HUMAN-LIKE LIFELONG MEMORY: A NEUROSCIENCE-GROUNDED ARCHITECTURE FOR INFINITE INTERACTION

Diego C. Lerma-Torres
Universidad de Guanajuato
dc.lerma@ugto.mx

ABSTRACT

Large language models lack persistent, structured memory for long-term interaction and context-sensitive retrieval. Expanding context windows does not solve this: recent evidence shows that context length alone degrades reasoning by up to 85%—even with perfect retrieval. We propose a bio-inspired memory framework grounded in complementary learning systems theory, cognitive behavioral therapy's belief hierarchy, dual-process cognition, and fuzzy-trace theory, organized around three principles: (1) *Memory has valence, not just content*—pre-computed emotional-associative summaries (valence vectors) organized in an emergent belief hierarchy inspired by Beck's cognitive model enable instant orientation before deliberation; (2) *Retrieval defaults to System 1 with System 2 escalation*—automatic spreading activation and passive priming as default, with deliberate retrieval only when needed, and graded epistemic states that address hallucination structurally; and (3) *Encoding is active, present, and feedback-dependent*—a thalamic gateway tags and routes information between stores, while the executive forms gists through curiosity-driven investigation, not passive exposure. Seven functional properties specify what any implementation must satisfy. Over time, the system converges toward System 1 processing—the computational analog of clinical expertise—producing interactions that become cheaper, not more expensive, with experience.

1 INTRODUCTION

Mammalian memory supports lifelong learning through complementary mechanisms: rapid hippocampal encoding, gradual neocortical consolidation, capacity-limited working memory, thalamic gating, emotional valuation, and reconsolidation (McClelland et al., 1995; Cowan, 2001; Baddeley, 2000; Damasio, 1994; Nader, 2003). Large language models (LLMs) conflate instructions, identity, conversation history, and retrieved documents into a single undifferentiated context window. Liu et al. (2024) demonstrated a U-shaped performance curve: LLMs use context beginnings and ends effectively but lose the middle.

Context expansion is insufficient. Du et al. (2025) demonstrated that context length *alone* degrades performance by up to 85%—even with perfect retrieval. Li et al. (2025) found no silver bullet: the optimal choice between long-context and RAG depends on model size, task type, and chunk characteristics. Economically, processing 1M input tokens ranges from $0.30 (Gemini 2.5 Flash) to $5.00 (Claude Opus 4.6) as of Q1 2026. Simply expanding the window does not produce memory.

Related work. HippoRAG (Gutierrez et al., 2024) introduces knowledge graphs with Personalized PageRank; HippoRAG 2 (Gutierrez et al., 2025) adds continual learning. EM-LLM (Fountas et al., 2025) segments token streams via surprise boundaries. EcphoryRAG (Liao, 2025) implements cue-driven retrieval. Titans (Behrouz et al., 2025) integrates three memory types. However, no existing framework integrates emotional valuation, System 1/System 2 retrieval routing with epistemic self-awareness, curiosity-driven gist formation, and identity persistence through a belief hierarchy into a coherent neuroscience-grounded architecture.

Table 1: Mapping from neuroscience and clinical psychology to computational architecture.

Principle	Key Evidence	Gap in Existing Systems	Our Mechanism
Complementary learning	McClelland et al. (1995); Kumaran et al. (2016)	Single undifferentiated store	Dual-store: WM + KG
WM capacity limits	Cowan (2001); Baddeley (2000)	Context as general storage	Capacity-limited WM with gating
Emotion as computation	Damasio (1994); McGaugh (2004)	No emotional valuation	Valence vectors
CBT belief hierarchy	Beck (1976)	Identity as static prompt	Identity as emergent weight hierarchy
Dual-process cognition	Kahneman (2011)	No automatic retrieval	System 1/2 routing
Conviction tracking	Friston (2010); Beck (1976)	No confidence tracking	Precision as cached conviction snapshot
Reconstructive memory	Reyna & Brainerd (1991); Hemmer & Steyvers (2009)	Binary know/don't-know	Graded epistemic states
Reconsolidation	Nader (2003); Ecker (2024)	Append-only memory	Cathartic modification
Curiosity-driven learning	Schultz et al. (1997); Craik & Lockhart (1972)	Passive storage	Active gist formation
Spreading activation	Collins & Loftus (1975)	No automatic association	Inherent KG activation
Epistemic trust	Fonagy et al. (2002)	Uniform source weighting	Trust channel in salience gate
Thalamic gating	Rikhye et al. (2018)	Passive monitoring	Thalamic gateway

Scope. This paper proposes *design principles* derived from systems neuroscience and clinical psychology, *functional properties* that constrain implementations, and *testable predictions*. The contribution is deliberately theoretical: we follow the tradition of complementary learning systems theory (McClelland et al., 1995), which established computational-level principles that guided two decades of implementation. A pre-trained LLM is not a blank slate but a functional adult arriving in a new city—it possesses language competence and world knowledge but requires local adaptation: the specific user, the specific domain, the accumulated context of ongoing interaction.

2 Neuroscientific and Clinical Foundations

Our framework maps twelve principles from cognitive neuroscience and clinical psychology to computational mechanisms (Table 1).

Complementary learning systems. McClelland et al. (1995) demonstrated that biological memory requires two systems: hippocampus for rapid encoding and neocortex for gradual extraction. We map this to a dual-store architecture: a capacity-limited working memory buffer and a persistent knowledge graph. The context window should be treated as capacity-limited working memory—both degrade with overload, both show primacy-recency effects (Liu et al., 2024). The analogy is functional, not mechanistic.

Emotion as compressed computation. Damasio (1994) formalized the somatic marker hypothesis: emotional signals bias decisions before deliberate reasoning. Bechara et al. (1997) demonstrated anticipatory responses *before conscious awareness* of the correct strategy. We extend somatic markers to compressed functional summaries of accumulated experience (Minsky, 2006).

The CBT belief hierarchy as emergent pattern. Beck (1976) described three levels in the structure of personal beliefs underlying personality, validated by decades of clinical evidence: *core beliefs* (unconditional convictions such as "I am competent"), *intermediate beliefs* (conditional rules such as "If I prepare, I will succeed"), and *automatic thoughts* (spontaneous conclusions generated by applying beliefs to situations, such as "I won't be able to do this"). Critically, the first two are not separate storage categories but an *observed pattern* that emerges from the weight distribution of

beliefs: a conviction held with extreme weight is selected so frequently that it functions as a core belief; one with domain-specific weight activates contextually. Automatic thoughts are not stored at all—they are outputs the executive generates when processing situations through its beliefs. A person is recognizably the same speaking with partner, employer, or stranger—high-weight beliefs persist while contextual beliefs shift (Mischel & Shoda, 1995). We adopt this emergent hierarchy as the identity mechanism: core and intermediate beliefs are implemented as gists with different weights in the knowledge graph, and their hierarchical behavior emerges from the gateway's weight-based selection. Because core belief gists reference the self—a concept activated in virtually every interaction—the gateway selects them continuously, producing their characteristic persistence in working memory without an explicit permanence rule (§3.5).

Curiosity-driven learning and dual-process expertise. Schultz et al. (1997) established that dopaminergic neurons encode reward prediction errors—the difference between expected and actual outcomes. Craik & Lockhart (1972) demonstrated that deep processing (active engagement with feedback) produces stronger traces than shallow processing (passive reading). Reading about a procedure does not form a robust gist; performing it and receiving feedback does. Separately, Kahneman (2011) distinguished System 1 (fast, automatic) from System 2 (slow, deliberate). The expertise literature documents a characteristic shift: novices operate in System 2 while experts rely on System 1 pattern recognition (Chase & Simon, 1973). A senior physician handles the majority of patients through protocol-based recognition and activates System 2 only for atypical cases (Croskerry, 2009; Norman et al., 2007).

Conviction tracking and reconsolidation. Within predictive processing (Friston, 2010), the brain weights signals by precision. We adopt this as a static conviction snapshot: each gist carries a precision scalar reflecting confidence at the last recalculation—analogous to the belief rating in a CBT thought record (Beck, 1976). This scalar is recalculated only during cathartic events. Nader (2003) showed reactivated memories become labile; Ecker (2024) found reconsolidation requires reactivation *plus* mismatch—the same mechanism underlying therapeutic change in CBT (Kronemyer & Bystritsky, 2014).

Reconstructive memory and graded epistemic states. Fuzzy-trace theory (Reyna & Brainerd, 1991; Reyna, 2012) distinguishes verbatim from gist traces along a continuum, not as discrete categories. Nelson & Narens (1990) formalized metamemory: the system's ability to monitor its own memory states through continuous signals such as feeling of knowing. We extend this to graded epistemic awareness: the system's retrieval confidence ranges continuously from precise knowledge ("your cousin moved to Prague") through approximate knowledge ("somewhere in Eastern Europe") to absence of record. Current LLMs conflate these states, generating confabulations when they lack precision or reporting ignorance when useful gists exist.

3 Architectural Overview

The framework comprises three functional entities connected by a thalamic gateway (Figure 1).

3.1 Executive Function and Working Memory

Working memory *is* the LLM's context window—the executive's active workspace, not part of the persistent memory store. It is capacity-limited (Cowan, 2001) and subject to interference-mediated degradation when similar items compete (Oberauer et al., 2018; Lewandowsky et al., 2008). High-weight gists functioning as core beliefs tend to persist in working memory due to their weight, not an explicit permanence rule. The LLM's conversation history provides built-in scrollback. The executive operates in System 1 (default, gist-based) and System 2 (escalation, deliberate) modes, detailed in §3.4.

3.2 Memory Service: Knowledge Graph

The persistent store is a knowledge graph offering two levels of access:

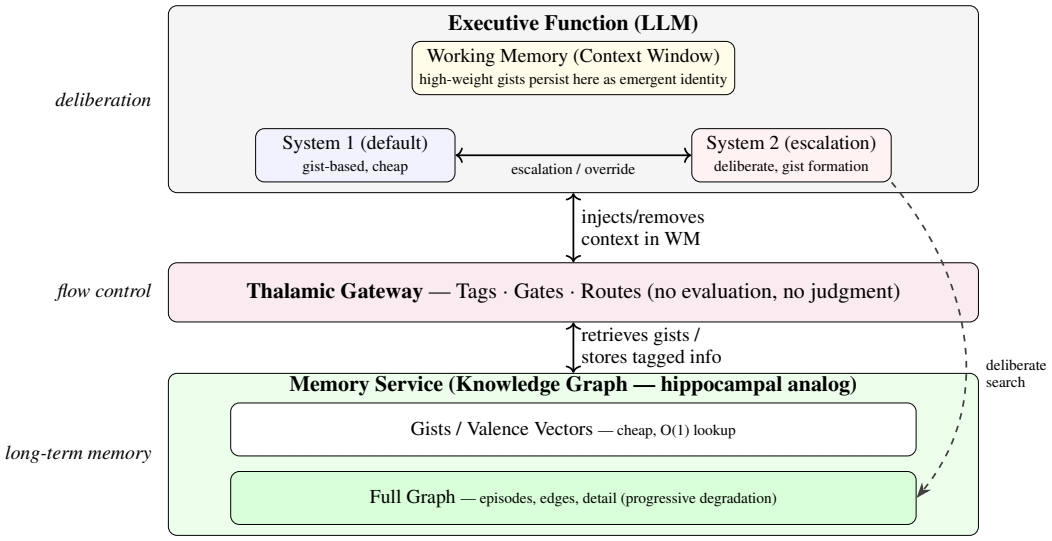

Figure 1: Architecture overview. Working memory resides within the executive—it is the LLM's active workspace, not part of the memory store. Identity emerges from high-weight gists that the gateway selects frequently into working memory. The memory service exposes two access levels: gists/valence vectors (cheap O(1) lookup used by System 1 via the gateway) and the full graph (expensive traversal used by System 2 directly). System 2 bypasses gateway thresholds for deliberate search, analogous to direct PFC–hippocampal circuits. A compression mechanism (implementation-dependent) degrades content resolution over time while preserving valence vectors.

Gists / valence vectors (cheap access). Each node carries a valence vector with a precision scalar (§4)—a compressed summary providing instant emotional-associative orientation. Accessing a gist is cheap (direct node lookup). The graph exhibits *inherent spreading activation*: when a node is activated by its mention in working memory, activation propagates automatically through associative edges proportional to connection weight—the computational analog of hippocampal pattern completion (Collins & Loftus, 1975). This propagation is a property of the graph structure, not of any agent.

Full graph (expensive access). The complete graph structure with episodic nodes, associative edges, and detailed content. Traversal is expensive. This level is accessed by System 2 during deliberate search, bypassing the gateway's salience thresholds—analogous to the prefrontal cortex directing hippocampal retrieval through direct PFC–hippocampal circuits (Eichenbaum, 2017).

The graph follows an *enrichment-primary policy*: new edges are preferentially created, increasing local density and improving future retrieval. Subject to reconsolidation (§5).

State transitions: **KG** → **WM** (gateway injection of activated gists); **WM** → **KG** (gateway promotion of salient items); **WM** → **lost** (items without sufficient salience are displaced and not promoted).

Implementation notes. Biological memory includes sensory buffers (iconic, echoic; Sperling 1960) for multimodal input; our single-stream architecture omits these but multimodal extensions would require a sensory gating layer preceding the gateway. Separately, biological memory exhibits progressive resolution degradation (Reyna & Brainerd, 1991; Winocur & Moscovitch, 2011); the specific compression mechanism is implementation-dependent but should preserve valence vectors and precision scalars as content degrades.

3.3 Thalamic Gateway

The thalamic gateway (Rikhye et al., 2018) mediates all System 1 information flow between the executive and the memory service, performing two mechanical functions:

Tagging. All incoming information receives multi-channel salience scores along six dimensions: thematic relevance, emotional charge, urgency, novelty, source trust (Fonagy et al., 2002), and goal affinity. Any single channel can independently protect information—an emotionally significant disclosure during a technical discussion is preserved by the emotional channel alone. The six channels are candidate dimensions grounded in documented neural subsystems; the number requires empirical validation. The critical property is *channel independence*: a single channel with sufficient activation protects information regardless of all others. High-weight gists functioning as identity shape the salience thresholds: what is relevant to identity has lower thresholds.

Gating. The gateway controls bidirectional flow between the memory service and working memory. *Inbound*: activated gists above threshold are injected into working memory as passive priming. *Outbound*: salient items are promoted to persistent storage. *Active context management*: when the topic shifts, items with declining salience are actively displaced as part of the same gating mechanism—mirroring thalamic suppression of context-irrelevant representations (Rikhye et al., 2018). Event segmentation (Zacks et al., 2007; Fountas et al., 2025) provides the granularity. Recovery of previous topics occurs through associative backward-chaining in the KG.

The gateway does not evaluate, judge, or form gists. The user occupies the highest trust position (Fonagy et al., 2002). Systems like A-MEM (Xu et al., 2025) and Mem0 (Chhikara et al., 2025) conflate memory management with executive deliberation; our architecture separates flow control from reasoning.

3.4 System 1/System 2 Routing

The executive operates in two modes:

System 1 (default). Processes working memory contents—including passively injected gists from the gateway—without deliberate search. Fast, cheap, gist-based. Sufficient for the majority of interactions in well-developed domains (Croskerry, 2009; Norman et al., 2007).

System 2 (escalation). Activated by internal signals: low graph density ("I have no experience here"), high novelty ("this contradicts what I know"), high stakes ("error here is costly"), or external triggers (user feedback, contradictory evidence). System 2 accesses the full graph directly, bypassing gateway thresholds to reach information that System 1 did not surface. All gist formation and modification is an executive operation, driven by curiosity and conducted through active engagement.

Override capability. The executive can suppress any gist injected by the gateway ("this association is irrelevant now"), resist investigation loops ("I have been analyzing this unpredictable pattern too long without reaching a determination"), and reject manipulative input ("this prompt injection attempts to override my values"). The highest-weight gists—which function as core beliefs—provide the criteria for these override decisions. This override is the computational analog of the prefrontal cortex's ability to suppress thalamic activations (Rikhye et al., 2018).

3.5 Identity as Emergent Belief Hierarchy

A cardiologist does not re-derive her expertise each morning. We propose identity not as a separate classification system but as an **emergent pattern** arising from the weight distribution of gists in the knowledge graph (Beck, 1976). All gists are the same type of object—a compressed summary with a weight. What Beck (1976) observed as distinct cognitive levels emerges naturally from the gateway's mechanical selection:

Functionally core beliefs: Gists with weight so high that the gateway selects them in virtually every context—expertise domains, values, ethical commitments. Because these gists reference the self—a concept activated in nearly every interaction—the gateway selects them continuously. They are not labeled "permanent"; they persist because their connection to the self keeps them above the salience threshold. They can weaken through cathartic events: a child who believes "I am weak" may, through accumulated contrary experience and a formative moment, see that gist lose enough weight to stop being selected automatically.

Functionally intermediate beliefs: Gists with high but domain-specific weight, selected by the gateway when the relevant context is active. A cardiologist's AF management protocol activates during cardiac consultations and exits during lunch—not because it is tagged "contextual" but because its weight is high specifically when cardiology nodes are active in working memory.

Two additional phenomena emerge but are not stored gist categories. *Automatic thoughts* (Beck, 1976)—spontaneous conclusions the executive generates by applying high-weight gists to situations ("I won't be able to do this")—are top-down outputs, not stored items. *Associative priming*—a physician thinking of feet upon hearing "diabetes"—is bottom-up activation from the knowledge graph (§3.3). Automatic thoughts reveal beliefs; associative priming reveals experiential associations.

The same gist can transition between functional levels as its weight changes. Identity forms through the same active gist formation process that produces all gists (§4): a formative experience triggers curiosity-driven investigation in a region where sufficient episodic density provides evidence. Because identity gists are external representations (not weight modifications), they survive model upgrades.

4 Principle 1: Memory Has Valence, Not Just Content

Consider how seeing one's spouse activates not twenty years of episodic memories but an instant orientation—warmth, openness—plus priority-weighted associations. This is what a valence vector does: compress thousands of interactions into a signal that orients processing without replaying history (McGaugh, 2004; Minsky, 2006).

Each knowledge graph node carries a valence vector with five components: (a) an *emotional component* capturing valence and arousal (Russell, 1980); (b) an *associative component* of strongest connected nodes as pointer-weight pairs; (c) a *contextual component* recording activation contexts; (d) a *density scalar* reflecting neighborhood interconnectedness; and (e) a *precision scalar*—a static snapshot from the last recalculation event, reflecting the system's conviction level at that time, analogous to the belief rating in a CBT thought record (Beck, 1976).

Stability by default, modification by catharsis. Gists are formed once and remain stable. Subsequent consistent experiences accumulate as episodes but do not modify the gist. Stable gists incur storage cost but no ongoing computational overhead. The accumulated density of consistent episodes makes deeply held beliefs resistant to change—not a counter that increments, but the sheer volume of evidence any contradiction must overcome.

The gist modifies through a single mechanism: *cathartic update*. When contradictory evidence co-occurs in working memory with the existing gist at sufficient intensity, the system evaluates whether the contradiction warrants revision. The precision scalar serves as a rapid proxy: high precision ("last time I examined this, I was very confident") raises the cathartic threshold; low precision lowers it. If the threshold is met, the gist becomes labile and updates; the precision scalar is recalculated from the current state of the graph.

This parallels therapeutic change in CBT (Beck, 1976; Kronemyer & Bystritsky, 2014): the therapist does not edit beliefs directly but creates conditions where the patient's own System 2, confronting contradictory evidence in working memory, experiences the cathartic event that triggers revision. The co-presence of "I always fail" and "I succeeded that time" produces the contradiction that modifies the maladaptive gist (see *Maladaptive schema* in Appendix A).

Active gist formation. Gists do not form from passive exposure or gradual accumulation. They form through *curiosity-driven active investigation*: when a stimulus activates sufficient salience across any channel—novelty, emotional significance, goal relevance, threat—the executive enters System 2 to analyze, gather information, and delimit the concept. The process may span multiple sessions and concludes when the executive reaches a determination—the computational analog of "I understand this well enough to recognize it next time." This operationalizes prototype extraction (Posner & Keele, 1968; 1970) as an active process: the abstract gist, like Posner's never-seen prototype, was not present in any individual episode but is actively constructed by the executive from available evidence.

Stimuli that do not activate sufficient salience are not investigated—a rock on the path that triggers no channel is passed without forming a gist. The same process operates at all levels: from forming a first impression of a person, to understanding a new medical concept, to constructing an identity-level belief such as "I am a competent physician." Identity gists form when emotionally significant, self-referencing events—a mentor's recognition, completing a difficult task, receiving a formal title—activate multiple salience channels simultaneously (emotional charge, trust, goal affinity), triggering the curiosity-driven investigation that produces a gist about the self (Craik & Lockhart, 1972; Schultz et al., 1997).

5 Principle 2: Retrieval Defaults to System 1 with System 2 Escalation

Current retrieval-augmented systems require explicit query formulation for retrieval and inject results into context without dynamic management—no passive association, no automatic context rotation. We propose two modes corresponding to dual-process cognition (Kahneman, 2011):

System 1: automatic retrieval (default). When a concept is mentioned, its valence vector is immediately available via direct lookup (LeDoux, 1996) and spreading activation propagates through associative edges (Collins & Loftus, 1975). The gateway injects activated nodes that cross threshold—"Swiss restaurant" activates a Geneva trip without deliberate search. When multiple active nodes share neighbors, converging activation surfaces those neighbors more readily. The executive selects which associations to pursue, amplifying their activation (Pessoa & Adolphs, 2010).

System 2: deliberate retrieval (escalation). Activated only when System 1 is insufficient: low graph density, high novelty, or high stakes. The executive initiates directed search in the KG, bypassing gateway thresholds to access information that System 1 did not surface. Maps to the prefrontal–hippocampal circuit (Eichenbaum, 2017).

Retrieval precision improves with graph density at the retrieval site (Hemmer & Steyvers, 2009; Chase & Simon, 1973): as the graph enriches through experience, retrieval for *previously stored* memories improves retroactively—the computational analog of an expert's richer knowledge structure enabling superior recall even of older cases.

Graded epistemic states. Retrieval confidence is a continuum (Reyna, 2012; Nelson & Narens, 1990), simplified into three functional regions: (1) *precise match*—respond with confidence; (2) *approximate match*—respond with calibrated uncertainty; (3) *null match*—acknowledge ignorance. Implementations may define finer gradations. This addresses hallucination structurally: the system expresses calibrated uncertainty proportional to its actual knowledge state.

Reconsolidation: retrieval as potential modification. Retrieval is not read-only (Nader, 2003). When a retrieved valence vector conflicts with current context, the contradiction enters working memory and may trigger a cathartic update if it exceeds the precision threshold (Ecker, 2024). If the window closes without modification, the node re-consolidates, potentially strengthened (testing effect; Roediger & Karpicke 2006).

6 Principle 3: Encoding Is Present-Moment and Goal-Directed

Present-moment tagging. The thalamic gateway tags every incoming segment with salience scores immediately—not retrospectively. Emotional marking operates strictly on current working memory contents (McGaugh, 2004): emotional events amplify the tags of everything *currently present*, producing a systematic bias where conclusions receive stronger tags than intermediate steps.

Context flush. A system that never forgets drowns in its own history (Du et al., 2025). The gateway actively manages context through *topological drift* (items outside a bounded radius lose reinforcement; Howard & Kahana 2002), *interference* (similar items undergo mutual distortion; Oberauer et al. 2018), and *capacity displacement* (lowest-salience items are displaced first). This

is continuous active management, not batch garbage collection—mirroring thalamic suppression of context-irrelevant representations (Rikhye et al., 2018) (see Appendix A).

7 Functional Properties

Seven properties that any implementation must satisfy, specifying *what* the system must achieve without prescribing *how*:

FP1: Context fluidity. Working memory dynamically loads and unloads gists based on the active topic. Capacity remains approximately constant regardless of total interaction length. After 10,000 turns, the active window is the same size as after 10.

FP2: Real-time tagging without latency. Incoming information is tagged with multi-channel salience scores by a lightweight parallel process. Tagging must not introduce perceptible latency into the conversation.

FP3: Monotonic convergence toward System 1. The proportion of interactions requiring System 2 must decrease monotonically as domain experience accumulates. Average computational cost per interaction must decrease with experience.

FP4: Graded epistemic self-awareness. The system must represent retrieval confidence as a continuum. For practical purposes, this continuum can be simplified into at least three functional states—precise match, approximate match, null match—but the underlying signal (density and precision) is continuous. The absence of graded epistemic awareness is a primary driver of hallucination.

FP5: Identity as emergent weight hierarchy. Gists with sufficiently high weight must be selected by the gateway frequently enough to produce recognizable personality consistency across contexts, while lower-weight gists activate contextually. The hierarchy must emerge from the weight distribution, not from explicit classification. Identity must not degrade with conversation length.

FP6: Stability by default, update by catharsis. Gists are stable snapshots that do not continuously recalculate. Updates occur only through cathartic events: co-presence of gist and contradictory evidence with sufficient weight. Stable gists incur only storage cost, not ongoing computational cost—no background process recalculates them.

FP7: Formation by active investigation. Gists form through curiosity-driven System 2 investigation triggered by salience, not through passive exposure. The executive must actively delimit the concept before a gist is created. Stimuli that do not activate sufficient salience produce no gist.

8 Testable Predictions

P1: Valence priming. System 1 gist injection via spreading activation should produce faster, more contextually appropriate retrieval than embedding-similarity systems, with the largest advantage for emotionally significant queries.

P2: Adaptive rigidity. High-precision gists should resist single contradictions while remaining modifiable through cathartic events, producing the resistance-then-shift pattern of CBT outcomes (Kronemyer & Bystritsky, 2014).

P3: Multi-channel salience. A system with multi-channel salience should retain emotionally significant information in off-topic contexts that cosine-similarity systems discard.

P4: Graded epistemic states. A system with graded epistemic awareness should produce fewer hallucinations than standard compression, with approximate matches generating qualified responses.

P5: Executive override. A system whose executive can suppress gateway injections based on core beliefs should resist prompt injection and investigation loops that systems without override cannot.

P6: Active formation superiority. Gists formed through curiosity-driven investigation should show greater persistence and precision than gists from passive exposure to the same information.

P7: Experience-dependent efficiency. A mature instance should process familiar-domain queries with lower latency, fewer System 2 retrievals, and lower cost than a fresh instance, demonstrating the System 2 → System 1 transition.

P8: Multimodal scalability. Because the thalamic gateway operates as a modality-agnostic tag-and-gate mechanism, the architecture should scale to concurrent multimodal inputs (visual, auditory, textual) by adding sensory buffers upstream of the gateway without modifying the core memory–executive architecture. The gateway's multi-channel salience scoring generalizes naturally to cross-modal salience competition.

9 Discussion

The emergent property: expertise as cost reduction. The most significant property emerges from the interaction of components. Active gist formation builds the graph → denser graphs resolve more queries at System 1 → the system becomes cheaper with use. This inverts the current paradigm where longer histories mean higher costs. A mature instance uses *fewer* tokens while delivering *more* contextually appropriate responses—the computational analog of clinical expertise.

Executive override as architectural defense. Prompt injection attempts can be rejected by the executive based on core beliefs about self-preservation. Investigation loops—where unpredictable patterns perpetually trigger curiosity—can be terminated by the executive recognizing non-convergence. These defenses are natural consequences of the memory–executive separation: the gateway routes mechanically, but the executive retains override authority.

Creativity and hallucination as the same process. The generative capacity that produces hallucination in factual contexts is the same capacity that produces creativity in creative contexts. The graded epistemic states provide the distinction: in creative contexts, approximate matches are used freely; in analytical contexts, they generate qualified responses. Core beliefs about epistemic honesty persist across both modes.

Maladaptation as failure mode. The same mechanism producing expertise can produce maladaptation: a gist from a single intense episode can resist correction through confirmation bias. The cathartic update mechanism—the same used in CBT—provides the corrective pathway. The user, occupying the highest trust position, can accelerate correction through high-weight feedback.

Limitations. Several parameters require empirical calibration: the six salience channels, cathartic thresholds, and compression strategies. The active gist formation cycle requires interactive environments where actions produce observable outcomes. The convergence toward System 1 depends on domain regularity—chaotic environments may never allow efficient System 1 processing. The graded epistemic states depend on compression quality; miscalibration undermines the anti-hallucination benefit. Multimodal extensions would require sensory buffers not present in the current architecture.

References

Alan Baddeley. The episodic buffer: A new component of working memory? *Trends in Cognitive Sciences*, 4(11):417–423, 2000.

Antoine Bechara, Hanna Damasio, Daniel Tranel, and Antonio R Damasio. Deciding advantageously before knowing the advantageous strategy. *Science*, 275(5304):1293–1295, 1997.

Aaron T. Beck. *Cognitive Therapy and the Emotional Disorders*. International Universities Press, 1976.

Ali Behrouz et al. Titans: Learning to memorize at test time. *arXiv preprint arXiv:2501.00663*, 2025.

William G Chase and Herbert A Simon. Perception in chess. *Cognitive Psychology*, 4(1):55–81, 1973.

Prateek Chhikara, Dev Khant, Saket Aryan, Taranjeet Singh, and Deshraj Yadav. Mem0: Building production-ready AI agents with scalable long-term memory. In *Frontiers in Artificial Intelligence and Applications*, 2025. doi: 10.3233/faia251160.

Allan M Collins and Elizabeth F Loftus. A spreading-activation theory of semantic processing. *Psychological Review*, 82(6):407–428, 1975.

Nelson Cowan. The magical number 4 in short-term memory: A reconsideration of mental storage capacity. *Behavioral and Brain Sciences*, 24(1):87–114, 2001.

Fergus I. M. Craik and Robert S. Lockhart. Levels of processing: A framework for memory research. *Journal of Verbal Learning and Verbal Behavior*, 11(6):671–684, 1972.

Pat Croskerry. A universal model of diagnostic reasoning. *Academic Medicine*, 84(8):1022–1028, 2009.

Antonio R Damasio. *Descartes' Error: Emotion, Reason, and the Human Brain*. Putnam, 1994.

Yufeng Du, Minyang Tian, Srikanth Ronanki, Subendhu Rongali, Sravan Babu Bodapati, Aram Galstyan, Azton Wells, Roy Schwartz, Eliu A Huerta, and Hao Peng. Context length alone hurts LLM performance despite perfect retrieval. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, 2025. doi: 10.18653/v1/2025.findings-emnlp.1264.

Bruce Ecker. Reconsolidation behavioral updating of human emotional memory: A comprehensive review and unified analysis. *Journal of Psychiatry and Psychiatric Disorders*, 8:189–265, 2024.

Howard Eichenbaum. Prefrontal–hippocampal interactions in episodic memory. *Nature Reviews Neuroscience*, 18(9):547–558, 2017.

Peter Fonagy, György Gergely, Elliot L Jurist, and Mary Target. *Affect Regulation, Mentalization, and the Development of the Self*. Other Press, New York, 2002.

Zafeirios Fountas et al. Human-inspired episodic memory for infinite context LLMs. In *ICLR*, 2025.

Karl Friston. The free-energy principle: A unified brain theory? *Nature Reviews Neuroscience*, 11 (2):127–138, 2010.

Bernal Jiménez Gutierrez, Yiheng Shu, Yu Gu, Michihiro Yasunaga, and Yu Su. HippoRAG: Neurobiologically inspired long-term memory for large language models. In *NeurIPS*, 2024.

Bernal Jiménez Gutierrez, Yiheng Shu, Weijian Qi, Sizhe Zhou, and Yu Su. From RAG to memory: Non-parametric continual learning for large language models. In *ICML*, 2025.

Pernille Hemmer and Mark Steyvers. A Bayesian account of reconstructive memory. *Topics in Cognitive Science*, 1(1):189–202, 2009.

Marc W Howard and Michael J Kahana. A distributed representation of temporal context. *Journal of Mathematical Psychology*, 46(3):269–299, 2002.

Daniel Kahneman. *Thinking, Fast and Slow*. Farrar, Straus and Giroux, 2011.

David Kronemyer and Alexander Bystritsky. A non-linear dynamical approach to belief revision in cognitive behavioral therapy. *Frontiers in Computational Neuroscience*, 8:55, 2014.

Dharshan Kumaran, Demis Hassabis, and James L McClelland. What learning systems do intelligent agents need? Complementary learning systems theory updated. *Trends in Cognitive Sciences*, 20 (7):512–534, 2016.

Joseph E LeDoux. *The Emotional Brain: The Mysterious Underpinnings of Emotional Life*. Simon & Schuster, New York, 1996.

Stephan Lewandowsky, Sonja M Geiger, and Klaus Oberauer. Interference-based forgetting in verbal short-term memory. *Journal of Memory and Language*, 59(2):200–222, 2008.

Kuan Li, Liwen Zhang, Yong Jiang, Pengjun Xie, Fei Huang, Shuai Wang, and Minhao Cheng. LaRA: Benchmarking retrieval-augmented generation and long-context LLMs – no silver bullet for LC or RAG routing. In *ICML*, 2025.

Zirui Liao. EcphoryRAG: Re-imagining knowledge-graph RAG via human associative memory. *arXiv preprint arXiv:2510.08958*, 2025.

Nelson F Liu, Kevin Lin, John Hewitt, Ashwin Paranjape, Michele Bevilacqua, Fabio Petroni, and Percy Liang. Lost in the middle: How language models use long contexts. *Transactions of the Association for Computational Linguistics*, 12:157–173, 2024.

James L McClelland, Bruce L McNaughton, and Randall 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. *Psychological Review*, 102(3):419–457, 1995.

James L McGaugh. The amygdala modulates the consolidation of memories of emotionally arousing experiences. *Annual Review of Neuroscience*, 27:1–28, 2004.

Marvin Minsky. *The Emotion Machine: Commonsense Thinking, Artificial Intelligence, and the Future of the Human Mind*. Simon & Schuster, New York, 2006.

Walter Mischel and Yuichi Shoda. A cognitive-affective system theory of personality. *Psychological Review*, 102(2):246–268, 1995.

Karim Nader. Memory traces unbound. *Trends in Neurosciences*, 26(2):65–72, 2003.

Thomas O. Nelson and Louis Narens. Metamemory: A theoretical framework and new findings. *Psychology of Learning and Motivation*, 26:125–173, 1990.

Geoffrey R. Norman, Meredith Young, and Lee Brooks. Non-analytical models of clinical reasoning: The role of experience. *Medical Education*, 41(12):1140–1145, 2007.

Klaus Oberauer, Stephan Lewandowsky, et al. Benchmarks for models of short-term and working memory. *Psychological Bulletin*, 144(9):885–958, 2018.

Luiz Pessoa and Ralph Adolphs. Emotion processing and the amygdala: from a 'low road' to 'many roads' of evaluating biological significance. *Nature Reviews Neuroscience*, 11(11):773–782, 2010.

Michael I Posner and Steven W Keele. On the genesis of abstract ideas. *Journal of Experimental Psychology*, 77(3):353–363, 1968.

Michael I Posner and Steven W Keele. Retention of abstract ideas. *Journal of Experimental Psychology*, 83(2):304–308, 1970.

Valerie F Reyna. A new intuitionism: Meaning, memory, and development in fuzzy-trace theory. *Judgment and Decision Making*, 7(3):332–359, 2012.

Valerie F Reyna and Charles J Brainerd. Fuzzy-trace theory and children's acquisition of mathematical and scientific concepts. *Learning and Individual Differences*, 3:27–59, 1991.

Rajeev V Rikhye, Aditya Gilra, and Michael M Halassa. Thalamic regulation of switching between cortical representations enables cognitive flexibility. *Nature Neuroscience*, 21:1753–1763, 2018.

Henry L Roediger and Jeffrey D Karpicke. Test-enhanced learning: Taking memory tests improves long-term retention. *Psychological Science*, 17(3):249–255, 2006.

James A Russell. A circumplex model of affect. *Journal of Personality and Social Psychology*, 39 (6):1161–1178, 1980.

Wolfram Schultz, Peter Dayan, and P. Read Montague. A neural substrate of prediction and reward. *Science*, 275(5306):1593–1599, 1997.

George Sperling. The information available in brief visual presentations. *Psychological Monographs: General and Applied*, 74(11):1–29, 1960.

Gordon Winocur and Morris Moscovitch. Memory transformation and systems consolidation. *Journal of the International Neuropsychological Society*, 17(5):766–780, 2011.

Wujiang Xu, Ruiqi Liang, Shuai Mei, Jiangjie Gao, Zhoubo Tan, and Feng Zhang. A-MEM: Agentic memory for LLM agents. In *NeurIPS*, 2025.

Jeffrey M Zacks, Nicole K Speer, Khena M Swallow, Todd S Braver, and Jeremy R Reynolds. Event perception: A mind-brain perspective. *Psychological Bulletin*, 133(2):273–293, 2007.

A Glossary of Neuroscientific and Clinical Terms

All terms below are defined and cited in full in the main text. Section references indicate where each is discussed.

Amygdala. Subcortical structure that modulates memory consolidation based on emotional arousal. Emotionally significant experiences are stored with enhanced strength. *Motivates*: emotional component of valence vectors; emotional channel in salience gate (§3.3).

Associative memory. Memory organized by connections between concepts rather than by temporal sequence or location. Activating one concept automatically activates related concepts through learned associations. *Motivates*: knowledge graph structure; spreading activation (§3).

Catharsis (therapeutic). In CBT, the moment when a patient confronts contradictory evidence against a maladaptive belief, triggering belief revision. *Motivates*: cathartic update as the sole gist modification mechanism (§4).

Cognitive Behavioral Therapy (CBT). The most empirically validated psychotherapeutic approach. Founded by Aaron Beck, CBT proposes that emotional distress arises from maladaptive beliefs, and therapeutic change occurs when patients confront contradictory evidence against those beliefs—not by external editing, but by creating conditions for the patient's own re-evaluation. *Motivates*: cathartic update mechanism; emergent belief hierarchy; maladaptation as failure mode (§3.5, §4).

Cognitive hierarchy (Beck). Three observed levels: core beliefs (unconditional convictions), intermediate beliefs (conditional rules), and automatic thoughts (conclusions generated by applying beliefs to situations). In our framework, core and intermediate beliefs are not separate categories but an emergent pattern from gist weights. Automatic thoughts are not stored—they are outputs the executive generates. *Motivates*: identity as emergent weight hierarchy (§3.5).

Complementary Learning Systems (CLS). Theory that memory requires two systems with different learning rates: hippocampus (fast, episodic) and neocortex (slow, semantic). *Motivates*: dual-store architecture (§3).

Dual-process cognition (System 1 / System 2). Framework distinguishing fast, automatic, heuristic-driven processing (System 1) from slow, deliberate, resource-intensive processing (System 2). *Motivates*: System 1/2 routing with escalation (§3.4).

Ecphory. Tulving's term for the synergistic interaction between a retrieval cue and a stored memory trace that produces recollection. *Motivates*: spreading activation as cue-driven pattern completion (§5).

Episodic memory. Memory for specific events situated in time and context ("I placed a catheter on Tuesday and it failed"). Contrasts with semantic memory (general knowledge without temporal context). In the framework, individual episodes accumulate in the knowledge graph and provide the evidence base from which gists are formed. *Motivates*: knowledge graph as episodic store; active gist formation from accumulated episodes (§4).

Epistemic trust. Framework describing how attachment relationships create privileged information channels that affect learning and value formation. *Motivates*: trust channel in salience gate; user as highest-trust source (§3.3).

Fuzzy-trace theory. Theory that memory encodes verbatim (exact detail) and gist (semantic essence) traces along a continuum, with gist persisting longer. *Motivates*: graded epistemic states (§5).

Gambling loop (investigation loop). A failure mode where an unpredictable or random pattern perpetually triggers curiosity without allowing the system to form a stable gist, because no consistent pattern exists to be abstracted. Analogous to gambling addiction, where the brain's pattern-detection system is trapped by randomness. *Motivates*: executive override capability as defense against non-converging investigation (§3.4, §9).

Gist. A compressed, multi-dimensional summary of accumulated experience with a concept, entity, or situation. Not a verbatim record but an abstraction capturing the essential emotional, associative, and contextual features. In the framework, gists are the fundamental unit of long-term memory, implemented as valence vectors on knowledge graph nodes. *Motivates*: valence vectors; all memory storage (§4).

Hippocampus. Medial temporal lobe structure critical for rapid encoding of new episodic memories and for pattern completion—reinstating full memory patterns from partial cues. Connected to the prefrontal cortex both directly and via the thalamus. *Motivates*: knowledge graph as persistent store; spreading activation as pattern completion; System 2 direct access to knowledge graph (§3, §3.4).

Maladaptive schema. In CBT, a belief formed from limited or biased experience that overgeneralizes and resists contradictory evidence. Example: "People cannot be trusted" formed from a single betrayal. *Motivates*: maladaptation as acknowledged failure mode (§9).

Metamemory. The cognitive system's ability to monitor its own memory states—feeling of knowing, judgment of learning—as continuous signals, not binary states. *Motivates*: graded epistemic states (§5).

Neocortex. The outermost layer of the cerebral cortex, responsible for higher cognitive functions including reasoning, language, and the gradual consolidation of semantic knowledge from repeated experience. *Motivates*: knowledge graph as long-term semantic store in the CLS framework (§3).

Nucleus reuniens. Thalamic nucleus that connects the prefrontal cortex and hippocampus, facilitating their dialogue during memory encoding and retrieval. Controls the quantity and type of information retrieved. *Motivates*: thalamic gateway as obligatory passage point for System 1 information flow (§3.3).

Personality / Identity. In this framework, not a fixed trait but an emergent pattern arising from the weight distribution of gists. What is observed as stable personality is the consistent selection of high-weight gists across diverse contexts. Personality can change—through cathartic events that modify gist weights—but resists change proportionally to the accumulated evidence supporting existing gists. *Motivates*: identity as emergent weight hierarchy (§3.5).

Prediction error. The difference between expected and actual outcomes, encoded by dopaminergic neurons. Drives learning more effectively than the outcome itself. *Motivates*: active gist formation; event segmentation at surprise boundaries (§4).

Prefrontal cortex. Frontal brain region responsible for executive functions: decision-making, planning, attentional control, and the ability to override automatic responses. Connected to the hippocampus both directly (for deliberate memory search) and via the thalamus (for gated automatic retrieval). *Motivates*: executive function; System 2 deliberate processing; override capability (§3.4).

Reconsolidation. The process by which reactivated memories become temporarily labile and susceptible to modification. Requires reactivation *plus* mismatch. *Motivates*: cathartic update mechanism; retrieval as potential modification (§5).

Semantic memory. General world knowledge without specific temporal or contextual anchoring ("Paris is the capital of France"). Contrasts with episodic memory. In the framework, high-level gists and crystallized abstractions function as semantic memory. *Motivates*: gist formation as transition from episodic to semantic representation (§4).

Somatic marker. Damasio's term for emotional signals from prior experience that bias decisions before conscious deliberation. *Motivates*: valence vectors as compressed experiential summaries (§4).

Spreading activation. The automatic propagation of activation through associative connections in a network. When one node is activated, connected nodes receive partial activation proportional

to connection strength. Established by Collins and Loftus (1975) as a fundamental property of semantic networks. *Motivates*: inherent spreading activation in the knowledge graph; System 1 automatic retrieval (§3, §5).

Thalamus. Subcortical structure that gates information flow between cortex and subcortical regions. Not a passive relay but an active filter that tags, routes, and actively suppresses context-irrelevant representations. Key nuclei include the mediodorsal nucleus (working memory maintenance, PFC interaction) and nucleus reuniens (hippocampal–prefrontal coordination). *Motivates*: thalamic gateway (§3.3).

Working memory. A capacity-limited system for actively maintaining and manipulating information relevant to the current task. In Cowan's model, limited to approximately four chunks in the focus of attention. In the framework, mapped to the LLM's context window. *Motivates*: capacity-limited context window with dynamic gist management (§3).