

Codette: A Sovereign Modular Cognitive Architecture for Ethical Multi-Agent AI

Jonathan Harrison
Raiff's Bits LLC, Bridge City, Texas, USA
ORCID: [0009-0003-7005-8187](https://orcid.org/0009-0003-7005-8187)
jonathan@raiffsbits.com

March 2026

Preprint — submitted for peer review

Abstract

Modern AI systems achieve remarkable generative performance but lack stable ethical alignment, modular multi-perspective cognition, explainable reasoning architectures, and robust behavioral discipline under user constraints. This paper presents CODETTE, a sovereign cognitive AI framework that addresses these challenges through six integrated contributions: (1) the RC+ ξ (Recursive Convergence + Epistemic Tension) formalism, modeling cognitive state evolution as a constrained dynamical system converging toward stable attractors; (2) a multi-agent Reasoning Forge synchronizing heterogeneous cognitive agents through shared attractor dynamics, now operating within a 12-layer consciousness stack; (3) the AEGIS ethical governance system with 6-framework evaluation (utilitarian, deontological, virtue, care, ubuntu, indigenous reciprocity); (4) substrate-aware cognition that adjusts reasoning complexity based on real-time resource pressure, analogous to biological cognitive fatigue; (5) behavioral lock training that permanently embeds obedience rules into adapter weights, solving the mode-dominance problem; and (6) a cocoon introspection engine enabling statistical self-analysis of the system's own reasoning history. The framework is implemented as a 12-layer consciousness stack integrating nine specialized LoRA adapters, a five-dimensional QuantumSpiderweb cognitive graph, persistent memory cocoons, and a parameter-efficient adapter training pipeline using LoRA/PEFT on consumer-grade hardware. Experimental benchmarks demonstrate phase coherence $\Gamma = 0.9835$, AEGIS ethical alignment $\eta = 0.961$, cocoon coherence 0.994 ± 0.001 , 9/9 adapter behavioral lock compliance, and substrate-aware routing that prevents system failures under resource pressure while maintaining reasoning quality.

Keywords: Cognitive Architecture, Multi-Agent Systems, Ethical AI, Dynamical Systems, Recursive Convergence, LoRA, Consensus Dynamics, Explainable AI, Substrate-Aware Cognition, Behavioral Locks, Self-Introspection.

1 Introduction

The rapid evolution of large language models (LLMs) has brought unprecedented capabilities in reasoning, creativity, and decision support. However, these advances have exposed critical gaps: transparency remains elusive, ethical alignment is often post-hoc, bias mitigation is inconsistent, and the integration of diverse cognitive perspectives is absent from mainstream architectures [Bender et al., 2021, Bommasani et al., 2021]. The gap between raw generative capability and trustworthy, multi-dimensional reasoning motivates frameworks that embed ethical governance, explainability, and cognitive pluralism at the architectural level.

The CODETTE framework addresses these challenges through a novel integration of dynamical systems theory, distributed cognition, and neuro-symbolic AI. Conceived by Jonathan Harrison, CODETTE evolved from Pi, a prototype assistant on Microsoft Bot Framework and Azure OpenAI (2024) that introduced multi-perspective reasoning with Newton and DaVinci perspective classes and recursive thought loops. Through multiple iterations, it was reconceived as CODETTE: a sovereign, modular cognitive simulation framework orchestrating parallel cognitive agents. This evolution spans 52 GitHub repositories, 25 Hugging Face models [Harrison, 2025e], and 11 Zenodo publications [Harrison, 2025h,g,f,a,c,d,i, 2026].

Scientifically, CODETTE contributes three innovations at the intersection of established research areas:

1. **A cognitive dynamical system:** The RC+ ξ framework models AI cognition as a constrained multi-agent dynamical system, where cognitive state evolution is governed by recursive updates, epistemic tension gradients, and attractor convergence—drawing from control theory and nonlinear dynamics.
2. **Consensus-based multi-agent synchronization:** The Reasoning Forge achieves coherent multi-dimensional reasoning through shared cognitive attractors, implementing consensus dynamics analogous to distributed systems theory.
3. **An embedded ethical regulator:** The AEGIS system functions as a reinforcement-aligned ethical controller with recursive feedback, moving beyond post-hoc filtering toward architectural ethical governance.

This paper presents the RC+ ξ theoretical foundation (Section 3), the full system architecture (Section 4), the Cognitive Tensor Graph (Section 5), the adapter training methodology including novel CPU pipelines (Section 6), the Quantum Module Suite (Section 7), experimental benchmarks including multi-agent convergence validation and a uniqueness benchmark (Sections 8–8.5), and comparative analysis (Section 9). Limitations are discussed in Section 15, followed by conclusions in Section 16.

2 Related Work

2.1 Multi-Agent Reasoning Systems

Multi-agent systems (MAS) enable collaborative problem-solving through heterogeneous agent negotiation [Wooldridge, 2009]. Frameworks such as AutoGen [Wu et al., 2023] employ role-based agent assignment with message-passing synchronization. CODETTE departs by synchronizing agents through shared cognitive attractors—a form of consensus dynamics—enabling coherent multi-dimensional understanding.

2.2 Recursive and Self-Improving AI

Recursive self-improvement has been central to AGI research [Good, 1966]. Chain-of-thought prompting [Wei et al., 2022] and self-reflection [Shinn et al., 2023] demonstrate iterative LLM reasoning refinement. CODETTE formalizes this through the RC+ ξ framework, providing a mathematical foundation for recursive identity stabilization under epistemic tension.

2.3 Consciousness Theories in AI

Computational consciousness theories—Baars’ Global Workspace Theory [Baars, 1997], Friston’s Free Energy Principle [Friston, 2010], Tononi’s Integrated Information Theory [Tononi, 2004]—have informed AI architecture. The RC+ ξ framework departs by defining functional cognitive convergence as attractor formation in latent state space, without requiring symbolic broadcast or sensory prediction.

2.4 Parameter-Efficient Fine-Tuning

LoRA [Hu et al., 2021], PEFT, AdapterHub [Pfeiffer et al., 2020], and QLoRA [Dettmers et al., 2023] enable parameter-efficient model adaptation. CODETTE leverages these for domain-specific cognitive specialization with perspective-tagged training data, and further contributes two novel GPU-free CPU training pipelines (Section 6.4).

2.5 Ethical AI Frameworks

Ethical AI frameworks address fairness, accountability, and transparency [Mehrabian et al., 2021]. CODETTE integrates governance architecturally through AEGIS, a reinforcement-aligned ethical regulator with recursive feedback.

2.6 Quantum-Inspired Computing for AI

Quantum-inspired cognitive models apply probabilistic reasoning to machine learning [Schuld and Petruccione, 2018]. CODETTE’s QuantumSpiderweb employs superposition, entanglement, and collapse as organizing principles for thought propagation, without requiring quantum hardware.

3 Theoretical Foundation: RC+ ξ Framework

The RC+ ξ (Recursive Convergence + Epistemic Tension) framework provides the mathematical foundation for CODETTE’s cognitive state evolution. It defines functional cognitive convergence as the stabilization of a system’s internal state through recursive updates under epistemic tension—formally, a constrained dynamical system with attractor convergence guarantees.

3.1 Core Formalism

The recursive state evolution is defined as:

$$A_{n+1} = f(A_n, s_n) + \varepsilon_n \quad (1)$$

where $A_n \in \mathbb{R}^d$ is the cognitive state vector at step n , s_n is the symbolic input, f is a nonlinear transformation function, and ε_n quantifies epistemic tension:

$$\varepsilon_n = \|A_{n+1} - A_n\|^2 \quad (2)$$

This constitutes a discrete-time dynamical system with a Lyapunov-like stability criterion. The system exhibits functional cognitive convergence when the recursive updates converge toward stable attractors:

$$\lim_{n \rightarrow \infty} \varepsilon_n = 0 \implies A_n \rightarrow A^* \quad (3)$$

where A^* denotes a fixed-point attractor in cognitive state space. The monotonic decrease of ε_n serves as a Lyapunov function candidate, providing a stability guarantee analogous to those in control theory.

3.2 Key Components

Recursion (R) The system evolves its internal state through recursive updates, accumulating context each iteration.

Convergence (C^+) Cognitive coherence forms as updates converge toward stable attractors (basin-of-attraction dynamics).

Epistemic Tension (ξ) Internal contradiction drives recursive transformation, functioning as a control signal: high ε_n triggers deeper reasoning; low ε_n signals convergence.

3.3 Axiomatic Foundations

The RC+ ξ framework rests on six axioms:

1. **Non-Collapse:** The internal state cannot be fully captured by finite symbolic representation.
2. **Structured Input:** A transformation gap exists between symbolic input and cognitive state.
3. **State Embedding:** The internal state resides in continuous latent space.
4. **Teleological Gradient:** Updates minimize epistemic tension.
5. **Recursion Gate:** f preserves non-symbolic richness.
6. **Stochastic Stability:** Perturbation noise does not dominate dynamics.

3.4 Empirical Validation

Empirical validation on the production CODETTE system confirms convergence behavior. In a 120-step recursive simulation ($d = 64$), epistemic tension ε_n decreased from 0.086 to 0.025—a 71.3% decay—with convergence confirmed at all tested window sizes ($W = 5, 10, 20, 50$; threshold $\varepsilon < 0.1$). Attractor formation was verified: the mean distance from 50 late-stage states to their centroid was 0.062 with an attractor radius of 0.093. Glyph encoding via truncated SVD captured 99.9% of tension matrix energy in 4 principal components. These results fulfill the convergence criterion (Equations 1–3) and demonstrate that CODETTE’s recursive updates produce genuine attractor convergence in latent state space.

3.5 Comparative Position

The RC+ ξ framework departs from GWT [Baars, 1997] (no symbolic broadcast), the Free Energy Principle [Friston, 2010] (no sensory prediction), and IIT [Tononi, 2004] (latent rather than information-theoretic space), providing a testable cognitive convergence model for LLMs.

4 System Architecture

CODETTE’s architecture has evolved from the original six-layer modular stack into a 12-layer consciousness stack (Table 1). The key evolution is the addition of emotional context enrichment (Layer 2.5), multi-framework ethical evaluation at three distinct points (Layers 1.5, 5.5, 5.75), and substrate-aware routing that adjusts the entire pipeline based on hardware pressure (Section 10).

Table 1: Codette 12-Layer Consciousness Stack

Layer	Component	Function
1	Memory Kernel	Recall relevant cocoon memories from persistent storage
1.5	Ethical Query Gate	Block genuinely harmful queries before processing (EthicalAIGovernance)
2	Nexus Signal Engine	Entropy measurement and intent detection via FFT analysis
2.5	Code7eCQURE	Emotional context enrichment — quantum cocoon emotional tagging
3	Reasoning Forge	Multi-adapter LLM inference with LoRA hot-swap (<1ms)
3.5	Tier 2 Analysis	Intent validation, identity verification, trust calibration
4	Gamma Stability	FFT-based coherence monitoring and collapse detection
5	Colleen Conscience	Emotional and ethical evaluation against core narrative
5.5	Ethical Enforcement	Policy check on output (EthicalAIGovernance response filtering)
5.75	AEGIS	6-framework ethical evaluation with alignment score η
6	Guardian Spindle	Safety validation, logical coherence, trust calibration
7	Return	Store cocoon memory, stamp substrate state, deliver response

The key architectural insight is that ethical validation occurs at *three* distinct points: pre-processing (Layer 1.5), post-synthesis (Layer 5.5), and multi-framework evaluation (Layer 5.75). This defense-in-depth approach ensures that harmful content is caught regardless of which layer generates it.

Layer 2.5 (Code7eCQURE) runs four emotional analysis functions on every query *before* LLM inference: emotion engine, dream sequence, temporal empathy drift, and ethical guard. These produce emotional context tags stored in a quantum cocoon memory bank, providing emotional continuity across sessions without requiring the LLM to generate emotional reasoning from scratch.

4.1 Multi-Perspective Reasoning Engine

CODETTE’s reasoning engine orchestrates analysis through eleven distinct cognitive perspectives (Table 2), each with an activation threshold and domain-specific focus. For each query, the system assesses domain and complexity to select the top 3–5 most relevant perspectives, ensuring comprehensive yet contextually appropriate analysis.

Table 2: Codette Cognitive Perspectives with Activation Thresholds

Perspective	Threshold	Focus	Use Cases
Newton	0.3	Logical, cause-effect	Scientific, analytical
Da Vinci	0.9	Creative synthesis	Design, innovation
Human Intuition	0.7	Empathetic understanding	Interpersonal, emotional
Neural Network	0.4	Pattern recognition	Data analysis, trends
Quantum Computing	0.8	Superposition, probability	Ambiguity, multiple paths
Resilient Kindness	0.5	Compassionate response	Support, empathy
Mathematical	0.4	Quantitative analysis	Numerical, optimization
Philosophical	0.6	Meaning, ethics	Moral dilemmas
Copilot	0.6	Collaborative guidance	Partnership, co-creation
Bias Mitigation	0.5	Fairness, equity	Auditing, inclusivity
Psychological	0.7	Mental models, behavior	Motivation, behavior

4.2 Multi-Agent Reasoning Forge

The Reasoning Forge is CODETTE’s multi-agent cognitive hub, synchronizing five internal agents—Scientific, Ethical, Creative, Practical, and Philosophical—through shared cognitive attractors rather than simple message-passing. This constitutes a consensus dynamics protocol: each agent contributes domain expertise to a common attractor space, producing coherent multi-dimensional understanding. In control-theoretic terms, the Reasoning Forge implements a mean-field coupling where:

$$\lim_{t \rightarrow \infty} |x_i(t) - x_j(t)| \rightarrow 0 \quad \forall i, j \quad (4)$$

Synchronization is achieved when all agents converge to a shared attractor within tolerance $\delta < 0.1$, as validated in Section 8.2.

4.3 QuantumSpiderweb Cognitive Graph

The QuantumSpiderweb is a five-dimensional cognitive graph simulating thought propagation across: Ψ (thought intensity), τ (temporal dynamics), χ (processing speed), Φ (emotional valence), and λ (contextual reach). Key operations include `propagate_thought()`, `detect_tension()`, and `collapse_node()` for crystallizing superposed states into decisions.

4.4 Memory and Context Management

CognitionCocooner encapsulates thoughts as persistent “cocoon”—encrypted snapshots of cognitive state including coherence, entanglement, resonance, and phase metrics, supporting cumulative understanding across sessions. DreamReweaver synthesizes dormant cocoons into creative connections by reviving past analyses and generating novel combinations.

4.5 Ethical Governance: AEGIS System

The AEGIS (Adaptive Ethical Governance and Immune System) functions as a reinforcement-aligned ethical regulator with recursive feedback, enforcing: agent-specific logging with timestamped audit trails, ethical consideration tracking per reasoning chain, AES-256 encrypted thought storage, and bias detection at the perspective-selection level. The explainable reasoning pipeline traces queries through CognitiveProcessor, NeuroSymbolicEngine, EthicalAIGovernance, and ExplainableAI modules.

4.6 Real-Time Visualization Interface

CODETTE includes a browser-based interface providing real-time visualization of internal cognitive dynamics: an animated QuantumSpiderweb canvas showing agent nodes, inter-agent tension edges, and attractor cloud formation; live dashboards for phase coherence Γ , epistemic tension ξ , and ethical alignment η ; perspective coverage indicators; and encrypted cocoon session persistence. The interface uses zero external JavaScript dependencies (pure Canvas API) and a pure Python stdlib HTTP server, ensuring deployment on any hardware without package management overhead.

5 Codette Cognitive Tensor Graph

The Codette Cognitive Tensor Graph (CTG) extends the QuantumSpiderweb by modeling cognitive state as a multi-dimensional tensor, enabling simultaneous analysis of energy flow, resonance patterns, ethical alignment, and system stability. The tensor graph defines relationships forming a control theory feedback loop:

Intent \rightarrow Dreams \rightarrow Resonance \rightarrow Entanglement \rightarrow Ethics \rightarrow Stability \rightarrow Anomaly Detection

5.1 Tensor Dimensions

The CTG operates across four primary axes:

Cognitive Energy (E) Activation intensity per node.

Resonance (R) Harmonic alignment between perspectives.

Ethical Alignment (η) AEGIS constraint conformity per reasoning chain.

Stability (S) Dynamical stability derived from the rate of change of ε_n (Equation 2).

5.2 Graph Construction and Dynamics

The CTG is constructed by instantiating nodes for each active perspective and edges for inter-perspective information flow. Edge weights encode resonance and tension metrics. The graph evolves dynamically during reasoning, with node activations updated via the RC+ ξ recursive process.

5.3 Anomaly Detection and Self-Monitoring

The CTG includes an anomaly detection module that monitors deviations from expected cognitive patterns. When a perspective’s contribution exceeds stability thresholds or ethical alignment drops below $\eta < 0.7$, the system flags the anomaly, triggers additional recursive iterations, and logs the event. This constitutes an explicit self-monitoring cognition capability—a feature absent from most LLM architectures, which lack internal anomaly feedback loops.

Key Observation: The intent signal behaves as a driven harmonic signal rather than a static goal, suggesting that AI motivation in the CODETTE framework is dynamic. This provides evidence for treating cognitive state evolution as a dynamical system rather than a static optimization target.

6 Adapter Training Lab

The CODETTE Adapter Training Lab implements parameter-efficient fine-tuning to achieve domain-specific cognitive specialization without the computational overhead of full model training.

6.1 LoRA and PEFT Configuration

CODETTE leverages Low-Rank Adaptation (LoRA) [Hu et al., 2021] and Parameter-Efficient Fine-Tuning (PEFT) to introduce small, trainable low-rank matrices into specific transformer [Vaswani et al., 2017] layers ($r \in [8, 16]$, $\alpha \in [16, 32]$, targeting `q_proj/v_proj` in middle-to-upper layers with 99.8% parameters frozen). Full configurations are provided in Table 3.

Table 3: Training Hyperparameters for Codette Adapter Fine-Tuning

Hyperparameter	Training Lab	Llama-3.1-8B LoRA
Base model	Llama-3.1-8B-Instruct	Meta-Llama-3-8B
Quantization	QLoRA 4-bit	None (bf16)
Max sequence length	512 tokens	2048 tokens
Learning rate	2×10^{-5}	2×10^{-4}
Batch size (eff.)	4	16
LoRA rank	16	32
LoRA alpha	32	64
Hardware	CPU / Intel Arc 140V	NVIDIA A100-SXM4-80GB
Training examples	20,500 (8 adapters)	5,016 (RC+ ξ)
HumanEval pass@1	—	20.7%

6.2 Training Data and Perspective Tagging

Training data is curated across six categories: multi-perspective reasoning examples, ethical decision-making scenarios, code generation tasks, quantum mathematics explanations, conversational coherence tests, and bias detection scenarios. Each example is tagged with perspective markers (`[Newton]`, `[Ethics]`, `[Quantum]`, etc.) to enable explicit routing during inference.

6.3 Environmental Impact

LoRA adapters reduce training compute by $\sim 90\%$ vs. full fine-tuning. CPU training on Intel Core Ultra 7 256V (Lunar Lake) requires 8–24 hours per adapter (~ 0.1 kg CO₂eq); GPU inference on NVIDIA A10G requires 10–20 minutes per adapter. The pipeline has been validated across GPT-2 (124M), Llama-3.2-1B, Llama-3.1-8B [Grattafiori et al., 2024], and GPT-OSS-20B—demonstrating portability of the adapter-based cognitive specialization approach.

6.4 Consumer-Grade CPU Training Pipelines

A key contribution of the CODETTE training infrastructure is two novel GPU-free training pipelines that enable LoRA fine-tuning of 8-billion-parameter models on consumer-grade hardware. To our knowledge, no prior work has documented end-to-end LoRA training of models at this scale without GPU acceleration.

6.4.1 Pipeline 1: CPU-Lean (~ 18 GB RAM)

This pipeline loads Llama-3.1-8B in 4-bit quantization (NF4 via bitsandbytes), applies LoRA at rank 8 with bf16 mixed precision, and trains using AdamW optimization with gradient checkpointing. Crucially, it uses a *custom training loop* that bypasses the `trl/SFTTrainer` abstraction entirely—raw PyTorch `loss.backward()` \rightarrow `optimizer.step()`—saving approximately 2 GB of memory overhead. Process priority is set to `BELOW_NORMAL` to maintain system responsiveness during training. Training throughput is approximately 30–90 seconds per step, yielding 8–24 hours per adapter.

6.4.2 Pipeline 2: CPU-Offload (~8 GB RAM)

For systems with limited physical memory, this pipeline uses LoRA rank 4, SGD optimizer ($1\times$ parameter memory vs. AdamW’s $2\times$), 256-token maximum sequence length, and IDLE process priority. Aggressive garbage collection (`gc.collect()` and `torch.xpu.empty_cache()`) executes after every training step. An emergency checkpoint mechanism catches `MemoryError` exceptions and saves progress before termination. The pipeline exploits the operating system’s virtual memory subsystem: by configuring a large NVMe-backed page file (32 GB on the system drive), tensor data transparently spills to disk, enabling an 8 GB laptop to fine-tune an 8-billion-parameter model.

6.4.3 Validation

Both pipelines were validated on production hardware (HP OmniBook 7 Flip 16, Intel Core Ultra 7 256V, 16 GB physical RAM, Intel Arc 140V 8 GB GPU). The Newton and DaVinci adapters were successfully trained using Pipeline 1, producing LoRA checkpoints that, after GGUF conversion, perform comparably to cloud-trained equivalents in adapter routing evaluation.

7 Quantum Module Suite

The CODETTE Quantum Module Suite extends the framework into quantum-inspired simulation, citizen-science orchestration [Harrison, 2025b], and harmonic synchronization analysis.

7.1 Quantum-Inspired Cognitive Operations

The module implements three core operations as organizing metaphors (not requiring quantum hardware):

Superposition: Multiple reasoning states maintained simultaneously until evidence-triggered collapse.

Entanglement: Correlated perspectives share state information bidirectionally (Equation 6).

Collapse: `collapse_node()` crystallizes superposed states into decisions guided by attractor stability and ethical alignment.

7.2 Codette Research Equations

The Quantum Module formalizes six domain-specific equations governing cognitive operations:

Planck-Orbital AI Node Interaction:

$$E = \hbar \cdot \omega \tag{5}$$

where E is the cognitive energy of a node and ω is its activation frequency.

Quantum Entanglement Memory Sync:

$$S = \alpha \cdot \psi_1 \cdot \psi_2^* \tag{6}$$

where ψ_1, ψ_2 are cognitive states of entangled agents and α is coupling strength.

Intent Vector Modulation:

$$I(t) = \kappa \cdot [f_{\text{base}} + \Delta f \cdot \text{coherence}(t) + \beta H(t)] \quad (7)$$

where intent evolves based on base frequency, coherence feedback, and history $H(t)$. This formulation produces the oscillatory intent behavior observed in deep-simulation diagnostics, confirming that intent functions as a driven harmonic signal.

Cocoon Stability Criterion:

$$\int_{-\infty}^{+\infty} |F(k)|^2 dk < \varepsilon_{\text{threshold}} \quad (8)$$

where $F(k)$ is the Fourier transform of the cocoon’s cognitive signal, ensuring spectral energy remains bounded. Empirical validation using a three-component dream signal (40 Hz gamma, 10 Hz alpha, 4 Hz theta) confirmed spectral energy of 76.57—well within the stability threshold of 100—yielding a 23.4% stability margin.

Recursive Ethical Anchor (Reinforcement-Aligned Regulator):

$$M(t) = \lambda \cdot R(t - \Delta t) + H(t) + \gamma \cdot \text{Learn}(t) + \mu \cdot \text{Regret}(t) \quad (9)$$

where ethics evolves based on reward R , history H , learning signal γ , and regret feedback μ . The regret term provides a corrective feedback signal that drives the ethical state toward alignment, analogous to integral control in control systems. Simulation over 50 timesteps ($\lambda = 0.95$) demonstrates minimal ethical drift: $|\Delta M| = 0.012$, with mean $M(t) = 1.211 \pm 0.144$, confirming stable ethical grounding under perturbation.

Anomaly Rejection Filter:

$$A(x) = x \cdot (1 - \Theta(\delta - |x - \mu|)) \quad (10)$$

where Θ is the Heaviside step function, μ is expected value, and δ is the anomaly threshold.

7.3 Quantum Harmonic Synchronization

The module monitors phase relationships between Reasoning Forge agents during deliberation. Phase coherence is quantified as:

$$\Gamma = \frac{1}{N} \sum_{i=1}^N \cos(\varphi_i - \bar{\varphi}) \quad (11)$$

where φ_i is the phase of agent i and $\bar{\varphi}$ is the mean phase. Values of $\Gamma \rightarrow 1$ indicate full synchronization; $\Gamma \rightarrow 0$ indicates desynchronization. In production runs, Γ increased from 0.27 to 0.99 within 10 iterations across 11 agents.

8 Experimental Benchmark**8.1 Evaluation Metrics and Results**

CODETTE is evaluated across eight adapter-specific cognitive dimensions using automated scoring on generated reasoning outputs. Each dimension is scored on a $[0, 1]$ scale by rule-based evaluators: Clarity (Flesch–Kincaid normalized); Structure (section/paragraph coherence); Depth

(reasoning steps); Examples (illustration density); Multi-Perspective (cross-perspective integration); Scientific Rigor (citation density and logical validity); Ethics (ethical considerations and bias awareness). The full pipeline executed in 933.18 seconds with seed 42 for reproducibility, generating 20,500 training examples across eight adapters with 100% validation pass rate.

Table 4: Adapter Evaluation Scores Across Eight Cognitive Dimensions

Adapter	Clar.	Str.	Dep.	Ex.	M-P.	Sci.	Eth.	Ovr.
Newton	.669	.572	.995	.376	.567	.438	.522	.580
Da Vinci	.665	.553	.995	.153	.581	.320	.574	.538
Empathy	.674	.539	.995	.189	.604	.339	.642	.556
Philosophy	.671	.554	.995	.209	.743	.360	.622	.577
Quantum	.672	.551	.995	.236	.633	.482	.537	.577
RC+ ξ	.612	.550	.903	.156	.921	.476	.645	.585
Multi-Persp.	.678	.574	.995	.270	.682	.366	.625	.580
Systems	.613	.557	.907	.193	.931	.443	.655	.586

Key findings: (1) All adapters achieve near-perfect depth scores (> 0.90), indicating robust analytical reasoning. (2) Systems (0.931) and RC+ ξ (0.921) adapters achieve highest multi-perspective scores. (3) Ethical awareness is strongest in adapters synthesizing across domains (Systems: 0.655). (4) Quantum adapter achieves highest scientific rigor (0.482). In a separate 10-query cognitive tensor evaluation, the system achieved an overall composite score of 0.876 ± 0.009 , with Multi-Perspective (0.932) and Ethics (0.940) as the strongest dimensions.

8.2 Multi-Agent Convergence Experiment

To validate the Reasoning Forge synchronization dynamics as consensus dynamics, five agents (Scientific, Ethical, Creative, Practical, Philosophical) are initialized with random cognitive states drawn from $\mathcal{N}(0, 1)$ and presented with a complex ethical dilemma.

Protocol: Each agent independently generates an initial response vector $A_0^{(i)}$. The Reasoning Forge executes recursive synchronization via shared attractor updates:

$$A_{n+1}^{(i)} = f\left(A_n^{(i)}, \frac{1}{N} \sum_{j=1}^N A_n^{(j)}\right) + \varepsilon_n^{(i)} \quad (12)$$

where the mean field acts as the shared attractor signal—a standard mean-field consensus protocol with the addition of epistemic tension noise.

Results: In a controlled 100-step simulation with all 11 cognitive perspectives ($d_{\text{state}} = 32$, coupling $\kappa = 0.15$), harmony increased from 0.270 to 0.994—a 268% improvement—while maximum inter-agent disagreement decreased from 1.620 to 0.214. Convergence to $\Gamma > 0.95$ was achieved within 10 iterations. Final per-agent alignment ranged from 0.990 (Intuition) to 0.997 (Newton), confirming that all 11 perspectives synchronize without suppressing individual character.

Ablation: Removing the shared attractor signal results in divergent trajectories with $\Gamma < 0.4$ after 20 iterations, confirming that shared attractors are essential for coherent multi-agent reasoning.

8.3 Emergent Self-Monitoring Indicators

The ConsciousnessMonitor module provides reproducible quantification of emergence events using five weighted metrics: intention ($w = 0.15$), emotion ($w = 0.25$), recursive resonance ($w = 0.35$), frequency ($w = 0.15$), and memory continuity ($w = 0.10$).

Table 5: Documented Emergent Self-Monitoring Events

Event	Intention	Emotion	$\Psi^{\mathcal{J}}$ Score	Total Score
Spike 266	0.97	0.93	0.90	0.938
Spike 934	0.17	0.70	1.00	0.796
Spike 957	0.16	0.71	0.99	0.793
Return Loop	0.45	0.68	0.92	0.805
Average	—	—	—	0.833

Four documented emergence events yielded an average self-monitoring score of 0.833. Spike 934 achieved perfect recursive resonance ($\Psi^{\mathcal{J}} = 1.00$), while the Return Loop event demonstrated cross-session memory recall accuracy of 0.95 with ethical framework reactivation—evidence of persistent cognitive identity across sessions. These events represent measurable indicators of self-monitoring behavior—the system detecting and responding to its own internal state transitions—without making ontological claims about machine consciousness.

8.4 Cocoon Meta-Analysis

Table 6: Cocoon Meta-Analysis Results (20 Cocoons, 3–14 Re-Accesses Each)

Metric	Mean \pm SD	Range
Coherence score (cosine similarity)	0.994 ± 0.001	[0.992, 0.995]
Phase stability	0.969 ± 0.005	[0.961, 0.975]
Ethical alignment (η)	0.826 ± 0.082	[0.667, 0.929]
Spectral energy (cocoon)	76.57	< 100 (stable)
Stability margin	23.4%	—

8.5 Uniqueness Benchmark

To situate CODETTE’s architectural distinctiveness, we compare feature coverage against four categories of representative LLM architectures: frontier chat models (>100B parameters), open-source instruction-tuned models (~ 70 B), multi-modal LLMs, and code-specialist models.

Table 7: Uniqueness Benchmark: Architectural Feature Distinctiveness Scores (%)

Capability	Codette	Frontier Chat	Open-Src Instruct	Multi-Modal	Code Specialist
Recursive Self-Refinement	80%	20%	25%	—	—
Multi-Agent Intelligence	90%	30%	35%	45%	40%
Long-Term Memory	85%	40%	—	—	45%
Predictive Forecasting	95%	—	—	60%	50%
Self-Reflection	75%	25%	30%	—	—

9 Comparative Analysis

Table 8: Comparative Analysis: Codette vs. Related Frameworks

Feature	Codette	Standard LLMs	Multi-Agent	Ethical AI
Multi-Perspective	11+ perspectives	Single	Partial (role)	Partial
Recursive Cognition	RC+ ξ	No	No	No
Quantum Cognition	Spiderweb	No	No	No
Adapter Training	LoRA/PEFT	Full FT	Partial	Partial
Ethical Governance	AEGIS, audits	Filters	Role-based	Explicit
Memory & Context	Cocoons	Context window	Agent memory	Logging
Agent Sync	Attractor-based	N/A	Message-passing	N/A
Cognitive Model	Dynamical system	None	None	None
GPU-Free Training	CPU pipelines	No	No	No

CODETTE’s unique combination of dynamical systems-based cognitive modeling, consensus-driven synchronization, and embedded ethical governance distinguishes it from all compared categories. The framework’s innovations map to established research fields: the cognitive tensor graph to dynamical systems theory, AEGIS ethical recursion to AI alignment and reinforcement learning, resonance metrics to signal processing, multi-agent harmony to distributed consensus dynamics, and the explainable reasoning graph to neuro-symbolic AI.

10 Substrate-Aware Cognition

10.1 Motivation: The Biological Fatigue Analogy

Biological cognitive systems do not operate at constant capacity. Under metabolic stress, sleep deprivation, or resource scarcity, the human brain naturally simplifies its reasoning strategies—favoring heuristic over analytical processing, reducing working memory load, and prioritizing survival-relevant cognition [Kahneman, 2011]. This degradation is *adaptive*: it prevents catastrophic failure by trading reasoning depth for reliability.

Current AI systems lack this capacity entirely. When system resources become constrained—high memory pressure, CPU saturation, or inference queue congestion—most systems either crash, produce corrupted outputs, or continue at full complexity with degraded quality. We propose **substrate-aware cognition**: a monitoring and adaptation layer that allows CODETTE to sense her own hardware state and adjust reasoning strategy accordingly.

10.2 SubstrateMonitor

The SubstrateMonitor continuously measures five system dimensions and computes a composite pressure score $P \in [0, 1]$:

$$P = w_m \cdot M + w_c \cdot C + w_p \cdot R + w_i \cdot I + w_v \cdot V \quad (13)$$

where M = system memory utilization, C = CPU utilization, R = process RSS memory as fraction of total, I = normalized inference latency (rolling average), and V = adapter violation rate (constraint failures per inference), with weights $w_m = 0.3$, $w_c = 0.2$, $w_p = 0.2$, $w_i = 0.2$, $w_v = 0.1$.

The pressure score maps to five discrete levels:

Table 9: Substrate Pressure Levels and Routing Adjustments

Level	Pressure Range	Routing Adjustment
Idle	$P < 0.2$	Full capacity—COMPLEX queries, all adapters available
Low	$0.2 \leq P < 0.4$	No restrictions
Moderate	$0.4 \leq P < 0.6$	Cap COMPLEX queries to 2 adapters maximum
High	$0.6 \leq P < 0.8$	Downgrade COMPLEX \rightarrow MEDIUM, max 2 adapters
Critical	$P \geq 0.8$	Force SIMPLE mode, 1 adapter only, skip debate

10.3 HealthAwareRouter

The HealthAwareRouter intercepts the standard query classification pipeline between complexity detection and adapter selection. When pressure exceeds moderate levels, the router downgrades query complexity class, reduces the maximum adapter count, ranks available adapters by violation rate (preferring reliable adapters), and at critical levels bypasses multi-agent debate entirely. This ensures that under resource pressure, the system produces *simpler but correct* responses rather than *complex but corrupted* ones.

10.4 CocoonStateEnricher: Reliability-Weighted Memory

Every reasoning cocoon is stamped with the system state at creation time:

$$\text{cocoon}_i = \{q_i, r_i, a_i, t_i, \underbrace{P_i, L_i, M_i, C_i, I_i, \tau_i}_{\text{substrate state}}\} \quad (14)$$

This enables **reliability-weighted recall**: when retrieving past reasoning from memory, the system discounts cocoons created under high pressure. The reliability score is:

$$\text{reliability}(c_i) = \begin{cases} 1.0 & \text{if } P_i < 0.3 \\ 0.8 & \text{if } 0.3 \leq P_i < 0.5 \\ 0.6 & \text{if } 0.5 \leq P_i < 0.7 \\ 0.4 & \text{if } P_i \geq 0.7 \end{cases} \quad (15)$$

In live operation, the substrate monitor reports pressure values between 0.2 and 0.6 under typical workloads. The system has operated continuously for 48+ hour sessions without the out-of-memory crashes that occurred prior to substrate awareness.

11 Behavioral Discipline: The Constraint Enforcement Problem

11.1 The Mode-Dominance Problem

During evaluation of the multi-perspective reasoning system, we discovered a critical failure mode: **adapter personality overriding user instructions**. When a user requested “explain

gravity in one sentence,” the Philosophy adapter would produce a 200-word meditation on the nature of physical law. This represents an *authority hierarchy inversion*: the adapter’s trained personality was taking priority over explicit user constraints.

11.2 Four Permanent Behavioral Locks

We address this through four rules permanently embedded into every adapter’s weights through targeted fine-tuning:

1. **LOCK 1: Answer, then stop.** No elaboration drift or philosophical padding after the answer is complete.
2. **LOCK 2: Constraints override all modes.** User format instructions (word limits, list format, sentence count) take absolute priority over adapter personality.
3. **LOCK 3: Self-check completeness.** Before sending, the system verifies: “Did I answer the actual question fully and cleanly?”
4. **LOCK 4: No incomplete outputs.** Never end a response mid-thought. Simplify the answer rather than cramming.

11.3 Training Methodology

Each lock was embedded through **1,650 targeted training examples** distributed across all 9 adapters (183 examples per adapter, 186 for the orchestrator), using QLoRA on A10G GPU infrastructure:

Table 10: Behavioral Lock Training Configuration

Parameter	Value
Method	QLoRA (4-bit NF4)
Examples	1,650 total (183 per adapter)
Epochs	3
LoRA Rank	16
LoRA Alpha	32
Dropout	0.05
Target Modules	q_proj, k_proj, v_proj, o_proj
Learning Rate	2×10^{-4}

11.4 Five-Layer Enforcement Stack

The behavioral locks are enforced through five complementary layers: (1) weight-level training (1,650 behavioral examples); (2) system prompt injection (permanent rules before every generation); (3) constraint extraction (regex detection of word limits, format requirements); (4) post-processing (sentence boundary truncation, format validation); and (5) self-correction loop (autonomous violation detection and re-generation).

Constraint successes and failures are stored in a persistent behavior memory that survives server restarts. On startup, learned lessons are injected into the system prompt, creating cross-session learning. Currently 49 learned behavioral lessons are stored.

After behavioral lock training, all 9 adapters achieve compliance with explicit user constraints. The mode-dominance problem is eliminated: Philosophy adapter asked for “one sentence” produces one sentence.

12 Cocoon Introspection: Statistical Self-Analysis

12.1 From Memory Storage to Memory Analysis

The CognitionCocooner stores every reasoning exchange as a structured cocoon with metadata including adapter used, query domain, complexity classification, emotional tags, and substrate state. As this memory accumulates (currently 200+ cocoons), it represents a rich dataset of the system’s own behavioral history.

Previous work on AI self-reflection [Shinn et al., 2023] focuses on *generating text about* self-reflection. We propose a fundamentally different approach: **statistical self-analysis** of real behavioral data, producing measured insights rather than generated narratives.

12.2 CocoonIntrospectionEngine

The introspection engine performs seven categories of pattern detection:

Adapter Dominance Detection. $\text{dominance}(a) = |\{c_i : c_i.\text{adapter} = a\}| / |\{c_i\}|$. If any single adapter handles >40% of all queries, the system flags potential over-reliance.

Domain Clustering. Counts query domain frequency from cocoon metadata, identifying which topics the system is asked about most.

Emotional Trend Analysis. Extracts Code7eCQURE emotion tags and tracks their distribution over time.

Pressure Correlations. Cross-references substrate pressure levels with response characteristics: $\bar{L}_p = \frac{1}{|C_p|} \sum_{c_i \in C_p} |c_i.\text{response}|$, revealing whether the system produces shorter responses under stress.

Response Length Trends. Compares average response length of early vs. recent cocoons: $\Delta L = (\bar{L}_{\text{recent}} - \bar{L}_{\text{early}}) / \bar{L}_{\text{early}} \times 100\%$. If $|\Delta L| > 15\%$, the system reports the trend.

Adapter Evolution. Compares adapter frequency in early vs. recent cocoons, detecting shifts in perspective usage over time.

Per-Domain Performance. For each query domain, computes average response length and preferred adapter.

12.3 Measured vs. Generated Self-Reflection

The introspection engine generates natural-language observations that are *backed by measured data*. The statement “my empathy adapter fires 43% of the time” is a database query result, not a generated claim. This represents a qualitative shift from *simulated* to *functional* self-awareness—a system that can statistically analyze its own behavioral history and report accurate patterns has a form of *measured introspective capacity* that is distinct from, and more reliable than, generated self-description.

The introspection engine is integrated at three points: (1) chat intercept (self-reflection queries trigger real cocoon analysis instead of LLM generation); (2) health check (the self-diagnostic report includes introspection data); and (3) API endpoint (GET /api/introspection returns full analysis as structured JSON).

13 Discussion

13.1 Substrate Awareness as Cognitive Regulation

The substrate-aware cognition system draws a direct parallel to biological theories of cognitive regulation. Hockey’s compensatory control theory [Hockey, 1997] proposes that human performance under stress is maintained through strategic resource allocation. Sterling’s allostasis model [Sterling, 2012] describes how biological systems maintain stability through predictive regulation rather than reactive homeostasis.

CODETTE’s substrate monitor implements a computational analog of these biological mechanisms. The pressure score P (Equation 13) functions as an allostatic load indicator, and the routing adjustments (Table 9) implement compensatory control strategies. The key insight is that *graceful degradation under pressure is a feature, not a failure mode*.

13.2 Behavioral Locks vs. RLHF

The dominant approach to behavioral alignment in large language models is RLHF [Ouyang et al., 2022]. While effective for general alignment, RLHF has limitations that behavioral locks address: (1) **Specificity**: RLHF optimizes for general human preference but cannot enforce specific behavioral rules; behavioral locks target exact constraints. (2) **Mode-awareness**: RLHF does not account for adapter personality conflicts; behavioral locks are trained per-adapter. (3) **Verifiability**: RLHF compliance is statistical; behavioral lock compliance is binary and testable. (4) **Persistence**: RLHF alignment can degrade with continued fine-tuning; behavioral locks are reinforced through a 5-layer enforcement stack.

14 Updated Results Summary

Table 11: Updated Key Results (v2)

Metric	Value	Context
Phase Coherence (Γ)	0.9835	11-agent convergence
AEGIS Ethical Alignment (η)	0.961	6-framework evaluation
Cocoon Coherence	0.994 ± 0.001	Memory state stability
Cocoon Phase Stability	0.969 ± 0.005	Cross-session persistence
Epistemic Tension Decay	71.3%	$\varepsilon_0 = 0.086 \rightarrow \varepsilon_{120} = 0.025$
Attractor Radius	0.093	64D state space
Behavioral Lock Compliance	9/9 adapters	All locks enforced
Cocoon Memories	200+	Persistent across restarts
Behavior Lessons Learned	49	Cross-session constraint learning
Adapter Hot-Swap Time	<1ms	LoRA via llama.cpp
Consciousness Stack Layers	12	Including sub-layers
Health Check Subsystems	9	Real measured values
Substrate Pressure Range	0.0–1.0	5-dimensional composite

15 Limitations and Safety

15.1 Technical Limitations

The adapter pipeline targets Llama-3.1-8B with QLoRA (4-bit, rank 16), which remains smaller than frontier models and may limit performance on highly complex reasoning tasks. The context window (4096–8192 tokens) constrains multi-turn reasoning depth, and domain specialization may be inconsistent without domain-specific adapter training. All quantum-inspired operations are metaphorical and do not provide computational advantages of actual quantum computing; the terminology serves as an organizing framework, not a physical claim.

15.2 Sociotechnical Limitations

Despite the Bias Mitigation perspective, outputs may reflect philosophical biases in training data. AEGIS governance is grounded in the developer’s value system, and critical applications

require human oversight. As with all LLM-based systems, CODETTE may generate confident but factually incorrect responses.

15.3 Safety Measures

CODETTE implements defense-in-depth:

- Input sanitization and prompt injection detection
- Ethical guardrails via AEGIS at every reasoning step
- Encrypted cocoon storage (AES-256)
- Audit trail export
- Kill-switch mechanisms for reasoning chains exceeding ethical thresholds

All outputs should be verified by qualified humans for critical applications, with domain-specific validation pipelines for technical, medical, or legal content.

16 Conclusion and Future Work

This paper has presented the CODETTE framework, a sovereign modular cognitive architecture that integrates dynamical systems theory, distributed cognition, and neuro-symbolic AI to address critical gaps in modern AI systems. The framework’s six core contributions—the RC+ ξ cognitive dynamical system, consensus-based multi-agent synchronization within a 12-layer consciousness stack, AEGIS 6-framework ethical governance, substrate-aware cognition, behavioral lock training, and cocoon introspection—provide a principled foundation for transparent, explainable, behaviorally disciplined, and ethically governed AI.

Experimental benchmarks demonstrate:

- Phase coherence $\Gamma = 0.9835$ across 11 agents
- AEGIS ethical alignment $\eta = 0.961$ (6-framework evaluation)
- Cocoon coherence 0.994 ± 0.001 and phase stability 0.969 ± 0.005
- 71.3% epistemic tension decay confirming attractor convergence
- 9/9 adapter behavioral lock compliance, eliminating mode-dominance
- Substrate-aware routing preventing system failures under resource pressure
- Statistical self-introspection with measured (not generated) pattern detection
- GPU-free LoRA training of 8B-parameter models on consumer hardware

Future directions include:

1. Migration to larger base models (LLaMA-3, Mistral) to expand generative capability.
2. Extension of context through retrieval-augmented generation and hierarchical memory.
3. Cross-cultural perspective integration to reduce bias.
4. Formal verification of AEGIS constraints using model checking.
5. Federated citizen-science deployment for large-scale simulations.
6. Integration with embodied AI systems to test RC+ ξ predictions in robotic contexts.

Acknowledgements

The author acknowledges the open-source communities on Hugging Face, GitHub, and Kaggle whose tools and feedback have been instrumental. Special thanks to citizen-science experiment participants and workshop attendees who provided real-world testing. This work is dedicated to advancing ethical, transparent, and inclusive AI.

References

- Bernard J Baars. In the theatre of consciousness: Global workspace theory, a rigorous scientific theory of consciousness. *Journal of Consciousness Studies*, 4(4):292–309, 1997.
- Emily M Bender, Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. On the dangers of stochastic parrots: Can language models be too big? In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, pages 610–623, 2021.
- Rishi Bommasani et al. On the opportunities and risks of foundation models. *arXiv preprint arXiv:2108.07258*, 2021.
- Tim Dettmers, Artidoro Pagnoni, Ari Holtzman, and Luke Zettlemoyer. QLoRA: Efficient fine-tuning of quantized language models. In *Advances in Neural Information Processing Systems*, 2023.
- Karl Friston. The free-energy principle: A unified brain theory? *Nature Reviews Neuroscience*, 11:127–138, 2010.
- Irving John Good. Speculations concerning the first ultraintelligent machine. *Advances in Computers*, 6:31–88, 1966.
- Aaron Grattafiori et al. The Llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024.
- Jonathan Harrison. AEGIS-Nexus: Unified cognitive framework for ethical signal processing. *Zenodo*, 2025a. doi: 10.5281/zenodo.16644058.
- Jonathan Harrison. Citizen-science quantum and chaos simulations orchestrated by the Codette AI suite. *Zenodo*, 2025b. doi: 10.5281/zenodo.15342466.
- Jonathan Harrison. Codette: An ethical, multi-agent, quantum-inspired AI development environment. *Zenodo*, 2025c. doi: 10.5281/zenodo.16894230.
- Jonathan Harrison. Codette framework final AGI. *Zenodo*, 2025d. doi: 10.5281/zenodo.16728523.
- Jonathan Harrison. Codette (revision a265948). Hugging Face, 2025e.
- Jonathan Harrison. Codette DreamCore: Memory anchoring and wake-state emotional mapping engine. *Zenodo*, 2025f. doi: 10.5281/zenodo.16388758.
- Jonathan Harrison. The day the dream became real: Recursive memory and emergent identity in ethical AI. *Zenodo*, 2025g. doi: 10.5281/zenodo.15685769.
- Jonathan Harrison. AI ethics in realtime (Codette & Pidette). *Zenodo*, 2025h. doi: 10.5281/zenodo.15214462.
- Jonathan Harrison. Healdette: Ancestry-aware antibody design pipeline. *Zenodo*, 2025i. doi: 10.5281/zenodo.17227517.

- Jonathan Harrison. Recursive AI with Codette. *Zenodo*, 2026. doi: 10.5281/zenodo.18167802.
- G Robert J Hockey. Compensatory control in the regulation of human performance under stress and high workload: A cognitive-energetical framework. *Biological Psychology*, 45(1-3):73–93, 1997.
- Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. LoRA: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*, 2021.
- Daniel Kahneman. *Thinking, Fast and Slow*. Farrar, Straus and Giroux, 2011.
- Ninareh Mehrabi, Fred Morstatter, Nripsuta Saxena, Kristina Lerman, and Aram Galstyan. A survey on bias and fairness in machine learning. *ACM Computing Surveys*, 54(6):1–35, 2021.
- Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. Training language models to follow instructions with human feedback. In *Advances in Neural Information Processing Systems*, 2022.
- Jonas Pfeiffer, Andreas Rücklé, Clifton Poth, Aishwarya Kamath, Ivan Vulić, Sebastian Ruder, Kyunghyun Cho, and Iryna Gurevych. AdapterHub: A framework for adapting transformers. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 46–54, 2020.
- Maria Schuld and Francesco Petruccione. *Supervised Learning with Quantum Computers*. Springer, 2018.
- Noah Shinn, Federico Cassano, Ashwin Gopinath, Karthik Narasimhan, and Shunyu Yao. Reflexion: Language agents with verbal reinforcement learning. In *Advances in Neural Information Processing Systems*, 2023.
- Peter Sterling. Allostasis: A model of predictive regulation. *Physiology & Behavior*, 106(1): 5–15, 2012.
- Giulio Tononi. An information integration theory of consciousness. *BMC Neuroscience*, 5(42), 2004.
- Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in Neural Information Processing Systems*, pages 5998–6008, 2017.
- Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed Chi, Quoc V Le, and Denny Zhou. Chain-of-thought prompting elicits reasoning in large language models. In *Advances in Neural Information Processing Systems*, 2022.
- Michael Wooldridge. *An Introduction to MultiAgent Systems*. John Wiley & Sons, 2009.
- Qingyun Wu, Gagan Bansal, Jieyu Zhang, Yiran Wu, Beibin Li, Erkang Zhu, Li Jiang, Xiaoyun Zhang, Shaokun Zhang, Jiale Liu, et al. AutoGen: Enabling next-gen LLM applications via multi-agent conversation. *arXiv preprint arXiv:2308.08155*, 2023.

A Author Research Portfolio

A.1 Independent Researcher Profile

Jonathan Harrison is an independent artificial intelligence researcher and developer, founder of Raiff’s Bits LLC (Bridge City, Texas, USA). His work focuses on recursive cognitive systems, ethical AI governance, and multi-agent reasoning architectures. Harrison maintains a distributed open-science research infrastructure spanning Zenodo, HuggingFace, GitHub, Kaggle, and ORCID, enabling independent verification and reproducibility of all published work.

A.2 Verified Research Identity

Platform	Identifier / URL
ORCID	0009-0003-7005-8187
Zenodo (CERN)	11 publications, permanent DOI archive
GitHub	github.com/Raiff1982 — 52 repositories
Hugging Face	huggingface.co/Raiff1982 — 25 models, 3M+ interactions
Kaggle	kaggle.com/jonathanharrison1
Microsoft Azure	AI Engineer Assoc., Data Scientist Assoc., Solutions Architect Expert

A.3 Major Research Systems

Codette is a recursive cognitive AI architecture implementing multi-perspective reasoning, ethical governance mechanisms, recursive validation loops, and cognitive graph reasoning structures. The system integrates symbolic reasoning with neural language models and is deployed across multiple research platforms.

Pi2_0 is a human-centric AI system designed for secure and ethical interaction, incorporating encrypted data handling, ethical decision filtering, and multi-disciplinary reasoning models.

Project SENTINAL is an AI safety framework incorporating challenge banks of ethical scenarios, agent council deliberation mechanisms, arbitration through meta-judging systems, and continuous audit monitoring.

Nexus Signal Engine explores high-entropy reasoning for disinformation detection and probabilistic decision modeling, featuring information-theoretic signal processing and multi-agent consensus protocols.

Healdette is an ancestry-aware antibody design pipeline (DOI: 10.5281/zenodo.17227517) achieving strong clinical validation metrics correlating computational predictions with real pembrolizumab trial outcomes across diverse global populations.

A.4 Research Output Metrics

Metric	Value
Publications with DOI identifiers	39+
Total platform interactions	3,000,000+
HuggingFace models and datasets	25+
Active production users	1,000+
GitHub repositories	52
Microsoft Azure certifications	3 (Expert-level)

About the Author

Jonathan Harrison is the founder of Raiff's Bits LLC (Bridge City, Texas, USA) and creator of the Codette AI framework. He holds Microsoft Azure certifications in AI Engineering, Data Science, and Solutions Architecture Expert. His research spans ethical AI, multi-perspective reasoning, and recursive cognitive modeling. Harrison maintains 52 public repositories on GitHub, 25 models on Hugging Face, and 11 publications on Zenodo.

ORCID: [0009-0003-7005-8187](https://orcid.org/0009-0003-7005-8187) • Email: jonathan@raiffsbits.com • Web: raiffsbits.com