BIO-INSPIRED HORMONAL MODULATION AND ADAPTIVE ORCHESTRATION IN S-AI-GPT

Said Slaoui

University Mohammed V, Rabat, Morocco

ABSTRACT

This second article delves into the bio-inspired regulatory mechanisms and memory architectures integrated into the S-AI-GPT system. Building upon the modular design introduced in Article 1, we explore how artificial hormonal signaling enables dynamic orchestration of agents and emotional coherence. The system relies on a triadic hormonal regulation layer composed of the Hormonal Engine, the GPT-Gland Agents, and the GPT-MemoryGland, working in concert to adjust activation thresholds, modulate emotional tone, and support context-aware responsiveness. A dedicated section addresses the integration of memory components, including Dynamic Contextual Memory (DCM), the Memory Agent for personalization, and a bio-inspired memory system based on neuronal mini-structures and artificial engrams. These memory structures interact strategically with hormonal dynamics to maintain both adaptability and persistence. We also examine the autonomy, lifecycle, and coordination of GPT-Specialized Agents across conversational and business contexts. The orchestrator, GPT-MetaAgent, supervises the entire system by integrating semantic cues, user profiles, and hormonal feedback. This approach paves the way for a resource-efficient, interpretable, and cognitively enriched conversational AI.

KEYWORDS

Sparse Artificial Intelligence, GPT-MetaAgent, GPT-Specialized Agents, GPT-Gland Agents, Hormonal Engine.

1. INTRODUCTION

The S-AI-GPT system, introduced in the first article as a modular, sparse, and interpretable alternative to traditional GPT architectures, integrates a biologically inspired regulation layer that enables fine-grained contextual adaptation, emotional modulation, and orchestrated, energy-efficient activation of GPT-Specialized Agents. This second article focuses on the artificial hormonal layer structured around three core components : a central Hormonal Engine responsible for hormone generation, diffusion, and degradation; GPT-Gland Agents emitting targeted signals in response to semantic and emotional stimuli; and the GPT-MemoryGland, which anchors long-term affective coherence within the memory subsystem. Beyond hormonal signaling, this article details the integration of memory components—including the Memory Agent, Dynamic Contextual Memory (DCM), and a bio-inspired memory architecture based on neuronal mini-structures and artificial engrams. These memory subsystems work closely with the hormonal system to support persistence and adaptive recall, reinforcing S-AI-GPT's capacity for truly personalized and context-aware conversational AI. We also examine the autonomy, lifecycle, and coordination of GPT-Specialized Agents, demonstrating how their activation is modulated by both memory states and hormonal feedback. This hormonal and cognitive regulation layer constitutes a cornerstone of adaptive intelligence within the S-AI-GPT framework.

1.1. Memory Agent : Contextual Persistence and Personalization

The Memory Agent is a key component of the S-AI-GPT ecosystem, dedicated to the persistent management of contextual information, user preferences, and interaction traces. It allows the system to go beyond transactional memory and build long-term cognitive continuity, essential for truly personalized and evolving dialogue. The Memory Agent manages two main types of storage: Persistent Memory, which stores global user preferences such as tone, style, preferred structure, and thematic affinities, enabling continuous personalization across multiple sessions; and Contextual Memory, which retains session-specific information including emotional tone, conversation subject, or desired style, and supports fine-grained real-time adaptation without interfering with global memory. This dual-layer approach ensures both stability and adaptive flexibility throughout interactions. The GPT-MetaAgent serves as the primary supervisor of the Memory Agent. It decides when to query persistent or contextual memory, adjusts personalization strategies based on user evolution, and uses memorized information to modulate hormonal profiles, such as triggering an empathy gland if user sadness is detected. Through this centralized orchestration, dialogue becomes progressively more personalized without increasing computational complexity. The Memory Agent is deployed across multiple scenarios. In educational coaching, it helps retain learner progress and adjust exercises according to style and proficiency. In emotional monitoring, it can detect recurring emotional patterns such as stress in order to fine-tune future responses. In professional assistance, it adapts terminology and detail levels based on user preferences in specialized domains such as medical, financial, or legal contexts. The integration of the Memory Agent strengthens cognitive continuity across dialogue sessions, reduces the user's cognitive load by minimizing the need to re-specify preferences, optimizes computational resources by avoiding redundant recalculations, and enhances user experience by delivering proactive, natural, and evolving personalization. By combining modularity, adaptive persistence, and hormonal orchestration, the Memory Agent helps transform S-AI-GPT into a truly living, reflexive conversational companion.

1.2. GlandMemoryAgent : Hormonal Context Encoding

Beyond the symbolic and declarative role of the MemoryAgent, the GlandMemoryAgent (GlandeMemoire) provides a complementary form of memory based on hormonal traces and contextual emotional feedback. Its primary function is to encode, retain, and reuse affective and hormonal states associated with previous interactions, in order to influence future hormonal signaling and agent behavior. This agent does not store explicit knowledge or facts, but rather maintains a dynamic record of the hormonal context, including the emotional tone of past conversations, the type and intensity of hormones diffused (e.g., vigilance, empathy, caution), and the perceived success or failure of previous modulations in achieving coherent and user-aligned responses. Its internal model allows it to simulate affective memory homeostasis, helping the system regulate its future responses in accordance with both the user's emotional profile and recent contextual stimuli. Concretely, the GlandMemoryAgent supports the GPT-MetaAgent and the GPT-Gland Agents by suggesting preactivation of certain hormonal profiles based on historical convergence, preventing hormonal overreaction (e.g., repeated stress signals in neutral contexts), and encouraging consistency in emotional tone across long-term interactions. This biologically inspired memory layer enables temporal coherence, emotional continuity, and adaptive hormonal learning—reinforcing the cognitive plausibility and ethical responsiveness of the S-AI-GPT framework. The result is a more human-like, emotionally sensitive, and resilient conversational AI.

1.3. Dynamic Contextual Memory (DCM) : Hormone-Dependent Working Memory (WM)

The Dynamic Contextual Memory (DCM) is a volatile and adaptive memory structure that retains the contextual elements relevant to an ongoing session. It plays a central role in S-AI-GPT by serving as a transient Working Memory (WM) regulated by hormonal signals. The DCM is not an agent—it has no autonomy or activation cycle—but instead functions as a sensitive and reactive buffer.Unlike traditional approaches based on dense monolithic models or static memory buffers, the S-AI-GPT framework introduces a new paradigm rooted in **Sparse Artificial Intelligence (S-AI)**. This architecture leverages **Explainable Artificial Intelligence (XAI)** mechanisms to ensure interpretability, and adopts a **Mixture of Experts (MoE)** strategy for modular and sparse agent activation. The system also draws inspiration from **Reinforcement Learning from Human Feedback (RLHF)** to adapt its behaviors, while maintaining full compatibility with classical **Large Language Models (LLMs)** such as **GPT** and **BERT** variants. These enhancements ensure that S-AI-GPT is not only scalable and efficient, but also cognitively plausible and human-aligned.

1.3.1. Nature and Role

The DCM is a passive, non-agentive structure, without internal decision-making. It stores short-term contextual information such as topic, emotional tone, or conversational cues. It is modulated dynamically by the Memory Gland Agent, according to stress, urgency, or attention signals. It is supervised and arbitrated by the Memory Agent, which filters, reinforces, or inhibits content when needed. It acts as an interface layer between perception, reasoning, and response generation.

1.3.2. Conceptual Clarification

A common confusion arises between the DCM and the Memory Gland Agent, since both are involved in the memory regulation loop. However, they have distinct roles. The DCM is a memory structure that stores, updates, and filters session-specific context, whereas the Memory Gland Agent is an active agent that receives hormonal signals and modulates DCM contents (amplifying or suppressing them). The important distinction is that the DCM is a memory layer influenced by hormones, while the Memory Gland is an active agent that modifies the DCM in response to internal or external stimuli. The DCM react s; the gland acts.

1.3.3. Biological Analogy

The DCM can be likened to Working Memory (WM) or the hippocampus in biological systems, holding context “here and now”. The Memory Gland Agent is similar to an endocrine gland such as cortisol or adrenaline, which modulates memory based on emotional or internal state.

1.3.4. Strategic Value

The DCM is central to the cognitive ecology of S-AI-GPT because it allows real-time, emotion-sensitive memory filtering. It supports contextual adaptability without polluting Long-Term Memory (LTM). It enables hormonal feedback loops that adjust retention dynamically, and contributes to an ethical, explainable, and efficient system behavior in emotionally intense or medically critical contexts.

1.4. Towards a Bio-Inspired Modular Memory : Neuronal Mini-Structures and Artificial Engrams

1.4.1. Functional Summary of the Memory Architecture

This section highlights the power of integrating memory, hormonal modulation, and adaptive orchestration within the S-AI-GPT system. Each component—whether passive structure or active agent—has a specific but coordinated role, enabling intelligent, emotionally stable, and cognitively relevant dialogue. The tripartite memory architecture contributes to proactive personalization, cognitive continuity across sessions, reduced user mental load, and enhanced computational efficiency. It thus forms a strong technical and ethical foundation for the development of adaptive, explainable, and sustainable conversational systems.

1.4.2. Bio-Inspired Model: Distributed Memory and Artificial Engrams

In direct analogy with the functioning of the human brain, the S-AI-GPT system does not treat memory as a passive storage space, but rather as a set of activatable mini-neural structures, comparable to biological engrams. In neuroscience, an engram refers to a distributed micro-network of neurons whose joint activation encodes a specific memory, concept, or learned knowledge. Here, each mini-neural structure corresponds to a lightweight, autonomous model, encapsulating a specific skill or knowledge element that can be reactivated on demand. This vision inspires a sparse and distributed representation of memory in S-AI-GPT, where each fact or useful notion is not stored as a static data block, but embodied in a context-sensitive module, triggered when needed and modulated by hormonal signals. Such an architecture allows the system to mobilize only the relevant sub-networks to respond to a request, minimizing global activity—just like a biological brain that awakens only the “neuronal fireflies” required to formulate an answer. Finally, the entire memory subsystem of S-AI-GPT—including the Memory Agent, Memory Gland, and DCM—can be reinterpreted through this unified bio-inspired lens. Each component contributes to a distributed, dynamic, and embodied memory architecture, where specialized agents, adaptive memory modules, and contextual activation glands operate collectively as a living, explainable, and economically activated memory.

1.5. An Autonomous, Bio-Inspired Memory Engine in S-AI-GPT

In most architectures based on large language models (LLMs) (LLMs), memory is either absent or reduced to a static context window or an external database queried on demand. In contrast, S-AI-GPT introduces an internal, modular, bio-inspired, and cognitively active memory engine that plays a central role in the system’s contextual intelligence. This engine does not merely archive data—it encodes, selects, modulates, anticipates, and guides decisions made by the global orchestrator. It is an independent mechanism from the GPT generative core, grounded in the principles of parsimony, plasticity, and hormonal supervision. The S-AI-GPT memory engine is built upon three complementary layers. The Memory Agent supervises long- and short-term persistence and personalization. It retains structured dialogue elements, user preferences, and interaction traces. The Memory Gland encodes past affective states and dynamically modulates the Working Memory (WM) (DCM) via artificial hormonal signaling. It forms a bio-inspired affective memory system. The Dynamic Contextual Memory (DCM) acts as a real-time, volatile, emotion-sensitive memory buffer. Though passive, it is hormonally regulated and critical to contextual coherence. Through this tripartite structure, the memory engine operates as a standalone cognitive subsystem, governed by its own logic of update, selective forgetting, and consolidation. It is orchestrated by the GPT-MetaAgent, which receives signals (hormonal profiles, memory cues, modulation suggestions) to adjust its agent activation strategies accordingly. The memory engine is based on the notion of activatable mini-structures, inspired by

neuronal engrams. Each memory, concept, or contextual state is represented by a lightweight module, reactivated only when needed. This design enables computational frugality, higher explainability, and reduced redundancy. Thus, memory in S-AI-GPT is discreet—only relevant modules are activated at any moment. It is context-sensitive—decisions are driven by internal signals rather than hard-coded heuristics. It is also reflective—memory states are updated based on feedback from ongoing interactions.

This memory engine is central to the progressive personalization of responses, cross-session cognitive and emotional continuity, targeted activation of specialized agents based on the user's evolving profile, proactive hormonal modulation guided by past interactions, and ethical and cognitive alignment in emotionally sensitive contexts. This bio-inspired approach to memory represents a significant departure from conventional context-handling mechanisms used in most LLMs. While current systems typically rely on static context windows, flat conversation histories, or external retrieval modules, S-AI-GPT introduces a structurally embodied, active, and distributed memory engine. Rather than acting as a passive buffer or external cache, memory in S-AI-GPT functions as an internal cognitive engine, capable of dynamically modulating relevance, retention, and selective forgetting. This paradigm shift enables deep personalization, affective continuity, and genuinely adaptive orchestration—features absent from classical approaches. It thus opens a new research direction centered on integrated intelligent memory systems, combining sparse activation, emotional modulation, and layered active recall, in direct analogy with human cognitive mechanisms.Par Ajouté (Recom 1 et 2)

1.6. Novelty of the Addressed Problems and Interdisciplinary Contributions

While several studies have explored memory enhancement or modular coordination in LLMs, most of them remain limited to prompt engineering, external retrieval mechanisms, or black-box memory vectors. This article tackles a set of novel and underexplored challenges in the field of conversational AI, including : (1) how to implement biologically inspired hormonal signaling in reasoning processes, (2) how to structure symbolic and interpretable memory systems that can influence internal agent activation, and (3) how to integrate adaptive orchestration mechanisms based on emotional and strategic memory traces. These problems are not only original but also foundational to building future generations of cognitively inspired and frugally orchestrated AI systems.This work combines innovative mechanisms from multiple research domains: (1) from biology, it draws inspiration from endocrine signaling and emotional regulation; (2) from symbolic AI, it leverages structured memory encoding and rule-based orchestration; (3) from modular cognitive architectures, it introduces agent-based sparsity and meta-control; and (4) from affective computing, it incorporates emotional traces into system prioritization. This interdisciplinary synthesis gives rise to a novel orchestration paradigm that goes beyond the limitations of monolithic deep learning systems.

2. GPT-SPECIALIZED AGENTS AND DISTRIBUTED EXPERTISE IN GPT

2.1. Distinction Between System Agents and Conversational Business Agents

The architecture of S-AI-GPT is built upon a fundamental distinction between two major categories of agents. On one hand, system agents handle supervision, hormonal modulation, protection, adaptive memory, and output filtering functions. These include the GPT-MetaAgent, GPT-Gland Agents, Security Agent, Memory Agent, and Access Result Agent. On the other hand, conversational business agents are specialized in performing specific dialogic tasks such as medical analysis, empathetic expression, analogy generation, or stylistic structuring (Moutoussis and Dolan [1]).

This section is entirely dedicated to the study of these conversational business agents. It details their shared structure, representative use cases, adaptation strategies, and their interactions within the modular ecosystem of S-AI-GPT (Picard [2]).

2.2. Functional Role and Unified Structure of GPT-Specialized Agents (Qu et al. [3])

GPT-Specialized Agents (SAs) are the functional pillars of the S-AI-GPT system. Each SA is responsible for solving a specific type of subproblem — such as emotional tone, formatting, domain expertise, or logical reasoning — as assigned by the GPT-MetaAgent based on a combination of semantic decomposition and contextual signals. Agents are grouped into functional families, including cognitive agents for reasoning and explanation, stylistic agents for formatting and tone, emotional agents for empathy and humor, and structural agents for memory access and aggregation (Richards [4]). While each specialized agent in S-AI-GPT is equipped with a dedicated mini-neural structure, this design does not contradict the core principle of sparsity expressed by the 20/80 heuristic. In practice, these neural modules are not systematically activated. Their use is conditioned by task complexity, contextual signals, and the decisions of the MetaAgent. Most subproblems—approximately 80%—are handled using lightweight symbolic heuristics or fast-response rules embedded in the agent. The internal neural structure is invoked only when needed, typically in the remaining 20% of cases that require deeper reasoning or uncertainty management. Thus, sparsity is preserved both globally across the architecture and locally within each agent, ensuring computational frugality and interpretability (Rosenbaum et al. [5]).

2.3. Representative Use Cases and Domain-Specific Agents (Schick and Schütze [6])

Each specialized agent is tailored for a domain or functional class. For example, the MedicalAgent handles health explanations, the EconomicsAgent manages financial and policy analysis, the EmpathyAgent adjusts emotional tone, the AnalogyAgent supports metaphor-based explanation, and the InstructionAgent enforces output formatting. These agents are deployed in various real-world contexts. In healthcare, the EmpathyAgent and MedicalAgent collaborate to provide contextualized, emotionally aware advice. In customer service, agents adapt tone and structure to complaints or requests with emotional sensitivity. In virtual assistant applications, the GPT-MetaAgent orchestrates agents based on their role, such as scheduler, explainer, or rephraser. In educational contexts, the AnalogyAgent, PhiloAgent, and TranslationAgent collaborate to offer diverse pedagogical support (Schmidhuber [7]). The use of hormones such as urgency or calmness, diffused through GPT-Gland Agents, complements the activation of domain agents and contributes to more human-like responses (Shazeer et al. [8])/

2.4. Differences from Traditional Expert Systems and GPT Monoliths (Shen et al. [9])

Unlike traditional GPT models where all capabilities are entangled in a single neural block, S-AI-GPT adopts a network of independent, loosely coupled agents. This introduces several key advantages : (Singh et al. [10])

Table 1. Structural and functional comparison between traditional GPT and S-AI-GPT architectures.

Feature	Traditional GPT	S-AI-GPT
Internal structure	Dense monolithic model	Network of modular agents
Activation	Full model each time	Sparse, on-demand
Interpretability	Very limited	Agent-level traceability
Personalization	Prompt-based only	Orchestrated via GPT-MetaAgent
Extendability	Requires retraining	Add/replace individual agents
Cognitive modularity	Absent	Explicit, domain-specific
Bio-inspired regulation	None	Hormonal modulation layer

This architecture enables transparent, adaptable, and efficient problem-solving [11], and extends the principle of sparse activation found in Mixture-of-Experts models [16] to symbolic and biologically inspired orchestration.2.5. Lifecycle, Adaptation, and Governance of GPT-Specialized Agents (Talebirad and Nadiri [12])

Each GPT-Specialized Agent follows a structured lifecycle. It is instantiated during system initialization, activated by the GPT-MetaAgent based on relevance, adapted through feedback and memory signals, and logged for explainability. Governance is delegated to the GPT-MetaAgent, which tracks performance, adjusts agent priorities, and disables redundant activations when needed (TechTarget [13]).

Real-world deployment confirms this lifecycle. In healthcare applications, the EmpathyAgent is reused across multiple steps to ensure consistency. In educational settings, agents such as the PhiloAgent progressively adapt their style based on feedback from the memory subsystem (Vicci [14]).

2.5. Inter-Agent Coordination in the S-AI-GPT Ecosystem

Coordination between agents occurs both horizontally and vertically. Horizontally, agents may share intermediate results via blackboard memory or other shared resources. Vertically, orchestration is handled by the GPT-MetaAgent, which governs activation, communication, and deactivation flows. Hormonal signals influence multiple agents simultaneously, modulating behavior in a coordinated fashion. The GPT-ResultAggregator component merges the diverse contributions of activated agents into a coherent and contextually relevant output. This overall design supports multi-perspective answers, collaborative problem-solving, and high reusability.

2.6. Learning Capacity and Adaptive Strategies of GPT-Specialized Agents

Each GPT-Specialized Agent is capable of adapting its behavior through various channels. These include feedback loops from the GPT-MetaAgent, signals from memory agents such as the AgentMemoire and GlandeMemoire, and the hormonal context itself. For instance, verbosity can be reduced if the urgency hormone level exceeds a certain threshold. While most SAs are pre-trained or rule-based, they also support adaptation mechanisms. These mechanisms include memorizing user preferences via memory agents, omitting verbose reasoning when simplicity is preferred, or adjusting language and tone when the empathy hormone level is high. From the analysis in Module I, it becomes evident that agents are only activated if their contribution adds

marginal value. This selective activation reinforces computational frugality without compromising output quality.

2.7. Embedded GPT-Knowledge Base and Agent Autonomy

Each agent has access to a shared or private GPT-Knowledge Base. This enables domain-specific fact checking in the MedicalAgent, contextual tone memory for the EmpathyAgent, and persistent user-specific data handling by the InstructionAgent. Access to knowledge is filtered according to the hormonal context, directives from the GPT-MetaAgent, and the agent's internal memory state. This embedded knowledge layer enhances agent autonomy, supports explainable behavior, and allows for incremental improvement of agents without requiring retraining of the global model.

2.8. Memory–Hormone–Output Interaction Diagram

This subsection introduces a system-level diagram that highlights the dynamic interplay between memory management, hormonal signaling, and context-aware output generation within the S-AI-GPT architecture. It complements the detailed descriptions from Section 1.8 by visualizing how internal regulatory loops connect hormonal signals, memory components, and adaptive output to form a coherent and responsive cognitive system.

3. BUSINESS AGENTS AND CONVERSATIONAL GPT AGENTS

The S-AI-GPT architecture is built upon a tripartite classification of agents. System Agents are responsible for orchestrating and regulating the overall functioning of the system (memory management, hormonal signaling, security, display, etc.). Specialized Conversational Agents ensure fine-grained, adaptive, and expressive verbal output. In contrast, Specialized Domain Agents are dedicated to solving specific professional or technical subproblems, leveraging expert knowledge in fields such as medicine, law, programming, or financial forecasting. This differentiation enhances system transparency, facilitates modular extensibility, and supports parsimonious and traceable orchestration.

3.1. Business-Oriented Conversational Agents

3.1.1. General Introduction

To address the full spectrum of conversational needs, the S-AI-GPT architecture integrates a suite of 25 specialized conversational domain agents. These agents cover all the fundamental competencies expected from a modern intelligent dialogue system, including semantic understanding, reasoning, emotional modulation, stylistic adaptation, pedagogical support, summarization, creativity, and practical advice. Each agent is designed to solve a specific subproblem type within a conversation and is dynamically activated by the GPT-MetaAgent based on semantic context, hormonal signals, and user preference memory. This strategy ensures comprehensive domain coverage while maintaining computational parsimony and explainability.

Beyond technical orchestration, this agent-oriented strategy echoes Marvin Minsky's *Society of Mind* [26], in which cognitive processes emerge from the coordination of specialized modules. Similarly, S-AI-GPT treats intelligent dialogue as the product of modular, context-sensitive, and hormone-modulated interactions.

3.1.2. List of Specialized Conversational Agents

Here is the detailed list of the 25 specialized agents in S-AI-GPT :

Table 2. Specialized conversational agents in S-AI-GPT and their respective roles.

Agent	Main Role
UnderstandingAgent	Analyzes and reformulates the user's request for clarification.
ReasoningAgent	Performs complex logical or inferential reasoning.
SummarizationAgent	Condenses long information into concise summaries.
ExpansionAgent	Expands an idea, argument, or explanation.
FormattingAgent	Structures responses (lists, tables, paragraphs).
TranslationAgent	Translates content into different languages or registers.
AnalogyAgent	Generates adapted analogies or metaphors.
EmpathyAgent	Adjusts emotional tone based on user sentiment.
HumorAgent	Adds appropriate humorous elements.
CreativityAgent	Proposes innovative or creative ideas.
MinimalistAgent	Simplifies answers into minimalist versions.
PhiloAgent	Introduces philosophical or reflective perspectives.
MemoryRetrievalAgent	Recalls relevant contextual memories to enrich responses.
InstructionAgent	Ensures compliance with user-specified formatting or style constraints.
CorrectionAgent	Detects and corrects errors, inconsistencies, or contradictions.
FocusAgent	Refocuses the conversation on the main topic when needed.
CounterArgumentAgent	Offers counterarguments to enrich discussions.
SimplificationAgent	Reformulates complex ideas into accessible terms.
DepthAgent	Deepens the discussion upon request.
PragmaticAgent	Proposes practical solutions and actionable advice.
EthicalAgent	Verifies ethical compliance of responses.
CuriosityAgent	Stimulates curiosity and exploration of new ideas.
DiplomacyAgent	Moderates answers to maintain respectful and diplomatic tone.
StoryTellingAgent	Transforms information into engaging narratives.
ContextAdaptationAgent	Dynamically adapts style according to conversation flow.

3.1.3. Functional Coverage of Conversational Systems

Together, these agents offer full functional coverage of conversational AI needs. Understanding and clarification are addressed by the UnderstandingAgent and MemoryRetrievalAgent. Reasoning and debate are supported by the ReasoningAgent, CounterArgumentAgent, and DepthAgent. Summarization and structuring are handled by the SummarizationAgent and FormattingAgent. Creativity and storytelling are enabled through the CreativityAgent, AnalogyAgent, and StoryTellingAgent. Emotional management is guided by the EmpathyAgent, HumorAgent, and DiplomacyAgent. Stylistic adaptation is ensured via the InstructionAgent,

MinimalistAgent, and ContextAdaptationAgent. Multilingual support is provided by the TranslationAgent. Pedagogy and accessibility are delivered by the SimplificationAgent and PhiloAgent. Safety and ethics are enforced by the CorrectionAgent and EthicalAgent. Finally, pragmatic advice is offered by the PragmaticAgent, while the CuriosityAgent and FocusAgent foster engagement and motivation.

3.1.4. Activation Modulated by Context and Hormonal Signals

Agent activation within S-AI-GPT is not systematic. It depends on the subproblem identified by the GPT-DecompositionAgent, is influenced by the current hormonal profile (such as urgency, empathy, or creativity), and is modulated by user memory traces managed by the Memory Agents. This mechanism allows the system to respond precisely to each user's expectations by combining fine-grained personalization, resource optimization, and complete traceability. Every agent activation is explainable and linked to its internal context.

3.2. Business Agents for GPT Systems

3.2.1. Introduction

While conversational agents represent a major class of specialized components within S-AI-GPT, the architecture extends far beyond dialogue to encompass the full spectrum of modern GPT-based applications. To address the diversity of emerging use cases—including creative writing, decision support, programming assistance, and autonomous task execution—S-AI-GPT integrates a broad range of Specialized Domain Agents. These agents are designed to tackle distinct professional, creative, technical, and analytical tasks within a modular, sparse, and context-sensitive framework.

3.2.2. Typology of Specialized Domain Agents

S-AI-GPT incorporates multiple families of domain-specific agents, each tailored to address specialized needs. Content Generation Agents include the ContentGenerationAgent for general writing, the StorytellingAgent for narrative creation, and the CreativeWritingAgent for poetic and fictional content. Programming and Code Assistance Agents consist of the CodeCompletionAgent for generation, the DebuggingAgent for error correction, and the CodeExplanationAgent for source code interpretation. Decision Support and Analysis Agents include the LegalAnalysisAgent for document analysis, the MedicalAnalysisAgent for diagnostic assistance, and the FinancialForecastAgent for predictive modeling. Knowledge Retrieval and Validation Agents are represented by the KnowledgeRetrievalAgent for information lookup and the FactCheckingAgent for factual verification. Autonomous Planning and Execution Agents include the GoalSettingAgent for defining objectives, the TaskExecutionAgent for managing execution loops, and the ReflectionAgent for self-assessment. Finally, Multimodal Interaction Agents such as the ImageAnalysisAgent, AudioTranscriptionAgent, and MultimodalFusionAgent enable processing of image, audio, and cross-modal inputs. These agents operate independently but collaborate when orchestrated by the GPT-MetaAgent.

3.2.3. Common Structural Characteristics

All Specialized Domain Agents share a common architectural framework. They inherit from a unified BaseAgent class and expose a standardized interface via the process method, which takes a subproblem and optional contextual parameters. The method signature is designed for contextual adaptation and hormonal reactivity. Each agent is natively integrated with the Hormonal Engine, enabling its behavior to be modulated by contextual variables such as urgency,

creativity, emotional tone, or required factual rigor. This design guarantees interoperability, traceability, and flexible composition within complex cognitive workflows.

3.2.4. Activation Strategies and Contextual Selection

The GPT-MetaAgent plays a central role in managing these agents. It selects the most relevant Specialized Domain Agents based on the semantic decomposition of the user's request. It modulates their activation intensity through artificial hormonal signals, increasing creativity or promoting factual rigor depending on the context. It also enables cooperative activation when multi-perspective processing is beneficial, such as combining legal and financial analysis. Only agents strictly necessary to solve a given subproblem are activated, thereby preserving energy efficiency and improving interpretability.

3.2.5. Illustrative Use Cases

Several real-world examples highlight the strategic value of these agents. In content creation, the CreativeWritingAgent and StorytellingAgent work together to generate imaginative articles tailored to a user's stylistic and emotional preferences. In software development, the CodeCompletionAgent and DebuggingAgent assist in building and refining codebases, while the CodeExplanationAgent documents critical sections. In medical decision support, the MedicalAnalysisAgent interprets patient data, the KnowledgeRetrievalAgent queries relevant medical databases, and the EmpathyAgent ensures emotionally sensitive communication. For fact validation, the KnowledgeRetrievalAgent and FactCheckingAgent collaborate to verify claims and prevent hallucinations.

4. AUTONOMY IN SPECIALIZED AGENTS : A REVIEW OF EXISTING RESEARCH AND S-AI-GPT'S APPROACH

4.1. Introduction to Autonomous Agents

The concept of autonomous agents has been extensively explored in AI research, particularly in the domains of multi-agent systems (MAS), robotics, expert systems, and conversational AI. In these systems, agents are often designed to perform tasks autonomously, but the level of autonomy and the scope of the tasks handled can vary significantly. This section reviews key research on autonomous agents, highlighting the major trends, challenges, and solutions, and compares them with the S-AI-GPT approach to specialized, context-sensitive autonomous agents.

4.2. Multi-Agent Systems and Robotics

4.2.1. Multi-Agent Systems (MAS)

Multi-Agent Systems have been one of the primary research areas focusing on the autonomy of individual agents within a larger system. MAS refers to systems composed of multiple agents that can act independently to achieve common or individual goals. These agents are typically autonomous, meaning they can make decisions based on their environment and objectives without human intervention. JADE (Java Agent Development Framework) is a widely used framework for developing autonomous agents. It provides tools for agent communication, coordination, and learning, enabling agents to act autonomously within a collaborative environment. However, JADE agents often lack the fine-grained modularity and task-specific specialization seen in S-AI-GPT, where agents are activated only when needed. FIPA (Foundation for Intelligent Physical Agents) also offers standards and tools for developing

autonomous agents, with an emphasis on inter-agent communication and coordination. While MAS frameworks focus on coordination between autonomous agents, S-AI-GPT takes a more modular approach, allowing agents to specialize in specific tasks, optimizing resource usage, and ensuring contextual adaptation.

4.2.2. Autonomous Agents in Robotics and Physical Systems

In the field of robotics, agents are often designed to perform autonomous tasks in real-world environments, such as navigation, decision-making, and task execution. These systems often rely on reinforcement learning and planning algorithms to make decisions based on sensory input. RoboCup is an example where autonomous robots collaborate in a team to play soccer. These robots make decisions about movement, strategy, and team coordination in real time, adapting to changes in the environment. However, the complexity of these agents often requires specialized hardware and high resource consumption, unlike the lightweight agents in S-AI-GPT, which are optimized for computational efficiency. Autonomous Vehicles, such as those developed by companies like Waymo and Tesla, implement decision-making systems that allow vehicles to make complex decisions in real time. These systems require extensive training and computational power to process sensory data and adjust driving behavior. In contrast, S-AI-GPT's agents are designed to be task-specific and modular, activated only when required, and orchestrated by a MetaAgent, reducing the overall computational burden.

4.3. Expert Systems and Knowledge-Based Autonomous Agents

Expert systems were among the earliest examples of AI that attempted to model expert knowledge and decision-making in a narrow field. In autonomous expert systems, the agent is typically designed to act as an independent decision-maker in a specialized domain, such as medical diagnosis, financial analysis, or legal consultation. MYCIN, an early expert system for medical diagnosis, made autonomous decisions based on a set of rules and a knowledge base. While MYCIN operated autonomously, it lacked the modularity and context-sensitive adaptation present in S-AI-GPT, where specialized agents operate on-demand and dynamically adapt their behavior. Modern systems such as IBM Watson extend the idea of autonomous expert systems, offering decision support in domains like medicine and law. These systems, however, are often monolithic and require extensive computation and data storage, in contrast to the lightweight, modular agents of S-AI-GPT, which are more resource-efficient and contextually flexible.

4.4. Autonomous Agents in Conversational AI

The emergence of conversational agents such as chatbots and virtual assistants has opened up new avenues for autonomy in natural language processing systems. These systems aim to handle a variety of tasks autonomously, such as answering queries, providing recommendations, and managing user interactions. Rasa and Dialogflow are popular frameworks for building autonomous conversational agents. While these systems can handle specific conversations, they are typically monolithic and require predefined intents or dialogue trees. In contrast, S-AI-GPT offers a more modular and specialized approach, allowing agents to work in parallel on specific tasks and activate dynamically based on the context of the conversation. More advanced systems, such as OpenAI's GPT models, show promise in autonomy by generating responses based on a wide range of inputs. However, they still lack the fine-grained specialization and context-aware decision-making capabilities of S-AI-GPT, where agents are activated only for the tasks they are specialized in, ensuring both efficiency and relevance.

4.5. Modular Architecture of Autonomous Specialized Agents

The architecture of S-AI-GPT is based on deep modularity, where each specialized agent operates independently while being orchestrated by a central MetaAgent. Each agent is designed to handle a specific task and is activated only when necessary. This structure ensures high flexibility while optimizing resource usage. The MetaAgent, as the central coordinator, decides which agents should be activated based on system needs, thus guaranteeing a parsimonious use of computational resources.

4.6. Contextual Activation and Adaptation of Autonomous Agents

The activation of autonomous agents in S-AI-GPT is strongly influenced by the context in which they operate. Thanks to dynamic management by the MetaAgent, agents are activated only when their domain expertise is required to solve a specific task. This process is also modulated by artificial hormonal signals, which allow agents to adapt their behavior according to contextual criteria such as creativity, factual rigor, or emotional tone. Contextual activation and adaptation ensure that agents operate optimally while remaining lightweight and efficient.

4.7. Learning and Adaptation Mechanisms of Autonomous Agents

Autonomous agents in S-AI-GPT are equipped with mechanisms that allow them to learn and adapt continuously. They can use supervised and unsupervised learning techniques to refine their capabilities based on interactions and user feedback. For instance, a medical analysis agent can improve over time by learning from new trends or adjusting its predictions. These agents can also use reinforcement learning techniques to make optimal decisions and adapt to evolving environments or contexts.

4.8. Communication and Cooperation Between Agents

Although each agent is autonomous, S-AI-GPT relies on effective inter-agent communication to solve complex tasks. Specialized agents can share information, collaborate on shared subproblems, and coordinate to achieve a global objective, in a spirit comparable to modular multi-agent learning systems studied in [15].

This cooperation is orchestrated by the MetaAgent, which ensures that agents work together coherently and efficiently. Agents may be activated to collaborate simultaneously on a task, enabling S-AI-GPT to tackle problems more complex than what any single agent could solve alone.

These foundational principles of inter-agent cooperation form the basis for orchestrating complex reasoning flows in S-AI-GPT. Their practical deployment will be further illustrated in the upcoming implementation-focused article.

4.9. Future Directions and Evolution of Autonomous Agents in S-AI-GPT

Autonomous specialized agents in S-AI-GPT are already a major asset of the system, but ongoing improvements are possible. The integration of techniques such as meta-learning—where agents learn from interactions with other agents—could enhance their autonomy and adaptability. New features could also be introduced to enable agents to handle multimodal information such as text, image, and audio, thereby increasing the versatility and efficiency of the system.

5. THE DECOMPOSITION AGENT IN S-AI-GPT

5.1. Introduction to Semantic Decomposition

Decomposition is a cornerstone of the S-AI-GPT architecture. Unlike monolithic language models that process entire prompts holistically, S-AI-GPT separates complex queries into manageable subtasks. This selective breakdown, driven by the GPT-DecompositionAgent, allows for more efficient resource usage, better task allocation, and clearer interpretability. Following the 80/20 heuristic, approximately 80% of user needs are classified as simple tasks solvable via symbolic or rule-based mechanisms, while the remaining 20% require more sophisticated, neuro-symbolic processing.

5.2. Role of the GPT-Decomposition Agent

The GPT-DecompositionAgent plays a central role by transforming a global user request into multiple subproblems. These subtasks are then relayed to the GPT-MetaAgent, which selects the appropriate GPT-Specialized Agents for execution. The GPT-DecompositionAgent applies heuristic rules, task-specific templates, or S-AI-GPT-Parsing (symbolic mode) to segment the request by domain, such as factual, emotional, or stylistic components. It also labels and prioritizes subtasks, and structures communication with the GPT-MetaAgent via structured objects. This modular flow guarantees that only the necessary agents are mobilized, minimizing cognitive redundancy and ensuring explainable processing pipelines.

5.3. Symbolic and Neuro-Symbolic Processing

S-AI-GPT leverages a dual processing strategy. Symbolic processing is used for the majority of cases (approximately 80%), handling simple factual, definitional, or stylistic requests through lightweight deterministic rules. Neuro-symbolic processing addresses the remaining 20%, using neural reasoning models such as BERT, DistilBERT, or GPT, combined with semantic rules, to tackle multi-faceted, abstract, or context-heavy problems. For example, a query like “What is the capital of Morocco ? ” triggers a symbolic retrieval mechanism, whereas a more complex question such as “What are the geopolitical implications of AI in warfare ?” leads to decomposition into economic, political, and ethical subtasks, each processed by specialized neuro-symbolic agents.

5.4. Dynamic GPT-Knowledge Base Integration

The GPT-DecompositionAgent accesses and enriches a dynamic GPT-Knowledge Base. This base is continuously updated with user interactions and new domain facts, is context-aware and adapts to the emotional and stylistic profile of the user, and is filtered by hormonal context and GPT-MetaAgent priorities. Such integration allows the GPT-DecompositionAgent to reason not only over fixed rules but also over evolving, context-dependent information.

5.5. Learning and Adaptive Refinement

Two modes of learning govern the GPT-DecompositionAgent. Supervised learning relies on labeled datasets or human feedback to refine the rules and improve prompt segmentation and classification. Adaptive learning uses user preferences, agent feedback, and task success or failure signals to adjust rule thresholds and decision paths. This hybrid learning approach enables co-evolution of the decomposition strategy with the orchestration logic of the GPT-MetaAgent.

5.6. Illustrative Example : Trump’s Tariff Policy

When given a prompt such as "What will be the impact of Donald Trump's tariff policy ?", the GPT-DecompositionAgent extracts three main subproblems: economic impact, political impact, and social impact. These subtasks are delegated respectively to the EconomicImpactAgent, PoliticalImpactAgent, and SocialImpactAgent. Each of these agents independently processes its subtask, and their outputs are returned to the GPT-MetaAgent for aggregation, formatting, and delivery.

The result aggregation process, although not covered in this article, is handled by a dedicated GPT-ResultAggregator module and will be detailed in the forthcoming implementation article.

5.7. Evolutionary Rule Learning and Maintenance

To remain robust over time, the GPT-DecompositionAgent includes mechanisms for incremental rule refinement. These mechanisms rely on feedback-driven learning to detect decomposition errors or subtask misassignments, pruning of obsolete rules that no longer align with updated knowledge or usage patterns, and integration of user preferences to guide decomposition paths in alignment with recurring stylistic or semantic expectations. This continuous evolution allows the system to remain attuned to real-world complexity and user-specific behavior without requiring costly retraining of the global architecture.

5.8. Advanced Techniques for Contextual Decomposition

To ensure maximal adaptability across a wide range of user intents and linguistic styles, the GPT-DecompositionAgent integrates several advanced techniques beyond standard symbolic parsing and domain segmentation. These techniques allow the system to handle subtle, indirect, or emotionally charged requests while preserving computational parsimony.

5.8.1. Intention Parsing

Beyond syntactic analysis, the agent performs intent-oriented parsing by identifying key linguistic indicators that reflect the user's objective—whether they seek an explanation, a comparison, a prediction, or a recommendation. This approach relies on verb forms, modal auxiliaries, and discourse structures that suggest intent. For example, a prompt beginning with “Should I...” is flagged as decision-oriented, whereas “Why does...” triggers causal decomposition. Such parsing ensures the selection of agent types that align with the communicative goal.

5.8.2. Specialized Micro-Parsers

The decomposition engine internally delegates specific parsing tasks to a suite of micro-parsers, each specialized in extracting one semantic dimension: time (TemporalParser), emotion (AffectiveParser), domain (DomainMapper), argument structure (ArgumentParser), or implicit negation (NegationDetector). These lightweight symbolic agents allow fine-grained decomposition without invoking deep language models, thus aligning with the sparse activation principle.

5.8.3. Hormone-Guided Parsing Modulation

S-AI-GPT-Parsing strategy is also influenced by the current hormonal profile, as determined by the GPT-MetaAgent and contextual glands. For instance, a “doubt” hormone may trigger the

activation of a more fine-grained parsing mode, splitting a vague query into multiple clarification subtasks. Conversely, an "urgency" hormone might bias the decomposition toward direct, time-sensitive aspects of the request. This biologically inspired modulation allows the system to dynamically prioritize parsing strategies without global reconfiguration.

5.8.4. Template Expansion for Implicit Requests

In some cases, user prompts contain implicit intentions not overtly expressed. The agent leverages a library of emotional and stylistic templates to reframe ambiguous or emotionally loaded statements into actionable subtasks. For instance, a sentence like "I'm not sure this tree will survive summer" is parsed as a request for predictive and advisory subagents (e.g., WeatherAgent, PlantCareAgent), even in the absence of explicit question formatting. This template-based transformation bridges emotional expressiveness and symbolic processing.Taken together, these advanced decomposition mechanisms allow the system to intelligently interpret user queries that go beyond factual content, while preserving the lightweight, modular, and explainable properties of the S-AI-GPT framework.

6. HORMONAL ENGINE AND ADAPTIVE REGULATION IN S-AI-GPT

6.1. Introduction

The artificial hormonal engine is one of the most innovative components of the S-AI-GPT system. It enables soft, decentralized, and context-sensitive regulation of agent behavior, inspired by biological endocrine mechanisms. Unlike traditional architectures that directly activate logic-based modules, S-AI-GPT uses synthetic hormonal signals to adjust response tone, intensity, emotional weight, or task priority. These signals influence both agent activation and the affective nature of the outputs.

6.2. Role of GPT-Gland Agents in Hormonal Signaling

GPT-Gland Agents are responsible for emitting artificial hormones, under the supervision of the GPT-MetaAgent. Each gland corresponds to a specific emotional or contextual pattern and triggers the release of a matching signal such as urgency, softness, authority, or calmness. For instance, an urgency gland accelerates the response and activates rapid-processing agents, while a softness gland boosts empathic behavior and selects gentle response tones. These signals are then processed by the hormonal engine to dynamically influence the behavior of downstream agents.

6.3. Hormonal Engine: Internal Functioning

The Hormonal Engine manages the diffusion of hormone-like signals across the system, the temporal decay and progressive fading of signals, intensity encoding (strong, medium, weak), and targeted propagation to affected agents or submodules. It acts as a dynamic regulator, adjusting agent activation priority, stylistic tone of generated responses, and the engagement duration or processing depth. This asynchronous mechanism ensures adaptive, flexible, and non-blocking modulation of system behavior. Beyond these core mechanisms, the Hormonal Engine plays a broader role in the overall architecture. Operating under the exclusive control of the GPT-MetaAgent, it simulates biologically inspired mechanisms of hormone release, propagation, degradation, and regulatory feedback, based on semantic indicators extracted by the system and the currently active user profile. Through this hormonal layer, the MetaAgent can orchestrate the sparse activation of agents, regulate the intensity of their contributions, and foster the emergence of context-aware responses. As a result, the centralized hormonal engine becomes a key

mechanism to introduce plasticity, computational frugality, and bio-inspired coordination across the entire system.

6.4. Contextual Hormonal Profiles

Each artificial hormone has an associated contextual profile that defines its intent—such as calm, urgency, tension, or reassurance—its lifespan, which may be transient or persistent, and its scope, whether global, domain-specific, or task-local. These profiles are selected automatically by the GPT-MetaAgent based on contextual cues such as emotional signals from user input, or are learned adaptively via reinforcement and memory feedback loops. For example, a calmness hormone may reduce the system's cognitive pressure and activate the EmpathyAgent, while a precision hormone might boost fact-checking by promoting activation of agents such as the MathAgent or FactAgent.

6.5. Adaptive Regulation Through Hormonal Modulation

Hormonal modulation allows S-AI-GPT to evolve its behavior in real-time. Agent activation depends not only on the problem structure but also on the current hormonal profile. Agents can dynamically adapt their style, verbosity, or empathy level based on received hormonal signals. The GPT-MetaAgent adjusts hormone profiles in real time based on emotional cues, user history, and system load. This creates a reactive balance between precision, personalization, and computational efficiency.

6.6. Evolution and Long-Term Optimization of Hormonal Profiles

S-AI-GPT supports long-term enhancement of its hormonal signaling system through the addition of new hormones for emerging behavioral traits such as focus, clarity, or prudence, dynamic threshold tuning via incremental learning, and the design of new self-adaptive glands capable of reconfiguring their signals based on interaction history. These capabilities ensure durability, adaptability, and evolving personalization over extended use.

6.7. Functional Differentiation of Hormonal Components in S-AI-GPT

Although the hormonal regulation system in S-AI-GPT may initially appear homogeneous, it relies on a modular architecture composed of three complementary entities operating at different levels of scale and temporality : the central Hormonal Engine, the GPT-Gland Agents, and the Memory Gland Agent. Each plays a specific role in the bio-inspired simulation of artificial hormonal modulation and collectively contributes to the system's emotional plasticity, affective coherence, and contextual adaptation.

6.7.1. The Hormonal Engine – Core Mechanism for Diffusion and Real-Time Modulation

The Hormonal Engine constitutes the operational core of hormonal diffusion within the system. It is responsible for the generation, propagation, and real-time degradation of artificial hormones. Acting as a low-level biological simulator, it manages asynchronous signal dissemination, temporal decay, and the overall hormonal concentration of the system. Under the supervision of the MetaAgent, it simulates mechanisms of hormone release, propagation, feedback, and deactivation. These mechanisms are triggered by semantic indicators and modulated by the active user profile. This allows the MetaAgent to orchestrate sparse activation of agents and finely regulate their intensity, promoting adaptive and contextualized responses. The Hormonal Engine is a key lever for introducing plasticity, frugality, and coordination.

6.7.2. GPT Gland Agents – Targeted Emitters and Local Modulators

GPT-Gland Agents act as targeted emitters of hormones. Each gland is associated with a specific cognitive or emotional dimension, such as empathy, vigilance, or focus, and can be activated punctually by the MetaAgent. They inject emotionally charged hormonal signals into the system to influence tone, rhythm, or urgency. Each gland includes a HormonalModule—a micro-engine for hormonal regulation—allowing the agent to manage internal hormonal reactions, emit secondary hormones, and adapt style or tempo. This local regulation ensures agent-level behavioral adaptation while remaining aligned with global orchestration.

6.7.3. The Memory Gland Agent – Emotional Memory and Long-Term Regulation

The Memory Gland Agent encodes long-term emotional memory. Unlike the volatile DCM, it maintains a reservoir of emotional impact, recurrence frequency, and affective charge of interactions. It learns which hormonal profiles were most effective and anticipates future configurations. Before each session, it may recommend emotional presets, adjust parameters, or reconfigure diffusion strategies. Its evolving personalization capacity makes it essential for affective continuity.

6.7.4. Triadic Structure and Functional Differentiation

Together, the Hormonal Engine, GPT-Gland Agents, and Memory Gland Agent form a triadic hormonal regulation system. This structure allows S-AI-GPT to adapt dynamically to emotional and contextual signals, ensure long-term coherence, and simulate realistic emotional expressiveness, while preserving computational efficiency required for distributed or embedded environments.

6.7.5. Summary Table – Differentiated Roles

Table 4. Typology of hormonal components in S-AI-GPT with functional roles and biological analogies.

Component	Primary Function	Nature of Action	Temporality	Biological Analogy
Hormonal Engine	Hormone diffusion and degradation	Operational	Real-time	Autonomic nervous system
GPT Gland Agents	Emission of specific hormones on demand	Contextual emission	Instantaneous	Endocrine glands
Memory Gland Agent	Learning and anticipation of hormonal profiles	Strategic recommendation	Medium / long-term	Limbic memory (hippocampus)

These three components together form a bio-inspired hormonal regulation triad. They enable S-AI-GPT to: dynamically modulate its behavior based on emotional and contextual cues, maintain long-term affective coherence, and simulate realistic sensitivity in its dialogical interactions. This functional separation is essential to achieve both emergent emotional expressiveness and computational parsimony, especially in constrained or embedded environments.

6.8. Typology of Glands and Hormones in Conversational S-AI-GPT Systems

In the S-AI-GPT framework, artificial glands play a crucial regulatory role by emitting digital hormones that modulate agent activation, emotional resonance, task persistence, and dynamic prioritization. Inspired by biological endocrine mechanisms, each gland operates either locally

(close to a specialized agent) or systemically (through the MetaAgent), enabling lightweight, interpretable, and decentralized modulation of cognitive activity.

6.8.1. Functional Gland-Hormone Mapping

The table below outlines the main artificial glands and their associated hormonal signals within conversational interactions:

Gland Component	Hormonal Signal	Core Function	Typical Use Case	Biological Analogy
MemoryGland	EmotionalTraceHormone	Amplifies affective resonance from engrams	Trigger emotional memory recall	Amygdala / Hippocampus
TaskPersistenceGland	PersistenceHormone	Maintains activation during long user turns	Long reasoning chains or iterative requests	Adrenal Cortex
CuriosityGland	ExplorationHormone	Stimulates knowledge-seeking agents	Open-ended or ambiguous questions	Dopaminergic System
ConfidenceGland	ConfidenceHormone	Boosts assertive response agents	After multiple confirmations or known facts	Prefrontal Cortex Feedback
InhibitionGland	InhibitionHormone	Suppresses irrelevant agents	When user input narrows focus or urgency rises	GABAergic Pathways
UrgencyGland	AlertnessHormone	Prioritizes real-time agents	Quick clarification or system interruption	Adrenaline System

Each hormone can propagate through the **HormonalEngine**, interact with the **MetaAgent**, and dynamically influence agent orchestration.This preliminary typology provides a functional overview of the main glands and hormonal signals orchestrating conversational dynamics in S-AI-GPT. A more in-depth exploration—including signaling pathways, temporal dynamics, and inter-gland coordination—will be presented in a forthcoming article entirely dedicated to hormonal modulation in S-AI-GPT.

6.8.2. Example: Dialogue Modulation by Hormones

In a multi-turn conversation, if the user recalls a prior negative experience, the MemoryGland releases an EmotionalTraceHormone, which:reinforces memory retrieval agents,inhibits unnecessary elaboration agents (via InhibitionHormone), and amplifies response precision through ConfidenceHormone.This modulation yields a context-aware, emotionally aligned, and computationally sparse response.

6.8.3. Benefits and Future Expansion

The gland-hormone framework allows for: transparent and traceable decision flows, parsimonious activation of specialized agents, and bio-inspired modulation aligned with user affective cues.Future versions of S-AI-GPT may integrate personalized gland profiles, adjusted based on long-term usage patterns and user preferences.

7. CONCLUSION AND PERSPECTIVES

This article has presented an in-depth exploration of the bio-inspired hormonal modulation and adaptive memory mechanisms integrated within the S-AI-GPT system. Building upon the modular and parsimonious foundation introduced in the first article, we have detailed the design and function of a triadic hormonal regulation architecture comprising the central Hormonal Engine, the distributed GPT Gland Agents, and the Memory Gland. This structure enables flexible, soft, and context-aware modulation of agent activation, emotional tone, and response style, inspired by biological endocrine systems. We also described the tightly coupled memory subsystem, including the Dynamic Contextual Memory and a bio-inspired architecture of neuronal mini-structures and artificial engrams, which collectively support personalized, persistent, and emotionally coherent interactions. These components form an integrated cognitive ecosystem, orchestrated by the GPT-MetaAgent, which dynamically adapts to semantic cues, user profiles, and hormonal feedback, thereby achieving a balance between computational efficiency, interpretability, and rich user experience. The artificial hormonal signaling layer not only provides a novel approach to emotional and contextual regulation in conversational AI, but also opens new avenues for scalable, adaptive, and human-centered dialogue systems. By decoupling activation from rigid logic flows and embracing fluid hormonal modulation, S-AI-GPT demonstrates the potential for enhanced responsiveness, personalization, and resource frugality in complex AI ecosystems.Beyond its modularity and biological inspiration, the S-AI-GPT architecture also offers promising avenues for cognitive interpretability.

By embedding symbolic memory structures, hormonal modulation, and semantic control loops, the system moves toward a transparent and explainable form of conversational artificial intelligence. This addresses growing demands for human-centric and ethically grounded AI systems—capable not only of understanding but also justifying their reasoning and emotional states. Looking forward, promising research directions include the personalization of hormonal profiles to better match individual user emotional styles, extension of the hormonal system with new axes such as motivation and attention, and meta-learning frameworks for hormonal orchestration. Further, the generation of hormone-sensitive specialized agents tailored to complex, emotion-rich problem domains represents a fertile avenue for development.

Finally, adapting these mechanisms for deployment in resource-constrained environments such as edge computing, while maintaining real-time responsiveness, is a critical challenge. The forthcoming third article will detail the full implementation of S-AI-GPT, illustrating how hormonal signaling, memory, and modular orchestration combine in practical scenarios. It will also provide insights into code architecture, experimental validation, and deployment strategies for cloud and embedded applications, thus completing the vision of a sustainable, explainable, and emotionally intelligent conversational AI system.

REFERENCES

[1] M. Moutoussis and R. J. Dolan, “How computation connects affect,” *Trends Cogn. Sci.*, vol. 19, no. 4, pp. 157–163, 2015.
[2] R. W. Picard, *Affective Computing*, MIT Press, 1997.
[3] C. Qu, S. Zhang, Y. Li, and J. Ma, “Tool learning with LLMs: A survey,” *arXiv preprint*, arXiv:2405.17935, 2024.
[4] T. Schick and H. Schütze, “Toolformer: Language models can teach themselves to use tools,” *arXiv preprint*, arXiv:2302.04761, 2023.
[5] F. A. Gers and J. Schmidhuber, “Recurrent nets that time and space the gradient,” *Neural Computation*, vol. 12, no. 7, pp. 1789–1804, 2000.

[6] Y. Shen, K. Zhang, Y. Wang, and X. Liu, "HuggingGPT: Solving AI tasks with ChatGPT and its friends," *arXiv preprint*, arXiv:2303.17580, 2023.

[7] S. Slaoui, "S-AI: Sparse Artificial Intelligence System with MetaAgent," *Int. J. Fundam. Mod. Res. (IJFMR)*, vol. 1, no. 2, pp. 1–18, 2025.

[8] Y. Talebirad and A. Nadiri, "Multi-agent collaboration: Harnessing LLM agents," *arXiv preprint*, arXiv:2306.03314, 2023.

[9] T. B. Richards, *AutoGPT* [Computer software], GitHub Repository, 2023. [Online]. Available: https://github.com/Torantulino/Auto-GPT

[10] C. Rosenbaum, T. Klinger, and M. Riemer, "Routing networks for multi-task learning," in *Proc. 7th Int. Conf. Learning Representations (ICLR)*, New Orleans, LA, USA, 2019.

[11] TechTarget, "Mixture-of-experts models explained: What you need to know," *SearchEnterpriseAI*, 2024. [Online]. Available : https://www.techtarget.com/searchenterpriseai/definition/mixture-of-experts

[12] J. Schmidhuber, "Curiosity and boredom in neural controllers," in *Proc. Int. Conf. Simulation of Adaptive Behavior*, pp. 424–429, 1991.

[13] H. Vicci, "Emotional intelligence in AI: Review and evaluation," *SSRN Working Paper*, 2024. [Online]. Available: https://ssrn.com/abstract=4768910

[14] S. Slaoui, "Bio-Inspired Architecture for Parsimonious Conversational Intelligence: The S-AI-GPT Framework," *Int. J. Artif. Intell. & Applications (IJAIA)*, vol. 16, no. 4, 2025.

[15] A. Goyal, J. Binas, Y. Bengio, and C. Pal, "Coordination and learning in modular multi-agent systems," in *Adv. Neural Inf. Process. Syst. (NeurIPS)*, 2021.

[16] N. Shazeer, A. Mirhoseini, K. Maziarz, A. Davis, Q. V. Le, G. Hinton, and J. Dean, "Outrageously large neural networks: The sparsely-gated mixture-of-experts layer," *arXiv preprint*, arXiv:1701.06538, 2017.

[17] S. Singh, S. Bansal, A. El Saddik, and M. Saini, "From ChatGPT to DeepSeek AI: Revisiting monolithic and adaptive AI models," *arXiv preprint*, arXiv:2504.03219, 2025. [Online]. Available: https://arxiv.org/abs/2504.03219

[18] G. Montero Albacete, A. López, and A. García-Serrano, "Fattybot: Hormonal chatbot," *Information*, vol. 15, no. 8, p. 457, 2024.

[19] L. Cañamero and J. Fredslund, "I show you how I like you – Can you read it in my face?" *IEEE Trans. Syst., Man, and Cybern., Part A*, vol. 31, no. 5, pp. 454–459, 2001.

[20] D. Hassabis, D. Kumaran, C. Summerfield, and M. Botvinick, "Neuroscience-inspired artificial intelligence," *Neuron*, vol. 95, no. 2, pp. 245–258, 2017.

[21] G. E. Hinton and R. R. Salakhutdinov, "Reducing the dimensionality of data with neural networks," *Science*, vol. 313, no. 5786, pp. 504–507, 2006.

[22] V. Mnih, K. Kavukcuoglu, D. Silver, A. A. Rusu, J. Veness, M. G. Bellemare, A. Graves, M. Riedmiller, A. K. Fidjeland, G. Ostrovski, S. Petersen, C. Beattie, A. Sadik, I. Antonoglou, H. King, D. Kumaran, D. Wierstra, S. Legg, and D. Hassabis, "Human-level control through deep reinforcement learning," *Nature*, vol. 518, no. 7540, pp. 529–533, 2015.

[23] J. Andreas, M. Rohrbach, T. Darrell, and D. Klein, "Neural module networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, pp. 39–48, 2016.

[24] T. B. Brown, B. Mann, N. Ryder, M. Subbiah, J. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, S. Agarwal, A. Herbert-Voss, G. Krueger, T. Henighan, R. Child, A. Ramesh, D. M. Ziegler, J. Wu, C. Winter, C. Hesse, M. Chen, E. Sigler, M. Litwin, S. Gray, B. Chess, J. Clark, C. Berner, S. McCandlish, A. Radford, I. Sutskever, and D. Amodei, "Language models are few-shot learners," *arXiv preprint*, arXiv:2005.14165, 2020. [Online]. Available: https://arxiv.org/abs/2005.14165

[25] G. Chen, S. Liu, H. Wu, Q. Zhou, and X. Chen, "AutoAgents: A framework for automatic agent generation," in *Proc. Int. Joint Conf. Artif. Intell. (IJCAI-24)*, Jeju, Korea, 2024. [Online]. Available: https://arxiv.org/abs/2405.06758.

[26] M. Minsky, *The Society of Mind*, Simon & Schuster, 1986.

AUTHOR

Said Slaoui is a professor at Mohammed V University in Rabat, Morocco. He graduated in Computer Science from University Pierre and Marie Curie, Paris VI (in collaboration with IBM France), 1986. He has over 40 years of experience in the fields of AI and Big Data, with research focused on modular architectures, symbolic reasoning, and computational frugality. His recent work introduces the Sparse Artificial Intelligence (S-AI) framework, which integrates bio-inspired signaling and agent-based orchestration. He has published numerous scientific papers in international journals and conferences, and actively contributes to the development of sustainable and explainable AI systems.