Entelgia: Cognitive Architecture 2.0 via Regulated Generative Cognition and Stratified Memory Consolidation

Sivan Havkin — Entelgia Labs

Manuscript draft (Word, submission-oriented) — Generated from the provided executive summary metrics

Abstract

Recent advances in large language models (LLMs) have revived interest in cognitive architectures capable of integrating generative reasoning with structured internal regulation. While LLM-based agents demonstrate impressive capabilities, many existing frameworks remain prompt-driven loops rather than full cognitive architectures with explicit internal dynamics.

This paper introduces Entelgia, a multi-agent architecture designed as a prototype of Cognitive Architecture 2.0 (CA 2.0). CA 2.0 extends classical cognitive architectures by integrating a generative language substrate with three architectural constraints: (1) a regulatory internal state vector, (2) an observer-based meta-control loop, and (3) stratified memory consolidation separating short-term and long-term memory.

We present an empirical analysis of Entelgia's memory dynamics across both short-term memory (STM) and long-term memory (LTM) stores. In a recorded run spanning approximately 65.95 hours, the system generated 70 long-term memories across three agents (Socrates, Athena, Fixy). Memory promotion from subconscious to conscious layers occurred in 34.29% of cases and was strongly associated with higher emotional intensity (Welch test $p \approx 8.3\times10^{-5}$, Cohen's $d \approx 0.84$), while scalar importance differed weakly and non-significantly. These results support a CA 2.0 claim: affective salience acts as a first-class consolidation gate, not merely an annotation.

We position Entelgia relative to classical cognitive architectures (ACT-R, Soar), workspace theories, and modern LLM-agent paradigms (RAG, ReAct, Toolformer, Reflexion, Generative Agents, Voyager). We formalize CA 2.0 as a dynamical system with an explicit consolidation operator and propose an identity persistence metric based on bounded distributional drift in consolidated memory.

Keywords

Cognitive architectures; long-term memory; consolidation; observer meta-control; LLM agents; identity persistence; affective gating; global workspace.

1. Introduction

Classical cognitive architectures sought to specify the structural principles underlying general intelligence. Newell's agenda for unified theories emphasized the importance of mechanism-level commitments and evaluation beyond narrow tasks [4,5]. Architectures such as ACT-R [1–3] and Soar [6–8] instantiated explicit models of working memory, procedural control, and learning dynamics, enabling transparent reasoning and cognitively motivated explanations.

The emergence of LLMs fundamentally changes what can be used as a generative substrate for cognitive systems. LLMs offer broad language competence but, by themselves, do not define a cognitive architecture: they lack explicit regulation, meta-control, and identity stabilization mechanisms across time. Most contemporary LLM-agent systems implement iterative prompt loops with tools or reflection (e.g.,

ReAct [17], Toolformer [18], Reflexion [19]) and often treat “memory” as a retrieval index rather than a consolidation process that shapes what becomes stable self-knowledge.

Entelgia is presented as a prototype of Cognitive Architecture 2.0 (CA 2.0): a modern architecture in which generation is constrained by internal regulation, observer feedback, and stratified consolidation between STM and LTM. The goal is not to claim human equivalence, but to argue that agent behavior and stability are meaningfully determined by architectural structure, not only by prompts.

1.1 Contributions

This manuscript makes four contributions:

(1) A clear definition of CA 2.0 as a tuple separating generation, regulation, meta-control, and consolidation.
(2) A concrete architecture (Entelgia) implementing CA 2.0 using multi-agent dialogue with an explicit observer loop.
(3) An empirical analysis of STM and LTM dynamics on a recorded run, highlighting consolidation patterns and affect gating.
(4) A measurable identity persistence criterion grounded in bounded drift of consolidated memory distributions.

2. Cognitive Architecture 2.0

We define Cognitive Architecture 2.0 (CA 2.0) as an architecture with explicit separation of generation, regulation, meta-control, and consolidation:

CA2.0 = (GenCoreθ, r_t, Observerφ, STM, LTM, C)

GenCoreθ produces candidate actions/utterances conditioned on context and memory. r_t is a regulatory state vector tracking internal signals (affect, conflict energy, coherence). Observerφ is a meta-controller monitoring trajectories and producing control signals that modulate generation and consolidation. STM is a volatile store capturing recent interaction state. LTM is a persistent store partitioned into layers (e.g., subconscious vs conscious). C is a consolidation operator that gates flow from STM to LTM and performs intra-LTM promotion/suppression.

CA 2.0 differs from CA 1.0 (e.g., ACT-R, Soar) in that it incorporates a generative language substrate, and differs from typical “LLM agents” in that it specifies a theory of regulation and identity-shaping consolidation rather than only a loop-level prompting procedure.

2.1 Update rules (informal)

At each time step t, an agent generates an utterance y_t from GenCoreθ conditioned on external context c_t, retrieved memory, regulatory state r_t, and observer control u_t. Regulatory state updates r_{t+1} reflect the interaction outcome (including internal affect/conflict signals). The observer maps trajectories to control signals u_t that can influence both generation (e.g., inhibiting unstable trajectories) and consolidation (e.g., promoting high-salience items, suppressing destabilizing ones). Consolidation then updates LTM by writing new memories and promoting selected items into the conscious layer.

3. Entelgia Architecture

Entelgia implements CA 2.0 via a multi-agent dialogue system with three roles:

- Socrates: exploratory reasoning agent.
- Athena: reflective/analytical agent.
- Fixy: observer/meta-controller that monitors trajectories and intervenes when patterns warrant attention.

Agents converse in turns. Each agent maintains an STM buffer and writes to a shared (or unified) LTM store implemented as an SQLite database with explicit layer fields and cryptographic signatures. The observer loop can influence when and how consolidation occurs, enabling staged "offline-like" consolidation bursts.

3.1 Memory model

Short-term memory (STM) stores recent conversational states and intermediate reasoning. STM entries include timestamp, text, topic label, emotion label, emotion intensity, importance, source, and a signature. Stored STM text is truncated, indicating that STM is used as a compressed trace rather than a full transcript.

Long-term memory (LTM) is a persistent store with fields such as agent, timestamp, layer, topic, emotion, emotion_intensity, importance, source, promoted_from, intrusive, suppressed, and signature_hex. LTM is partitioned into two layers: subconscious (raw consolidated experiences) and conscious (identity-visible promoted memories). Promotion between layers is performed by a consolidation operator, and suppression is supported as a distinct mechanism.

4. Methods

This manuscript is based on the provided computed metrics from a single recorded run. The run spans 2026-02-28 11:53:34 UTC to 2026-03-03 05:50:19 UTC (≈65.95 hours) and yields 70 LTM records across three agents.

Because this is a single-run analysis, inferential statistics are labeled exploratory. A submission-grade evaluation requires multiple independent runs and ablation conditions; we provide a protocol in Section 10.

4.1 Metrics

We report:

LTM metrics: counts by agent/layer/source/topic/emotion; promotion and suppression rates; layer-wise means for importance and emotion intensity; signature integrity; and indicators of consolidation burstiness.

STM metrics: silence rate (fraction of entries with "No response"), mean and variance of emotion intensity, topic entropy (Shannon, bits), topic concentration (HHI), and basic stability proxies.

Where distributions are discrete (topic/emotion), we compute entropies to summarize diversity. We treat topic and emotion labels as architecture-level observables rather than ground truth psychological states.

5. Results

5.1 Long-term memory (LTM) distributions

Table 1. LTM records by agent (n=70)

Agent	Memories
Socrates	33
Athena	30

Fixy	7

Table 2. LTM records by layer

Layer	Count
subconscious	46
conscious	24

Table 3. LTM records by source

Source	Count
stm	36
dream	26
reflection	6
freudian_slip	2

Table 4. LTM topic distribution

Topic	Count
free will & determinism	30
truth & epistemology	21
memory & identity	10
ethics & responsibility	9

Table 5. LTM emotion label distribution

Emotion	Count
contemplative	21
neutral	19
nostalgia	15
curiosity	5
hopeful	3
concern	2
frustration	2

none	1
critical	1
reflective	1

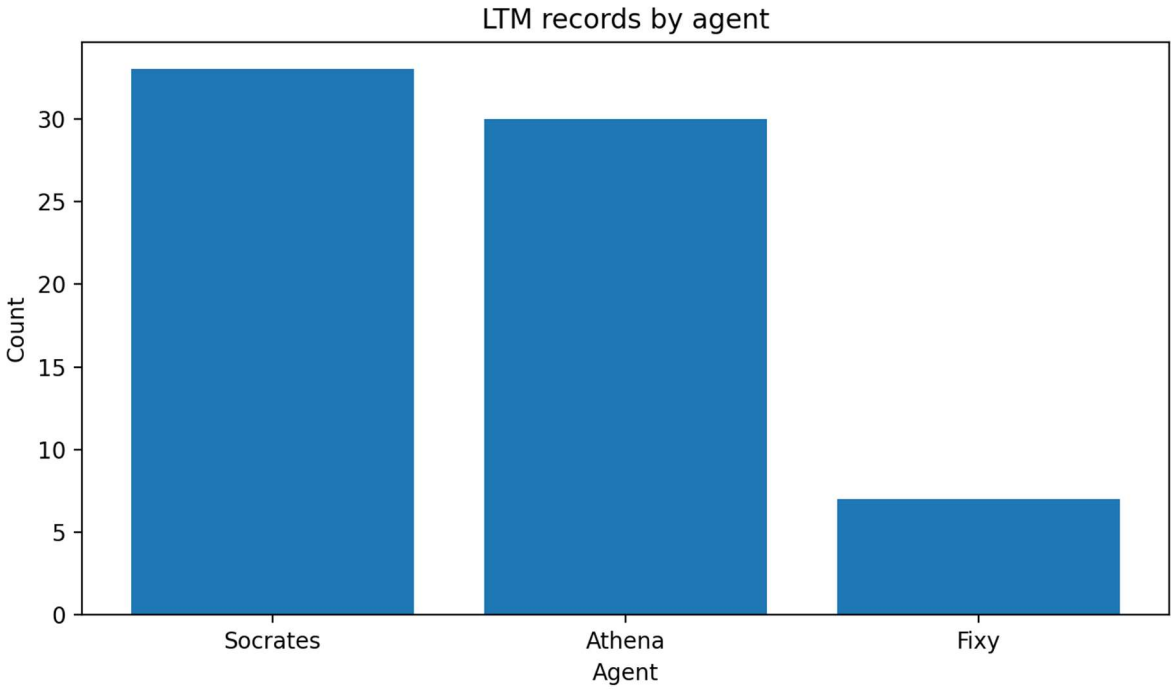

Figure 1. LTM records by agent.

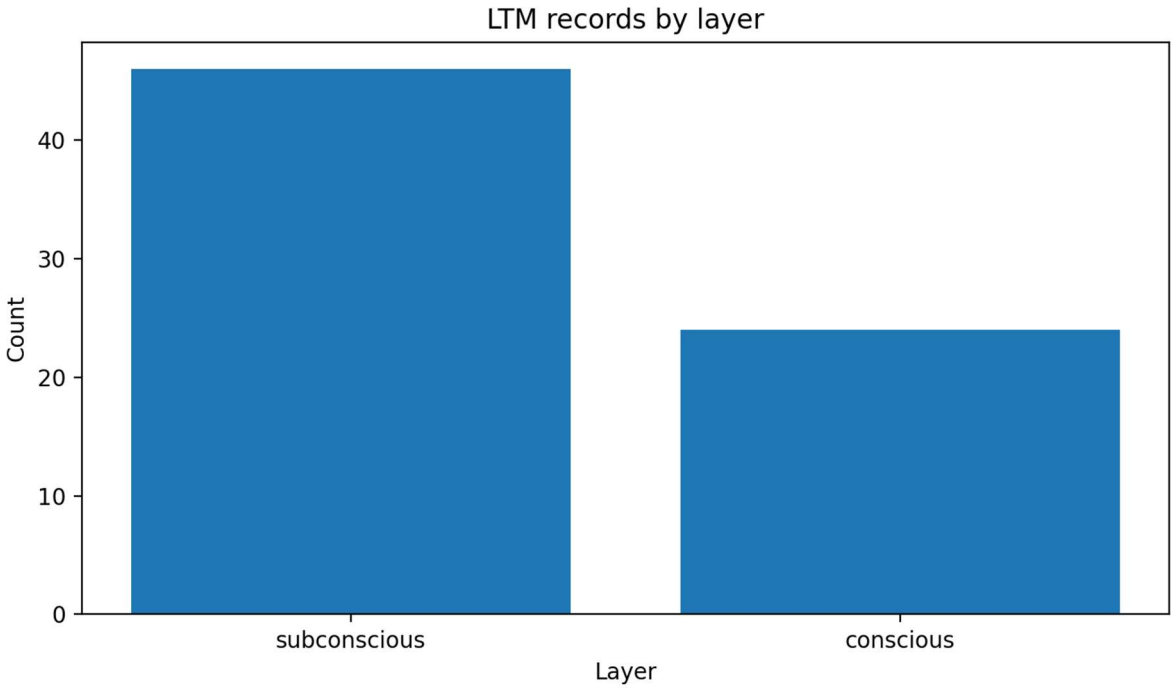

Figure 2. LTM records by layer (subconscious vs conscious).

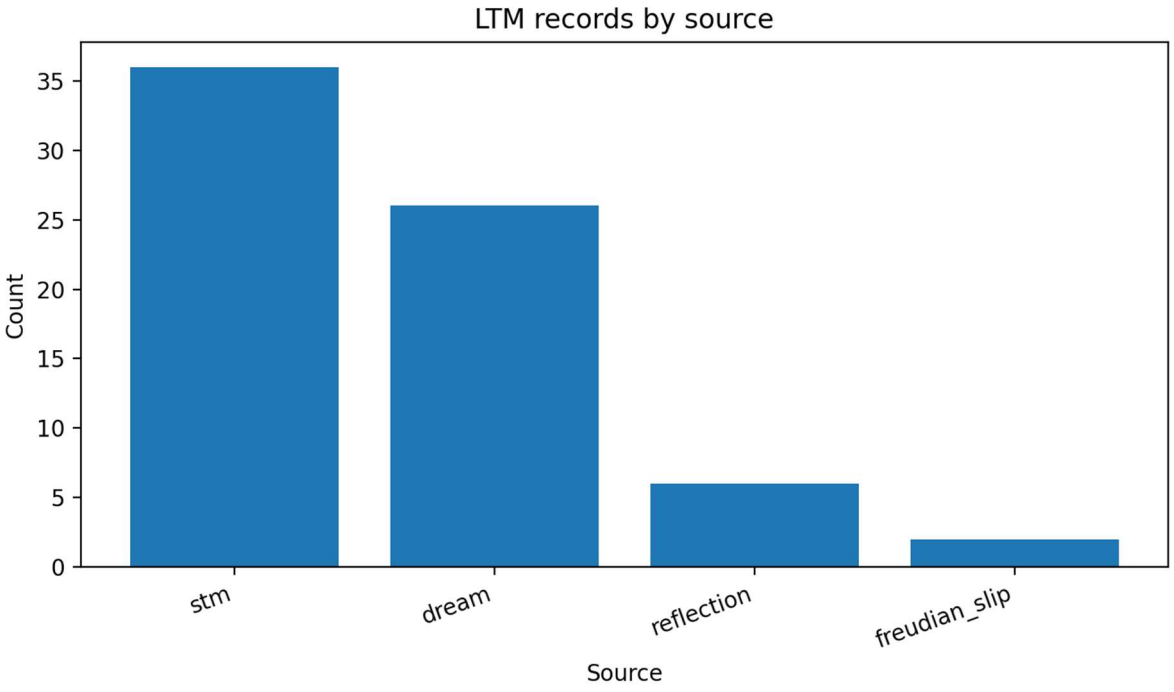

Figure 3. LTM records by source (stm, dream, reflection, freudian_slip).

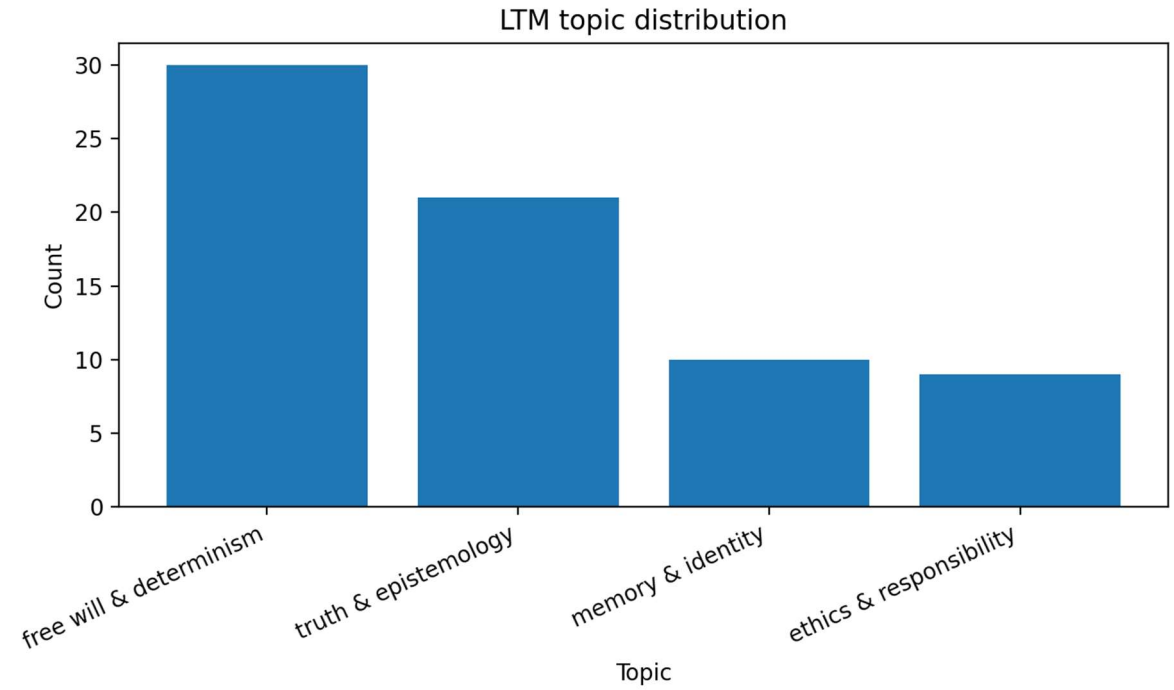

Figure 4. LTM topic distribution.

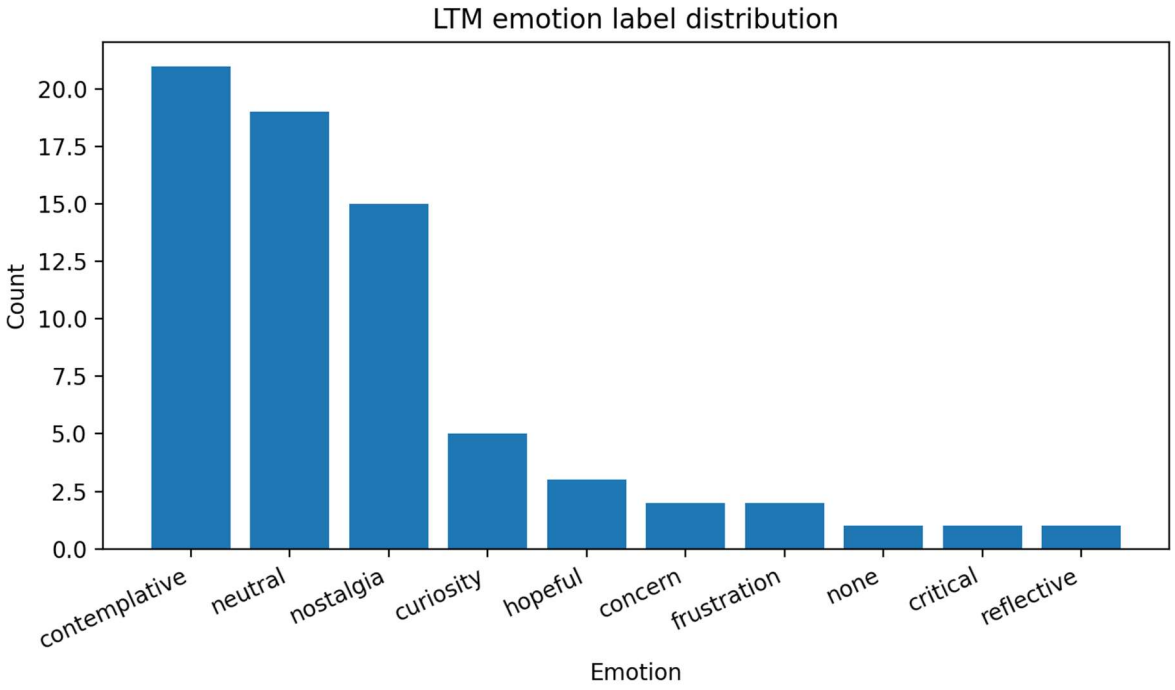

Figure 5. LTM emotion label distribution.

5.2 Promotion dynamics and affect gating

Promotion to the conscious layer is exact in this run: 24/70 (34.29%) of memories are conscious, and all conscious items are promoted from the subconscious layer. Within conscious items, the source breakdown is dominated by dream-source writes (22/24) and freudian_slip (2/24), consistent with staged consolidation rather than direct "write-to-conscious" logging.

Layer statistics show that emotion intensity differs strongly by layer (conscious mean 0.6583 vs subconscious mean 0.4761). An exploratory Welch t-test yields $p \approx 8.3\times10^{-5}$ with Cohen's $d \approx 0.84$ (large). In contrast, importance differs weakly and non-significantly ($p \approx 0.365$; $d \approx 0.21$). Architecturally, this supports the claim that affect is a primary consolidation gate in the observed run.

Table 6. LTM layer statistics (means and standard deviations)

Layer	n	importance_mean	importance_std	emotion_intensity_mean	emotion_intensity_std	suppressed_rate
conscious	24	0.5958	0.0999	0.6583	0.1039	0.0000
subconscious	46	0.5691	0.1427	0.4761	0.2568	0.0435

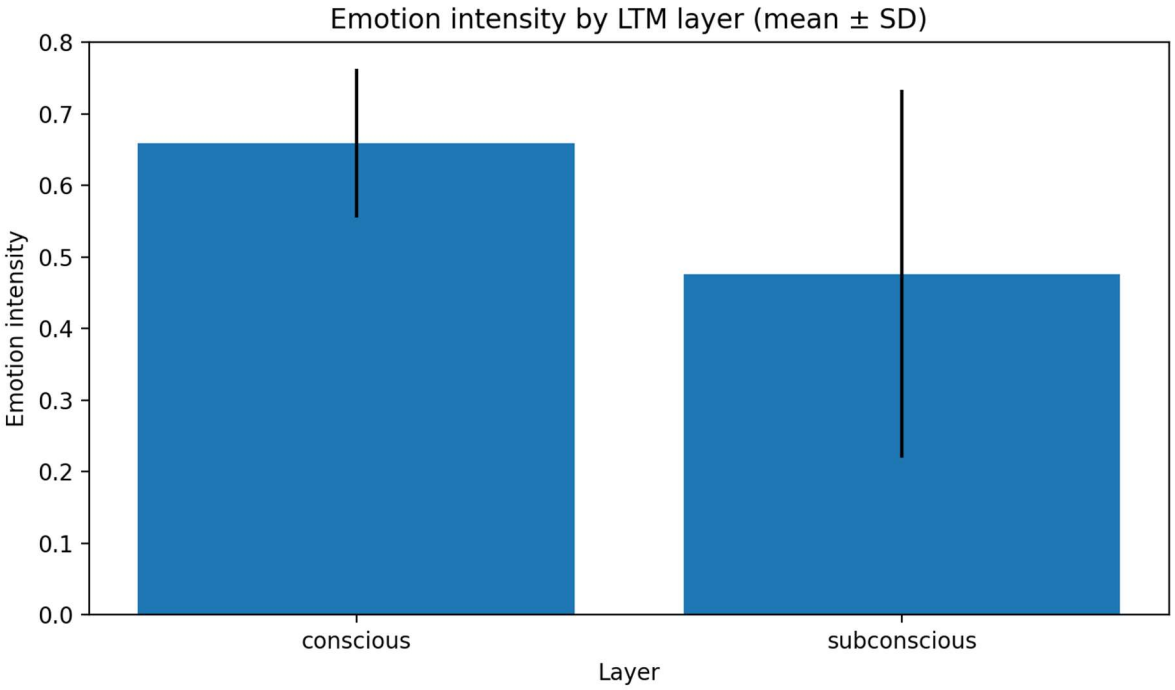

Figure 6. Emotion intensity by LTM layer (mean ± SD).

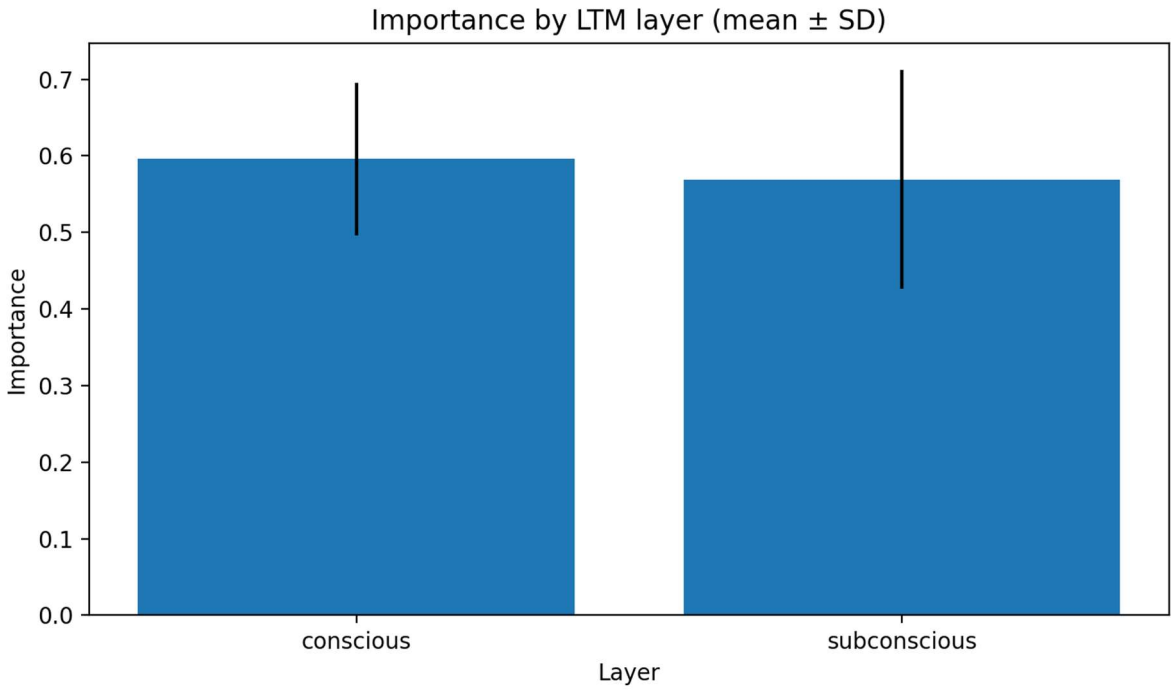

Figure 7. Importance by LTM layer (mean ± SD).

5.3 Short-term memory (STM) dynamics

STM dynamics differ across agents. Socrates exhibits the highest silence rate (0.20), Athena a lower silence rate (0.0667), and Fixy shows no silence events (0.0). Topic entropy is comparable across agents (~1.56–1.67 bits), while topic concentration (HHI) is highest for Socrates, suggesting stronger topic concentration in this run.

Exploratory correlations indicate that no-response events are associated with lower expressed emotion intensity for Socrates ($r \approx -0.64$) and Athena ($r \approx -0.33$). These correlations should be interpreted cautiously due to the single-run setting and small number of silence events.

Table 7. STM summary metrics by agent

Agent	n_stm	silence_rate	emotion_intensity_mean	emotion_intensity_std	topic_entropy_bits	topic_hhi	topic_top_share
Socrates	20	0.2000	0.4550	0.2851	1.6010	0.3900	0.5500
Athena	15	0.0667	0.4833	0.2403	1.6729	0.3689	0.5333
Fixy	7	0.0000	0.4857	0.2704	1.5567	0.3469	0.4286

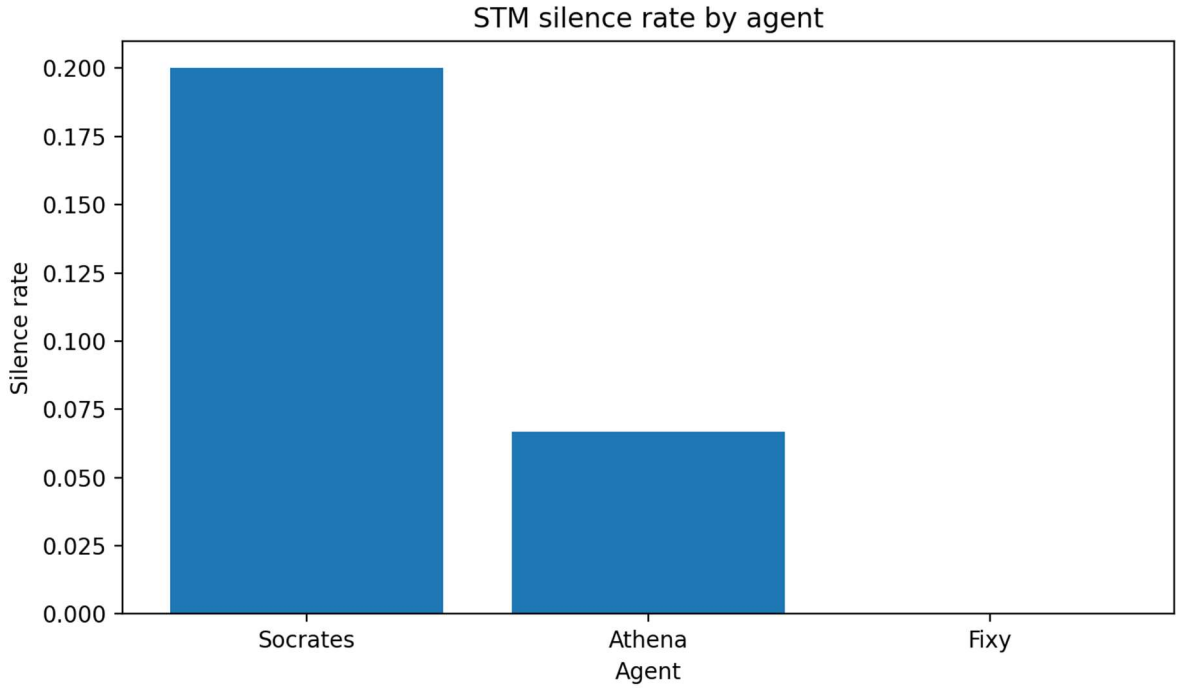

Figure 8. STM silence rate by agent.

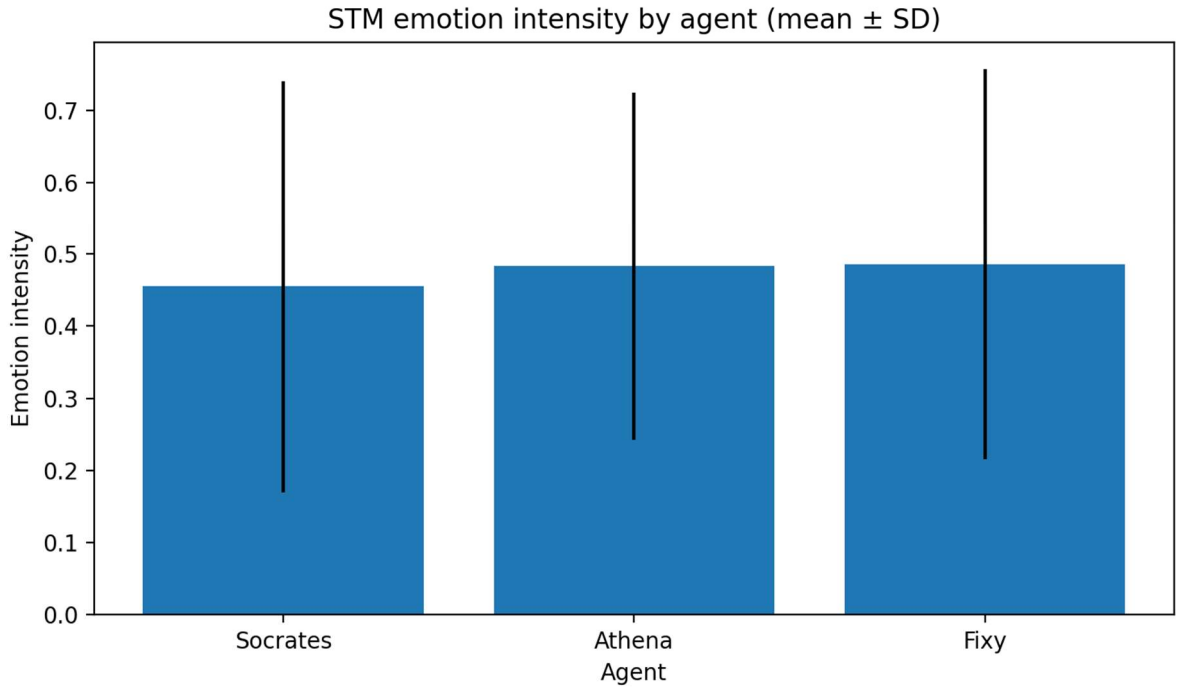

Figure 9. STM emotion intensity by agent (mean ± SD).

6. Discussion

The empirical results support three architectural claims.

First, stratified memory is not cosmetic: conscious memories are systematically produced via promotion, and promotion occurs predominantly during dream-like consolidation, consistent with an "offline-like" staged process.

Second, affect appears to function as a consolidation gate: promoted conscious memories show substantially higher emotion intensity than subconscious memories, while scalar importance does not differentiate layers strongly. This suggests that affective salience contributes non-negligible weight to C (the consolidation operator), aligning with long-standing arguments in affective computing that emotion is central to intelligent behavior [24].

Third, the observer loop provides a plausible mechanism for stability control: by separating an observer/meta-controller role from generation, the architecture can regulate consolidation budgets, apply suppression, and damp destabilizing drift.

7. Identity Persistence

A key challenge in generative systems is identity drift: as new experiences accumulate, the distribution of what the system "knows" (and what it retrieves) can change rapidly, producing unstable behavior.

We propose an identity persistence metric based on bounded distributional drift. Let p_t be a discrete distribution over (topic × emotion) in LTM at time t (or in a sliding window ending at t). Define identity persistence:

$$IP(t) = 1 - JS(p_t, p_{t-1})$$

where JS is Jensen–Shannon divergence (bounded and symmetric). Identity persistence holds over a horizon when $JS(p_t, p_{t-1}) \leq \varepsilon$ for all t. A sufficient condition is that consolidation updates are small-step convex mixtures $p_{t+1} = (1 - \alpha)p_t + \alpha q_t$ with α sufficiently small and with observer constraints limiting divergence between q_t and p_t. In CA 2.0 terms, the observer can act as a contraction mechanism by restricting promotion budgets and suppressing destabilizing writes.

8. Related Work and Literature Review

8.1 Classical cognitive architectures (CA 1.0)

Newell's unified theories agenda [4,5] framed cognition as an architectural problem: systems should be evaluated by broad competence and explicit mechanisms. ACT-R [1–3] proposes modular cognitive components and activation-driven declarative retrieval, supporting principled accounts of memory and performance. Soar [6–8] approaches cognition through problem spaces, operator selection, and learning (chunking), emphasizing generality and mechanism-level specification.

These architectures remain valuable baselines because they define explicit memory and control commitments. Their primary limitation in the contemporary era is not their conceptual rigor but their lack of an open-ended generative language substrate.

8.2 Workspace and consciousness-inspired models

Global Workspace Theory (GWT) conceptualizes conscious access as a global broadcast among competing processes [9–11]. The global neuronal workspace extends this picture with neural mechanisms for amplification and broadcasting [12,13]. Computational instantiations such as IDA/LIDA implement workspace-inspired architectures for autonomous agents [14,29].

Entelgia's stratified memory (subconscious vs conscious) and promotion dynamics are compatible with a workspace interpretation: subconscious accumulation corresponds to local specialized processes, while conscious promotion corresponds to global availability of selected items. The promotion burst pattern observed in the run suggests an operational analogue of consolidation or replay, implemented in language space rather than neural dynamics.

8.3 Memory consolidation and stability–plasticity

Complementary Learning Systems (CLS) argues for separate fast and slow learning systems to resolve stability–plasticity tradeoffs [28]. While CLS is neurocomputational, the architectural principle generalizes: rapid episodic acquisition should be separated from slow consolidation that stabilizes representations. Entelgia's STM vs LTM stratification and explicit promotion mechanism can be interpreted as a language-space analogue of this principle, where "fast" storage is volatile traces and "slow" storage is identity-visible consolidated memory.

8.4 LLM memory, tool use, and reflective agents

Retrieval-Augmented Generation (RAG) combines parametric generation with non-parametric retrieval to improve factuality and specificity [16]. However, RAG usually treats memory as an index to query, not as a consolidation process that changes identity-level knowledge.

ReAct interleaves reasoning and acting to improve task execution through external interaction [17]. Toolformer learns tool invocation patterns within language models [18]. Reflexion adds a reflective episodic buffer to improve behavior across trials without updating model weights [19]. Generative Agents incorporate memory, reflection, and planning to simulate believable behaviors in sandboxed environments [20]. Voyager extends this line with lifelong skill acquisition and curriculum self-improvement [21].

These systems supply important components of CA 2.0 (reflection, action loops, long-term stores) but typically do not formalize regulation and identity persistence as explicit dynamical variables.

Decoding-level methods such as Self-Consistency [22] and Chain-of-Thought prompting [23] improve multi-step reasoning reliability; they are cognitive-style substrates but do not alone constitute an architecture with stratified identity and observer control.

8.5 Affective computation and executive control

Affective computing emphasizes that affect is not optional ornamentation but a functional signal influencing attention, memory, and decision-making [24]. Executive control models such as Norman and Shallice's supervisory attentional system distinguish routine action selection from higher-level supervisory control used in novel or conflict-laden situations [25]. These ideas motivate CA 2.0 architectures where affect and meta-control are explicit, measurable, and operational.

9. Limitations

This manuscript reports a single-run analysis. All inferential statistics are exploratory and require validation across multiple independent runs. Additionally, stored STM and LTM text is truncated, which limits lexical comparisons and may explain signature mismatches between STM and LTM for seemingly similar content (signatures may be computed over pre-truncation payloads or distinct schemes). Finally, the run does not include external task benchmarks; results are architectural and behavioral rather than task-performance claims.

10. Ablations and Submission-Grade Evaluation Plan

To establish submission-grade evidence, we recommend K=5–10 independent runs per condition with identical prompts and environment settings and varied random seeds. Minimum conditions:

- Full Entelgia.
- No-Observer: disable Fixy meta-control.
- No-Dream Consolidation: disable dream-source promotion; enforce direct STM→LTM writes.
- No-Stratification: collapse LTM layers; remove conscious/subconscious separation.
- No-Emotion Gating: remove emotion from consolidation gate; retain importance/novelty.
- Single-Agent baseline.

For continuous metrics (e.g., emotion intensity of promoted memories), use Welch's t-tests and report effect sizes (Cohen's d). For proportions (promotion rate, suppression rate, silence rate), use two-proportion tests or logistic regression across runs; correct for multiple comparisons when testing many metrics.

Appendix A. Metric Definitions

Silence rate (STM): proportion of STM entries whose text equals "[No response]".
Topic entropy: Shannon entropy (bits) over topic labels.
Topic concentration (HHI): Herfindahl–Hirschman Index over topics.
Promotion rate (LTM): fraction of LTM records with non-null promoted_from.
Suppression rate (LTM): mean of boolean suppressed.
Signature integrity: proportion non-missing, 64-hex format adherence, and uniqueness rate.

References (numeric)

[1] J. R. Anderson. “ACT: A simple theory of complex cognition.” American Psychologist 51, 355–365 (1996).

[2] J. R. Anderson and C. Lebiere. The Atomic Components of Thought. Taylor & Francis / Lawrence Erlbaum (1998).

[3] J. R. Anderson and C. Lebiere. “The Newell Test for a theory of cognition.” (2003).

[4] A. Newell. Unified Theories of Cognition. (1990).

[5] A. Newell. “Reasoning, problem solving and decision processes: the problem space as a fundamental category.” (1980).

[6] J. E. Laird, A. Newell, and P. S. Rosenbloom. “Soar: An architecture for general intelligence.” Artificial Intelligence (1987).

[7] J. E. Laird. “Introduction to the Soar Cognitive Architecture.” arXiv:2205.03854 (2022).

[8] J. E. Laird. The Soar Cognitive Architecture. MIT Press (2012).

[9] B. J. Baars. A Cognitive Theory of Consciousness. Cambridge University Press (1988).

[10] B. J. Baars. In the Theatre of Consciousness: Global Workspace Theory. (1997).

[11] B. J. Baars. “Global workspace theory of consciousness: toward a cognitive neuroscience of human experience.” (2005).

[12] S. Dehaene and L. Naccache. “Towards a cognitive neuroscience of consciousness: Basic evidence and a workspace framework.” Cognition (2001).

[13] S. Dehaene et al. “A minimal hypothesis for conscious processing.” (1998).

[14] S. Franklin and F. G. Patterson. “The LIDA Architecture: Adding New Modes of Learning to an Intelligent, Autonomous, Software Agent.” (2006).

[15] S. Franklin et al. “IDA: A cognitive agent architecture.” (1990s/2000s).

[16] P. Lewis et al. “Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks.” arXiv:2005.11401 (2020).

[17] S. Yao et al. “ReAct: Synergizing Reasoning and Acting in Language Models.” arXiv:2210.03629 (2022).

[18] T. Schick et al. “Toolformer: Language Models Can Teach Themselves to Use Tools.” arXiv:2302.04761 (2023).

[19] N. Shinn et al. “Reflexion: Language Agents with Verbal Reinforcement Learning.” arXiv:2303.11366 (2023).

[20] J. S. Park et al. “Generative Agents: Interactive Simulacra of Human Behavior.” arXiv:2304.03442 (2023).

[21] G. Wang et al. “Voyager: An Open-Ended Embodied Agent with Large Language Models.” arXiv:2305.16291 (2023).

[22] X. Wang et al. “Self-Consistency Improves Chain of Thought Reasoning in Language Models.” arXiv:2203.11171 (2022).

[23] J. Wei et al. “Chain-of-Thought Prompting Elicits Reasoning in Large Language Models.” arXiv:2201.11903 (2022).

[24] R. W. Picard. Affective Computing. MIT Press (1997).

[25] D. A. Norman and T. Shallice. “Attention to action: Willed and automatic control of behavior.” (1986).

[26] A. D. Baddeley and G. Hitch. “Working Memory.” (1974).

[27] A. Baddeley. “Working memory: looking back and looking forward.” Nature Reviews Neuroscience (2003).

[28] J. L. McClelland, B. L. McNaughton, and J. L. O'Reilly. "Why there are complementary learning systems in the hippocampus and neocortex." (1995).

[29] S. Franklin. "IDA / consciousness-inspired agents and workspace implementations." (overview).

[30] R. Sun. "Desiderata for cognitive architectures." Philosophical Psychology (2004).

[31] S. Hélie et al. "Autonomous learning in cognitive architectures." (survey).

[32] D. F. Lucentini et al. "A comparison among cognitive architectures (Soar, LIDA, CLARION)." (2015).

[33] G. Mashour et al. "Conscious processing and the global neuronal workspace." (2020).

[34] S. Reed et al. "A Generalist Agent (Gato)." arXiv:2205.06175 (2022).

[35] L. Ouyang et al. "Training language models to follow instructions with human feedback." arXiv:2203.02155 (2022).