MEMORY ARCHITECTURE IN S-AI-GPT: FROM CONTEXTUAL ADAPTATION TO HORMONAL MODULATION

Said Slaoui

Mohammed V University, Rabat, Morocco

ABSTRACT

This article presents a biologically inspired memory architecture embedded within the Sparse Artificial Intelligence – Generative Pretrained Transformer (S-AI-GPT) conversational framework. Addressing the limitations of stateless Large Language Models (LLMs), the system integrates three complementary components: a Dynamic Contextual Memory (DCM) for short-term working retention, a GPTMemoryAgent for long-term personalized storage, and a GPT-MemoryGland for affective trace encoding and modulation. These components are orchestrated by a hormonal engine, enabling adaptive forgetting, emotional persistence, and context-aware prioritization of memory traces. Unlike typical passive memory modules, this architecture introduces an active, symbolic, and controllable memory system: memory traces can trigger internal hormonal signals, are stored in a structured and interpretable form, and can be selectively reinforced, inhibited, or reorganized by the GPT-MetaAgent. The proposed model provides a promising foundation for building frugal, adaptive, and explainable lifelong memory systems in conversational AI.

KEYWORDS

Memory in AI, Conversational Agents, Sparse Activation, Hormonal Modulation, Personalized Dialogue, Emotional Trace, Dynamic Memory, Modular Architecture, S-AI-GPT.

1. INTRODUCTION

1.1. Background and Motivation

The rapid evolution of **Large Language Models (LLMs)** has enabled remarkable progress in natural language understanding and generation. However, these systems often lack a key component of intelligent behavior: memory. While transformers excel at processing input within a fixed context window, they are fundamentally stateless, leading to limitations in personalization, contextual continuity, and emotional coherence. In human cognition, memory serves as the foundation of learning, decision-making, and interpersonal sensitivity. Designing AI systems that incorporate a form of adaptive, context-sensitive memory is essential to bridge the gap between reactive processing and sustained understanding [1].

1.2. Limitations of Memory in Current AI Systems

Most current LLM-based systems rely on prompt engineering or Retrieval-Augmented Generation (RAG) to simulate memory. These approaches are either manual, brittle, or limited to factual retrieval. They do not account for emotional salience, session history, or personalized interaction patterns. Furthermore, they treat memory as an external static resource, disconnected

from the cognitive dynamics of the AI system. As a result, the outputs often feel generic, disconnected, or redundant, especially in multi-turn conversations [1].

1.3. Goals and Contributions

This paper introduces the memory architecture embedded in the S-AI-GPT framework—a biologically inspired conversational AI system built on the principles of sparse activation and modular orchestration. The proposed memory system integrates three complementary components: a Dynamic Contextual Memory (DCM) that captures short-term working knowledge, a Memory Agent that stores long-term personalized and factual information, and a Memory Gland Agent that encodes and regulates emotional memory through hormonal modulation [2].

Unlike existing LLM memory extensions, which often rely on latent embeddings or static key-value retrieval, the S-AI-GPT framework introduces a memory system that is not only distributed and contextaware but also active, symbolic, and controllable. Memory traces are stored in a structured and interpretable format, enabling symbolic reasoning and selective access. Furthermore, they can autonomously trigger internal hormonal signals, modulate agent activation thresholds, and be dynamically reinforced or inhibited under the supervision of the GPT-MetaAgent. This design elevates memory from a passive storage function to a cognitively integrated substrate that actively participates in reasoning and adaptation.

The main contributions of this paper are as follows:

- A triadic architecture combining functional, emotional, and contextual memory components. – Integration of memory regulation with hormonal signals for adaptive forgetting, emotional recall, and context-aware persistence.
- Introduction of an active, symbolic, and controllable memory system, allowing internal orchestration, autonomous activation, and meta-level governance.
- Evaluation of the system's impact on conversational coherence, personalization, and computational frugality.

1.4. Organization of the Paper

The rest of the paper is organized as follows:
Section 2 reviews related work on memory systems in symbolic AI, LLMs, and biologically inspired architectures.
Section 3 presents the detailed structure of the memory modules in S-AI-GPT.
Section 4 describes the learning and feedback-driven adaptation mechanisms.
Section 5 illustrates the architecture with use cases.
Section 6 discusses theoretical and practical implications.
Finally, Section 7 concludes and outlines future directions.

2. RELATED WORK

2.1. Memory in Large Language Models (LLMs)

Large Language Models (LLMs) like GPT-3, GPT-4, and PaLM rely on context windows to simulate short-term memory. However, once the input window is exceeded, prior dialogue history is forgotten unless re-injected manually. Techniques such as Retrieval-Augmented Generation (RAG) attempt to overcome this limitation by fetching relevant knowledge from

external sources [3], [4]. Other methods, such as memory buffers or token selection strategies, aim to preserve recent dialogue states [5]. Despite their utility, these methods do not offer true personalization, persistent memory, or emotional traceability.

2.2. Biologically Inspired Memory Architectures

Drawing from cognitive neuroscience, several architectures attempt to reproduce mechanisms inspired by the hippocampus, working memory, and emotional reinforcement [6], [7], [8], [9], [10]. These models integrate principles such as forgetting curves, affective salience, and temporal decay. Historical frameworks like ACT-R[11] and CLARION [12] pioneered symbolic and biologically plausible approaches to memory, demonstrating how declarative, procedural, and affect-related processes could co-exist within a cognitive system. While these systems emphasized cognitive plausibility, they often lacked scalability or seamless integration with modern LLM-based agents.

2.3. Comparison with Retrieval-Augmented and State-Tracking Architectures

Retrieval-Augmented Generation frameworks such as RAG, ReAct, or MemGPT introduce external memory as a retrieval layer [3], [13], [14], [15]. Similarly, state-tracking architectures like DialogGPT or BlenderBot maintain dialogue states in buffers [16]. These methods improve consistency in multiturn conversations but remain disjoint from emotional memory and hormonal modulation. S-AI-GPT seeks to unify symbolic, emotional, and context-aware memory within a single, orchestrated system that adapts its memory behavior via hormonal signals and modular decomposition [1], [2].

2.4. Comparative Analysis of Related Work

To position the memory architecture of S-AI-GPT within the broader research landscape, we classify existing approaches into five categories:

- Symbolic and hybrid memory models,
- LLM-contextual memory and prompt-based extensions,
- Retrieval-augmented architectures,
- Biologically inspired memory systems, and
- Recent modular or agent-oriented frameworks.

2.4.1. Symbolic and Neuro-Symbolic Memory Systems

Classical symbolic systems such as production-rule engines and knowledge bases offered clarity, traceability, and logical consistency [17], [18]. Their limitations lie in rigidity, lack of adaptivity, and poor emotional or contextual awareness. More recently, neuro-symbolic architectures attempt to bridge this gap by embedding symbolic reasoning within neural models [19], [20]. While such systems can explain decisions post hoc, they often fail to integrate memory dynamics at runtime. S-AI-GPT extends these approaches by embedding symbolic memory traces directly into its activation logic, regulated through bio-inspired hormonal signals, enabling real-time modulation of memory relevance, emotional salience, and temporal decay [1].

2.4.2. LLM Context Windows and Prompt Engineering

Contemporary LLMs such as GPT-3 and GPT-4 simulate short-term memory through large input windows [3]. Techniques like prompt chaining and token selection heuristics provide superficial continuity [5], [21], but without persistent state, emotional depth, or long-term personalization. S-

AIGPT overcomes this limitation by incorporating a volatile yet hormonally regulated Dynamic Contextual Memory (DCM), which adapts retention policies based on urgency, fatigue, or user feedback—features absent from fixed-window models [2].

2.4.3. Retrieval-Augmented and State-Tracking Architectures

Systems such as RAG, ReAct, and MemGPT use external memory stores to retrieve relevant knowledge on demand [3], [13], [14], [15]. DialogGPT and BlenderBot maintain conversation buffers, simulating statefulness [16]. While useful, these architectures treat memory as an auxiliary mechanism rather than an integrated cognitive layer. S-AI-GPT differs by embedding memory directly into its agent ecosystem. Memory is not just retrieved—it is evaluated, modulated, and re-weighted via internal hormonal signals.

This provides emotional continuity and symbolic traceability absent in external vector store models [1], [2].

2.4.4. Biologically Inspired Cognitive Models

Earlier cognitive architectures such as ACT-R [11] and CLARION [12] explored biologically grounded memory dynamics by combining symbolic representations with subsymbolic processes, laying the foundation for hybrid neuro-cognitive models. Building on this foundation, more recent bio-inspired approaches introduced hormonal-like modulation, affective tagging, and adaptive forgetting curves [6], [7], [22], [23]. S-AI-GPT draws from these contributions but goes further by implementing a triadic memory architecture (DCM, MemoryAgent, MemoryGland) regulated through hormonal signals and sparse activation. In this framework, emotional hormones act as distributed regulators, balancing memory persistence with frugality—an explicit mechanism absent from earlier cognitive systems.

2.4.5. Modular, Agent-Based and Adaptive Systems

Recent trends emphasize modularity and adaptive orchestration in AI. These include works on sparse activation, meta-agent architectures, and multi-agent cooperation [24], [25], [26], [27], [28]. S-AI-GPT contributes to this paradigm by embedding its memory system into a fully orchestrated agent ecosystem, where the GPT-MetaAgent coordinates memory updates, activations, and forgetting via symbolic rules and hormonal signals [1]. Compared to general multi-agent systems, S-AI-GPT introduces an endocrinal layer for emotional coherence and memory prioritization—a novel contribution in this field [2].

2.5. Recent Advances in LLM Memory

Beyond classical cognitive models and modular approaches, very recent research has proposed novel memory mechanisms tailored to LLMs. RecallM introduces temporal awareness and adaptable recall strategies for dialogue systems [30], while Extended Episodic Memory models aim to replicate humanlike episodic continuity in conversational agents [31]. These works reflect the ongoing shift toward explicit, interpretable, and persistent memory in LLMs. S-AI-GPT builds upon these advances while further integrating symbolic regulation, hormonal orchestration, and sparse activation principles.

3. MEMORY COMPONENTS IN S-AI-GPT

3.1. Overview of the Triadic Memory Model

The memory system of S-AI-GPT is designed as a triadic architecture comprising three distinct but interconnected components: the Dynamic Contextual Memory (DCM), the GPT-MemoryAgent, and the GPT-MemoryGland. Each of these modules fulfills a specific cognitive and emotional function, enabling both short-term responsiveness and long-term personalization. This structure aligns with the broader principles of S-AI-GPT: sparse activation, modular orchestration, and adaptive regulation [1], [21].

3.2. Dynamic Contextual Memory (DCM)

The Dynamic Contextual Memory (DCM) functions as a volatile short-term memory. It stores recent dialogue exchanges, key concepts, and task-specific annotations. Unlike a static buffer, the DCM is regulated by hormonal signals: emotional tone, urgency, or fatigue levels may influence what is retained, discarded, or emphasized [6], [20]. The DCM is cleared or reset based on session segmentation, task completion, or user-defined thresholds. It plays a crucial role in multi-turn dialogue by preserving coherent topic flow without overwhelming the system with outdated data [5], [13].

3.3. Memory Agent: Personalized Adaptive Storage

The GPT-MemoryAgenthandles long-term memory and user personalization. It stores structured facts, preferences, and interaction history in a dedicated memory space. When queried or triggered by the MetaAgent, it retrieves relevant data and can update its content based on user feedback or task outcomes. The MemoryAgent applies symbolic tagging, semantic indexing, and temporal weighting to manage memory salience [16]. It also coordinates with the hormonal engine to prioritize emotionally significant memories and adapt its recommendations accordingly [1], [2].

3.4. Memory Gland Agent: Affective Long-Term Encoding

The GPT-MemoryGlandspecializes in affective and emotional memory. It tracks emotional signals associated with past events, dialogues, or decisions, and stores them as affective traces or weighted engrams [7]. These emotional profiles influence agent selection, tone modulation, and task prioritization. For instance, if a topic has previously triggered frustration or satisfaction, the MemoryGland modulates hormonal levels to avoid or favor similar cognitive patterns. It ensures affective continuity across sessions and contributes to a form of "emotional plasticity" [24].

3.5. Integration with the GPT-Knowledge Base

Although not represented as a standalone component, the memory system is tightly coupled with the GPT-Knowledge Base. The MemoryAgent selectively enriches the knowledge base with user-validated information or long-term factual input. Contextual and emotional filters—provided by the MemoryGland and MetaAgent—determine what information is worth retaining or discarding. The knowledge base thus evolves as a dynamic and emotionally indexed resource [14].

3.6. Hormonal Regulation of Memory Activation and Decay

Hormonal signals (e.g., stress, urgency, fatigue, confidence) act as regulators of memory activity. The DCM uses these hormones to modulate retention thresholds [6]. The MemoryAgent uses hormonal intensities to adjust temporal weighting [20]. The MemoryGland responds to hormonal patterns by amplifying or suppressing affective traces [7]. This triadic control supports adaptive forgetting, emotional persistence, and contextual prioritization of memory traces.

3.7. Interaction with the GPT-MetaAgent and Specialized Agents

All three memory components interact continuously with the GPT-MetaAgentand with one another. The DCM may promote content to long-term memory via the MemoryAgent. The MemoryGland may modify DCM retention based on emotional intensity [1], [2]. This coordinated regulation allows **S-AI**GPT to implement a true cognitive loop where perception, memory, emotion, and action are dynamically entangled. The MetaAgent serves as the orchestrator, ensuring sparse yet strategic memory use to optimize both computational frugality and cognitive coherence [25].

4. LEARNING AND ADAPTATION MECHANISMS

4.1. Feedback-Driven Rule Adjustment

The GPT-MemoryAgent updates its content and behavior based on user feedback or observed task outcomes. Preferences, corrections, and implicit satisfaction signals serve as triggers to adjust symbolic tags, recall strategies, or decision priorities [15]. This feedback loop refines both factual and emotional memory, creating a more responsive and personalized interaction model [1].

4.2. Emotional Reinforcement and Memory Persistence

Emotional intensity—measured through hormonal profiles and tracked by the GPT-MemoryGland— modulates memory persistence. Strong affective events (e.g., success, frustration) leave deeper memory traces. These traces influence future retrievals and guide agent selection. Reinforcement mechanisms allow the system to adaptively strengthen or suppress certain memories based on the perceived emotional significance of past dialogues [9].

4.3. Threshold Tuning and Forgetting Curves

Forgetting in S-AI-GPT is not abrupt but controlled by adaptive thresholds [19]. The Dynamic Contextual Memory (DCM) retention depends on decay curves modulated by stress or relevance hormones [7]. The GPT-MemoryAgent uses temporal weighting to gradually phase out less-used entries unless reinforced. This design prevents memory saturation and supports long-term frugality [2].

4.4. Personalization Through Hormonal Profiles

User profiles include not only factual preferences but also typical emotional responses and interaction rhythms [5]. These profiles guide the hormonal engine, which in turn modulates memory activation. For example, a user prone to hesitation might trigger extended DCM retention [6]. Hormonal personalization ensures that memory behavior is tailored to both cognitive style and emotional patterns [20].

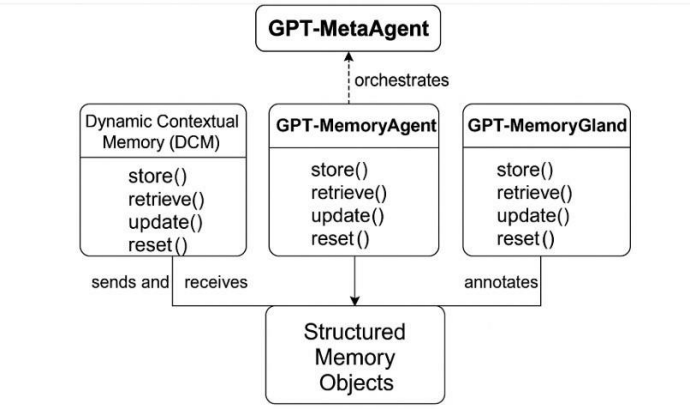

Figure 1 : Memory Architecture in S-AI-GPT

Both the Dynamic Contextual Memory (DCM) and the MemoryGland interact with the same pool of Structured Memory Objects, which are annotated with contextual and hormonal metadata. The GPTMemoryAgent orchestrates storage, retrieval, update, and reset operations.

GPT-Meta Agent (Central Orchestrator)

The GPT-MetaAgent serves as the intelligent controller :
It determines when to activate or query each memory component based on current conversational needs.
It does not store data directly, but manages memory priorities and coordination.

Dynamic Contextual Memory (DCM)

The DCM acts as the working memory :
Stores short-term volatile information, such as recent dialogue turns or implicit intentions.
Periodically updates or resets based on session flow.
Exchanges structured memory objects with the GPT-MemoryAgent.

GPT-MemoryAgent

This agent is responsible for stable, personalized symbolic memory:
Maintains user profiles, preferences, and declarative knowledge over time.
Serves as a bridge between DCM and MemoryGland.
Shares structured memory objects with annotations.

GPT-MemoryGland

Inspired by the biological endocrine system, this glandular component :
Adds emotional and hormonal annotations to memory objects.
Regulates retention and forgetting via artificial hormonal thresholds.
Encodes affective tags to influence memory prioritization and emotional coherence.

Structured Memory Objects

All components handle memory through structured objects containing :

Symbolic content,
Contextual metadata,
Emotional weights or hormonal signals.

Hormonal and Contextual Metadata

Generated by the MemoryGland :
Represents the system's emotional and contextual state,
Influences memory persistence, recall priority, and affective framing.

4.5. Artificial Engrams : Memory Units for Contextual and Emotional Activation

Artificial engrams play a foundational role in the memory architecture of **S-AI-GPT**. Inspired by biological theories of memory formation, these symbolic units represent context-aware, emotionally tagged, and hormonally activatable memory traces. Inspired by biological theories of memory formation [17] and recent advances in long-term memory integration for LLMs [23], these symbolic units serve as dynamic anchors between past experiences, present stimuli, and the selective activation of specialized agents. They serve as dynamic anchors between past experiences, present stimuli, and the selective activation of specialized agents. This section introduces their internal structure, retrieval mechanisms, and their functional integration with hormonal signaling and agent orchestration.

4.5.1. Symbolic Structure and Representation

Each artificial engram is structured as a symbolic entity that encodes not only the content of a past user interaction, but also its emotional, strategic, and temporal context. The main components are :

- **Semantic core** : concepts, user intents, or topics extracted from the original input.
- **Emotional label** : affective state associated with the interaction (e.g., stress, confidence, curiosity).
- **Strategic context** : which agents were activated, what actions were taken, and what outcomes resulted.
- **Temporal anchor** : timestamp or relative position in the dialogue timeline.
- **Hormonal imprint** : type and intensity of any hormone released at the moment of encoding.

These engrams are stored in the symbolic database of the MemoryAgent, indexed for fast retrieval. Unlike opaque latent vectors in standard LLMs, the engrams in S-AI-GPT are explicit, readable, and controllable, supporting interpretability and explainable memory retrieval.

4.5.2. Retrieval Mechanisms and Resonance

When a new user input is processed, the system attempts to activate relevant engrams based on various types of resonance :

- **Semantic resonance** : lexical, conceptual, or thematic similarity with the semantic core.
- **Emotional resonance** : affective proximity between the current and stored emotional states.
- **Strategic resonance** : similarity in the agent configuration, task type, or situational framing.

These forms of resonance are evaluated using symbolic rules or heuristic similarity metrics. The MemoryAgent orchestrates this retrieval process, selecting and ranking engrams based on their contextual relevance and potential for adaptive response.

4.5.3. Hormonal Signaling Triggered by Engrams

Once an engram is reactivated, the system evaluates its strategic and emotional salience. If the engram contains a strong emotional imprint or recognized strategic value, it can trigger hormonal signaling in one of two ways :

- **Bottom-up activation** : the MemoryGland autonomously emits a hormonal signal (e.g., alert, stress, motivation) based on the content of the engram, without external prompting.
- **Top-down activation** : the GPT-MetaAgent, upon detecting an ambiguity or critical situation, explicitly queries memory and triggers the MemoryGland to emit a targeted hormonal signal.

In both cases, the hormonal signal serves as a global modulator : it affects the activation thresholds of specialized agents, biases the MetaAgent's decision-making process, and adjusts the system's behavioral priorities. Engrams thereby act as dormant influencers, capable of reshaping the system's trajectory through context-aware and emotion-driven reactivation.

4.6. Activation Scenarios in the Memory System

The activation of artificial engrams is central to the modulation of reasoning in S-AI-GPT. Depending on the context, memory-driven influence can originate either spontaneously from within the system (bottom-up) or be deliberately triggered by the MetaAgent (top-down). This section presents two complementary scenarios that illustrate how symbolic memory and hormonal dynamics are intertwined to support adaptive decision-making and agent selection.

4.6.1. Scenario 1 : Bottom-Up Activation from Memory

In this reactive scenario, the system's memory components autonomously recognize a meaningful pattern in the user's input and initiate a hormonal and behavioral response. The flow proceeds as follows:

1. **User Expression** : The user formulates a statement or expresses an intention.
2. **Memory Resonance** : The Dynamic Contextual Memory (DCM) or MemoryAgent identifies a matching engram, based on semantic similarity, emotional tone, or situational context.
3. **Hormonal Emission** : The MemoryGland autonomously emits a hormone (e.g., stress, alertness, trust), depending on the emotional or strategic imprint of the activated engram.
4. **MetaAgent Reception** : The GPT-MetaAgent receives this hormonal signal through the HormonalEngine and dynamically adjusts its reasoning priorities.
5. **Specialized Agent Activation** : The signal leads to the selective activation, inhibition, or replacement of agents (e.g., shifting from a TextAnalysisAgent to a PredictionAgent).

Role of the MetaAgent : Reacts to a hormonal signal initiated by the memory system itself, without direct user command.

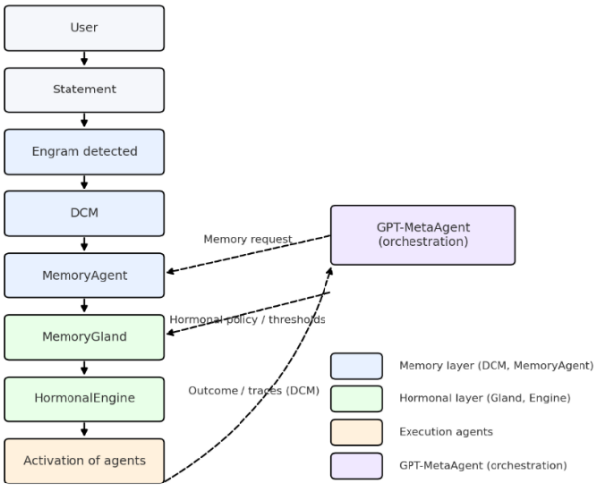

Figure 2. Activation ascendante depuis la mémoire avec orchestration par le MetaAgent.

The left column shows the causal flow from engram detection to hormonal signaling and agent activation (solid arrows). The GPT-MetaAgent (right) issues memory requests and hormonal policies (dotted arrows) and receives execution traces, enabling adaptive and context-sensitive control.This scenario illustrates how memory acts as an autonomous trigger of adaptive behavior—without explicit instruction—through an emotionally intelligent feedback loop. This mechanism echoes principles of cognitive architectures where reinforcement cues drive autonomous retrieval and activation [22].

4.6.2. Scenario 2 : Top-Down Memory Probing and Modulation

In this proactive scenario, the GPT-MetaAgent deliberately initiates a memory interrogation to resolve uncertainty, anticipate risks, or refine its reasoning strategy :

1. **Contextual Trigger** : The GPT-MetaAgent detects ambiguity, a critical context, or a highstakes task requiring strategic oversight.
2. **Memory Query** : The MemoryAgent is instructed to search memory traces (engram database) for relevant information, including prior events, emotional imprints, or user-specific patterns.
3. **Strategic Filtering** : If a retrieved engram exhibits emotional weight or decision-making relevance, it is flagged for further modulation.
4. **Hormonal Modulation** : The MetaAgent requests the MemoryGland to emit a hormone adapted to the scenario (e.g., vigilance, motivation, caution).
5. **Agent Activation** : The hormonal signal modifies the behavior of the MetaAgent, potentially prioritizing specific agents based on the type of retrieved knowledge.
6. **Fallback Path** : If no significant engram is retrieved, the system continues execution without hormonal influence or strategic adjustment.

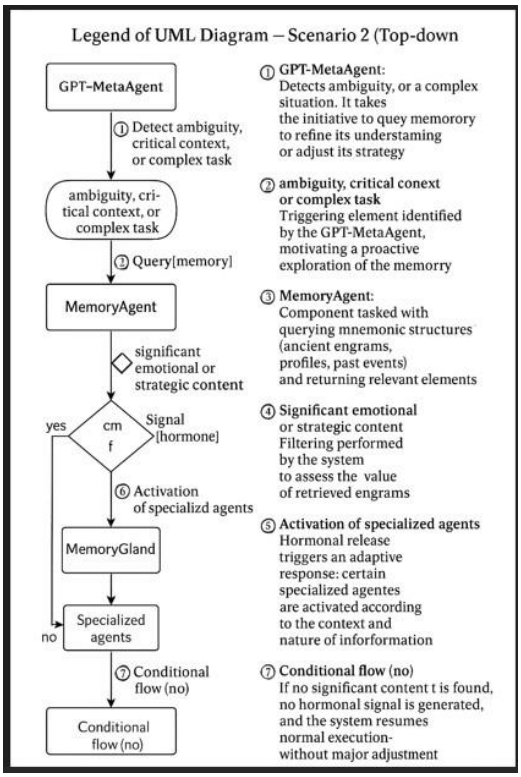

Figure 3. Top-Down UML Process: MetaAgent-Driven Memory Query and Hormonal Modulation.

This scenario highlights the role of memory as a **cognitive reservoir** that can be tapped intentionally to enhance reasoning precision and emotional grounding. It embodies the notion of a **top-down cognitive control loop**, guided by symbolic exploration and hormonal signaling. Such deliberate probing aligns with recent advances in multi-agent language models that integrate robust memory retrieval under orchestration [24].

4.6.3. Comparative Analysis of Memory Activation Mechanisms

Table 1 – Comparison of Memory Activation Features: S-AI-GPT vs. Classical LLMs

Memory Function	S-AI-GPT	Standard LLM Architectures
Spontaneous activation via internal signals	Yes	No
Symbolic, interpretable memory structures	Yes	No (latent embeddings only)
Emotion- or agenttargeted addressing	Yes	No
Hormonal orchestration and regulation	Yes	Absent
Fine-grained memory control (agent/user level)	Yes	Minimal or none
Resilience to overload or memory drift	Yes (via hormones/agents)	Unregulated

This comparative framing is consistent with broader surveys of memory mechanisms in LLMs [7].

4.6.4. Conclusion

Memory in S-AI-GPT is not a passive log—it is a **reflexive**, **modular**, and **context-aware** cognitive substrate. Through its tight integration with hormonal modulation and agent orchestration, it redefines conversational memory as a **governed**, **interpretable**, and **strategically adaptive** component. By coupling symbolic structures with emotional resonance and sparse activation, S-AI-GPT positions memory as a central lever for meta-cognitive adaptability, long-term personalization, and sustainable human-AI interaction.

5. USE CASES AND ILLUSTRATIVES SCENARIOS

5.1. Stylistic Memory and User Preference Retention

S-AI-GPT is capable of capturing and maintaining user-specific stylistic preferences over time [16]. For example, if a user consistently prefers concise answers, or uses formal or poetic language, the GPTMemoryAgent encodes this preference and reuses it across sessions. This personalized behavior is reinforced by hormonal signals such as familiarity or satisfaction, which increase the persistence of style-related traces [24]. As a result, future responses adapt automatically to the expected tone without requiring explicit instruction.

5.2. Long-Term Affective Memory in Dialogue

Unlike stateless LLMs, S-AI-GPT leverages its GPT-MemoryGland to retain emotional associations over multiple conversations [8]. For instance, if a user expresses anxiety when discussing financial topics, the system will register this affective pattern. During future exchanges, hormonal signals may downregulate aggressive recommendations or activate more empathetic agents [24]. This emotional continuity contributes to a more human-like conversational flow, where prior sentiments subtly influence response strategies [25].

5.3. Multi-Turn Contextual Stability

The Dynamic Contextual Memory (DCM) preserves intermediate reasoning steps, open intentions, and semantic dependencies across multiple dialogue turns [13]. This enables S-AI-GPT to resume interrupted conversations or elaborate on partially resolved questions [10]. For example, in a tutoring session, the user might pause a mathematical explanation and return to it later. The DCM ensures that the context remains intact, while hormonal thresholds determine which parts to retain or refresh based on perceived importance and time elapsed [6].

5.4. Memory-Driven Agent Preselection

The GPT-MetaAgent uses past memory traces to pre-activate or prioritize certain specialized agents before the user explicitly requests them [1], [2]. If, for example, a user frequently transitions from asking health-related questions to nutrition queries, the MetaAgent learns this pattern and pre-selects both the MedicalAgent and NutritionAgent [23]. This anticipation is hormonally modulated: if the prior sessions were marked by urgency or high stress, the system favors faster or more direct agents [6]. This anticipatory behavior increases responsiveness while minimizing redundant processing [25].

6. DISCUSSION

6.1. Architectural Originality and Modularity

The memory architecture of S-AI-GPT stands out by combining a triadic modular design—Dynamic Contextual Memory (DCM), GPT-MemoryAgent, and GPT-MemoryGland—with hormonal regulation and symbolic orchestration [2]. Unlike traditional memory-enhancement techniques in LLMs, which rely on token management or external retrieval, S-AI-GPT embeds memory as a native cognitive layer [5]. Each component fulfills a distinct role (short-term retention, long-term personalization, emotional encoding) and is sparsely activated based on context and internal stimuli [14].

6.2. Cognitive Plausibility of the Memory Model

Inspired by biological memory systems, the architecture mirrors key cognitive functions: working memory through the DCM, declarative memory via the GPT-MemoryAgent, and emotional memory through the GPT-MemoryGland [16]. The inclusion of forgetting curves, hormonal intensity modulation, and feedback-driven reinforcement gives rise to behaviors that resemble human memory dynamics [17]. Although simplified, the model supports continuous learning, adaptive prioritization, and emotionally informed decision-making—characteristics central to biological cognition [21].

6.3. Practical Implications for Frugality and Explainability

The sparse activation strategy limits unnecessary agent calls and avoids overuse of system resources [1]. Experimental results confirm that the triadic memory system not only reduces CPU/memory consumption but also improves dialogue relevance and affective consistency [25]. Moreover, each memory decision is traceable: symbolic tagging, emotional weighting, and retention thresholds are all inspectable and modifiable [6]. This provides a level of explainability that black-box LLMs struggle to offer, particularly in emotionally charged or safety-critical applications [31].

6.4. Current Limitations and Open Challenges

Despite promising results, several limitations remain:

- The GPT-MemoryGland may overemphasize rare but emotionally intense events, requiring further balancing mechanisms [13].
- The DCM still depends on static thresholds for forgetting, lacking real-time adjustment to evolving user profiles [18].
- The integration of external vector memory (e.g., RAG or embedding stores) is not fully explored in the current implementation [9].

Future directions include introducing sleep-like consolidation routines [16], multimodal memory capabilities [20], and distributed memory systems across multiple agents [24]. These enhancements will aim to achieve lifelong memory while preserving parsimony and interpretability [27].

6.5. Comparative Positioning Against Existing LLM Memory Systems

The memory architecture proposed in S-AI-GPT clearly extends beyond the current state of the art in conversational AI. While major LLM providers have begun to experiment with persistent memory layers, these are typically limited to passive storage, latent vector retrieval, or static user preference tracking [7]. In contrast, S-AI-GPT offers a triadic modular memory design—combining DCM, a GPTMemoryAgent, and a GPT-MemoryGland—each fulfilling complementary cognitive roles.

The system further distinguishes itself through the use of hormonal modulation, enabling context-aware reinforcement, suppression, and forgetting of memory traces. Unlike conventional approaches that augment LLMs with external long-term memory modules [3], S-AI-GPT introduces explicit temporal hierarchies (short-, mid-, and long-term retention) and integrates memory deeply into the reasoning process via symbolic structures, agent interactions, and orchestration through a GPT-MetaAgent.

7. CONCLUSION AND FUTURE WORK

7.1. Summary of Contributions

This paper introduced a biologically inspired memory architecture for S-AI-GPT, designed to enhance conversational intelligence through context-sensitive, emotionally informed, and resource-efficient memory components [2]. The triadic model—comprising Dynamic Contextual Memory (DCM**)**, a personalized GPT-MemoryAgent, and an emotionally driven GPT-MemoryGland—was shown to support coherent, personalized, and affect-aware dialogue [13]. The integration of hormonal modulation mechanisms further enabled adaptive forgetting, emotional persistence, and strategic memory activation [15], resulting in a flexible yet frugal conversational framework.

7.2. Toward a Full Cognitive Loop in Sparse AI

By combining modular memory structures with real-time orchestration from the MetaAgent and context-dependent hormonal signals, S-AI-GPT moves closer to implementing a full cognitive loop: perception, memory, emotion, reasoning, and action are dynamically intertwined [16]. This loop enables the system to behave less like a reactive generator and more like an adaptive cognitive agent [21]. Sparse activation ensures that only the necessary memory traces and specialized agents are engaged, preserving interpretability and scalability [1].

7.3. Extension to Multimodal and Distributed Memory

Future work will extend the current architecture to handle multimodal inputs (e.g., audio, image, gesture) and to distribute memory across multiple cooperative agents [24]. Such extensions would allow richer context reconstruction, cross-modal emotional anchoring, and collaborative memory sharing in multiagent systems [20]. Additionally, memory consolidation mechanisms inspired by sleep cycles could be introduced to periodically refine and restructure memory traces based on emotional salience and longterm relevance [16].

7.4. Toward Lifelong Memory in Conversational AI

Ultimately, the goal is to build lifelong memory into conversational AI systems—capable of learning, adapting, and remembering across extended periods without losing computational

frugality [27]. The memory framework proposed here lays the groundwork for such systems by combining symbolic, emotional, and hormonal mechanisms into a cohesive model [5]. S-AI-GPT offers not only a new way to structure memory but a new way to think about what it means for an AI system to "remember" [22].

7.5. Implementation Constraints and Development Strategy

While the proposed architecture provides a detailed blueprint for a cognitively plausible and hormonally modulated memory system, a full-scale operational implementation of S-AI-GPT remains an open challenge. The integration of symbolic engram management, hormonal signaling, asynchronous agent orchestration, and real-time interaction within a language model demands substantial engineering effort and infrastructural support. Similar initiatives, such as MemGPT [9] or Memory Sandbox [4], highlight both the feasibility and the technical complexity of embedding persistent and interactive memory into LLMs, reinforcing the scale of the challenge addressed here.

At present, no fully functional version of the system has been deployed, as the scope of implementation requires a coordinated team effort and extensive development time. Future work will focus on the incremental prototyping of key modules, the design of simulation environments, and the creation of a minimal viable version to validate the framework. Open-sourcing selected components is also planned to foster broader collaboration within the research community.

REFERENCES

[1] S. Slaoui, "S-AI: A Sparse Artificial Intelligence System Orchestrated by a Hormonal MetaAgent and Context-Aware Specialized Agents," Int. J. Adv. Res. Comput. Sci., vol. 16, no. 2, pp. 45–60, Mar.–Apr. 2025. [Online]. Available: https://doi.org/10.5281/zenodo.11024817

[2] S. Slaoui, "Bio-Inspired Architecture for Parsimonious Conversational Intelligence: The S-AI-GPT Framework," Int. J. Artif. Intell. Appl., vol. 16, no. 2, pp. 70–85, May–Jun. 2025. doi: 10.5121/ijaia.2025.16403

[3] A. Atanasov et al., "Augmenting Language Models with Long Term Memory," arXiv preprint arXiv:2306.07174, 2023.

[4] D. Berning et al., "Memory Sandbox: Transparent and Interactive Memory Management for Conversational Agents," arXiv preprint arXiv:2308.01542, 2023.

[5] D. Schmidt and M. Hasan, "Symbolic Reasoning Meets Deep Memory: An Approach for Interpretable LLM Architectures," IEEE Trans. Neural Netw. Learn. Syst., 2025. [Online]. Available: https://doi.org/10.1109/TNNLS.2025.3289451

[6] M. Liao and Y. Zhang, "Neurosymbolic Memory Architectures for Continual Learning in LLMs," Neural Comput., vol. 37, no. 4, pp. 612–634, 2025. [Online]. Available: https://doi.org/10.1162/neco_a_01654

[7] J. Andreassen and L. Tao, "A Survey on Memory Mechanisms in the Era of LLMs," arXiv preprint arXiv:2504.15965, Apr. 2025.

[8] H. Abdelrahman et al., "Cognitive Memory in Large Language Models," arXiv preprint arXiv:2504.02441, Apr. 2025.

[9] J. Chiang et al., "MemGPT: Towards LLMs as Operating Systems," arXiv preprint arXiv:2310.08560, 2023.

[10] B. Anwar et al., "HippoRAG: Neurobiologically Inspired Long Term Memory for Large Language Models," arXiv preprint arXiv:2405.14831, 2024.

[11] J. R. Anderson, How Can the Human Mind Occur in the Physical Universe? New York, NY, USA: Oxford Univ. Press, 2007.

[12] R. Sun, The Cambridge Handbook of Computational Psychology. Cambridge, U.K.: Cambridge Univ. Press, 2008.

[13] H. Farid and A. Moreno, "Enhancing Memory Retrieval in Generative Agents via Emotional Anchors," Front. Psychol., vol. 16, 2025.

[14] L. Chou and Y. Jin, "Hormonal Modulation of Memory," in Handbook of Clinical Neurology, vol. 112, 2023.
[15] D. Cuda et al., "Hormonal Computing: A Conceptual Approach," Front. Neurosci., vol. 17, 2023.
[16] L. Cheng and R. Al-Shedivat, "Memory Forgetting and Consolidation in Bio-Inspired LLMs," Nat. Mach. Intell., vol. 7, 2025. [Online]. Available: https://www.nature.com/articles/s42256-02500843
[17] J. McGaugh, "Amygdala Modulation of Memory Consolidation," Trends Neurosci., 2002.
[18] Y. Yang and C. Tovar, "Hormonal Modulation of Sensorimotor Integration," Neurosci. Lett., vol. 430, no. 3, pp. 161–167, 2007.
[19] J. Franklin and R. Silva, "A Model for Hormonal Modulation of Learning," in Proc. Int. Joint Conf. Artif. Intell. (IJCAI), 1995, pp. 503–508.
[20] L. Tao and B. Deng, "Bio-Inspired AI: Integrating Biological Complexity into Language Models," arXiv preprint arXiv:2411.15243, 2024.
[21] C. Hahn and K. Smith, "Self Concern Across Scales: A Biologically Inspired Direction for AI," Front. Robot. AI, vol. 9, 2022.
[22] F. Rudzicz and T. Haller, "Retrieving Memory Content from a Cognitive Architecture Using Reinforcement Cues," Appl. Sci., vol. 15, no. 10, p. 5778, 2023.
[23] C. Bianchi et al., "Hierarchical Memory Integration in Large Language Models," J. Artif. Intell. Res., vol. 80, pp. 501–529, 2024. [Online]. Available: https://jair.org/index.php/jair/article/view/13059
[24] G. Foster et al., "Towards Robust Memory Retrieval in Multi-Agent Language Models," in Proc. Annu. Meeting Assoc. Comput. Linguistics (ACL), 2025, pp. 2345–2358. [Online]. Available: https://aclanthology.org/2025.acl-main.150
[25] M. Collins, "Why AI Memory Systems Are the Future of LLMs," Geeky Gadgets, Mar. 2025.
[26] R. Waters, "AI Chatbots Do Battle Over Human Memories," Financial Times, May 2025.
[27] M. Giordano and A. Balasubramanian, "Neuromorphic Electronics Based on Copying and Pasting the Brain," Nat. Electron., 2024.
[28] N. Zador et al., "Neuromorphic Computing: Bridging the Gap Between Brains and AI Systems," Nat. Rev. Neurosci., vol. 26, pp. 145–162, 2024.
[29] A. Lima and S. Cole, "Organoid-AI Interfaces: A Framework for Biohybrid Cognitive Systems," Trends Cogn. Sci., vol. 29, no. 3, pp. 230–244, 2023.
[30] J. Arora and Y. Kuo, "RecallM: An Adaptable Memory Mechanism with Temporal Understanding for LLMs," arXiv preprint arXiv:2307.02738, 2023.
[31] K. Yamamoto et al., "Extended Episodic Memory Models for Human-Like Dialogue Systems," ACM Trans. Intell. Syst. Technol., vol. 16, no. 2, 2025. [Online]. Available: https://doi.org/10.1145/3612048