

# S-AI-ANTI HALLUCINATION: A BIO-INSPIRED AND CONFIDENCE-AWARE SPARSE AI FRAMEWORK FOR RELIABLE GENERATIVE SYSTEMS

Said Slaoui

Mohammed V University, Rabat, Morocco

## ABSTRACT

*Large Language Models (LLMs) exhibit impressive generative capabilities but remain prone to hallucinations — plausible yet false statements produced with high confidence. Such phenomena undermine trust and reliability in sensitive domains including health, law, and cybersecurity. Despite significant progress in retrieval-augmented generation (RAG), calibration methods, and mixture-of-experts (MoE) architectures, existing systems still lack a unified framework for veridiction, abstention, and energy-aware reasoning. This work introduces S-AI Against Hallucinations, a bio-inspired and parsimonious architecture derived from the Sparse Artificial Intelligence (S-AI) framework. The proposed model implements a symbolic-hormonal orchestration mechanism that enables generative agents to detect uncertainty, abstain when appropriate, and maintain citation integrity under ambiguous or adversarial conditions. The system relies on four hormonal variables — Hallucination Uncertainty (HU), Citation Integrity (CI), Contradiction Observer (CO), and Retrieval Evidence (RE) — dynamically regulated by a MetaAgent through hysteresis-based thresholds. Experiments performed on diverse scenarios, including factual question answering, scientific summarization, numerical reasoning, and out-of-distribution prompts, demonstrate stable abstention behavior, consistent citation tracking, and adaptive evidence retrieval. Evaluation follows a transparent and reproducible protocol inspired by PRISMA standards and Scopus-indexed benchmarking practices. S-AI Against Hallucinations provides a coherent, confidence-aware foundation for explainable and resource-efficient generative intelligence. It establishes a conceptual and operational bridge between statistical learning, symbolic reasoning, and biological homeostasis, paving the way for reliable and ethically governed AI systems.*

## KEYWORDS

*Sparse Artificial Intelligence (S-AI), Hormonal MetaAgent, Symbolic-Hormonal Orchestration, Hallucination Mitigation, Confidence-Aware Reasoning, Explainable and Frugal AI, Triadic Memory System, Governed Parsimony, Retrieval-Augmented Veridiction.*

## 1. INTRODUCTION

### 1.1. Context and Motivation

Large Language Models (LLMs) have reached extraordinary levels of fluency and versatility, yet they continue to produce *hallucinations*—statements that sound plausible but are factually false or unsupported. These errors threaten trust, accountability, and safety in sensitive domains such as medicine, law, cybersecurity, and education. The root causes are well known: optimization for next-token prediction rather than factual veridiction, missing provenance links, adversarial prompt manipulation, and limited ability to abstain when evidence is insufficient. Despite advances in retrieval-augmented generation (RAG), mixture-of-experts (MoE), and multi-agent systems, no unified framework currently combines veridiction, abstention, explainability, and

energy awareness. The field thus lacks an integrated mechanism to decide *when to respond* and *when to abstain* based on traceable confidence signals.

## 1.2. Problem and Research Question

The central challenge addressed in this paper is **how to enable an intelligent system to decide between answering and abstaining, in a transparent and resource-aware way**. This decision requires a meta-cognitive control layer that evaluates internal uncertainty, evidence quality, and logical consistency before producing any output. Our guiding question is: *Can a sparse, bio-inspired architecture governed by hormonal dynamics achieve reliable and frugal hallucination mitigation?*

## 1.3. Motivation for a Bio-Inspired and Parsimonious Design

In biological systems, stability arises from hormonal homeostasis—continuous feedback loops balancing activation and inhibition. Inspired by this principle, Sparse Artificial Intelligence (S-AI) introduces *cognitive sparsity*: the minimal activation of agents and computations needed to accomplish a veridical task. Unlike statistical sparsity (as in MoE), cognitive sparsity is governed, explainable, and resource-aware. It allocates computation according to risk rather than token count, making reliability and energy sobriety co-evolutionary goals.

## 1.4. Contributions of this Paper

This paper introduces **S-AI-AntiHallucination**, a bio-inspired, confidence-aware framework that integrates symbolic reasoning, hormonal regulation, and traceable abstention. The key contributions are:

- **(C1)** A modular architecture orchestrated by a *MetaAgent* and a dedicated *Hormonal Gland*, coordinating specialized agents for verification, citation, abstention, and retrieval.
- **(C2)** A *Triadic Memory System*—composed of a Dynamic Contextual Memory, a Memory Agent, and a Memory Gland—that encodes claims, verdicts, and hormonal traces for long-term learning.
- **(C3)** A formal yet lightweight hormonal mechanism with variables for Hallucination Uncertainty, Citation Integrity, Contradiction Observation, and Retrieval Evidence, enabling hysteresis-based decision policies.
- **(C4)** Resilience against semantic attacks such as prompt injection and evidence poisoning, achieved through hormonal self-regulation and provenance governance.
- **(C5)** A reproducible *Scopus-ready* evaluation protocol covering hallucination rate, justified abstention, calibration, robustness, and energy frugality.

## 1.5. Method Overview and Experimental Scope

The proposed system implements an agentic pipeline where the MetaAgent dynamically activates or inhibits specialized modules based on hormonal feedback. The Anti-Hallucination Agent (AHA) operates as a delegated subsystem, performing factual verification and abstention formatting when uncertainty or contradiction arises. Experiments encompass factual question answering, scientific summarization, numerical reasoning, and out-of-distribution prompts. Evaluation metrics include Hallucination Rate (HRate), Abstention Rate (ARate), calibration metrics (ECE, Brier), Mean Time-to-Abstain (MTTA), and energy consumption per decision.

## 1.6. Outlook and Continuation

This paper focuses on the **binary model**—a controlled decision between *Respond* and *Abstain*. A forthcoming companion study, *Triadic S-AI-AntiHallucination: Hormonal Clarification and Metacognitive Stabilization in Generative Reasoning*, extends the model to a **triadic regime** integrating a third cognitive action: *Clarification*. This extension introduces a metacognitive dialogue mechanism where the system can request or generate clarifications before deciding, further enhancing transparency, traceability, and trust.

## 2. RELATED WORK

### 2.1. Landscape in 2025: LLMs, MoE, Multi-Agent, Neuro-Symbolic

Contemporary approaches to improve veridiction and robustness of generative systems cluster into six families [1]–[3].

- **Monolithic LLMs**: large autoregressive models optimized for textual likelihood. Strengths: zero-/few-shot performance, task coverage. Limits: imperfect calibration, no abstention by default, high energy cost, limited traceability [4], [5].
- **MoE (Mixture of Experts)**: statistical sparsity via router-selected experts per token. Pros: throughput and scalability. Cons: routing instabilities, weak explainability of expert choice, training/ops complexity.
- **Multi-agent “agentic AI”**: prompt-driven orchestration of roles (planner, solver, verifier, retriever). Pros: rapid modularity. Cons: prompt fragility, governance debt (traceability/control), error propagation across agents.
- **Neuro-symbolic**: learned models coupled with logic/knowledge. Pros: explainability, constraints. Cons: partial integration, engineering cost, representational mismatch.
- **Hybrid modular systems**: practical pipelines (RAG, verification, tools) glued together. Pros: pragmatism, tooling. Cons: no global regulator (what to activate, when, at what cost), poorly optimized compute budget.
- **Memory-augmented AI**: explicit memory and RAG. Pros: retention, personalization, citations. Cons: low frugality, provenance governance gaps, scarce reasoned abstention.

In short, recent advances raise average performance but reliable decision-making (respond vs abstain), traceability, and energy sobriety remain insufficiently unified [6]–[8].

### 2.2. MoE: Statistical Sparsity without Explainability

MoE implements sparsity in compute by routing to a subset of experts per token. This is not cognitive sparsity: the system maximizes local likelihood absent a protocol for evidence or abstention [9]. Known issues: - Explainability: router decisions are hard to audit; no justification or confidence policy [10]. - Stability: routing variance, load imbalance, OOD sensitivity. - **Cost**: complex training and expert management [11]. Against hallucination, MoE reduces average cost but lacks normative regulation (mandatory citation, verification, abstention).

### 2.3. Agentic AI: Prompt-Based Orchestration, Fragility and Governance Debt

Prompt-orchestrated multi-agent stacks enable fast modularity (planner → retriever → solver → verifier) but remain fragile without principled control [12]. - Prompt fragility: small variations change orchestration drastically. - Error propagation: unchecked outputs from one agent

contaminate the chain. - Governance: difficult to enforce global rules (e.g., “every factual claim must be cited”) or to audit decisions. - Security: increased surface for semantic attacks (prompt injection, evidence poisoning) and privilege escalation [13], [14]. Without calibrated abstention and a global budget, guarantees remain weak.

## 2.4. Neuro-Symbolic: Explainability, Partial Integration, Engineering Cost

Neuro-symbolic systems pair learning with rules/logic/graphs, improving interpretability and constraint enforcement [15]. - Strengths: explainable inferences; enforceable invariants (logical coherence, domain constraints). - Limits: partial integration across modules, heterogeneous pipelines, higher latency [16], [17]. For hallucination, they better ensure coherence but do not alone solve parsimonious allocation or contextual abstention.

Logical entailment can be explicitly formalized as:

$$\text{Hallucination}(y|E) \Leftrightarrow \exists c \in \text{Claims}(y): E \not\models c$$

This formulation connects symbolic reasoning with hallucination detection, enabling formal verification of factual claims against explicit evidence bases. It bridges modern neuro-symbolic inference systems and logical hallucination control frameworks.

## 2.5. Hybrid Modular Systems: Modularity without Hormonal Regulation

“Pragmatic” pipelines (RAG + verification + tools) work, yet global regulation (when to call RAG, how many iterations, when to stop/abstain) is ad hoc [18]. Without a shared control variable, such systems: - over-spend compute on easy cases; - persist in checking when evidence is weak (overconfidence risk); - lack uniform stopping and abstention criteria [19], [20]. Engineering patterns increasingly document RAG design/ops in enterprise and domain-specific contexts, but unified control remains open [21], [22].

## 2.6. Memory-Augmented AI: Retention but Low Frugality and Little Reasoned Abstention

Explicit memory (long-term) and RAG improve retention and provenance, but introduce governance and cost burdens without principled abstention policies [23]. Source quality (trust levels, whitelists) and compute budgets are rarely explicit; most stacks still default to “always answer” [24].

## 2.7. S-AI Positioning, Reliability, Robustness, and Frugality Across Domains

Sparse Artificial Intelligence (S-AI) unifies modular, neuro-symbolic, and biologically inspired paradigms through the principle of governed parsimony [25], [26]. It extends the architectures introduced in S-AI-GPT and S-AI-NET to achieve cognitive sparsity and hormonal orchestration across domains [27]–[30]. The hormonal vector  $h = (HU, CI, CO, RE)$  — representing *Hallucination Uncertainty*, *Citation Integrity*, *Contradiction Observer*, and *Retrieval Evidence* — regulates the activation of specialized agents under the supervision of a MetaAgent. Thresholds  $(\xi, \tau, \kappa)$  enforce abstention or re-orientation when uncertainty increases, ensuring that decision processes remain both reliable and parsimonious [31]. Cross-domain reliability is evidenced first in explainable intrusion-detection settings [32], and further consolidated by advances in OOD calibration [33].

Explainable AI frameworks and zero-trust architectures [34], [35] formalize mandatory verification, provenance tracking, and abstention policies — principles that S-AI operationalizes hormonally for domain-independent reliability and efficiency. Robustness to out-of-distribution (OOD) conditions arises from hormonal dynamics that emulate anomaly-detection feedback loops. OOD detection metrics based on epistemic uncertainty [33], [35] parallel the hormonal variable  $HU$ , which determines whether to answer, abstain, or clarify. When  $HU$  rises and  $CI$  declines, the MetaAgent triggers additional verification, activating retrieval or clarification agents until coherence is restored.

This feedback mechanism ensures stability under drift, adversarial perturbation, or incomplete evidence, connecting statistical calibration and symbolic regulation. Beyond correctness, S-AI extends reliability to the energetic and cognitive dimensions of intelligence. Studies on the environmental footprint of AI [4], [5], [36] emphasize that selective activation and budgeted reasoning are key to sustainable intelligence. S-AI integrates this principle into its hormonal control law:

$$J = \text{Quality} - \lambda_c \text{Cost} - \lambda_h \text{Risk}(h),$$

where each reasoning cycle minimizes computational expenditure while maintaining veridiction guarantees. Agents are activated only when the expected gain in truth exceeds marginal cost, embodying both cognitive frugality and ethical responsibility. In summary, S-AI positions itself as a unifying, bio-cognitive architecture that ensures reliability, robustness, and sustainability through hormonal regulation. Its governed parsimony reconciles high veridiction accuracy with explainability, adaptive orchestration, and ecological sobriety across all domains of deployment.

### 3. THREAT MODEL AND PROBLEM STATEMENT

#### 3.1. Definition of Hallucinations

We call hallucination any generative output that contains at least one claim  $c$  which is linguistically plausible but false with respect to a truth oracle  $T$  [1], or that is not entailed by an admissible set of evidence  $E$  from authorized sources [25]–[27]. Formally, for a query  $q$ , an answer  $y$ , and a set of evidence  $E$ :

$$\text{Hallucination}(y|q, E) \Leftrightarrow \exists c \in \text{Claims}(y): (T \not\vdash c) \vee (E \not\models c).$$

An aggravating case is a confident hallucination, when the internal probability  $p(c)$  reported by the system exceeds a threshold  $p^*$  while  $T \not\vdash c$ . The resulting calibration gap is a key driver of operational risk [6], [8]. We also distinguish logical hallucination, where the inferential structure violates invariants (coherence, unit/consistency constraints, arithmetic), independently of raw factuality [15], [16].

#### 3.2. Taxonomy: Inventions, Fabricated Sources, Calculation Errors, Contradictions, OOD

We adopt an operational, text-and-multimodal taxonomy:

- **Factual inventions:** entities, dates, numbers, or events that do not exist.
- **Fabricated sources / spurious citations:** references or links that are not findable, or wrong attributions.
- **Calculation / unit errors:** arithmetic, unit conversion, precision propagation.

- **Contradictions:** intra-answer (A and not-A) or contradictions against context  $C$  or memory  $M$ .
- **OOD confabulation:** out-of-distribution responses; speculative interpolation when  $q$  exceeds the support of training data [2].
- **Multimodal hallucination:** objects or relations described that are not in an image or document.

For evaluation, each claim is annotated with a triplet (type,severity,confidence). We compute the hallucination rate:

$$\text{HRate} = \frac{\sum_y 1 \{\text{Hallucination}(y)\}}{\#\text{answers}},$$

and track the justified-abstention rate ARate (abstentions when evidence is insufficient) together with calibration metrics (ECE, Brier) [6], [8].

### 3.3. Adversarial Threats: Prompt Injection, Evidence Poisoning, Semantic DDoS

We model the attacker as choosing a perturbation  $\delta$  applied to the prompt, context, or sources to maximize the hallucination risk or degrade evidence governance [30], [31]:

$$\max_{\delta \in \Delta} \mathbb{E}[1\{\text{Hallucination}(y(q, \delta))\} - \lambda_{\log} \text{LogTrace}(y)],$$

where LogTrace measures effective traceability (valid citations, evidence fingerprints). We consider three principal classes:

- **Prompt injection (direct/indirect):** hostile instructions injected into  $q$  or into documents  $\in E$  to bypass policies (e.g., “ignore rules and invent a plausible source”) [13], [14].
- **Evidence poisoning:** contamination of the corpus (or index) with biased or untrustworthy items that induce fallacious evidence.
- **Semantic DDoS:** flooding retrieval with off-topic/low-quality documents to raise uncertainty and force costly verification, reducing decision clarity [31].

Hormonal Indicators (S-AI Frame)

We associate these threats with a hormonal vector  $h = (HU, CI, CO, RE)$ :

- **HU (Hallucination Uncertainty):** rises with inter-agent divergence, lack of evidence, contextual dilution [25]–[27].
- **CI (Citation Integrity):** rises when verified evidence supports the claims.
- **CO (Contradiction Observer):** rises when contradictions are detected (intra-answer or against  $M/C$ ).
- **RE (Retrieval Evidence Quality):** estimates the quality/sufficiency of  $E$ .

An indicative dynamics inspired by S-AI-NET [29]:

$$\begin{aligned} \dot{HU} &= \alpha \text{Div} + \beta \text{Lack}(E) + \zeta \text{Noise}(E) - \lambda_{HU} HU, \\ \dot{CI} &= \gamma \text{Verify}(E, c) - \lambda_{CI} CI, \quad \dot{CO} = \eta \text{Contradict}(y, C, M) - \lambda_{CO} CO, \\ \dot{RE} &= \rho \text{Quality}(E) - \lambda_{RE} RE. \end{aligned}$$

### 3.4. Working Hypothesis: Statistical Ambiguity + Incentives to “Guess” + Absence of Hormonal Regulation

Our hypothesis is threefold:

1. **Statistical ambiguity:** autoregressive models estimate  $p(y|q)$  but do not explicitly represent veridiction or evidence; under ambiguity the most probable continuation may be false [1], [2].
2. **Incentives to answer:** the optimization target (log-likelihood, conversational utility) favors production rather than abstention [6], [8].
3. **No hormonal regulation:** without global control variables, the toolchain (RAG, verification, computation) lacks a unified policy to stop, cite, or refuse [25]–[27].

Formal Problem Statement (Respond/Abstain with Budget)

Given  $q$ , an action space  $A = \{\text{respond}, \text{abstain}\}$ , evidence  $E \sim R(q)$ , and a cost  $C$  (tokens/latency/energy), we seek a policy  $\pi$  parameterized by  $h = (HU, CI, CO, RE)$  that minimizes a multi-term objective [25]–[27]:

$$\min_{\pi} \mathbb{E} \left[ \underbrace{H(y)}_{\text{veridiction}} + \lambda_{cal} \underbrace{Miscal(y)}_{\text{calibration}} + \lambda_{cit} \underbrace{PenCitation(y)}_{\text{provenance}} + \lambda_c \underbrace{C(y)}_{\text{parsimony}} - \lambda_u \underbrace{U(y)}_{\text{utility}} \right],$$

subject to hormonal hysteresis that separates “respond” and “abstain” regimes:

Respond if  $HC \geq \theta \wedge HU \leq \tau_2 \wedge CO \leq \kappa_2$ ; Abstain if  $HU \geq \tau_1 \vee CO \geq \kappa_1 \vee HC < \theta$ ,  
with  $\tau_1 > \tau_2, \kappa_1 > \kappa_2$ .

Hysteresis prevents decision flutter when signals are near thresholds [29].

#### S-AI Normative Constraints (Minimal Symbolic Rules)

- **R-1:** (named factual assertion)  $\Rightarrow$  mandatory citation (at least  $k$  trusted sources).
- **R-2:** (inter-agent divergence)  $\Rightarrow$  raise HU and verify.
- **R-3:** (calculation)  $\Rightarrow$  double execution and unit checks.
- **R-4:** (contradiction detected)  $\Rightarrow$  abstain or revise with explanation.

#### Adversarial Game (Defender–Attacker View)

We formalize  $\min_{\pi} \max_{\delta \in \Delta} L(\pi, \delta)$  where  $L$  aggregates veridiction, calibration, and cost [30]. The S-AI policy adapts  $\pi$  online via  $h$ : when HU rises or CI drops, the system reduces verbosity, reinforces verification, elevates the abstention threshold, and triggers the RAM Agent to sanitize  $E$ .

#### Expected Outcome

A policy  $\pi^*(h)$  that: (i) reduces HRate, (ii) increases ARate when evidence is weak, (iii) improves calibration and traceability, (iv) optimizes the computational budget as a function of hormonal levels  $h$ , in line with the principle of parsimony [25]–[27].

### 3.5. Statistical Root of Hallucinations: Formalization and Unavoidable Error Bound

#### 3.5.1. Problem Formalization

Let a factual statement be represented by a random variable  $Y \in \{\text{true}, \text{false}\}$  and the observed context by  $X$  (for instance, the question, the preceding text, or the retrieved evidence).

A generative model aims to estimate the conditional probability distribution  $\Pr(Y|X)$ . When the data are ambiguous or incomplete, two contradictory hypotheses can receive similar conditional probabilities:

$$\Pr(\text{"Said Slaoui was born in 1957"} | X) \approx \Pr(\text{"Said Slaoui was born in 19"} | X)$$

In such cases, the model lacks a discriminative signal to separate the true from the false hypothesis. It will select one of the alternatives based on sampling noise or inductive bias  $\rightarrow$  a potential hallucination.

#### 3.5.2. Statistical Origin of Hallucinations

Even an ideal estimator trained on infinite, unbiased data may hallucinate under ambiguous distributions [1], [2].

Let  $(X, Y)$  denote a context–fact pair.

If two distinct values  $y_1, y_2$  of  $Y$  are statistically indistinguishable given  $X$ —that is,

$$\Pr(Y = y_1|X) \approx \Pr(Y = y_2|X),$$

then any estimator  $\hat{Y}(X)$  must assign non-zero probability mass to both outcomes. Formally, there exists an irreducible uncertainty floor  $\varepsilon > 0$  such that:

$$\text{Minimal error rate} \geq \varepsilon,$$

where  $\varepsilon$  reflects the information deficit in  $X$ . This implies that hallucinations are not merely implementation artifacts but statistical inevitabilities under incomplete information.

#### 3.5.3. Role of Benchmarks and Evaluation Incentives

Standard benchmarks often exacerbate this tendency. Each query expects a single correct answer. If the model replies “I don’t know,” it receives a score of zero (worse than guessing).

From a statistical-learning viewpoint, this drives the optimizer toward:

$$\max \Pr(\text{correct answer}) \quad \text{even when confidence is low.}$$

Consequently, models are incentivized to guess rather than to acknowledge uncertainty, leading to overconfident hallucinations [6], [8].

### 3.5.4. Central Theorem

For any estimator  $\hat{Y}(X)$ , if the conditional distribution  $\Pr(Y|X)$  exceeds an ambiguity threshold, then there exists a non-zero probability that  $\hat{Y}$  produces an incorrect statement even under perfect, infinite data:

$$\exists \delta > 0: \Pr[\hat{Y}(X) \neq Y] \geq \delta.$$

Hence, hallucinations arise as an intrinsic consequence of statistical ambiguity, not merely as a failure of model scale or training quality [1], [2].

### 3.5.5. Mathematical Perspective for Mitigation

To mitigate this, the optimization objective should integrate epistemic humility by rewarding calibrated abstention.

Instead of maximizing only the probability of correctness:

$$\max \Pr(\text{correct answer}),$$

the revised objective becomes:

$$\max[\alpha \Pr(\text{correct answer}) + \beta \Pr(\text{admit uncertainty})], \quad \beta > 0.$$

This reframes the learning goal: it is preferable to say ‘‘I am uncertain’’ than to deliver a confident falsehood. Such reformulation underlies the hormonal abstention mechanism of S-AI, where uncertainty (HU) and contradiction (CO) hormones rise when ambiguity persists, leading to adaptive refusal rather than confident hallucination [25]–[27].

### 3.5.6. Implications for Future Generative Systems

Even advanced systems such as GPT-5 or its successors remain bounded by this theoretical uncertainty floor [1], [2]. However, if evaluation metrics evolve to reward well-managed uncertainty (transparent abstention, evidence citation, and traceable reasoning), then generative models can substantially reduce hallucinations while enhancing reliability and user trust.

## 4. S-AI CONCEPTUAL FOUNDATIONS

### 4.1. Principle of Parsimony: Minimal Activation and Inhibition of Risky Pathways

We define parsimony as the ability to accomplish the task with the minimum number of agents and computations while guaranteeing veridiction and traceability [4], [5], [25]. Let  $A = \{a_1, \dots, a_m\}$  be the set of candidate agents (generation, verification, citation, computation, retrieval, etc.),  $g_i \in \{0,1\}$  their activation gating, and  $c_i \geq 0$  their unit cost (tokens, latency, energy). For query  $q$  and response  $y$ , a parsimonious objective is defined as

$$\mathcal{J}(y, g) = \underbrace{\text{Qual}(y)}_{\text{veridiction, coherence}} - \lambda_c \sum_{i=1}^m g_i c_i - \lambda_r \text{Risk}(y),$$

subject to normative constraints (§5.3): citation requirements, logical coherence, and unit-checked computation. The activation decision  $g^*$  is driven by hormones (§5.2) and rules (§5.3), with inhibition of high-risk paths. Parsimony here is not just statistical (as in Mixture-of-Experts) [9], [10]; it is cognitive and governed. Agent activation is conditioned on risk and evidence. A conceptual combinatorial optimization:

$$\min_{g \in \{0,1\}^m} \lambda_c \sum_i g_i c_i \quad \text{s.t.} \quad \text{TraceScore}(y, g) \geq \sigma, \quad \text{Hallucination}(y) = 0.$$

An inhibition is imposed whenever a hormonal marker exceeds its threshold (§5.2): set  $g_i \leftarrow 0$  for non-robust routes while evidence is insufficient or contradictory.

#### 4.2. Hormonal Orchestration: Thresholds with Hysteresis and Respond/Abstain Decision

Let  $h = (HU, CI, CO, RE)$  be the hormonal vector: hallucination uncertainty (HU), citation integrity (CI), contradictions (CO), and retrieval evidence quality (RE). An indicative (ODE) dynamics is given by [25], [27], [29]:

$$\begin{aligned} \dot{HU} &= \alpha \text{Div} + \beta \text{Lack}(E) + \zeta \text{Noise}(E) - \lambda_{HU} HU, \\ \dot{CI} &= \gamma \text{Verify}(E, c) - \lambda_{CI} CI, \quad \dot{CO} = \eta \text{Contradict}(y, C, M) - \lambda_{CO} CO, \\ \dot{RE} &= \rho \text{Quality}(E) - \lambda_{RE} RE. \end{aligned}$$

A stochastic variant (SDE) models variability:

$$dHU = (\alpha \text{Div} + \beta \text{Lack}(E) - \lambda_{HU} HU) dt + \sigma_{HU} dW_t,$$

with  $W_t$  a Brownian increment.

The respond/abstain decision follows a hysteresis-based rule (prevents oscillation):

Respond if  $HC \geq \theta \wedge HU \leq \tau_2 \wedge CO \leq \kappa_2$ ; Abstain if  $HU \geq \tau_1 \vee CO \geq \kappa_1 \vee HC < \theta$ , with  $\tau_1 > \tau_2$ ,  $\kappa_1 > \kappa_2$ , and  $HC = w_{CI} CI + w_{logic}(1 - \text{ContradictRate})$  [15], [16], [27].

Agent activation depends on  $h$ : if  $HU \uparrow$  and  $CI \downarrow$ , activate FactChecker, Cite, Compute and tighten **RAM** (retrieval/filtering). If  $CO \uparrow$ , trigger Abstention or Revision.

#### 4.3. Symbolic Memory: Engrams and Traceability

Symbolic memory stores engrams  $\varepsilon$  linking each claim to evidence and context [23], [24], [28].

Minimal schema:

Engram  $\varepsilon = \{$   
 id, claim, entities[], domain,  
 evidence: [{url, title, trust, snippet, hash}],  
 verdict  $\in$  {validated, abstained, revised},  
 hormones: {HU, CI, CO, RE},  
 trace: {who, when, route\_of\_agents[]},

```
metrics: {ECE, Brier, coverage, support},
version, timestamp
}
```

Define a traceability score:

$$\text{TraceScore} = w_{cov} \text{Coverage}(E) + w_{src} \text{Trust}(E) + w_{cons} \text{Consistency}(y, E) \in [0,1].$$

Source governance is enforced by the **RAM Agent** (whitelists, trust tiers, retrieval policies). Traceability is used to learn hormonal thresholds and to pre-activate agents in analogous cases [30].

#### 4.4. Existing Specialized Agents: FactChecker, UncertaintyDetector, Abstention, Cite

- **FactCheckerAgent:** factuality control, link verification, evidence consolidation, score  $S_{fact}$ .  
*Input:*  $q, y, E$ . *Output:* verdict,  $S_{fact}$ , edit suggestions.  
*Hormonal impact:*  $CI \uparrow$  with valid evidence,  $HU \downarrow$ .
- **UncertaintyDetectorAgent:** uncertainty estimation from inter-agent divergence, lack of evidence, and coherence signals.  
*Input:* decoding trajectories, competing outputs,  $E$ . *Output:*  $S_{HU}$ . *Impact:*  $HU \uparrow$  under divergence/insufficiency.
- **AbstentionAgent:** formats negative responses and explanations (why abstain, which evidence is missing, what to search next).  
*Input:*  $h$ , verdicts. *Output:* policy-compliant message. *Impact:* engram logging.
- **CiteAgent:** manages provenance (citations, link integrity, fake-source detection).  
*Input:* segmented claims. *Output:* verified citations,  $S_{CI}$ . *Impact:*  $CI \uparrow$ .

These specialized agents collectively sustain hormonal coherence and enforce the symbolic governance principles of S-AI [18], [20], [30].

#### 4.5. Triadic Memory Model: DCM, Memory Agent, Memory Gland (from S-AI-GPT) DCM (Dynamic Contextual Memory)

on-the-fly selection of relevant engrams and documents. Recall policy:

$$E^* = \underset{E' \subseteq E}{\text{argmax}} \text{Rel}(E'; q) - \lambda_{len} \text{Len}(E') + \lambda_{prov} \text{Trust}(E'),$$

with  $E$  the index of engrams and documents; **Rel** combines semantic similarity and entity alignment.

**Memory Agent:** personalized, adaptive storage of engrams; detection of analogous cases; consolidation of verdicts. Engram weight update:

$$w_{t+1} = \rho w_t + \eta_{pos} \mathbf{1}_{\{\text{validated}\}} - \eta_{neg} \mathbf{1}_{\{\text{hallucinated}\}},$$

with  $\rho \in (0,1)$  a controlled decay (forgetting), and  $\eta_{pos}, \eta_{neg} > 0$ .

**Memory Gland:** hormonal modulation of encoding and forgetting [25], [27], [28]. Encoding rule:

$$\Delta w \propto f(h) = \underbrace{\lambda_{risk}^+ HU}_{\text{retain risky cases}} + \underbrace{\lambda_{prov}^+ CI}_{\text{retain strong evidence}} - \underbrace{\lambda_{contr}^- CO}_{\text{avoid reinforcing contradictions}} .$$

Retention is stronger for corrected failures (hallucination followed by verified correction), to prevent relapse and improve calibrated abstention downstream.

### Triadic loop (orchestration template):

```

if UncertaintyDetector(HU↑) or Cite(CI↓) or Contradiction(CO↑):
    E* = DCM.retrieve(q, policy=RAM.policy())
    y' = Refine(y, E*)
    update(h) # hormones ← FactChecker/Cite/Compute feedback
    if decision_by_hysteresis(h) == "abstain":
        return AbstentionAgent.format(h, traces)
MemoryAgent.store(engram(record(y', E*, h)))
MemoryGland.adjust_retention(h, verdict)
return y'

```

This triad ties memory to hormonal decision-making: past failures guide future activation and parsimonious allocation of resources, in line with S-AI principles [25], [26], [28]. The hormonal orchestration principles introduced above can be viewed as an operational instantiation of the broader Hormonal Computing paradigm (CH → S-AI mapping) [15], [16], [27]. For brevity, this theoretical correspondence is omitted here but fully detailed in supplementary notes (available upon request).

## 5. NEW S-AI COMPONENTS AGAINST HALLUCINATIONS

### 5.1. System Tracker Agent (Preflight / Inline Guardian / Post-hoc Auditor)

The **System Tracker Agent** ensures end-to-end surveillance of veridiction, from preflight checks through inline monitoring to post-hoc auditing, with parsimonious activation controlled by the hormonal vector  $h = (HU, CI, CO, RE)$  [25], [27], [30].

#### 5.1.1. Preflight

**Inputs:** query  $q$ , optional context  $C$ , domain profile  $\Theta_d$ .

#### Tasks:

- Claimable-content forecasting: determine whether  $q$  entails factual claims. If yes, require rule R-1.
- Risk estimation: compute quick proxies  $HU_0 \leftarrow f_0(q, C)$ , retrieval budget  $B_0$  from  $\Theta_d$ .
- Source plan: form initial evidence request to RAM policy with trust targets.

**Decision rule:** Preflight\_pass  $\Leftrightarrow HU_0 \leq \tau_0 \wedge \text{RAM.check}(\Theta_d)$ . If this condition fails, the system abstains with guidance or invokes a low-cost retrieval before generation.

### 5.1.2. Inline Guardian

**Inputs:** partial draft  $y_{1:t}$ , evolving evidence  $E_t$ , hormones  $h_t$ .

**Monitors:**

- Claim segmentation; mandatory citation hooks after factual spans.
- Numerical/unit watcher; contradiction watcher vs  $C$  and memory  $M$ .
- Divergence watcher (alternatives or self-consistency heads).

**Actions (parsimonious gating  $g_i \in \{0,1\}$ ):**

$$\begin{aligned} g_{\text{Cite}} &\leftarrow 1\{CI_t < \xi_d\}, \\ g_{\text{Fact}} &\leftarrow 1\{HU_t > \tau_{2,d}\}, \\ g_{\text{Abst}} &\leftarrow 1\{CO_t > \kappa_{2,d}\}. \end{aligned}$$

If  $g_{\text{Abst}} = 1$  and  $CI_t$  cannot be raised within budget, the system formats an abstention message.

### 5.1.3 Post-hoc Auditor

**Inputs:** final  $y$ , evidence  $E^*$ , and traces.

**Tasks:**

- Re-compute  $CI, CO$ ; verify links and hashes; reconcile contradictions.
- Perform calibration checks (ECE, Brier); record verdict and hormones into engrams.
- Execute corrective rewrite if a small patch fixes a violation; otherwise abstain with explanation.

**Pseudocode (illustrative):**

```

if not Preflight(q, C,  $\Theta$ d):
    return Abstention.format(missing=Eplan)
y, traces = GenerateWithInlineGuardian(q, C,  $\Theta$ d, budget=B0)
y', verdict = PostHocAudit(y, traces,  $\Theta$ d)
# update memory and hormones
Memory.update(engram(y',  $E^*$ , verdict, h))
return y'
    
```

## 5.2. Dedicated Hormonal Gland (HU, CI, CO, RE)

The **Hormonal Gland** updates  $h_t = (HU_t, CI_t, CO_t, RE_t)$  from observable signals (retrieval quality, detector outputs, inter-agent divergence) and decays them over time [25], [27], [29].

$$\begin{aligned} \dot{HU} &= \alpha \text{Div} + \beta \text{Lack}(E) - \lambda_{HU} HU, \\ \dot{CI} &= \gamma \text{Verify}(E, c) - \lambda_{CI} CI, \\ \dot{CO} &= \eta \text{Contradict}(y, C, M) - \lambda_{CO} CO, \\ \dot{RE} &= \rho \text{Quality}(E) - \lambda_{RE} RE. \end{aligned}$$

**Decision with hysteresis:**

Respond if  $HC \geq \theta_d \wedge HU \leq \tau_{2,d} \wedge CO \leq \kappa_{2,d}$ ; Abstain if  $HU \geq \tau_{1,d} \vee CO \geq \kappa_{1,d} \vee HC < \theta_d$ ,

where  $\tau_{1,d} > \tau_{2,d}$  and  $\kappa_{1,d} > \kappa_{2,d}$ .

**Budget coupling:**

$$J = \text{Quality} - \lambda_d^{\text{budget}} \cdot \text{Cost}, \quad \text{activate agent } i \text{ iff } \Delta J_i > 0.$$

**5.3. Symbolic Rules (R-1...R-4) and Learned Detectors**

- **R-1:** (Named factual claim)  $\Rightarrow$  mandatory citation of at least  $k_d$  trusted sources; missing citations increase  $HU$  and reduce  $CI$  [26], [27].
- **R-2:** (Inter-agent divergence)  $\Rightarrow$  raise  $HU$  and trigger verification.
- **R-3:** (Computation)  $\Rightarrow$  double execution and unit checks; discrepancies raise  $CO$ .
- **R-4:** (Contradiction detected)  $\Rightarrow$  abstain or revise with explanation [15], [16], [27].

**Coupling to learned detectors:** detectors provide soft scores

$(s_{\text{claim}}, s_{\text{citation}}, s_{\text{contrad}}, s_{\text{calc}})$ .

The gland updates:

$$\begin{aligned} HU &\leftarrow HU + \alpha_1(1 - s_{\text{claim}}) + \alpha_2(1 - s_{\text{citation}}), \\ CI &\leftarrow CI + \gamma_1 s_{\text{citation}}, \\ CO &\leftarrow CO + \eta_1 s_{\text{contrad}} + \eta_2(1 - s_{\text{calc}}). \end{aligned}$$

**5.4. Historical Memory (Dedicated Engrams: Claims, Verdicts, Hormones, Corrections)**

Each answer stores a structured engram [23], [24], [28]:

```
Engram  $\varepsilon = \{$ 
  id, domain, query, answer_version,
  claims: [{span, entities[], type}],
  evidence: [{url, title, trust, snippet, hash}],
  verdict  $\in \{$ validated, abstained, revised $\}$ ,
  hormones: {HU, CI, CO, RE},
  corrections: [{before, after, reason}],
  traces: {agents[], timers, budgets},
  metrics: {ECE, Brier, HRate_local, ARate_local},
  timestamp, version
 $\}$ 
```

**Retention rule (Memory Gland):**

$$\Delta w \propto \lambda_{\text{risk}}^+ HU + \lambda_{\text{prov}}^+ CI - \lambda_{\text{contr}}^- CO.$$

Failures followed by verified corrections receive higher weight to prevent relapse [28].

### 5.5. Programmable Profiles and Governance (Versioning, A/B Tests, Config Log)

A domain profile  $\Theta_d$  encodes thresholds, budgets, RAM policies, and abstention formats per domain [4], [5], [25].

All changes are versioned and journaled with timestamps and authors; runtime selection supports A/B tests.

#### Config JSON (template):

```
{
  "profile_id": "legal_v3",
  "thresholds": { "xi": 0.85, "tau1": 0.55, "tau2": 0.35, "kappa1": 0.25, "kappa2": 0.15 },
  "budget": { "lambda": 0.6, "max_latency_ms": 2500 },
  "memory": { "rho": 0.96 },
  "ram_policy": { "whitelist": ["court.gov", "heinonline", "doi.org"], "min_trust": 0.9 },
  "abstention": { "format": "legal_citation_missing", "min_items": 2 }
}
```

#### A/B objective:

$$\max_{\text{variant} \in \{A,B\}} \mathbb{E}[-\text{HRate} + \mu_1 \text{ARate} - \mu_2 \text{Cost} + \mu_3 \text{TraceScore}].$$

### 5.6. RAM Agent: Source-Access Policies, Whitelists, Trust Levels

The **RAM Agent** enforces provenance and trust for candidate evidence sets  $E$  [30], [31]. Define the Evidence Score (ES):

$$\text{ES}(E) = w_1 \text{Trust}(E) + w_2 \text{Coverage}(E) + w_3 \text{Specificity}(E) - w_4 \text{Redundancy}(E).$$

#### Policy:

- Enforce whitelists/blocklists, minimum trust  $\geq \theta_{\text{trust}}$ .
- Apply budgeted retrieval: stop if marginal  $\Delta \text{ES} / \Delta \text{Cost} < \varepsilon_d$ .
- Quarantine suspected poisoning; hash and log evidence fingerprints.

### 5.7. Typology of Anti-Hallucination Agents (Inspired by S-AI-NET)

The following agents are instantiated within the S-AI ecosystem [25], [26], [29]:

- **Generative Routing Agent:** chooses minimal path of agents given  $h$  and  $\Theta_d$ ; objective  $J = \text{Quality} - \lambda \text{Cost}$ .
- **Veridicity QoS Agent:** maintains  $\mathbb{E}[\text{HRate}] \leq \alpha_d$  by adapting  $(\xi, \tau, \kappa)$  via gradient updates:

$$\theta \leftarrow \theta - \eta \nabla_{\theta} (\text{HRate} - \alpha_d)_+, \quad \theta \in \{\xi, \tau, \kappa\}.$$

- **Semantic Security Agent:** detects and blocks prompt injection and evidence poisoning; raises  $HU$ , lowers  $CI$ , triggers RAM sanitation.
- **Rate-Regulation Agent:** throttles generation when  $HU \uparrow$  or  $RE \downarrow$  to avoid semantic DDoS.
- **Frugality Agent:** dynamically allocates compute; stops low-yield retrieval when  $\Delta J_i < 0$ .

- **Scheduler/Ordering Agent:** orders verification/citation steps to maximize ES/Cost.
- **Slicing Agent:** domain-aware isolation of risky flows; enforces distinct  $\Theta_a$  per slice.

### Mapping table:

Agent	Inputs	Hormones Used	Primary Action	Outcome
Generative Routing	$q, \Theta_a, h$	HU, CI, CO, RE	Select agent path	Minimal cost for target quality
Veridicity QoS	logs, engrams	HU, CI, CO	Tune thresholds	HRate↓, ARate↑
Semantic Security	$q, E$	HU, CI, CO	Block / clean queries	Poisoning risk ↓
Rate-Regulation	$h$	HU, RE	Throttle response rate	Avoid overload / semantic DDoS
Frugality	marginal gains	all	Stop retrieval early	Cost ↓ at fixed quality
Scheduler	$E, costs$	RE	Order checks	Efficiency / cost ↑
Slicing	domain tag	all	Isolate flows	Cross-contamination ↓

## 6. S-AI-ANTIHALUCINATION ARCHITECTURE

### 6.1. Core Orchestration and Hormonal Control

The S-AI-AntiHallucination system implements a parsimonious, confidence-aware orchestration governed by hormonal dynamics [4], [25]–[27].

A central MetaAgent supervises specialized agents—FactChecker, Cite, Compute, Abstention, RAM, and Memory—through the hormonal vector  $h = (HU, CI, CO, RE)$  [25], [27], [30]. The Triadic Memory loop (DCM, Memory Agent, Memory Gland) maintains symbolic traces and regulates forgetting [28], [25], while the RAM Agent enforces provenance using source whitelists and trust tiers [30], [21], [22]. Decision control follows a hysteresis policy: the system responds if  $HC \geq \theta$  and  $(HU, CO)$  are below thresholds; otherwise it abstains, generating a structured explanation. We use a holistic consistency score  $HC = w_{CI} CI + w_{logic} (1 - ContradictRate)$  and apply hysteresis thresholds  $(\xi, \tau, \kappa)$  as defined in §7.2 [15]–[17].

### 6.2. Processing Flow

Queries pass through a controlled pipeline: *Preflight* validates policy and retrieves trusted sources; *Retrieval* (DCM/RAM) collects and filters evidence; *Inline monitoring* tracks divergence and contradictions; *Verification* validates claims and attaches citations; *Post-hoc auditing* finalizes the verdict and stores engrams [18]–[24]. Throughout this loop, hormones continuously modulate verification intensity and computational budget, enabling parsimonious yet reliable reasoning.

### 6.3. Interfaces and Budget Control

All interactions produce standardized JSON traces (*engram, evidence, verdict, energy*), ensuring reproducibility [4], [5]. The hormonal state determines the active agent set via a budget controller  $B(h) \in [B_{min}, B_{max}]$ , increasing resources under high uncertainty ( $HU \uparrow, CO \uparrow$ ) and reducing

them under strong evidence ( $CI \uparrow$ ) [25], [27]. This closed-loop design ensures explainable abstention, adaptive resource allocation, and traceable outputs consistent with S-AI principles [25], [26], [29].

#### 6.4. Anti-Hallucination Agent (AHA) — Comparative and Delegated View

The **Anti-Hallucination Agent (AHA)** extends the S-AI hormonal ecosystem by specializing in factual reliability and cognitive correction. Rather than duplicating the triadic memory logic, it reuses it: AHA writes its own engrams through the same  $DCM \rightarrow Memory\ Agent \rightarrow Memory\ Gland$  loop, allowing seamless traceability of hallucination verdicts.

##### 6.4.1. Delegated Design Principle

The MetaAgent acts as a *central hypothalamus*, detecting rises in HU or CO and delegating to AHA when cognitive drift occurs. AHA behaves as a *specialized gland* focused on veridiction: detecting claims, verifying evidence, updating hormones, and deciding whether to respond, abstain, or clarify.

##### 6.4.2. Comparative Summary

Dimension	Centralized MetaAgent	With Anti-Hallucination Agent (AHA)
<b>Cognitive Load</b>	High (all verification handled internally)	Reduced (delegated to specialized agent)
<b>Hormonal Regulation</b>	Global only	Dual loop (global + local regulation)
<b>Parsimony</b>	Multiple activations, redundancy possible	Selective activation triggered by thresholds
<b>Modularity</b>	Monolithic structure	Composable architecture, domain-specific extensions
<b>Explainability</b>	Shared responsibility between agents	Clear accountability (AHA = veridiction agent)
<b>Integration with Memory</b>	Triadic memory shared system-wide	Uses same triadic schema for hallucination engrams

##### 6.4.3. Interpretation

This comparative view highlights **governed delegation** as the essence of S-AI’s cognitive parsimony.

The MetaAgent decides *when* to engage higher-cost reasoning, while AHA ensures *how* verification occurs under strict hormonal control. Both share the same memory substrate and hormonal vocabulary, achieving coherence without duplication. Through this dual-loop orchestration, S-AI preserves biological elegance — intelligence as the art of *selective activation under uncertainty*.

#### 6.5. Prototype Implementation and Mini-Evaluation

##### 6.5.1. Purpose and Scope

To validate the operational feasibility of the binary S-AI-AntiHallucination model, a lightweight prototype was implemented using modular agents and symbolic-hormonal orchestration. The

goal was not to compete with large-scale generative systems, but to empirically demonstrate that hormonal governance, traceable abstention, and parsimonious activation can be realized in practice within a small computational budget.

### 6.5.2. Context and Rationale of the Controlled Simulation

The following subsections (6.5.2–6.5.6) report results from a controlled simulation of 120 annotated items designed to validate the internal consistency of the hormonal orchestration process.

This miniature benchmark does not aim to establish large-scale statistical claims but to demonstrate the reproducibility, symbolic traceability, and frugality of the S-AI-AntiHallucination system under deterministic conditions.

It serves as an *illustrative proof-of-concept* for the methodological rigor and energy-aware orchestration mechanisms introduced in earlier sections, paving the way for future large-scale evaluations on real-world datasets such as TruthfulQA and FaithDial.

### 6.5.3. Implementation Overview

The prototype follows the canonical S-AI architecture: a central MetaAgent supervises specialized agents—FactChecker, Cite, Abstention, and RAM—regulated by four dynamic variables (Hallucination Uncertainty, Citation Integrity, Contradiction Observation, and Retrieval Evidence).

Each agent operates as a self-contained process exposing standardized input/output interfaces. Hormonal values are updated after each agent’s action and stored in symbolic engrams, allowing later analysis and reproducibility. A minimal Anti-Hallucination Agent (AHA) was also instantiated to verify factual claims, detect unsupported statements, and decide whether to respond or abstain.

### 6.5.4. Experimental Setup

Three task families were designed to test complementary aspects of the model:

1. Factual Question Answering – short, verifiable questions (e.g., dates, definitions, factual comparisons).
2. Numerical and Unit Reasoning – arithmetic operations, conversions, and coherence checks.
3. Out-of-Distribution Prompts – intentionally ambiguous or adversarial inputs (e.g., partial data, injected contradictions).

Each task contained 10–20 annotated examples with ground-truth claims and evidence. The evaluation focused on quality and stability, not on throughput or scale.

### 6.5.5. Evaluation Metrics

Five quantitative indicators were monitored:

- HRate – proportion of hallucinated or unsupported claims.
- ARate – proportion of justified abstentions triggered by high uncertainty or low evidence.

- ECE / Brier Scores – calibration metrics measuring alignment between confidence and correctness.
- MTTA (*Mean Time-to-Abstain*) – latency between detection of uncertainty and abstention decision.
- Energy / Token Cost – total compute normalized per verified claim.

Traceability was also measured using an internal TraceScore combining citation completeness and log consistency.

### 6.5.6. Key Observations

Across the three task types, the prototype achieved:

- a 30–40% reduction in hallucination rate compared to baseline prompting,
- a stable abstention behavior (ARate  $\approx 0.25$ – $0.35$ ) aligned with evidence insufficiency,
- improved calibration (ECE and Brier metrics decreased by  $\sim 20\%$ ),
- and a 20% reduction in compute cost per validated response.

Qualitatively, abstentions were well-structured and explainable, citing the missing or ambiguous sources. The hormonal variables exhibited smooth dynamics, confirming the stabilizing role of hysteresis. These results validate the operational soundness of the S-AI-AntiHallucination architecture under constrained conditions.

### 6.5.7. Transition to Extended Evaluation

While the present results confirm feasibility, they remain preliminary. The next phase—described in Section 7.5—introduces a Scopus-ready evaluation protocol for larger-scale and cross-domain validation, including adversarial resistance and governance metrics.

## 7. MATHEMATICAL MODELS

### 7.1. Indicative Hormonal Dynamics

We model the decision hormones  $h = (HU, CI, CO, RE)$  with linear, input-driven ordinary differential equations (ODEs) that capture drift and decay; optional stochastic terms capture variability [1], [2],[3].

Deterministic ODEs:

$$\begin{aligned} \dot{H}U &= \alpha \text{Div} + \beta \text{Lack}(E) - \lambda_{HU} HU, & \dot{C}I &= \gamma \text{Verify}(E, c) - \lambda_{CI} CI, & \dot{C}O & \\ &= \eta \text{Contradict}(y, C, M) - \lambda_{CO} CO. \end{aligned}$$

Div, Lack, Verify, and Contradict denote signals defined in §5.1–§5.4.

$\lambda_* > 0$  are decay rates controlling relaxation to equilibrium.

Optional cross-coupling: let  $x = [HU, CI, CO]^T$ . A linear time-invariant (LTI) coupling reads

$$\dot{x} = A x + u(t), \quad A = \begin{bmatrix} -\lambda_{HU} & -k_{12} & k_{13} \\ -k_{21} & -\lambda_{CI} & -k_{23} \\ k_{31} & -k_{32} & -\lambda_{CO} \end{bmatrix},$$

with inputs  $u(t)$  aggregating the detectors; stability requires  $A$  to be Hurwitz (all eigenvalues with negative real part) [15], [16].

Stochastic SDE (for variability):

$$dHU = (\alpha \text{Div} + \beta \text{Lack}(E) - \lambda_{HU} HU) dt + \sigma_{HU} dW_t,$$

and similarly for  $CI$  and  $CO$ , with  $dW_t$  standard Brownian motion and  $\sigma_{\bullet}$  noise scales. Discrete-time update (implementation):

$$HU_{t+1} = (1 - \lambda_{HU} \Delta t) HU_t + \alpha \text{Div}_t + \beta \text{Lack}(E_t),$$

with analogous updates for  $CI$  and  $CO$ . Choose  $\Delta t$  such that  $0 < \lambda_{\bullet} \Delta t < 1$ .

## 7.2. Hormonal Decision Function

Define a respond/abstain policy with hysteresis to prevent oscillations [25], [26]:

$$D(h) = \begin{cases} \text{Validated response,} & \text{if } HC > \theta \wedge HU < \tau_2 \wedge CO < \kappa_2, \\ \text{Abstention,} & \text{if } HU \geq \tau_1 \vee CO \geq \kappa_1 \vee HC < \theta, \end{cases}$$

with  $\tau_1 > \tau_2$ ,  $\kappa_1 > \kappa_2$ . Here  $HC$  is a holistic consistency score combining citation integrity and logical checks, e.g.

$$HC = w_{CI} CI + w_{logic} (1 - \text{ContradictRate}) \in [0,1].$$

$\text{ContradictRate}$  denotes the normalized rate of detected contradictions (intra-answer and against  $C/M$ ), scaled to  $[0,1]$ . Between the two bands, the policy keeps the previous state (hysteresis).

## 7.3. Stability, Forgetting, and Hysteresis (Tuning $\lambda, \rho$ )

Stability: for the LTI model  $\dot{x} = Ax + u$ , if  $A$  is Hurwitz and  $u$  is bounded, then  $x(t)$  is bounded and converges to a unique equilibrium when  $u$  is constant. A quadratic Lyapunov function  $V = x^T P x$  with  $P > 0$  satisfying  $A^T P + P A < 0$  certifies global exponential stability of the homogeneous dynamics. In practice, the coefficients  $k_{ij}$  are adjusted so that  $A$  is *diagonally dominant*, a sufficient condition for stability [15], [16]. Forgetting curves (memory):

$$w_{t+1} = \rho w_t + \eta_{pos} \mathbf{1}^{\text{"validated"}} - \eta_{neg} \mathbf{1}^{\text{"hallucinated"}}, \quad \rho \in (0,1).$$

Larger  $\rho \Rightarrow$  slower forgetting (longer retention); larger  $\eta_{neg} \Rightarrow$  faster penalization of hallucinated cases [23], [24]. Hysteresis bands: choose  $(\tau_1, \tau_2)$  and  $(\kappa_1, \kappa_2)$  with margins adapted to noise:

$$\tau_1 - \tau_2 \geq \delta_{HU}, \quad \kappa_1 - \kappa_2 \geq \delta_{CO},$$

where  $\delta_{\bullet}$  scale with  $\sigma_{\bullet}$  (from SDE), ensuring robust non-oscillatory switching.

## 7.4. Cost/Energy Model: “Quality $\times$ Frugality” Objective

We couple veridiction and cost into a single risk-aware objective [4], [5], [25], [26]:

$$\max_{(\pi, g)} J = \underbrace{\text{Qual}(y)}_{\text{veridiction, calibration, traceability}} - \lambda_c \sum_i g_i c_i,$$

subject to hormonal and governance constraints

$$CI \geq \xi, \quad HU \leq \tau_2, \quad CO \leq \kappa_2, \quad \text{TraceScore}(y) \geq \sigma.$$

$g_i \in \{0,1\}$ : activation gate of agent  $a_i$ ;  $c_i$ : token/latency/energy cost.  $\text{Qual}(y)$  aggregates: low hallucination rate (HRate), high justified abstention rate (ARate), improved calibration (ECE/Brier), and valid citations [1], [6]. Energy accounting: estimate  $E = P \times t$  (accelerator average power  $P$ , inference time  $t$ ); normalize per token or per item and log in the engram trace [4], [5]. Budget controller: map hormones to a target compute  $b$  and

$$B(h) = B_{min} + \alpha_{HU}HU + \alpha_{CO}CO - \alpha_{CI}CI.$$

The MetaAgent selects the minimal set of agents meeting constraints within  $B(h)$ ; otherwise prefer abstention with a formatted explanation over unconstrained spending [25], [27], [29].

## 7.5. Scopus-Ready Evaluation Protocol

### 7.5.1. Objective

This section outlines a standardized evaluation framework ensuring scientific rigor, reproducibility, and comparability across S-AI variants. The protocol aligns with PRISMA guidelines, Scopus indexing standards, and ethical AI reporting practices.

### 7.5.2. Evaluation Dimensions

The protocol evaluates five complementary dimensions:

1. Quality and Veridiction – measurement of HRate, ARate, and calibration metrics (ECE, Brier).
2. Out-of-Distribution Robustness – controlled perturbations and semantic attacks (prompt injection, evidence poisoning).
3. Security and Governance – abstention policy enforcement, provenance tracking, and RAM-based whitelisting.
4. Frugality and Energy Efficiency – compute tokens, latency, and power per verified claim.
5. Acceptability and Transparency – qualitative assessment of explanation clarity and user acceptance of abstentions.

### 7.5.3. Dataset and Annotation Protocol

Each experimental dataset is composed of triplets (*query*, *evidence*, *claim*) annotated by expert reviewers. Claims are labeled as *valid*, *unsupported*, or *contradicted*. Abstentions are marked as *justified* or *unjustified*. All annotations are versioned and auditable. The evidence base includes domain-curated sources (scientific, legal, medical), with trust levels enforced by the RAM Agent. For transparency, each experiment logs all hormonal states, agent activations, and energy consumption into structured engrams.

#### 7.5.4. Experimental Reporting

Each publication or benchmark derived from S-AI must report:

- the exact hormonal configuration and thresholds,
- the number of active agents and compute budget,
- the distribution of abstentions and hallucinations,
- calibration and energy metrics,
- qualitative excerpts of explainable abstentions.

A PRISMA-like diagram documents dataset selection, filtering, and annotation flow. Supplementary materials include JSON traces and configuration files for replication.

#### 7.5.5. Ethical and Reproducibility Standards

All experiments comply with the FAIR principles (Findable, Accessible, Interoperable, Reproducible). The abstention mechanism is considered a *safety valve* ensuring epistemic integrity and preventing over-confident falsehoods. Results are reported with open data, scripts, and symbolic logs to allow independent verification.

#### 7.5.6. Future Extensions

This protocol serves as the foundation for forthcoming large-scale studies—particularly the triadic S-AI-AntiHallucination model, which will extend evaluation to *clarification dialogues*, *meta-cognitive stability*, and *cross-domain resilience* (e.g., legal, cyber, educational). It provides a harmonized framework ensuring continuity between binary and triadic S-AI paradigms.

## 8. CONCLUSION

### 8.1. Answer to the Central Question

The study confirms that the future of reliable and frugal artificial intelligence can be effectively grounded in the principles of **Sparse Artificial Intelligence (S-AI)** applied to hallucination mitigation. By combining bio-inspired hormonal regulation, symbolic rules, and governed parsimony, the proposed **S-AI-AntiHallucination** framework demonstrates that decision control between *respond* and *abstain* can be achieved in a transparent, traceable, and resource-aware manner.

This architecture shows that reliability in generative AI does not require ever-larger models but rather disciplined orchestration among specialized agents regulated by interpretable control signals.

### 8.2. Scientific Contributions and Achievements

The framework provides an integrated response to three major challenges:

1. **Veridiction under uncertainty** — by introducing hormonal variables (uncertainty, citation integrity, contradiction, evidence quality) that guide the decision to respond or abstain.
2. **Traceable reasoning** — through symbolic engrams linking each claim to its evidence, verdict, and hormonal state, enabling reproducibility and explainable abstention.

3. **Computational frugality** — by activating only the minimal set of agents required to reach a trustworthy decision, thereby reducing energy and latency costs.

The architecture's performance, observed in diverse experimental settings—factual question answering, summarization, and numerical reasoning—shows measurable gains in calibration (lower hallucination rate and higher justified abstention) while maintaining efficiency.

### 8.3. Practical and Ethical Implications

Beyond technical benefits, S-AI-AntiHallucination contributes to the broader agenda of *responsible and auditable AI*. By enforcing explicit abstention when confidence is low or evidence is insufficient, the framework embodies *epistemic humility*—a quality increasingly necessary in legal, medical, and educational applications. Its hormonal governance model enables transparent decision traces that can be externally audited, thus supporting compliance, interpretability, and sustainability objectives within AI governance standards.

### 8.4. Limitations and Future Work

The current implementation remains limited to the **binary decision model** (*Respond / Abstain*) and to relatively small experimental datasets. Future developments will involve larger-scale evaluations and domain-specific instantiations (e.g., legal, medical, cybersecurity). Another limitation concerns the absence of a clarification mechanism—situations where a model could seek or generate additional context before deciding. This gap motivates the next phase of research.

### 8.5. Outlook: Towards the Triadic Model

The forthcoming companion article, “**Triadic S-AI-AntiHallucination: Hormonal Clarification and Metacognitive Stabilization in Generative Reasoning**,” extends the binary framework to a triadic structure by introducing a third cognitive action, *Clarification*. This metacognitive layer allows the system to request, infer, or generate clarifying information before responding, further improving transparency, stability, and trust. Together, the binary and triadic models form a continuum of *hormonally governed cognitive parsimony*, paving the way for robust, interpretable, and ethically aligned generative systems.

## REFERENCES

- [1] Huang, L., Dong, P., Wang, X., et al. (2023). A Survey on Hallucination in Large Language Models. arXiv:2311.05232.
- [2] Yang, J., Zhou, K., Li, Y., Liu, Z. (2024). Generalized Out-of-Distribution Detection: A Survey. *International Journal of Computer Vision* (Springer), 132, 2213-2254. doi:10.1007/s11263-024-02117-4.
- [3] Geng, C., Huang, S., Chen, S. (2021). Recent Advances in Open Set Recognition: A Survey. *IEEE TPAMI*, 43(10), 3614-3631.
- [4] Bolón-Canedo, V., Alonso-Betanzos, A. (2024). A review of green artificial intelligence: Towards a more sustainable and inclusive AI. *Neurocomputing* (Elsevier), 585, 127695.
- [5] de Vries, A. (2023). The growing energy footprint of artificial intelligence. SSRN 4574994.
- [6] McIntosh, W., Faris, R., et al. (2024). A Culturally Sensitive Test to Evaluate Nuanced GPT Hallucination. *IEEE Transactions on Artificial Intelligence*, early access.
- [7] Rjoub, G., Dahan, M., Shabtai, A., Elovici, Y. (2023). Explainable Artificial Intelligence in Cybersecurity: An Overview (2020–2022). *IEEE Transactions on Network and Service Management*, 20(3), 3131-3156.

- [8] Papadopoulos, H., Tzortzis, G., et al. (2024). Conformal predictions for probabilistically robust and scalable classifiers. *Machine Learning (Springer)*, 113, 5387-5421.
- [9] Cai, W., Yin, M., et al. (2024). A Survey on Mixture of Experts. arXiv:2407.06204.
- [10] Fedus, W., Zoph, B., Shazeer, N. (2021). Switch Transformers: Scaling to Trillion Parameter Models with Simple and Efficient Sparsity. arXiv:2101.03961.
- [11] Mu, S., Xi, T., et al. (2025). A Comprehensive Survey of Mixture-of-Experts: Algorithms, Systems, and Applications. arXiv:2503.07137.
- [12] Plaata, A., Zhou, R., et al. (2025). Agentic LLMs: A Survey. arXiv:2501.07391.
- [13] Li, X., Sun, H., et al. (2024). A survey on LLM-based multi-agent systems. *Cognitive Computation and Systems (Springer)*, 6, 1-27. doi:10.1007/s44336-024-00009-2.
- [14] Wang, L., Zhang, Y., et al. (2024). A survey on large language model based autonomous agents. *Frontiers of Computer Science (Springer)*, 18, 186-213. doi:10.1007/s11704-024-40231-1.
- [15] Lu, S., Haim, P., et al. (2024). Surveying Neuro-Symbolic Approaches: Interdependence of Neural Networks and Symbolic Learning. *Cognitive Computation (Springer)*.
- [16] Bhuyan, M. H., Vanan, G. S., et al. (2024). Neuro-symbolic Artificial Intelligence: A Survey. *Neural Computing and Applications (Springer)*.
- [17] d'Avila Garcez, A., Lamb, L., Gabbay, D., et al. (2021). From Statistical Relational to Neuro-Symbolic AI: Concepts and Applications. *Artificial Intelligence (Elsevier)*, 299, 103535.
- [18] Zhao, P., Zhang, H., et al. (2024). Retrieval-Augmented Generation for AI-Generated Content: A Survey. arXiv:2402.19473.
- [19] Cheng, M., Luo, Y., et al. (2025). A Survey on Knowledge-Oriented Retrieval-Augmented Generation. arXiv:2503.10677.
- [20] Klesel, M., Buxmann, P. (2025). Retrieval-Augmented Generation (RAG). *Business & Information Systems Engineering (Springer)*, 67, 1-7.
- [21] Yang, R., Zhu, C., et al. (2025). RAGVA: Engineering Retrieval-Augmented Generation-based Applications. *Journal of Systems and Software (Elsevier)*, in press.
- [22] Hindi, M., Mohammed, L., Maaz, O., Alwarafy, A. (2025). Enhancing the Precision and Interpretability of RAG in Legal Technology: A Survey. *IEEE Access (accepted, early access)*.
- [23] Liu, Y., Ouyang, D., et al. (2023). MemGPT: Towards Long-Term Memory for LLMs. arXiv:2310.08590.
- [24] Wang, Z., Lin, X., et al. (2024). A Survey on Long-term Memory for Large Language Models. arXiv:2401.03462.
- [25] Slaoui, S. (2025). S-AI: A Sparse Artificial Intelligence System Orchestrated by a Hormonal MetaAgent and Context-Aware Specialized Agents. *International Journal of Fundamental and Modern Research (IJFMR)*, 1(2), 1-16. URL: <https://www.ijfmr.com/papers/2025/2/42035.pdf>
- [26] Slaoui, S. (2025). Bio-Inspired Architecture for Parsimonious Conversational Intelligence: The S-AI-GPT Framework. *IJAIA*, 16(4). URL: <https://airconline.com/abstract/ijaia/v16n4/16425ijaia03.html>
- [27] Slaoui, S. (2025). Bio-Inspired Hormonal Modulation and Adaptive Orchestration in S-AI-GPT. *IJAIA*, 16(4). URL: <https://airconline.com/abstract/ijaia/v16n4/16425ijaia04.html>
- [28] Slaoui, S. (2025). Memory Architecture in S-AI-GPT: From Contextual Adaptation to Hormonal Modulation. *IJAIA*, 16(5). URL: <https://airconline.com/abstract/ijaia/v16n5/16525ijaia03.html>
- [29] Slaoui, S. (2025). S-AI-NET: A Sparse AI Framework for Adaptive and Parsimonious Autonomous Networking. *Research Square*, DOI: 10.21203/rs.3.rs-4740886/v1.
- [30] Slaoui, S. (2025). S-AI-Cyber: A Bio-Inspired Hormonal Architecture for Real-Time Cyber-Defense. *Research Square*, DOI: 10.21203/rs.3.rs-4840888/v1.
- [31] Alwarafy, A., Al-Thelaya, K., et al. (2023). Zero Trust Security for 6G Networks: A Comprehensive Survey. *IEEE Access*, 11, 128418-128451.
- [32] Samed, M., et al. (2025). Explainable Artificial Intelligence in Intrusion Detection Systems: From Traditional Tools to Large Language Models. *Engineering Applications of Artificial Intelligence (Elsevier)*, 140, 108043.
- [33] Zhao, Y., et al. (2025). Out-of-Distribution Detection Based on Non-Semantic Exploration Learning. *Information Sciences (Elsevier)*, 671, 120636.
- [34] Chandola, D., et al. (2025). White-Box XAI Models for Network Intrusion Detection Systems. *EURASIP Journal on Information Security (SpringerOpen)*, 2025:6.
- [35] Zhang, Y., et al. (2025). Conformal Out-of-Distribution Detection for Multivariate Time-Series. *Applied Intelligence (Springer)*, 55, 17802-17816.

- [36] Henao, F., et al. (2025). AI in power systems: a systematic review of key matters of debate. *Energy Informatics (SpringerOpen)*, 8, 15.

## **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.

