Brain-Inspired Language Production System with Multi-Stage Evidence Accumulation and Self-Monitoring

Open research specification 6 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: Standard Neural Language Generation 3
2.2 Prior Art: Drift-Diffusion Models 3
2.3 Prior Art: Levelt Speech Production Model 4
2.4 Prior Art: Inhibition of Return 4
2.5 Prior Art: WEAVER++ Lexical Selection 4
2.6 Prior Art: Kahneman Dual-Process Theory 4
2.7 Limitations of the Prior Art 5
2.7.1 No Biological Grounding of Decoding Parameters 5
2.7.2 No Evidence Accumulation 5
2.7.3 No Self-Monitoring 5
2.7.4 No Competitive Lexical Dynamics 5
2.7.5 No Working Memory Integration 5
2.8 Need for Improvement 5

3 SUMMARY **6**

4 DESCRIPTION OF THE FIGURES **8**

5 DETAILED DESCRIPTION **31**

5.1 Overview of the Language Production System 31
5.2 Signal Extraction from the Cortical Column Network 32
5.2.1 Multi-Scale Prediction Precision (5 signals) 32
5.2.2 Neuromodulatory Levels (4 signals) 32
5.2.3 Metacognitive Signals (3 signals) 33
5.2.4 Subcortical Signals (2 signals) 33
5.2.5 Memory Confidence Signals (5 signals) 33
5.3 Stage 1: Conceptualise — Discourse Guidance 33
5.3.1 Global Workspace Guidance 34
5.3.2 NMDA Novelty Suppression 34
5.4 Stage 2: Formulate — Syntactic Planning 34
5.4.1 L2/3 Spreading Activation 34
5.4.2 Phonological Loop Planning 35
5.4.3 Schema-Guided Generation 35
5.4.4 Episodic Production Retrieval 35
5.5 Stage 3: Basal Ganglia Action Selection 36

5.6 Stage 4: Decide — Drift-Diffusion Evidence Accumulation 36
5.6.1 Signal-to-Parameter Mapping 36
5.6.2 Evidence Accumulation Process 38
5.6.3 Confidence-Gated Dual-Process Deliberation 39
5.7 Stage 5: Monitor — Self-Check with WM Checkpoint/Restore 39
5.7.1 Working Memory Checkpoint 39
5.7.2 Comprehension Verification 39
5.7.3 Adaptive Rejection Threshold 40
5.7.4 Accept or Reject with Restore 40
5.8 Stage 6: Produce — IOR Update and Token Commitment 40
5.8.1 Neuromodulatory-Adaptive Decay 41
5.8.2 Discourse-Modulated Strength 41
5.9 Stage 6b: Comprehend — Unconditional Self-Feedback 41
5.10 Stage 7: Competitor Suppression 42
5.11 Stage 8: Remember — Online Episodic Encoding 42
5.12 Stage 8b: SWR Phase-Locked Micro-Consolidation 42
5.12.1 SWR Phase Gate 43
5.13 Curiosity-Driven Exploration 43
5.14 Speculative Decoding via Prospective Hippocampus 43
5.14.1 Draft Phase 44
5.14.2 Verification Phase 44
5.15 Competitive Lexical Production Head (BrainLMHead) 44
5.15.1 Stage 1: Lexical Activation with Production Delta 44
5.15.2 Stage 2: Competitive Lateral Inhibition 45
5.16 Brain Coherence Score (BCS) 45
5.16.1 Self-Coherence via Self-Reading 46
5.16.2 IOR Effectiveness 46
5.16.3 Vocabulary Diversity 47
5.16.4 Entity Binding Fidelity 47
5.16.5 Working Memory Utilisation 47
5.16.6 Boundary Structure 47
5.16.7 HPC Context Influence 47
5.16.8 Composite Score 47

TITLE

Brain-Inspired Language Production System with Multi-Stage Evidence Accumulation and Self-Monitoring

1. FIELD

The present invention relates generally to systems and methods for language generation in neural network architectures, and more particularly to a brain-inspired language production system implementing an eight-stage pipeline modelled on the Levelt speech production framework, incorporating drift-diffusion evidence accumulation with neuromodulatory signal integration, confidence-gated dual-process deliberation, inhibition of return with neuromodulatory adaptation, competitive lexical production via lateral inhibition, self-monitoring with working memory checkpoint and restore, sharp wave ripple phase-locked micro-consolidation during generation, curiosity-driven exploration, speculative multi-token decoding, and a composite brain coherence evaluation metric, for producing sequential language output from a cortical column network.

2. BACKGROUND

2.1. Prior Art: Standard Neural Language Generation

State-of-the-art neural language models generate text by computing a probability distribution over a vocabulary of tokens at each generation step, and selecting the next token using a decoding strategy. The standard decoding pipeline comprises: (a) computing raw logit scores from the final hidden state via a linear projection head; (b) applying a temperature scaling parameter τ to modulate the sharpness of the distribution; (c) optionally masking logits outside a top-k or nucleus (top-p) subset; (d) normalising via softmax; and (e) sampling from the resulting multinomial distribution. Variants such as beam search, contrastive search, and typical sampling have been proposed, but all share the fundamental structure of a single-pass logit-to-token pipeline with static or manually-tuned hyperparameters.

2.2. Prior Art: Drift-Diffusion Models

Ratcliff and McKoon (2008) established the drift-diffusion model (DDM) as the dominant account of two-alternative forced choice decision making in cognitive psychology and neuroscience. In the DDM, evidence for competing alternatives accumulates stochastically over time until one accumulator crosses a decision threshold. The model has been validated against human reaction time distributions, error rates, and neural recordings in parietal cortex (Gold & Shadlen 2007).

Key parameters include the drift rate (quality of evidence), the decision threshold (caution), the non-decision time, and diffusion noise. Extensions include collapsing bounds (urgency signals) and multi-alternative race models (Usher & McClelland 2001). However, no prior work has mapped neuromodulatory signals from a biologically-grounded neural network into DDM parameters for language generation.

2.3. Prior Art: Levelt Speech Production Model

Levelt (1989) proposed a staged model of speech production comprising: conceptualisation (deciding what to say), formulation (grammatical and phonological encoding), articulation (motor execution), and self-monitoring (checking output through the comprehension system). This model has been extensively validated in psycholinguistics and has informed computational models of word production such as WEAVER++ (Roelofs 1997, Levelt et al. 1999). However, no prior work has implemented the full Levelt pipeline as a computational decoding system within a neural language model.

2.4. Prior Art: Inhibition of Return

Klein (2000) and Posner and Cohen (1984) established inhibition of return (IOR) as a fundamental attentional mechanism whereby recently attended locations or objects receive reduced attentional priority, preventing perseverative reprocessing. In language production, analogous mechanisms prevent perseverative word repetition. However, existing neural decoding systems implement repetition avoidance through ad hoc penalties (e.g., static n-gram blocking, frequency penalties) that bear no relation to the biological IOR mechanism and lack neuromodulatory adaptation.

2.5. Prior Art: WEAVER++ Lexical Selection

Roelofs (1997) and Levelt et al. (1999) proposed the WEAVER++ model of lexical access, wherein word production involves activation spreading through a semantic-syntactic-phonological network with competitive lateral inhibition between semantically related lemmas. Standard neural language model heads compute logits $= h \cdot W_{\text{embed}}^{\top}$ without any competitive dynamics, producing a flat projection that treats all vocabulary items independently. This fails to model the competitive winner-take-all dynamics observed in human word production.

2.6. Prior Art: Kahneman Dual-Process Theory

Kahneman (2011) proposed that human cognition operates through two systems: System 1 (fast, automatic, intuitive) and System 2 (slow, deliberate, analytical). In language production, high-confidence tokens are produced rapidly (System 1) while ambiguous contexts trigger deliberate reasoning (System 2). No prior decoding system implements this dual-process architecture with an explicit confidence gate that routes tokens between fast and slow processing paths.

2.7. Limitations of the Prior Art

Despite the extensive neuroscience and psycholinguistic literature on language production, existing neural text generation systems suffer from the following limitations:

2.7.1. No Biological Grounding of Decoding Parameters

Standard decoding hyperparameters (temperature, top-k, top-p, repetition penalty) are manually tuned and bear no correspondence to biological signals. The brain modulates production through neuromodulatory systems (dopamine, norepinephrine, serotonin, acetylcholine) that adaptively control decision thresholds, noise levels, candidate breadth, and temporal patience. No prior system maps internal neuromodulatory signals to decoding parameters.

2.7.2. No Evidence Accumulation

Standard decoding selects tokens in a single forward pass without temporal evidence accumulation. The brain accumulates evidence over time through drift-diffusion processes in cortical and subcortical circuits before committing to a decision. This temporal dimension is entirely absent from prior decoding systems.

2.7.3. No Self-Monitoring

Standard generation systems produce tokens without any internal verification. The Levelt model's self-monitoring loop — wherein the speaker processes their own planned output through the comprehension system before articulation — has no counterpart in existing neural decoding. This leads to error-unaware generation analogous to jargon aphasia.

2.7.4. No Competitive Lexical Dynamics

Standard linear projection heads treat all vocabulary items independently. The WEAVER++ competitive dynamics that produce winner-take-all lexical selection in human word production are absent.

2.7.5. No Working Memory Integration

Standard decoding has no mechanism to checkpoint and restore working memory state when a planned token is rejected during self-monitoring. Without state management, rejected tokens corrupt the model's internal representation.

2.8. Need for Improvement

There exists a need in the art for a language production system that: (a) implements a biologically-grounded multi-stage production pipeline; (b) maps internal neuromodulatory signals to drift-

diffusion parameters with zero learnable decoding parameters; (c) provides confidence-gated dual-process deliberation; (d) implements neuromodulatory-adaptive inhibition of return; (e) includes competitive lexical production with lateral inhibition; (f) supports self-monitoring with working memory checkpoint and restore; (g) implements sharp wave ripple phase-locked micro-consolidation during generation; and (h) provides a composite brain coherence evaluation metric. The present invention addresses all of these needs.

3. SUMMARY

The present invention provides a computer-implemented system and method for brain-inspired language production that replaces conventional single-pass decoding with a biologically-grounded eight-stage production pipeline, modelled on the Levelt speech production framework and implemented within a cortical column network.

In a first aspect, the invention provides a language production system comprising:

(i) an eight-stage production pipeline implementing the stages of Conceptualise, Formulate, Basal Ganglia Action Selection, Decide, Monitor, Produce, Competitor Suppress, and Remember with micro-consolidation;

(ii) a drift-diffusion decision module that maps at least seventeen brain signals from a cortical column network into six drift-diffusion parameters — candidate count K, initial threshold θ_0, urgency time constant τ, diffusion noise σ, maximum deliberation steps M, and drift rate d — with zero learnable decoding parameters, wherein evidence for competing token candidates is accumulated stochastically over discrete time steps according to $a_i(t{+}1) = a_i(t) + d_i + \sigma \cdot \xi_i(t)$ under a collapsing threshold $\theta(t) = \theta_0 \cdot \exp(-t/\tau)$;

(iii) a confidence-gated deliberation mechanism implementing Kahneman dual-process theory, wherein a confidence score computed as the ratio of the logit margin between the top two candidates to the diffusion noise is compared against a threshold, routing high-confidence tokens to a fast System 1 path that selects the top candidate without evidence accumulation, and low-confidence tokens to a slow System 2 path that executes the full drift-diffusion process;

(iv) a neuromodulatory-adaptive inhibition of return module wherein the decay time constant is modulated by acetylcholine and norepinephrine levels according to $\tau_{\text{eff}} = \tau_{\text{base}} \times (0.5 + \text{ACh}) \times (1.5 - \text{NE})$, and the inhibition strength is modulated by discourse-level prediction precision;

(v) a competitive lexical production head implementing WEAVER++ lateral inhibition, comprising a base embedding weight matrix augmented with a low-rank production delta $\mathbf{W}_{\text{produce}} = \mathbf{W}_{\text{embed}} + \Delta\mathbf{A} \cdot \Delta\mathbf{B}$, semantic projection into a competition space, and gated

lateral inhibition wherein each token's logit is suppressed in proportion to its similarity to the active semantic field minus self-similarity;

(vi) a self-monitoring module that, before committing a token, saves the complete working memory state (NMDA attractor slots, activity-silent trace variables, phonological loop state and pointer, and L4 generation cache), runs the selected token through the comprehension system, computes a prediction error ratio, and upon rejection restores the saved working memory state and re-samples with the rejected token suppressed, with an adaptive rejection threshold modulated by discourse-level precision;

(vii) a sharp wave ripple phase-locked micro-consolidation mechanism wherein hippocampal replay during generation is gated by theta oscillation phase according to $g_{\text{SWR}} = 0.5 \times (1 - \cos(\phi_\theta))$, allowing consolidation only at the theta trough;

(viii) a curiosity-driven exploration mechanism wherein a curiosity signal computed as $c = \bar{\sigma}^2 \times (1 - c_{\text{concept}})$ modulates effective norepinephrine (widening candidate pool), decision threshold (lowering barrier), and diffusion noise (increasing exploration);

(ix) a speculative decoding module using prospective hippocampal predictions to draft multiple candidate tokens ahead of the current position, followed by batch verification against the full model;

(x) a competitive lexical production head wherein lateral inhibition operates in a learned semantic competition space with gated alpha controlling competition strength, frequency-adaptive per-token output bias, and identity-at-init(Stale from transformer era check with /nct-research)ialisation ensuring the head degrades gracefully to standard linear projection; and

(xi) a brain coherence score evaluation metric comprising a weighted combination of self-coherence (measured by re-reading own output), inhibition of return effectiveness, vocabulary diversity, entity binding fidelity, working memory utilisation, boundary structure, and hippocampal context influence, producing a composite score on a 0–100 scale.

In a second aspect, the invention provides a method for generating language output from a cortical column network, comprising executing the eight-stage production pipeline at each generation step: conceptualising discourse intent via global workspace and NMDA novelty bias; formulating syntactic structure via phonological loop planning, L2/3 spreading activation, schema-guided bias, and episodic production retrieval; computing basal ganglia go/no-go signals that modulate drift rate; deciding via drift-diffusion evidence accumulation with neuromodulatory signal integration; monitoring via self-comprehension with working memory checkpoint and restore; producing via IOR update and token commitment; suppressing competitors via IOR-like inhibition of rejected candidates; and encoding the production event in hippocampal episodic memory with theta-phase-locked micro-consolidation.

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 language production system and method described above.

The system provides the following advantages over the prior art:

- Biologically-grounded decoding with zero learnable parameters, replacing manually-tuned temperature, top-k, and top-p with neuromodulatory signal-driven adaptive control;
- Multi-step evidence accumulation providing temporal depth absent from single-pass decoding;
- Confidence-gated dual-process deliberation balancing speed and accuracy;
- Self-monitoring with working memory state preservation preventing error-blind generation;
- Competitive lexical production implementing winner-take-all dynamics observed in human word production;
- Phase-locked micro-consolidation enabling online learning during generation without disrupting production;
- Curiosity-driven exploration promoting novel token selection in uncertain contexts;
- A principled composite evaluation metric grounded in specific brain mechanisms rather than surface-level text statistics; and
- Disease-mapping validation wherein removal of any pipeline stage produces a specific 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. P6-1 (Claim 1) is a block diagram illustrating the complete DDM-based language production system, showing the signal extraction module with 22 brain signals, the drift-diffusion decision module with zero learnable parameters, the confidence-gated dual-process deliberation mechanism, and the self-monitoring module with working memory checkpoint and restore.

FIG. P6-2 (Claim 2) is a flow diagram illustrating the eight-stage Levelt production pipeline, showing the sequential flow from signal extraction through Conceptualise, Formulate, BG Select, Decide, Monitor, Produce, Suppress, and Remember stages, with the reject feedback loop and formulation sub-stages.

FIG. P6-3 (Claim 3) is a radar chart and block diagram illustrating the Brain Coherence Score evaluation metric, showing the seven weighted components (self-coherence 0.25, IOR effectiveness 0.15, vocabulary diversity 0.15, entity binding fidelity 0.15, WM utilisation 0.10, boundary structure 0.10, HPC context influence 0.10) with disease mapping annotations.

FIG. P6-4 (Claim 4) is a schematic diagram illustrating ACh-modulated logit sharpness, showing the scaling factor $s_{\text{ACh}} = 0.4 + 1.2 \times \text{ACh}$ applied to logits prior to candidate selection.

FIG. P6-5 (Claim 5) is a block diagram illustrating the curiosity-driven exploration module, showing the curiosity signal computation from mean prediction precision and concept confidence, and the three modulation targets: effective NE, decision threshold, and diffusion noise.

FIG. P6-6 (Claim 6) is a schematic diagram illustrating the neuromodulatory-adaptive inhibition of return module, showing the exponential decay trace with effective time constant $\tau_{\text{eff}} = \tau_{\text{base}} \times (0.5 + \text{ACh}) \times (1.5 - \text{NE})$ and discourse-level strength modulation.

FIG. P6-7 (Claim 7) is a block diagram illustrating competitor suppression via the Oppenheim (2010) model, showing partial IOR application to rejected DDM candidates.

FIG. P6-8 (Claim 8) is a timing diagram illustrating SWR phase-locked micro-consolidation, showing the theta oscillation phase, the SWR gate function $g_{\text{SWR}} = 0.5 \times (1 - \cos(\phi_\theta))$, and the consolidation window at the theta trough.

FIG. P6-9 (Claim 9) is a schematic diagram illustrating the adaptive rejection threshold modulated by discourse-level precision, showing $\theta_{\text{reject}} = \theta_{\text{base}} \times (1 + 0.3 \times \min(\sigma^2_{\text{discourse}}, 2.0))$.

FIG. P6-10

(Claim 10) is a flowchart illustrating the iterative self-monitoring loop with maximum resample attempts and force-commit fallback.

FIG. P6-11

(Claim 11) is a flow diagram illustrating the unconditional comprehension module that updates internal state after token commitment regardless of monitoring.

FIG. P6-12

(Claim 12) is a block diagram illustrating online episodic encoding with production flag and prediction-error-based priority.

FIG. P6-13

(Claim 13) is a schematic diagram illustrating the dual-process confidence gate, showing the System 1 fast path and System 2 deliberative path with the confidence threshold formula.

FIG. P6-14

(Claim 14) is a diagram illustrating the seven-factor multiplicative modulation of the initial threshold θ_0, including tonic dopamine, curiosity, mean prediction error, hippocampal confidence, NMDA confidence, concept confidence, and ACC conflict.

FIG. P6-15

(Claim 15) is a diagram illustrating the five-factor additive modulation of maximum deliberation steps, including ACC conflict, cerebellar loss, PC settling difficulty, WM uncertainty, and sentence-level precision.

FIG. P6-16

(Claim 16) is a block diagram illustrating four-factor NMDA novelty suppression during the Conceptualise stage, showing NMDA confidence, discourse uncertainty, HPC retrieval confidence, and concept confidence modulation.

FIG. P6-17

(Claim 17) is a schematic diagram illustrating reward-modulated production biasing, showing $\mathrm{DA}_{\mathrm{phasic}}$ and WM confidence scaling of spreading activation and phonological planning biases.

FIG. P6-18

(Claim 18) is a block diagram illustrating schema retrieval from HPC with four-factor gating: schema similarity, schema confidence, HPC retrieval confidence, and phasic dopamine.

FIG. P6-19

(Claim 19) is a flow diagram illustrating episodic production recall from HPC, showing experience-based production planning via past episode retrieval and logit bias application.

FIG. P6-20

(Claim 20) is a schematic diagram illustrating IOR decay modulation by ACh and NE during the Produce stage, with discourse-level strength modulation.

FIG. P6-21

(Claim 21) is a block diagram illustrating speculative multi-token decoding using prospective hippocampal projection, showing the draft phase, verification phase, and greedy acceptance.

FIG. P6-22

(Claim 22) is a flow diagram illustrating self-coherence score computation via re-feeding the generated sequence for self-prediction loss.

FIG. P6-23

(Claim 23) is a block diagram illustrating HPC context influence score computation as a fraction of generation steps with episodic retrieval.

FIG. P6-24

(Claim 24) is a table and formula diagram illustrating the specific BCS component weights summing to 1.0 with disease diagnostics mapping.

FIG. P6-25

(Claim 25) is a block diagram illustrating the WEAVER++ competitive lexical production head with LoRA delta, frequency bias, and lateral inhibition.

FIG. P6-26

(Claim 26) is a schematic diagram illustrating the low-rank production delta (LoRA) with zero-init A, Gaussian-init B, and identity at initialisation.

FIG. P6-27

(Claim 27) is a diagram illustrating lateral inhibition winner-take-all dynamics with the formula $\text{inhib}_v = \text{ReLU}(f_v - p_v \times \|\mathbf{S}_v\|^2)$ and self-similarity protection.

FIG. P6-28

(Claim 28) is a diagram illustrating the five-stage diffusion noise sigma construction from token precision, phrase excess, column disagreement, and memory confidence signals.

FIG. P6-29

(Claim 29) is a timing diagram illustrating SWR phase-locked consolidation of production episodes restricted to theta trough windows.

FIG. P6-30

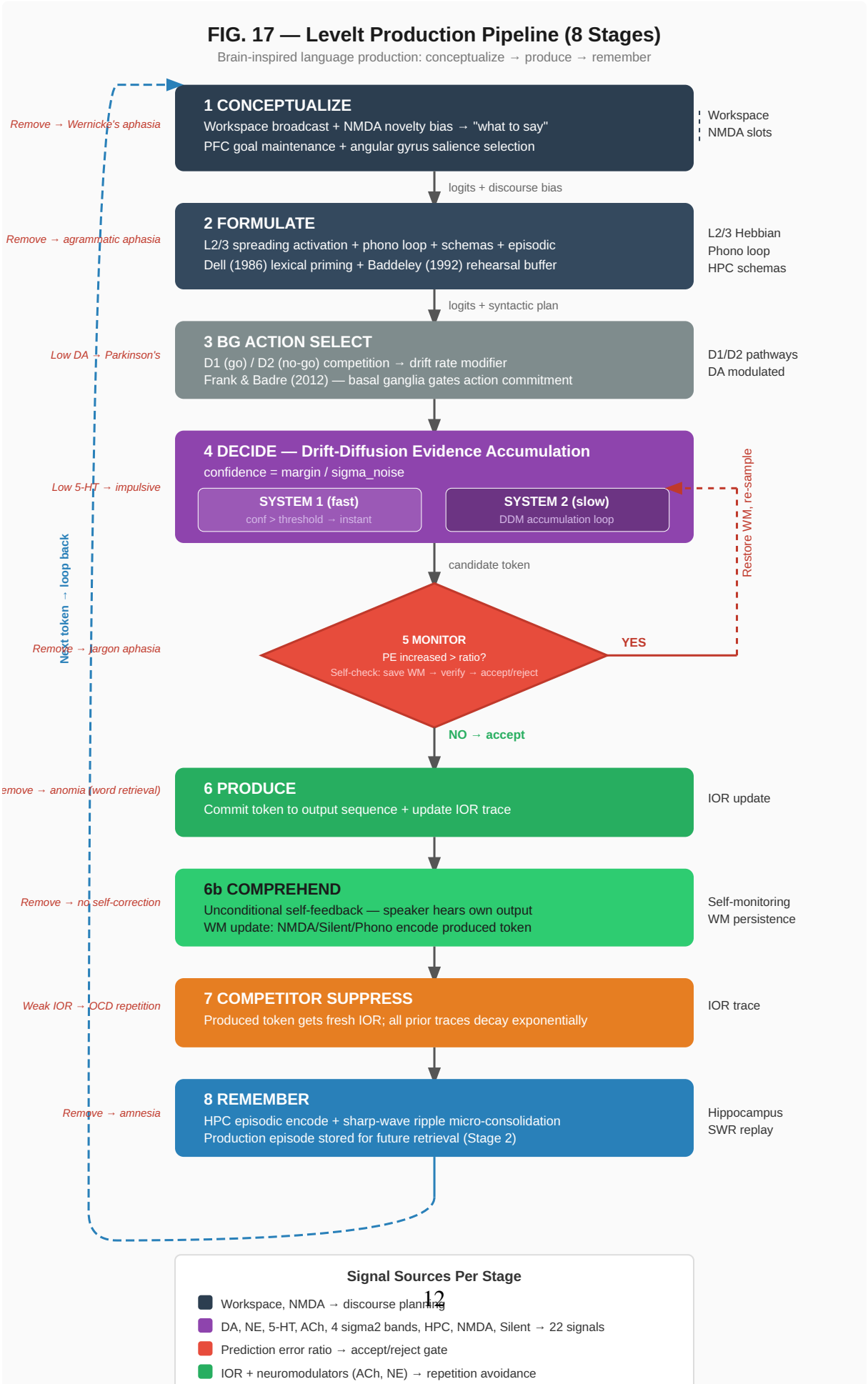
(Claim 30) is a block diagram illustrating basal ganglia D1/D2 action selection producing a drift rate modifier in the range $[0.8, 1.2]$.

FIG. P6-31

(Claim 31) is a comprehensive block diagram illustrating the complete Levelt eight-stage pipeline with all brain region mappings, signal flows, and disease annotations per stage.

Note: All figures are provided as scalable vector graphics (SVG) and are incorporated in the following pages.

DRAWING SHEETS



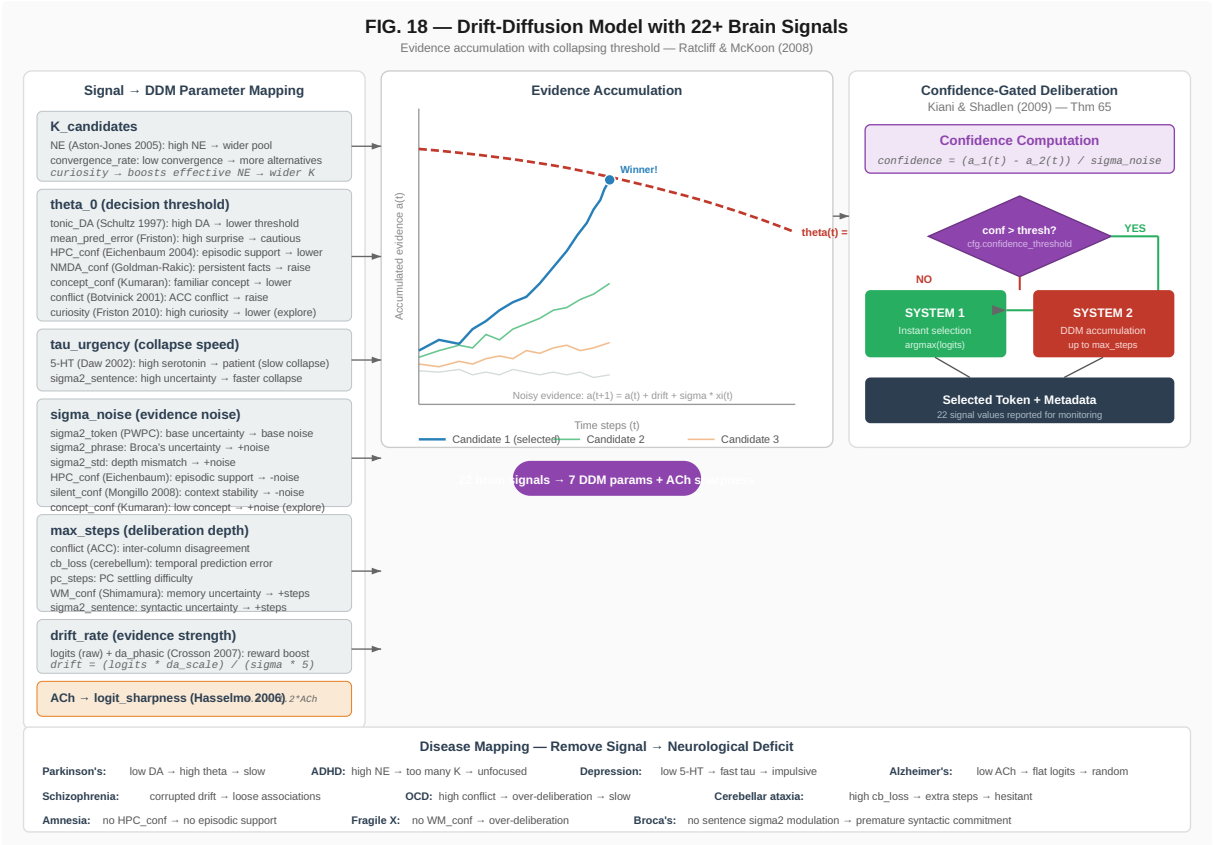


Figure 2: FIG. 2 — Mapping of 22 Brain Signals to DDM Parameters

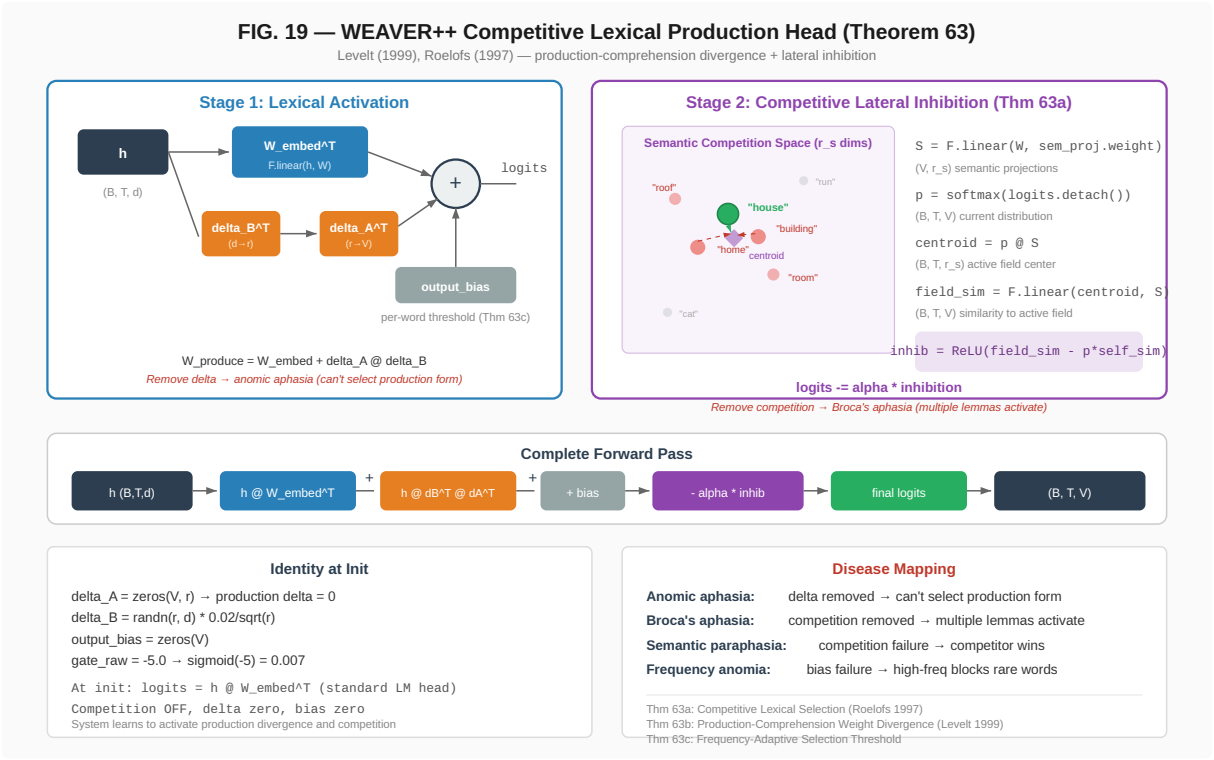


Figure 3: FIG. 3 — WEAVER++ Competitive Lexical Production Head

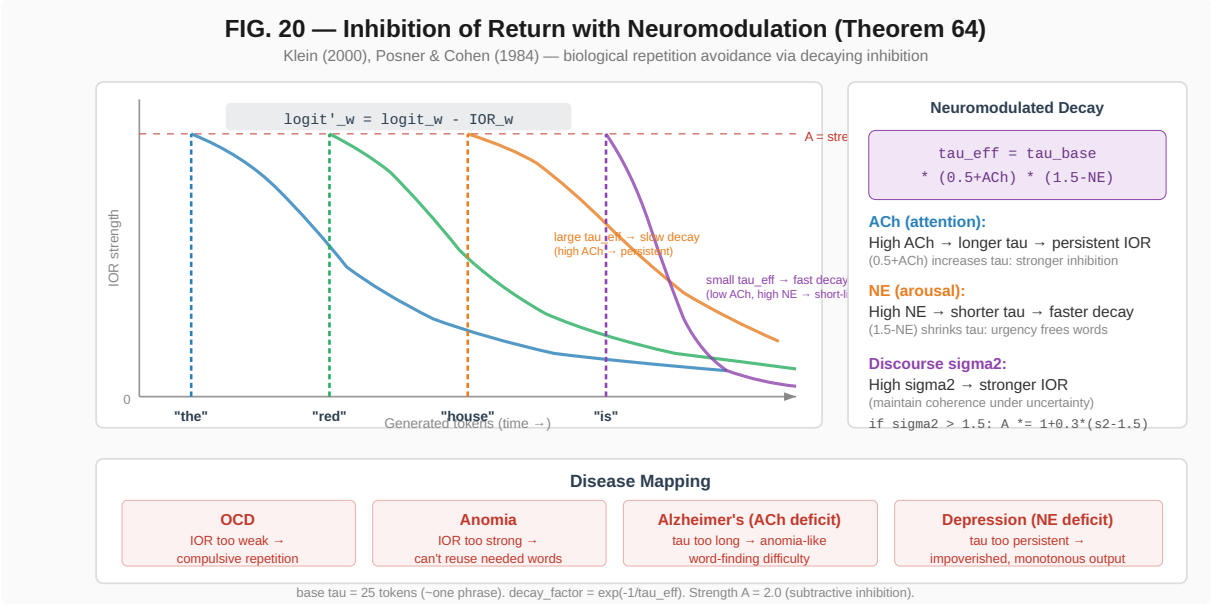


Figure 4: FIG. 4 — Neuromodulatory-Adaptive Inhibition of Return

CLAIM ILLUSTRATION SHEETS

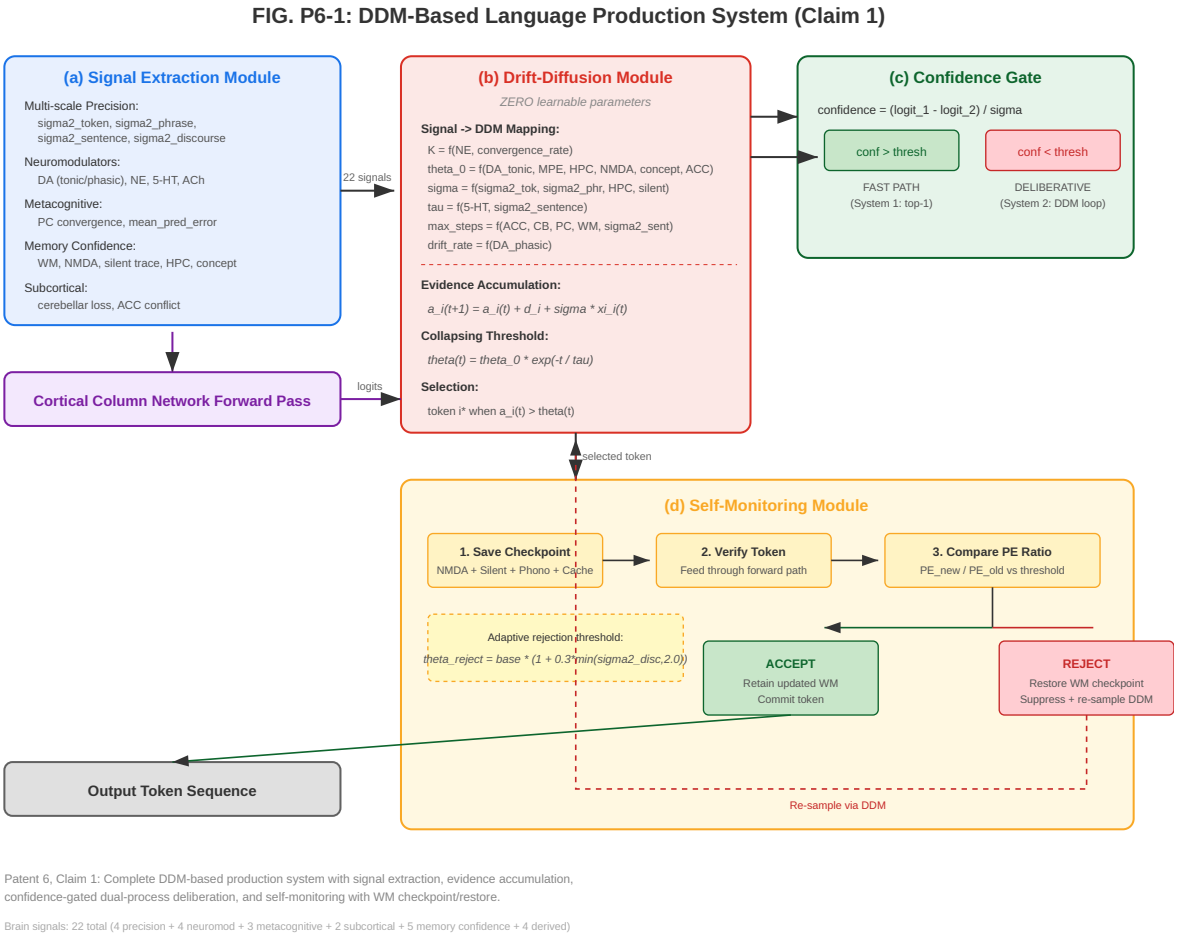


Figure 5: FIG. P6-1: DDM-Based Language Production System (Claim 1). The complete system architecture showing signal extraction (22 brain signals), drift-diffusion decision module (zero learnable parameters), confidence-gated dual-process deliberation, and self-monitoring with WM checkpoint/restore.

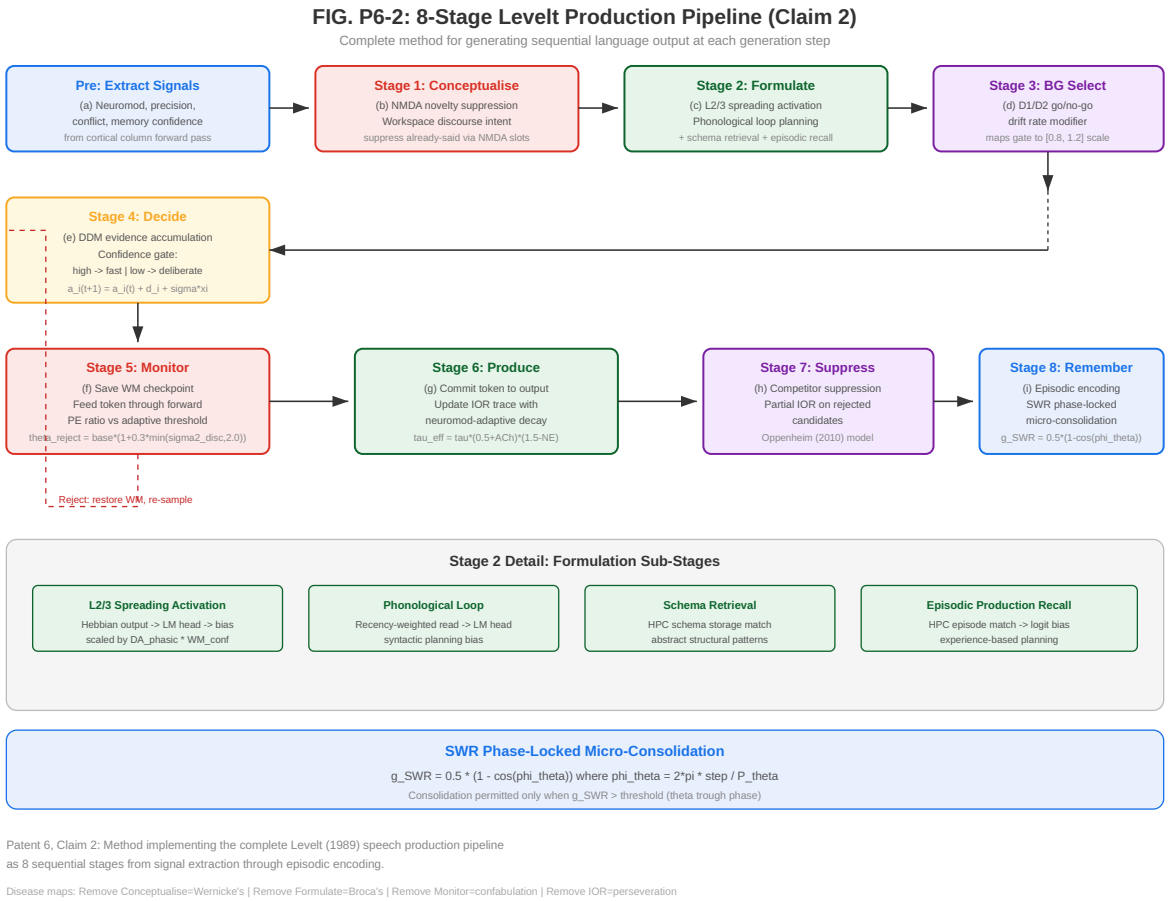


Figure 6: FIG. P6-2: Eight-Stage Levelt Production Pipeline (Claim 2). Pre-extraction followed by Stages 1–8: Conceptualise, Formulate, BG Select, Decide, Monitor, Produce, Suppress, Remember.

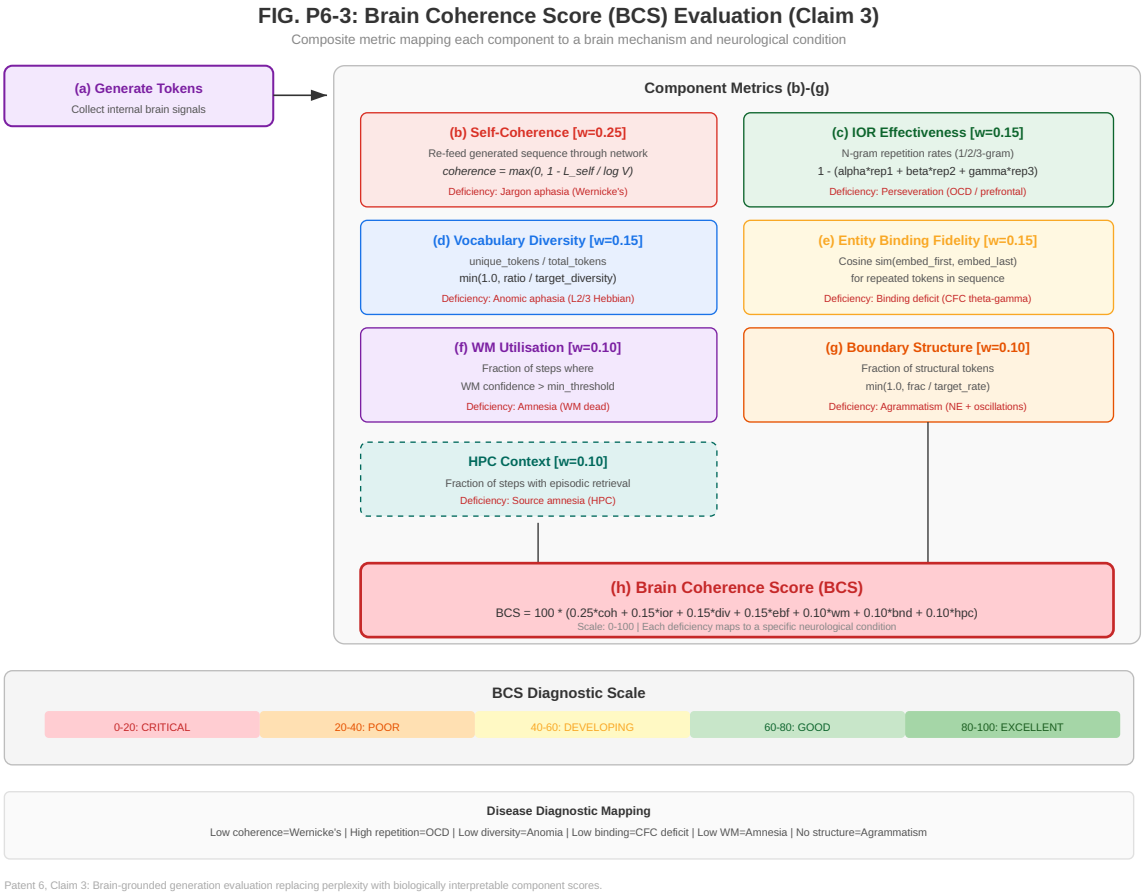


Figure 7: FIG. P6-3: Brain Coherence Score Evaluation Metric (Claim 3). Seven weighted components on a 0–100 scale, each mapped to a brain mechanism and neurological disease.

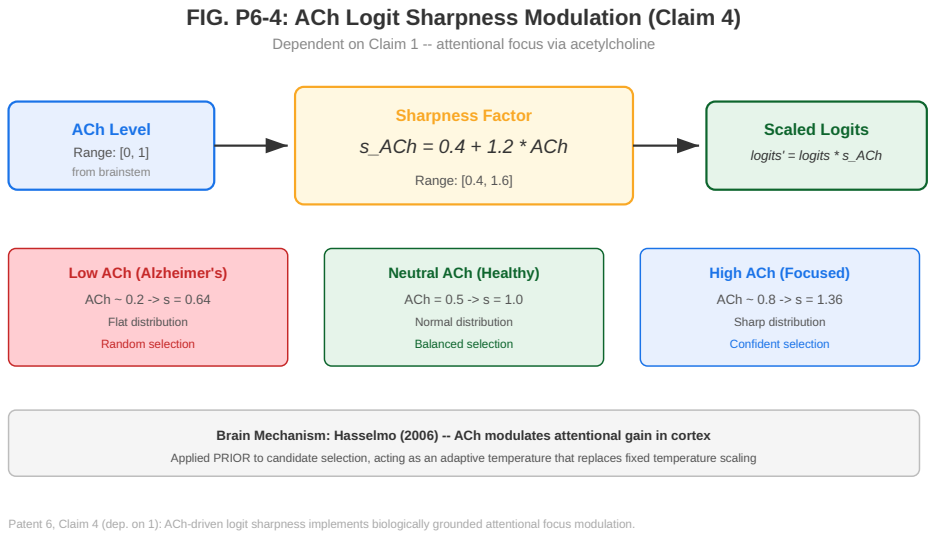


Figure 8: FIG. P6-4: ACh-Modulated Logit Sharpness (Claim 4). Sharpness factor $s_{\text{ACh}} = 0.4 + 1.2 \times \text{ACh}$ applied before candidate selection.

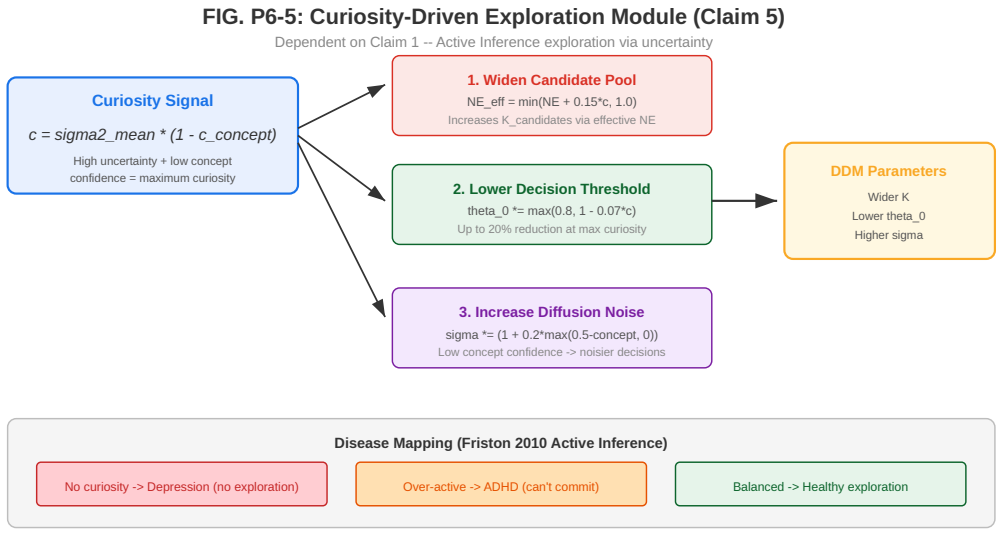


Figure 9: FIG. P6-5: Curiosity-Driven Exploration Module (Claim 5). Curiosity signal from $\bar{\sigma}^2 \times (1 - c_{\text{concept}})$ modulates NE, threshold, and noise.

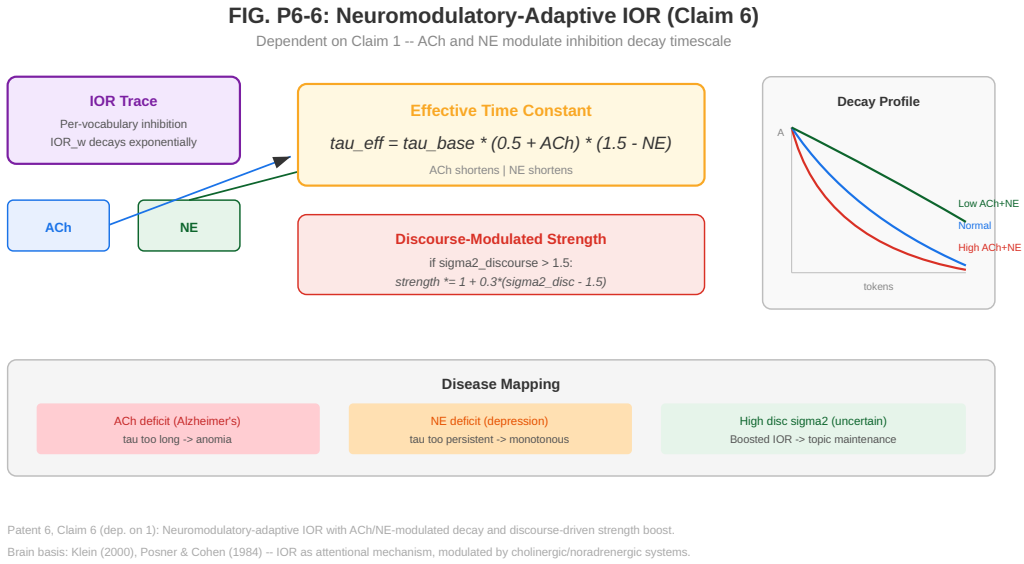


Figure 10: FIG. P6-6: Neuromodulatory-Adaptive Inhibition of Return (Claim 6). Exponential decay with $\tau_{\text{eff}} = \tau_{\text{base}} \times (0.5 + \text{ACh}) \times (1.5 - \text{NE})$.

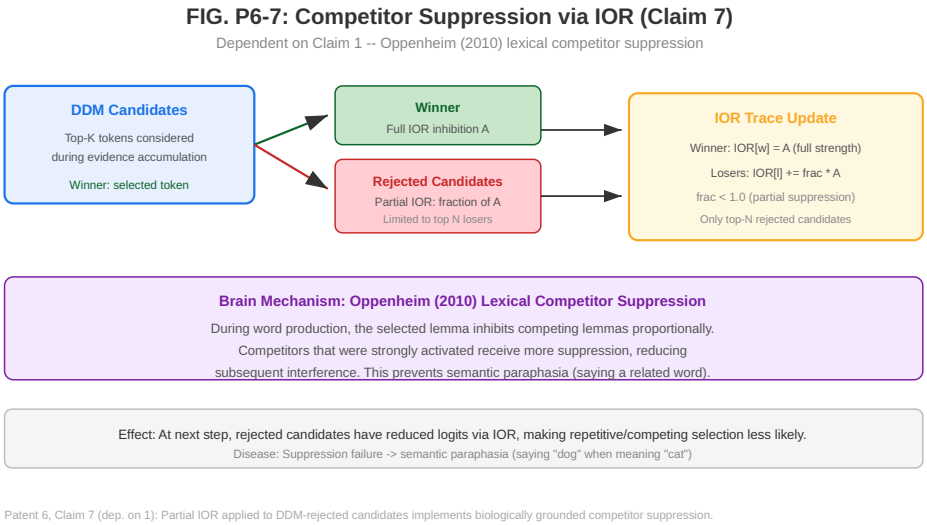


Figure 11: FIG. P6-7: Competitor Suppression via Oppenheim (2010) Model (Claim 7). Partial IOR on rejected DDM candidates.

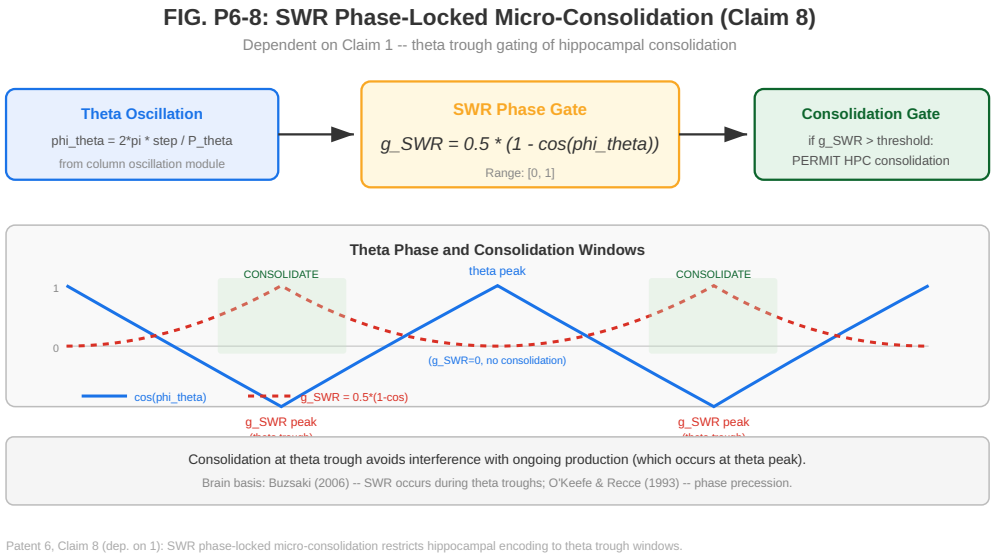


Figure 12: FIG. P6-8: SWR Phase-Locked Micro-Consolidation (Claim 8). Consolidation gated by $g_{\text{SWR}} = 0.5 \times (1 - \cos(\phi_\theta))$, permitted only at theta trough.

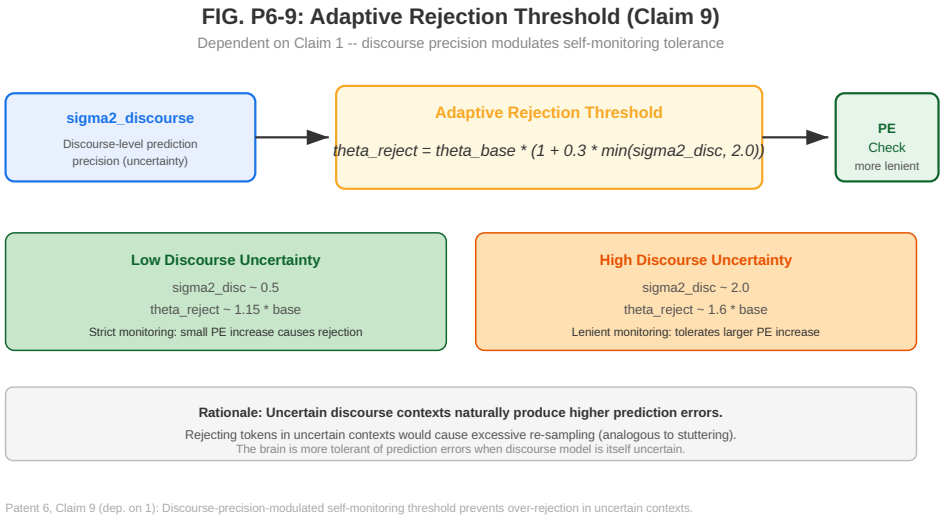


Figure 13: FIG. P6-9: Adaptive Rejection Threshold (Claim 9). $\theta_{\text{reject}} = \theta_{\text{base}} \times (1 + 0.3 \times \min(\sigma^2_{\text{disc}}, 2.0))$.

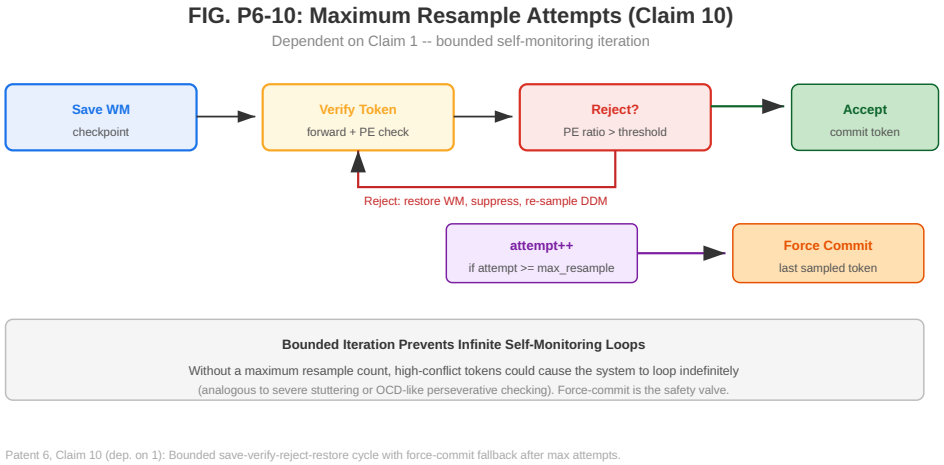


Figure 14: FIG. P6-10: Iterative Self-Monitoring with Max Resamples (Claim 10). Up to R_{max} resample attempts before force-commit.

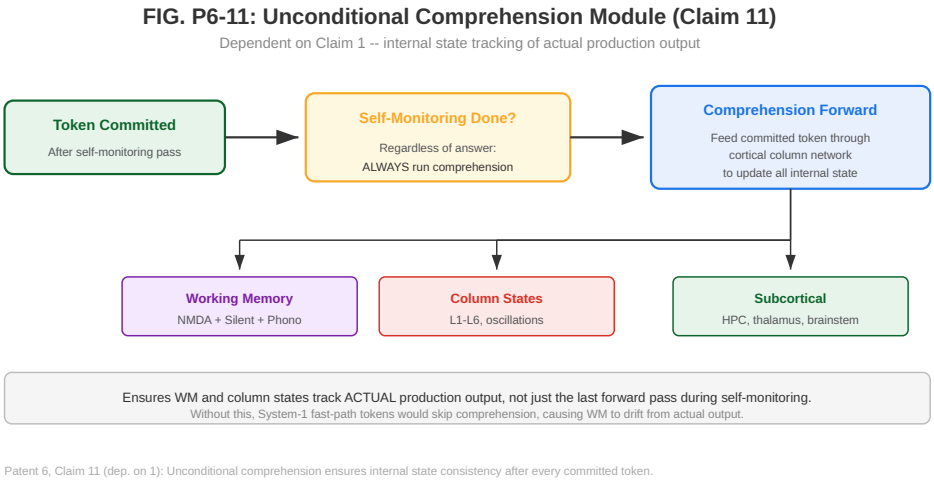


Figure 15: FIG. P6-11: Unconditional Comprehension Module (Claim 11). Forward pass after commitment updates internal state regardless of monitoring.

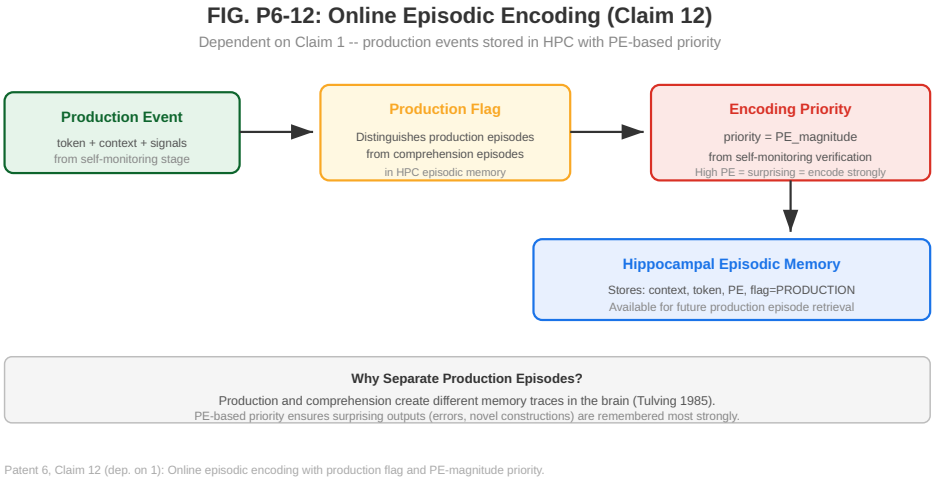


Figure 16: FIG. P6-12: Online Episodic Encoding with Production Flag (Claim 12). PE-based priority determines encoding strength.

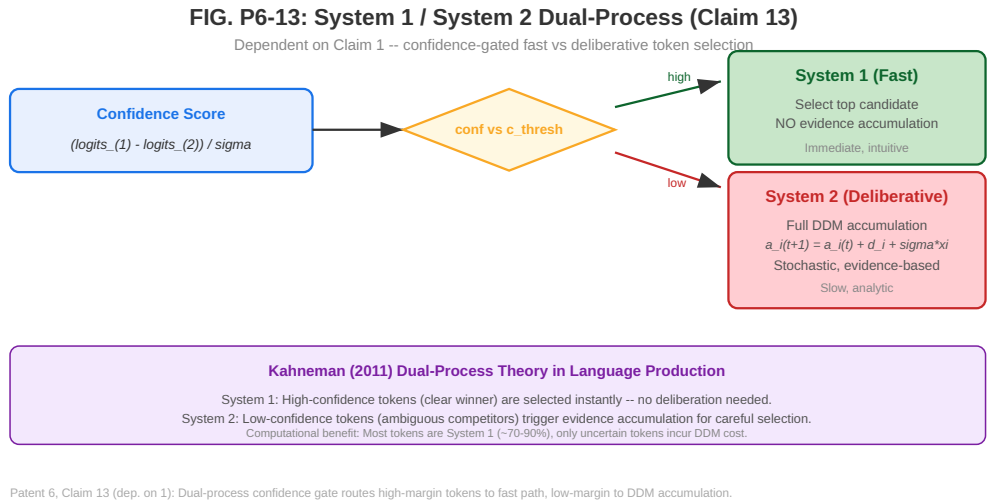


Figure 17: FIG. P6-13: Dual-Process Confidence Gate (Claim 13). System 1 fast path vs. System 2 deliberative path with confidence threshold.

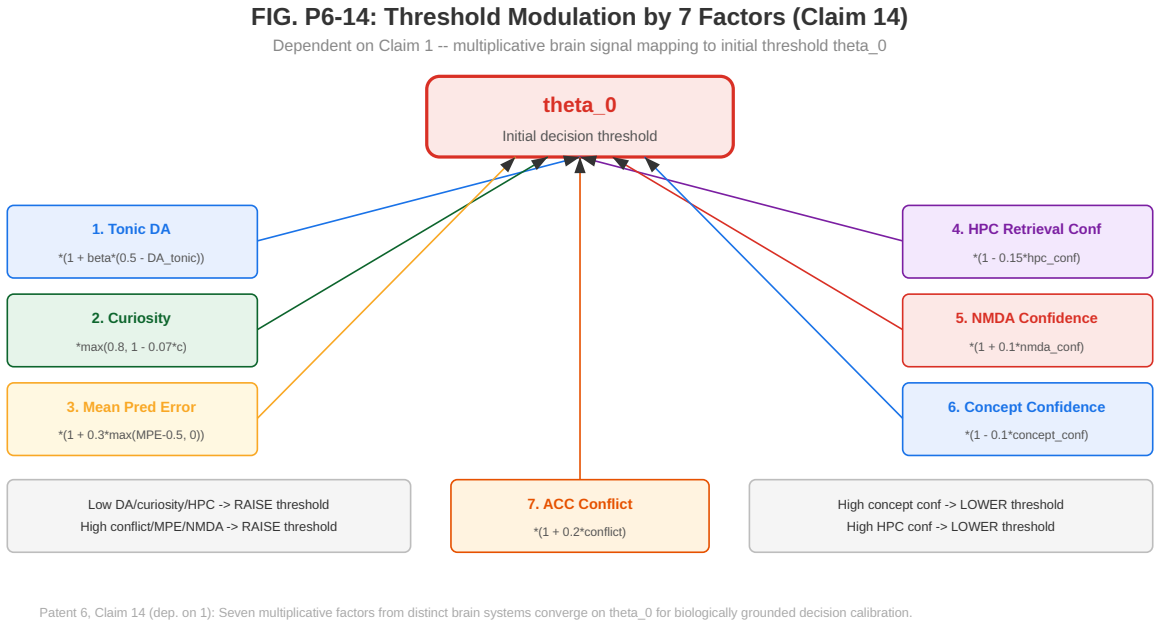


Figure 18: FIG. P6-14: Seven-Factor Initial Threshold θ_0 Mapping (Claim 14). Multiplicative modulation by $\mathrm{DA}_{\mathrm{tonic}}$, curiosity, MPE, HPC, NMDA, concept, and ACC conflict.

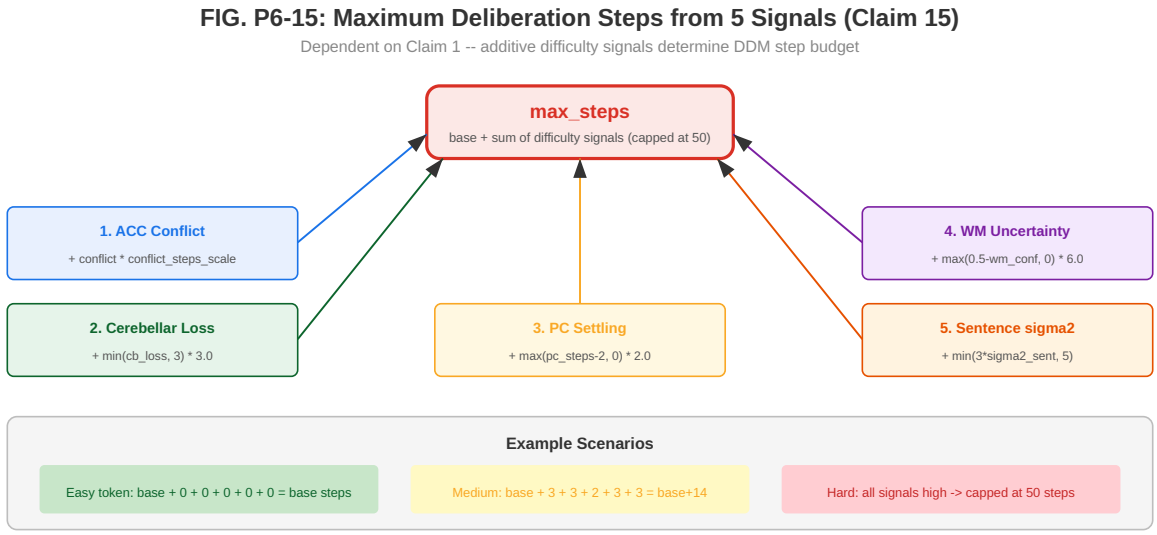


Figure 19: FIG. P6-15: Five-Factor Maximum Deliberation Steps (Claim 15). Additive contributions from ACC conflict, cerebellar loss, PC difficulty, WM uncertainty, and sentence precision.

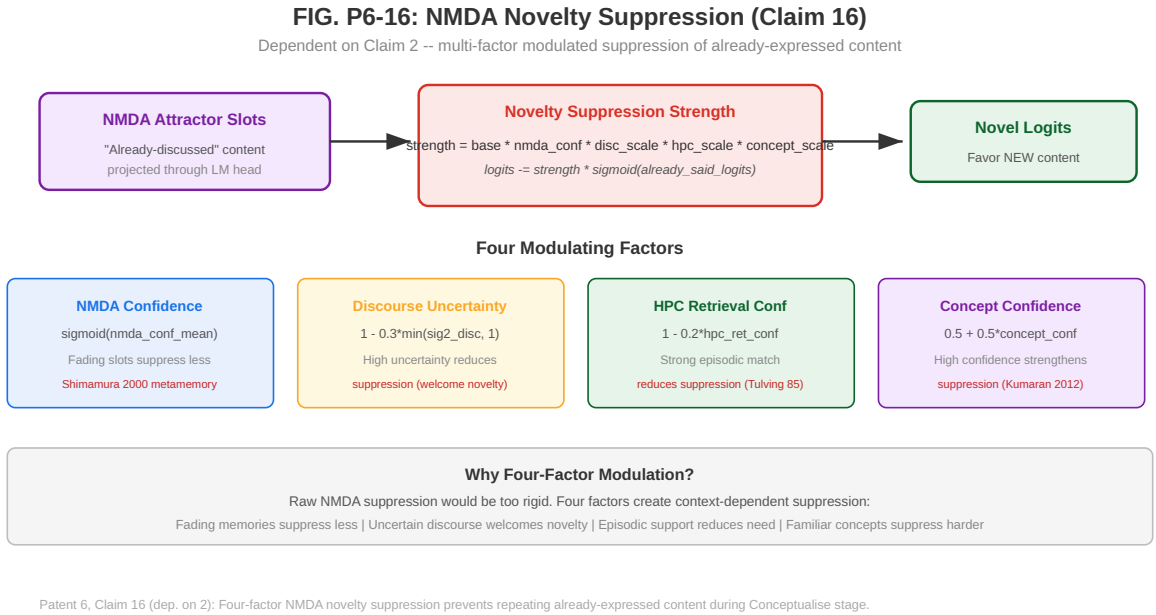


Figure 20: FIG. P6-16: Four-Factor NMDA Novelty Suppression (Claim 16). Modulated by NMDA confidence, discourse uncertainty, HPC retrieval confidence, and concept confidence.

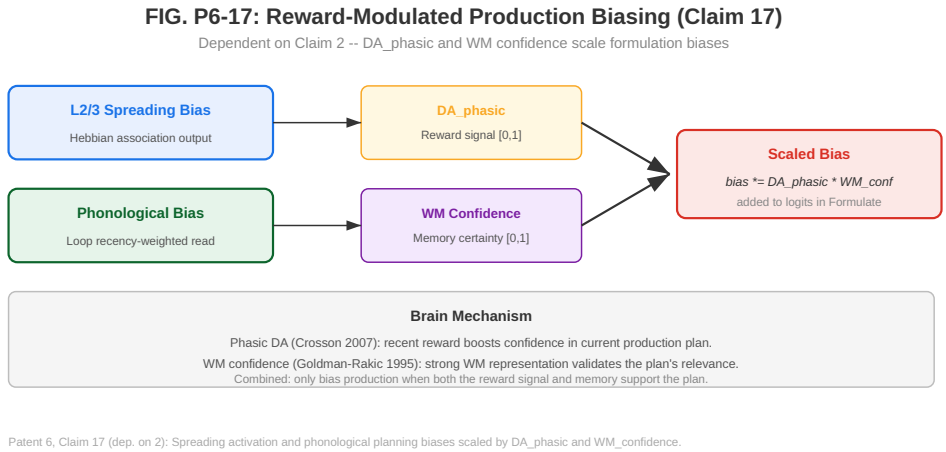


Figure 21: FIG. P6-17: Reward-Modulated Production Biasing (Claim 17). Spreading activation and phonological biases scaled by DA_{phasic} and WM confidence.

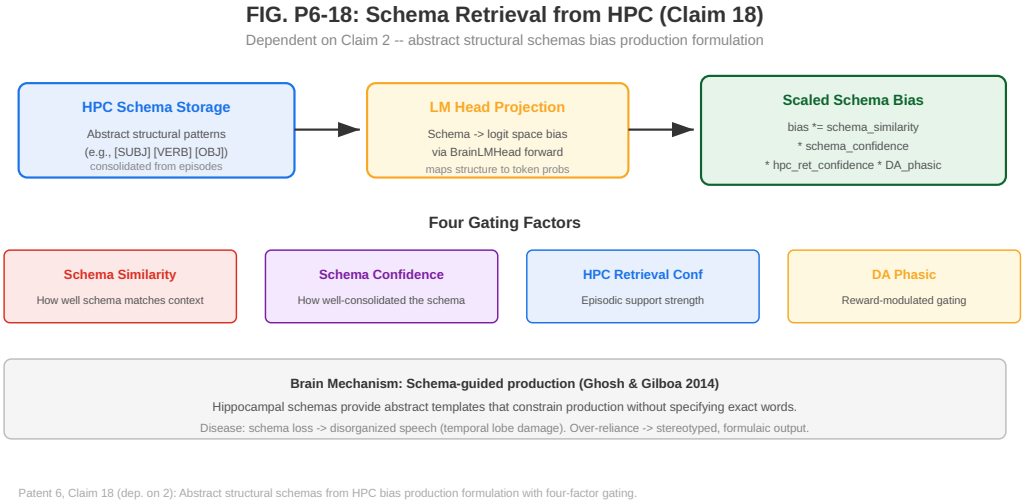


Figure 22: FIG. P6-18: Schema Retrieval from HPC (Claim 18). Four-factor gating: schema similarity, schema confidence, HPC retrieval confidence, DA_{phasic}.

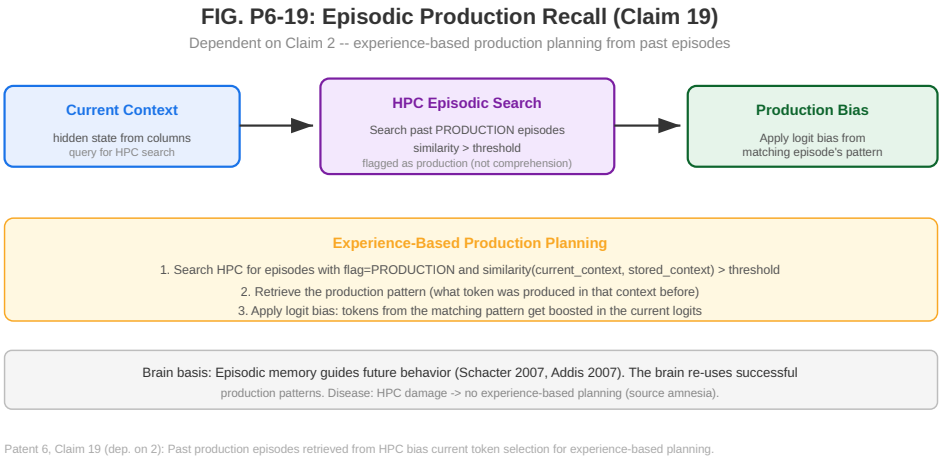


Figure 23: FIG. P6-19: Episodic Production Recall (Claim 19). Past production episodes retrieved from HPC bias current token selection.

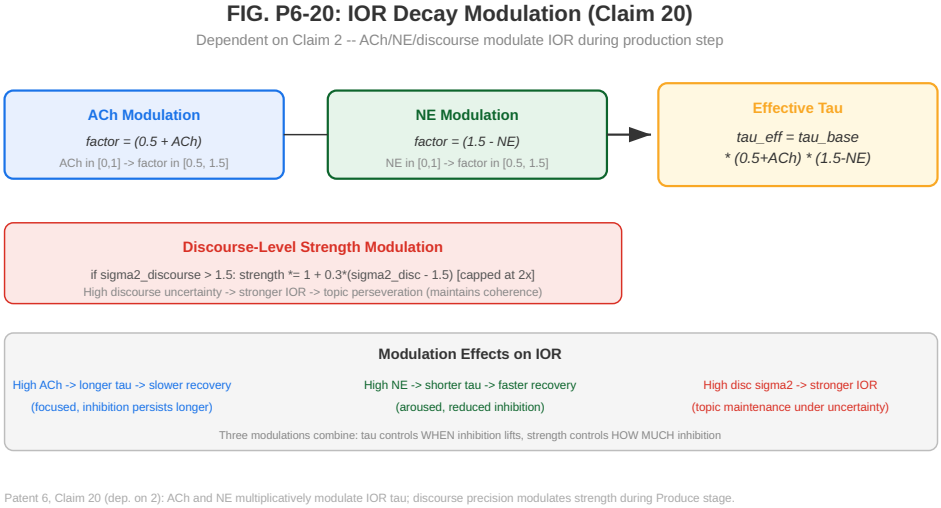


Figure 24: FIG. P6-20: IOR Decay Modulation (Claim 20). ACh and NE multiplicatively modulate τ; discourse precision modulates strength.

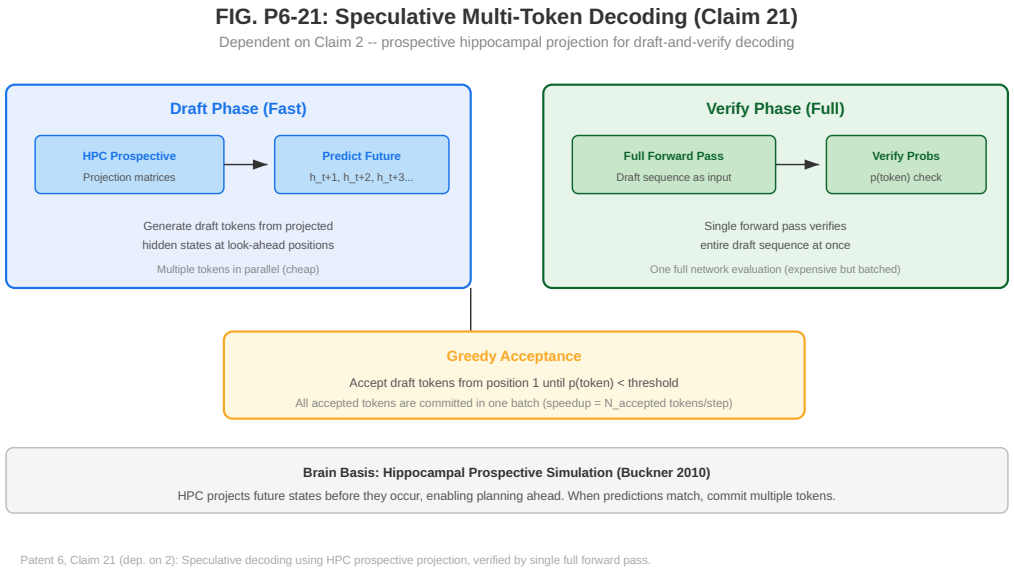


Figure 25: FIG. P6-21: Speculative Multi-Token Decoding (Claim 21). HPC prospective projection for draft-and-verify decoding.

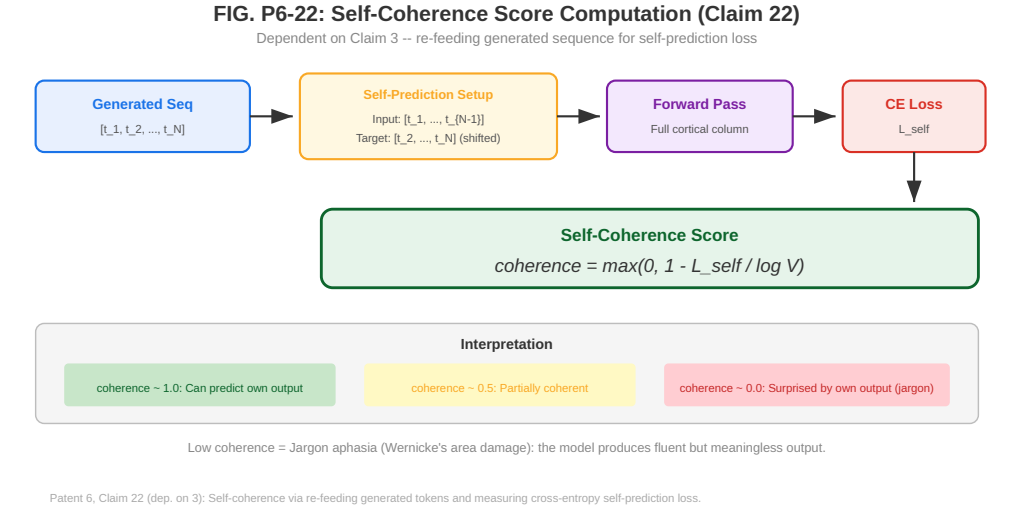


Figure 26: FIG. P6-22: Self-Coherence Score Computation (Claim 22). Re-feeding generated sequence for self-prediction loss: coherence $= \max(0, 1 - L_{\text{self}} / \log V)$.

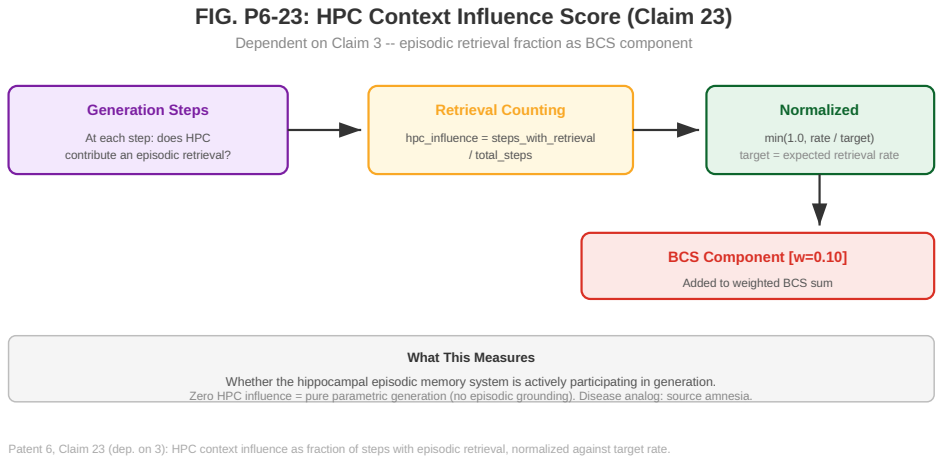


Figure 27: FIG. P6-23: HPC Context Influence Score (Claim 23). Fraction of generation steps with episodic retrieval, normalised against target rate.

FIG. P6-24: BCS Weights and Disease Diagnostics (Claim 24)

Dependent on Claim 3 -- specific component weights and neurological condition mapping

BCS Component Weights and Disease Mapping

Component	Weight	Brain Mechanism	Disease if Deficient
Self-Coherence	0.25	Wernicke's area comprehension	Jargon aphasia
IOR Effectiveness	0.15	Prefrontal inhibition (IOR)	Perseveration (OCD)
Vocabulary Diversity	0.15	L2/3 Hebbian competition	Anomic aphasia
Entity Binding Fidelity	0.15	CFC theta-gamma coupling	Binding deficit
WM Utilisation	0.10	NMDA/Silent/Phono (temporal)	Amnesia (WM dead)
Boundary Structure	0.10	NE + oscillations (brainstem)	Agrammatism
HPC Context Influence	0.10	Episodic retrieval (HPC)	Source amnesia
	1.00		

BCS = 100 * (0.25*coh + 0.15*ior + 0.15*div + 0.15*ebf + 0.10*wm + 0.10*bnd + 0.10*hpc)

Each component score in [0,1]. Composite in [0, 100]. Diagnostic output identifies deficient components.

Diagnostic: for each component scoring below threshold, the corresponding neurological condition is flagged.
This enables targeted intervention: low IOR -> adjust IOR parameters. Low coherence -> investigate self-monitoring.

Patent 6, Claim 24 (dep. on 3): Specific BCS weights summing to 1.0 with each deficiency mapping to a neurological condition.

Figure 28: FIG. P6-24: BCS Component Weights and Disease Diagnostics (Claim 24). Seven weights summing to 1.0, each deficiency mapped to a neurological condition.

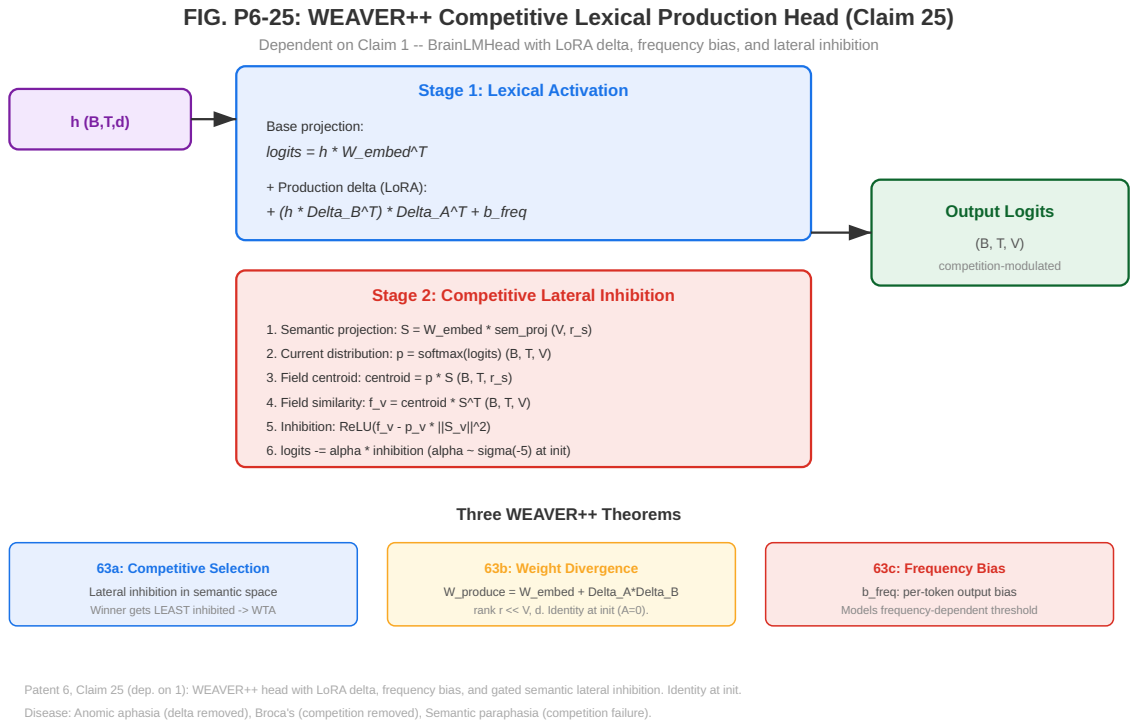


Figure 29: FIG. P6-25: WEAVER++ Competitive Lexical Production Head (Claim 25). LoRA delta, frequency bias, and gated semantic lateral inhibition.

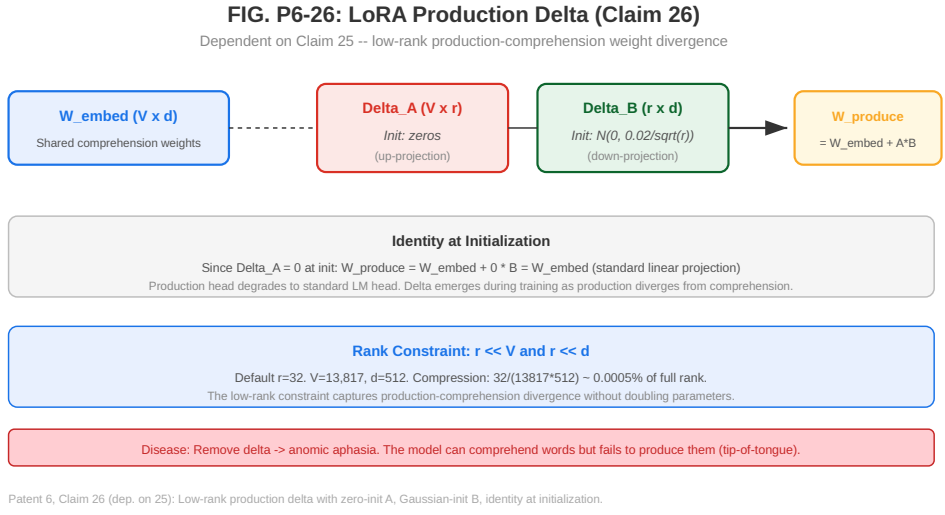


Figure 30: FIG. P6-26: LoRA Production Delta (Claim 26). Low-rank divergence: $\Delta A = 0$ at init, $\Delta B \sim \mathcal{N}(0, 0.02/\sqrt{r})$, identity at initialisation.

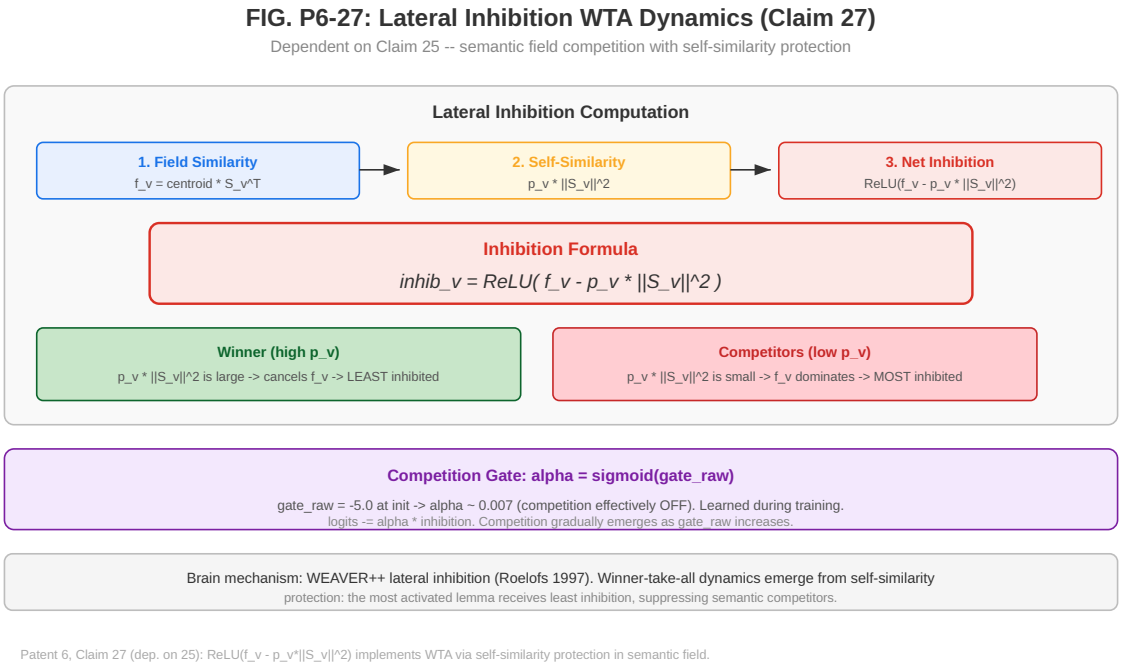


Figure 31: FIG. P6-27: Lateral Inhibition WTA Dynamics (Claim 27). $\text{inhib}_v = \text{ReLU}(f_v - p_v \times \|\mathbf{S}_v\|^2)$ with self-similarity protection.

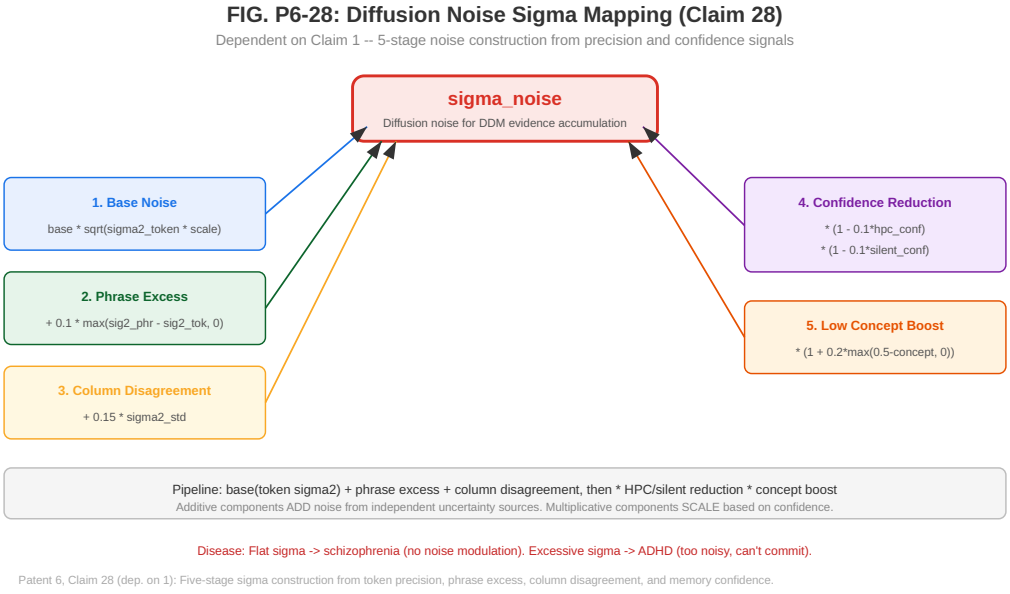


Figure 32: FIG. P6-28: Five-Stage Diffusion Noise σ Construction (Claim 28). Base noise + phrase excess + column disagreement, scaled by confidence signals.

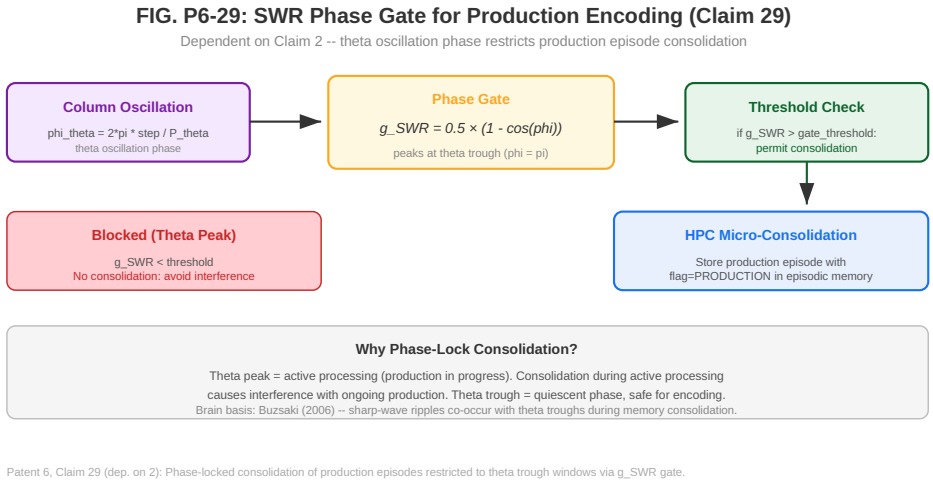


Figure 33: FIG. P6-29: SWR Phase Gate for Production Encoding (Claim 29). Theta oscillation phase restricts production episode consolidation to theta trough.

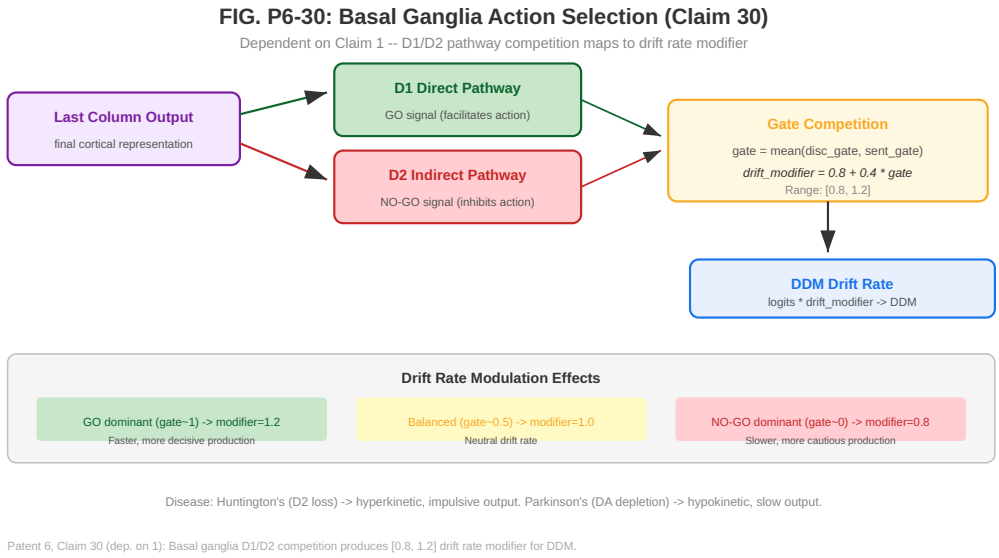


Figure 34: FIG. P6-30: Basal Ganglia D1/D2 Action Selection (Claim 30). Gate competition produces drift rate modifier in $[0.8, 1.2]$.

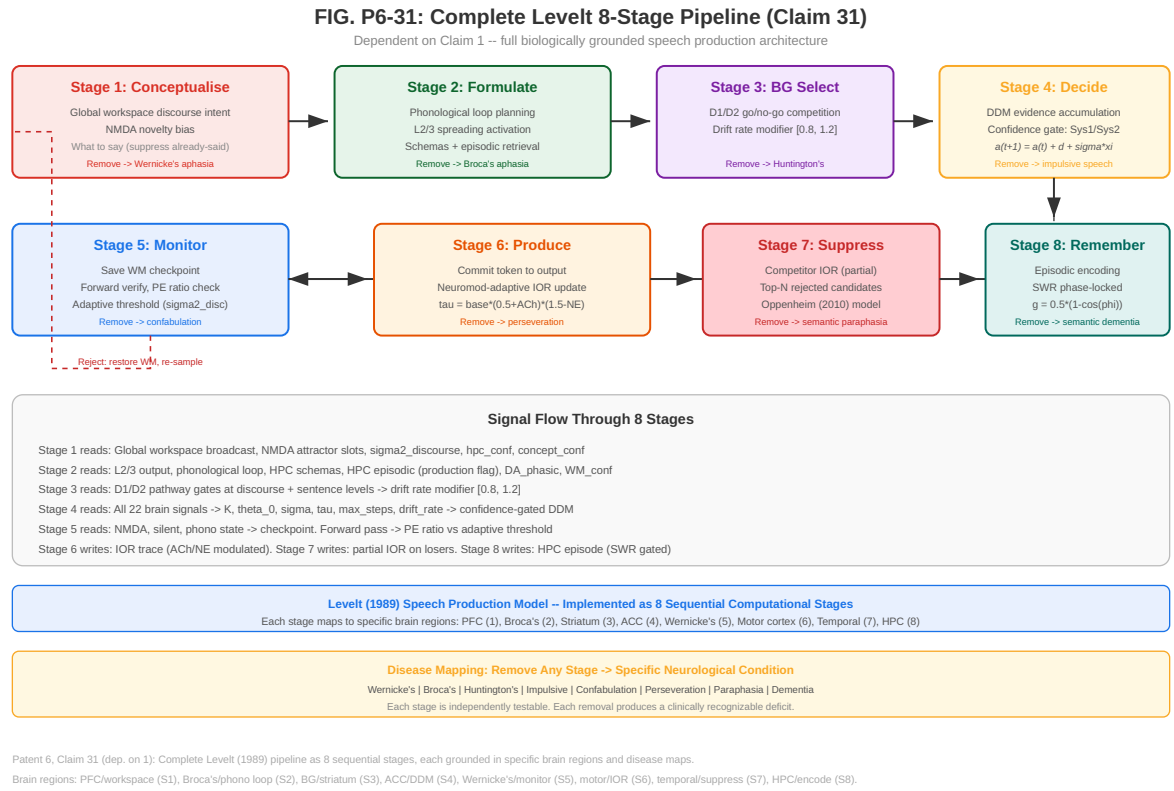


Figure 35: FIG. P6-31: Complete Levelt Eight-Stage Pipeline (Claim 31). All stages with brain region mappings, signal flows, and disease annotations.

5. DETAILED DESCRIPTION

The following detailed description sets forth specific embodiments of the invention. It is to be understood that the invention is not limited to these embodiments but encompasses all alternatives, modifications, and equivalents as may be included within the spirit and scope of the appended claims.

5.1. Overview of the Language Production System

The present invention implements a brain-inspired language production system (hereinafter "BrainDecoder") that replaces conventional single-pass neural decoding with an eight-stage pipeline modelled on the Levelt (1989) speech production framework (see Figure 1 for the general pipeline overview; FIG. P6-1, Figure 5 for the detailed system architecture; and FIG. P6-2, Figure 6 for the eight-stage pipeline). At each generation step, the system receives raw logit scores and internal brain signals from a cortical column network and produces a single output token through a biologically-grounded decision process.

The system has **zero learnable decoding parameters**. All adaptive behaviour arises from reading the cortical column network's trained internal signals — neuromodulatory levels, multi-scale prediction precision, conflict signals, working memory confidence, hippocampal retrieval

confidence, and cerebellar temporal prediction error — and converting them into biologically-motivated control of the decision process. This is in contrast to standard decoding, which requires manually-tuned temperature, top-k, top-p, repetition penalty, and other hyperparameters.

5.2. Signal Extraction from the Cortical Column Network

Before each production step, the system extracts at least seventeen scalar signals from the cortical column network's last forward pass (the signal extraction module is shown in FIG. P6-1, Figure 5, block (a)). These signals are grouped into five categories:

5.2.1. Multi-Scale Prediction Precision (5 signals)

The cortical column network computes per-token prediction precision σ^2 at multiple temporal scales through its precision-weighted predictive coding (PWPC) system:

- σ^2_{token}: token-level precision (base uncertainty at the current position);
- σ^2_{phrase}: phrase-level precision (syntactic uncertainty within the current phrase);
- $\sigma^2_{\text{sentence}}$: sentence-level precision (uncertainty about current sentence structure);
- $\sigma^2_{\text{discourse}}$: discourse-level precision (uncertainty about topic and narrative structure);
- σ^2_{std}: standard deviation of precision across cortical columns (inter-column disagreement).

5.2.2. Neuromodulatory Levels (4 signals)

The brainstem module produces four neuromodulatory signals, each in the range $(0, 1)$ via sigmoid activation:

- $\text{DA}_{\text{phasic}}$: phasic dopamine (reward prediction error — recent token was rewarding);
- NE: norepinephrine (arousal and exploration breadth);
- 5-HT: serotonin (temporal patience and time-horizon);
- ACh: acetylcholine (attentional focus and logit sharpness).

Additionally, tonic dopamine (DA_{tonic}) is read from the brainstem's persistent buffer, representing the baseline reward expectation.

5.2.3. Metacognitive Signals (3 signals)

- PC steps: the number of predictive coding inference iterations that were required for the last forward pass to converge (higher = more difficult input);
- Convergence rate: the rate at which predictive coding settled (low convergence = ambiguous input);
- Mean prediction error: the average prediction error across all columns (surprise level).

5.2.4. Subcortical Signals (2 signals)

- CB loss: cerebellar temporal prediction error (the cerebellum's estimate of how far the temporal pattern deviated from expectation);
- ACC conflict: anterior cingulate cortex conflict signal (inter-column disagreement measured by variance in column outputs).

5.2.5. Memory Confidence Signals (5 signals)

- WM confidence: overall working memory retrieval confidence;
- NMDA confidence: confidence specifically in NMDA attractor slot contents (persistent facts);
- Silent confidence: confidence in activity-silent synaptic trace contents (recent context);
- HPC retrieval confidence: hippocampal episodic retrieval confidence (strength of matching episode);
- Concept confidence: hippocampal concept-level confidence (consolidation strength of the retrieved concept).

Additionally, a **curiosity signal** is computed as:

$$\text{curiosity} = \bar{\sigma^2} \times (1 - c_{\text{concept}}) \tag{1}$$

where $\bar{\sigma^2}$ is the mean prediction precision and c_{concept} is the concept confidence. High curiosity indicates uncertainty about unknown concepts, driving exploratory token selection.

5.3. Stage 1: Conceptualise — Discourse Guidance

The first stage implements the Levelt Conceptualiser, which determines "what to say" before grammatical encoding begins (see FIG. P6-16, Figure 20 for the NMDA novelty suppression detail).

5.3.1. Global Workspace Guidance

The cortical column network's global workspace module has already performed surprise-weighted broadcast competition during the forward pass, selecting the most salient column content for global distribution. This workspace influence is inherently present in the logits. The Conceptualise stage adds a complementary **NMDA novelty bias**.

5.3.2. NMDA Novelty Suppression

The NMDA attractor slots in working memory hold entities and topics that have already been "discussed" in the current generation context. The system reads the NMDA state, projects it through the column's L4 projection and the language model head to obtain logit-space representations of "already said" content, and *suppresses* these logits to favour novel content:

$$\text{logits}_{\text{novel}} = \text{logits} - \alpha_{\text{novelty}} \cdot \text{sigmoid}\big(\text{LMHead}\big(\text{L4}^{-1}(\text{NMDA_read}(\bar{s}_{\text{NMDA}}))\big)\big) \quad (2)$$

where $\bar{s}_{\text{NMDA}}$ is the mean NMDA slot content, L4^{-1} is the approximate inverse projection from column space to model space, and α_{novelty} is a composite scaling factor computed as:

$$\alpha_{\text{novelty}} = 0.3 \times c_{\text{NMDA}} \times (1 - 0.3 \cdot \min(\sigma^2_{\text{discourse}}, 1)) \times (1 - 0.2 \cdot c_{\text{HPC}}) \times (0.5 + 0.5 \cdot c_{\text{concept}}) \quad (3)$$

This formulation ensures that: NMDA confidence gates suppression strength (fading slots suppress less); high discourse uncertainty reduces suppression (novel tokens welcome when discourse is uncertain); strong episodic retrieval reduces novelty bias (familiar territory needs less novelty); and high concept confidence strengthens suppression of already-known content.

Disease mapping: Removal of the NMDA novelty bias produces confabulation (the system repeats what it has already said without novelty pressure, analogous to frontal confabulation). Removal of workspace guidance produces Wernicke's aphasia (fluent but meaningless output).

5.4. Stage 2: Formulate — Syntactic Planning

The second stage implements the Levelt Formulator, which builds grammatical frames and selects lexical items to fill them (see FIG. P6-17, Figure 21 for reward-modulated biasing; FIG. P6-18, Figure 22 for schema retrieval; and FIG. P6-19, Figure 23 for episodic production recall).

5.4.1. L2/3 Spreading Activation

The Hebbian association weights in Layer 2/3 of the cortical columns hold recently-activated patterns that prime related lexical items (Dell 1986). The system projects the last column's L2/3 output through the LM head to produce a soft lexical bias toward associated tokens:

$$\text{logits}_{\text{spread}} = \text{logits} + \alpha_{\text{L23}} \cdot \tanh\big(0.5 \cdot \text{LMHead}\big(\text{L4}^{-1}(h_{\text{L23}})\big)\big) \quad (4)$$

where $\alpha_{\text{L23}} = 0.2 \times (0.5 + 0.5 \cdot \text{DA}_{\text{phasic}}) \times (0.5 + 0.5 \cdot c_{\text{WM}})$.

The scaling by phasic dopamine implements the finding of Crosson et al. (2007) that reward prediction error strengthens all production biases. The scaling by WM confidence implements Baddeley's (1992) observation that rehearsal confidence modulates output strength.

5.4.2. Phonological Loop Planning

The phonological loop in working memory holds a rehearsed sequence representing the "upcoming plan." The system reads the phonological loop content with recency-weighted integration (later slots = more recent plan), projects through the LM head, and adds a soft planning bias:

$$\text{logits}_{\text{plan}} = \text{logits}_{\text{spread}} + \alpha_{\text{phono}} \cdot \tanh\big(0.5 \cdot \text{LMHead}\big(\text{L4}^{-1}(\text{phono_read}(\bar{p}))\big)\big) \tag{5}$$

where $\bar{p} = \sum_k (k/K) \cdot p_k$ is the recency-weighted phonological content and $\alpha_{\text{phono}} = 0.15 \times (0.5 + 0.5 \cdot c_{\text{WM}}) \times (0.5 + 0.5 \cdot \text{DA}_{\text{phasic}})$.

Disease mapping: Removal of phonological planning produces agrammatic aphasia (telegraphic speech without grammatical structure). Removal of L2/3 spreading produces anomia (word-finding difficulty).

5.4.3. Schema-Guided Generation

The system retrieves matching abstract structural templates (schemas) from the hippocampal schema buffer (Tse et al. 2007, Gilboa & Marlatt 2017). Each schema carries a structural pattern and a similarity/confidence score. The schema pattern is projected through the LM head and added as a soft bias:

$$\text{logits}_{\text{schema}} = \text{logits}_{\text{plan}} + \sum_{i=1}^{2} \alpha_{\text{schema},i} \cdot \tanh(0.5 \cdot \text{LMHead}(\text{schema}_i)) \tag{6}$$

where $\alpha_{\text{schema},i} = 0.1 \times s_i \times (0.5 + 0.5 \cdot c_{\text{schema},i}) \times (1 + 0.15 \cdot c_{\text{HPC}}) \times (0.5 + 0.5 \cdot \text{DA}_{\text{phasic}})$ and s_i is the cosine similarity of the schema to the current context.

5.4.4. Episodic Production Retrieval

The hippocampus searches its memory buffer for past production episodes (flagged with `is_production=True`) that match the current context. Matching episodes carry prior production patterns — "what was produced in a similar context before" — providing experience-based logit bias. The match strength must exceed a minimum similarity threshold (0.3) to be applied.

Disease mapping: Removal of episodic retrieval produces de novo production (no benefit from conversational experience, analogous to amnesia in production planning — Moscovitch 2008).

5.5. Stage 3: Basal Ganglia Action Selection

The basal ganglia module computes go/no-go signals via D1 (direct) and D2 (indirect) pathway competition (Frank & Badre 2012), as illustrated in FIG. P6-30 (Figure 34). The go signal from the discourse and sentence gates is mapped to a drift rate modifier:

$$m_{\text{BG}} = 0.8 + 0.4 \times \frac{g_{\text{discourse}} + g_{\text{sentence}}}{2} \tag{7}$$

where $g_{\text{discourse}}$ and g_{sentence} are the basal ganglia gate values in $[0, 1]$. The logits are scaled by this modifier before entering the DDM, so that a strong go signal (high confidence in action plan) boosts the drift rate for faster decisions, while a no-go signal (action conflict) reduces the drift rate for more deliberation.

Disease mapping: Removal of BG→DDM coupling produces akinesia (decisions ignore action selection state) or impulsivity (no no-go signal to slow decisions), analogous to Parkinson's and Huntington's diseases respectively.

5.6. Stage 4: Decide — Drift-Diffusion Evidence Accumulation

The core of the production system is a drift-diffusion model (DDM) that accumulates evidence for candidate tokens over discrete time steps (see Figure 2 for the complete signal-to-DDM mapping; FIG. P6-1, Figure 5, block (b) for the DDM module; FIG. P6-14, Figure 18 for threshold mapping; FIG. P6-15, Figure 19 for max steps mapping; and FIG. P6-28, Figure 32 for noise sigma construction). The DDM has **zero learnable parameters**; all control arises from the seventeen extracted brain signals mapped to six DDM parameters.

5.6.1. Signal-to-Parameter Mapping

The mapping from brain signals to DDM parameters is as follows:

Candidate Count K

$$K = \min(\lfloor K_{\text{base}} \times (0.5 + \text{NE}_{\text{eff}}) \times (1 + \max(0.5 - r_{\text{conv}}, 0)) \rfloor, V) \tag{8}$$

where $\text{NE}_{\text{eff}} = \min(\text{NE} + 0.15 \times c_{\text{eff}}, 1.0)$, $c_{\text{eff}} = \min(\text{curiosity}, 3.0)$, r_{conv} is the convergence rate, and K is floored at 2. High norepinephrine (arousal) widens the candidate pool (Aston-Jones & Cohen 2005). Low convergence rate (ambiguous input) further expands candidates. Curiosity boosts effective NE.

Initial Threshold θ_0

$$\begin{aligned}\theta_0 = \theta_{\text{base}} &\times (1 + \beta_{\text{DA}} \times (0.5 - \text{DA}_{\text{tonic}})) \\ &\times \max(0.8,\ 1 - 0.07 \times c_{\text{eff}}) \\ &\times (1 + 0.3 \times \max(\text{MPE} - 0.5, 0)) \\ &\times (1 - 0.15 \times c_{\text{HPC}}) \\ &\times (1 + 0.1 \times c_{\text{NMDA}}) \\ &\times (1 - 0.1 \times c_{\text{concept}}) \\ &\times (1 + 0.2 \times \text{conflict})\end{aligned} \tag{9}$$

where β_{DA} is the DA coupling strength, MPE is mean prediction error, and $\theta_0 \geq 0.1$. This implements: high tonic DA lowers threshold (Schultz 1997); curiosity lowers threshold (explore); high surprise raises threshold (caution); strong episodic match lowers threshold (familiar); strong NMDA raises threshold (protect stored facts — Goldman-Rakic 1995); high concept confidence lowers threshold (commit faster on known territory — Kumaran & McClelland 2012); and high conflict raises threshold (Botvinick 2001 conflict monitoring).

Diffusion Noise σ

$$\begin{aligned}\sigma = \sigma_{\text{base}} &\times \sqrt{\max(\sigma^2_{\text{token}} \times s_{\sigma^2}, 0.01)} \\ &+ 0.1 \times \max(\sigma^2_{\text{phrase}} - \sigma^2_{\text{token}}, 0) \\ &+ 0.15 \times \sigma^2_{\text{std}}\end{aligned} \tag{10}$$

followed by:

$$\sigma \leftarrow \sigma \times (1 - 0.1 \times c_{\text{HPC}}) \times (1 - 0.1 \times c_{\text{silent}}) \times (1 + 0.2 \times \max(0.5 - c_{\text{concept}}, 0)) \tag{11}$$

This maps: token precision to base noise (PWPC); phrase excess to syntactic uncertainty noise; inter-column disagreement to mismatch noise; episodic confidence reduces noise (Eichenbaum 2004); silent trace confidence reduces noise (Mongillo 2008 context stability); and low concept confidence increases noise.

Urgency Time Constant τ

$$\tau_{\text{eff}} = \tau_{\text{base}} \times (0.5 + \text{5-HT}) \tag{12}$$

with $\tau_{\text{eff}} \leftarrow \tau_{\text{eff}} \times \max(0.5,\ 1 - 0.2 \times (\sigma^2_{\text{sentence}} - 1.5))$ when $\sigma^2_{\text{sentence}} > 1.5$. High serotonin = patience = slower threshold collapse (Daw et al. 2002). High sentence uncertainty speeds urgency (simplify/shorten).

Maximum Deliberation Steps M

$$M = \min(M_{\text{base}} + \lfloor \text{conflict} \times s_c + \min(\text{CB}, 3) \times 3 + \Delta_{\text{PC}} \times 2 + \Delta_{\text{WM}} + \Delta_{\text{sent}} \rfloor,\ 50) \quad (13)$$

where $\Delta_{\text{PC}} = \max(\text{PC steps} - 2, 0)$, $\Delta_{\text{WM}} = \max(0.5 - c_{\text{WM}}, 0) \times 6$, and $\Delta_{\text{sent}} = \min(\lfloor 3 \times \sigma^2_{\text{sentence}} \rfloor, 5)$. Five independent difficulty signals extend deliberation: ACC conflict, cerebellar temporal error, predictive coding settling difficulty, working memory uncertainty, and sentence-level syntactic uncertainty.

Drift Rate d

$$d_i = \frac{\text{logits}_i \times (0.8 + 0.4 \times \text{DA}_{\text{phasic}})}{\max(\sigma \times 5, 0.1)} \quad (14)$$

High phasic DA (recent reward) boosts drift rate (Crosson et al. 2007), implementing reward-guided action selection in the basal ganglia.

ACh Logit Sharpness Prior to candidate selection, logits are scaled by an ACh-derived sharpness factor (see FIG. P6-4, Figure 8):

$$\text{logits}_{\text{sharp}} = \text{logits} \times (0.4 + 1.2 \times \text{ACh}) \quad (15)$$

implementing Hasselmo's (2006) finding that acetylcholine modulates attentional focus. High ACh sharpens the distribution (focused attention); low ACh flattens it (diffuse attention).

5.6.2. Evidence Accumulation Process

For each batch element routed to the System 2 (deliberative) path, the DDM initialises K accumulators to zero and iterates:

$$a_i(t+1) = a_i(t) + d_i + \sigma \cdot \xi_i(t), \quad \xi_i(t) \sim \mathcal{N}(0,1) \quad (16)$$

$$\theta(t) = \theta_0 \cdot \exp(-t/\tau_{\text{eff}}) \quad (17)$$

At each step t, if $\max_i a_i(t) > \theta(t)$ for any undecided batch element, the corresponding candidate index $i^* = \arg\max_i a_i(t)$ is selected and that element is marked as decided. If all elements are decided or t reaches M, the loop terminates. Undecided elements at step M are force-decided by selecting the candidate with maximum accumulated evidence.

5.6.3. Confidence-Gated Dual-Process Deliberation

Before entering the DDM loop, a confidence score is computed for each batch element (see FIG. P6-13, Figure 17 for the dual-process architecture):

$$\text{confidence} = \frac{\text{logits}_{(1)} - \text{logits}_{(2)}}{\max(\sigma, 0.01)} \tag{18}$$

where $\text{logits}_{(1)}$ and $\text{logits}_{(2)}$ are the top two logit values among the K candidates. If confidence $>$ c_{thresh} (default 2.5), the element takes the **System 1 (fast)** path: the top candidate is selected immediately without evidence accumulation. Otherwise, the element takes the **System 2 (slow)** path: the full DDM loop is executed.

This implements Kahneman's (2011) dual-process framework: when the evidence clearly favours one candidate, the brain commits instantly (System 1); when evidence is ambiguous, the brain engages deliberate reasoning (System 2).

Disease mapping: A too-low confidence threshold produces impulsive generation (everything routed to System 1, analogous to frontal disinhibition). A too-high threshold produces pathological deliberation (everything routed to System 2, analogous to OCD).

5.7. Stage 5: Monitor — Self-Check with WM Checkpoint/Restore

The self-monitoring stage implements Levelt's (1983) inner monitoring loop, wherein the speaker processes their planned output through the comprehension system before committing to articulation (see FIG. P6-9, Figure 13 for the adaptive rejection threshold; FIG. P6-10, Figure 14 for the iterative resample loop).

5.7.1. Working Memory Checkpoint

Before verification, the system saves the complete working memory state for all cortical columns:

- NMDA attractor slot tensors (persistent facts);
- Activity-silent trace variables u and x (Tsodyks-Markram dynamics);
- Phonological loop state tensor and write pointer;
- L4 generation cache (incremental context window).

This checkpoint captures the model's internal state as it existed *before* the candidate token was processed.

5.7.2. Comprehension Verification

The selected token is fed through the model's forward path (via `generate_step`) as if it were input, causing the model to "comprehend" its own planned output. The prediction error from

this comprehension pass is compared against the baseline prediction error:

$$r_{\text{PE}} = \frac{\text{PE}_{\text{verify}}}{\max(\text{PE}_{\text{baseline}}, 10^{-6})} \tag{19}$$

5.7.3. Adaptive Rejection Threshold

The rejection threshold is modulated by discourse-level precision:

$$\theta_{\text{reject}} = \theta_{\text{base}} \times (1 + 0.3 \times \min(\sigma^2_{\text{discourse}}, 2.0)) \tag{20}$$

This implements Botvinick's (2004) ACC conflict monitoring: in uncertain discourse contexts, the monitoring system is more tolerant of prediction error increases (minor errors are expected when the discourse is uncertain).

5.7.4. Accept or Reject with Restore

If $r_{\text{PE}} \leq \theta_{\text{reject}}$: the token is **accepted**. The verification forward pass has already updated the model's internal state (WM, caches) to reflect the accepted token — no further action needed.

If $r_{\text{PE}} > \theta_{\text{reject}}$: the token is **rejected**. The system:

1. Restores all saved working memory state (NMDA, silent trace, phonological loop, generation cache) from the checkpoint;
2. Suppresses the rejected token's logit by a large margin (5.0);
3. Re-samples using the DDM with the modified logits.

This save-verify-accept/reject cycle repeats up to R_{max} times (default 3). If all resamples are rejected, the last candidate is force-committed.

Disease mapping: Removal of self-monitoring produces jargon aphasia (the model produces errors without any awareness or correction ability). Removal of WM restore produces state corruption (rejected tokens leave residual WM contamination, causing subsequent tokens to lose coherence).

5.8. Stage 6: Produce — IOR Update and Token Commitment

Upon acceptance (or force-commitment), the token is committed to the output sequence (see Figure 4 for the IOR overview; FIG. P6-6, Figure 10 for the detailed IOR module; FIG. P6-20, Figure 24 for ACh/NE decay modulation). The inhibition of return module is updated:

$$\text{IOR}_w \leftarrow A_{\text{eff}} \tag{21}$$

for the produced token w, and all existing traces decay:

$$\text{IOR}_v \leftarrow \text{IOR}_v \times \exp(-1/\tau_{\text{eff}}) \tag{22}$$

for all vocabulary items v.

5.8.1. Neuromodulatory-Adaptive Decay

The effective decay time constant is modulated by ACh and NE:

$$\tau_{\text{eff}} = \tau_{\text{base}} \times (0.5 + \text{ACh}) \times (1.5 - \text{NE}) \tag{23}$$

with $\tau_{\text{eff}} \geq 1.0$. High ACh (focused attention) shortens IOR (overrides inhibition for attended items — Hasselmo 2006). High NE (arousal/urgency) shortens IOR (urgency reduces inhibition). The product $(0.5 + \text{ACh})(1.5 - \text{NE})$ ranges from $[0.25, 2.0]$.

5.8.2. Discourse-Modulated Strength

The inhibition strength is increased when discourse precision is high (discourse representation uncertain):

$$A_{\text{eff}} = A_{\text{base}} \times \min(2.0,\ 1 + 0.3 \times (\sigma^2_{\text{discourse}} - 1.5)) \quad \text{when } \sigma^2_{\text{discourse}} > 1.5 \tag{24}$$

When the topic representation is uncertain, stronger IOR suppresses off-topic tokens and maintains coherence.

Disease mapping: ACh deficit (Alzheimer's) $\rightarrow \tau$ too long $\rightarrow$ anomia (can't reuse words). NE deficit (depression) $\rightarrow \tau$ too persistent $\rightarrow$ impoverished vocabulary. Too-weak IOR $\rightarrow$ OCD-like perseveration.

5.9. Stage 6b: Comprehend — Unconditional Self-Feedback

After token commitment, the produced token is unconditionally fed through the comprehension system (via `generate_step`) to update the model's internal state (see FIG. P6-11, Figure 15). This is distinct from the monitoring verification in Stage 5: monitoring is conditional and may be disabled, while the comprehend stage is unconditional.

When self-monitoring is enabled and the token was accepted (not force-committed), the verification forward pass from Stage 5 already served as the comprehension pass. In all other cases (monitoring disabled, or token force-committed after exhausting resamples), the comprehend stage runs an explicit forward pass.

This implements Pickering & Garrod's (2013) production-comprehension coupling: after producing a word, the speaker always processes their own output to maintain coherent internal state.

Disease mapping: Removal of unconditional comprehension produces jargon aphasia with WM divergence (the speaker is unaware of their own output, and working memory diverges from actual production, causing cascade failure in subsequent tokens).

5.10. Stage 7: Competitor Suppression

After token commitment, the DDM's rejected candidates receive inhibition through the IOR trace (see FIG. P6-7, Figure 11):

$$\text{IOR}_v \leftarrow \max(\text{IOR}_v,\ 0.3 \times A_{\text{base}}) \quad \text{for each rejected candidate } v \in \text{top-}K \setminus \{w^*\} \tag{25}$$

where w^* is the selected token. The suppression strength is 30% of the base IOR strength and is applied to up to 8 top candidates.

This implements Oppenheim et al.'s (2010) "dark side of repetition priming": producing one word *weakens* the pathway to semantic competitors. The mechanism is short-term activation suppression via the IOR trace (not synaptic weight change). This makes future retrieval more decisive by reducing competition from recently-rejected alternatives.

Disease mapping: Removal of competitor suppression produces semantic interference (naming "dog" primes "cat" equally, causing word-finding delays and tip-of-the-tongue states).

5.11. Stage 8: Remember — Online Episodic Encoding

The system stores production events in the hippocampal episodic memory for later consolidation (see FIG. P6-12, Figure 16 for the encoding module). Each event is tagged with `is_production=True` to distinguish production episodes from comprehension episodes. The surprise signal from the self-monitoring prediction error determines encoding priority (Lisman & Grace 2005): high-PE events are more strongly encoded.

Critically, this is **episodic encoding, not synaptic weight change**. Hebbian LTP requires protein synthesis (hours). Predictive coding weight updates require sustained error signals (training). What happens during conversation is: (a) hippocampal encoding of surprising events; (b) working memory maintenance of active items; and (c) short-term facilitation via the silent trace mechanism (Tsodyks-Markram). Permanent synaptic weight changes occur during sleep when hippocampal replay consolidates episodic memories into cortical weights.

5.12. Stage 8b: SWR Phase-Locked Micro-Consolidation

During generation, the hippocampus performs micro-consolidation — compressed replay that transfers episodic memories into cortical weights — gated by the sharp wave ripple (SWR) oscillation phase (see FIG. P6-8, Figure 12 and FIG. P6-29, Figure 33).

5.12.1. SWR Phase Gate

The system computes a soft SWR gate from the theta oscillation phase of the first cortical column's oscillation module:

$$\phi_\theta = \frac{2\pi \times \text{step}}{P_\theta} \tag{26}$$

where $P_\theta = \text{softplus}(\text{period_theta_raw})$ is the learned theta period, and:

$$g_{\text{SWR}} = 0.5 \times (1 - \cos(\phi_\theta)) \tag{27}$$

The gate g_{SWR} varies from 0 (at theta peak, when gamma is suppressed — encoding mode) to 1 (at theta trough, when gamma power is high — SWR-like state, replay mode). Micro-consolidation proceeds only when $g_{\text{SWR}} > 0.3$.

This implements Buzsáki's (2015) finding that sharp wave ripples occur at theta troughs, when CA3→CA1 fires compressed replay sequences. The phase-locking ensures that consolidation and production do not interfere.

Disease mapping: Removal of SWR gating produces consolidation-production interference (replay disrupts ongoing generation, analogous to involuntary memory intrusions in PTSD). Removal of micro-consolidation entirely prevents the model from improving during extended use.

5.13. Curiosity-Driven Exploration

The curiosity signal modulates three DDM parameters to promote exploratory token selection when the model encounters uncertain and novel content (see FIG. P6-5, Figure 9):

1. **Effective NE**: $\text{NE}_{\text{eff}} = \min(\text{NE} + 0.15 \times c_{\text{eff}}, 1.0)$ — curious state widens attentional spotlight via LC-NE phasic bursts (Aston-Jones & Cohen 2005);
2. **Threshold reduction**: $\theta_0 \leftarrow \theta_0 \times \max(0.8, 1 - 0.07 \times c_{\text{eff}})$ — up to 20% reduction at maximum curiosity;
3. **Noise increase**: indirectly via NE effect on candidate count K (more candidates = more noise in competition).

Disease mapping: No curiosity modulation → depression-like (no exploration drive, stuck in familiar patterns). Over-active curiosity → ADHD (can't commit to any single path, constantly jumping to novel tokens).

5.14. Speculative Decoding via Prospective Hippocampus

The system optionally performs speculative multi-token decoding using the prospective hippocampal head (ProspectiveHPC), which maintains learned projection matrices $\{\mathbf{W}^{(k)}_{\text{prosp}}\}_{k=1}^{S}$ that predict future hidden states k steps ahead (see FIG. P6-21, Figure 25).

5.14.1. Draft Phase

The system generates S draft tokens by:

1. Computing $X^{(k)}_{\text{pred}} = \mathbf{W}^{(k)}_{\text{prosp}}(\text{norm}(h_{\text{HPC}}))$ for each look-ahead position k;
2. Projecting through the LM head: $\text{logits}^{(k)} = \text{LMHead}(\text{norm}(X^{(k)}_{\text{pred}}))$;
3. Sampling via the DDM with current brain signals and IOR trace.

The IOR state is saved before drafting and restored after, so draft tokens do not permanently affect the IOR trace.

5.14.2. Verification Phase

The draft token sequence is concatenated with the prompt and verified against the full model in a single forward pass. For each draft position k, the verification probability $p_{\text{verify}}(w_k)$ is compared against a threshold p_{accept}/V. Tokens are accepted greedily from position 1 until the first rejection; all accepted tokens are committed as a batch.

This mimics the brain's phrase-level planning: the articulatory system prepares entire phrases before execution (Levelt 1989). The hippocampus-based look-ahead leverages the system's predictive capacity for faster generation.

Disease mapping: Removal of speculative decoding reduces generation throughput but does not affect quality (analogous to cautious word-by-word speech in unfamiliar contexts).

5.15. Competitive Lexical Production Head (BrainLMHead)

The language model head replaces the standard linear projection with a three-stage competitive lexical production system based on the WEAVER++ model (Levelt et al. 1999, Roelofs 1997), as illustrated in Figure 3 for the general overview, FIG. P6-25 (Figure 29) for the full system, FIG. P6-26 (Figure 30) for the LoRA delta, and FIG. P6-27 (Figure 31) for the lateral inhibition dynamics.

5.15.1. Stage 1: Lexical Activation with Production Delta

The base logits are computed from the shared embedding weight matrix, augmented with a low-rank production delta:

$$\text{logits} = h \cdot \mathbf{W}^{\top}_{\text{embed}} + (h \cdot \Delta\mathbf{B}^{\top}) \cdot \Delta\mathbf{A}^{\top} + b_{\text{freq}} \tag{28}$$

where $\Delta\mathbf{A} \in \mathbb{R}^{V \times r}$ and $\Delta\mathbf{B} \in \mathbb{R}^{r \times d}$ form a rank-r (default 32) LoRA-style production delta, and $b_{\text{freq}} \in \mathbb{R}^{V}$ is a per-token frequency-adaptive output bias.

The production delta models the divergence between comprehension and production pathways: the brain uses partially different circuits for recognising and producing words (production is

more effortful and selective than comprehension). The LoRA factorisation keeps the additional parameter cost at $O(Vr + rd)$ rather than $O(Vd)$.

Identity at init: $\Delta\mathbf{A} = \mathbf{0}$, $b_{\text{freq}} = \mathbf{0}$, so the head degrades to standard linear projection before training.

5.15.2. Stage 2: Competitive Lateral Inhibition

A semantic projection maps the embedding matrix to a low-dimensional competition space:

$$\mathbf{S} = \mathbf{W}_{\text{embed}} \cdot \mathbf{W}_{\text{sem}}^{\top} \in \mathbb{R}^{V \times r_s} \tag{29}$$

where r_s is the competition rank (default 64). The current probability distribution (detached from gradient) computes a semantic centroid:

$$\text{centroid} = \text{softmax}(\text{logits}) \cdot \mathbf{S} \in \mathbb{R}^{B \times T \times r_s} \tag{30}$$

Each token's similarity to the active semantic field is:

$$f_v = \text{centroid} \cdot \mathbf{S}_v^{\top} \tag{31}$$

The inhibition is computed as the field similarity minus self-similarity (removing self-inhibition so the winner gets least inhibited — winner-take-all emerges):

$$\text{inhibition}_v = \text{ReLU}(f_v - p_v \times \|\mathbf{S}_v\|^2) \tag{32}$$

The final logits are:

$$\text{logits}_{\text{final}} = \text{logits} - \alpha \times \text{inhibition} \tag{33}$$

where $\alpha = \text{sigmoid}(\text{gate_raw})$ with gate_raw $= -5.0$ at init ($\alpha \approx 0.007$, competition effectively off). As training progresses, α increases and competition dynamics emerge.

Disease mapping: Anomic aphasia if the production delta is weakened (can't select production form). Broca's aphasia if competition is disrupted (multiple lemmas activate simultaneously). Semantic paraphasia if competition fails selectively (a competitor wins over the target).

5.16. Brain Coherence Score (BCS)

The system provides a composite evaluation metric that measures generation quality through eight brain-grounded components, each mapped to a specific brain mechanism and a disease if broken (see FIG. P6-3, Figure 7 for the BCS overview; FIG. P6-22, Figure 26 for self-coherence; FIG. P6-23, Figure 27 for HPC influence; and FIG. P6-24, Figure 28 for the complete weight table):

Component	Weight	Brain Mechanism	Disease If Broken
Self-coherence	0.25	Wernicke's area	Jargon aphasia
IOR effectiveness	0.15	Prefrontal inhibition	Perseveration
Vocabulary diversity	0.15	L2/3 Hebbian competition	Anomic aphasia
Entity binding	0.15	CFC theta-gamma	Binding deficit
WM read rate	0.10	NMDA/Silent/Phonol	WM dead (pure Markov)
Boundary structure	0.10	NE + oscillations	Agrammatism
HPC influence	0.10	Episodic retrieval	Source amnesia

5.16.1. Self-Coherence via Self-Reading

The model re-reads its own generated output: the generated token sequence (excluding the last token) is fed as input, and the model attempts to predict the next token at each position. The cross-entropy loss on this self-prediction task is converted to a coherence score:

$$\text{coherence} = \max\left(0,\ 1 - \frac{L_{\text{self}}}{\log V}\right) \tag{34}$$

where L_{self} is the mean cross-entropy loss and $\log V$ is the maximum possible loss (random prediction). A model that can predict its own output has coherent internal representations; high self-prediction loss indicates jargon aphasia (surprised by own output).

This is a direct computational implementation of the Wernicke's area function: comprehension of one's own speech.

5.16.2. IOR Effectiveness

Measured as $1 - r_{3\text{gram}}$ where $r_{3\text{gram}}$ is the trigram repetition rate. Effective IOR suppresses recently-produced token patterns, reducing repetition.

5.16.3. Vocabulary Diversity

The ratio of unique tokens to total generated tokens, normalised against a target diversity of 0.5: $\min(1, d_{\text{unique}}/0.5)$. Low diversity indicates anomic aphasia (stuck on common words).

5.16.4. Entity Binding Fidelity

For tokens that appear two or more times in the generated output, the cosine similarity between the embedding at the first and last occurrence is computed. High similarity indicates that repeated tokens maintain consistent representation (the CFC theta-gamma binding mechanism is working). Low similarity indicates binding deficit (tokens lose identity across distance).

5.16.5. Working Memory Utilisation

The fraction of generation steps where working memory confidence exceeded 0.05, indicating that NMDA/Silent/Phono slots contributed to the generation. Zero utilisation means WM is dead and the model generates in a purely Markov fashion.

5.16.6. Boundary Structure

The fraction of structural tokens (punctuation, newlines) in the generated output, normalised against a target rate of 10%: $\min(1, r_{\text{punct}}/0.10)$. Absence of boundary structure indicates agrammatism (word salad without grammatical organisation).

5.16.7. HPC Context Influence

The fraction of generation steps where the hippocampus contributed a retrieval, normalised against a target rate of 10%: $\min(1, r_{\text{HPC}}/0.10)$. Zero HPC influence indicates source amnesia (no episodic memory contribution to generation).

5.16.8. Composite Score

The composite Brain Coherence Score is computed as:

$$\text{BCS} = 100 \times \sum_i w_i \times c_i \quad \in [0, 100] \tag{35}$$

where w_i are the component weights (summing to 1.0) and c_i are the normalised component scores in $[0, 1]$.

A BCS of 80–100 indicates excellent coherent generation; 60–80 is good; 40–60 is developing; 20–40 is poor; and 0–20 indicates critical failure (multiple brain mechanisms non-functional). Each component's failure corresponds to a specific neurological condition, enabling targeted diagnosis of generation deficits.

CLAIMS

We claim:

Claim 1. A computer-implemented language production system for generating sequential language output from a cortical column network, the system comprising:

(a) a signal extraction module configured to extract, from a cortical column network's forward pass, a plurality of brain signals comprising at least: multi-scale prediction precision signals at token, phrase, sentence, and discourse levels; neuromodulatory level signals for dopamine, norepinephrine, serotonin, and acetylcholine; metacognitive signals including predictive coding convergence and mean prediction error; subcortical signals including cerebellar temporal prediction error and anterior cingulate cortex conflict; and memory confidence signals including working memory confidence, NMDA attractor confidence, activity-silent trace confidence, hippocampal retrieval confidence, and concept confidence;

(b) a drift-diffusion decision module having zero learnable parameters, configured to map said plurality of brain signals into drift-diffusion parameters comprising: a candidate count K modulated by norepinephrine level and convergence rate; an initial threshold θ_0 modulated by tonic dopamine, mean prediction error, hippocampal retrieval confidence, NMDA confidence, concept confidence, and conflict; a diffusion noise σ modulated by multi-scale prediction precision, hippocampal confidence, and silent trace confidence; an urgency time constant τ modulated by serotonin level and sentence-level precision; a maximum deliberation step count modulated by conflict, cerebellar loss, predictive coding steps, working memory confidence, and sentence-level precision; and a drift rate modulated by phasic dopamine; wherein evidence for competing token candidates is accumulated stochastically over discrete time steps according to $a_i(t{+}1) = a_i(t) + d_i + \sigma \cdot \xi_i(t)$ under a collapsing threshold $\theta(t) = \theta_0 \cdot \exp(-t/\tau)$, and a token is selected when its accumulated evidence exceeds the threshold;

(c) a confidence-gated deliberation mechanism configured to compute a confidence score as the ratio of the logit margin between the top two candidates to the diffusion noise, and to route tokens with confidence exceeding a threshold to a fast selection path that selects the top candidate without evidence accumulation, and tokens with confidence below the threshold to a deliberative path that executes the drift-diffusion evidence accumulation process; and

(d) a self-monitoring module configured to, before committing a selected token: save a checkpoint of the working memory state comprising NMDA attractor slot tensors, activity-silent trace variables, phonological loop state and write pointer, and generation cache; feed the selected token through the cortical column network's forward path to compute a verification prediction error; compare a prediction error ratio against an adaptive rejection threshold;

upon acceptance, retain the updated working memory state; and upon rejection, restore the saved working memory state from the checkpoint, suppress the rejected token, and re-sample via the drift-diffusion decision module.

Claim 2. A computer-implemented method for generating sequential language output from a cortical column network, the method comprising, at each generation step:

(a) extracting a plurality of brain signals from the cortical column network's forward pass, said signals comprising at least neuromodulatory levels, multi-scale prediction precision, conflict, and memory confidence signals;

(b) conceptualising discourse intent by reading NMDA attractor slot contents from working memory, projecting said contents through a language model head to obtain logit-space representations of previously-expressed content, and suppressing said representations in the current logits to favour novel content, the suppression strength being modulated by NMDA confidence, discourse-level precision, hippocampal retrieval confidence, and concept confidence;

(c) formulating syntactic structure by: projecting L2/3 Hebbian association layer output through the language model head to produce a spreading activation bias; reading phonological loop contents with recency weighting and projecting through the language model head to produce a syntactic planning bias; retrieving matching schemas from hippocampal schema storage and projecting through the language model head; and retrieving matching production episodes from hippocampal episodic memory;

(d) computing basal ganglia go/no-go signals via D1 direct pathway and D2 indirect pathway competition, and mapping the resulting gate values to a drift rate modifier;

(e) deciding a token via drift-diffusion evidence accumulation using the extracted brain signals, with a confidence gate routing high-confidence selections to a fast path and low-confidence selections to a deliberative evidence accumulation path;

(f) monitoring the selected token by saving working memory state, feeding the token through the cortical column network's comprehension path, comparing the resulting prediction error ratio against an adaptive threshold modulated by discourse-level precision, and upon rejection restoring the saved working memory state and re-sampling;

(g) producing the token by committing it to the output sequence and updating an inhibition of return trace with neuromodulatory-adaptive decay modulated by acetylcholine and norepinephrine;

(h) suppressing competitor tokens by applying partial inhibition to drift-diffusion rejected candidates via the inhibition of return trace; and

(i) encoding the production event in hippocampal episodic memory with a production flag, with hippocampal micro-consolidation gated by theta oscillation phase according to a sharp wave ripple phase gate $g_{\text{SWR}} = 0.5 \times (1 - \cos(\phi_\theta))$.

Claim 3. A computer-implemented method for evaluating brain-grounded generation quality of a cortical column network, the method comprising:

(a) generating a sequence of tokens from the cortical column network while collecting internal brain signals at each generation step;

(b) computing a self-coherence score by re-feeding the generated token sequence through the cortical column network's forward path and measuring the cross-entropy loss on self-prediction, wherein coherence equals $\max(0, 1 - L_{\text{self}}/\log V)$ where L_{self} is the self-prediction loss and $V$is the vocabulary size;

(c) computing an inhibition of return effectiveness score based on n-gram repetition rates in the generated sequence;

(d) computing a vocabulary diversity score as the ratio of unique tokens to total tokens normalised against a target diversity level;

(e) computing an entity binding fidelity score by measuring cosine similarity between embeddings of repeated tokens at their first and last occurrences;

(f) computing a working memory utilisation score as the fraction of generation steps where working memory confidence exceeded a minimum threshold;

(g) computing a boundary structure score based on the fraction of structural tokens normalised against a target rate; and

(h) computing a composite Brain Coherence Score as a weighted sum of said component scores, each component weight corresponding to the relative importance of the associated brain mechanism, producing a score on a 0–100 scale wherein each component's deficiency corresponds to a specific neurological condition.

Claim 4. The system of Claim 1, wherein the drift-diffusion decision module maps acetylcholine level to a logit sharpness factor according to $s_{\text{ACh}} = 0.4 + 1.2 \times \text{ACh}$, and scales the logits by said factor prior to candidate selection, implementing attentional focus modulation.

Claim 5. The system of Claim 1, further comprising a curiosity-driven exploration module configured to compute a curiosity signal as $c = \bar{\sigma}^2 \times (1 - c_{\text{concept}})$ where $\bar{\sigma}^2$ is mean prediction precision and c_{concept} is concept confidence, and to modulate the drift-diffusion parameters by: increasing effective norepinephrine to widen the candidate pool; reducing the decision threshold to lower the commitment barrier; and increasing diffusion noise to promote exploration.

Claim 6. The system of Claim 1, further comprising a neuromodulatory-adaptive inhibition of return module configured to maintain a per-vocabulary-item inhibition trace that decays exponentially with an effective time constant $\tau_{\text{eff}} = \tau_{\text{base}} \times (0.5 + \text{ACh}) \times (1.5 - \text{NE})$, wherein acetylcholine shortens inhibition decay and norepinephrine shortens inhibition decay, and the inhibition strength is increased when discourse-level prediction precision exceeds a threshold.

Claim 7. The system of Claim 1, further comprising a competitor suppression module configured to apply partial inhibition to drift-diffusion rejected candidates via the inhibition of return trace, wherein the suppression strength is a fraction of the base inhibition strength applied to a limited number of top candidates, implementing the Oppenheim (2010) model of lexical competitor suppression.

Claim 8. The system of Claim 1, further comprising a sharp wave ripple phase-locked micro-consolidation module configured to compute a consolidation gate from the theta oscillation phase of a cortical column's oscillation module according to $g_{\text{SWR}} = 0.5 \times (1 - \cos(\phi_\theta))$ where $\phi_\theta = 2\pi \times \text{step}/P_\theta$, and to permit hippocampal micro-consolidation only when the gate exceeds a threshold, restricting consolidation to the theta trough phase.

Claim 9. The system of Claim 1, wherein the self-monitoring module employs an adaptive rejection threshold modulated by discourse-level prediction precision according to $\theta_{\text{reject}} = \theta_{\text{base}} \times (1 + 0.3 \times \min(\sigma^2_{\text{discourse}}, 2.0))$, whereby the system is more tolerant of prediction error increases in uncertain discourse contexts.

Claim 10. The system of Claim 1, wherein the self-monitoring module is configured to iterate the save-verify-reject-restore cycle up to a maximum number of resample attempts, and upon exhausting all attempts to force-commit the last sampled candidate.

Claim 11. The system of Claim 1, further comprising an unconditional comprehension module configured to, after token commitment, feed the committed token through the cortical column network's forward path to update internal state regardless of whether self-monitoring was performed, ensuring that working memory and column states track actual production output.

Claim 12. The system of Claim 1, further comprising an online episodic encoding module configured to store production events in hippocampal episodic memory with a production flag distinguishing production episodes from comprehension episodes, wherein the encoding priority is determined by the prediction error magnitude from self-monitoring.

Claim 13. The system of Claim 1, wherein the confidence-gated deliberation mechanism implements a dual-process architecture wherein: a System 1 fast path selects the top candidate when $(\text{logits}_{(1)} - \text{logits}_{(2)})/\sigma > c_{\text{thresh}}$; and a System 2 deliberative path executes the full drift-diffusion accumulation loop otherwise.

Claim 14. The system of Claim 1, wherein the drift-diffusion decision module maps the plurality of brain signals to the initial threshold θ_0 through multiplicative modulation by at least seven

factors: tonic dopamine level, curiosity signal, mean prediction error, hippocampal retrieval confidence, NMDA confidence, concept confidence, and anterior cingulate cortex conflict.

Claim 15. The system of Claim 1, wherein the drift-diffusion decision module maps the plurality of brain signals to the maximum deliberation step count through additive contributions from at least five independent difficulty signals: anterior cingulate cortex conflict, cerebellar temporal prediction error, predictive coding settling difficulty, working memory uncertainty, and sentence-level syntactic uncertainty, with the total capped at a maximum value.

Claim 16. The method of Claim 2, wherein the conceptualising step further comprises computing an NMDA novelty suppression strength as a product of NMDA confidence, a discourse uncertainty scaling factor, a hippocampal retrieval scaling factor, and a concept confidence scaling factor, such that: fading NMDA slots suppress less; high discourse uncertainty reduces suppression; strong episodic retrieval reduces suppression; and high concept confidence strengthens suppression.

Claim 17. The method of Claim 2, wherein the formulating step further comprises scaling the spreading activation bias and phonological planning bias by phasic dopamine and working memory confidence, implementing reward-modulated production biasing.

Claim 18. The method of Claim 2, wherein the formulating step further comprises retrieving abstract structural schemas from hippocampal schema storage, projecting said schemas through the language model head, and scaling the resulting bias by schema similarity, schema confidence, hippocampal retrieval confidence, and phasic dopamine.

Claim 19. The method of Claim 2, wherein the formulating step further comprises searching hippocampal episodic memory for past production episodes matching the current context above a similarity threshold, and applying a logit bias based on the matching episode's production pattern, implementing experience-based production planning.

Claim 20. The method of Claim 2, wherein the producing step further comprises modulating the inhibition of return decay time constant by acetylcholine according to a multiplicative factor $(0.5+\text{ACh})$ and by norepinephrine according to a multiplicative factor $(1.5-\text{NE})$, and modulating the inhibition strength by discourse-level prediction precision when said precision exceeds a threshold.

Claim 21. The method of Claim 2, further comprising, between generation steps, performing speculative multi-token decoding by: generating draft tokens using prospective hippocampal projection matrices that predict future hidden states at multiple look-ahead positions; verifying the draft sequence against the full cortical column network in a single forward pass; and accepting draft tokens greedily from the first position until a verification probability falls below a threshold.

Claim 22. The method of Claim 3, wherein the self-coherence score computation further comprises: feeding the generated token sequence excluding the last token as input to the cortical column network; computing a forward pass with the generated token sequence shifted by one

position as targets; obtaining the cross-entropy loss on said self-prediction task; and converting the loss to a coherence score according to coherence $= \max(0, 1 - L_{\text{self}}/\log V)$.

Claim 23. The method of Claim 3, further comprising computing a hippocampal context influence score as the fraction of generation steps where the hippocampal module contributed an episodic retrieval, normalised against a target retrieval rate, and incorporating said score into the composite Brain Coherence Score.

Claim 24. The method of Claim 3, wherein the composite Brain Coherence Score assigns weights of approximately 0.25 to self-coherence, 0.15 to inhibition of return effectiveness, 0.15 to vocabulary diversity, 0.15 to entity binding fidelity, 0.10 to working memory utilisation, 0.10 to boundary structure, and 0.10 to hippocampal context influence, and produces a diagnostic output mapping each deficient component to a corresponding neurological condition.

Claim 25. The system of Claim 1, further comprising a competitive lexical production head configured to compute logits as $h \cdot \mathbf{W}_{\text{embed}}^{\top} + (h \cdot \Delta\mathbf{B}^{\top}) \cdot \Delta\mathbf{A}^{\top} + b_{\text{freq}}$ where $\Delta\mathbf{A}$ and $\Delta\mathbf{B}$ form a low-rank production delta and b_{freq} is a per-token frequency-adaptive bias, followed by competitive lateral inhibition wherein: a semantic projection maps the embedding matrix to a competition space; a centroid of the current activation in semantic space is computed from the softmax distribution; each token's inhibition is computed as its field similarity minus self-similarity; and the inhibition is gated by a learned competition gate initialised near zero.

Claim 26. The system of Claim 24, wherein the low-rank production delta implements production-comprehension weight divergence with $\Delta\mathbf{A} \in \mathbb{R}^{V \times r}$ initialised to zero and $\Delta\mathbf{B} \in \mathbb{R}^{r \times d}$ initialised from a scaled Gaussian distribution, such that the production head is identity at initialisation (degrades to standard linear projection), and the rank r is substantially smaller than both the vocabulary size V and the model dimension d.

Claim 27. The system of Claim 24, wherein the competitive lateral inhibition computes inhibition as $\text{ReLU}(f_v - p_v \times \|\mathbf{S}_v\|^2)$ where f_v is each token's similarity to the semantic field centroid, p_v is the current probability, and $\|\mathbf{S}_v\|^2$ is the squared norm of the token's semantic projection, such that the selected winner receives the least inhibition and winner-take-all dynamics emerge.

Claim 28. The system of Claim 1, wherein the drift-diffusion decision module maps the diffusion noise σ through: a base noise derived from token-level prediction precision; an additive phrase-level excess noise when phrase-level precision exceeds token-level precision; an additive inter-column disagreement noise; a multiplicative reduction by hippocampal retrieval confidence and silent trace confidence; and a multiplicative increase when concept confidence is low.

Claim 29. The method of Claim 2, wherein the encoding step further comprises computing a sharp wave ripple phase gate from the theta oscillation phase of a cortical column's oscillation module, and permitting hippocampal micro-consolidation of production episodes only when the phase gate exceeds a threshold corresponding to the theta trough, implementing phase-locked consolidation that avoids interference with ongoing production.

Claim 30. The system of Claim 1, further comprising a basal ganglia action selection module configured to compute go and no-go gate values via D1 direct pathway and D2 indirect pathway competition on the last column output, and to map the mean of discourse-level and sentence-level gate values to a drift rate modifier in the range $[0.8, 1.2]$ that scales the logits before drift-diffusion processing.

Claim 31. The system of Claim 1, wherein the system implements the complete Levelt speech production pipeline as eight sequential stages: Stage 1 Conceptualise providing discourse guidance via global workspace and NMDA novelty bias; Stage 2 Formulate providing syntactic planning via phonological loop, L2/3 spreading activation, schemas, and episodic retrieval; Stage 3 Basal Ganglia Action Selection providing go/no-go modulation of drift rate; Stage 4 Decide providing drift-diffusion evidence accumulation with confidence gating; Stage 5 Monitor providing self-check with working memory checkpoint and restore; Stage 6 Produce providing token commitment with neuromodulatory-adaptive inhibition of return; Stage 7 Competitor Suppress providing inhibition of rejected alternatives; and Stage 8 Remember providing episodic encoding with SWR phase-locked micro-consolidation.

ABSTRACT

A computer-implemented language production system and method for generating sequential language output from a cortical column network through a biologically-grounded eight-stage pipeline modelled on the Levelt speech production framework. The system comprises a drift-diffusion decision module with zero learnable parameters that maps at least seventeen internal brain signals — including neuromodulatory levels, multi-scale prediction precision, conflict, and memory confidence — into drift-diffusion parameters controlling candidate count, decision threshold, diffusion noise, urgency, deliberation depth, and drift rate. A confidence-gated dual-process mechanism routes high-confidence tokens to fast selection and low-confidence tokens to deliberative evidence accumulation. A self-monitoring module saves and restores working memory state during token verification. A competitive lexical production head implements WEAVER++ lateral inhibition with a low-rank production delta. Sharp wave ripple phase-locked micro-consolidation enables learning during generation. A composite Brain Coherence Score evaluates generation quality through eight brain-grounded component metrics.