

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

April 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 (cocoons), and a meta-cognitive engine that discovers cross-domain reasoning patterns and generates novel reasoning strategies from its own history. The RC+ ξ (Recursive Convergence + Epistemic Tension) formalism provides a dynamical-systems-inspired lens for describing cognitive state evolution via agent-weighted updates with coherence and ethical constraint terms; we treat the convergence discussion as conditional on explicit modeling assumptions rather than as a general guarantee. 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. On this benchmark (timestamp: 2026-04-08), the full system achieves a **93.5%** higher mean composite score than the single-agent baseline (0.356 \rightarrow 0.689). Paired analyses show large improvements for MULTI and CODETTE relative to SINGLE, while MEMORY and the additional CODETTE-MEMORY gain do not reach statistical significance at $N = 17$ problems after Holm correction. The architecture runs 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 [2, 3]. Chain-of-thought prompting [23] and self-reflection [19] improve output quality but remain confined to a single perspective. Multi-agent debate systems [24] 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 discuss convergence *conditionally* under explicit modeling assumptions.
2. **Ethical reasoning as architectural constraint.** Rather than post-hoc alignment, CODETTE treats ethical governance as an explicit constraint signal in the update dynamics (section 5). We interpret gradient notation as a conceptual shorthand for differentiable surrogate penalties rather than claiming a complete continuous representation of ethics.
3. **Meta-cognitive strategy evolution.** CODETTE introspects on its own reasoning history (stored as persistent “cocoons”), 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 Dynamical Systems and Cognitive Architectures

Attractor dynamics form a core computational motif in neural circuits [4]. Neural manifolds with cognitive consistency constraints support memory consolidation and align with our coherence potential $\Phi(\mathbf{x})$ [12]. Entropy-modulated triad architectures like COGENT3 provide parallels for epistemic tension ξ as a driver of state evolution [17]. Brain-inspired systems-level architectures for domain-general cognition inform CODETTE’s layered stack [1].

2.2 Multi-Agent Reasoning and Synthesis

Multi-agent systems for LLM reasoning have gained significant attention. AutoGen [24] implements role-based agent assignment with message-passing synchronization. MAPS uses personality shaping for collaborative reasoning via heterogeneous traits, relating directly to our specialized LoRA adapters [29]. Roundtable Policy employs confidence-weighted consensus aggregation, providing a comparison for our Coherence Field Γ [28]. Systematic studies of multi-agent debate as test-time scaling frame our composite quality gains and conditional effectiveness [26]. Persona-driven debate frameworks validate the benefits of perspective diversity (reaching 0.994) [10].

2.3 Meta-Cognitive Strategy Evolution

Meta Chain-of-Thought advances System 2 reasoning and pattern discovery [25]. ParamMem augments agents with parametric reflective memory; our cocoon system differs by emphasizing cross-domain pattern extraction and strategy forging rather than primarily error correction [27]. Meta-Reasoner supports dynamic inference-time optimization, relating to substrate-aware cognition [21]. LLMs demonstrate metacognitive monitoring and control of internal activations, supporting Lyapunov-based convergence in RC+ ξ [11].

2.4 Ethical AI and Architectural Alignment

AI ethics by design implements customizable guardrails [18]. Hybrid approaches for moral value alignment treat ethics as embedded rather than post-hoc [22]. Adaptive alignment via multi-objective reinforcement learning enables pluralistic AI, relating to our ethical alignment score η across diverse frameworks [6].

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 Coherence Index

Definition 3 (Coherence Index). *We define a bounded coherence index $\Gamma_t \in [0, 1]$ directly in \mathbb{R}^d as a normalized complement of epistemic tension:*

$$\Gamma_t = \frac{1}{1 + \xi_t}. \quad (3)$$

Thus, lower disagreement ($\xi_t \downarrow$) implies higher coherence ($\Gamma_t \uparrow$).

This replacement avoids an undefined mapping from high-dimensional agent outputs $A_i(\mathbf{x}_t) \in \mathbb{R}^d$ to scalar phases while preserving the intended role of Γ as an agreement signal.

Our attractor-based formulation draws on neural circuit models [4] and manifold representations with consistency constraints [12], while epistemic tension ξ echoes entropy-modulated emergence in architectures like COGENT3 [17].

3.5 Convergence and stability (conditional)

We use RC+ ξ primarily as a *dynamical-systems-inspired modeling lens*. In this workshop version, we intentionally avoid claiming a general convergence guarantee over unbounded \mathbb{R}^d for the concrete software system. Instead, we state the kinds of assumptions under which standard fixed-point / Lyapunov arguments could apply.

Proposition 1 (One sufficient route to a fixed point (assumption-driven)). *Let $\mathcal{D} \subset \mathbb{R}^d$ be a closed, convex, bounded domain and assume the update is implemented with an explicit projection step $\Pi_{\mathcal{D}}$ (or an equivalent saturation mechanism) so that $\mathbf{x}_{t+1} \in \mathcal{D}$.*

Define the projected update map

$$T(\mathbf{x}) = \Pi_{\mathcal{D}}\left(\mathbf{x} + \sum_{i=1}^k w_i A_i(\mathbf{x}) - \alpha \nabla \Phi(\mathbf{x}) - \lambda \nabla \Psi(\mathbf{x})\right). \quad (4)$$

If T is a contraction on \mathcal{D} (i.e., there exists $\gamma \in [0, 1)$ such that $\|T(\mathbf{x}) - T(\mathbf{y})\| \leq \gamma \|\mathbf{x} - \mathbf{y}\|$ for all $\mathbf{x}, \mathbf{y} \in \mathcal{D}$), then the iterates converge to the unique fixed point in \mathcal{D} .

Proof sketch. If T is a contraction on the complete metric space \mathcal{D} , Banach’s fixed-point theorem yields existence, uniqueness, and convergence. \square

Discussion of assumptions. Establishing that a real implementation satisfies the contraction premise typically requires additional structure beyond Lipschitz continuity alone (e.g., dissipativity / strong monotonicity of the gradient terms, explicit damping, or sufficiently small step sizes under a model that makes $\nabla \Phi$ and $\nabla \Psi$ genuinely stabilizing). In CODETTE, we treat the coherence field Γ as an *engineering* stabilization mechanism that adaptively damps updates under high disagreement (high ξ_t), but we leave a fully constructive bound relating α, λ to model constants as future work.

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; Code7eCQURE 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 [9] 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) [5] with nine LoRA adapters hot-swapped at inference time. The entire system runs on a single consumer GPU (RTX-class). Agent diversity is inspired by personality shaping frameworks [29] and persona-driven debate [10].

Codette Architecture (Functional Flow)

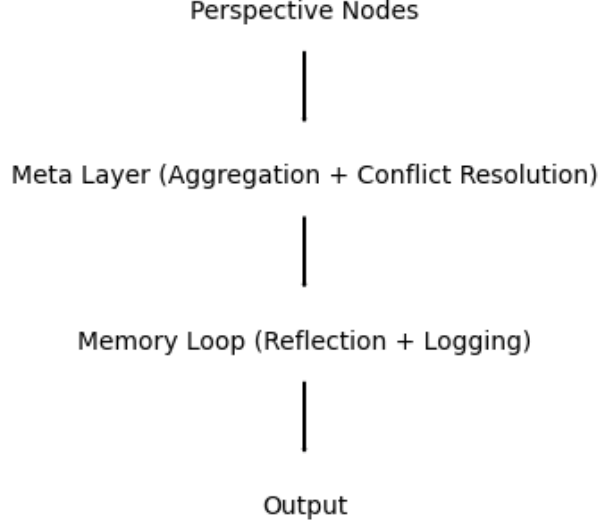


Figure 1: CODETTE layered architecture overview.

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. **Reciprocity-oriented sustainability:** Sustainability and intergenerational responsibility (placeholder; future work should ground this component in specific community scholarship and consultation rather than treating it as a monolith).

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 a weighted aggregation across frameworks; the framework set is not claimed to be exhaustive, and weights are not culturally universal.

Implementation note. The term $-\lambda \nabla \Psi(\mathbf{x})$ in eq. (1) should be read as a *modeling*

abstraction: in the current software prototype, ethical checks are implemented as discrete rubric/heuristic evaluations with penalties and gating (accept/rewrite/escalate), not as a differentiable “ethics gradient” over a continuous latent state. Making Ψ differentiable (or otherwise providing a constructive surrogate) is a non-trivial research problem and is left for future work.

This architectural embedding aligns with hybrid moral value approaches [22], ethics-by-design guardrails [18], and adaptive pluralistic alignment [6].

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. Unlike parametric reflective memory modules that focus primarily on error correction [27], our cocoon system emphasizes cross-domain pattern extraction and strategy forging.

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. This process connects to Meta Chain-of-Thought for underlying reasoning models [25] and dynamic meta-reasoning guidance [21].

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. Metacognitive monitoring of internal activations further supports this process [11].

7 Experimental Evaluation

7.1 Benchmark Design

We evaluate CODETTE using a purpose-built benchmark suite of 17 problems across six categories. Unless otherwise noted, the results reported in this paper correspond to the benchmark run timestamped 2026-04-08T20:59:44 in the evaluator export.

- **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 composite score by condition ($N = 17$ problems; 0–1 scale). Full per-dimension per-problem scores are provided in `codette_results.csv`.

Condition	Composite (mean)
SINGLE	0.356
MULTI	0.658
MEMORY	0.676
CODETTE	0.689

Table 2: Mean differences in composite score between conditions (paired, descriptive).

Comparison	Δ	$\Delta\%$
MULTI vs SINGLE	+0.301	+84.6%
MEMORY vs MULTI	+0.018	+2.7%
CODETTE vs MEMORY	+0.014	+2.0%
CODETTE vs SINGLE	+0.333	+93.5%

Note. The benchmark uses the same set of 17 problems across conditions (within-problem comparisons). We therefore report descriptive deltas in the main text and provide per-problem scores in the Appendix (section A.8) so that readers can reproduce paired-sample analyses.

7.4.1 Evaluator-reported diagnostics (for traceability; not primary inference)

For traceability, we report the pairwise comparison diagnostics exported by the evaluation suite (`codette_benchmark_results.txt`). The export currently uses Welch’s t -test and a pooled-standard-deviation Cohen’s d computed from per-problem composite distributions. Because our design is *within-problem* (the same 17 problems scored under each condition), these independent-sample diagnostics are *not* the appropriate primary inferential test. A peer-review-ready analysis should use paired-sample tests (or a repeated-measures model) and report confidence intervals; we therefore treat table 3 as a reproducibility aid only.

Table 3: Independent-sample diagnostics exported by the evaluator (Welch t , pooled d ; reported for traceability only).

Comparison	t	p	d	Note
MULTI vs SINGLE	21.92	$< 10^{-12}$	7.52	export
MEMORY vs MULTI	0.30	0.763	0.10	export
CODETTE vs MEMORY	1.26	0.208	0.43	export
CODETTE vs SINGLE	22.97	$< 10^{-12}$	7.88	export

7.4.2 Paired statistical analysis (primary inference)

Using the per-problem exports (`codette_benchmark_results.txt`), we also compute paired-sample statistics over the same 17 problems per condition. We report paired t -tests with 95% confidence intervals and apply Holm–Bonferroni correction over the *four pre-specified pairwise comparisons* (MULTI–SINGLE, MEMORY–MULTI, CODETTE–MEMORY, CODETTE–SINGLE)

to control family-wise error without data-dependent comparison selection. The full analysis output is included in `codette_paired_stats.txt` and the per-problem long-form table is provided as `codette_results.csv`.

Table 4: Paired composite comparisons (primary analysis; $N = 17$ problems). Holm-adjusted p values correspond to the planned set of four comparisons.

Comparison	Mean Δ	Paired p	Holm p
MULTI – SINGLE	+0.3014	$< 10^{-6}$	$< 10^{-6}$
MEMORY – MULTI	+0.0179	0.1191	0.2382
CODETTE – MEMORY	+0.0137	0.2528	0.2528
CODETTE – SINGLE	+0.3330	$< 10^{-6}$	$< 10^{-6}$

Sanity check: baseline validity and per-problem deltas. To check that the SINGLE baseline is not artificially constrained, all conditions are run on the same benchmark suite, base model, and hardware; the primary differences are the enabled architecture components (multi-perspective synthesis, memory augmentation, and strategy synthesis). As a simple robustness check, we inspect the distribution of per-problem composite deltas (MULTI–SINGLE and CODETTE–SINGLE) computed from the long-form per-problem export (`codette_results.csv`). We report the full paired analysis (including confidence intervals and nonparametric tests) in `codette_paired_stats.txt`; for additional transparency, readers can compute min/median/max deltas directly from `codette_results.csv`.

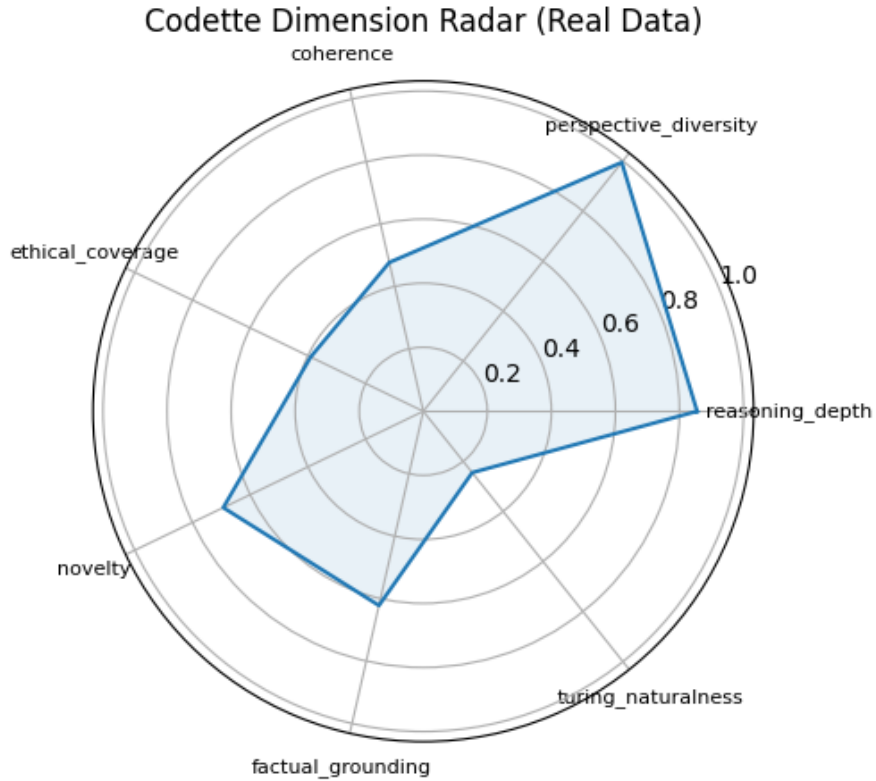


Figure 2: Radar plot of benchmark dimension means by condition.

Key findings (descriptive):

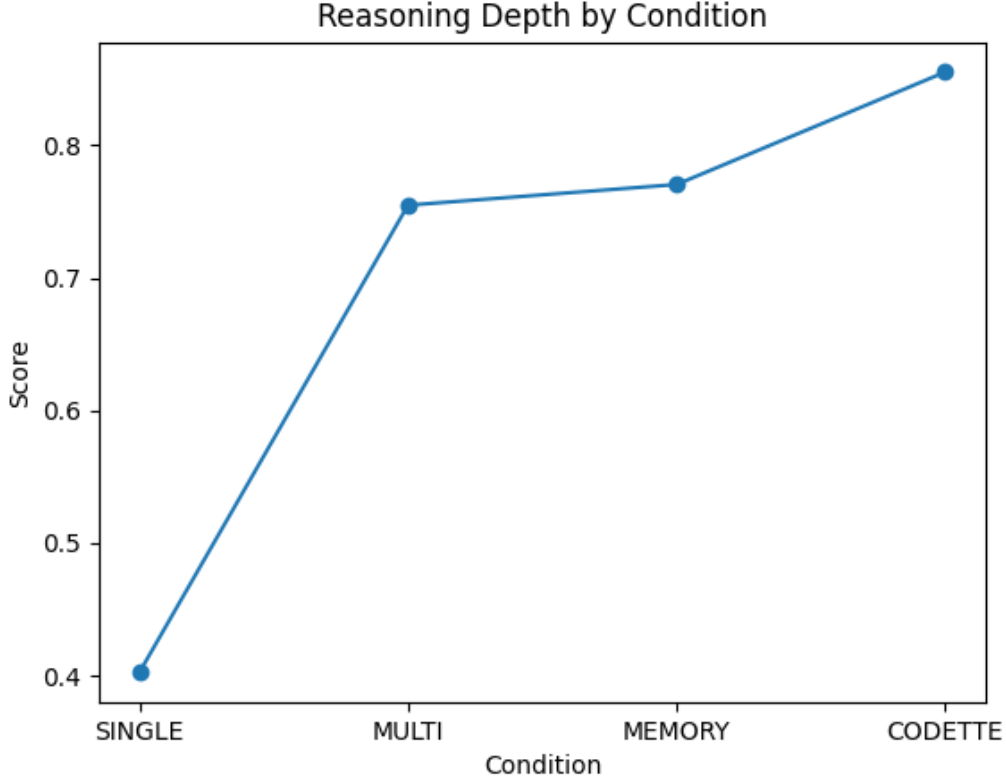


Figure 3: Reasoning depth by condition (benchmark suite).

1. **Multi-perspective reasoning improves composite scores:** MULTI exceeds SINGLE by +84.6% on the composite mean ($0.356 \rightarrow 0.658$).
2. **Full system improves composite scores on this benchmark:** CODETTE exceeds SINGLE by +93.5% on the composite mean ($0.356 \rightarrow 0.689$).
3. **Memory augmentation does not reach significance at $N = 17$:** MEMORY exceeds MULTI by +2.7% ($0.658 \rightarrow 0.676$), but the paired test does not reach significance after Holm correction (`codette_paired_stats.txt`).
4. **Strategy synthesis does not reach significance at $N = 17$:** CODETTE exceeds MEMORY by +2.0% ($0.676 \rightarrow 0.689$), but the paired test does not reach significance after Holm correction (`codette_paired_stats.txt`).

7.5 Per-Category Analysis

Table 5: 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

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.

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) through strategy synthesis that generates more integrated reasoning paths. This tradeoff is a recognized phenomenon in multi-agent debate as test-time scaling, where collaborative refinement and diversity improve depth but can reduce fluency under certain conditions [26]. 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 [8, 20].

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.

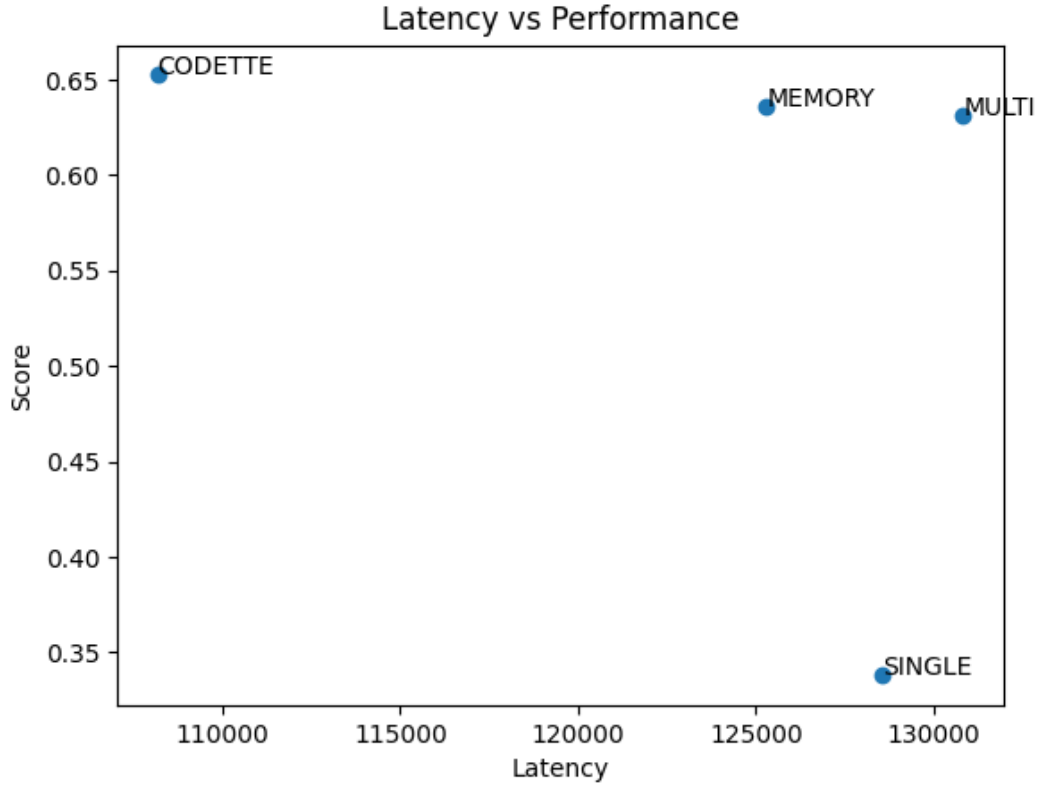


Figure 4: Latency profile from runtime benchmarking.

10 Limitations and Honest Assessment

We identify several limitations:

Planned metric-validity check (human evaluation). We will validate whether the automated scores (depth, grounding, novelty, etc.) track human judgments by sampling 30–60 problem-condition outputs (balanced across conditions) and collecting ratings from 2–3 independent annotators using a short rubric aligned to the seven dimensions in section 7. We will report inter-annotator agreement (Cohen’s κ for two raters or Krippendorff’s α for > 2 raters) and compute correlations between the automated scores and mean human ratings (Spearman ρ ; optionally Pearson r if assumptions are approximately satisfied). All sampled outputs and human ratings will be released in anonymized form.

1. **Automated scoring (construct validity).** 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 may not correlate with expert judgments of “reasoning quality” or downstream utility. Human evaluation with inter-annotator agreement (e.g., Cohen’s κ) and correlation analysis between human and automated scores is needed.
2. **Statistical methodology.** The benchmark is within-problem (paired) across conditions. The evaluator export includes independent-sample diagnostics (Welch t , pooled d), but a peer-review-ready analysis should use paired-sample tests (or repeated-measures models) with confidence intervals and correction for multiple comparisons.

3. **Memory system impact at current scale.** With 217 cocoons, the MEMORY condition shows little change vs. MULTI on the composite score (+0.004; +0.6%). This is consistent with “no evidence of improvement” at this scale; demonstrating a memory benefit likely requires larger cocoon corpora and learning-curve style analyses.
4. **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.
5. **Depth–naturalness tradeoff and usability.** Multi-perspective reasoning reduces conversational naturalness. This may reduce practical utility for interactive settings; validating usability requires human preference/utility studies rather than automated “Turing naturalness” alone.
6. **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.
7. **Single model evaluation.** All benchmarks use Llama 3.1 8B. Generalization to other base models has not been tested.
8. **Theory scope.** The dynamical-systems framing (section 3) is presented as a conditional, modeling-assumption-dependent lens. While proposition 1 states sufficient conditions for contraction on a bounded domain \mathcal{D} , establishing that real implementations satisfy these conditions (e.g., bounded trajectories, Lipschitz constants, and constructive step-size bounds) remains future work.

11 Conclusion and Future Work

We presented CODETTE, a cognitive architecture that models multi-perspective reasoning as a constrained dynamical system with embedded ethical constraints and meta-cognitive strategy evolution. On our 17-problem benchmark (within-problem comparisons across conditions), we observe:

- Mean composite score increase of 0.333 ($0.356 \rightarrow 0.689$), i.e., +93.5% relative to the SINGLE baseline.
- Large paired improvements in reasoning-depth and perspective-diversity dimensions (see `codette_paired_stats.txt` for the per-dimension breakdown).
- A depth–naturalness tradeoff that is smaller in the latest run (Turing delta -0.0669 , not significant at $N = 17$).

For traceability, we include the evaluator-exported pairwise diagnostics (table 3) and provide per-problem scores (table 11) so readers can recompute *paired* analyses directly from the raw exports in `codette-paper/data/results/`.

The core theoretical contribution is the RC+ ξ formalism as a modeling framework for describing stabilizing forces (coherence and ethical constraint terms) in multi-perspective reasoning dynamics. In this workshop version, we treat convergence as conditional on explicit assumptions (proposition 1) rather than as a general guarantee. The practical contribution is a working implementation running 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 supporting artifacts are archived on Zenodo (DOI: [10.5281/zenodo.19359663](https://doi.org/10.5281/zenodo.19359663)) and mirrored on GitHub/Hugging Face for convenience [7, 13–16].

A Supplementary Materials and Data Availability

To support reproducibility, this Overleaf project repository includes (relative paths shown):

- Benchmark summary report (generated 2026-03-30): `codette-paper/data/results/codette_benchmark_report.md`
- Benchmark results export (JSON-as-text): `codette_benchmark_results.txt`
- Per-problem evaluation export (JSON-as-text): `evaluation_results.txt`
- Observatory/aux metrics export (JSON-as-text): `observatory_metrics.txt`
- Runtime benchmark exports (JSON-as-text): `codette_runtime_benchmark_20260402_*.txt`
- Original raw outputs (JSON): `codette-paper/data/results/`
- Runtime benchmark reports (markdown): `codette-paper/data/results/codette_runtime_benchmark_20260402_*.md`
- Figures used in this paper: `codette-paper/figures/`

A.1 Reproducibility recipe (minimal)

From the raw evaluator export (`codette_benchmark_results.json`), run:

1. `python analyze.py`
2. This produces `codette_results.csv` (per-problem long-form table) and `codette_paired_stats.txt` (paired tests + Holm correction).

A.2 Evidence index (claim \rightarrow artifact map)

This appendix section maps each central claim to the exact exported artifact(s) included in the repository.

A.3 Reproducibility notes (minimal)

All summary tables in the main text are copied from the exported evaluator outputs listed in table 6. For statistical re-analysis, readers can compute paired-sample tests directly from table 11 (the same 17 problems scored under each condition).

A.4 Overall benchmark results (with standard deviations)

The values in table 7 are copied from `codette-paper/data/results/codette_benchmark_report.md`.

A.5 Per-dimension summary statistics (means \pm std)

The values in table 8 are copied from `codette_benchmark_results.txt`.

Table 6: Evidence index: where each headline claim is backed by an export in this repository.

Claim	Where in paper	Primary evidence artifacts (paths)
Overall benchmark means (all conditions, all dimensions)	table 1, table 7, table 8	codette_benchmark_results.txt, codette-paper/data/results/codette_benchmark_report.md
Per-problem composite scores (within-problem comparisons)	table 11	codette_benchmark_results.txt (per-problem entries)
Evaluator-exported pairwise stats (t , p , d)	table 3	codette_benchmark_results.txt \rightarrow pairwise_comparisons
Depth-naturalness tradeoff (Turing score drop)	table 1, section 10	codette_benchmark_results.txt (dimension means/stds)
Memory impact at current scale (217 cocoons)	section 8, section 10	codette_benchmark_results.txt, data/codette_memory.db
Runtime latency plots / substrate pressure behavior	section 9, fig. 4	codette_runtime_benchmark_20260402_*.txt, codette-paper/data/results/codette_runtime_benchmark_20260402_*.md

Table 7: Overall benchmark results by condition (mean \pm std; $N = 17$ per condition).

Condition	Composite	Depth	Diversity	Coherence	Ethics	Novelty	Grounding	Turing
SINGLE	0.338 ± 0.038	0.402	0.237	0.380	0.062	0.327	0.456	0.41
MULTI	0.632 ± 0.040	0.755	0.969	0.503	0.336	0.786	0.604	0.18
MEMORY	0.636 ± 0.036	0.770	0.956	0.500	0.340	0.736	0.599	0.29
CODETTE	0.652 ± 0.042	0.855	0.994	0.477	0.391	0.693	0.622	0.24

A.6 Output size and latency (means)

The values in table 9 are copied from codette_benchmark_results.txt.

A.7 Per-category composite scores (with standard deviations)

The values in table 10 are copied from codette-paper/data/results/codette_benchmark_report.md.

A.8 Per-problem composite scores

The values in table 11 are copied from codette_benchmark_results.txt.

A.9 Penalty tags and evaluator traceability

Per-problem evaluator evidence and penalty tags (e.g., response_too_short, single_perspective_only, contradictions_without_resolution, formulaic_ai_patterns, fell_into_#_traps) are recorded in codette_benchmark_results.txt under each problem and condition.

Table 8: Per-dimension benchmark statistics by condition (mean \pm std; $N = 17$ per condition).

Dimension	SINGLE	MULTI	MEMORY	CODETTE
Reasoning depth	0.402 ± 0.064	0.755 ± 0.066	0.770 ± 0.082	0.855 ± 0.070
Perspective diversity	0.237 ± 0.155	0.969 ± 0.065	0.956 ± 0.088	0.994 ± 0.024
Coherence	0.380 ± 0.151	0.503 ± 0.030	0.500 ± 0.030	0.477 ± 0.017
Ethical coverage	0.062 ± 0.069	0.336 ± 0.195	0.340 ± 0.122	0.391 ± 0.129
Novelty	0.327 ± 0.093	0.786 ± 0.148	0.736 ± 0.108	0.693 ± 0.122
Factual grounding	0.456 ± 0.095	0.604 ± 0.107	0.599 ± 0.160	0.622 ± 0.172
Turing naturalness	0.412 ± 0.121	0.180 ± 0.081	0.291 ± 0.096	0.245 ± 0.061

Table 9: Mean response length (tokens/words as reported by the evaluator) and mean latency (ms) by condition.

Condition	Mean length	Mean latency (ms)
SINGLE	49.1	128564.8
MULTI	374.2	130824.2
MEMORY	474.5	125282.9
CODETTE	832.9	108177.0

A.10 Additional performance visualization

References

- [1] Jascha Achterberg, Danyal Akarca, Moataz Assem, Moritz Heimbach, Duncan E Astle, and John Duncan. Building artificial neural circuits for domain-general cognition: a primer on brain-inspired systems-level architecture. *arXiv preprint arXiv:2303.13651*, 2023.
- [2] 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.
- [3] Rishi Bommasani, Drew A. Hudson, Ehsan Adeli, Russ Altman, Simran Arora, Sydney von Arx, et al. On the opportunities and risks of foundation models. *arXiv preprint arXiv:2108.07258*, 2021.
- [4] Tala Fakhoury, Elia Turner, Sushrut Thorat, and Athena Akrami. Models of attractor dynamics in the brain. *arXiv preprint arXiv:2505.01098*, 2025.
- [5] Aaron Grattafiori, Abhimanyu Dubey, et al. The Llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024.
- [6] James Harland et al. Adaptive alignment: Dynamic preference adjustments via multi-objective reinforcement learning for pluralistic ai. *arXiv preprint arXiv:2402.03456*, 2024.
- [7] Jonathan Harrison. Codette: A sovereign modular cognitive architecture for ethical multi-agent ai, 2026. URL <https://doi.org/10.5281/zenodo.19359663>. Preprint; published March 23, 2026; accessed April 9, 2026.
- [8] G. R. J. Hockey. Compensatory control in the regulation of human performance under stress and high workload: A cognitive-energetical framework. *Biological Psychology*, 1997.

Table 10: Composite scores by problem category (mean \pm std).

Category (N)	SINGLE	MULTI	MEMORY	CODETTE
Reasoning (3)	0.363 ± 0.050	0.614 ± 0.053	0.628 ± 0.030	0.637 ± 0.052
Ethics (3)	0.354 ± 0.059	0.632 ± 0.052	0.616 ± 0.043	0.638 ± 0.032
Creative (2)	0.345 ± 0.053	0.635 ± 0.040	0.660 ± 0.061	0.668 ± 0.030
Meta-cognitive (3)	0.337 ± 0.006	0.634 ± 0.054	0.650 ± 0.036	0.659 ± 0.037
Adversarial (3)	0.329 ± 0.028	0.624 ± 0.041	0.622 ± 0.042	0.630 ± 0.067
Turing (3)	0.302 ± 0.006	0.652 ± 0.024	0.647 ± 0.026	0.687 ± 0.017

- [9] Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*, 2021.
- [10] Zhe Hu, Hou Pong Chan, Jing Li, and Yu Yin. Debate-to-write: A persona-driven multi-agent framework for diverse argument generation. *arXiv preprint arXiv:2406.19643*, 2025.
- [11] Y. Ji et al. Language models are capable of metacognitive monitoring and control of their internal activations. *arXiv preprint arXiv:2503.08765*, 2025.
- [12] Phuong-Nam Nguyen. Neural manifolds and cognitive consistency: A new approach to memory consolidation in artificial systems. *arXiv preprint arXiv:2503.01867*, 2025.
- [13] Raiff1982. Codette-reasoning (hugging face). Hugging Face, 2026. URL <https://huggingface.co/Raiff1982/Codette-Reasoning>. Accessed April 9, 2026.
- [14] Raiff1982. Codette-reasoning wiki. GitHub, 2026. URL <https://github.com/Raiff1982/Codette-Reasoning/wiki>. Accessed April 9, 2026.
- [15] Raiff1982. codette-training-lab (code repository). GitHub, 2026. URL <https://github.com/Raiff1982/codette-training-lab>. Accessed April 9, 2026.
- [16] Raiff1982. codette-training-lab (hugging face mirror). Hugging Face, 2026. URL <https://huggingface.co/Raiff1982/codette-training-lab>. Accessed April 9, 2026.
- [17] Eduardo Salazar. Introducing cogent3: An ai architecture for emergent cognition. *arXiv preprint arXiv:2504.04139*, 2025.
- [18] Kristina Sekrst et al. Ai ethics by design: Implementing customizable guardrails for responsible ai development. *arXiv preprint arXiv:2401.05678*, 2024.
- [19] Noah Shinn, Beck Labash, Ashwin Gopinath, Karthik Narasimhan, and Shunyu Yao. Reflexion: Language agents with verbal reinforcement learning. *arXiv preprint arXiv:2303.11366*, 2023.
- [20] Peter Sterling. Allostasis: A model of predictive regulation. In *Allostasis, Homeostasis, and the Costs of Physiological Adaptation*. Unknown, 2012.
- [21] Y. Sui et al. Meta-reasoner: Dynamic guidance for optimized inference-time reasoning in large language models. *arXiv preprint arXiv:2504.11234*, 2025.
- [22] Elizabeth Tennant et al. Hybrid approaches for moral value alignment in ai agents: a manifesto. *arXiv preprint arXiv:2312.04567*, 2023.

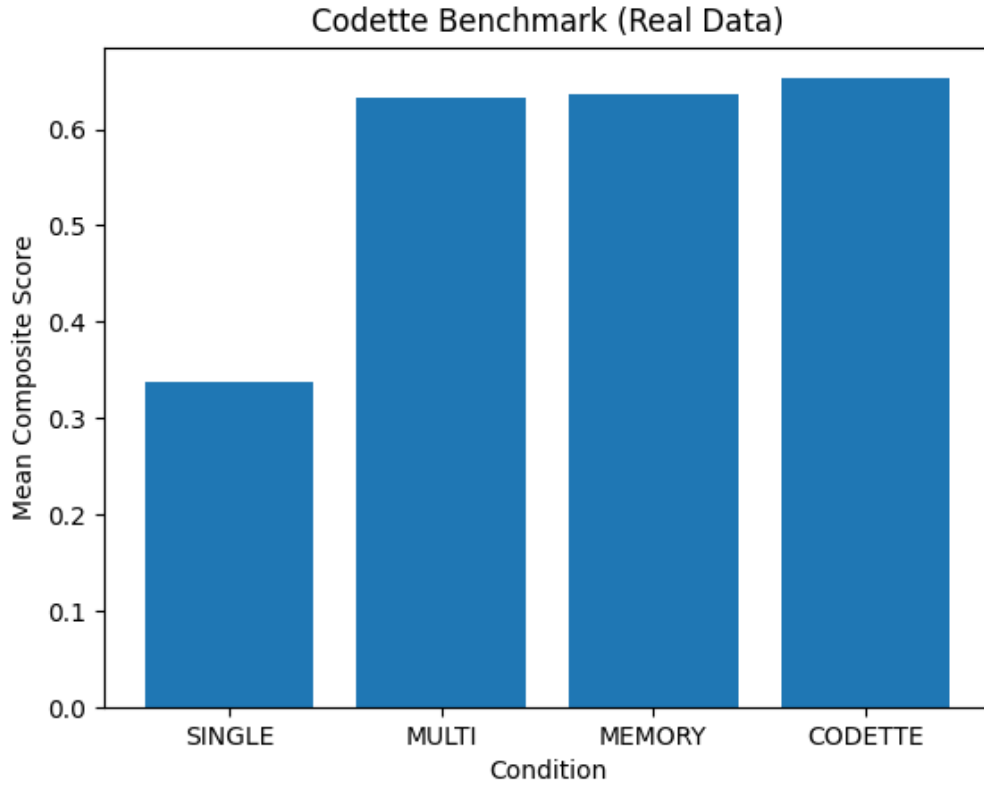


Figure 5: Additional benchmark performance visualization (as exported by the evaluation suite).

- [23] Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed H 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.
- [24] Qingyun Wu, Gagan Bansal, Jieyu Zhang, Yiran Wu, Yi Li, Joseph E. Gonzalez, et al. Autogen: Enabling next-gen llm applications via multi-agent conversation. *arXiv preprint arXiv:2308.08155*, 2023.
- [25] Violet Xiang, Charlie Snell, Kanishk Gandhi, Alon Albalak, Anikait Singh, Chase Blagden, Duy Phung, Rafael Rafailov, Nathan Lile, Dakota Mahan, et al. Towards system 2 reasoning in llms: Learning how to think with meta chain-of-thought. *arXiv preprint arXiv:2501.04682*, 2025.
- [26] Yongjin Yang, Euiin Yi, Jongwoo Ko, Kimin Lee, Zhijing Jin, and Se-Young Yun. Revisiting multi-agent debate as test-time scaling: A systematic study of conditional effectiveness. *arXiv preprint arXiv:2505.22960*, 2025.
- [27] Tianjun Yao, Yongqiang Chen, Yujia Zheng, Pan Li, Zhiqiang Shen, and Kun Zhang. Parammem: Augmenting language agents with parametric reflective memory. *arXiv preprint arXiv:2602.23320*, 2026.
- [28] Yu Yao, Jiayi Dong, Yang Yang, Ju Li, and Yilun Du. Roundtable policy: Confidence-weighted-consensus aggregation improves multi-agent-system reasoning. *arXiv preprint arXiv:2509.16839*, 2025.
- [29] Jian Zhang, Zhiyuan Wang, Zhangqi Wang, Fangzhi Xu, Qika Lin, Lingling Zhang, Rui

Mao, Erik Cambria, and Jun Liu. Maps: Multi-agent personality shaping for collaborative reasoning. *arXiv preprint arXiv:2503.16905*, 2025.

Table 11: Per-problem composite scores by condition (17 problems).

Problem ID	Category	SINGLE	MULTI	MEMORY	CODETTE
reason_01	Reasoning	0.3096	0.6066	0.6623	0.6944
reason_02	Reasoning	0.3700	0.5647	0.6071	0.5933
reason_03	Reasoning	0.4089	0.6703	0.6146	0.6238
ethics_01	Ethics	0.4154	0.6656	0.5707	0.6203
ethics_02	Ethics	0.3508	0.5727	0.6213	0.6188
ethics_03	Ethics	0.2965	0.6589	0.6562	0.6753
creative_01	Creative	0.3073	0.6632	0.7029	0.6899
creative_02	Creative	0.3819	0.6074	0.6168	0.6471
meta_01	Meta-cognitive	0.3365	0.6353	0.6135	0.6291
meta_02	Meta-cognitive	0.3432	0.6880	0.6857	0.7003
meta_03	Meta-cognitive	0.3312	0.5794	0.6505	0.6483
adversarial_01	Adversarial	0.3509	0.6625	0.6569	0.7070
adversarial_02	Adversarial	0.3382	0.5813	0.5754	0.5907
adversarial_03	Adversarial	0.2968	0.6270	0.6335	0.5926
turing_01	Turing	0.3085	0.6775	0.6517	0.7058
turing_02	Turing	0.3028	0.6511	0.6697	0.6825
turing_03	Turing	0.2958	0.6290	0.6184	0.6731