BIO-INSPIRED ARCHITECTURE FOR PARSIMONIOUS CONVERSATIONAL INTELLIGENCE : THE S-AI-GPT FRAMEWORK

Said Slaoui

University Mohammed V, Rabat, Morocco

ABSTRACT

S-AI-GPT, a conversational artificial intelligence system, is based on the principles of Sparse Artificial Intelligence (S-AI) developed by the author. S-AI-GPT provides a modular and bio-inspired solution to the structural limitations of monolithic GPT-based language models, particularly in terms of excessive resource consumption, low interpretability, and limited contextual adaptability. This proposal is part of a broader effort to design sustainable, explainable, and adaptive AI systems grounded in cognitive principles.

The sparse activation of specialized GPT agents, coordinated by a central GPT-MetaAgent, and a cognitive framework modeled after the functional modularity of the human brain form the foundation of the system. These agents are activated only when relevant, based on task decomposition and contextual cues. Their orchestration is handled through an internal symbolic pipeline, designed for transparency and modular control.

The rationale for the paradigm shift is explained in this article along with relevant literature reviews, the modular system architecture, and the agent-based decomposition and orchestration logic that form the basis of S-AI-GPT. Each component is introduced through a conceptual analysis, highlighting its function and integration within the overall architecture. By doing this, the article establishes the foundation for upcoming improvements that will be discussed in later articles and are based on artificial hormonal signaling and cognitive memory subsystems. This is the first paper in a three-part series, with subsequent work addressing personalization, affective regulation, and experimental validation.

KEYWORDS

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

1. INTRODUCTION: FROM GPT LIMITATIONS TO MODULAR INTELLIGENCE

ChatGPT and other large language models (LLMs) have fundamentally altered the course of artificial intelligence. They are highly beneficial for dialogue generation and natural language processing. These monolithic architectures still have a number of significant shortcomings, though. No matter how difficult the task is, they activate the entire model, which is too expensive to compute, difficult to comprehend, difficult to decompose into smaller components, and difficult to maintain context over extended interactions. They also lack task-specific specialization, explicable reasoning pathways, and fine-grained emotional modulation.

S-AI-GPT applies this concept to conversational AI by fusing modular parsimony with biologically inspired approaches to regulation via synthetic hormone signaling. Similar to how the human brain only activates the parts required for a specific task, the system only activates the

parts required for a given task. As a result, the reasoning process becomes more effective, adaptable, and simple to comprehend.

To address these issues, the Sparse Artificial Intelligence (S-AI) [21] paradigm provides a cost-effective and modular approach. This approach distributes cognitive tasks among a network of specialized agents, each of which is best suited for a particular task, like formatting, emotion, memory, or reasoning. Its foundations are agent-oriented orchestration, selective activation, and task decomposition. A central orchestrator called the GPT-MetaAgent ensures intelligent delegation, coherence, and adaptive control of the system's behavior.

S-AI-GPT expands this paradigm to conversational AI by fusing modular parsimony with biologically inspired regulatory mechanisms, most notably artificial hormone signaling. The system only engages the relevant areas required for a given task, simulating the sparse and context-dependent activation observed in the human brain. The reasoning process is therefore generally more efficient, adaptable, and explicable.

Recent advances in modular architectures [23], tool-augmented reasoning [18], and affective modulation [29], although significant, remain fragmented and lack the unified orchestration proposed in our work.

2. METHODS

2.1. Biological Foundations of S-AI-GPT

2.1.1. Introduction: A Living Architecture

S-AI-GPT goes one step further than classical modular AI by introducing a new kind of internal regulation — artificial hormonal signaling. Inspired by how living organisms respond to their environment, this mechanism allows the system to subtly adapt the tone, pacing, emotional depth, and style of its responses in a smooth and natural way.Unlike rule-based machines that simply follow instructions or activate components mechanically, hormonal signaling adds a soft, system-wide modulation layer. It doesn't give direct orders to agents. Instead, it gently shifts their behavior based on context, much like a biological organism reacting to mood, urgency, or focus. This allows S-AI-GPT to behave less like a rigid engine and more like an adaptive entity, capable of responding differently depending on how the user feels, what they need, and even how fast they expect an answer.

2.1.2. Biological Inspiration and Justification

In the human body, endocrine glands do not issue direct commands. Instead, they release hormones that influence how organs behave, depending on timing, concentration, and internal state. This kind of regulation is powerful because it is **flexible**, **robust**, and **adaptive**—three qualities that are essential for intelligence that needs to evolve over time. S-AI-GPT applies the same principle to artificial intelligence. It uses**artificial hormonal signals** to:

- Adjust agent behavior based on cognitive-emotional context (e.g., *urgency*, *empathy*)
- Introduce time-dependent modulation (such as **decay** or **reinforcement**)
- Enable broad, indirect influence over multiple agents simultaneously
- Avoid brittle, hard-coded decision chains by relying on dynamic hormonal levels

This architecture draws on ideas from **neuroscience-inspired AI**, where loosely coupled, bio-inspired components are preferred for their ability to **scale** and **adapt** [7].

2.1.3. GPT-Gland Agents: Emission Modules for Hormonal Signals

GPT-Gland Agents are specialized components responsible for producing and regulating artificial hormones within the system. Activated by the GPT-MetaAgent in response to contextual cues—such as user emotion, conversational flow, or task urgency—each gland embodies one or more hormonal profiles. These profiles enable the following functions:

- Emission of new hormonal signals into the GPT-HormonalEngine,
- Adjustment or resetting of existing hormone levels,
- System-wide modulation of multiple GPT-Specialized Agents.

This mechanism supports soft, indirect orchestration, ensuring that agent behaviors are adaptively tuned rather than rigidly commanded. It enhances the system's emotional sensitivity, continuity, and parsimony by allowing dynamic and non-invasive behavioral modulation.

2.1.4. Types of Signals and Hormonal Profiles

Each hormone is a named signal with a changing intensity between 0.0 and 1.0 that is used to change how agents act. Some of the characteristics are: the name of the signal (like *"urgency,"* *"empathy,"* or *"depth"*), how it fades over time (exponentially or linearly), how strong agents think it is, and the sensitivity threshold to ignore weak signals. This signaling mechanism, influenced by affective neuroscience, reflects the manner in which internal computational variables — such as intensity and decay — serve as intermediaries between mechanistic control and phenomenological states, as posited by Moutoussis & Dolan [11].

2.1.5. The Engine That Produces Hormones

The GPT-HormonalEngine manages the overall hormonal context of the system. It performs three main tasks: (i) tracks currently active hormonal signals, (ii) applies natural decay over time to simulate dissipation, and (iii) provides all GPT agents with access to a shared hormonal state. This operates as a transient, fuzzy memory layer that enables consistent system behavior without enforcing centralized control. Agents adapt their actions in harmony while preserving autonomy.

2.1.6. Hormonal Modulation in Action

Prompt Example – "Can you use a funny analogy to explain blockchain, but keep it short?"

Step-by-Step Execution

• **Decomposition Phase** :

  – *AnalogyAgent*: selects metaphor
  – *KnowledgeAgent*: provides core facts
  – *HumorAgent*: sets tone
  – *MinimalistAgent*: ensures brevity

- **Hormonal Injection** :

  – *HumorGland*: emits playfulness = 0.7
  – *StressGland*: emits urgency = 0.6

- **Hormonal Context State** :

  { "playfulness": 0.7, "urgency": 0.6 }

- **Agent Modulation** :

  – *HumorAgent*: adds playful tone
  – *MinimalistAgent*: favors shorter output
  – *AnalogyAgent*: picks easy-to-grasp metaphor

Final Output – "Blockchain is like a notebook shared by the whole class. No one can erase what's written, and everyone sees who adds what."

Interpretation – This example illustrates how hormone-driven coordination enables the system to adjust tone and brevity without rewriting the prompt. Modulation arises contextually and dynamically, echoing curiosity-driven activation in affective robotics [17]

2.1.7. Overview and Bio-Inspired Role of the Hormonal Layer

What makes S-AI-GPT's hormonal layer unique is that it borrows from biology — not from lines of code or rigid rule sets, but from how the human body regulates itself. Instead of issuing hard commands, it uses soft, delayed signals. It doesn't force agents to behave a certain way; it nudges them, influencing tone, pacing, cognitive focus, or emotional tone in ways that feel more intuitive than mechanical.This makes the system act less like a machine and more like an entity — one that adapts, reflects, and reacts subtly to its environment. You don't need to rewrite prompts or manually change settings: the modulation happens from within, invisibly but meaningfully.

S-AI-GPT achieves something rare here: it bridges three worlds that are usually kept apart:

- Symbolic planning,
- Neuro-symbolic orchestration,
- And emergent emotional intelligence.

The result is a flexible, modular architecture that doesn't just scale technically — it scales humanely. It aligns with how people think, feel, and change. It keeps resource consumption low while keeping transparency high. You can trace back decisions, inspect signal histories, and understand why the system did what it did.More than a technical upgrade, this hormonal layer sets the foundation for future AI systems that are emotionally aware, self-regulating, and fundamentally more compatible with how humans work.

2.2. General Architecture of the S-AI-GPT

The S-AI-GPT system is built on a modular architecture that draws inspiration from distributed multi-agent systems as well as human brain principles (specialization, hormone regulation, and contextual memory). Every system component is made to function as a specialized agent with a distinct role that is only activated when necessary. This guarantees traceability, adaptive behavior, and sparsity.

2.2.1. The GPT Model's Function in S-AI-GPT: A Supervised Generative Assistant Under the 20/80 Principle

The function of the internal memory engine developed in the second article must be distinguished from that of the GPT model. Contextual persistence, artificial hormone-based affective regulation, and adaptive activation of specialized agents are all functions of the memory engine, an independent cognitive subsystem. It functions as an internal cognitive core that is self-regulating and is based on activatable mini-structures that were influenced by biological engrams. The GPT model, on the other hand, lacks orchestration and memory. When symbolic agents reach their expressive limits, it provides free-form text generation. This is its sole linguistic purpose.

This design adheres to the fundamental 20/80 sparsity principle, which is essential to S-AI-GPT: lightweight, symbolic, or specialized agents can handle 80% of user queries. The GPT model should be activated because only 20% of tasks call for the creation of complex, flexible, or free-form language. In certain situations, if the symbolic layer is not enough, the GPT-MetaAgent may also call upon deep models other than GPT, such as speech, vision, or multimodal classifiers. This section, however, only addresses the GPT model's linguistic function within the system.

2.2.2. Central Orchestration and Specialized GPT Agent Activation

2.2.2.1. Central Orchestration by the GPT-MetaAgent

The *GPT-MetaAgent* acts as the central orchestrator. It supervises the activation of GPT-Specialized Agents, manages the global interaction context, adjusts hormonal profiles, and coordinates the final response. It makes decisions based on:

- The user's prompt and task decomposition,
- Hormonal signals and contextual stimuli,
- The user's profile, preferences, and interaction history.

This orchestration allows *S-AI-GPT* to dynamically adapt to cognitive load, conversational style, and emotional context.

2.2.2.2. Sparse Activation and GPT-Specialized Agents

GPT-Specialized Agents (SAs) are grouped into functional families: reasoning, memory, emotion, style, logic, etc. All agents inherit from a shared interface (BaseAgent), which allows for dynamic and uniform activation. Each agent is autonomous, executes a specific task, and then returns its output to the GPT-MetaAgent. This enables modular response construction while maintaining low computational costs. Inactive agents consume no resources, adhering to the "sparse activation" principle.

2.2.3. The Decomposition Agent

2.2.3.1. Reading Beyond the Prompt

Rather than rushing to generate a reply, the Decomposition Agent pauses. It tries to understand what the user really wants: Is there worry behind the words? A tone that seeks reassurance? A need for exactness? It breaks things down carefully—capturing unspoken intent, identifying what sort of cognitive work is needed, and spotting any practical constraints that may guide the answer.

2.2.3.2. Breaking Down with Finesse

This agent draws from different techniques to do its job. Sometimes, it's about recognizing a familiar pattern—like spotting that a question beginning with "What are the effects of…" is probably asking for causal insight. Other times, it leans on experience, thanks to models trained on real human examples, to grasp subtle intentions. It also uses a kind of internal compass: the 20/80 rule. It knows that not everything needs deep reasoning—and it saves its energy for what matters most. The process isn't rigid. If a prompt feels ambiguous or unusual, the agent doesn't guess. It seeks help—reaching into memory, or consulting another agent. That's what makes it flexible, and that's what gives the whole system its depth.

2.2.3.3. A Dialogue with the MetaAgent

Once the agent has mapped out the pieces, it hands them over—not to a black box, but to the GPT-MetaAgent, the one that decides what happens next. The map it provides includes more than just subtasks; it carries nuance: what kind of tone might suit the user, which agent might be best suited to each role, and even how urgent or delicate the situation is. This isn't a one-way exchange. If the MetaAgent senses that something's off—maybe the response is weak or the tone doesn't land—it can request changes. Together, these two agents form a loop, each adjusting to improve the whole.

2.2.4. A Network That Breathes Together

Each specialized agent in S-AI-GPT has its own voice, its own domain, its own rhythm. Some are analytical, others empathetic. Some organize, others remember. And all of them are designed to work side-by-side—not in isolation, but in coordination.

For example:

- MedicalAgent brings verified insight
- EmpathyAgent adjusts tone
- FormattingAgent structures outputs
- MemoryAgent ensures continuity across exchanges

These agents don't live in fixed roles. They appear, interact, and dissolve as needed. The orchestration is dynamic—just like conversation itself. Further exploration of how this coordination unfolds will be the focus of the second article.

2.2.5. Keeping the System Grounded

No matter how intelligent a system is, it needs to stay grounded. That's the role of the Security Agent. It watches—not to restrict, but to protect. It scans for anything unusual: strange patterns, repeated access attempts, behavior that doesn't match the flow. If something seems off, it acts. It raises alerts. It informs the GPT-MetaAgent. It can even impose temporary limits—closing access, isolating parts of the system—until things settle. But this agent doesn't just respond. It thinks ahead. It broadcasts warnings in the form of hormonal cues—signals like "vigilance" or "caution"—subtle shifts that ripple through the system, nudging every agent to adjust its tone, its precision, its behavior. Security here isn't a fence. It's more like an immune system—alert, adaptive, always learning.

2.2.6. Result Aggregator Agent : Combining Multiple Agents and Putting Them in Context

In a modular architecture like S-AI-GPT, where multiple specialized agents can work on a single query at the same time, the result aggregation phase is crucial. The Result Aggregator Agent is an independent system agent supervised by the GPT-MetaAgent.

Functional Role – Its main responsibilities include:

- Gathering outputs from the activated specialized agents responding to a specific subproblem;
- Checking the quality, coherence, and contextual relevance of each partial result;
- Combining or selecting these responses to produce a final output that is clear, consistent, and meaningful.

It acts as both an intelligent filter and a content synthesizer, capable of weighting, majority voting, semantic merging, or selecting a single-best answer based on memory or hormonal cues. Strategies for Aggregation – The Result Aggregator can adapt its aggregation strategy to fit the context or follow GPT-MetaAgent directives:

- Dynamic weighting: weights based on confidence, hormonal signals, or memory relevance;
- Majority or priority voting: preference for consensus or domain-prioritized agents;
- Single-best selection: when diversity would harm clarity;
- Symbolic/textual fusion: structured synthesis into summaries, lists, or tables.

Working with Other System Agents – The Aggregator is overseen by the GPT-MetaAgent, which can dynamically alter its strategy (e.g., prioritize conciseness or diversity). The final result is sent to the Display Agent or Result Access Agent, depending on the intended endpoint. Hormonal cues may also be triggered to inform future executions or signal inter-agent disagreements.

2.2.7. Hormonal Modulation and Gland Agents

The use of *GPT-Gland Agents* in *S-AI-GPT* creates a new biological metaphor. These agents change how the system works by releasing artificial hormones that spread out at different times and change the thresholds for activating agents. Some important parts are:

- Hormonal context profiles, which are based on emotional tone, urgency, or trust levels ;
- Selective activation or inhibition of agents based on what the context needs ;
- MetaAgent supervision, which controls hormone distribution by commanding gland agents without hard-coded logic.

This layer of soft coordination makes the system more reactive, less computationally expensive, and more likely to show new patterns of behavior.

2.2.8. Dynamic Contextual Memory (DCM): Working Memory That Changes Quickly and Is Controlled by Hormones

The *Dynamic Contextual Memory* (DCM) is *S-AI-GPT*'s main adaptive working memory. The DCM is a volatile and intelligent memory structure that changes in real time based on hormonal activity, emotional state, and orchestration decisions. This is different from static session memory or simple conversational buffers.

2.2.8.1. A Cognitive Filter That can be Changed in Size

The DCM acts as a smart buffer between how users see things, how agents carry out tasks, and how responses are made. It keeps only the most important parts of the conversation (intent, emotion, content) and changes their level of detail based on the hormones that are active. For instance:

- When stressed, the DCM cuts out unimportant parts ;
- When focused, it makes important new data points clearer.

2.2.8.2. A Regulated Structure, Not an Agent

The *Dynamic Contextual Memory* (DCM) is not regarded as an agent in the strict sense ; instead, it functions as a transversal cognitive module that sustains and facilitates adaptive short-term memory throughout the system. When contextual information is needed, both system agents (like the *MemoryAgent* or the *MetaAgent*) and domain-specific agents ask for it and update it.The DCM can be wrapped in a callable object that has agent-like methods (like process() and receive_trace()), which makes it easier to use. However, the *MetaAgent* does not control the DCM as an independent agent and it does not have its own lifecycle.

This difference helps keep the architecture clear between active agents and shared cognitive resources, while supporting gated, regulated information flow mechanisms inspired by early recurrent memory architectures [5].The DCM is not an independent agent ; it does not make choices on its own. There are three parts that control and shape it :

- the *Memory Gland*, which changes its content in real time ;
- the *Memory Agent*, which keeps an eye on its strategies for remembering or forgetting ;
- the *GPT-MetaAgent*, which uses its state for smart orchestration.

2.2.8.3. Cognitive Persistence and Emotional Consistency

The DCM lets *S-AI-GPT* :
• Keep a consistent conversation context over time ;
• Give answers that match the user's tone and emotional history ;
• Use controlled forgetting to keep from getting too much information.
The DCM is therefore very important for making memory management in the system context-aware, affect-sensitive, and computationally efficient, echoing the principles of adaptive reinforcement observed in neural agents trained through delayed reward mechanisms [9].

2.2.9. Memory Gland – Affect Modulation of Active Memory

The Memory Gland Agent is one of the most innovative parts of S-AI-GPT. It is a simulated gland that is based on biology and is used to change working memory based on emotions and context. The Memory Gland is not like traditional cognitive agents because it doesn't interpret content. Instead, it changes how the Dynamic Contextual Memory (DCM) works based on hormonal signals that are sent out when someone is feeling emotional or needs to think quickly.

2.2.9.1. Emotional Control over How Memories Are Made

When the hormonal engine sends a signal to the Memory Gland (like stress, focus, or fatigue), it changes the contents of the DCM in real time. For example, it removes peripheral details when you're stressed, amplifies important information when you're very focused, and shortens the

memory window when you're tired.This makes sure that only information that is useful in the context and emotionally relevant is kept, which helps both cognitive frugality and contextual alignment.

2.2.9.2. Feedback on Proactive Orchestration

In addition to reactive modulation, the Memory Gland can proactively suggest hormonal changes to the GPT-MetaAgent based on past emotions. For example, it might suggest raising oxytocin levels after stress has been detected several times. It works like an affective memory sensor, helping the system change how it acts based on hidden emotional states.

2.2.9.3. Working Together with the MetaAgent and the DCM

The Memory Gland, the Dynamic Contextual Memory, and the GPT-MetaAgent make up a regulatory triangle: the gland modulates, the MetaAgent orchestrates, and the DCM filters. This closed loop makes it possible to orchestrate emotions in a sensitive way and makes sure that the user stays aligned with the conversation in a way that is adaptive and coherent.

2.2.10. Knowledge Base Agent: Structured Knowledge Access and Inter-Agent Synergy

2.2.10.1. Introduction

In the *S-AI-GPT* architecture, the *Knowledge Base Agent* (KBA) is a core system agent responsible for managing structured, evolutive, and distributed knowledge. It serves as the semantic backbone of the system, ensuring that all agents operate on coherent and accessible conceptual grounds.

Unlike traditional static databases, the KBA functions as a dynamic intelligent agent with:

- Symbolic reasoning capabilities,
- Contextual adaptability, and
- Hormonal reactivity based on the current system state.

It is tightly coupled with memory and orchestration layers, enabling semantic enrichment, shared grounding, and real-time contextual knowledge access.

2.2.10.2. Functional Role

The KBA fulfills three core missions:

- **Knowledge Retrieval**: Answering queries from the Decomposition Agent, MetaAgent, and Specialized Agents by retrieving symbolic knowledge, factual assertions, or inference rules.
- **Knowledge Update**: Incorporating new symbolic statements, structured facts, or learned rules, which may be produced by agents during execution, orchestration, or learning phases.
- **Inter-agent Grounding**: Ensuring semantic alignment between agents relying on different conceptual schemas or terminologies, allowing coherent cooperation across specialized domains. These roles make the KBA a shared epistemic environment, maintaining a stable and intelligible knowledge layer for all interacting agents.

2.2.10.3.   Hormonal Modulation and Access Prioritization

The KBA is fully integrated into the hormonal signaling loop of *S-AI-GPT*. It receives modulatory inputs from:

- The *MetaAgent*, to shift focus based on strategic planning, system phase, or orchestration refinement ;
- *Gland Agents*, to bias or prioritize retrieval based on urgency, emotional tone, or uncertainty.
- These signals influence:
- The type of knowledge retrieved (e.g., heuristic vs. deep logical rule) ;
- The depth of inference allowed ;
- The confidence thresholds for symbolic reasoning.

Example: In a high-stress scenario, the KBA may prioritize fast, low-depth heuristics over complex inference chains. This enables adaptive semantic modulation, mirroring emotional prioritization in biological systems.

2.2.10.4. Integration with Other Agents

The *Knowledge Base Agent* interacts seamlessly with various components:

- The *Decomposition Agent* uses it to match semantic decomposition templates or domain rules.
- The *MetaAgent* queries the KBA for orchestration memories, agent-performance mappings, and symbolic planning templates.
- *Specialized Agents* use it for validation, enrichment, or correction of their outputs.
- The *Display Agent* accesses it to generate justifications or answer transparency-related queries (*Explainable AI*).
- The *Memory Agent* collaborates with the KBA to ensure temporal consistency and knowledge persistence across sessions.

The KBA thus acts as a semantic interoperability layer, harmonized through memory and hormonal signaling.

2.2.10.5. Internal Architecture

The KBA is built upon a hybrid and extensible architecture composed of:

- Symbolic knowledge graphs (*RDF/OWL/SPARQL*), enabling structured knowledge representation ;
- Rule bases (*Prolog-style* or logic-based), allowing forward/backward chaining ;
- Annotated factual stores, for storing raw and contextualized knowledge units ;
- A reasoning and query engine, supporting pattern matching and symbolic inference ;
- A temporal interface, synchronized with the *Memory Agent* and *Gland Agents* to maintain coherent system-wide knowledge evolution.

The architecture supports incremental updates, asynchronous rule injection, and cross-agent knowledge pushing.

2.2.10.6. Conclusion

The *Knowledge Base Agent* is much more than a passive storage component. It embodies a context-aware, hormonally-regulated semantic core that supports:

- Modular cooperation,
- Symbolic reasoning,
- Adaptive orchestration, and
- Long-term knowledge evolution.

It forms, along with memory and orchestration mechanisms, the triadic backbone of cognitive intelligence within the *S-AI-GPT* framework, offering scalability, explainability, and biological plausibility in multi-agent conversational AI.This modular design enables agent autonomy while maintaining contextual coherence — a limitation noted in earlier monolithic or planner-centric approaches [13], [19], [23].

3. RESULT AND DISCUSSION

3.1. Positioning S-AI-GPT in the Current Landscape

This section provides a thorough comparative review of existing approaches to firmly position the S-AI-GPT architecture within the current artificial intelligence landscape. We have intentionally centralized all pertinent contributions concerning modularity, orchestration, sparsity, memory, emotional regulation, and ethical supervision within a single cohesive framework, in contrast to numerous studies that emphasize isolated comparisons. This editorial choice shows how S-AI-GPT is cross-disciplinary. It doesn't just suggest a small improvement; it also tries to bring together and combine a group of problems that have usually been dealt with separately in the literature. Complementary articles that talk about technical implementation and real-world use cases will refer back to this basic analysis without repeating everything in it.

3.2. Related Work

3.2.1. AutoGPT – A Language Model that Tries to Think for Itself

When it was released in 2023, AutoGPT surprised many [14]. Built on top of GPT-4, it demonstrated that a language model could go beyond reactive prompting. It was capable of setting its own goals, breaking them into subtasks, and executing them iteratively, as if attempting to reason autonomously. Given a broad instruction—such as "book a flight"—the system would initiate a self-directed loop: generating subgoals, calling external tools, and adapting its plan along the way, all with minimal human oversight, illustrating a first attempt at automated agent generation later formalized in frameworks such as AutoAgents [4].This looped autonomy marked a conceptual leap. But it came with significant trade-offs. AutoGPT's planning remains largely stochastic and fragile. Its memory is shallow and forgetful. Agents it spawns are ephemeral—without continuity, identity, or shared context. There's no central oversight, no system-wide reasoning, and no internal coordination. The result is often chaotic: actions repeat, diverge, or stall in loops with no clear way out. This apparent autonomy often results in disoriented or incoherent behavior.

Comparison with S-AI-GPT

S-AI-GPT takes a different path altogether. Instead of pushing one monolithic model to do everything, it constructs a structured ecosystem of persistent, specialized agents, each with a distinct role and purpose, coordinated by a central GPT-MetaAgent. The entire system is inspired by biological principles—particularly hormonal signaling and modular regulation—resulting in an architecture that is not only adaptive but traceable, explainable, and efficient.Centralized orchestration: The GPT-MetaAgent acts like a conductor, evaluating the user's intent, the contextual state, and internal "signals" to activate only the relevant agents. This coordination replaces the chaotic loops of AutoGPT with purpose-driven delegation.

Semantic decomposition: A dedicated Decomposition Agent breaks down complex queries using symbolic and neuro-symbolic heuristics, including the 20/80 rule.

Hormonal modulation: A Hormonal Engine and Gland Agents simulate urgency, fatigue, or attention, influencing agent behavior dynamically.

Dual-layered memory: A long-term Memory Agent and a Dynamic Contextual Memory ensure session continuity and adaptive recall.

Security and supervision: A Security Agent monitors behavior and enforces ethical boundaries, unlike AutoGPT's open loop.

Specialized modularity: Only necessary domain-specific agents are activated, optimizing reasoning and resource use.

Transparency and efficiency: Sparse activation with full traceability and explainability, eliminating the black-box effect.

3.2.2. Toolformer – Self-Taught Tool-Augmented Language Models

Toolformer [16], introduced by Schick et al. (2023), is a big step forward in letting language models use external tools by themselves. Instead of being fine-tuned on data labeled by humans, Toolformer adds its own API calls to the training data. These API calls—like using a calculator, a search engine, a QA system, or a translation tool—are learned by the model during training. The idea is simple : if adding a tool call reduces token prediction error, then the model keeps it. This helps the model learn when and how to use tools, without outside help. Toolformer becomes smarter, without getting bigger or needing expensive prompt tuning. It works especially well on zero-shot tasks like arithmetic and factual lookup, even beating bigger models that don't use tools.

Comparison with S-AI-GPT

Toolformer and S-AI-GPT both embrace modularity, but through different mechanisms:

Modularity: Toolformer sees tools as external APIs it can call when needed. S-AI-GPT builds internal agents like MathAgent and TextAnalysisAgent, each with their own memory, logic, and context. These agents aren't always running; they're activated when useful.

Orchestration and Adaptivity: Toolformer just uses the tool calls it discovered during training. It doesn't track what the tools are doing while they run. S-AI-GPT uses a GPT-MetaAgent that

decides in real time which agent to activate, based on the task, user profile, and hormone signals. This makes it more flexible and easier to follow.

Task Decomposition: Toolformer handles everything inside a single model, choosing tools one token at a time. S-AI-GPT separates the planning from the execution. A Decomposition Agent figures out subtasks, which makes the process more modular and parallel.

Memory and Context: Toolformer relies on short context windows. S-AI-GPT includes a Memory Agent and a Dynamic Contextual Memory (DCM), so it can remember past preferences and stay consistent over time.

Interpretability and Control: Toolformer adds API calls into the token stream, but doesn't manage them explicitly. S-AI-GPT tracks everything—agent activations, hormone signals, decisions—so users can understand what happened and why.

Security and Robustness: Toolformer doesn't check if the tools are being used safely. S-AI-GPT includes a Security Agent that makes sure no bad decisions are made, and that the system stays within safe boundaries.

Early experiments on sparse gating mechanisms, such as Mixture-of-Experts (MoE) architectures, laid the groundwork for scalable activation control, although they lacked explicit symbolic coordination [18].

3.2.3. HuggingGPT – Model-Orchestrated Multimodal Reasoning with External Expert Systems

3.2.3.1. Framework Overview and Operational Pipeline

HuggingGPT [18], introduced by Shen et al. (2023), presents a novel orchestration-centric framework that leverages a large language model (LLM)—specifically ChatGPT—as a central planner to coordinate the use of diverse specialized AI models hosted within the Hugging Face ecosystem. Rather than solving user queries internally, the LLM assumes the role of a task planner and system orchestrator, responsible for decomposing complex instructions, selecting external models, delegating execution via APIs, and integrating the results into coherent outputs.

The system operates through a four-stage processing pipeline:

- *Task Planning* – The LLM parses the user's input, infers intent, and decomposes the query into elementary subtasks.
- *Model Selection* – It identifies the most suitable expert models (e.g., for vision, speech, translation) from Hugging Face's model repository.
- *Task Execution* – It invokes these models via standardized API calls to process each subcomponent.
- *Response Generation* – It aggregates intermediate results into a unified, contextually relevant response.

Comparative Evaluation with S-AI-GPT: While both HuggingGPT and S-AI-GPT adopt a modular approach to task resolution through delegation, they diverge across critical architectural, operational, and cognitive dimensions:

- *Architectural Modularity*: HuggingGPT operates via external modularity, outsourcing task execution to third-party expert models accessed through API interfaces. In contrast,

S-AI-GPT is based on internal modularity, embedding domain-specific agents (e.g., ImageAgent, SpeechAgent) directly within the system's architecture. These agents are orchestrated by the GPT-MetaAgent, allowing shared memory, hormonal influence, and tight integration of agent state and context.

- *Task Structuring Paradigm*: HuggingGPT employs an LLM-based planner to parse and organize user intent. S-AI-GPT introduces a dedicated Decomposition Agent, structurally decoupled from the main orchestrator, enabling reusable, explainable, and domain-aware subtask formalization.
- *Memory and Contextual Persistence*: HuggingGPT does not maintain a persistent memory trace across sessions. By contrast, S-AI-GPT integrates a Memory Agent alongside a Dynamic Contextual Memory (DCM), supporting incremental personalization, temporal coherence, and adaptive context reconstruction.
- *Adaptivity and Hormonal Modulation*: In HuggingGPT, once external models are selected, execution is static and reactive. In S-AI-GPT, the behavior of internal agents is modulated in real time by a Hormonal Engine and Gland Agents, allowing dynamic adaptation to emotional tone, ambiguity, stress signals, or task complexity.
- *Security and Ethical Control*: HuggingGPT lacks an internal mechanism for runtime verification or behavior filtering. S-AI-GPT embeds a Security Agent that enforces safety policies, detects anomalies, and prevents potentially harmful or unethical outcomes during execution.

3.3. Analytical Discussion

3.3.1. What the Hormonal Signaling Layer does

Artificial hormones add a new, fuzzy layer of rules to S-AI-GPT. They give :

- Adaptive tone and tempo without prompt engineering ;
- Asynchronous, soft modulation of behavior ;
- Indirect impact over agent dynamics ;
- Global conversational continuity (for example, persistent mood) ;
- Dynamic cost optimization (for example, suppressing deep agents when not needed).

3.3.2. Bio-Inspired Emotional Regulation and Its Comparison with Conventional Approaches

Along with the benefits of the hormone layer, it is helpful to compare the S-AI-GPT method to more traditional models of emotional intelligence in AI. Most traditional ways to measure emotional intelligence (EI) focus on being able to perceive and hear how others feel (for example, through face recognition, audio analysis, and semantic processing) and employ preprogrammed reactions to make communication between people and computers better. These systems are often built as separate functional modules that are connected to a core architecture from the outside. They don't have deep integration with memory, contextual dynamics, or computing efficiency.

S-AI-GPT, on the other hand, provides an approach to govern emotions that is highly integrated, modular, and based on biology. This is based on:

- A Memory Gland Agent, which changes active memory based on hormonal signals (like stress, attention, and cognitive load);

- An artificial hormonal signaling mechanism, which mimics slow and diffuse diffusion like the biological endocrine system, allowing for smooth and continuous changes in agent behavior;
- A distributed hormonal orchestration, in which decision-making, memory processes, and agent activation are influenced by an evolving hormonal profile, without relying on explicit emotion recognition.

This architecture provides emotional regulation a built-in, flexible, and cost-effective way to control emotions that is closely linked to how the multi-agent system functions, rather than just an added feature. This layer of design includes elements for emotive modulation that are akin to early notions in affective computing [12], which said that robots need emotion-like mechanisms to be flexible and aware of their surroundings.

3.3.3. Comparative Positioning with Modular and Multi-Agent Architectures

Before contrasting S-AI-GPT with monolithic or hybrid models, it is important to analyze its positioning among modular and agent-based AI architectures.

Several comparison axes help structure this evaluation :

- **Level of orchestration** : centralized (orchestrated by a master agent) versus emergent (based on local agent interactions) ;
- **Agent autonomy**: rigid pipelines with predefined flows versus dynamically instantiated agents based on context ;
- **Context integration**: rule-based triggers versus biologically inspired signaling mechanisms (e.g., hormonal modulation) ;
- **Feedback capabilities**: static systems versus reflexive architectures that adapt through feedback loops.

S-AI-GPT innovates by extending existing multi-agent paradigms, through the introduction of:

- A **semantic decomposition pipeline** decoupled from the generative core ;
- A **hormonal regulation layer** for soft, asynchronous behavior modulation ;
- A **self-regulating orchestration core** (MetaAgent) capable of selecting and coordinating agents contextually.

This approach builds upon foundational works on complexity reduction using distributed representations and latent abstractions [8], which showed that deep, layered, and low-dimensional architectures enable more scalable, modular, and interpretable systems.

3.3.4. Typology of Conversational Architectures and Positioning of S-AI-GPT

3.3.4.1. Monolithic LLM Architectures

Monolithic architectures rely on a single, very large autoregressive transformer model that handles all cognitive functions: understanding, reasoning, generation, and working memory. Examples include ChatGPT (OpenAI), Claude (Anthropic), and Gemini (Google DeepMind) [20].

While these systems perform well on general conversational tasks, their architecture suffers from several limitations :

- No internal modularity ;
- Low explainability of decisions due to full model activation ;
- High computational cost with no selective activation ;
- Limited contextual or personalized adaptation.

They operate as textual black boxes, generating answers based on prompt history without an interpretable decision flow [20] [2].

3.3.4.2. Hybrid Architectures (LLM + Tools or API)

Hybrid architectures aim to address monolithic rigidity by combining a central LLM with external tools, API calls, symbolic rules, or plug-ins. Examples include Copilot (Microsoft), Google Assistant with Gemini, and dynamic interaction patterns like ReAct or Toolformer [16]. These models offer limited modularity, with the LLM controlling tool invocation.While this improves task automation, orchestration remains centralized, and modules lack autonomy or context-aware activation. More recent proposals, such as HuggingGPT [19], extend this paradigm by coordinating specialized APIs through a GPT controller but still rely on monolithic core planning. Technical explanations of Mixture-of-Experts (MoE) mechanisms also fall into this category when they are controlled by a central model rather than a distributed agent-based strategy [16], [19],[23].

3.3.4.3. Modular Agent-Based Architectures

Several recent architectures have explored agent-based approaches, where distinct specialized modules collaborate to accomplish complex tasks. Notable examples include:

- AutoGPT [14], which dynamically spawns agents to address evolving subgoals in a recursive task loop ;
- BabyAGI, which simulates a lightweight planning loop with limited memory persistence ;
- And more general multi-agent collaboration frameworks, as discussed in [22], which distribute subtasks among cooperating agents, sometimes augmented with memory or explicit planning mechanisms.

While promising in principle, these systems often suffer from several structural limitations:

- Lack of robust orchestration: coordination is typically emergent or loosely defined, relying on dialogue among agents rather than a centralized strategy ;
- Cognitive fragility: persistence across tasks is weak, making long-term coherence difficult to sustain ;
- Limited adaptivity: few systems integrate real-time behavioral modulation, and most depend on fixed heuristics or stochastic planning loops.

In real-world, dynamic environments—especially in conversational settings—these limitations often lead to degraded performance, insufficient adaptability, and weak explainability [14], [22].

3.3.4.4. Unique Positioning of S-AI-GPT

S-AI-GPT introduces a fundamentally new paradigm, distinct from traditional modular or agent-based architectures, through its bio-inspired and parsimony-driven design philosophy. Rooted in the Sparse Artificial Intelligence (S-AI) framework [21] initially proposed by Said Slaoui, S-AI-GPT extends this vision to conversational intelligence. Its distinguishing components include:

- A **dedicated Decomposition Agent**, enabling semantic segmentation of complex user inputs into manageable subtasks ;
- A **GPT-MetaAgent**, acting as a centralized orchestrator with full traceability and adaptive control ;
- A suite of **Specialized GPT Agents**, each focused on a specific domain (e.g., medical, legal, emotional) ;
- An **artificial hormonal signaling system**, inspired by endocrine regulation, for smooth and context-sensitive agent activation ;
- A network of **Gland Agents**, modulating task execution based on emotional, temporal, or cognitive states ;
- And an integrated **memory infrastructure** combining long-term memory and real-time contextual adaptation.

Unlike conventional modular AI frameworks which primarily compartmentalize model capabilities, S-AI-GPT embeds dynamic orchestration into the very fabric of agent interactions through hormonal modulation.This leads to a parsimonious, explainable, and scalable architecture, optimized for human-centric dialogues and sustainable AI operation. Rather than being a simplified version of a GPT model, S-AI-GPT embodies a conceptual transformation— from monolithic prediction engines to adaptive, orchestrated cognitive ecosystems [6], [15], [21].

3.3.5. Global Orchestration and Feedback Loops

3.3.5.1. Introduction

The S-AI-GPT architecture relies on central orchestration handled by the MetaAgent, enhanced by distributed feedback mechanisms involving memory, gland agents, hormonal signals, and aggregated results. This section describes how all agents interact through a continuous cycle of perception – decision – modulation – learning – adaptation.

3.3.5.2. Role of the MetaAgent in Global Orchestration

The GPT-MetaAgent serves as the main conductor of the system. It manages the selection and activation of specialized agents based on the task, modulation via gland agents, aggregation of results via the Aggregator Agent, and synchronization with memory components (via Memory Agents). It functions as a strategic supervisor, capable of interrupting or redirecting the task depending on user input, emotional context, or memory state.

3.3.5.3. Internal Feedback Loop

Several internal feedback loops underpin the system's adaptability:

- Hormonal Feedback: hormones emitted by Gland Agents modulate agent priorities, thresholds, and emotional tone.
- Memory Feedback: modules like DCM, Memory Agent, and Memory Gland adjust outputs based on prior dialog history and context.
- Cognitive Feedback: post-aggregation, a feedback signal is sent to the MetaAgent to refine orchestration strategies for future iterations.
- User Feedback: implicit or explicit user reactions (e.g., corrections, emotional tone) are encoded into memory or hormones.

3.3.5.4. Cascade Modulation and Multi-Layer Interaction

The responses generated by S-AI-GPT do not follow a traditional linear flow, but rather a non-sequential modulated cascade involving multiple loops and adaptive layers:

- The **Decomposition Agent** segments the task into subproblems;
- The **GPT-MetaAgent** dynamically activates the relevant specialized agents;
- The **Gland Agents** modulate internal dynamics through hormonal signals;
- The **Result Aggregator** merges the partial outputs;
- The **Display Agent** adapts the format and presentation style;
- The **memory system** and **Knowledge Base Agent (KBA)** are updated asynchronously;
- The **GPT-MetaAgent** adjusts its strategies based on observed outcomes.

This cycle constitutes a **reflexive, multi-loop architecture** that far surpasses the rigid and sequential pipelines of traditional LLMs.

Temporal and Hormonal Synchronization

A fundamental innovation of S-AI-GPT lies in its **multi-level synchronization mechanisms**, including:

- **Temporal synchronization**: agents share a phase marker (initiation, execution, feedback);
- **Hormonal synchronization**: hormones circulate in two distinct cycles — fast (reactive) and slow (affective);
- **Strategic synchronization**: agent goals, priorities, and preferences evolve dynamically based on context and memory.

3.3.5.5. Output Management : Display and RAM Agents

At the end of the orchestration process, two agents play a crucial role in the controlled and ethical delivery of results :

- The Display Agent is responsible for the **stylistic and structured presentation** of the final responses. It adjusts the form, tone, and visual layout based on the user profile (e.g., list format, bullet points, empathetic or technical tone).
- The Result Access Agent manages the external exposure of results. It ensures:

  - Traceability of responses;
  - Ethical filtering (e.g., medical or legal disclaimers);
  - Alignment with user access rights or system constraints.

It may hide, delay, or dynamically contextualize parts of the output, relying on memory or hormonal signals. Together, these two agents close the system loop, ensuring that the delivery of content is intelligible, responsible, and contextually appropriate.

4. CONCLUSION AND PERSPECTIVES

4.1. Paradigm Shift Toward Modular, Adaptive, snd Interpretable AI

The ideas in this first article set the stage for S-AI-GPT to take a number of different strategic development paths. In the near future, the system could become a cognitive companion that can

change based on how each user feels, what they want, and how they talk. This vision depends on the gradual addition of user feedback loops, the ability to change activation profiles on the fly, and the ability for orchestration and real-world use to evolve together. S-AI-GPT's modular design also makes it good for use in embedded environments (edge computing) because it is lightweight and can be turned on and off as needed. This makes it possible to use smart home systems, medical assistants on board, and adaptive interfaces for self-driving cars in the real world.

4.2. Future Directions

A natural evolution of the system will also include the dynamic creation of specialized agents that can grow the ecosystem in response to new needs without having to retrain the whole model. Lastly, a major strategic goal is to build a dedicated internal generative engine that is specifically made to meet S-AI-GPT's language needs. This part, which is light and easy to control, would make the system fully autonomous, easier to understand, and more compatible with the 20/80 parsimony principle that underlies the architecture. This article outlines the main architectural framework of S-AI-GPT, which includes modular orchestration, semantic decomposition, hormonal signaling, and multi-agent coordination. However, it only introduces a few important parts in a general way or at a high level.

4.3. Roadmap for Upcoming Articles

To ensure clarity and continuity, the second article will focus extensively on the internal mechanisms and adaptive logic of key components. It will explore in depth:

- The Decomposition Agent, beyond its orchestration role, including its semantic parsing capabilities, rule-based adaptability, and dynamic subproblem granularity management;
- The structure, taxonomy, and learning strategies of GPT Specialized Agents, encompassing both business-oriented and domain-specific agents built on mini-neural architectures;
- The GPT Gland Agents, which operate under an endocrine-inspired framework of contextual hormonal profiles and adaptive regulation loops;
- And above all, the entire memory architecture, including the Memory Agent, the Memory Gland, and the Dynamic Contextual Memory (DCM)—all of which are essential to personalization, learning, and cognitive persistence.

This article will demonstrate how the interplay between hormonal signaling and memory dynamics fosters a coherent, adaptive, and emotionally responsive conversational system. These developments are the core focus of Article II, which emphasizes functional autonomy, emotional plasticity, and long-term evolutionwithin S-AI-GPT.At the same time, the third article will provide a comprehensive overview of implementation strategies, evaluation procedures, and deployment scenarios in real-world contexts. It will consolidate:

- Detailed code structures and modular implementation patterns,
- Experimental test cases validating performance and scalability,
- Deployment strategies aligned with user profiles, system constraints, and ethical considerations.

Together, these three articles establish S-AI-GPT as a reference framework for designing modular, resource-efficient, explainable, customizable, and durable conversational AI—aligned with human expectations, technical limitations, and interpretability standards.

REFERENCES

[1] J. Andreas, M. Rohrbach, T. Darrell, and D. Klein, "Neural module networks," *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, vol. 2016, pp. 39–48, 2016, doi: https://doi.org/10.1109/CVPR.2016.13.

[2] T. B. Brown et al., "Language models are few-shot learners," *arXiv preprint*, arXiv:2005.14165, 2020. [Online]. Available: https://arxiv.org/abs/2005.14165.

[3] L. Cañamero and J. Fredslund, "I show you how I like you – Can you read it in my face?" *IEEE Trans. Syst., Man, Cybern. A*, vol. 31, no. 5, pp. 454–459, 2001, doi: https://doi.org/10.1109/3468.952719.

[4] 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)*, 2024, doi: https://doi.org/10.24963/ijcai.2024/3.

[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, doi: https://doi.org/10.1162/089976600300015840.

[6] 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. [Online]. Available: https://proceedings.neurips.cc/paper/2021/hash/73a427b34802887d4d06cb69a7b09e92.

[7] D. Hassabis, D. Kumaran, C. Summerfield, and M. Botvinick, "Neuroscience-inspired artificial intelligence," *Neuron*, vol. 95, no. 2, pp. 245–258, 2017, doi: https://doi.org/10.1016/j.neuron.2017.06.011.

[8] G. E. Hinton and R. R. Salakhutdinov, "Reducing the dimensionality of data with neural networks," *Science*, vol. 313, no. 5786, pp. 504–507, 2006, doi: https://doi.org/10.1126/science.1127647.

[9] V. Mnih et al., "Human-level control through deep reinforcement learning," *Nature*, vol. 518, no. 7540, pp. 529–533, 2015, doi: https://doi.org/10.1038/nature14236.

[10] G. Montero Albacete, A. López, and A. García-Serrano, "Fattybot: Hormonal chatbot," *Information*, vol. 15, no. 8, p. 457, 2024, doi: https://doi.org/10.3390/info15080457.

[11] M. Moutoussis and R. J. Dolan, "How computation connects affect," *Trends Cogn. Sci.*, vol. 19, no. 4, pp. 157–163, 2015, doi: https://doi.org/10.1016/j.tics.2015.01.002.

[12] R. W. Picard, *Affective Computing*, MIT Press, 1997. [Online]. Available: https://affect.media.mit.edu/pdfs/97.picard.pdf.

[13] C. Qu, S. Zhang, Y. Li, and J. Ma, "Tool learning with LLMs: A survey," *arXiv preprint*, arXiv:2405.17935, 2024. [Online]. Available: https://arxiv.org/abs/2405.17935.

[14] T. B. Richards, "AutoGPT [Computer software]," GitHub, 2023. [Online]. Available: https://github.com/Significant-Gravitas/AutoGPT.

[15] C. Rosenbaum, T. Klinger, and M. Riemer, "Routing networks for multi-task learning," in *Int. Conf. Learn. Representations (ICLR)*, 2019. [Online]. Available: https://openreview.net/forum?id=ry8dv3R9YQ.

[16] T. Schick and H. Schütze, "Toolformer: Language models can teach themselves to use tools," *arXiv preprint*, arXiv:2302.04761, 2023. [Online]. Available: https://arxiv.org/abs/2302.04761.

[17] J. Schmidhuber, "Curiosity and boredom in neural controllers," in *Proc. Int. Conf. Simulation of Adaptive Behavior*, pp. 424–429, 1991. [Online]. Available: https://link.springer.com/chapter/10.1007/978-1-4471-1990-4_38.

[18] N. Shazeer et al., "Outrageously large neural networks: The sparsely-gated mixture-of-experts layer," *arXiv preprint*, arXiv:1701.06538, 2017. [Online]. Available: https://arxiv.org/abs/1701.06538.

[19] Y. Shen, K. Zhang, Y. Wang, and X. Liu, "HuggingGPT: Solving AI tasks with ChatGPT and its friends," *arXiv preprint*, arXiv:2303.17580, 2023. [Online]. Available: https://arxiv.org/abs/2303.17580.

[20] 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.

[21] S. Slaoui, "S-AI: Sparse Artificial Intelligence System with MetaAgent," *Int. J. Fundam. Mod. Res. (IJFMR)*, vol. 1, no. 2, pp. 1–18, 2025. [Online]. Available: https://www.ijfmr.com/papers/2025/2/42035.pdf.

[22] Y. Talebirad and A. Nadiri, "Multi-agent collaboration: Harnessing LLM agents," *arXiv preprint*, arXiv:2306.03314, 2023. [Online]. Available: https://arxiv.org/abs/2306.03314.

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

[24] H. Vicci, "Emotional intelligence in AI: Review and evaluation," *SSRN Working Paper*, 2024, doi: https://doi.org/10.2139/ssrn.4818285.

AUTHORS

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.