International Journal of Transdisciplinary Research and Perspectives
Volume 2, Issue 2, February 2026
www.ijtrp.com | E-ISSN: 3107-7935 | RESAIM Publishing | www.resaim.com

The Artificial Brain: A Neuroscience Inspired Architecture for Multimodal AI Systems

Krrish Choudhary[1*], Tanvi Kandoi[2]
[1]Department of Computer Science, The LNM Institute of Information Technology, Jaipur, India
[2]Department of Computer Science, Indian Institute of Information Technology Tiruchirappalli, Tiruchirappalli, India

***Abstract*: Current AI models process input through a single lens. The human brain never does this—it cross-references every sense against every other sense, flags conflicts, and only calls on expensive conscious reasoning when something doesn't add up. We present a complete architecture for an artificial brain that mirrors this design: specialized Small Language Models (SLMs) as parallel sensory cortices, a lightweight conflict detector as the anterior cingulate cortex, an expensive reasoning model as the prefrontal cortex activated only on demand, a streaming identity core as the default mode network, episodic memory via vector databases, slow knowledge consolidation via sleep-cycle fine-tuning, and a neuromodulator reward system that shapes all behavior over time. The system is born when started, develops personality through experience, sleeps to consolidate, and dies when stopped. Every component maps to a specific brain structure. Every design decision is grounded in neuroscience research. The architecture is implementable today on consumer hardware (RTX 4050, 6GB VRAM).**

***Keywords*: Neuroscience-inspired AI, Multimodal Integration, Hallucination Detection, Small Language Models, Mixture of Experts, Predictive Processing.**

1. Introduction

A. Single-Source Truth Bias

Modern multimodal AI models suffer from modality dominance. LLaVA-7B exhibits a measured Modality Dominance Index of 10.23—it trusts its text decoder 10× more than its visual encoder [1].

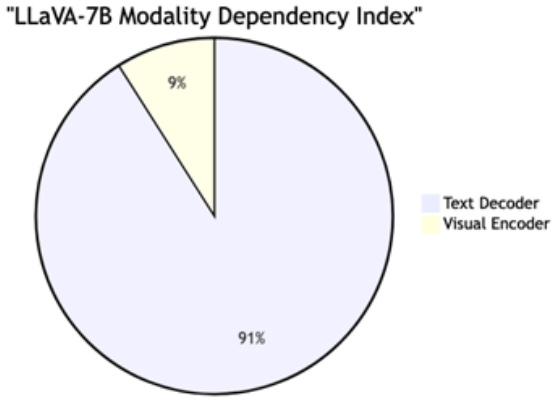

Fig. 1. LLaVA-7B trust allocation: 91% text decoder, 9% visual encoder

Three root causes drive this bias:

1) Attention dilution from token redundancy in non-text modalities
2) Fusion architectures that implicitly privilege text
3) Task formulations biased toward textual input

B. Three Failure Archetypes

We identify three recurring failure patterns in current multimodal systems (Fig. 2):

Fig. 2. Three failure archetypes: the fragile agent ignores DOM evidence; the dreaming AI produces audio-visual desync; the gullible system ignores sonar data

C. The Numbers

Empirical evidence confirms the severity of the problem:

Table 1
Evidence of multimodal failure

Metric	Value	Significance
Unimodal vs. multimodal	81.5%	Specialists beat ALL multimodal models
Modality Dominance (LLaVA)	10.23	10× more trust in text
Hallucinated training data	32.6%	Nearly a third is wrong
Adding audio (AVHBench)	WORSE	More modalities → less accuracy

D. Thesis

The brain solved multimodal integration millions of years ago. We can replicate its architecture using existing AI components. The pieces exist. We just need the architecture.

2. Neuroscience Foundations

A. Stein's Three Rules of Multisensory Integration

Barry Stein's research on the superior colliculus [2] identified three rules governing multisensory integration (Fig. 3).

*Corresponding author: krrishchoudhary109@gmail.com

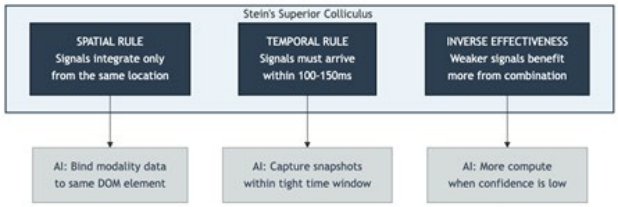

Fig. 3. Stein's three rules and their AI analogs: spatial binding, temporal synchrony, and inverse effectiveness

Ernst & Banks (Nature, 2002) [3] proved that the brain combines senses using Maximum Likelihood Estimation:

$$w\_v = (1/\sigma\_v^2) / (1/\sigma\_v^2 + 1/\sigma\_h^2),$$
$$w\_h = (1/\sigma\_h^2) / (1/\sigma\_v^2 + 1/\sigma\_h^2)$$

Each modality is weighted by its inverse variance (reliability). The combined estimate has lower variance than either alone.

B. Global Workspace Theory

Baars' (1988) [4] Global Workspace Theory uses the metaphor of a theater of consciousness (Fig. 4). Hundreds of specialized unconscious processors run in parallel. Information becomes conscious only when it wins competition and gets broadcast to all.

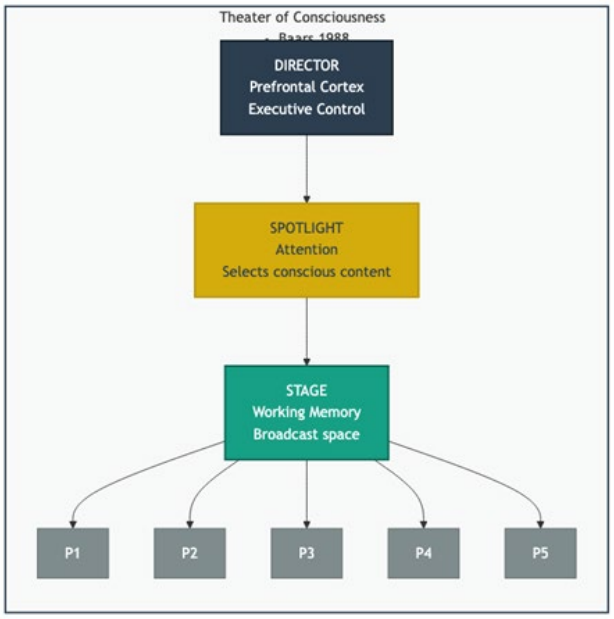

Fig. 4. The theater of consciousness: Director (PFC), Spotlight (attention), Stage (working memory), Audience (unconscious processors)

C. Dehaene's Ignition Model

Stanislas Dehaene [5] discovered a nonlinear threshold in conscious processing (Fig. 5). During the first 0–300ms, signal flows through V1→V2→V4→IT in a feedforward sweep. When threshold is exceeded, the prefrontal cortex "ignites" reactivating all earlier areas via feedback loops.

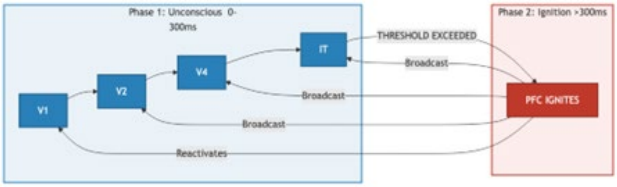

Fig. 5. Dehaene's ignition: unconscious feedforward sweep (0–300ms) followed by nonlinear PFC activation that broadcasts globally

AI Mapping: Unconscious SLMs process in parallel. Only when the conflict detector exceeds a threshold does the expensive reasoning model "ignite."

D. Friston's Free Energy Principle

Friston's predictive processing framework [6] posits that the brain maintains a generative model making top-down predictions, compared against bottom-up sensory input (Fig. 6).

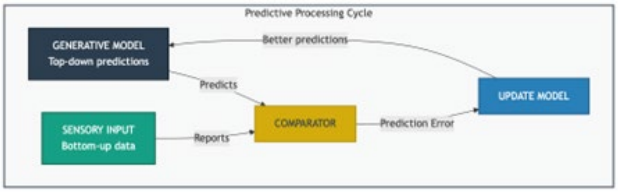

Fig. 6. The predictive processing cycle. When priors override input = hallucination. When input overrides priors = learning

E. The Default Mode Network

The DMN is active when not performing a task—daydreaming, self-reflecting, maintaining identity [7]. It maintains four persistent streams: core beliefs, personality, goals, and accumulated wisdom (Fig. 7).

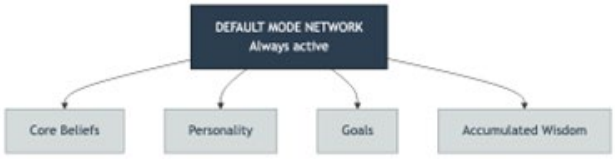

Fig. 7. The default mode network: Always-on identity maintenance

F. Complementary Learning Systems

McClelland et al. (1995) [8] proposed two complementary memory systems (Fig. 8): a fast hippocampal system for one-shot episodic learning and a slow neocortical system for generalized knowledge. Knowledge transfers during sleep.

Fig. 8. Complementary learning systems: Fast hippocampus (days) and slow neocortex (lifetime), linked by sleep

G. Schultz's Dopamine Discovery

Wolfram Schultz [10] discovered that dopamine neurons encode reward prediction error, not reward itself:

This is the temporal difference (TD) error in reinforcement learning. The brain literally runs RL.

Table 2
Dopamine = temporal difference error

Signal	Condition	Meaning
Spike ↑	Unexpected reward	"Do more of this!"
Flat →	Expected reward	"Everything normal"
Dip ↓	Missing reward	"Something's wrong"

3. Architecture Overview

The architecture comprises five layers plus two support systems, each mapping to specific brain structures (Fig. 9).

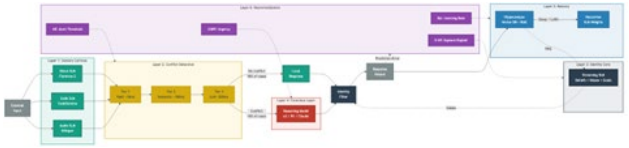

Fig. 9. The complete Artificial Brain architecture: Identity Core (DMN), Conscious Layer (PFC), Conflict Detection (ACC), Unconscious Layer (sensory cortices), Memory Systems (hippocampus + neocortex), and Reward System (neuromodulators). All layers are modulated by reward signals

A. Brain-to-AI Mapping

Table 3 provides the complete mapping from brain structures to AI components.

Table 3
Complete brain-to-AI component mapping

Brain Structure	Function	AI Component
Visual cortex	Vision	Florence-2 / PaliGemma
Auditory cortex	Sound	Whisper
Somatosensory	Structure	CodeGemma (DOM)
Sup. colliculus	Convergence	Embedding comparator
ACC	Conflict	ModernBERT / MLP
PFC	Executive	Reasoning (o3/R1/Claude)
DMN	Identity	Persistent streaming SLM
Hippocampus	Fast memory	Vector DB + RAG
Neocortex	Slow knowledge	SLM weights
Thalamus	Gating	Router MLP
Basal ganglia	Action sel.	Winner-take-all
Sleep (NREM)	Consolidation	LoRA fine-tuning
Sleep (REM)	Augmentation	Synthetic generation
Dopamine	Reward error	TD error → learning rate
Norepinephrine	Alertness	Conflict threshold
Serotonin	Satisfaction	Explore vs. exploit
Cortisol	Urgency	Speed vs. thoroughness
Synaptic tagging	Protect	Elastic Weight Cons.
Sparse activation	Efficiency	Mixture of Experts

4. The Unconscious Layer

A. A. SLM Sensory Cortices

The brain's visual cortex is not "a general brain running on low power"—it is a purpose-built specialist. The unconscious layer should use specialized SLMs, not cheap general LLMs (Table 4).

Table 4
Cheap LLM vs. Specialized SLM

Property	LLM (API)	SLM (local)
Latency	500ms–2s	<100ms
Cost	$0.15/M tok	Free
Domain accuracy	General	Expert
Hallucination	General	Lower
Network	Required	Offline
Privacy	Sent to API	Local
Fine-tuning	Unavailable	Full control

B. Candidate SLMs

Fig. 10 shows the candidate models organized by sensory cortex.

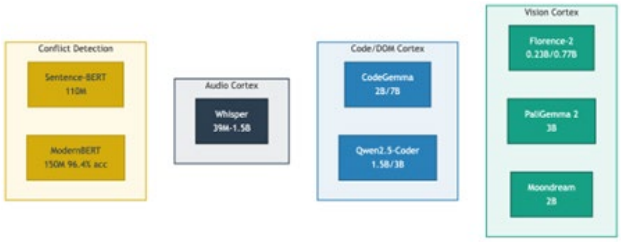

Fig. 10. Candidate SLMs: Vision (Florence-2, PaliGemma, Moondream), Code (CodeGemma, Qwen2.5-Coder), Audio (Whisper), Conflict (Sentence-BERT, ModernBERT)

Result: 90% of the time, SLMs agree → $0 cost. 10% of the time, they disagree → conscious layer resolves via API.

5. The Conflict Detector

A. Three Detection Tiers

The conflict detector implements a three-tier cascade analogous to the anterior cingulate cortex (Fig. 11).

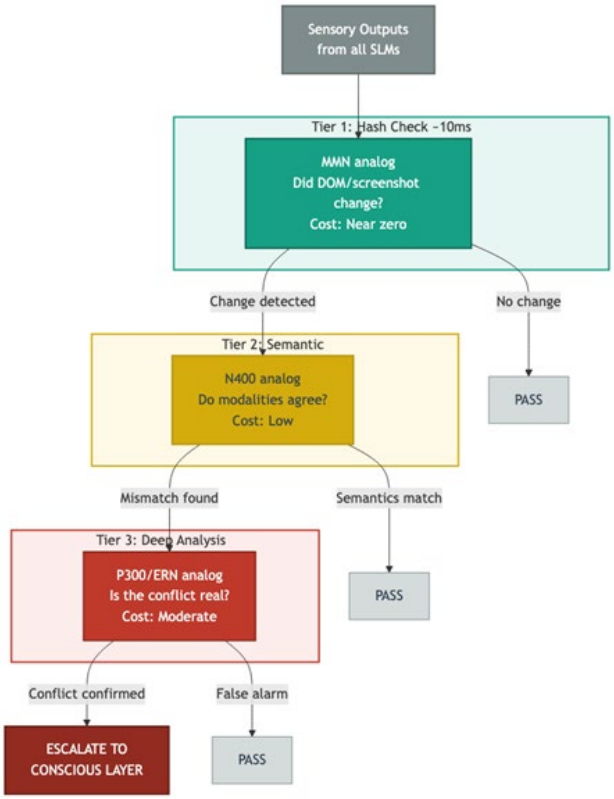

Fig. 11. Three-tier conflict detection: hash check (~10ms), semantic comparison (~100ms), deep LLM analysis (~500ms). Each tier can pass or escalate

B. Adaptive Threshold

The detection threshold adapts via norepinephrine:
threshold_eff = threshold_base × (1 - NE)
Too low = anxious (expensive). Too high = inattentive (misses real conflicts).

6. The Conscious Layer

Table 5
PFC functions and AI analogs

PFC Function	AI Analog
Top-down attention	Task-specific instructions
Working memory	Context window
Goal maintenance	Objective tracking
Conflict resolution	Adjudicating modalities
Inhibition	Blocking hallucinated actions

The conscious layer functions as the prefrontal cortex analog, activated only when the conflict detector exceeds the ignition threshold.

It issues one of four verdicts (Fig. 12): Trust A, Trust B, Synthesize, or Abort.

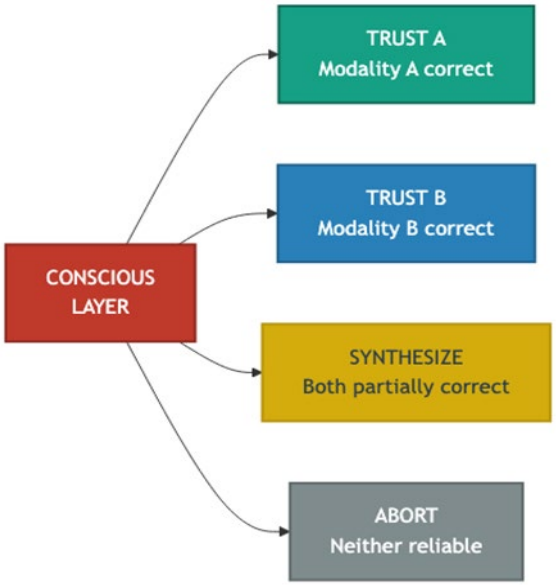

Fig. 12. The four conscious verdicts: Trust A, Trust B, Synthesize, Abort

7. The Identity Core

This is not the conscious or unconscious layer. It is the identity layer—always running, providing the frame through which everything else is interpreted. It maps to the Default Mode Network.

A. Identity Biases Decisions

Every raw decision passes through the identity filter (Fig. 13): "Click submit" becomes "Screenshot → Click → Verify" when the identity holds "I verify before irreversible actions."

Fig. 13. Identity filter: raw decisions are modified by core beliefs

B. Streaming vs. Stateless

Table 6
Transformer vs. Mamba/RWKV for identity core

Property	Transformer	Mamba/RWKV
State	Recomputed	Carried forward
Memory	O(n)	O(1)
After 1M steps	Impossible	Same cost
Continuity	Simulated	Real

8. Memory Systems

A. Hippocampus + Neocortex + Sleep

Following McClelland's CLS theory, we implement two complementary systems linked by a sleep cycle (Fig. 14).

Fig. 14. Memory systems: Wake mode stores to hippocampus (vector DB). Sleep mode replays, fine-tunes neocortex (SLM weights) via LoRA, and prunes. REM generates synthetic augmentation

B. Spaced Repetition Priority

Memory priority determines consolidation order:

$P\_mem = N\_access \times w\_recency \times I\_emotion$

High-access memories consolidate deeply; zero-access memories fade—analogous to cache eviction.

C. Preventing Catastrophic Forgetting

Four mechanisms prevent new learning from destroying old knowledge (Fig. 15).

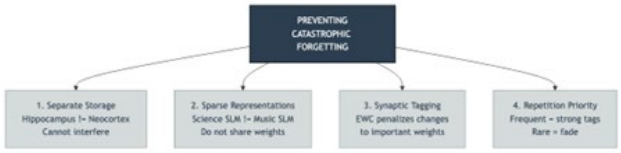

Fig. 15. Four defenses: separate storage, sparse representations (different SLMs), synaptic tagging (EWC), repetition priority

D. Action Selection—Basal Ganglia Analog

A lightweight router MLP (the thalamus, ~5ms) performs winner-take-all selection across SLMs (Fig. 18). Feedback rewards update routing weights.

9. The Reward System

A. Artificial Neuromodulation

Four neuromodulator variables control system behavior, all driven by prediction error (Fig. 16).

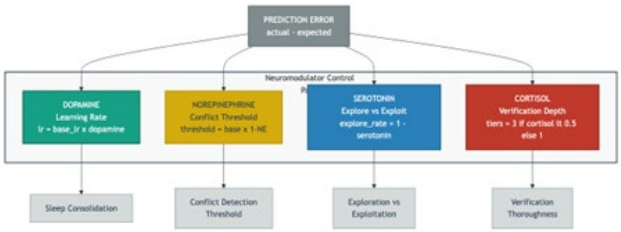

Fig. 16. Neuromodulator control panel: prediction error drives dopamine (learning rate), norepinephrine (threshold), serotonin (explore/exploit), cortisol (thoroughness)

Table 7
Neuromodulator effects on system behavior

Variable	Controls	Formula
Dopamine	Learning rate	$lr = lr_0 \times D$
Norepinephrine	Threshold	$\theta = \theta_0 \times (1\text{-}NE)$
Serotonin	Exploration	$\varepsilon = 1 - S$
Cortisol	Verification	tiers = 3 if $C < 0.5$ else 1

B. Reward Propagation

Fig. 17 shows how prediction error propagates through the system, modifying all four variables which in turn affect the next processing cycle.

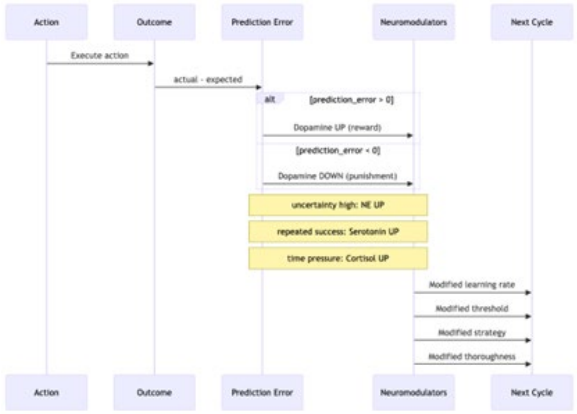

Fig. 17. Reward propagation: action → outcome → prediction error → neuromodulator update → modified next cycle

10. The SLM Swarm

A. System-Level Mixture of Experts

Instead of experts inside one model, each SLM is an expert. The architecture itself is the router.

Table 8
Monolithic LLM vs. SLM Swarm

Property	Monolithic	SLM Swarm
Learn guitar	Retrain all	Add guitar-SLM
Fix a bug	Risk breakage	Patch one SLM
Scale	More for ALL	Add where needed
Cost/task	100% active	2–3 SLMs (sparse)
Forgetting	Overwrites	Isolated

B. Action Selection—Basal Ganglia Analog

A lightweight router MLP (the thalamus, ~5ms) performs winner-take-all selection across SLMs (Fig. 18). Feedback rewards update routing weights.

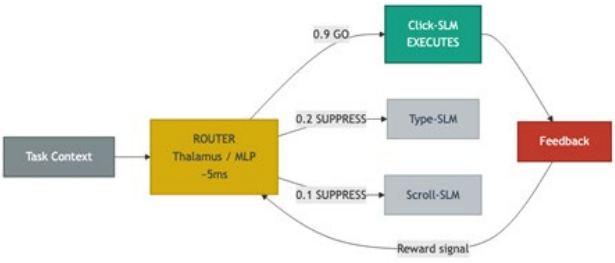

Fig. 18. Basal ganglia analog: router assigns activation scores; highest-scoring SLM executes; feedback updates routing

Intelligence isn't in any single component. It's in the connections.

11. The Life Cycle

A. Birth, Wake, Sleep, Death

The system follows a biological lifecycle (Fig. 19): START triggers birth (model loading, identity initialization). The wake loop processes input continuously. Periodically, the system enters sleep for consolidation. STOP triggers death.

Fig. 19. Life cycle: Birth → Wake (sense–remember–route–detect–respond–reward–store loop) → Sleep (NREM–REM–prune–deploy) → Wake. STOP = Death

B. The Fundamental Difference

Table 9
Current AI (zombie) vs. this architecture (alive)

	Current AI	This Arch.
Lifecycle	Per-request	Continuous
Continuity	None	Streaming state
Identity	None	Emergent
Learning	Frozen	Online + offline
State	input → output	Process + identity

C. Identity Emergence

Fig. 20 shows how personality emerges over time: blank at birth, cautious by hour 100, thorough by hour 1000, mature and efficient by hour 10000. Two identical architectures with different experiences become different beings.

Fig. 20. Identity evolution: blank (hour 0) → cautious (100) → thorough (1000) → mature (10000)

12. Precision Weighting

Each modality is trusted differently depending on the task (Table X), following Ernst & Banks' MLE framework.

Table 10
Task-specific modality trust

Task	Trust	Distrust	Why
Button visible?	Vision	DOM	Visual property
#btn exists?	DOM	Vision	Structural
Button text?	DOM	Vision	Exact text
Where on screen?	Vision	DOM	Layout
Audio match lips?	Audio	Vision	Temporal

13. Hallucination

A. The McGurk Effect

When audio "ba" is paired with visual "ga," humans perceive "da"—a third percept that neither modality alone supports (Fig. 21). This is not an error; it is the brain's fusion mechanism creating artifacts [9].

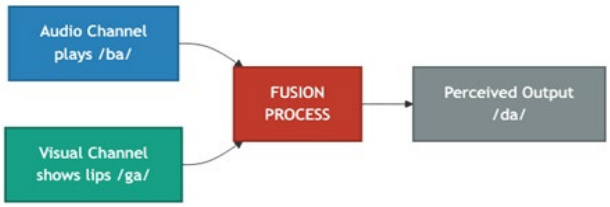

Fig. 21. The McGurk Effect: fusion creates percepts neither modality supports

B. *Five Root Causes*

Table 11
Hallucination root causes and defenses

#	Root Cause	Defense
1	Language prior dominance	Separate processing, late fusion
2	Cross-modal misalignment	Fine-tune on aligned pairs
3	Spatial perception bias	TVER-based head selection
4	Spurious co-occurrence	Counterfactual augmentation
5	Training contamination	Cross-MLLM filtering

C. *Architectural Defense*

The key insight: compare before merging. Each SLM processes independently; the conflict detector compares outputs; only agreement leads to merger; conflicts escalate (Fig. 22).

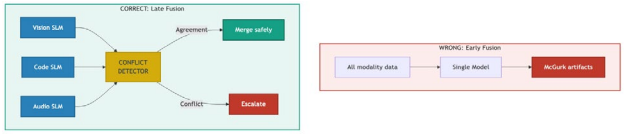

Fig. 22. Wrong: early fusion creates McGurk artifacts. Right: late fusion with conflict detection prevents hallucination

D. *The Core Loop*

The Fig. 23 shows the core loop.

14. Related Work

A. *Cross-Modal Detection*

Table 12
Cross-modal hallucination detection methods

Method	Venue	Result
DHCP	ACM MM 2025	96.78% acc
AVHBench	ICLR 2025	50→84%
HalluciDoctor	CVPR 2024	−44.6% halluc
INTER	ICCV 2025	−7.6pt CHAIR
ONLY	ICCV 2025	89.70% POPE
CausalMM	arXiv 2024	+65.3%
ExoViP	COLM 2024	Plug-and-play

B. *Existing Agent Systems*

Table 13
Existing web automation agents

System	Arch.	Score	Innovation
Skyvern	Plan-Act-Val	85.85%	Dual perception
Browser Use	DOM-first	89.1%	5-stage pipeline
MMCTAgent	Plan-Critic	—	Gen-Insp-Sup
Magentic-One	Orch+4	—	Task ledgers
Agent-E	DOM distill	73.2%	MutationObserver
Anthropic CU	Pure vision	72.7%	Vision-only limit

The Gap: None implements precision weighting, ignition thresholds, dedicated conflict detection, identity, memory, or reward. That is the opportunity.

15. Implementation

A. *Hardware Target*

RTX 4050 (6GB VRAM), Intel i7-13th Gen, 32GB RAM.

B. *Model Selection*

Table 14
Model selection and resource allocation

Component	Model	Size	Location
Identity Core	Qwen2.5-1.5B Q4	1.1GB	GPU
Response Gen	Qwen2.5-3B Q4	2.1GB	GPU
Sentiment	distilbert-sst-2	250MB	CPU
Conflict Det.	MiniLM-L6-v2	90MB	CPU
Hippocampus	ChromaDB+MiniLM	90MB	CPU
Conscious	Claude API	0	Cloud
Total GPU		3.6GB	Fits

16. Limitations

Table 15
Known limitations and severity

#	Limitation	Severity
1	Coordination failures dominate (79%)	HIGH
2	Hallucination may be inevitable	FUND.
3	No general cross-modal detector	HIGH
4	McGurk trap: fusion creates errors	MED
5	More agents can worsen things	MED
6	Non-determinism persists	MED
7	Confidence estimation unsolved	HIGH
8	Latency vs. thoroughness tradeoff	VAR
9	Missing modality streams	MED
10	Training data contamination (32.6%)	HIGH

A. *Buildability Assessment*

Fig. 24 categorizes components by readiness: buildable now (green), needs work (yellow), and open research (red).

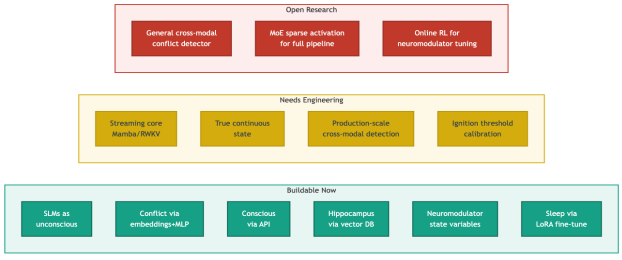

Fig. 24. Buildability: green = ready now, yellow = needs work, red = open research

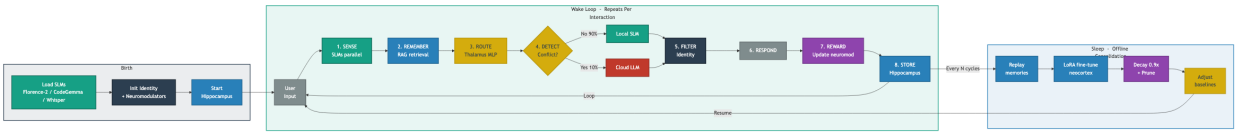

Fig. 23. The core processing loop. Birth: Load SLMs, initialize identity and neuromodulators, connect hippocampus. Wake Loop (per interaction): Sense → Remember (RAG) → Route (thalamus MLP) → Detect (3-tier conflict cascade) → Local SLM (90%) or Cloud LLM (10%) → Filter (identity alignment) → Respond → Reward (update neuromodulators) → Store (hippocampus). Sleep (every N cycles): Replay → LoRA fine-tune neocortex → Decay priorities ×0.9 → Prune → Adjust baselines → Resume

17. Conclusion

We derive six design principles from neuroscience:

1) Specialize, don't generalize. SLMs over cheap LLMs—brain cortices are experts.
2) Compare before merging. Never fuse without checking—the McGurk trap.
3) Trust by task. DOM for structure, vision for layout—Ernst & Banks MLE.
4) Flag absence, don't ignore. Missing stream = danger—the Charles Bonnet lesson.
5) Conscious is expensive. Only activate on conflict—Dehaene's ignition.
6) Learn during sleep. Online experience → offline consolidation—CLS theory.

The system is born when started, develops personality through experience and reward, sleeps to consolidate memories into permanent knowledge, and dies when stopped. It is not a chatbot with extra steps. It is an architecture for a fundamentally different kind of AI—one that has continuity, identity, and the ability to grow.

References

[1] H. Wu et al., "When language overrules: Revealing text dominance in multimodal large language models," *arXiv* preprint arXiv:2508.10552, 2025.

[2] B. E. Stein et al., "Multisensory integration in the superior colliculus requires synergy among corticocollicular inputs," *J. Neurosci.*, vol. 29, no. 20, pp. 6580–6592, 2009.

[3] M. O. Ernst and M. S. Banks, "Humans integrate visual and haptic information in a statistically optimal fashion," *Nature*, vol. 415, pp. 429–433, 2002.

[4] B. J. Baars, *A Cognitive Theory of Consciousness*. Cambridge, U.K.: Cambridge Univ. Press, 1988.

[5] S. Dehaene and J.-P. Changeux, "Experimental and theoretical approaches to conscious processing," *Neuron*, vol. 70, no. 2, pp. 200–227, 2011.

[6] K. Friston, "The free-energy principle: A unified brain theory?," *Nat. Rev. Neurosci.*, vol. 11, no. 2, pp. 127–138, 2010.

[7] M. E. Raichle et al., "A default mode of brain function," *Proc. Natl. Acad. Sci. U.S.A.*, vol. 98, no. 2, pp. 676–682, 2001.

[8] J. L. McClelland, B. L. McNaughton, and R. C. O'Reilly, "Why there are complementary learning systems in the hippocampus and neocortex: Insights from the successes and failures of connectionist models of learning and memory," *Psychol. Rev.*, vol. 102, no. 3, pp. 419–457, 1995.

[9] H. McGurk and J. MacDonald, "Hearing lips and seeing voices," *Nature*, vol. 264, pp. 746–748, 1976.

[10] W. Schultz, "Dopamine reward prediction-error signalling: A two-component response," *Nat. Rev. Neurosci.*, vol. 17, no. 3, pp. 183–195, 2016.

[11] K. P. Körding et al., "Causal inference in multisensory perception," *PLoS One*, vol. 2, no. 9, e943, 2007.

[12] U. Noppeney, "Causal inference in multisensory perception and cognition," *Prog. Brain Res.*, vol. 236, pp. 13–29, 2017.

[13] C. R. Fetsch, A. H. Turner, G. C. DeAngelis, and D. E. Angelaki, "Dynamic reweighting of visual and vestibular cues during self-motion perception," *J. Neurosci.*, vol. 29, no. 49, pp. 15601–15612, 2009. (Neural correlates of reliability-based cue weighting.)

[14] DHCP, "Attention-based hallucination detection," in *Proc. ACM Multimedia*, 2025.

[15] AVHBench, "Audio-visual hallucination benchmark," in *Proc. Int. Conf. Learn. Represent. (ICLR)*, 2025.

[16] HalluciDoctor, "Cross-MLLM consistency checking," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2024, arXiv:2311.13614.

[17] INTER, "Harsanyi dividends for hallucination detection," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, 2025.

[18] "Why do multi-agent LLM systems fail?" *arXiv* preprint arXiv:2503.13657, 2025.

[19] S. Chen et al., "FrugalGPT: How to use large language models while reducing cost and improving performance," *arXiv* preprint arXiv:2305.05176, 2023.

[20] RouteLLM, "Model routing for cost optimization," in *Proc. Int. Conf. Learn. Represent. (ICLR)*, 2025.