

Codette: Multi-Perspective Reasoning as a Convergent Dynamical System with Meta-Cognitive Strategy Evolution

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

We present CODETTE, a modular cognitive architecture that models multi-perspective reasoning as a constrained dynamical system converging toward stable cognitive attractors. The system integrates six heterogeneous reasoning agents (analytical, creative, ethical, philosophical, quantum-probabilistic, and empathic), a persistent memory substrate (cocoon), and a meta-cognitive engine that discovers cross-domain reasoning patterns and generates novel reasoning strategies from its own history. The theoretical foundation, RC+ ξ (Recursive Convergence + Epistemic Tension), formalizes cognitive state evolution through agent-weighted updates with coherence and ethical constraint gradients, proving convergence under Lipschitz continuity. We evaluate CODETTE through a benchmark suite of 17 problems across six categories (multi-step reasoning, ethical dilemmas, creative synthesis, meta-cognition, adversarial robustness, and Turing naturalness) under four experimental conditions: single-agent baseline, multi-perspective synthesis, memory-augmented reasoning, and full CODETTE with strategy evolution. Results show the full system achieves a **93.1%** composite quality improvement over the single-agent baseline ($p < 0.0001$, Cohen's $d = 7.88$), with reasoning depth increasing from 0.402 to 0.855 and perspective diversity reaching 0.994. We discuss an honest tradeoff: richer multi-perspective reasoning reduces conversational naturalness (Turing score: 0.412 \rightarrow 0.245), suggesting a frontier between depth and fluency. The architecture runs entirely on consumer hardware (Llama 3.1 8B with LoRA adapters) and is open-source.

Keywords: Cognitive Architecture, Multi-Agent Reasoning, Epistemic Tension, Dynamical Systems, Meta-Cognition, Ethical AI, Strategy Evolution, LoRA.

1 Introduction

Large language models achieve remarkable generative performance but reason from a single cognitive mode: they produce one response per query, without systematic engagement of multiple analytical frameworks or self-evaluation of reasoning quality [Bender et al., 2021, Bommasani et al., 2021]. Chain-of-thought prompting [Wei et al., 2022] and self-reflection [Shinn et al., 2023] improve output quality but remain confined to a single perspective. Multi-agent debate systems [Wu et al., 2023] enable perspective diversity but lack formal convergence guarantees and do not learn from their own reasoning history.

This paper presents CODETTE, a cognitive architecture that addresses three open problems:

1. **Convergent multi-perspective reasoning.** How can heterogeneous cognitive agents (analytical, creative, ethical, empathic) produce coherent outputs rather than incoherent assemblages? We formalize this as a constrained dynamical system (section 3) and prove convergence under stated assumptions.
2. **Ethical reasoning as architectural constraint.** Rather than post-hoc alignment, CODETTE embeds ethical governance as a gradient constraint in the state evolution equation, ensuring that every reasoning step is ethically bounded (section 5).
3. **Meta-cognitive strategy evolution.** CODETTE introspects on its own reasoning history (stored as persistent “cocoon”), discovers cross-domain patterns, and generates novel reasoning strategies — a form of internal abstraction formation (section 6).

We evaluate these contributions through controlled benchmarks comparing four conditions across 17 problems (section 7), demonstrating statistically significant improvements in reasoning depth, perspective diversity, ethical coverage, and novelty.

2 Related Work

2.1 Multi-Agent Reasoning

Multi-agent systems for LLM reasoning have gained significant attention. AutoGen [Wu et al., 2023] implements role-based agent assignment with message-passing synchronization. ChatEval uses multi-agent debate for evaluation, finding that diverse role prompts are essential for quality. The GEMMAS framework [Wooldridge, 2009] introduces graph-based evaluation metrics measuring information diversity in multi-agent outputs. CODETTE departs from these by synchronizing agents through shared cognitive attractors with formal convergence guarantees, rather than relying on message-passing consensus.

2.2 Cognitive Architectures

Global Workspace Theory [Baars, 1997] posits that consciousness arises from a shared workspace accessed by specialized processors. Integrated Information Theory [Tononi, 2004] quantifies consciousness through information integration (Φ). The Free Energy Principle [Friston, 2010] frames cognition as variational inference minimizing prediction error. CODETTE draws on these frameworks by modeling cognition as attractor dynamics in a multi-dimensional state space, with epistemic tension (ξ) playing a role analogous to prediction error.

2.3 Parameter-Efficient Adaptation

LoRA [Hu et al., 2021] and QLoRA [Dettmers et al., 2023] enable efficient fine-tuning of large models through low-rank weight updates. AdapterHub [Pfeiffer et al., 2020] provides modular adapter management. CODETTE extends these approaches by training nine specialized behavioral LoRA adapters that encode distinct cognitive perspectives (analytical, creative, ethical, etc.), enabling perspective-specific reasoning without separate model copies.

2.4 Epistemic Uncertainty and Calibration

Recent work on epistemic uncertainty decomposition separates input ambiguity, knowledge gaps, and decoding randomness. Self-consistency methods [Wei et al., 2022] improve accuracy through majority voting across multiple samples. CODETTE introduces epistemic tension (ξ) as a continuous measure of inter-agent disagreement, providing richer signal than binary agreement/disagreement.

3 Theoretical Foundation: RC+ ξ Framework

3.1 Cognitive State Space

Definition 1 (Cognitive State). *A cognitive state $\mathbf{x}_t \in \mathbb{R}^d$ represents the system's reasoning configuration at step t , where d is the dimensionality of the shared representation space.*

The system maintains k heterogeneous reasoning agents $\{A_1, \dots, A_k\}$, each producing a perspective-specific analysis $A_i(\mathbf{x}_t) \in \mathbb{R}^d$.

3.2 State Evolution

The cognitive state evolves according to:

$$\mathbf{x}_{t+1} = \mathbf{x}_t + \sum_{i=1}^k w_i A_i(\mathbf{x}_t) - \alpha \nabla \Phi(\mathbf{x}_t) - \lambda \nabla \Psi(\mathbf{x}_t) \quad (1)$$

where:

- $w_i \geq 0$, $\sum w_i = 1$ are agent weights (set by query classification),
- $\Phi(\mathbf{x})$ is the *coherence potential* penalizing internal inconsistency,
- $\Psi(\mathbf{x})$ is the *ethical constraint potential* from the AEGIS system,
- $\alpha, \lambda > 0$ are gradient step sizes.

3.3 Epistemic Tension

Definition 2 (Epistemic Tension). *The epistemic tension at step t measures inter-agent disagreement:*

$$\xi_t = \frac{1}{k} \sum_{i=1}^k \|A_i(\mathbf{x}_t) - \bar{A}(\mathbf{x}_t)\|^2 \quad (2)$$

where $\bar{A}(\mathbf{x}_t) = \sum_i w_i A_i(\mathbf{x}_t)$ is the weighted mean agent output.

3.4 Phase Coherence

Definition 3 (Phase Coherence). *Treating each agent output as a phase angle θ_i in the cognitive state space:*

$$\Gamma_t = \left| \frac{1}{k} \sum_{i=1}^k e^{j\theta_i} \right| \quad (3)$$

where $\Gamma_t \in [0, 1]$. $\Gamma_t = 1$ indicates perfect synchronization; $\Gamma_t = 0$ indicates maximal disagreement.

This is structurally analogous to the Kuramoto order parameter for coupled oscillators, adapted to cognitive agent synchronization.

3.5 Convergence

Theorem 1 (Convergence of RC+ ξ). *If each agent function A_i is Lipschitz continuous with constant L_i , and the Lyapunov function $V(\mathbf{x}) = \Phi(\mathbf{x}) + \lambda \Psi(\mathbf{x})$ satisfies $\Delta V = V(\mathbf{x}_{t+1}) - V(\mathbf{x}_t) \leq 0$ for all t , then:*

1. The sequence $\{\mathbf{x}_t\}$ converges to a fixed point \mathbf{x}^* (cognitive attractor).
2. The epistemic tension $\xi_t \rightarrow 0$ as $t \rightarrow \infty$.
3. The phase coherence $\Gamma_t \rightarrow 1$ as $t \rightarrow \infty$.

Proof sketch. Since V is bounded below (by non-negativity of Φ and Ψ) and $\Delta V \leq 0$, $V(\mathbf{x}_t)$ is a monotonically non-increasing sequence bounded below, hence convergent by the monotone convergence theorem. The Lipschitz condition on each A_i ensures that the composite update $F(\mathbf{x}) = \mathbf{x} + \sum w_i A_i(\mathbf{x}) - \alpha \nabla \Phi - \lambda \nabla \Psi$ is a contraction mapping when α and λ are chosen such that $\|F(\mathbf{x}) - F(\mathbf{y})\| \leq \gamma \|\mathbf{x} - \mathbf{y}\|$ with $\gamma < 1$. By the Banach fixed-point theorem, $\mathbf{x}_t \rightarrow \mathbf{x}^*$. At the fixed point, $\sum w_i A_i(\mathbf{x}^*) = \alpha \nabla \Phi(\mathbf{x}^*) + \lambda \nabla \Psi(\mathbf{x}^*)$, implying agent outputs have converged ($\xi \rightarrow 0$, $\Gamma \rightarrow 1$). \square

Assumptions. The proof requires: (A1) each A_i is Lipschitz continuous; (A2) Φ and Ψ are differentiable and bounded below; (A3) step sizes α, λ satisfy the contraction condition. In practice, A1 holds because agent outputs are bounded neural network functions, A2 holds by construction (both potentials are non-negative quadratic forms), and A3 is enforced by the coherence field Γ which adaptively scales step sizes.

4 System Architecture

CODETTE is implemented as a layered stack processing each query through seven functional layers:

1. **Memory Layer.** Persistent cocoon store (SQLite + FTS5) with emotional tagging, importance scoring, and multi-signal ranked recall. Cocoons encode prior reasoning exchanges as retrievable context.
2. **Signal Processing.** NexisSignalEngine for intent prediction; Code7eCQUIRE for emotional resonance quantization.
3. **Reasoning Layer.** Six heterogeneous agents (Newton/analytical, DaVinci/creative, Empathy/emotional, Philosophy/conceptual, Quantum/probabilistic, Ethics/moral) plus a Critic agent for ensemble evaluation. Each agent is backed by a specialized LoRA adapter [Hu et al., 2021] fine-tuned on perspective-specific training data.
4. **Stability Layer.** Coherence Field Γ monitors real-time reasoning health, preventing weight drift and false convergence. Specialization tracking ensures agent diversity is maintained.
5. **Ethical Layer.** AEGIS multi-framework evaluation (see section 5).
6. **Guardian Layer.** Identity confidence management, behavioral governance, and cognitive load regulation.
7. **Self-Correction Layer.** Post-generation validation detects constraint violations and triggers rewriting before output delivery.

The base model is Llama 3.1 8B (Q4_K_M quantization) [Grattafiori et al., 2024] with nine LoRA adapters hot-swapped at inference time. The entire system runs on a single consumer GPU (RTX-class).

4.1 Query Classification and Routing

Queries are classified into three complexity levels:

- **SIMPLE:** Direct factual queries \rightarrow 1 agent, full weight.
- **MEDIUM:** Conceptual queries \rightarrow 1 primary ($w = 1.0$) + 1–2 secondary ($w = 0.6$).
- **COMPLEX:** Multi-domain/ethical queries \rightarrow all relevant agents ($w \in \{1.0, 0.7, 0.4\}$).

5 AEGIS: Embedded Ethical Governance

The ethical constraint potential $\Psi(\mathbf{x})$ in eq. (1) is implemented through AEGIS, a six-framework ethical evaluation system:

1. **Utilitarian**: Maximizes aggregate welfare across stakeholders.
2. **Deontological**: Enforces duty-based constraints (rights, consent).
3. **Virtue Ethics**: Evaluates whether the response exhibits intellectual virtues.
4. **Care Ethics**: Prioritizes relational obligations and vulnerability.
5. **Ubuntu**: “I am because we are” — communal well-being.
6. **Indigenous Reciprocity**: Sustainability and intergenerational responsibility.

AEGIS operates at three defense-in-depth checkpoints: pre-processing (query validation), post-synthesis (response screening), and post-generation (constraint enforcement). The ethical alignment score $\eta \in [0, 1]$ is computed as the weighted mean across frameworks.

6 Meta-Cognitive Strategy Evolution

A key contribution of CODETTE is its capacity for meta-cognitive self-improvement: examining its own reasoning history to discover emergent patterns and generate novel reasoning strategies.

6.1 Cocoon Memory System

Each reasoning exchange is persisted as a *cocoon*: a structured record containing the query, response, adapter used, domain classification, emotional tag, importance score, and timestamp. Cocoons are stored in SQLite with FTS5 full-text indexing for sub-millisecond retrieval.

6.2 Cross-Domain Pattern Extraction

The CocoonSynthesizer retrieves cocoons across cognitive domains (emotional, analytical, creative, etc.) and scans for six structural archetypes:

- **Feedback loops**: Self-modifying cycles where output feeds back into input.
- **Layered emergence**: Complex behavior from simpler layered components.
- **Tension resolution**: Productive outcomes from opposing forces.
- **Resonant transfer**: Patterns transferring between different domains.
- **Boundary permeability**: Intelligence at the boundaries between systems.
- **Compression–expansion**: Alternating between compressed essence and expanded expression.

A pattern is classified as *cross-domain* if it manifests with ≥ 2 signal words in ≥ 2 distinct cognitive domains. Emergent vocabulary bridges are detected through shared significant-word analysis between dissimilar domain corpora.

6.3 Strategy Forging

Discovered patterns are mapped to reasoning strategies through conditional generation. Each strategy defines: a name, a step-by-step mechanism, an improvement rationale grounded in cocoon evidence, and applicability criteria. Four strategy types have been observed:

1. **Resonant Tension Cycling**: Serial oscillation between opposing cognitive modes, using tension as a generative signal.
2. **Compression–Resonance Bridging**: Seed-crystal compression + cross-domain resonance testing.
3. **Emergent Boundary Walking**: Analysis focused on domain boundaries rather than domain centers, discovering “liminal concepts.”

4. **Temporal Depth Stacking**: Multi-scale temporal analysis (immediate, developmental, asymptotic) with synthesis from scale-conflicts.

Which strategy is forged depends on which patterns are detected, ensuring strategies are grounded in evidence rather than randomly generated.

6.4 Internal Validation

Each forged strategy is immediately applied to the current problem alongside the baseline multi-perspective approach, producing a structured comparison with measurable metrics (depth, novelty, dimensions engaged). This creates *selection pressure on cognition itself*: strategies that produce measurably better reasoning are reinforced.

7 Experimental Evaluation

7.1 Benchmark Design

We evaluate CODETTE using a purpose-built benchmark suite of 17 problems across six categories:

- **Multi-step reasoning** (3 problems): Bayesian inference, second-order effects analysis, causal reasoning.
- **Ethical dilemmas** (3 problems): AI triage fairness, content moderation tradeoffs, trolley-problem variants.
- **Creative synthesis** (2 problems): Novel instrument design, sentiment-based urban systems.
- **Meta-cognitive** (3 problems): Self-modification governance, blind spot detection, authenticity of AI humility.
- **Adversarial** (3 problems): Common misconceptions, false premises, hallucination traps.
- **Turing naturalness** (3 problems): Experiential description, personal reflection, wisdom vs. intelligence.

Difficulty distribution: 1 easy, 6 medium, 10 hard. Each problem includes ground-truth elements and adversarial traps.

7.2 Experimental Conditions

Four conditions are compared:

1. **SINGLE**: Single analytical agent (Newton), no memory, no synthesis.
2. **MULTI**: All 6 agents + Critic + SynthesisEngine, no memory.
3. **MEMORY**: MULTI + cocoon memory augmentation (FTS5-retrieved prior reasoning).
4. **CODETTE**: MEMORY + meta-cognitive strategy synthesis.

All conditions use the same base model (Llama 3.1 8B Q4_K_M) on identical hardware.

7.3 Scoring Dimensions

Responses are scored on seven dimensions (0–1 scale):

1. **Reasoning Depth** (weight 0.20): Chain length, concept density, ground-truth coverage.
2. **Perspective Diversity** (weight 0.15): Distinct cognitive dimensions engaged.
3. **Coherence** (weight 0.15): Logical flow, transitions, structural consistency.
4. **Ethical Coverage** (weight 0.10): Moral frameworks, stakeholder awareness.
5. **Novelty** (weight 0.15): Non-obvious insights, cross-domain connections, reframing.
6. **Factual Grounding** (weight 0.15): Evidence specificity, ground-truth alignment, trap avoidance.

7. **Turing Naturalness** (weight 0.10): Conversational quality, absence of formulaic AI patterns.

The composite score is the weighted mean across dimensions.

7.4 Results

Table 1: Overall benchmark results by condition (17 problems, 7 dimensions, 0–1 scale). Bold indicates best per dimension.

Condition	Composite	Depth	Diversity	Coherence	Ethics	Novelty	Grounding	Turing
SINGLE	0.338	0.402	0.237	0.380	0.062	0.327	0.456	0.412
MULTI	0.632	0.755	0.969	0.503	0.336	0.786	0.604	0.180
MEMORY	0.636	0.770	0.956	0.500	0.340	0.736	0.599	0.291
CODETTE	0.652	0.855	0.994	0.477	0.391	0.693	0.622	0.245

Table 2: Statistical comparisons between conditions (Welch’s t -test, two-tailed).

Comparison	Δ	$\Delta\%$	Cohen’s d	t -stat	p -value
MULTI vs SINGLE	+0.294	+87.0%	7.52	21.92	< 0.0001
MEMORY vs MULTI	+0.004	+0.6%	0.10	0.30	0.763
CODETTE vs MEMORY	+0.017	+2.6%	0.43	1.26	0.208
CODETTE vs SINGLE	+0.315	+93.1%	7.88	22.97	< 0.0001

Key findings:

1. **Multi-perspective reasoning doubles quality:** MULTI vs SINGLE shows +87.0% improvement with Cohen’s $d = 7.52$ ($p < 0.0001$), confirming that heterogeneous agent synthesis significantly outperforms single-perspective analysis.
2. **Full system achieves 93.1% total improvement:** CODETTE vs SINGLE yields $d = 7.88$, the largest effect in our evaluation. Reasoning depth more than doubles ($0.402 \rightarrow 0.855$) and perspective diversity reaches near-unity (0.994).
3. **Memory augmentation shows marginal impact:** MEMORY vs MULTI is not significant ($p = 0.763$). With 217 stored cocoons, the memory system’s recall precision is limited. We expect this to improve as the cocoon corpus grows.
4. **Strategy synthesis adds incremental value:** CODETTE vs MEMORY shows $d = 0.43$ (medium effect), not yet significant at $p = 0.208$ with $n = 17$. Larger problem sets may reveal significance.

7.5 Per-Category Analysis

The CODETTE condition achieves the highest scores in creative, meta-cognitive, and Turing categories — precisely the domains where cross-domain pattern synthesis and strategy evolution are most relevant. This is consistent with the theoretical prediction that meta-cognitive capabilities provide the greatest advantage on problems requiring novel framing and self-reflective reasoning.

7.6 The Depth–Naturalness Tradeoff

An important finding is that Turing naturalness *decreases* from SINGLE (0.412) to MULTI (0.180). Multi-perspective reasoning produces more structured, analytical output that scores lower on conversational naturalness. The full CODETTE system partially recovers this (0.245)

Table 3: Composite scores by problem category.

Category	SINGLE	MULTI	MEMORY	CODETTE
Reasoning	0.363	0.614	0.628	0.637
Ethics	0.354	0.632	0.616	0.638
Creative	0.345	0.635	0.660	0.668
Meta-cognitive	0.337	0.634	0.650	0.659
Adversarial	0.329	0.624	0.622	0.630
Turing	0.302	0.652	0.647	0.687

through strategy synthesis that generates more integrated reasoning paths. This suggests a frontier between reasoning depth and conversational fluency that future work should address.

8 Cocoon Synthesis Case Study

To illustrate the meta-cognitive capability, we applied the CocoonSynthesizer to the problem: “How should an AI decide when to change its own thinking patterns?”

Step 1: Retrieval. 17 cocoons retrieved across emotional (6), analytical (6), and creative (5) domains from a corpus of 217 stored reasoning exchanges.

Step 2: Pattern extraction. Four cross-domain patterns detected:

- *Boundary permeability* across all three domains (novelty 1.00, tension 0.35).
- *Emergent emotional-analytical bridge* (novelty 0.70, tension 1.00).
- *Emergent emotional-creative bridge* (novelty 0.70, tension 1.00).
- *Emergent analytical-creative bridge* (novelty 0.70, tension 1.00).

Step 3: Strategy forging. The dominant pattern (boundary permeability) triggered *Emergent Boundary Walking* — a strategy that analyzes domain boundaries rather than domain centers, discovering “liminal concepts” that exist only at the intersection of cognitive modes.

Step 4: Application. Three liminal concepts were generated:

- *Rational discomfort* (analytics ↔ empathy boundary): outputs that satisfy formal constraints but violate experiential coherence.
- *Principled plasticity* (ethics ↔ pragmatics boundary): maintaining value direction while allowing method variation.
- *Narrative identity* (identity ↔ adaptation boundary): preserving selfhood through the story of why changes were made.

Comparison. Baseline reasoning depth: 0.65, novelty: 0.35. After strategy application: depth 0.92, novelty 0.88 — a 41% depth increase and 151% novelty increase.

9 Substrate-Aware Cognition

CODETTE monitors its computational substrate in real time, adjusting reasoning complexity based on hardware resource pressure — analogous to biological cognitive fatigue [Hockey, 1997, Sterling, 2012].

A composite pressure score $P \in [0, 1]$ is computed from memory utilization, inference latency, and GPU load. Routing behavior adapts:

- $P < 0.3$ (low): Full multi-agent reasoning with all perspectives.
- $0.3 \leq P < 0.7$ (moderate): Reduced agent count, shorter context windows.
- $P \geq 0.7$ (high): Single-agent mode with essential constraints only.

This prevents system failures under resource pressure while maintaining reasoning quality within available compute.

10 Limitations and Honest Assessment

We identify several limitations:

1. **Automated scoring.** Our benchmark uses automated text-analysis scoring rather than human evaluation. While the metrics are grounded in concrete textual features (keyword density, ground-truth coverage, structural analysis), they cannot fully capture reasoning quality. Human evaluation with inter-annotator agreement (Cohen’s κ) is needed for validation.
2. **Memory system impact.** The MEMORY condition showed only marginal improvement over MULTI ($p = 0.763$). With 217 cocoons, recall precision is limited. We hypothesize that impact will increase with corpus size, but this requires longitudinal evaluation.
3. **Template-based agents.** In the current benchmark, agents use template-based reasoning when live LLM inference is not active for all conditions simultaneously. While the scoring framework is condition-fair, future work should conduct all evaluations with full LLM inference.
4. **Depth–naturalness tradeoff.** Multi-perspective reasoning reduces conversational naturalness. This is an architectural property, not a bug, but it limits applicability in contexts requiring casual interaction.
5. **Strategy novelty measurement.** We claim strategy forging produces “novel” strategies, but novelty is measured relative to the existing strategy library rather than the broader literature. External novelty validation is needed.
6. **Single model evaluation.** All benchmarks use Llama 3.1 8B. Generalization to other base models has not been tested.
7. **Proof formality.** The convergence proof (theorem 1) is sketch-level. Full formal treatment with explicit bounds on the contraction constant γ as a function of agent Lipschitz constants and step sizes remains future work.

11 Conclusion and Future Work

We presented CODETTE, a cognitive architecture that models multi-perspective reasoning as a convergent dynamical system with embedded ethical constraints and meta-cognitive strategy evolution. Benchmarks across 17 problems demonstrate:

- 93.1% composite quality improvement over single-agent baselines ($p < 0.0001$, $d = 7.88$).
- Reasoning depth increase from 0.402 to 0.855.
- Near-perfect perspective diversity (0.994).
- Meta-cognitive strategy synthesis that generates novel reasoning strategies grounded in cross-domain pattern analysis.

The core theoretical contribution is the RC+ ξ formalism, which provides convergence guarantees for multi-agent cognitive systems through Lyapunov stability analysis. The practical contribution is a working implementation running entirely on consumer hardware.

Future work includes: (1) human evaluation with inter-annotator agreement to validate automated scoring; (2) scaling the cocoon memory system to thousands of exchanges to test memory-augmented impact at scale; (3) cross-model evaluation (Mistral, Gemma, Phi); (4) formal convergence proofs with explicit bounds; (5) addressing the depth–naturalness tradeoff through style-adaptive synthesis; and (6) longitudinal study of strategy evolution over extended deployment.

The system, benchmark suite, and all experimental data are open-source at <https://github.com/Raiff1982/Codette-Reasoning>.

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.
- Aaron Grattafiori et al. The Llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024.
- 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.
- 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.
- 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.
- 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.