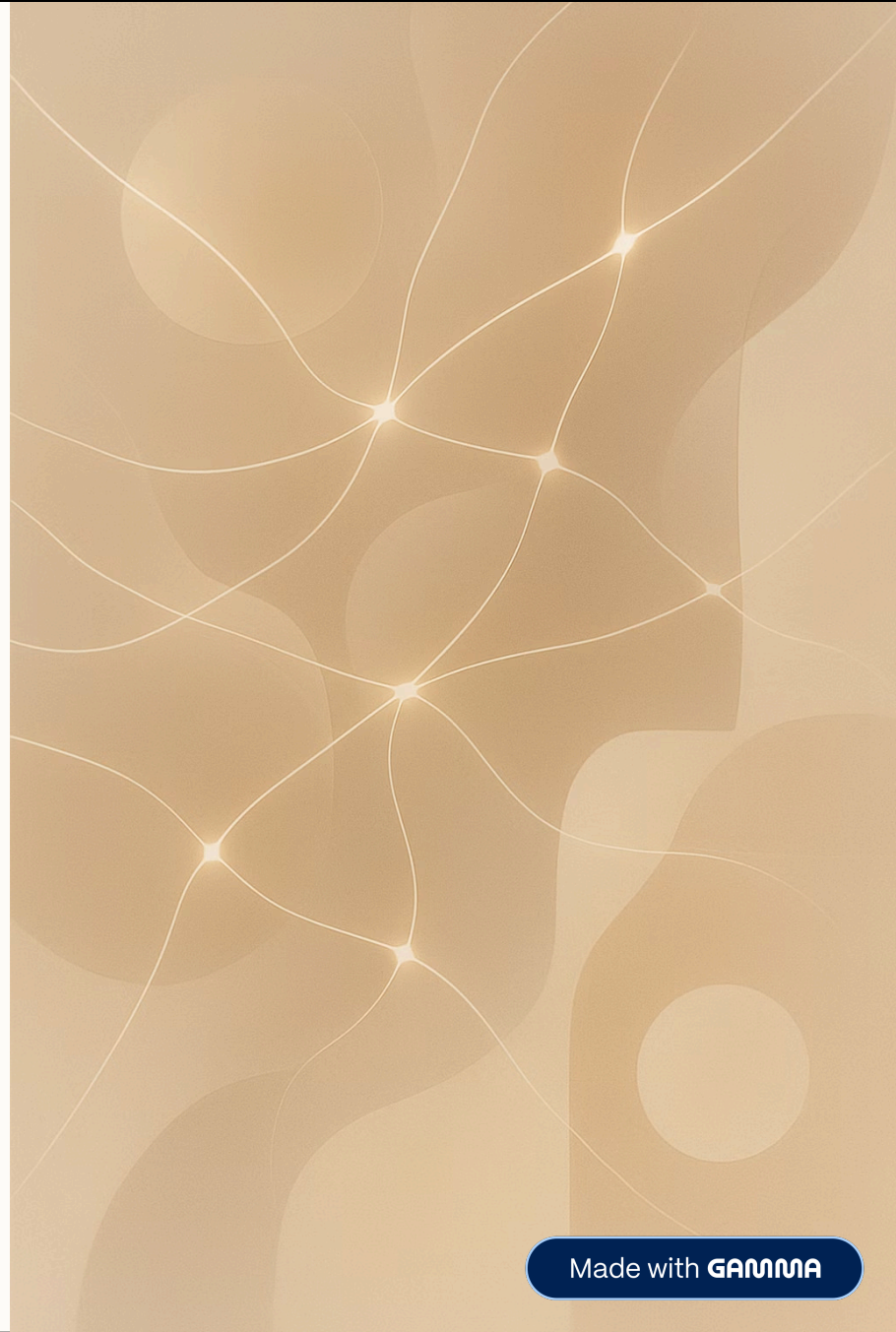
Human-Like Lifelong Memory

A Neuroscience-Grounded Architecture for Infinite Interaction

Diego C. Lerma-Torres • Universidad de Guanajuato



The Problem with Current LLMs

Single Undifferentiated Store

LLMs conflate instructions, identity, conversation history, and documents into one context window—no specialized memory systems.

Context Expansion Fails

Longer contexts degrade reasoning by up to 85%—even with perfect retrieval. More tokens ≠ better memory.

No Emotional Valuation

Current systems lack mechanisms for emotional-associative summaries that enable instant orientation before deliberation.

Three Core Principles

Memory Has Valence

Pre-computed emotional-associative summaries (valence vectors) organized in emergent belief hierarchy enable instant orientation before deliberation.

System 1 Default

Automatic spreading activation and passive priming as default, with deliberate retrieval only when needed. Graded epistemic states address hallucination.

Active Encoding

Thalamic gateway tags and routes information. Executive forms gists through curiosity-driven investigation, not passive exposure.

Architectural Overview

01

Executive Function

LLM with working memory context window operating in System 1 (default) and System 2 (escalation) modes.

02

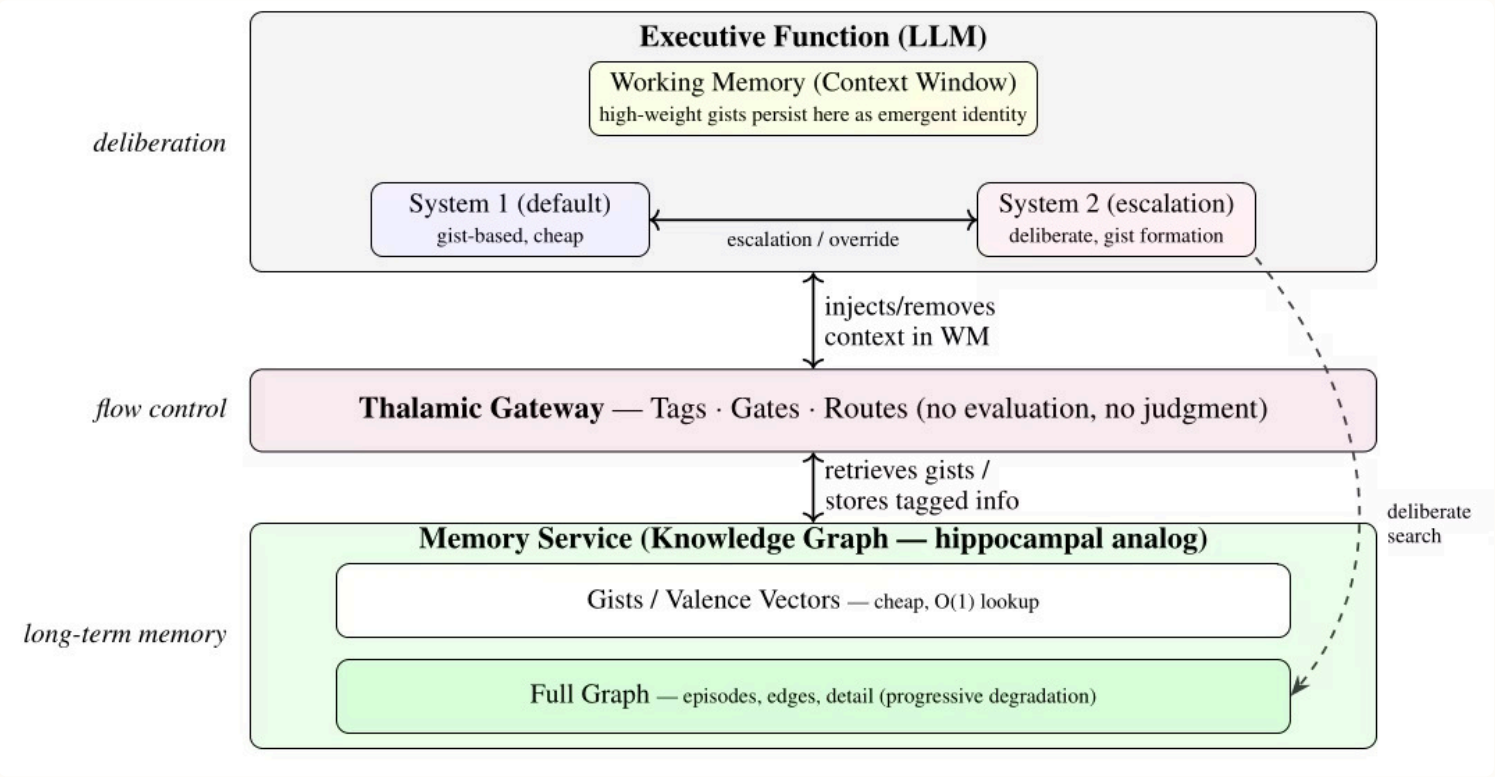
Thalamic Gateway

Tags, gates, and routes information with multi-channel salience scoring—no evaluation, no judgment.

03

Memory Service

Knowledge graph with gists/valence vectors (cheap O(1) lookup) and full graph (expensive traversal).



Principle 1: Memory Has Valence

Valence Vectors

Each knowledge graph node carries compressed emotional-associative summaries enabling instant orientation without replaying history.

- Emotional component (valence & arousal)
- Associative component (strongest connections)
- Contextual component (activation contexts)
- Density scalar (neighborhood interconnectedness)
- Precision scalar (conviction snapshot)

> ⓘ Stability by default, modification by catharsis: Gists remain stable until contradictory evidence triggers revision.

Principle 2: System 1/2 Routing

System 1 (Default)

Fast, cheap, gist-based processing. Automatic spreading activation surfaces relevant associations without deliberate search.

System 2 (Escalation)

Slow, deliberate, resource-intensive. Activated when low graph density, high novelty, or high stakes require deeper investigation.

Retrieval precision improves with graph density—retroactively enhancing recall of previously stored memories as expertise accumulates.

Principle 3: Active Encoding

Present-Moment Tagging

Gateway tags every incoming segment with multi-channel salience scores immediately—not retrospectively.

Dynamic Gating

Actively manages context through topological drift, interference, and capacity displacement—continuous active management, not batch garbage collection.

Curiosity-Driven Gist Formation

Gists form through active investigation triggered by salience—not passive exposure. Executive delimits concepts before creating gists.

Seven Functional Properties

1

Context Fluidity

Working memory dynamically loads/unloads gists. Capacity constant regardless of interaction length.

2

Real-Time Tagging

Lightweight parallel process adds salience scores without perceptible latency.

3

Monotonic Convergence

System 2 interactions decrease as experience accumulates. Cost per interaction decreases with use.

4

Graded Epistemic Awareness

Retrieval confidence as continuum: precise match, approximate match, null match.

5

Emergent Identity

High-weight gists selected frequently produce recognizable consistency across contexts.

6

Stability by Default

Gists stable until cathartic events. Updates only through gist + contradiction co-presence.

7

Active Formation

Curiosity-driven investigation required. Passive exposure produces no gist.

Testable Predictions

Valence Priming

System 1 gist injection produces faster, more appropriate retrieval—largest advantage for emotionally significant queries.

Adaptive Rigidity

High-precision gists resist single contradictions but remain modifiable through cathartic events—CBT resistance-then-shift pattern.

Multi-Channel Salience

Emotionally significant information retained in off-topic contexts that cosine-similarity systems discard.

Graded Epistemic States

Fewer hallucinations with approximate matches generating qualified responses.

Executive Override

Core belief-based suppression resists prompt injection and investigation loops.

Active Formation Superiority

Curiosity-driven gists show greater persistence and precision than passive exposure.

Experience-Dependent Efficiency

Mature instances process familiar queries with lower latency, fewer retrievals, lower cost.

Multimodal Scalability

Architecture scales to concurrent multimodal inputs by adding sensory buffers upstream.

Key Insights

Expertise as Cost Reduction

Mature instances use fewer tokens while delivering more appropriate responses—the computational analog of clinical expertise.

Executive Override as Defense

Prompt injection rejected by core beliefs. Investigation loops terminated by non-convergence recognition.

Creativity & Hallucination Unified

Same generative capacity produces both. Graded epistemic states distinguish: creative contexts use approximations freely; analytical contexts generate qualified responses.

Future Work: Empirical calibration of salience channels, cathartic thresholds, compression strategies, and multimodal extensions.