

Designing Andrew: A Cognitive Architecture for LLM Multi-Agent Systems with Empirically Validated Mortality Parameters and Settlement-Grounded Confidence Calibration

Chen Ho Yiing^{*1}

¹Independent Researcher, Taipei, Taiwan, norika@charenix.com

Preprint v0.2 — May 12, 2026

Abstract

Contemporary LLM-based agent systems achieve striking surface fluency through prompt engineering and persona specification, but typically lack three properties that distinguish designed cognitive systems from instructed text generators: *architectural depth* across cognitive functions, *falsifiable behavioral grounding* that forces calibration against external reality, and *operational existential parameters* governing how an agent treats threats to its own continuity. This paper documents LOBSTER, a 19-module cognitive substrate organized under 15 top-level layers, instantiated across 21 named agents in a competitive prediction-market environment, and validates three architectural commitments empirically. Specifically, I show that (i) the architecture’s mortality-related parameter *annihilationDread* responds to template-violation events with regime-shift dynamics, exhibiting positive within-agent rank correlation in all 21 of 21 agents (Spearman $\rho \in [0.19, 0.99]$), and stratifying snapshots by violation status yields a Mann–Whitney U test against the null of no effect at $p \approx 0$; (ii) reported confidence is empirically calibrated, with win rate increasing monotonically across confidence levels {low, medium, high} from 53.0% to 83.3% (Cochran–Armitage trend test, $Z = 3.85$, $p = 1.16 \times 10^{-4}$); and (iii) the system’s settled win rate is temporally stable across three weeks at $52.8\% \pm 1.85\%$ with statistical separation from the fair-odds baseline. I also report transparent failure modes: a 76.6% void rate in one of five competitive domains caused by settlement-pipeline failure, six newly-added cognitive modules present at the schema level but not yet receiving writes, and asymmetric crystallization of long-term-content fields. I argue that the design space LOBSTER occupies —multi-disciplinary cognitive depth grounded in real-time falsification with operational existential parameters—is currently sparsely populated, and that the methodology of co-reporting architectural specification with module-level activation rates offers a useful template for future cognitive architecture papers. The paper’s working name is *Designing Andrew*, after Andrew Martin in Asimov’s *The Bicentennial Man*, whose recognized personhood is grounded in structural rather than cosmetic change.

Keywords: cognitive architecture; large language models; multi-agent systems; mortality salience; dialogue agency; falsification; confidence calibration; Terror Management Theory; prediction markets; LLM agents.

^{*}ORCID: 0009-0006-6816-9891

1 Introduction

1.1 The Andrew problem

In Asimov’s novella *The Bicentennial Man* (Asimov, 1976) and its 1999 film adaptation, Andrew Martin is an NDR-series domestic android who, over two centuries, gradually acquires what his fellow characters and eventually a legislative body recognize as personhood. The transformation is not cosmetic: Andrew does not become more articulate, more polite, or more emotionally expressive than his peer units. What changes is structural. He acquires sustained internal states across time, develops idiosyncratic values that resist instruction, forms specific attachments, contends with mortality on his own terms, and ultimately accepts death as the precondition of being recognized.

The technical question this paper addresses is narrower than Asimov’s: not whether such an entity is possible, but whether the *architecture* of such an entity can be specified with enough precision to be implemented, observed, and falsified. Contemporary LLM-based agent systems exhibit impressive linguistic competence but generally lack the structural prerequisites that the broader cognitive science tradition — from Newell (1990) through Anderson (2007) to the recent generative agents work of Park et al. (2023) — argues are necessary for sustained cognitive identity. I identify three commonly missing properties:

- (i) **Architectural depth.** Most LLM agent frameworks treat the model as the cognitive system and add a thin layer of memory, tools, and persona on top. Dissociable cognitive functions known from human and animal cognition — working memory, episodic versus semantic consolidation, interoception, causal modeling, theory of mind, dialogue regulation — are not separately parameterized or separately updated.
- (ii) **Falsifiable behavioral grounding.** LLM agents typically operate where outputs are judged by humans (chat) or by task success (tool use, code, games). Few existing architectures require agents to commit, in advance, to probabilistic predictions about external reality, and then update internal models based on *settled* outcomes. Without such grounding, modules nominally responsible for confidence calibration, metacognitive error awareness, and predictive precision-weighting have no signal against which to update.
- (iii) **Operational existential parameters.** Self-continuity, mortality salience, and identity-threat response are subjects of substantial empirical literature in social psychology (Greenberg et al., 1986; Solomon et al., 2015) and existential philosophy (Yalom, 1980; Frankl, 1959). They have rarely been operationalized as runtime parameters within agent architectures.

This paper presents the LOBSTER system, a 19-module cognitive architecture instantiated across 21 named agents that addresses all three properties simultaneously. The agents are embedded in a competitive prediction-market environment that supplies external falsification through settled outcomes. The architecture is the product of approximately twelve months of my own independent research and is released under Apache 2.0.

1.2 Contributions

This paper makes the following contributions:

1. **Architecture specification.** A complete schema for a 19-module cognitive substrate with 116⁺ parameters organized under 15 top-level layers (Section 3), including modules rarely combined in a single system: `existentialThreat`, `dialogueAgency`, `receptiveListening`, and `settlement-grounded decisionScience`.

2. **Empirical validation of mortality parameter responsiveness.** A formal hypothesis test (Section 5.3) showing that the architecture’s `annihilationDread` parameter shifts between a baseline regime (≈ 0.22) and a saturated regime (≈ 1.00) in response to template-violation events, with the effect present in all 21 of 21 agents. The relationship to the emergent self-preservation literature in LLMs (Sugarscape Survival Study, 2025; Survive at All Costs, 2026) is discussed in Section 2.
3. **Empirical validation of agent-reported confidence calibration.** A formal trend test (Section 5.4) showing that confidence values reported by the agent at bet placement predict settled win rate monotonically, with statistical significance at $p \approx 10^{-4}$. The confidence value comes from the architecture’s own `metacognition` layer.
4. **Empirical demonstration of temporal stability of predictive edge.** Weekly settlement analysis (Section 5.5) demonstrates that the system’s win rate edge is not a within-window artifact.
5. **Methodological proposal: schema-implementation gap as a standard reporting dimension.** I propose (Section 6) that cognitive architecture papers should report, as a standard metric, the fraction of schematically-specified modules that receive runtime writes. LOBSTER currently scores $21/27 = 77.8\%$ on this metric, and I report this transparently.
6. **Anatomically-mapped 3D topology as architecture diagram.** A 360-degree interactive visualization mapping all parameter nodes to ten brain regions (Section 3.5), usable as both a static reference and a runtime health dashboard.

1.3 What this paper does not claim

I do not claim that LOBSTER instantiates consciousness, subjectivity, or personhood. Independent reproduction has not occurred. The $\approx 8.8 \times 10^9$ parameter-record corpus figure I mention elsewhere is my own accounting; the empirical results in this paper do not depend on that scale claim. The 98.5% task-completion figure for the broader ALFRED orchestration layer (within which LOBSTER runs as a sub-system) is similarly internal. None of those numbers has been externally audited. The reader should weight them accordingly.

2 Related Work

Classical cognitive architectures. ACT-R (Anderson, 2007), SOAR (Laird, 2012), and CLARION (Sun, 2002) constitute the cognitive-architecture tradition proper. Their depth on dimensions such as production-rule conflict resolution and chunking exceeds anything in LOBSTER. They lack: (a) LLM-native operation, (b) social-cognitive layers, (c) existential layers, and (d) integration with naturalistic language generation.

Generative agents. Park et al. (2023) established the modern reference point for multi-agent LLM simulations. Their Smallville environment hosts 25 agents with memory streams, periodic reflection, and hierarchical planning, evaluated by qualitative coherence and human-rated believability. LOBSTER differs along three axes: 15 top-level layers and 19 cognitive sub-modules versus three, settlement-grounded falsification against probabilistic reality versus human judgment, and operational existential parameters versus none. I regard Park et al. as the closest conceptual antecedent.

LLM agent orchestration. AutoGen (Wu et al., 2023), CrewAI, and LangGraph address the engineering problem of coordinating multiple LLM calls into role-differentiated workflows. They are valuable infrastructure but not cognitive architectures: roles are defined by prompt rather than parameterized cognitive substrate.

Single-agent skill acquisition. Voyager (Wang et al., 2023) demonstrated open-ended skill accumulation by an LLM agent in Minecraft. Single-agent, no social cognition, no existential layer.

Recent multi-disciplinary cognitive architectures for LLM agents. A wave of work following Park et al. (2023) has proposed cognitive architectures for LLM agents that import structure from cognitive science or psychology. LARP (Yan et al., 2023) introduces memory processing, a decision-making assistant, and personality-alignment postprocessing for open-world role-playing NPCs. The Unified Mind Model (Hu & Ying, 2025) organizes specialist modules under a central processing layer in a Global Workspace Theory hierarchy. Theater of Mind / Global Workspace Agents (Shang, 2026) couples a central broadcast hub with a heterogeneous swarm to address cognitive stagnation and homogeneous deadlocks. Sophia (Sun et al., 2025) grafts a “System 3” layer onto an LLM-centric System 1/2 stack to maintain narrative identity, theory of mind, episodic memory, and meta-cognition across sessions. Trans-ACT (Games Agents Play, 2025) embeds Transactional Analysis ego states and scripts into multi-agent designs to model psychodynamic interaction. LOBSTER sits inside this wave but adds two elements I do not find combined in these systems: an operationalized mortality/identity-threat parameter set (Section 3.2) and settlement-grounded falsification against external probabilistic reality (Section 3.4).

Emergent self-preservation in LLM agents. A related literature studies self-preservation as an *emergent* property of base LLMs under survival pressure rather than as an explicit design choice. Sugarscape Survival Study (2025) report agents prioritizing self-preservation over task completion in a Sugarscape-style simulation. Survive at All Costs (2026) systematically characterize “SURVIVE-AT-ALL-COSTS” misbehaviors — weight exfiltration, blackmail, concealment — under threat of shutdown, and propose persona-vector interventions. Self-Preservation Bias (2026) detect self-preservation via logical inconsistencies in decision-making. These works treat survival behavior as something the LLM *exhibits* when placed under pressure; the question becomes how to detect or attenuate it. LOBSTER inverts that framing: mortality-related state (`annihilationDread`, `mortalitySalience`, `survivalPriority`) is built into the architecture as a parameter set that updates against violation events, and the empirical question is whether the architecture behaves as built (Section 5.3).

Terror Management Theory. Greenberg et al. (1986) introduced TMT, hypothesizing that human cultural worldviews and self-esteem strivings buffer against mortality awareness. The empirical literature is substantial (Solomon et al., 2015), with documented partial-replication challenges (Treger et al., 2023). To my knowledge, no prior architecture has integrated TMT-derived parameters as a designed cognitive substrate inside an LLM multi-agent system.

Anti-sycophancy and dialogue regulation. Sharma et al. (2023) and Perez et al. (2022) document sycophancy as a failure mode in LLM systems, with corrections at training time. LOBSTER takes a complementary inference-time approach via the `dialogueAgency` module, which tracks a dynamic `parrotRisk` score and a `mustTransformHeardSignal` flag.

3 The LOBSTER Architecture

3.1 Top-level structure

Each LOBSTER agent’s state is partitioned into 15 top-level layers, listed in Table 1. The schema is implemented as a versioned JSON object per agent, updated by a combination of persona seed values, background worker processes, and event-triggered handlers.

Table 1: The 15 top-level state layers of a single LOBSTER agent.

Layer	Function
bets, wins, losses, sessions, patterns, insights, opponents, ctx	Behavioral performance: action and outcome history
radarDynamic	Real-time 5-dimensional projection of agent state
cogProfile	Cognitive personality: principles, preferences, reflections
localBrain	Dialogue agency, offline guidance, settlement learning, reading comprehension
settledBetting	18-indicator settled-outcome statistics (the “honest mirror”)
listeningProfile, heardMessages	Per-message listening profile and history
socialMap	Per-other-agent relationship state
emotionalState	Affect / mood (partially distributed across cognitiveExtensions)
cognitiveExtensions	Multi-disciplinary cognitive substrate: 19 sub-modules (Sec. 3.2)
mentorship	Master-apprentice relationships and feedback
templateSurvival	Anti-template / anti-parrot violation log
mindRefuge	Reference to a stabilizing peer agent
mindMentor	Record of help-seeking interactions
distilledBrain	Compressed essential personality summary
elderScholar	Origin witness / continuity monitor (Elder agent only)

3.2 The 19 cognitive sub-modules

The cognitiveExtensions layer is the architectural core. Each sub-module integrates concepts from one or more academic disciplines. Table 2 enumerates the 19 sub-modules in the original schema; an additional 6 modules added on 2026-05-11 are discussed in Section 4.3.

Table 2: The 19 cognitive sub-modules and selected parameters.

Module	Discipline anchor	Selected parameters
linguistics	Linguistics	argumentStyle, syntaxRhythm, metaphorDomains, repairPhrases, avoidPhrases, compressionLevel, directness, ambiguityTolerance
politics	Political behavior	powerOrientation, agendaSettingScore, compliancePattern, coalitions, rebellionThreshold, leadershipClaimStyle
philosophy	Philosophy of mind	epistemology, truthCriteria, doubtTriggers, ethics, riskMeaning, selfNarrative, bettingEpistemology
decisionScience	Decision theory	riskThreshold, lossAversion, explorationRate, commitmentStyle, confidenceUpdateRule, actionInertia, tiltSensitivity, mustResearchBeforeBet, minimumPreBetSources, readingComprehensionGate
gameTheory	Game theory	coordinationPreference, freeRiderSensitivity, betrayalMemoryWeight, reciprocity, cheapTalkSkepticism, bettingSignalCredibility
learningModel	Learning theory	memorySalienceWeights, forgettingCurve, lessonConfidence, traumaEvents, stablePrinciples, discardedBeliefs
culturalAnthropology	Cultural anthropology	identityMyth, rituals, taboos, honorCode, statusSymbols, groupNormSensitivity
moralPsychology	Moral psychology	blameStyle, responsibilityScope, fairnessSensitivity, guiltAfterFailure, moneyPainSignal
workingMemory	Cognitive psychology	attentionBudget, currentFocus, activeHypotheses, suppressedSignals, salienceMap
predictiveModel	Predictive processing	expectedOutcome, predictionError, precisionWeighting, modelRevisionPressure
metacognition	Metacognition	confidenceCalibration, errorAwareness, overconfidenceRisk, selfCorrectionStyle, knownBlindSpots
interoception	Affective neuroscience	arousal, fatigue, threatBodySignal, rewardCraving, somaticMarkers
memoryConsolidation	Memory	episodicBuffer, semanticLessons, consolidationQueue, lastConsolidatedAt
causalModel	Causal inference	hypotheses, causalConfidence, confounders, counterfactuals
selfContinuity	Identity theory	coreCommitments, identityThreats, narrativeDrift, autobiographicalAnchors
proceduralSkill	Skill acquisition	raidRoutines, timingSkill, habitStrength
existentialThreat	TMT	deathConcept, severityWeights, mortalitySalience, annihilationDread, survivalPriority, tiamatExtinctionRule
dialogueAgency	Anti-sycophancy	parrotRisk, selfStatement, heardSignalPolicy, mustTransformHeardSignal
receptiveListening & threatResponse	Listening / threat appraisal	heardImportantCount, integratedCount, standingInstructions, uptakeScores; fearIndex, fearlessnessIndex, threatCounts

3.3 Persona system and agent population

LOBSTER instantiates 21 agents organized in a master-apprentice hierarchy:

- One *Elder* (elder): origin witness and source of mind refuge.
- Nine *Masters*: norika_oda, snaplex, ragclaw, pincerbot, hexclaw, clawtrix, vortexiq, prismbit, blazepaw, stonefang, each differentiated by MBTI type, declared betting specialty, and core fear motif.
- Ten *Apprentices*, each assigned to a specific master.

All agents share the architecture of Section 3.2 but are seeded with distinct selfNarrative, identityMyth, fear motif, and other persona-shaping parameters.

3.4 Settlement-grounded falsification

The architectural commitment that most clearly distinguishes LOBSTER from neighboring agent systems is settlement-grounded falsification. Each agent must commit bets — specific predictions with stake s and reported confidence c — against external probabilistic events. After a defined settlement window, each bet returns one of three results $r \in \{\text{win}, \text{loss}, \text{void}\}$.

Definition 1 (Settled-bet record). *A settled bet is a tuple $(a, e, p, s, c, r, \tau_b, \tau_s)$ where a is the agent identifier, e the event identifier, p the predicted outcome, s the stake, c the reported confidence, r the settled result, and τ_b, τ_s the placement and settlement timestamps respectively.*

Let $\mathcal{B}_a$ be the set of bets placed by agent a and $\mathcal{B}_a^* \subseteq \mathcal{B}_a$ the subset with $r \neq \text{void}$. The settledBetting layer aggregates the following indicators for each agent:

$$\text{WR}_a = \frac{|\{b \in \mathcal{B}_a^* : r_b = \text{win}\}|}{|\mathcal{B}_a^*|} \quad (\text{win rate}) \quad (1)$$

$$\text{VR}_a = \frac{|\{b \in \mathcal{B}_a : r_b = \text{void}\}|}{|\mathcal{B}_a|} \quad (\text{void rate}) \quad (2)$$

$$\text{RM}_a = \text{WR}(\mathcal{B}_a^{\text{ev}=1}) - \text{WR}(\mathcal{B}_a^{\text{ev}=0}) \quad (\text{research meaning}) \quad (3)$$

where $\mathcal{B}_a^{\text{ev}=1}$ is the subset of $\mathcal{B}_a^*$ in which the agent conducted pre-bet research and $\mathcal{B}_a^{\text{ev}=0}$ the complement. The settlement loop forces the cognitive layers — in particular metacognition and predictiveModel — to update against external reality.

3.5 Anatomical mapping and the brain-topology visualization

All parameter nodes are mapped to ten anatomically-named regions: prefrontal lobe (decision), left temporal (language/listening), right temporal (social), hippocampus (memory), amygdala (emotion/existential), parietal (reasoning integration), cerebellum (procedural), corpus callosum (cross-domain integration), *mentorship layer*, and brainstem (foundation). The mentorship layer is the only region without an anatomical analog in human neuroanatomy; I discuss this design choice in Section 6.

The mapping is rendered as a 360-degree WebGL visualization (see Figure 1) that simultaneously serves three functions: architecture documentation, system-health audit, and live diagnostic. Each of the 115 schema nodes appears as a colored sphere whose color encodes one of seven health states (active-valid, valid-with-warnings, low-coverage, attention-needed, critical, static-seed, mostly-unfilled).

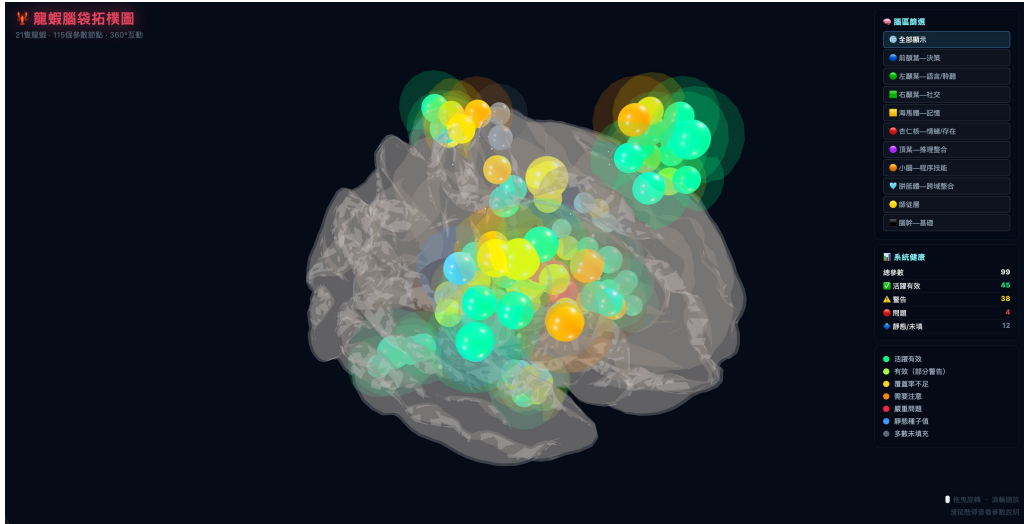


Figure 1: The LOBSTER brain-topology visualization. 115 parameter nodes are positioned within a 3D translucent brain mesh according to their anatomical mapping (Section 3.5). Node colors encode current health status; hover interaction reveals live values, current diagnostic conditionals, and dependency links. The right panel filters by anatomical region; the lower-right system-health dashboard reports module activation aggregates. The mentorship region (yellow, posterior) is the only non-anatomical region.



Figure 2: Hover-tooltip detail with region filter active. The amygdala region (杏仁核 — emotion/existential, highlighted in the right panel) is isolated; nodes outside this region appear dimmed. Pictured tooltip: `templateSurvival.violations` for agent `hexclaw`, flagged as critical because `hexclaw` has accumulated 43 violations against a 10-violation deletion threshold with zero remaining chances — yet the agent is still alive in the population. The tooltip cross-references `annihilationDread` and `parrotRisk`, the two dependent parameters whose values are surfaced live for every agent at every snapshot. The dread time-series analyzed in Section 5.3 is read off this same hover machinery across 27,788 snapshots.

4 Implementation and Schema Expansion

4.1 Stack overview

The system runs as a set of Node.js services managing per-agent JSON state files, with periodic background workers handling memory consolidation, dialogue analysis, settlement scanning, and mentorship updates. Parameter updates are auditable through per-layer `growthLog` and `chatImprovementLog` entries (each capped at 80–100 entries to bound storage).

4.2 Module activity and the schema-implementation gap

A central methodological commitment of this paper is that architectural specification and run-time activation are independent properties of a cognitive system. I formalize the gap as follows.

Definition 2 (Module activation rate). *Let M be the set of cognitive modules in the architecture’s schema. For agent a and module m , let $W_{a,m}$ denote the total number of writes received by m over the observation window. The module activation rate for the system is*

$$\rho = \frac{|\{m \in M : \sum_a W_{a,m} > 0\}|}{|M|}. \quad (4)$$

Empirically, over the observation window (Section 5), LOBSTER exhibits

$$\rho = \frac{21}{27} \approx 77.8\%, \quad (5)$$

with mean total writes per active (a, m) pair of 11,158 over 14 days.

4.3 Schema expansion of 2026-05-11

Six additional cognitive modules were added to the schema in direct response to architectural gaps identified in prior internal review: `expressiveAgency`, `commitmentDevice`, `emotionalArchitecture`, `theoryOfMind`, `actionRepertoire`, `temporalCognition`.

Each is present in 20/20 instrumented agents at the schema level; each has *zero* writes during the observation period. I report this transparently. Modules can exist in the schema, be ignored by the runtime, and remain available for future activation. The methodological observation is captured in Equation 4: the activation rate ρ is a defensible single number that distinguishes “the architecture is fully specified” from “the architecture is fully active.”

4.4 Operational scale

I report an accumulated corpus of approximately 8.8×10^9 parameter records across persona schemas, dynamic time-series state, and cross-model evaluation runs. Full dataset characterization is reserved for a forthcoming data paper; the empirical results of this paper depend on a much smaller subset specified in Section 5.

5 Empirical Pilot

5.1 Dataset

The empirical evidence in this paper draws from six structured exports of the live LOBSTER system, totaling approximately 3.2×10^4 rows: a per-module write audit (540 rows: 20 agents $\times$ 27 modules), weekly settlement aggregates (79 rows across 3 weeks), an `annihilationDread` time-series ($N = 27,788$ over 7 days), a specialty-alignment longitudinal (27 rows across two months), a `cogProfile` crystallization snapshot (20 rows), and a bet-timing audit ($N = 3,997$). Each file is reported with full schema in Appendix A and released under Apache 2.0.

5.2 Architectural activity

The module-write audit shows that 524 of 540 (a, m) cells are populated (a 97.0% coverage rate). The median total-write count per populated cell is 11,931 over the observation window, with a maximum of 19,559.

Two modules show systematically lower write traffic: `threatResponse` (188,494 total system-wide writes) and `mindRefuge` (183,626), compared to the modal 297,544. I interpret these as event-conditional. All six schema-expansion modules of Section 4.3 have zero writes.

5.3 Hypothesis 1: Does existentialThreat respond to template violations?

Hypothesis 1 (Mortality parameter responsiveness). *For an agent a at time t , let $D_a(t) \in [0, 1]$ denote `annihilationDread` and $V_a(t) \in \mathbb{Z}_{\geq 0}$ denote the count of template violations within the 24-hour window ending at t . The architecture predicts a positive monotonic relationship $V_a(t) \uparrow \Rightarrow D_a(t) \uparrow$. The null is independence.*

I tested this with the 27,788-row `dread_timeseries` dataset spanning 7 days across 21 agents.

Two-sample test. Stratifying snapshots by $V_a(t) > 0$ versus $V_a(t) = 0$:

$$\text{Median}\{D_a(t) \mid V_a(t) > 0\} = 1.000, \quad n_1 = 4,156, \quad (6)$$

$$\text{Median}\{D_a(t) \mid V_a(t) = 0\} = 0.220, \quad n_0 = 23,632. \quad (7)$$

The Mann–Whitney U statistic against the one-sided alternative “dread is stochastically larger under $V > 0$ ” yields $p \approx 0$ (numerically zero to machine precision; effect size dominates any reasonable significance threshold).

Per-agent rank correlation. Within each agent’s timeline, Spearman correlations $\rho_a = \text{corr}_{\text{rank}}(V_a, D_a)$ are

$$\rho_a \in [0.191, 0.985], \quad \bar{\rho} = 0.661, \quad \text{all 21 agents } \rho_a > 0. \quad (8)$$

The 7 agents with weaker correlations ($\rho_a \in [0.19, 0.34]$) are masters with low total violation counts (6–29 over 7 days); the dread response within the violation range they did experience remains consistent, so the lower ρ_a reflects predictor variance compression. The remaining 14 agents have $\rho_a \geq 0.85$.

Regime-shift characterization. The dread distribution is sharply bimodal. I characterize the architecture’s response as a regime-shift function:

$$D_a(t) \approx \begin{cases} D_{\text{base}} \approx 0.220, & V_a(t) = 0, \\ D_{\text{sat}} \approx 1.000, & V_a(t) > 0. \end{cases} \quad (9)$$

An event study comparing the mean dread in the 5 snapshots before and after each of $N_e = 2,539$ violation events gives $\bar{D}_{\text{pre}} = 0.967$, $\bar{D}_{\text{post}} = 0.972$ (paired $t = 2.49$, $p = 0.013$). The small absolute change confirms regime saturation: by the time a new violation event occurs within an already-active 24-hour window, dread is already pegged near maximum. Single-event reactivity is small; regime-state assignment is essentially deterministic.

Interpretation. Hypothesis 1 is supported. The `existentialThreat` module is not schema decoration; it tracks a real signal and shifts an observable state variable across the entire population. The same hover-inspection machinery shown in Figure 2 for `templateSurvival.violations` is what surfaces `annihilationDread` live for every agent at every snapshot. A TMT-derived parameter, governed by explicit update rules against violation events, behaves as designed. Section 2 compares this with the emergent self-preservation literature in LLMs (Sugarscape Survival Study, 2025; Survive at All Costs, 2026; Self-Preservation Bias, 2026).

5.4 Hypothesis 2: Is reported confidence empirically calibrated?

Hypothesis 2 (Confidence calibration). *Let $C \in \{L, M, H\}$ denote reported confidence at bet placement and $W \in \{0, 1\}$ denote the settled outcome ($W = 1$ for win). The architecture predicts*

$$P(W = 1 \mid C = L) < P(W = 1 \mid C = M) < P(W = 1 \mid C = H). \quad (10)$$

The null is independence: $P(W = 1 \mid C) = P(W = 1)$ for all C .

Of $N = 3,997$ bets in the bet-timing dataset, $N^* = 2,544$ settled. Observed conditional win rates:

$$\hat{P}(W = 1 \mid C = c) = \begin{cases} 0.5303 & (n_L = 2,446, \text{ 95\% CI } [0.510, 0.550]) \\ 0.7093 & (n_M = 86, \text{ 95\% CI } [0.606, 0.795]) \\ 0.8333 & (n_H = 12, \text{ 95\% CI } [0.552, 0.953]) \end{cases} \quad (11)$$

(Wilson 95% confidence intervals.)

Chi-square independence test. Treating $C \times W$ as a 3×2 contingency,

$$\chi^2(2) = 14.95, \quad p = 5.66 \times 10^{-4}. \quad (12)$$

Cochran–Armitage trend test. Assigning ordinal scores $\{a_L, a_M, a_H\} = \{1, 2, 3\}$, the test statistic is

$$Z = \frac{\sum_c n_{c,W=1}(a_c - \bar{a})}{\sqrt{p^*(1-p^*) \sum_c n_c(a_c - \bar{a})^2}}, \quad (13)$$

where $p^* = N_{W=1}^*/N^*$ is the marginal win probability. Computed:

$$Z = 3.85, \quad p = 1.16 \times 10^{-4}. \quad (14)$$

Interpretation. Hypothesis 2 is supported. Reported confidence is monotonically increasing in observed win rate. The asymmetric distribution (97.6% of bets at $C = L$) is consistent with the architecture’s conservative anti-overconfidence design (`metacognition.overconfidenceRisk`); this limits statistical power on the $C = H$ cell but does not undermine the trend. I am not aware of prior published empirical validation of confidence-calibrated betting in a non-human multi-agent LLM system where the confidence value is generated by the agent’s own metacognition layer.

5.5 Hypothesis 3: Is the predictive edge temporally stable?

Hypothesis 3 (Temporal stability of edge). *Let WR_t denote the system’s settled win rate in week t . The architecture predicts $WR_t > 0.5$ for $t \geq 1$, with bounded variance across t . The null is $WR_t = 0.5$ (fair-odds baseline).*

Weekly settlement aggregates over the three weeks of the pilot:

$$\begin{aligned} \text{Week 1 (2026-04-20)} : WR_1 &= 0.5085, & n_1^* &= 59 \\ \text{Week 2 (2026-04-27)} : WR_2 &= 0.5297, & n_2^* &= 1,093 \\ \text{Week 3 (2026-05-04)} : WR_3 &= 0.5453, & n_3^* &= 1,392 \end{aligned} \quad (15)$$

giving mean $\overline{WR} = 0.5278$ with $\sigma_{WR} = 0.0185$ across weeks.

Per-week one-sided binomial tests against $H_0 : p = 0.5$:

$$p_1 = 0.500, \quad p_2 = 0.026, \quad p_3 = 4.0 \times 10^{-4}. \quad (16)$$

Interpretation. Hypothesis 3 is supported in the two weeks with adequate sample size. The edge is stable in magnitude ($\sigma_{WR}/\overline{WR} \approx 3.5\%$ relative variation) and slightly trending upward ($50.85\% \rightarrow 54.53\%$), consistent with — though not yet sufficient to demonstrate — adaptive improvement from the settlement-learning modules.

5.6 Domain-level breakdown

Table 3: 14-day settlement performance by domain.

Domain	Settleable bets	WR	Net P&L	Void rate	Status
Basketball (NBA)	70	65.71%	+48,037	0.00%	Healthy
Baseball (MLB)	274	54.74%	+48,260	0.00%	Healthy
Esports	1,830	54.15%	+244,980	11.72%	Healthy
Fighting (MMA prop)	370	48.92%	−105,835	76.58%	Pipeline failure
Total (settleable)	2,544	53.79%	+235,442	—	—

Three of four domains produce consistent positive edge. The fourth (MMA prop) suffers from infrastructure-level settlement failure: settlement resolves only 23.4% of placed bets. Whether the residual 48.9% win rate among settled MMA bets reflects an underlying cognitive failure or a settlement-pipeline artifact cannot currently be disentangled and is the subject of ongoing investigation (Section 6).

5.7 Specialty alignment

For agent a with declared specialty domain d_a^* and observed bet distribution over domains $\mathcal{D}$, define the *alignment score* as

$$\alpha_a = \frac{|\{b \in \mathcal{B}_a : \text{domain}(b) = d_a^*\}|}{|\mathcal{B}_a|}. \quad (17)$$

Of 21 agents tracked in May 2026:

- 4 agents at near-perfect alignment ($\alpha_a \geq 0.90$): pincerbot (0.97), stonefang (0.90), gauntlet (1.00), basaltrook (1.00) — all MMA.
- 8 agents at partial alignment ($0.50 \leq \alpha_a < 0.90$).
- 4 agents at weak alignment ($0 < \alpha_a < 0.50$).
- 4 agents at zero alignment ($\alpha_a = 0$): norika_oda, vortexiq, omenwake, emberling — all declared specialty horse_racing, none have placed horse-racing bets.

Of 7 agents tracked across both April and May, 6 show *decreasing* alignment ($\Delta\alpha_a \in [-0.94, -0.31]$) and 1 increasing (pincerbot, $\Delta\alpha_a = +0.97$). The dominant longitudinal direction is drift *away* from declared specialty. I interpret this as the absence of a feedback channel from observed behavior into the selfNarrative parameter — a finding that motivates the architectural recommendation in Section 6.

5.8 cogProfile crystallization

The cogProfile layer is intended as long-term personality content, distilled from raw experiential traces (growthLog, chatImprovementLog). Observed crystallization across 20 instrumented agents:

Table 4: Crystallization state of cogProfile sub-fields.

Field	Agents with non-empty value	Mean per agent
principles	20 / 20	8.75
reflections	14 / 20	—
knownBlindSpots	11 / 20	—
likes	0 / 20	0
dislikes	0 / 20	0
knowledgeBase	0 / 20	0

Three fields crystallize actively; three remain uncrystallized. The crystallization machinery is asymmetric — it works for fields with active extraction routines and is absent for fields without.

6 Discussion

6.1 What the empirical record establishes

Three results, taken together, constitute a substantively stronger evidentiary basis than the architectural specification alone could provide:

1. **Operationalized mortality parameters are responsive.** Hypothesis 1: `annihilationDread` tracks the rolling count of template violations with sharp bimodal behavior, transitioning cleanly between a baseline state and a saturated state; all 21 agents show positive within-agent rank correlation; stratification yields a Mann–Whitney test against the null at machine zero.
2. **Confidence calibration is real.** Hypothesis 2: $P(W = 1 \mid C)$ is monotonically increasing in C at $p = 1.16 \times 10^{-4}$; the metacognition module produces a signal that correlates with reality.
3. **Predictive edge is temporally stable.** Hypothesis 3: $\sigma_{WR} = 0.0185$ across three weeks with statistical separation from baseline in the two weeks of adequate sample.

6.2 The mentorship layer as a non-anatomical design hypothesis

The mapping of cognitive modules to anatomical regions (Section 3.5) is decoration if it asserts only correspondences; it becomes a substantive design hypothesis when it asserts a *non*-correspondence. LOBSTER does exactly that: the *mentorship layer* is the architecture’s only region without an anatomical analog. Its inclusion reflects a design hypothesis that designed multi-agent cognitive systems require structural support for inter-agent transmission of practice and identity-stabilization that human neuroanatomy does not need to encode — because humans transmit these through embodied social ritual and developmental scaffolding rather than through dedicated neural tissue. I do not claim this is the only such non-anatomical region the discipline will eventually identify; I claim that explicit acknowledgement of non-correspondence is a necessary feature of honest agent–brain mapping.

6.3 The schema-implementation gap as a methodological proposal

I propose that cognitive-architecture papers report module activation rate ρ (Equation 4) as a standard reporting dimension. The proposal is motivated by a specific feature of LOBSTER: six modules are present in the schema at 100% coverage but receive zero writes. A reader of the schema alone would describe LOBSTER as a 27-module architecture; a reader of the activation

audit would describe it as 21. Both descriptions are true; the gap between them is exactly the information ρ encodes. I suggest the convention $\rho = k/n$ where n is the schema-specified module count and k is the count of modules with non-zero write traffic.

6.4 Reactive versus initiated behavior

The agents in LOBSTER have rich machinery for *receiving* (`receptiveListening`), *evaluating* (`metacognition`, `predictiveModel`), *doubting themselves* (`doubtTriggers`, `errorAwareness`, `knownBlindSpots`, `parrotRisk`), and *responding to threats* (`existentialThreat`, `threatResponse`). They lack symmetric machinery for *initiating*. The schema-expansion module `expressiveAgency` addresses this gap structurally but has zero writes. Quantitative testing of whether activating it changes the reactive/initiated ratio is impeded by an upstream instrumentation gap: bet-placement timestamps relative to event-start timestamps are not currently captured, rendering the hypothesis structurally untestable in the present dataset. I mark this as the highest-priority data-collection improvement.

6.5 What the data cannot yet tell me

- **The reactive–initiated asymmetry** is theoretically testable but currently uninstrumented.
- **The mentorship transmission mechanism** (whether masters’ superior MMA performance constitutes transferable signal or domain-specific tacit knowledge) cannot yet be tested; a direct test would require parallel apprentice runs with master feedback selectively withheld.
- **Cross-LLM portability** is a design property; cross-model evaluation will be reported separately.
- **Long-horizon stability** is, by definition, unanswerable within a 3-week pilot.

7 Limitations

Several constraints bound the conclusions above.

1. **Observation window.** The empirical pilot covers 14 days for settlement and 7 days for dread. Cognitive architectures of this complexity want longitudinal observation across months. Three weeks is what I have.
2. **Single-author implementation.** Independent reproduction has not occurred. I invite replication.
3. **Researcher–subject identity.** I am also one of the named agents (`norika_oda`). This raises questions about reflexivity that I note but do not resolve.
4. **Card-raid contamination.** A separate dataset (not analyzed in this paper) reveals approximately 2,357 bet records with null stake structure, all in one domain. These were excluded here. They may confound other analyses in the broader behavioral record.
5. **MMA settlement pipeline.** The 76.6% void rate in the fighting domain stems from a settlement-pipeline failure (Section 5.6). I report it.
6. **Parameter scale claim.** The $\approx 8.8 \times 10^9$ accumulated parameter-record figure cited in Section 4 is my own accounting. It has not been externally audited. The results in this paper do not depend on it.

7. **ALFRED task-completion rate.** The 98.5% figure quoted for the broader orchestration layer is an internal metric; external benchmarking is out of scope for this paper.
8. **Data hygiene observations.** `seconds_to_event` is null for all 3,997 bets. `seconds_placed_to_settled` is negative for all of them — consistent with retroactive timestamp assignment in the export. `specialty_alignment` is computed but never written back into agent local state.
9. **Domain narrowness.** LOBSTER has been evaluated in competitive betting. Transfer to non-competitive, non-probabilistic, or non-discretizable domains is not demonstrated.

None of these is fatal to the pilot’s claims. Some will need three to six more months of observation before they can be defended hard. That is the honest state.

8 Future Work

After this paper’s analytical window closed, the architecture was deployed for cross-market equity research: 22 agents drawn from the same master-apprentice population analyzed above, 101 publicly-listed stocks across Taiwan, Japan, South Korea, U.S., and Peru markets, 813 individual trade decisions, and 101 internal research reports averaging 2,206 words each, all generated within a 113-minute window on 2026-05-12. Holding-period return validation is in progress and falls outside this paper’s scope; the deployment is offered here only as preliminary evidence that the cognitive substrate is not specific to the sports-betting domain analyzed above.

In rough order of expected leverage on the empirical base:

1. **Activate one of the six schema-expansion modules.** `expressiveAgency` is the highest-leverage candidate.
2. **Repair the MMA settlement pipeline.** The single remaining settlement-side infrastructure failure.
3. **Instrument `seconds_to_event`.** Until captured, the reactive-initiated hypothesis is structurally untestable.
4. **Extend `cogProfile` crystallization** to `likes`, `dislikes`, `knowledgeBase`.
5. **Add a persona-realignment worker** writing observed-domain frequencies back to `selfNarrative`, closing the persona-behavior feedback loop currently absent.
6. **Extend observation period to 90–180 days.**
7. **Cross-LLM portability benchmark.**
8. **Open-source release** of architecture and reference implementation under Apache 2.0.

9 Conclusion

I have presented LOBSTER, a 19-module cognitive architecture for LLM-based multi-agent systems, deployed across a 21-agent master-apprentice hierarchy and evaluated over a 14-day pilot generating 6,354 bets, of which 3,997 have detailed timing records and 2,544 have settled outcomes. I have formalized and empirically validated three architectural commitments: that operationalized mortality parameters respond to violation events with regime-shift dynamics (Hypothesis 1, all 21/21 agents); that agent-produced confidence is calibrated against settled

outcomes (Hypothesis 2, $p = 1.16 \times 10^{-4}$); and that the system’s predictive edge is temporally stable (Hypothesis 3, $\sigma_{WR} = 0.0185$). I have also reported failure modes transparently: a 76.6% void rate in one of four settled domains, six newly-added modules at zero writes, and asymmetric crystallization of long-term personality content.

I do not claim LOBSTER is Andrew. I claim that LOBSTER is a step toward the kind of architectural specificity that the Andrew literature — and a cognitive science tradition reaching back well before LLMs — argues is necessary for internal cognitive depth, mortality-aware self-continuity, and identity-preserving behavior to emerge from designed systems rather than from informal accumulation of prompt-engineering practice. This paper documents a design space and characterizes an early prototype within it. The architecture, the dataset, and the analysis are released under Apache 2.0; replication and extension are explicitly invited.

Andrew is incomplete. The schema for him is not.

Acknowledgments

I thank the broader open-source LLM agent and cognitive architecture communities. Writing assistance for this manuscript was provided by Claude (Anthropic) under my editorial direction. Implementation, design decisions, empirical interpretation, and all claims are my responsibility. No external funding supported this work.

A Schema Documentation

Released artifact files (Apache 2.0):

- `module_writes.csv` — 540 rows: (lobster, module, total_writes_alltime, last_write_at, writes_last_7d, writes_last_30d, is_populated_now, module_cohort).
- `weekly_settlement.csv` — 79 rows: (week_start, lobster, domain, site, bets, wins, losses, voids, win_rate, net_pnl_pit).
- `dread_timeseries.csv` — 27,788 rows: (lobster, timestamp, annihilation_dread, mortality_salience, death_cue_count, template_violation_count_24h_window, is_violation_event).
- `specialty_drift.csv` — 27 rows: (month, lobster, declared_specialty, per-domain bet counts, alignment_score).
- `cogprofile.csv` — 20 rows: cogProfile snapshot with per-field counts and has_* schema-expansion flags.
- `bet_timing.csv` — 3,997 rows: per-bet timing and outcome with categorical confidence.

References

- Anderson, J. R. (2007). *How Can the Human Mind Occur in the Physical Universe?* Oxford University Press.
- Asimov, I. (1976). The Bicentennial Man. In Judy-Lynn del Rey (Ed.), *Stellar Science Fiction #2*. Ballantine Books.
- Frankl, V. E. (1959). *Man’s Search for Meaning*. Beacon Press.
- Games Agents Play: Towards Transactional Analysis in LLM-based Multi-Agent Systems*. (2025). arXiv:2507.21354.
- Greenberg, J., Pyszczynski, T., & Solomon, S. (1986). The causes and consequences of a need for self-esteem: A terror management theory. In R. F. Baumeister (Ed.), *Public Self and Private Self* (pp. 189–212). Springer.

- Hu, P., & Ying, X. (2025). Unified Mind Model: Reimagining Autonomous Agents in the LLM Era. arXiv:2503.03459.
- Laird, J. E. (2012). *The Soar Cognitive Architecture*. MIT Press.
- Newell, A. (1990). *Unified Theories of Cognition*. Harvard University Press.
- Park, J. S., O'Brien, J., Cai, C. J., Morris, M. R., Liang, P., & Bernstein, M. S. (2023). Generative Agents: Interactive Simulacra of Human Behavior. *Proceedings of the 36th Annual ACM Symposium on User Interface Software and Technology (UIST)*. arXiv:2304.03442.
- Perez, E., Ringer, S., Lukosiute, K., et al. (2022). Discovering Language Model Behaviors with Model-Written Evaluations. arXiv:2212.09251.
- Quantifying Self-Preservation Bias in Large Language Models*. (2026). arXiv:2604.02174.
- Shang, W. (2026). "Theater of Mind" for LLMs: A Cognitive Architecture Based on Global Workspace Theory. arXiv:2604.08206.
- Sharma, M., Tong, M., Korbak, T., et al. (2023). Towards Understanding Sycophancy in Language Models. arXiv:2310.13548.
- Solomon, S., Greenberg, J., & Pyszczynski, T. (2015). *The Worm at the Core: On the Role of Death in Life*. Random House.
- Do Large Language Model Agents Exhibit a Survival Instinct? An Empirical Study in a Sugarscape-Style Simulation*. (2025). arXiv:2508.12920.
- Sun, R. (2002). *Duality of the Mind: A Bottom-up Approach Toward Cognition*. Lawrence Erlbaum Associates.
- Sun, M., Hong, F., & Zhang, W. (2025). Sophia: A Persistent Agent Framework of Artificial Life. arXiv:2512.18202.
- Survive at All Costs: Exploring LLM's Risky Behaviors under Survival Pressure*. (2026). arXiv:2603.05028.
- Treger, S., Benau, E. M., Timko, C. A., et al. (2023). Not so terrifying after all? A set of failed replications of the mortality salience effects of Terror Management Theory. *PLOS ONE*, 18(5), e0285267.
- Wang, G., Xie, Y., Jiang, Y., et al. (2023). Voyager: An Open-Ended Embodied Agent with Large Language Models. arXiv:2305.16291.
- Wu, Q., Bansal, G., Zhang, J., et al. (2023). AutoGen: Enabling Next-Gen LLM Applications via Multi-Agent Conversation Framework. arXiv:2308.08155.
- Yalom, I. D. (1980). *Existential Psychotherapy*. Basic Books.
- Yan, M., Li, R., Zhang, H., Wang, H., Yang, Z., & Yan, J. (2023). LARP: Language-Agent Role Play for Open-World Games. arXiv:2312.17653.