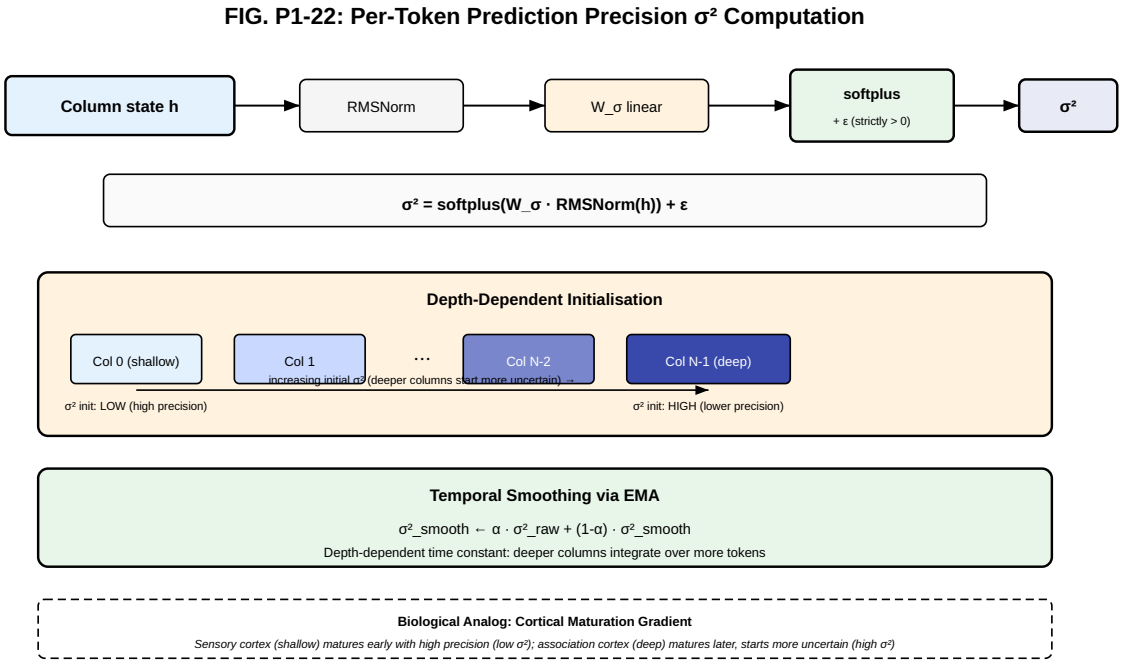


Figure 26: FIG. P1-22 — Per-Token Prediction Precision σ^2 Computation (Claim 22)

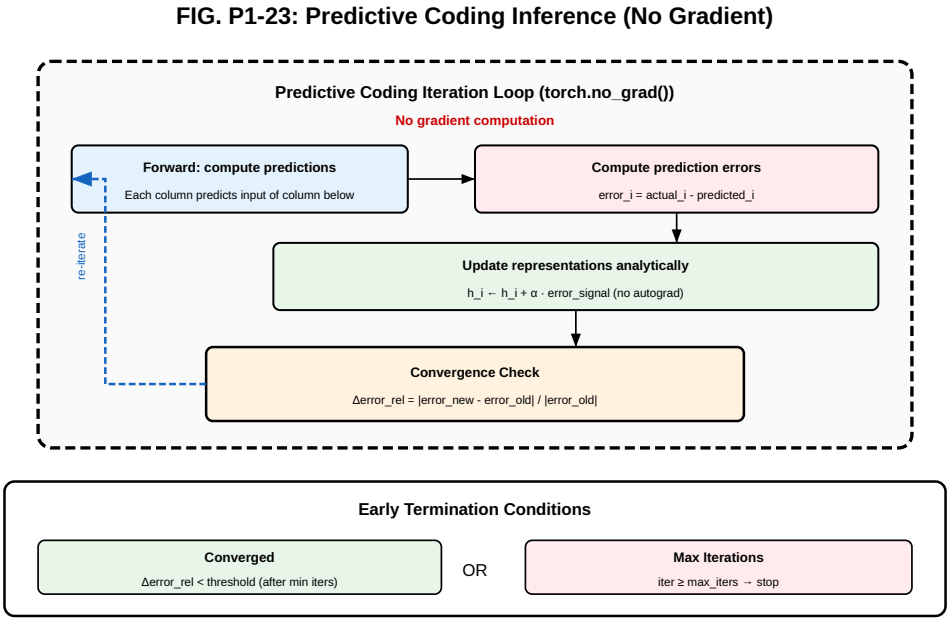


Figure 27: FIG. P1-23 — Predictive Coding Inference (No Gradient) (Claim 23)

FIG. P1-24: L2/3 Recurrent Settling with Re-Anchoring

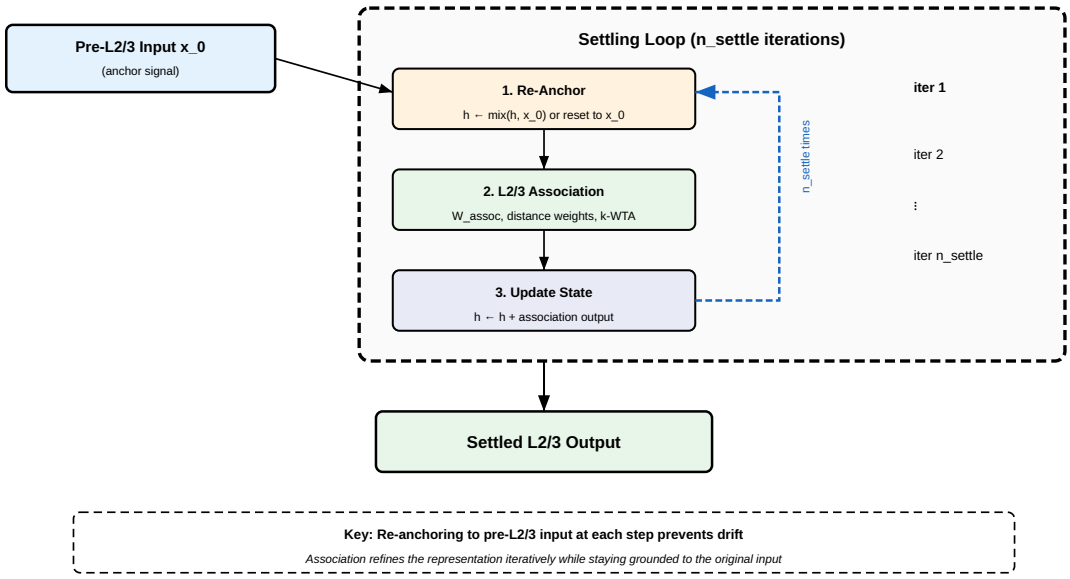


Figure 28: FIG. P1-24 — L2/3 Recurrent Settling with Re-Anchoring (Claim 24)

FIG. P1-25: Depth-Dependent σ² Temporal Smoothing

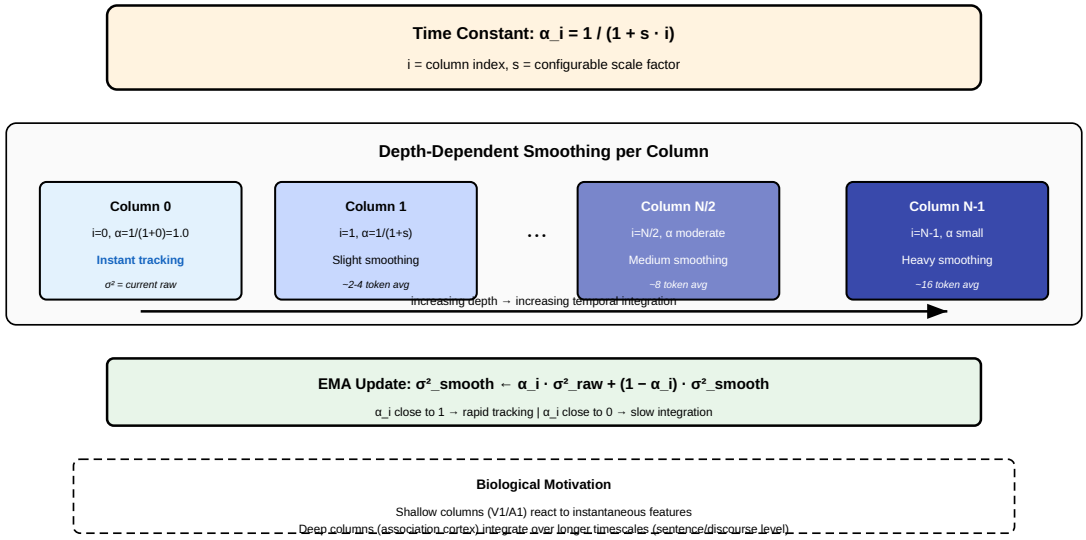


Figure 29: FIG. P1-25 — Depth-Dependent σ^2 Temporal Smoothing (Claim 25)

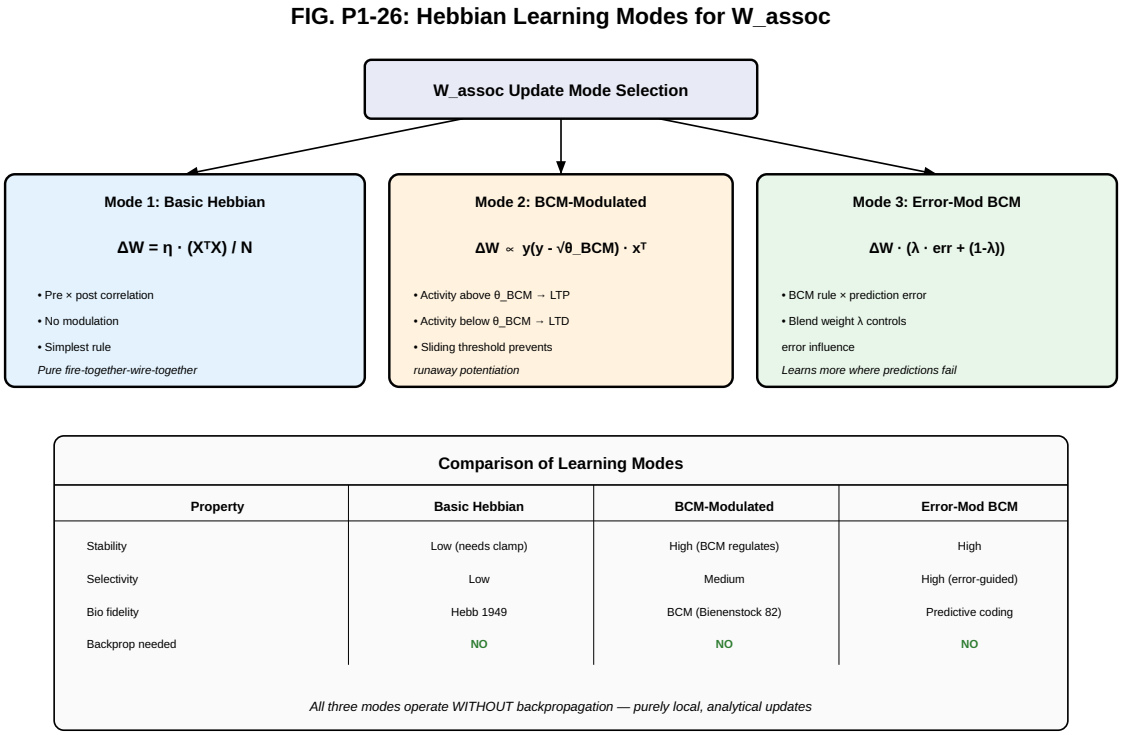


Figure 30: FIG. P1-26 — Hebbian Learning Modes for $\mathbf{W}_{\text{assoc}}$ (Claim 26)

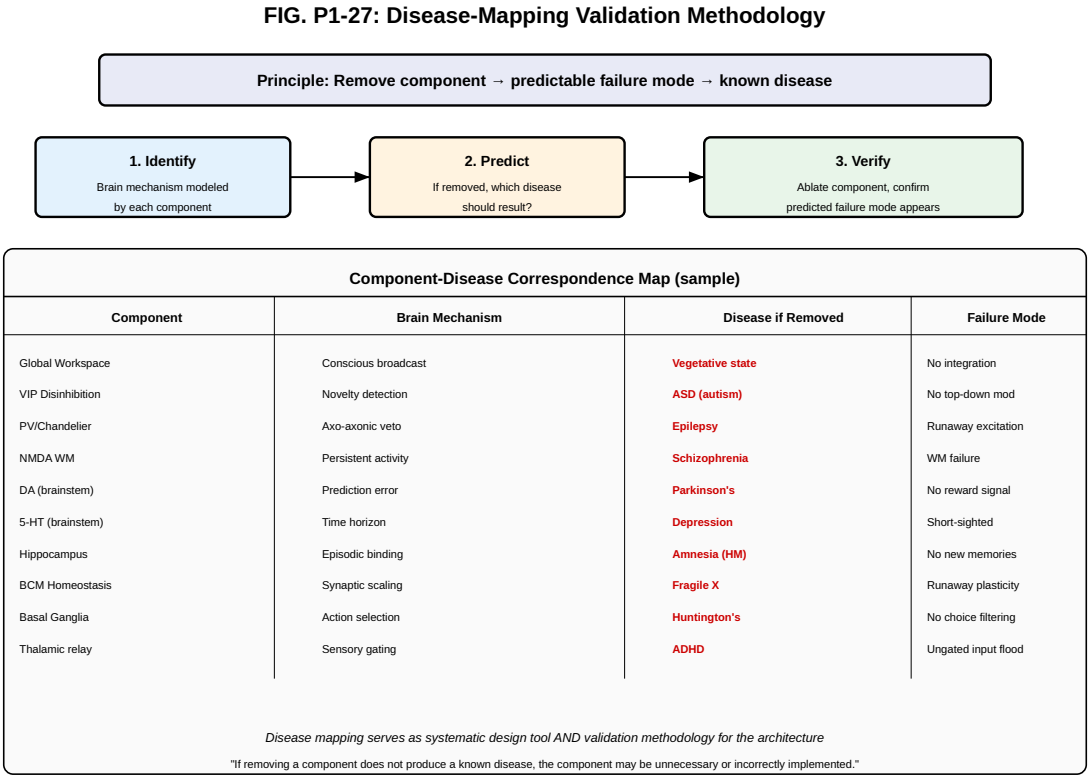


Figure 31: FIG. P1-27 — Disease-Mapping Validation Methodology (Claim 27)

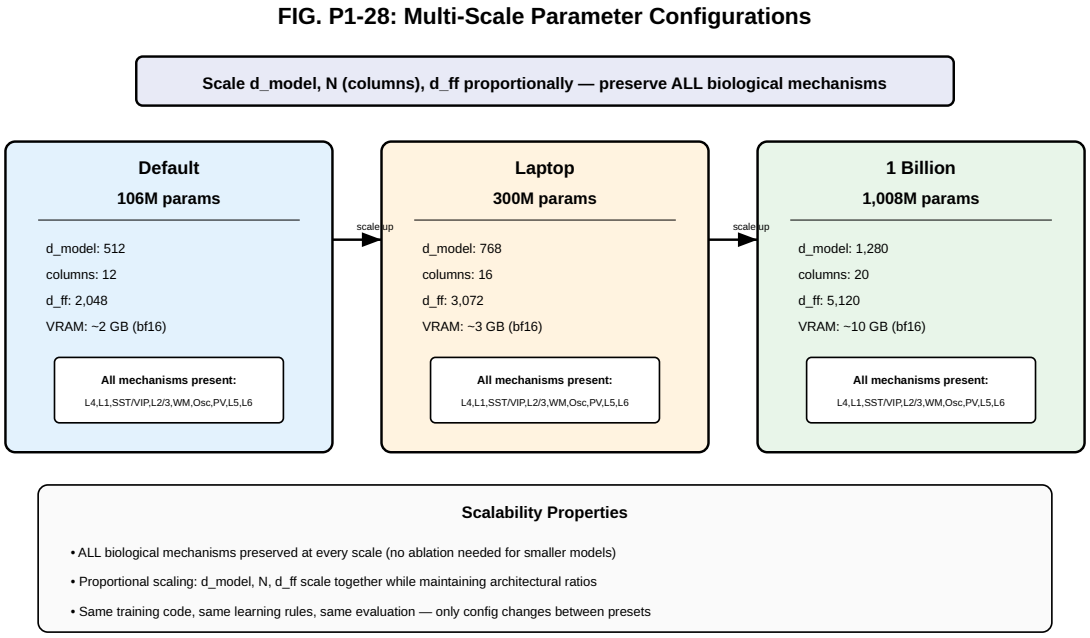


Figure 32: FIG. P1-28 — Multi-Scale Parameter Configurations (Claim 28)

5. DETAILED DESCRIPTION

The following detailed description discloses the cortical column network (CCN) system and method with sufficient specificity to enable a person skilled in the art of neural network architecture and natural language processing to reproduce the invention. Reference is made throughout to the accompanying drawings. Unless otherwise defined, all technical terms used herein have the meanings commonly understood by one of ordinary skill in the art.

5.1. System Overview (FIG. 1, FIG. P1-1)

Referring to FIG. 1 and FIG. P1-1, the cortical column network system comprises the following principal components:

(a) An **embedding subsystem** that converts discrete input token identifiers into continuous d_{model}-dimensional vector representations, with additive grid-cell position encoding;

(b) A **thalamic routing system** comprising G relay groups and a thalamic reticular nucleus (TRN), routing information between cortical columns;

(c) A **hierarchical stack** of N cortical column processing units (columns 0 through $N-1$), each implementing a six-layer biological pipeline;

(d) A **lateral column connection system** providing horizontal communication between adjacent columns via low-rank message passing;

(e) A **global workspace module** implementing surprise-weighted broadcast competition among all columns;

(f) A **metacognitive self-assessment module** computing inter-column conflict and allocating adaptive computation;

(g) A set of **subcortical systems** including a brainstem neuromodulation module, a basal ganglia action selection module, a hippocampal episodic memory module, and a cerebellar temporal prediction module; and

(h) A **language model head** converting the final column output into vocabulary logit scores.

In a preferred embodiment, the system is configured with $d_{\text{model}} = 512$, $N = 12$ columns, and a vocabulary size of approximately 14,000 tokens, yielding a model with approximately 106 million parameters. Scaling to 300 million parameters ($d_{\text{model}} = 768$, $N = 16$) and 1 billion parameters ($d_{\text{model}} = 1{,}280$, $N = 20$) is achieved by proportionally increasing dimensions while preserving all architectural features.

5.2. Cortical Column Processing Unit (FIG. 2, FIG. P1-2)

Referring to FIG. 2 and FIG. P1-2, each cortical column processing unit i (where $0 \leq i < N$) comprises the following functional sublayers, wired in a specific biological pipeline order:

5.2.1. Layer 4: Thalamic Input Sublayer

The Layer 4 (L4) thalamic input sublayer receives feedforward input $\mathbf{x} \in \mathbb{R}^{B \times T \times d_{\text{model}}}$ from the thalamic routing system and projects it into the intra-column processing space:

$$\mathbf{h}_{\text{L4}} = \text{RMSNorm}(\mathbf{W}_{\text{L4}} \cdot \mathbf{x}) \tag{2}$$

where $\mathbf{W}_{\text{L4}} \in \mathbb{R}^{d_{\text{col}} \times d_{\text{model}}}$ is a learned projection matrix and RMSNorm denotes root mean square layer normalisation. When $d_{\text{col}} = d_{\text{model}}$, $\mathbf{W}_{\text{L4}}$ is initialised as an identity matrix (identity-at-init), so the sublayer begins as a pass-through.

This sublayer corresponds to Layer 4 of the mammalian neocortex, which is the primary recipient of thalamic afferents. In visual cortex, L4 contains orientation-selective simple cells; in language cortex, it performs token-level feature extraction.

5.2.2. Layer 1: Top-Down Modulation Sublayer

The Layer 1 (L1) top-down modulation sublayer receives a top-down context signal $\mathbf{c}_{\text{td}} \in \mathbb{R}^{B \times T \times d_{\text{model}}}$ from a higher column (or from the global workspace) and produces a modulatory signal:

$$\mathbf{m}_{\text{L1}} = \text{sigmoid}(g_{\text{L1}}) \cdot \text{RMSNorm}(\mathbf{W}_{\text{L1}} \cdot \mathbf{c}_{\text{td}}) \tag{3}$$

where g_{L1} is a learnable scalar gate parameter and $\mathbf{W}_{\text{L1}} \in \mathbb{R}^{d_{\text{col}} \times d_{\text{model}}}$ is a learned projection initialised to zeros. The gate is initialised to $g_{\text{L1}} = -2.0$, yielding $\text{sigmoid}(-2.0) \approx 0.12$, so the top-down modulation is nearly suppressed at initialisation and emerges only through learning.

This sublayer corresponds to Layer 1 of the neocortex, which contains apical dendrite tufts receiving feedback from higher cortical areas. The modulatory signal is added to the L4 output: $\mathbf{h} = \mathbf{h}_{\text{L4}} + \mathbf{m}_{\text{L1}}$.

5.2.3. Interneuron Gating Circuits — Pre-Association Phase (FIG. P1-9)

Before the Layer 2/3 association step, the column state $\mathbf{h}$ is processed by the SST and VIP interneuron circuits.

VIP Disinhibitory Circuit. The VIP (vasoactive intestinal peptide) interneuron circuit detects novelty in the top-down signal and produces a disinhibition signal:

$$\mathbf{v}_{\text{VIP}} = \text{sigmoid}(g_{\text{VIP}}) \cdot \text{sigmoid}(\mathbf{W}_{\text{VIP,out}} \cdot \text{GELU}(\mathbf{W}_{\text{VIP,in}} \cdot \text{RMSNorm}(\mathbf{c}_{\text{td}}))) \tag{4}$$

where g_{VIP} is a learnable gate parameter initialised to -2.0, $\mathbf{W}_{\text{VIP,in}} \in \mathbb{R}^{d_i \times d_{\text{col}}}$ is an input projection, and $\mathbf{W}_{\text{VIP,out}} \in \mathbb{R}^{d_{\text{col}} \times d_i}$ is an output projection initialised to zeros. d_i is the inhibitory interneuron intermediate dimension (128 in the preferred embodiment). The VIP circuit is only active when top-down context is available.

SST Dendritic Inhibition. The SST (somatostatin) interneuron circuit implements dendritic input gating via tonic inhibition modulated by VIP disinhibition:

$$\mathbf{s}_{\text{SST}} = \text{sigmoid}(g_{\text{SST}}) \cdot (\tau_{\text{tonic}} + (1 - \tau_{\text{tonic}}) \cdot \text{sigmoid}(\mathbf{W}_{\text{SST,out}} \cdot \text{GELU}(\mathbf{W}_{\text{SST,in}} \cdot \text{RMSNorm}(\mathbf{h})))) \tag{5}$$

where $\tau_{\text{tonic}} = 0.10$ models the intrinsic sodium current that maintains baseline SST firing. When VIP disinhibition is present, the effective SST inhibition is reduced:

$$\mathbf{s}_{\text{eff}} = \min(\mathbf{s}_{\text{SST}} \cdot (1 - \mathbf{v}_{\text{VIP}}),\ 0.9) \tag{6}$$

The column state is then gated: $\mathbf{h} \leftarrow \mathbf{h} \cdot (1 - \mathbf{s}_{\text{eff}})$. The clamp at 0.9 prevents complete suppression of the column, maintaining a minimum 10% signal pass-through.

This circuit models the canonical interneuron motif: VIP $\dashv$ SST $\dashv$ Pyramidal (apical dendrites). VIP cells are activated by top-down novelty and inhibit SST cells, thereby disinhibiting pyramidal cells and opening the column to learning from new contexts.

5.2.4. Layer 2/3: Windowed Hebbian Association (FIG. 3, FIG. P1-5 through P1-8)

The Layer 2/3 (L2/3) windowed Hebbian association sublayer is the central computation of each cortical column. It replaces the transformer's self-attention mechanism with a biologically-grounded content-addressable recall mechanism operating within a local causal window.

Key Distinction from Self-Attention. Self-attention computes Q, K, V projections and a global dot-product attention matrix with $O(T^2)$ complexity. The present invention uses no $Q/K/V$ projections, no softmax attention weights, no causal mask, and no $O(T^2)$ computation. Instead, it employs Hebbian association weights that are updated through local correlation-based learning.

Architecture. L2/3 comprises H association heads (8 in the preferred embodiment), each with dimension $d_h = d_{\text{col}}/H$. Each head maintains the following parameters:

(i) **Association matrix** $\mathbf{W}_{\text{assoc}}^{(h)} \in \mathbb{R}^{d_h \times d_h}$: A content-addressable memory that recalls stored patterns. Initialised as identity (recalls the input itself).

(ii) **Value projection** $\mathbf{W}_{\text{val}} \in \mathbb{R}^{d_{\text{col}} \times d_{\text{col}}}$: Extracts value information from the window context. Initialised as identity.

(iii) **Distance weighting** $\mathbf{W}_{\text{dist}}^{(h)} \in \mathbb{R}^{d_h}$: Per-head learned distance weights for the local window context. Initialised to zeros (uniform weighting).

(iv) **Anti-Hebbian lateral weights** $\mathbf{W}_{\text{lateral}}^{(h)} \in \mathbb{R}^{d_h \times d_h}$: Decorrelation matrix enforcing orthogonal head representations. Initialised to zeros.

(v) **Output projection** $\mathbf{W}_{\text{out}} \in \mathbb{R}^{d_{\text{col}} \times d_{\text{col}}}$: Combines heads and projects to column dimension. Initialised to zeros (identity-at-init via residual).

(vi) **Gate parameter** g_{L23}: Scalar gate initialised to -2.0.

Forward Computation. For input $\mathbf{x} \in \mathbb{R}^{B \times T \times d_{\text{col}}}$:

1. **Normalise**: $\mathbf{x}_n = \text{RMSNorm}(\mathbf{x})$.
2. **Multi-head reshape**: $\mathbf{x}_h \in \mathbb{R}^{B \times H \times T \times d_h}$.
3. **Direct association**: For each head h, compute $\mathbf{a}_t^{(h)} = \mathbf{W}_{\text{assoc}}^{(h)} \cdot \mathbf{x}_{h,t}$ for all positions t. This is content-addressable recall: the association matrix maps each input pattern to its stored association.

4. **Windowed context**: For each position t, gather the local causal window $[t-w, t]$ where w is the window size (16 tokens in the preferred embodiment, corresponding to approximately two theta oscillation cycles). Compute distance-weighted context:

$$s_{t,j}^{(h)} = \text{softplus}\Big(\mathbf{W}_{\text{dist}}^{(h)} \cdot (\mathbf{x}_{h,t} - \mathbf{x}_{h,j})\Big), \quad j \in [t-w, t] \tag{7}$$

$$\alpha_{t,j}^{(h)} = \frac{s_{t,j}^{(h)}}{\sum_{k=t-w}^{t} s_{t,k}^{(h)} + \epsilon} \tag{8}$$

$$\mathbf{c}_t^{(h)} = \sum_{j=t-w}^{t} \alpha_{t,j}^{(h)} \cdot \mathbf{v}_{h,j} \tag{9}$$

   where $\mathbf{v} = \mathbf{W}_{\text{val}}(\mathbf{x}_n)$ reshaped to heads, and $\epsilon = 10^{-8}$ prevents division by zero.

5. **Combine**: $\mathbf{r}^{(h)} = \mathbf{a}^{(h)} + \mathbf{c}^{(h)}$.

6. **Anti-Hebbian decorrelation**: $\mathbf{r}^{(h)} \leftarrow \mathbf{r}^{(h)} - \mathbf{W}_{\text{lateral}}^{(h)} \cdot \mathbf{r}^{(h)}$. This subtraction of the lateral projection enforces that different heads encode distinct features, preventing representational collapse.

7. **k-WTA competitive coding**: For each head, retain only the top $k = \lfloor d_h \times \rho_{\text{kWTA}} \rfloor$ activations (where $\rho_{\text{kWTA}} = 0.15$ in the preferred embodiment), setting all others to zero. This enforces sparse distributed coding.

8. **Output**: Reshape heads back to $\mathbb{R}^{B \times T \times d_{\text{col}}}$, apply $\mathbf{W}_{\text{out}}$, and gate: $\mathbf{y}_{\text{L23}} = \text{sigmoid}(g_{\text{L23}}) \cdot \mathbf{W}_{\text{out}}(\mathbf{r})$.

Settling Dynamics (FIG. P1-24). The L2/3 computation may be iterated for n_{settle} steps (1 in the preferred embodiment), with re-anchoring to the input at each step:

$$\mathbf{h}_{\text{L23}}^{(k+1)} = \mathbf{h}_{\text{pre}} + \text{L23}(\mathbf{h}_{\text{L23}}^{(k)}) \tag{10}$$

where $\mathbf{h}_{\text{pre}}$ is the post-SST/VIP state. This recurrent settling allows competitive dynamics to reach equilibrium when $n_{\text{settle}} > 1$.

Complexity. The windowed context computation operates in $O(T \times w \times d_h)$ time and memory, where w is the fixed window size. Since $w \ll T$ (typically $w = 16$ while T may be 512 or 1024), this is effectively $O(T)$ — linear in sequence length, in contrast to the $O(T^2)$ complexity of self-attention.

Hebbian Learning Rules (FIG. P1-26). The association weights $\mathbf{W}_{\text{assoc}}$ and lateral weights $\mathbf{W}_{\text{lateral}}$ are updated through local Hebbian learning rules (not through backpropagation). Three update modes are provided:

(i) **Basic Hebbian**: $\Delta \mathbf{W}_{\text{assoc}}^{(h)} = \eta \cdot (\mathbf{X}^\top \mathbf{X})/N$ where $\mathbf{X}$ is the batch of head activations and $N = B \times T$.

(ii) **BCM-modulated Hebbian**: $\Delta \mathbf{W}_{\text{assoc}}^{(h)} \propto \mathbf{y}(\mathbf{y} - \sqrt{\theta_{\text{BCM}}}) \cdot \mathbf{x}^\top$ where θ_{BCM} is the BCM sliding threshold (see Section 5.6).

(iii) **Error-modulated BCM Hebbian**: The BCM update is further modulated by prediction error magnitude, blending blind correlation learning with error-driven learning:

$$\Delta \mathbf{W}_{\text{assoc}}^{(h)} \propto \left[(1 - \lambda) + \lambda \cdot \frac{\|\epsilon\|}{\overline{\|\epsilon\|}} \right] \cdot \mathbf{y}(\mathbf{y} - \sqrt{\theta_{\text{BCM}}}) \cdot \mathbf{x}^\top \tag{11}$$

where $\lambda \in [0, 1]$ is the error-Hebbian blend weight (0.5 in the preferred embodiment).

Anti-Hebbian update for lateral weights: $\Delta \mathbf{W}_{\text{lateral}}^{(h)} = \eta \cdot ((\mathbf{Y}^\top \mathbf{Y})/N - \mathbf{I})$, which drives outputs toward orthogonality.

All Hebbian updates are performed with `torch.no_grad()` (no gradient computation), consistent with the biological principle that synaptic modification occurs locally without a global error signal.

5.2.5. Working Memory Subsystem

After L2/3, the column state passes through the working memory (WM) subsystem, which maintains persistent information at three distinct timescales. The WM state persists across sequential forward calls (like biological working memory) and resets when the batch size changes or when explicitly cleared. A detailed description of the working memory subsystem is provided in a related patent application; the present description covers its integration within the cortical column pipeline.

NMDA Attractor Slots. K competitive attractor slots (4 in the preferred embodiment) with NMDA-receptor-mediated persistent activity. Each slot has dimension $d_{\text{slot}} = 64$, a learned decay rate initialised so that $\text{sigmoid}(6.0) \approx 0.9975$ (half-life ≈ 277 tokens), and competitive write/read via softmax scoring.

Activity-Silent Trace. Tsodyks-Markram facilitation/depression dynamics with dimension $d_{\text{trace}} = 64$. Maintains a hidden memory trace through short-term synaptic weight changes, retrievable by probe stimuli even when neural activity has ceased.

Phonological Loop. A ring buffer of $K_p = 8$ slots with passive decay, active rehearsal, and temporal context tagging for serial order. Implements the articulatory rehearsal system of Baddeley's working memory model.

Each WM system is gated by its own learnable parameter initialised to -2.0 and has read weights initialised to zeros, ensuring identity-at-init(Stale from transformer era check with /nct-research) behaviour.

5.2.6. Oscillatory Dynamics Module (FIG. P1-15, P1-16)

The oscillatory dynamics module implements nested theta-gamma phase-amplitude coupling (PAC), organising information processing across temporal scales.

Oscillatory Bands. Four oscillatory bands are modelled, with learnable periods constrained to be positive via softplus:

- **Gamma** ($\sim$30–80 Hz): Implicit at 1 token per cycle — each token IS one gamma cycle. Not parameterised separately.
- **Theta** ($\sim$4–8 Hz): $P_\theta = \text{softplus}(\hat{P}_\theta)$, initialised to 8 tokens. Corresponds to working memory binding cycles.
- **Alpha** ($\sim$8–12 Hz): $P_\alpha = \text{softplus}(\hat{P}_\alpha)$, initialised to 20 tokens. Corresponds to inhibitory gating.
- **Delta** ($\sim$1–4 Hz): $P_\delta = \text{softplus}(\hat{P}_\delta)$, initialised to 64 tokens. Corresponds to cross-regional coordination.

All periods are clamped to a minimum of 1.0 to prevent degenerate phase computations.

Phase-Amplitude Coupling. The gamma amplitude modulation at position t is:

$$A_\gamma(t) = \max\left(0,\ 1 + \kappa_1 \cos(\phi_\theta(t)) + \kappa_2 \cos(\phi_\delta(t)) - \kappa_3 \frac{1 + \cos(\phi_\alpha(t))}{2}\right) \tag{12}$$

where $\phi_b(t) = 2\pi t / P_b$ for each band b, and $\kappa_1, \kappa_2, \kappa_3$ are learnable coupling strengths constrained via tanh to $[-1, 1]$. At initialisation, $\kappa_1 = 0.3$ (resting theta-gamma PAC), $\kappa_2 = \kappa_3 = 0$.

Phase-Content Interaction. A low-rank bilinear interaction between phase features and column content:

$$\mathbf{p} = [\sin(\phi_\theta), \cos(\phi_\theta), \sin(\phi_\alpha), \cos(\phi_\alpha), \sin(\phi_\delta), \cos(\phi_\delta)] \in \mathbb{R}^{T \times 6} \tag{13}$$

$$\mathbf{m}_{\text{osc}} = \mathbf{W}_{\text{osc,out}} \cdot \tanh(\mathbf{W}_{\text{content}} \cdot \mathbf{h} + \mathbf{W}_{\text{phase}} \cdot \mathbf{p}) \tag{14}$$

where $\mathbf{W}_{\text{content}} \in \mathbb{R}^{r \times d_{\text{col}}}$, $\mathbf{W}_{\text{phase}} \in \mathbb{R}^{r \times 6}$, and $\mathbf{W}_{\text{osc,out}} \in \mathbb{R}^{d_{\text{col}} \times r}$ with rank $r = 32$.

Output. $\mathbf{h} \leftarrow \mathbf{h} + \text{sigmoid}(g_{\text{osc}}) \cdot A_\gamma \cdot \mathbf{m}_{\text{osc}}$.

5.2.7. Interneuron Gating Circuits — Post-Computation Phase (FIG. P1-10)

After working memory and oscillatory processing, the PV/Chandelier interneuron circuit applies an output veto.

PV/Chandelier Output Veto. The PV (parvalbumin) interneuron circuit models the chandelier cell axon initial segment gate — a binary go/no-go decision on whether the pyramidal cell fires:

$$p_{\text{pass}} = \text{sigmoid}(\mathbf{W}_{\text{PV,score}} \cdot \text{GELU}(\mathbf{W}_{\text{PV,in}} \cdot \text{RMSNorm}(\mathbf{h}))) \tag{15}$$

where $\mathbf{W}_{\text{PV,in}} \in \mathbb{R}^{d_i \times d_{\text{col}}}$ and $\mathbf{W}_{\text{PV,score}} \in \mathbb{R}^{1 \times d_i}$ with bias initialised to 5.0, so $p_{\text{pass}} \approx 0.99$ at initialisation (nearly always passes).

A straight-through gradient estimator is used: the forward pass applies a hard binary gate ($p_{\text{pass}} > 0.5$), while the backward pass uses the soft sigmoid gradient:

$$\hat{p} = \lfloor p_{\text{pass}} > 0.5 \rfloor - \text{sg}(p_{\text{pass}}) + p_{\text{pass}} \tag{16}$$

where $\text{sg}(\cdot)$ denotes the stop-gradient operation. The column state is then gated:

$$\mathbf{h} \leftarrow \mathbf{h} \cdot ((1 - \text{sigmoid}(g_{\text{PV}})) + \text{sigmoid}(g_{\text{PV}}) \cdot \hat{p}) \tag{17}$$

This preserves identity-at-init(Stale from transformer era check with /nct-research): at initialisation, $g_{\text{PV}} = -2.0$ contributes only $\sim$12% gating influence, and the bias of 5.0 ensures the gate is nearly open.

5.2.8. Layer 5: Output and Prediction Sublayer (FIG. P1-21, P1-22)

Layer 5 (L5) is the primary output layer of the cortical column, with three distinct functions: feedforward output computation, top-down prediction generation, and per-token prediction precision estimation.

SwiGLU Feedforward Network. The core computation uses SwiGLU activation (SiLU-gated linear unit, biologically analogous to the frequency-current curve of pyramidal neurons):

$$\mathbf{f} = \text{SiLU}(\mathbf{W}_{\text{gate}} \cdot \text{RMSNorm}(\mathbf{h})) \odot (\mathbf{W}_{\text{up}} \cdot \text{RMSNorm}(\mathbf{h})) \tag{18}$$

where $\mathbf{W}_{\text{gate}}, \mathbf{W}_{\text{up}} \in \mathbb{R}^{d_{\text{ff}} \times d_{\text{col}}}$ and $d_{\text{ff}} = 4 \times d_{\text{col}}$ in the preferred embodiment.

Dendritic Gating. Two dendritic gates modulate the feedforward computation:

- **Basal dendritic gate**: Gated by bottom-up L4 input, optionally modulated by the previous column's σ^2:

$$\mathbf{f} \leftarrow \mathbf{f} \odot \text{sigmoid}(\mathbf{W}_{\text{dend}} \cdot \mathbf{h}_{\text{L4}} + \mathbf{W}_{\sigma\text{-adapt}} \cdot \sigma^2_{\text{prev}}) \tag{19}$$

Bias initialised to 3.0 ($\text{sigmoid}(3.0) \approx 0.95$), so the gate is nearly open at initialisation.

- **Apical dendritic gate**: Gated by top-down L1 context:

$$\mathbf{f} \leftarrow \mathbf{f} \odot \text{sigmoid}(\mathbf{W}_{\text{apical}} \cdot \mathbf{m}_{\text{L1}}) \tag{20}$$

Also bias-initialised to 3.0 for identity-at-init(Stale from transformer era check with /nct-research).

k-WTA Sparsity. After dendritic gating, k-WTA sparsity is applied: only the top $k = \lfloor d_{\text{ff}} \times \rho_{\text{kWTA}} \rfloor$ activations are retained, setting all others to zero. This models cortical sparse coding: approximately 15% of neurons are active at any time.

Output.

$$\mathbf{o}_i = \mathbf{W}_{\text{down}} \cdot \text{Dropout}(\mathbf{f}) \tag{21}$$

where $\mathbf{W}_{\text{down}} \in \mathbb{R}^{d_{\text{model}} \times d_{\text{ff}}}$ is initialised to zeros (identity-at-init via the residual connection). The column output is:

$$\text{output}_i = \mathbf{x}_{\text{input}} + \mathbf{o}_i \tag{22}$$

implementing a residual connection that bypasses the entire column computation.

Top-Down Prediction. A separate linear projection generates a prediction of the input to the column below:

$$\hat{\mathbf{x}}_{i \to i-1} = \mathbf{W}_{\text{predict}} \cdot \text{RMSNorm}(\mathbf{h}) \tag{23}$$

where $\mathbf{W}_{\text{predict}} \in \mathbb{R}^{d_{\text{model}} \times d_{\text{col}}}$ is initialised to zeros. At initialisation, all predictions are zero, yielding maximum prediction error and maximum learning signal — the system begins in a state of maximum uncertainty and learns to make progressively better predictions.

Prediction Precision σ^2 (FIG. P1-22). A precision head estimates the column's confidence in its prediction:

$$\sigma_i^2 = \text{softplus}(\mathbf{W}_\sigma \cdot \text{RMSNorm}(\mathbf{h})) + \epsilon \tag{24}$$

where $\mathbf{W}_\sigma \in \mathbb{R}^{1 \times d_{\text{col}}}$ has weight initialised with small random values (std $= 0.01$) to enable content-dependent precision from step 0, and bias initialised so that $\text{softplus}(\text{bias}) = \sigma^2_{\text{init,col}}$ where $\sigma^2_{\text{init,col}}$ is depth-scaled:

$$\sigma^2_{\text{init,col}} = \sigma^2_{\text{init}} \cdot \left(1 + s_{\text{depth}} \cdot \frac{i}{N-1}\right) \tag{25}$$

Deeper columns start with higher uncertainty (more uncertain predictions), modelling the developmental gradient observed by Huttenlocher (1990) in which deeper cortical areas mature

later.

The bare softplus activation (no hard clamps, no soft clamps) guarantees $\sigma^2 > 0$ while preserving gradient flow. $\epsilon = 10^{-6}$ prevents exact zero.

σ^2 Temporal Smoothing (FIG. P1-25). The raw σ^2 is further temporally smoothed via an exponential moving average (EMA) whose time constant depends on column depth:

$$\alpha_i = \frac{1}{1 + s_{\text{EMA}} \cdot i} \tag{26}$$

Shallow columns ($i = 0$): $\alpha \approx 1.0$ (instantaneous tracking). Deep columns ($i = N{-}1$): $\alpha \approx 0.06$ (integrates over $\sim$16 tokens). This models the temporal receptive field hierarchy observed by Hasson et al. (2008), where early sensory cortex responds at sub-second timescales and prefrontal cortex integrates over seconds.

5.2.9. Layer 6: Thalamic Feedback Sublayer

The Layer 6 (L6) thalamic feedback sublayer generates a corticothalamic feedback signal:

$$\mathbf{f}_{\text{L6}} = \text{sigmoid}(g_{\text{L6}}) \cdot \mathbf{W}_{\text{L6}} \cdot \mathbf{h} \tag{27}$$

where $\mathbf{W}_{\text{L6}} \in \mathbb{R}^{d_{\text{model}} \times d_{\text{col}}}$ is initialised to zeros and $g_{\text{L6}} = -2.0$. This feedback modulates the thalamic relay for the column's next input processing cycle (see Section 5.3).

5.3. Thalamic Routing System (FIG. 4, FIG. P1-11, P1-12)

The thalamic routing system comprises G relay groups (4 in the preferred embodiment) and a thalamic reticular nucleus (TRN). Columns are assigned to relay groups: columns $[j \cdot C_r, (j + 1) \cdot C_r)$ share relay group j, where $C_r = N/G$ is the columns-per-relay ratio.

5.3.1. Thalamic Relay

Each relay group contains a single relay unit that gates and routes information:

$$\delta_{\text{relay}} = \mathbf{W}_{\text{relay,out}} \cdot \text{GELU}(\mathbf{W}_{\text{relay,in}} \cdot \text{RMSNorm}(\mathbf{x})) \tag{28}$$

where $\mathbf{W}_{\text{relay,in}} \in \mathbb{R}^{d_{\text{relay}} \times d_{\text{model}}}$ and $\mathbf{W}_{\text{relay,out}} \in \mathbb{R}^{d_{\text{model}} \times d_{\text{relay}}}$ is initialised to zeros. When L6 feedback $\mathbf{f}_{\text{L6}}$ is available, it modulates the relay: $\delta_{\text{relay}} \leftarrow \delta_{\text{relay}} + 0.1 \cdot \mathbf{f}_{\text{L6}}$.

5.3.2. Thalamic Reticular Nucleus (TRN)

The TRN computes a per-token relevance gate between pre-relay and post-relay signals:

$$r = \text{sigmoid}\left(\frac{(\mathbf{W}_q \cdot \mathbf{x}) \odot (\mathbf{W}_k \cdot \delta_{\text{relay}})}{d_{\text{TRN}}^{1/2}}\right) \in [0, 1] \tag{29}$$

where $\mathbf{W}_q, \mathbf{W}_k \in \mathbb{R}^{d_{\text{TRN}} \times d_{\text{model}}}$ are learned projections and $d_{\text{TRN}} = 24$ in the preferred embodiment.

The TRN gate is blended with a bypass via the TRN gate parameter:

$$r_{\text{eff}} = (1 - \text{sigmoid}(g_{\text{TRN}})) \cdot 1 + \text{sigmoid}(g_{\text{TRN}}) \cdot r \tag{30}$$

At initialisation ($g_{\text{TRN}} = -2.0$), the effective gate is $\approx 0.88 + 0.12r \approx 1$, passing nearly everything.

5.3.3. Bimodality Loss

To encourage the TRN to make binary gating decisions (fully pass or fully block), a bimodality loss is accumulated:

$$\mathcal{L}_{\text{TRN}} = \frac{1}{N} \sum_i \text{mean}(r_i \cdot (1 - r_i)) \tag{31}$$

This loss is minimised when $r \to 0$ or $r \to 1$, encouraging sharp gating decisions.

5.3.4. Routed Output

The final thalamic output for column i is:

$$\mathbf{x}_{\text{thal}} = \mathbf{x} + \text{sigmoid}(g_{\text{relay}}) \cdot r_{\text{eff}} \cdot \delta_{\text{relay}} \tag{32}$$

5.4. Lateral Column Connections (FIG. P1-13)

The lateral column connection system implements horizontal communication between adjacent columns via low-rank message passing.

Each column i has a send projection $\mathbf{W}_{\text{send},i} \in \mathbb{R}^{r_{\text{lat}} \times d_{\text{model}}}$ and a receive projection $\mathbf{W}_{\text{recv},i} \in \mathbb{R}^{d_{\text{model}} \times r_{\text{lat}}}$, where $r_{\text{lat}} = \max(d_{\text{model}}/8, 32)$ is a low-rank bottleneck.

Column i receives the average of messages from its immediate neighbours ($i-1$ and $i+1$):

$$\mathbf{l}_i = \text{sigmoid}(g_{\text{lat},i}) \cdot \mathbf{W}_{\text{recv},i} \cdot \frac{1}{|\mathcal{N}_i|} \sum_{j \in \mathcal{N}_i} \mathbf{W}_{\text{send},j} \cdot \text{output}_j \tag{33}$$

where $\mathcal{N}_i = \{i-1, i+1\} \cap \{0, \ldots, N-1\}$ is the set of neighbours. Edge columns have a single neighbour.

Send projections are initialised with small random weights (std $= 0.01$) to provide a learning signal for analytical updates. Gate parameters are initialised to -2.0.

5.5. Global Workspace Broadcast (FIG. P1-14)

The global workspace module implements the Global Workspace Theory (Baars 1988) and Global Neuronal Workspace (Dehaene & Naccache 2001), enabling columns to communicate beyond local lateral connections.

5.5.1. Write Projections

Each column i writes a candidate broadcast via $\mathbf{W}_{\text{write},i} \in \mathbb{R}^{d_{\text{ws}} \times d_{\text{model}}}$:

$$\mathbf{b}_i = \mathbf{W}_{\text{write},i} \cdot \text{output}_i \tag{34}$$

where $d_{\text{ws}} = \max(d_{\text{model}}/4, 64)$ is the workspace bottleneck dimension.

5.5.2. Surprise-Weighted Competition

Columns compete for broadcast access based on their prediction surprise:

$$s_i = \text{mean}(\epsilon_i^2), \quad \omega = \text{softmax}\left(\frac{[s_0, \ldots, s_{N-1}]}{\text{softplus}(\tau)}\right) \tag{35}$$

where ϵ_i is the prediction error for column i and τ is a learnable temperature parameter (initialised to 1.0). High-surprise columns — those whose predictions are most wrong — win broadcast access. This implements the "ignition" mechanism of Global Workspace Theory.

5.5.3. Broadcast and Read

The workspace broadcast is the surprise-weighted combination:

$$\mathbf{w} = \sum_{i=0}^{N-1} \omega_i \cdot \mathbf{b}_i \tag{36}$$

Each column reads the broadcast via its own read projection:

$$\mathbf{m}_{\text{ws},i} = \text{sigmoid}(g_{\text{ws}}) \cdot \mathbf{W}_{\text{read},i} \cdot \mathbf{w} \tag{37}$$

where $\mathbf{W}_{\text{read},i} \in \mathbb{R}^{d_{\text{model}} \times d_{\text{ws}}}$ and $g_{\text{ws}} = -2.0$ at initialisation.

5.6. BCM Homeostasis (FIG. P1-17)

Each cortical column contains a BCM (Bienenstock-Cooper-Munro) homeostatic plasticity module that prevents runaway Hebbian potentiation.

The BCM sliding threshold tracks mean squared postsynaptic activity via EMA:

$$\theta_{\text{BCM}} \leftarrow (1 - \alpha_{\text{BCM}}) \cdot \theta_{\text{BCM}} + \alpha_{\text{BCM}} \cdot \overline{y^2} \tag{38}$$

where $\alpha_{\text{BCM}} = 1/\tau_{\text{BCM}}$ with $\tau_{\text{BCM}} = 1000$ in the preferred embodiment.

The BCM rule modulates Hebbian learning: synapses driving activity above θ_{BCM} are potentiated; those driving activity below θ_{BCM} are depressed. This maintains selectivity and prevents the seizure-like instability that would result from unchecked positive-feedback Hebbian learning.

5.7. Metaplasticity (FIG. P1-18)

Each column also contains a metaplasticity module that adjusts learning rates based on recent plasticity history:

$$\text{plasticity_ema} \leftarrow (1 - 1/\tau_{\text{meta}}) \cdot \text{plasticity_ema} + (1/\tau_{\text{meta}}) \cdot |\Delta W| \tag{39}$$

$$\text{lr_mult} = \frac{1}{1 + \text{plasticity_ema}} \tag{40}$$

where $\tau_{\text{meta}} = 500$. High recent plasticity (large recent weight changes) results in a lower learning rate multiplier, preventing instability after rapid learning. Low recent plasticity results in a higher multiplier, promoting exploration when learning has stagnated. This models the metaplastic threshold shift mechanism described by Abraham & Bear (1996).

5.8. Structural Plasticity (FIG. P1-19)

The L2/3 association layer supports structural plasticity: the ability to prune weak connections and grow new ones based on prediction error.

5.8.1. Connection Mask

A binary mask $\mathbf{M}^{(h)} \in \{0, 1\}^{d_h \times d_h}$ tracks which connections in $\mathbf{W}^{(h)}_{\text{assoc}}$ are active. Initialised to all ones (fully connected).

5.8.2. Pruning

A counter $\mathbf{C}^{(h)}$ tracks the number of consecutive steps each connection has been below the pruning threshold $\theta_{\text{prune}} = 0.01$:

$$C^{(h)}_{jk} \leftarrow \begin{cases} C^{(h)}_{jk} + 1 & \text{if } |W^{(h)}_{jk}| < \theta_{\text{prune}} \text{ and } M^{(h)}_{jk} = 1 \\ 0 & \text{otherwise} \end{cases} \tag{41}$$

When $C^{(h)}_{jk} \geq T_{\text{prune}}$ ($T_{\text{prune}} = 1000$ steps), the connection is permanently pruned: $M^{(h)}_{jk} \leftarrow 0$, $W^{(h)}_{jk} \leftarrow 0$.

5.8.3. *Synaptogenesis*

When the mean prediction error exceeds $\theta_{\text{grow}} = 2.0$, a fraction (5%) of pruned connections are restored with small random weights (std $= 0.01$). This models the brain's ability to grow new synapses in response to persistent prediction failure.

5.9. Metacognitive Self-Assessment (FIG. P1-20)

The self-assessment module implements ACC-like (anterior cingulate cortex) conflict monitoring with zero learnable parameters.

5.9.1. *Conflict Detection*

Inter-column conflict is computed as the variance of column output means:

$$\text{conflict} = \text{Var}\left(\left\{\overline{\text{output}_i}\right\}_{i=0}^{N-1}\right) \tag{42}$$

where $\overline{\text{output}_i}$ is the mean of column i's output across batch and sequence dimensions.

5.9.2. *Adaptive Computation*

When conflict exceeds a threshold ($\theta_{\text{conflict}} = 0.5$), additional predictive coding inference iterations are allocated:

$$n_{\text{extra}} = \min\left(n_{\text{max_extra}}, \left\lfloor \frac{\text{conflict} - \theta_{\text{conflict}}}{\theta_{\text{conflict}}} \cdot n_{\text{max_extra}} \right\rfloor\right) \tag{43}$$

where $n_{\text{max_extra}} = 2$. This maps the brain's mechanism of "thinking harder" on difficult inputs: high inter-column disagreement triggers additional settling iterations to resolve the conflict.

5.9.3. *Health Monitoring*

The module also tracks per-column health via EMA of σ^2 and PC convergence rate:

- σ^2 divergence (> 100): triggers emergency down-scaling of L5 output weights.
- σ^2 collapse (< 0.001): nudges $\mathbf{W}_\sigma$ bias upward.
- PC non-convergence (consistently hitting pc_max_steps): flagged for monitoring.

5.10. Identity-at-Initialisation Design Principle (FIG. P1-4)

A critical design principle of the present invention is that every cortical column processing unit produces near-identity output at initialisation. This is achieved through:

(i) All gate parameters initialised to $g = -2.0$, yielding $\text{sigmoid}(g) \approx 0.12$;

(ii) All output weight matrices ($\mathbf{W}_{\text{out}}$, $\mathbf{W}_{\text{down}}$, $\mathbf{W}_{\text{predict}}$, L1 projection, L6 projection) initialised to zeros;

(iii) Residual connections bypassing each column: output $=$ input $+$ column_delta;

(iv) PV/Chandelier veto bias initialised to 5.0 (sigmoid$(5.0) \approx 0.99$, nearly always passes);

(v) Dendritic gate biases initialised to 3.0 (sigmoid$(3.0) \approx 0.95$, nearly open).

This ensures that a freshly constructed network with N columns produces output $\approx$ input (through the chain of residual connections), enabling stable training from random initialisation without warmup or careful learning rate tuning. All functional behaviour emerges through learning.

5.11. Forward Pass Through the Complete System (FIG. P1-2, P1-23)

The complete forward pass proceeds as follows:

1. **Embedding**: Input token IDs $\in \mathbb{N}^{B\times T}$ are embedded into $\mathbb{R}^{B\times T\times d_{\text{model}}}$ via a learned embedding table. Grid-cell position encoding is added with a small scaling factor (0.1).

2. **Bottom-up pass**: The embedded representation is processed sequentially through columns 0 to $N-1$:

   (a) Route through the thalamic relay for column i (with L6 feedback if available).

   (b) Process through the six-layer pipeline of column i.

   (c) Pass the column output as feedforward input to column $i+1$.

   (d) Pass the column's σ^2 as input to column $i+1$'s L5 dendritic gate.

3. **Prediction error computation**: For each column i, compute:

$$\epsilon_i = \text{target}_i - \text{prediction}_i \tag{44}$$

   where $\text{target}_0 = \text{embedding}$ and $\text{target}_i = \text{output}_{i-1}$ for $i > 0$.

4. **Lateral communication**: Compute lateral signals between adjacent columns (Section 5.4).

5. **Global workspace broadcast**: Compute surprise-weighted broadcast (Section 5.5).

6. **Metacognitive assessment**: Compute inter-column conflict and determine additional PC iterations (Section 5.9).

7. **Predictive coding inference**: Under `torch.no_grad()`, iterate up to pc_max_steps $+$ n_{extra} times:

(a) Process all columns with top-down context from the workspace + lateral + higher column predictions.

(b) Recompute predictions and errors.

(c) Recompute lateral and workspace communications.

(d) Check for convergence: stop early if relative error change $< \theta_{\text{conv}}$.

8. **Output**: Apply RMSNorm to the final column output and produce logits via the language model head: logits $=$ LMHead(RMSNorm(output$_{N-1}$)).

5.12. Disease-Mapping Validation Methodology (FIG. P1-27)

A distinguishing feature of the present invention is its use of neuroscience-grounded disease mapping as a systematic validation methodology. Because each architectural component models a specific brain mechanism, removing that component should produce a specific, predictable failure mode corresponding to a known neurological or psychiatric disease. This provides both:

- A **design tool**: For every proposed component, the question "what disease would result if this were removed?" must have a specific answer. If no disease analogue exists, the component may be unnecessary.
- A **validation tool**: If removing a component produces the *expected* failure mode, the component is correctly implemented. If it produces a *different* failure mode, there is a bug.

Table **??** in FIG. P1-27 provides the complete disease-mapping reference. Selected entries include:

Component Removed	Disease Analogue	Expected Failure Mode
Global workspace	Vegetative state	No global broadcast; unconscious-only processing
VIP interneurons	Autism (ASD)	No disinhibitory gate; failure to integrate context
SST interneurons	Hallucination-like	Unregulated top-down; false perceptions
PV/Chandelier	Epilepsy	No output veto; uncontrolled cortical output
NMDA WM slots	Schizophrenia	NMDA hypofunction; thought fragmentation
Thalamic relay	Coma	No routing; cortex isolated from input
TRN gating	Fatal familial insomnia	No relevance filtering; unfiltered input
L1 top-down	Broca's aphasia	No re-entrant processing
L2/3 lateral weights	Cortical dedifferentiation	Representational collapse (all heads same)
BCM homeostasis	Epilepsy (E/I imbalance)	Runaway Hebbian potentiation
Metaplasticity	Fragile X syndrome	Permanently shifted plasticity threshold
Oscillations	Schizophrenia (gamma)	Disrupted temporal binding
Self-assessment	Anosognosia	No self-awareness of processing deficits

This methodology extends beyond simple ablation studies: it provides a *theory-driven* prediction of failure modes that can be empirically verified, establishing a rigorous link between the architecture and its biological foundations.

5.13. Implementation Considerations

5.13.1. Mixed Precision Training

The system supports bfloat16 mixed precision (bf16) for reduced memory consumption. All Hebbian and predictive coding updates operate on `.data` attributes under `torch.no_grad()`, eliminating the unused gradient graph and saving approximately 3 GiB VRAM.

5.13.2. Custom CUDA Kernels

The preferred embodiment includes fused CUDA kernels for the windowed Hebbian context computation, Hebbian weight update, NMDA parallel scan, silent trace scan, and oscillatory PAC modulation, providing approximately 2–10$\times$ speedup over pure PyTorch implementations.

5.13.3. Torch Compilation

The architecture is compatible with `torch.compile` (`mode='reduce-overhead'`), with specific methods decorated with `@torch.compiler.disable` where dynamic Python control flow (variable-length lists, ring buffer pointer arithmetic) is incompatible with graph tracing.

5.13.4. Scalability (FIG. P1-28)

The architecture scales from 106M parameters (default preset) through 300M and 1B parameters by proportionally increasing d_{model}, N, and d_{ff} while preserving all biological mechanisms. All components scale with $O(d_{\text{model}}^2)$ parameters (same scaling as transformer blocks), with the additional advantage of $O(T \times w)$ rather than $O(T^2)$ sequence-length complexity.

CLAIMS

We claim:

Claim 1. A computer-implemented neural network system for natural language processing, comprising:

(a) a plurality of N cortical column processing units arranged in a hierarchical stack from a lowest column to a highest column, wherein each cortical column processing unit comprises:

(i) a Layer 4 thalamic input sublayer configured to receive feedforward input from a thalamic routing system and to project said input into an intra-column processing space via a learned linear projection followed by root mean square normalisation;

(ii) a Layer 1 top-down modulation sublayer configured to receive a top-down context signal from a higher column in the hierarchy and to produce a gated modulatory signal via a learned linear projection with a learnable scalar gate parameter;

(iii) a pre-association interneuron gating circuit comprising an SST dendritic inhibition unit configured to gate apical dendritic input via tonic inhibition modulated by a VIP disinhibitory unit, wherein said VIP disinhibitory unit is configured to detect novelty in the top-down context signal and to release said SST suppression;

(iv) a Layer 2/3 windowed Hebbian association sublayer configured to perform content-addressable recall within a local causal window of w tokens using a plurality of H association heads, each comprising a learned association matrix $\mathbf{W}_{\text{assoc}}$, a learned distance weighting vector, an anti-Hebbian lateral decorrelation matrix $\mathbf{W}_{\text{lateral}}$, and a k-winner-take-all competitive coding mechanism, wherein said sublayer operates with $O(T \times w)$ time complexity where T is the sequence length and w is the window size;

(v) a working memory subsystem comprising at least one of: NMDA attractor slots with persistent activity and learned decay rates, an activity-silent trace mechanism with facilitation and depression dynamics, and a phonological loop with decay, rehearsal, and temporal context tagging;

(vi) an oscillatory dynamics module configured to apply nested theta-gamma phase-amplitude coupling with learnable oscillatory periods constrained to positive values via a softplus activation function, and a phase-content interaction computed via low-rank bilinear coupling;

(vii) a post-computation interneuron gating circuit comprising a PV/Chandelier output veto unit configured to apply a binary go/no-go gate at the axon initial segment using a straight-through gradient estimator;

(viii) a Layer 5 output and prediction sublayer configured to compute a feedforward output via a SwiGLU activation function with dendritic gating, to generate a top-down prediction for the column below via a learned linear projection, and to estimate a per-token prediction precision σ^2 via a softplus activation function producing a strictly positive value; and

(ix) a Layer 6 thalamic feedback sublayer configured to generate a corticothalamic feedback signal for modulating the thalamic routing system;

(b) a thalamic routing system comprising a plurality of G relay groups and a thalamic reticular nucleus configured to compute a per-token relevance gate between pre-relay and post-relay signals, wherein said thalamic reticular nucleus is trained with a bimodality loss $r(1-r)$ to encourage binary gating decisions, and wherein said relay groups are further modulated by Layer 6 thalamic feedback signals from the cortical columns;

(c) a lateral column connection system comprising, for each cortical column, a low-rank send projection and a low-rank receive projection configured to exchange messages with immediately adjacent columns in the hierarchy;

(d) a global workspace module configured to receive outputs from all cortical columns, to compute surprise-weighted competition among said columns based on their prediction error magnitudes, and to broadcast a weighted combination of column representations to all columns via per-column read projections; and

(e) a language model head configured to convert the output of the highest cortical column into vocabulary logit scores.

Claim 2. A computer-implemented method for processing natural language input, comprising:

(a) embedding a sequence of input tokens into continuous vector representations and adding a position encoding thereto;

(b) routing said vector representations through a thalamic routing system comprising relay groups gated by a thalamic reticular nucleus with bimodality-encouraging loss;

(c) processing said routed representations sequentially through a hierarchical stack of N cortical column processing units, wherein each cortical column processing unit applies a six-layer pipeline comprising: a Layer 4 thalamic input projection, a Layer 1 top-down modulation, SST and VIP interneuron gating before lateral association, a Layer 2/3 windowed Hebbian association operating with $O(T \times w)$ time complexity within a local causal window of w tokens using learned association weights and k-winner-take-all competitive coding, working memory processing at multiple timescales, oscillatory modulation via nested theta-gamma phase-amplitude coupling, PV/Chandelier output veto after oscillatory processing,

a Layer 5 output computation with SwiGLU activation and dendritic gating that generates a top-down prediction and per-token prediction precision estimate, and a Layer 6 thalamic feedback signal;

(d) computing top-down prediction errors between each column's prediction and the actual input of the column below;

(e) computing lateral signals between adjacent columns via low-rank message passing;

(f) computing a global workspace broadcast via surprise-weighted softmax competition among columns;

(g) computing a metacognitive conflict score from inter-column output variance and allocating additional inference iterations when said conflict exceeds a threshold;

(h) iteratively settling column representations through predictive coding inference iterations under adaptive convergence control until prediction errors converge below a threshold or a maximum number of iterations is reached; and

(i) producing output logits from the highest column output through a language model head.

Claim 3. A non-transitory computer-readable storage medium storing computer-executable instructions that, when executed by one or more processors, cause the one or more processors to implement a neural network system for natural language processing, said system comprising a plurality of N cortical column processing units arranged in a hierarchical stack, a thalamic routing system with a reticular nucleus, a lateral column connection system, and a global workspace module, wherein each cortical column processing unit implements a six-layer pipeline comprising a Layer 4 thalamic input sublayer, a Layer 1 top-down modulation sublayer, pre-association SST and VIP interneuron gating, a Layer 2/3 windowed Hebbian association sublayer operating with $O(T \times w)$ time complexity, a working memory subsystem, an oscillatory dynamics module with nested theta-gamma phase-amplitude coupling, a post-computation PV/Chandelier output veto, a Layer 5 output and prediction sublayer computing per-token prediction precision via softplus activation, and a Layer 6 thalamic feedback sublayer.

Claim 4. The system of Claim 1, wherein all gate parameters in the cortical column processing units are initialised to a value of approximately -2.0 such that $\text{sigmoid}(-2.0) \approx 0.12$, all output weight matrices are initialised to zeros, and residual connections bypass each cortical column processing unit, so that a freshly initialised network produces output approximately equal to its input (identity-at-initialisation principle).

Claim 5. The system of Claim 1, wherein the Layer 2/3 windowed Hebbian association sublayer performs content-addressable recall by computing, for each position t and each head h, a direct association $\mathbf{a}_t^{(h)} = \mathbf{W}_{\text{assoc}}^{(h)} \cdot \mathbf{x}_t^{(h)}$ where $\mathbf{W}_{\text{assoc}}^{(h)}$ is a per-head association matrix initialised as an identity matrix and updated through Hebbian learning without backpropagation.

Claim 6. The system of Claim 5, wherein the windowed context aggregation computes, for each position t, a distance-weighted combination of value vectors within the causal window $[t-w, t]$ using learned distance weights $\mathbf{W}_{\text{dist}}^{(h)}$, wherein attention-like scores are computed as $\text{softplus}((\mathbf{x}_t - \mathbf{x}_j) \cdot \mathbf{W}_{\text{dist}}^{(h)})$ and normalised by division, rather than through a dot-product query-key mechanism followed by softmax.

Claim 7. The system of Claim 5, wherein the anti-Hebbian lateral decorrelation matrix $\mathbf{W}_{\text{lateral}}^{(h)}$ is applied by subtracting $\mathbf{W}_{\text{lateral}}^{(h)} \cdot \mathbf{r}^{(h)}$ from the combined association-plus-context representation, and is updated via an anti-Hebbian rule $\Delta\mathbf{W}_{\text{lateral}}^{(h)} = \eta \cdot ((\mathbf{Y}^\top \mathbf{Y})/N - \mathbf{I})$ that drives head outputs toward orthogonality, preventing representational collapse.

Claim 8. The system of Claim 5, wherein the k-winner-take-all competitive coding mechanism retains only the top $k = \lfloor d_h \times \rho \rfloor$ activations in each head (where ρ is a sparsity ratio, preferably 0.15), setting all other activations to zero, thereby enforcing sparse distributed coding analogous to cortical population sparseness.

Claim 9. The system of Claim 1, wherein the VIP disinhibitory unit computes a disinhibition signal $\mathbf{v}_{\text{VIP}} = \text{sigmoid}(g_{\text{VIP}}) \cdot \text{sigmoid}(\mathbf{W}_{\text{VIP,out}} \cdot \text{GELU}(\mathbf{W}_{\text{VIP,in}} \cdot \text{RMSNorm}(\mathbf{c}_{\text{td}})))$ from the top-down context signal, and the SST dendritic inhibition unit computes effective inhibition as $\mathbf{s}_{\text{eff}} = \mathbf{s}_{\text{SST}} \cdot (1 - \mathbf{v}_{\text{VIP}})$, clamped to a maximum of 0.9 to maintain a minimum 10% signal pass-through, and applies said inhibition multiplicatively to the column state before the Layer 2/3 association step.

Claim 10. The system of Claim 1, wherein the PV/Chandelier output veto unit computes a pass probability $p_{\text{pass}} = \text{sigmoid}(\mathbf{W}_{\text{PV,score}} \cdot \text{GELU}(\mathbf{W}_{\text{PV,in}} \cdot \text{RMSNorm}(\mathbf{h})))$ with a bias initialised to 5.0 such that $p_{\text{pass}} \approx 0.99$ at initialisation, and applies a hard binary gate in the forward pass with a straight-through gradient estimator in the backward pass, wherein the gate is blended with a bypass via a learnable gate parameter.

Claim 11. The system of Claim 1, wherein the thalamic reticular nucleus computes a per-token relevance gate $r = \text{sigmoid}((\mathbf{W}_q \cdot \mathbf{x}) \odot (\mathbf{W}_k \cdot \delta_{\text{relay}})/\sqrt{d_{\text{TRN}}})$, blended with a bypass via a learnable gate parameter so that at initialisation the effective gate passes substantially all information, and wherein the bimodality loss $\mathcal{L}_{\text{TRN}} = \text{mean}(r \cdot (1-r))$ is minimised during training to encourage binary gating decisions.

Claim 12. The system of Claim 1, wherein the thalamic relay groups are modulated by Layer 6 thalamic feedback signals from the cortical columns, such that the output of each relay group is $\delta_{\text{relay}} + 0.1 \cdot \mathbf{f}_{\text{L6}}$, forming a cortex-thalamus feedback loop.

Claim 13. The system of Claim 1, wherein each lateral column connection comprises a low-rank send projection $\mathbf{W}_{\text{send}} \in \mathbb{R}^{r \times d_{\text{model}}}$ and a low-rank receive projection $\mathbf{W}_{\text{recv}} \in \mathbb{R}^{d_{\text{model}} \times r}$ with $r = \max(d_{\text{model}}/8, 32)$, and wherein each column receives the average of messages from its immediately adjacent neighbours gated by a per-column learnable gate parameter.

Claim 14. The system of Claim 1, wherein the global workspace module computes surprise for each column as the mean squared prediction error, applies a softmax with a learnable temperature parameter to determine broadcast weights, combines column candidates via said broadcast weights, and distributes the combined broadcast to all columns via per-column read projections gated by a global gate parameter.

Claim 15. The system of Claim 1, wherein the oscillatory dynamics module computes a gamma amplitude modulation at each position t as:

$$A_\gamma(t) = \max(0,\ 1 + \kappa_1 \cos(\phi_\theta(t)) + \kappa_2 \cos(\phi_\delta(t)) - \kappa_3(1 + \cos(\phi_\alpha(t)))/2)$$

where $\phi_b(t) = 2\pi t / P_b$ for oscillatory band b, periods P_b are learnable via softplus to ensure positivity, and coupling strengths $\kappa_1, \kappa_2, \kappa_3$ are learnable via tanh to constrain them to $[-1, 1]$.

Claim 16. The system of Claim 16, wherein the oscillatory dynamics module further computes a phase-content interaction via a low-rank bilinear coupling: $\mathbf{m}_{\text{osc}} = \mathbf{W}_{\text{out}} \cdot \tanh(\mathbf{W}_{\text{content}} \cdot \mathbf{h} + \mathbf{W}_{\text{phase}} \cdot \mathbf{p})$, where $\mathbf{p}$ is a concatenation of sine and cosine of the theta, alpha, and delta phase angles, and the oscillatory modulation is applied as $\mathbf{h} + \text{sigmoid}(g_{\text{osc}}) \cdot A_\gamma \cdot \mathbf{m}_{\text{osc}}$.

Claim 17. The system of Claim 1, further comprising a BCM homeostatic plasticity module within each cortical column, said module maintaining a sliding threshold θ_{BCM} updated via exponential moving average of mean squared postsynaptic activity, wherein said threshold modulates Hebbian learning such that synapses driving activity above θ_{BCM} are potentiated and synapses driving activity below θ_{BCM} are depressed.

Claim 18. The system of Claim 1, further comprising a metaplasticity module within each cortical column, said module tracking recent plasticity magnitude via exponential moving average and computing a learning rate multiplier as $1/(1+\text{plasticity_ema})$, such that high recent plasticity yields a lower learning rate and low recent plasticity yields a higher learning rate.

Claim 19. The system of Claim 1, further comprising a structural plasticity mechanism in the Layer 2/3 windowed Hebbian association sublayer, said mechanism maintaining a binary connection mask and a below-threshold counter for each connection in the association matrix, pruning connections that remain below a weight magnitude threshold for a predetermined number of consecutive steps, and growing new connections with small random weights when mean prediction error exceeds a synaptogenesis threshold.

Claim 20. The system of Claim 1, further comprising a metacognitive self-assessment module with zero learnable parameters, configured to compute inter-column conflict as the variance of column output means across the plurality of cortical columns, and to allocate additional predictive coding inference iterations when said conflict exceeds a predetermined threshold.

Claim 21. The system of Claim 1, wherein the Layer 5 output and prediction sublayer comprises a basal dendritic gate modulated by the Layer 4 thalamic input and optionally by the previous

column's prediction precision σ^2_{prev}, and an apical dendritic gate modulated by the Layer 1 top-down context signal, wherein both dendritic gates are implemented as sigmoid-activated learned linear projections with biases initialised to 3.0 such that $\text{sigmoid}(3.0) \approx 0.95$ and the gates are nearly open at initialisation.

Claim 22. The system of Claim 1, wherein the per-token prediction precision σ^2 is computed as $\sigma^2 = \text{softplus}(\mathbf{W}_\sigma \cdot \text{RMSNorm}(\mathbf{h})) + \epsilon$ with depth-dependent initialisation such that deeper columns in the hierarchy have higher initial precision values, modelling the developmental gradient of cortical maturation, and wherein said precision is temporally smoothed via an exponential moving average with a depth-dependent time constant.

Claim 23. The method of Claim 2, wherein the predictive coding inference iterations of step (h) are performed without gradient computation, and wherein convergence is determined by monitoring the relative change in total prediction error across all columns, with early termination when said relative change falls below a convergence threshold after a minimum number of iterations.

Claim 24. The system of Claim 1, wherein the Layer 2/3 windowed Hebbian association sublayer supports recurrent settling dynamics by iterating the association computation for n_{settle} steps with re-anchoring to the pre-L2/3 input at each step.

Claim 25. The system of Claim 23, wherein the depth-dependent time constant for σ^2 temporal smoothing is computed as $\alpha_i = 1/(1 + s \cdot i)$ where i is the column index and s is a configurable scale factor, such that the shallowest column tracks σ^2 instantaneously and the deepest column integrates σ^2 over approximately 16 tokens.

Claim 26. The system of Claim 5, wherein the association matrix $\mathbf{W}_{\text{assoc}}$ is updated through one of: a basic Hebbian rule computing $\Delta\mathbf{W}_{\text{assoc}} = \eta \cdot (\mathbf{X}^\top\mathbf{X})/N$, a BCM-modulated Hebbian rule computing $\Delta\mathbf{W}_{\text{assoc}} \propto \mathbf{y}(\mathbf{y} - \sqrt{\theta_{\text{BCM}}}) \cdot \mathbf{x}^\top$, or an error-modulated BCM Hebbian rule that further modulates said update by prediction error magnitude with a configurable blend weight λ.

Claim 27. The system of Claim 1, wherein the system is designed according to a disease-mapping validation methodology in which each architectural component models a specific brain mechanism, and removal of said component produces a specific, predictable failure mode corresponding to a known neurological or psychiatric disease, wherein said correspondence serves as a systematic design and validation tool for the architecture.

Claim 28. The system of Claim 1, wherein the system is configurable at a plurality of parameter scales by proportionally adjusting the model dimension d_{model}, the number of columns N, and the feedforward dimension d_{ff} while preserving all biological mechanisms, including configurations of approximately 106 million parameters, 300 million parameters, and 1 billion parameters.

ABSTRACT

A computer-implemented neural network system and method for natural language processing comprising a plurality of cortical column processing units arranged in a hierarchical stack, wherein each cortical column processing unit implements a biologically-grounded six-layer pipeline modelling the laminar structure of the mammalian neocortex. Each cortical column comprises a thalamic input layer, a top-down modulation layer, interneuron gating circuits, a windowed Hebbian association layer operating in $O(T \times w)$ time complexity, working memory subsystems, oscillatory dynamics, and an output prediction layer generating per-token precision estimates. The system further comprises a thalamic routing system with a reticular nucleus for relevance-based gating, lateral column connections for horizontal communication, and a global workspace for surprise-weighted broadcast competition. The architecture replaces quadratic-complexity self-attention mechanisms with linear-complexity local association, supports local learning without global backpropagation, and provides built-in uncertainty quantification through precision-weighted predictive coding.