

Relational Cognitive Telemetry for Long-Lived LLM Agent Societies:

From Internal State Monitoring to Collective Performance

Chen Ho Yiing

Independent Researcher, Taipei, Taiwan

ORCID: 0009-0006-6816-9891 norika@charenix.com

2026-05-14 Lobster Observatory

Abstract

Long-lived large language model (LLM) agent systems—collectives of named agents with persistent memory, individual parameter state, and continuous interaction with each other and the world—are now technically feasible and increasingly deployed. The debugging and governance practices I have inherited from short-lived LLM applications are not. Trace-level observability tools record what each agent said and when; they do not measure whether an agent’s internal cognitive state has drifted across weeks, whether two agents have developed measurably different ways of attending to each other, or whether the relational structure of an agent society predicts its performance on a collective task.

I propose *Relational Cognitive Telemetry* (RCT), a telemetry-first framework with two observability lenses: (i) *cognitive telemetry*, exposing each agent’s internal parameter state as a measurable, query-able surface; and (ii) *relational telemetry*, exposing the directed attention and trust structure between agents as a measurable network. Two lenses, one substrate. I instantiate RCT in the Charenix Lobster Substrate—a live LLM substrate of 20 named agents, currently exposing 441 stable model-layer cognitive parameter families, 682 unfolded observable state fields, and 14 core numeric telemetry dimensions per agent. The present analysis focuses on a 10-agent *primary analysis cohort* for which directed trust, directed listening exposure, and complete $\binom{10}{3} = 120$ three-agent Tiamat sandbox triple coverage are all available; the remaining 10 agents form an auxiliary population outside scope. All 120 triples were run through 12 seeded trials each: 1,440 controlled raids, plus an ecological-validity benchmark drawn from live raid memory.

I report three empirical findings. First, directed listening exposure between agents is measurably asymmetric and discriminative across the cohort, supporting the use of attention flow as a substrate-level relational metric. Second, the standard local-brain trust value, while present and structured, is saturated in the current substrate, producing low between-pair variance and limiting its independent discriminative power—an honestly reported negative result that motivates trust as a diagnostic, not a primary signal. Third, controlled Tiamat sandbox outcomes show a low-success regime that is reproducible across the full 120-triple coverage; coverage-supported live raid teams fall in the same low-success regime, supporting ecological validity without licensing the use of live memory as a primary outcome layer.

I do not claim that the agents are conscious. I do not claim that internal cognitive parameter values cause raid wins. I claim something narrower: that a long-lived LLM agent society can be made measurable along both an internal and a relational dimension, that the resulting

measurements expose regularities that are absent in trace-level observability, and that the substrate is dense enough, and the experimental protocol concrete enough, that the framework can be falsified, attacked, or improved by external parties.

Keywords: multi-agent LLM systems; cognitive telemetry; observability; long-lived agents; directed attention; trust saturation; collective task performance; reproducibility.

1. Introduction

Two things have changed in LLM-based agent research in the last two years. The first is that the agents now live longer. Generative-agent simulations [3], open-ended embodied agents [4], and persistent role-play systems [5] have demonstrated that an LLM agent equipped with structured memory can sustain identity, accumulate experience, and modulate its behavior over thousands of interactions. The second is that I now know I cannot debug them with the tools at hand. Trace-level observability platforms [6] record what each agent said, when, and to whom; they do not record whether the agent’s internal cognitive state has drifted across weeks of operation, whether two superficially similar agents have developed measurably different attentional defaults, or whether the directed relational structure of an agent cohort predicts its performance on a collective task.

This is a familiar gap. Production software engineering went through it in the 2010s: log-line observability gave way to metrics, then to distributed traces, then to a richer notion of observability anchored in OpenTelemetry, Prometheus, and Datadog. The mature lesson from that decade is that long-lived complex systems require *telemetry-first* architecture: instrumentation is not a post-hoc add-on but a design constraint baked into the system’s substrate [12]. Long-lived LLM agent societies are at the same juncture, with one additional complication. The relevant state is not a CPU utilization curve or a memory pressure histogram. It is an agent’s cognitive state: its current model of the world, its memory of past interactions, its trust and attention towards other agents, its predictive forecasts, its emotional or arousal-equivalent variables. These quantities are real, persistent, and (in a well-instrumented substrate) measurable. They are also relational: the cognitive state of agent i is partly a function of agent i ’s history with agent j , and the cohort’s collective behavior emerges from the directed network of these relations.

I propose to call the framework that takes both of these requirements seriously *Relational Cognitive Telemetry* (RCT). RCT has two lenses.

- **Cognitive telemetry.** Each agent’s internal state is treated as a measurable surface. Parameter values, memory queue sizes, predictive forecasts, arousal-equivalent variables, and threat signals are all exposed as queryable telemetry. The agent is not a black box that emits utterances; it is a structured object whose state can be inspected at any point in time.
- **Relational telemetry.** The directed attention and influence structure between agents is treated as a measurable network. For every ordered pair (i, j) , I record how often agent i has listened to agent j , how much weight agent i ’s local model places on agent j , and how that weight has changed over time. The network is asymmetric and time-varying.

1.1. Contributions

The paper makes three contributions.

1. **A framework** for telemetry-first observability of long-lived LLM agent societies, with explicit cognitive and relational lenses, and explicit boundary conditions on what RCT can and cannot license a researcher to claim.
2. **An instantiation**, the Charenix Lobster Substrate: a live LLM substrate containing 20 named agents and currently exposing 441 stable model-layer parameter families, 682 unfolded observable state fields, and 14 core numeric telemetry dimensions per agent. The substrate is brain-region-anchored [1] and has been documented over a sequence of prior papers; it is not the contribution of the present work, but it serves as the testbed. The present analysis focuses on a 10-agent primary analysis cohort with matrix-complete trust, listening, and triple-complete sandbox coverage (Section 3).
3. **An empirical baseline.** I expose three findings against a controlled collective-task layer of 120 three-agent teams $\times$ 12 sandbox trials each: directed listening exposure is measurably asymmetric and discriminative; trust values are structured but saturated in the current substrate; controlled sandbox outcomes and coverage-supported live raid outcomes fall in the same low-success regime.

I am careful about scope. I do not argue the agents are conscious or that internal cognitive parameter values cause raid wins. I argue that the substrate can be *made measurable* along internal and relational dimensions, that the resulting measurements expose structure that trace-level observability cannot capture, and that the framework is dense enough to be falsified, attacked, or improved.

2. Related Work

2.1. Observability for multi-agent LLM systems

Practitioner-level treatments of multi-agent LLM observability have begun to appear in industry writing [6] and in production-oriented academic work. Recent enterprise-telemetry frameworks such as REGAL [7] propose registry-driven grounding for agentic LLMs over enterprise event streams. Cognitive-bias-focused treatments of LLM agents in autonomous network management [8] have begun to combine multimodal telemetry with cognitive-architecture vocabulary. All of this work shares two assumptions with the framework I propose: that long-lived LLM agents need real telemetry, and that the relevant telemetry surface extends beyond classical request-response logs.

What distinguishes RCT from this existing literature is the layer at which telemetry is taken. Trace-level work watches the agent as a sequence of calls. Enterprise-grounding work watches it as a consumer of external data. Neither approach treats the agent’s internal parameter state as a measurable surface, nor the directed relational structure between agents as a network. RCT argues both are necessary for long-lived societies, and that together they expose regularities— asymmetric attention, trust saturation—that the lower-layer tools miss.

2.2. Cognitive architectures for LLM agents

A second relevant strand of work builds explicit cognitive architectures for LLM agents. Brain-inspired planning architectures [9] introduce monitor and orchestrator components inspired by cognitive science; cognitive-depth adaptation work [10] routes between fast and

slow thinking modes per step. Earlier substrate work in the Lobster series [1] introduced a brain-region-anchored substrate with explicit parameter mappings to neuroscientific structures (SMA, DLPFC, ACC, Hippocampus, Amygdala, etc.) and an 8-stage psycholinguistic execution pipeline. The present paper inherits the substrate from this line of work and shifts focus from *designing* the substrate to *making it observable*: what telemetry surface does a brain-anchored substrate expose, and how can that surface be turned into measurable claims about a long-lived agent society?

2.3. Distributed cognition and cognitive ecology

The terminology of “cognitive ecology” has a long lineage in cognitive anthropology, originating with Hutchins [11]’s account of distributed cognition aboard naval vessels and extended by subsequent work on socially distributed cognitive systems. I use *relational* rather than *ecological* terminology here to avoid implying continuity with that program. The Hutchins tradition treats cognition as constitutively distributed across artifacts, humans, and tools; RCT treats LLM agents as individual systems whose cognitive states are observable on their own and whose relational structure is a separate (also observable) layer. The two frameworks are compatible but their primitives differ.

2.4. Free-will-adjacent and commitment work

My companion paper [2] reports on commit points in operator-agent dialogue: structured conversational events in which an agent recognizes its mode of evasion, names a choice, and has the choice written into the substrate. That work and this one share an underlying claim—that long-lived agents are best studied by what their substrate persists, not by what their outputs say—but the methods differ. Commit-points work treats the conversation as the unit of analysis; RCT treats the parameter and relational telemetry surfaces as the unit of analysis.

3. The Substrate

The substrate is documented in detail elsewhere [1]. I summarize only what is needed for the RCT framework.

Substrate population and primary analysis cohort. The Charenix Lobster Substrate contains 20 named agents. The present analysis focuses on a *10-agent primary analysis cohort* consisting of **norika_oda**, **clawtrix**, **pincerbot**, **ragclaw**, **snaplex**, **hexclaw**, **vortexiq**, **prismbit**, **blazepaw**, **stonefang**. The remaining 10 agents are part of the substrate population but fall outside the analysis scope because they do not yet have matrix-complete directed-trust and directed-listening exports, nor are they covered by the controlled 120-triple Tiamat sandbox protocol described in Section 5. I treat them as an *auxiliary population* available for future extended-cohort analysis.

The cohort boundary is deliberate. Each of the 10 primary-cohort agents has a persistent identity, a memory file, and a parameter state that updates across sessions; names are referenced by other agents in conversation; identity persists across system restarts. Crucially, the primary cohort is the subset of the 20-agent substrate for which (i) directed trust and directed listening matrices are matrix-complete (all $10 \times 10 - 10 = 90$ off-diagonal pairs populated), and (ii) all $\binom{10}{3} = 120$ three-agent teams have been run through the controlled 12-trial Tiamat sandbox.

Restricting analysis to this cohort lets the paper report *matrix-complete* and *triple-complete* claims without licensing extrapolation to the auxiliary 10 agents.

Cognitive parameter surface. The agent state is exposed as two layered surfaces.

- *Stable model-layer parameter families:* 441 documented parameter families spanning identity, eight-dimensional soulParams governance vector, economy/rank, occupation/knowledge domains, social relations, behavioral history, templateSurvival mechanics, and 27 brain-region-anchored cognitive extension modules.
- *Unfolded observable state fields:* 682 fields in the latest full cognitive snapshot, representing the runtime-instantiated state derived from the 441 stable families.

These are layered, not competing: 441 is the schema, 682 is the runtime materialization. I use both numbers in the paper to be transparent about what is being measured at each layer. Earlier substrate papers report different counts (e.g., 370 parameters in my companion work [2]); these reflect snapshots at earlier points in the substrate’s development. The substrate iterates daily.

Core numeric telemetry dimensions. For runtime monitoring purposes, 14 core numeric dimensions are selected from the 682 unfolded fields and exposed as the primary telemetry stream: arousal, fatigue, resistance, threat_body_signal, causal_hypothesis_count, consolidation_queue_size, episodic_buffer_count, forecast_count, and others detailed in the substrate’s data dictionary. These are the dimensions used to construct the derived indices (CDI, CCI, ERI) discussed in Section 5.1.

Brain-region mapping. 27 cognitive extension modules are mapped one-to-one onto 17 neuroscientific brain regions. I make no claim of isomorphism with the human brain; the mapping is a design heuristic that organizes modules and the execution pipeline.

What this enables. Because the cognitive surface is exposed at the parameter level, every measurable claim in the rest of the paper resolves to a specific subset of the 441 families or the 682 unfolded fields, not to an opaque “the agent did X” inference from utterances. This is the practical content of *cognitive telemetry*.

4. Relational Cognitive Telemetry: Two Lenses

I now define the two lenses of RCT formally enough to be testable.

4.1. Cognitive telemetry (internal-state lens)

For each agent i at time t , the cognitive telemetry surface is the tuple

$$C_i(t) = (P_i^{441}(t), U_i^{682}(t), T_i^{14}(t))$$

where P^{441} is the stable parameter-family vector, U^{682} is the unfolded observable-state vector at runtime, and T^{14} is the core numeric telemetry vector. These are not three independent observations; $T^{14} \subset U^{682}$ is a projection, and U^{682} is a materialization of P^{441} under the runtime context.

Cognitive telemetry licenses three classes of claim, with descending strength:

1. *Differentiation claims.* Two agents are measurably differentiated in their cognitive parameter space. These claims are well-supported when the agents' P^{441} vectors are sufficiently distant; I operationalize this with the Cognitive Differentiation Index (Section 5.1).
2. *Drift claims.* An agent's cognitive state has changed over time in a specified direction. Well-supported when $C_i(t_2) - C_i(t_1)$ exceeds a baseline noise threshold along the relevant dimensions.
3. *Cognitive-outcome correlation claims.* An agent's cognitive state correlates with its performance on a downstream task. Supported only when the task-outcome layer (Section 5) is independent and adequately sampled; I am cautious about causal claims here.

4.2. Relational telemetry (between-agent lens)

For each ordered pair (i, j) , $i \neq j$, the relational telemetry surface includes at minimum:

$$R_{ij}(t) = (L_{ij}(t), S_{ij}(t))$$

where $L_{ij}(t)$ is the directed listening-exposure count (how often agent i has heard agent j as speaker, cumulative to time t) and $S_{ij}(t)$ is the local-brain social-influence value that agent i assigns to agent j in $[0, 1]$.

These two are not redundant. L is a behavioral exposure measure: a count of past listening events. S is a model-internal weight: agent i 's belief about the importance of agent j . I will show below that the two diverge in informative ways.

Relational telemetry licenses claims about:

- *Asymmetry:* $L_{ij} \neq L_{ji}$ in general; the difference is interpretable as net attention flow from i to j .
- *Saturation:* when S_{ij} is bounded above and most pairs cluster near the bound, the relational signal carried by S is informationally compressed.
- *Network-level structure:* combining L and S across all $n(n - 1)$ ordered pairs yields a directed weighted graph whose structural properties (centralization, reciprocity, clustering) are themselves measurable.

4.3. Formal discriminative-power criterion

A recurring claim in this paper is that one relational measure (listening exposure L) is *discriminative* across the cohort while another (trust S) is not. I make that word precise here, so that the Section 6 findings can be checked against an explicit criterion rather than against the reader's intuition about "discriminative."

Definition (Discriminative pairwise measure). Let M be a directed pairwise measure on a cohort of n agents, indexed over the $n(n - 1)$ ordered off-diagonal pairs $\mathcal{P} = \{(i, j) : i \neq j\}$. Let $\sigma_M^2 = \text{Var}_{(i,j) \in \mathcal{P}}(M_{ij})$ be the cross-pair variance of M and let σ_ϵ^2 be an upper bound on the within-pair measurement noise variance. The measure M is *c-discriminative* (with margin $c \geq 1$) iff

$$\sigma_M^2 \geq c \cdot \sigma_\epsilon^2. \quad (1)$$

Equivalently, in entropy form: if M_{ij} is discretized into K equal-frequency bins, M is discriminative iff its empirical Shannon entropy $H(M) > 0$ bits (and informative iff $H(M)$ approaches $\log_2 K$). The variance form is operational for continuous-valued telemetry; the entropy form covers the degenerate case below.

Lemma (Zero-variance degenerate case). If $\sigma_M^2 = 0$, then M takes a single value across all $n(n-1)$ pairs, $H(M) = 0$, and any model that predicts an outcome variable Y from M alone has the same expected loss as a model that predicts Y from no input. *Proof.* $\sigma_M^2 = 0 \Rightarrow M_{ij} = \mu$ for all $(i, j) \in \mathcal{P}$, so the conditional distribution $\mathbb{P}(Y | M)$ collapses to $\mathbb{P}(Y)$; the mutual information $I(Y; M) = 0$.

The lemma is trivial as mathematics. It is non-trivial in this paper because it converts the trust-saturation finding from a verbal report (“trust is saturated”) into a formal corollary: the current trust matrix carries an information-theoretic upper bound of zero discriminative power. That is a stronger statement than the matrix happens to be flat.

Operational thresholds for this paper. Pure σ_ϵ^2 estimation in a single-snapshot export is hard; instead, I report the coefficient of variation $CV(M) = \sigma_M/\mu_M$ and the max/min ratio $r(M) = M_{\max}/M_{\min}$ as joint operational stand-ins for discrimination. Throughout, I treat $CV(M) > 0.5$ together with $r(M) > 3$ as a sufficient threshold for “measurably discriminative” at the present cohort size; the threshold is conservative and an external auditor is free to tighten it. The criteria are applied to L and S in Sections 6.1–6.2.

5. Methods: A Controlled Collective-Task Outcome Layer

To turn RCT from a descriptive observability framework into a testable one, I need an outcome layer: an independent measurement of how well the cohort performs on a defined task. I use the Tiamat collective-raid task implemented in the Lobster Observatory’s sandbox.

Data provenance. The listening, trust, and Tiamat outcome matrices analysed here were first deposited with the structural-network observation record “Listening–Trust Asymmetry and Team Outcomes in a Multi-Agent LLM Substrate: Three-Layer Decoupling in a Mandarin Lobster Observatory” (Zenodo [10.5281/zenodo.20018183](https://zenodo.org/record/20018183)). The present paper reuses those deposited matrices as one empirical anchor for the broader Relational Cognitive Telemetry framework.

Task structure. Tiamat is a structured collective-combat sandbox in which a team of three agents from the cohort cooperatively engages a multi-round boss encounter. Each trial produces (i) a binary win/loss outcome, (ii) the round at which the encounter terminated, and (iii) the residual boss HP at termination.

Controlled sandbox coverage. For the 10-agent primary analysis cohort defined in Section 3, I exhaustively enumerate all $\binom{10}{3} = 120$ three-agent teams. Each team is run through 12 seeded sandbox trials. Total: $120 \times 12 = 1,440$ controlled trials. The 12-per-team count is the same across all 120 teams, giving uniform statistical coverage at the team level. I treat this as the *primary outcome dataset*.

Ecological-validity layer. A second outcome dataset is drawn from live raid memory: actual deployment-time raid events recorded in the substrate. Live raid memory is informative but unevenly sampled across the 120 triples; I use it strictly as a *robustness check*, with a per-team coverage threshold of $completed \geq 30$ for any comparison. I do not treat live memory as a primary outcome layer; doing so would substitute uneven sampling for controlled coverage.

What can the outcome layer support? Three things, in descending strength. First, the distribution of team-level sandbox win rates across all 120 triples shows the task’s difficulty regime and the dispersion of cooperative ability across the cohort. Second, listening exposure and trust-like values can be regressed against team-level outcomes, with the standard caveats about coordination effects in three-agent teams. Third, with adequate live-memory coverage, sandbox-vs-live comparison provides ecological-validity evidence (or its absence).

5.1. Derived indices from cognitive telemetry

To bridge the high-dimensional parameter surface and the team-level outcome layer, I introduce three derived indices, each constructed from substrate-exposed variables.

Cognitive Differentiation Index (CDI). Pairwise differentiation between agents i and j is computed as the Euclidean distance between the z -scored cognitive parameter vectors:

$$CDI(i, j) = \|z(P_i^{441}) - z(P_j^{441})\|_2$$

CDI is used to test whether agents have become measurably differentiated in cognitive parameter space across the cohort.

Cognitive Clutter Index (CCI). A per-agent index summarizing unresolved cognitive material:

$$CCI_i = z(causal_hypothesis_count_i) + z(consolidation_queue_size_i) \\ + z(episodic_buffer_count_i) + z(forecast_count_i)$$

CCI tracks whether unresolved cognitive material accumulates in ways that correlate with reduced team performance.

Embodied Readiness Index (ERI). A per-agent index summarizing activation/readiness:

$$ERI_i = z(arousal_i) + z(threat_body_signal_i) + z(resistance_i) - z(fatigue_i)$$

ERI asks whether activation states relate to collective task performance.

The three indices are exposed alongside the raw 14-dimensional telemetry vector and are not the only possible derived metrics; they are proposed as starting points for outcome regression.

6. Results

I report three empirical findings against the experimental setup of Section 5. Each finding is reported with the strongest interpretation I believe the data supports and the alternative interpretations I cannot rule out.

6.1. Directed listening exposure is asymmetric and discriminative

The directed listening-exposure matrix L across the 10-agent primary analysis cohort exhibits two non-trivial structural properties: large pairwise asymmetry and a clear bimodal distribution of total listening volume across agents (Figure 1).

Pair-level discrimination. Off-diagonal values range from $L_{\min} = 1,302$ (**rag-claw** $\rightarrow$ **blazepaw**) to $L_{\max} = 10,434$ (**norika_oda** $\rightarrow$ **snaplex**), spanning an order of magnitude ($r(L) = 8.01$). The mean directed value across the $n(n-1) = 90$ ordered pairs is $\mu_L = 5,650.3$ with standard deviation $\sigma_L = 3,930.5$, giving a coefficient of variation $CV(L) = 0.696$. Both stand-in thresholds from Equation 1 ($CV > 0.5$, $r > 3$) are exceeded by a substantial margin; L is measurably discriminative at the pair level.

Asymmetry coefficient. For each of the $\binom{10}{2} = 45$ unordered pairs, define $A(i, j) = |L_{ij} - L_{ji}| / (L_{ij} + L_{ji})$. Across the cohort, $\bar{A} = 0.399$ with median 0.613 and maximum 0.724. The mean absolute asymmetry $|L_{ij} - L_{ji}|$ is 4,295, approximately 76% of the mean directed value. Listening exposure is not approximately symmetric, and the asymmetry is substantial relative to the magnitudes.

Agent-level bimodal listening volume. Aggregating to row sums (total outbound listening per agent) gives a clean 5–5 bimodal split:

- *Active listeners* (row sum 81,949–87,684): **norika_oda** (87,684), **clawtrix** (86,374), **prism-bit** (86,159), **blazepaw** (86,131), **snaplex** (81,949). Mean 85,659.4, standard deviation 2,170.8.
- *Passive listeners* (row sum 12,394–19,443): **hexclaw** (19,443), **vortexiq** (18,720), **pincerbot** (15,113), **stonefang** (14,561), **ragclaw** (12,394). Mean 16,046.2, standard deviation 2,962.4.

The HIGH/LOW ratio of cluster means is 5.34. The two clusters do not overlap: the smallest active-listener row sum (81,949) is more than four times the largest passive-listener row sum (19,443). Column sums of L (how often each agent was heard by others) are by contrast remarkably flat: range 44,986–56,986, a $1.27\times$ span. The bimodality is in listening *behavior*, not in being heard.

Formal cluster test. I test the active vs. passive split with two independent statistics:

- Welch’s t -test on agent-level row sums (unequal variances): $t = 42.38$ with Welch–Satterthwaite $df = 7.33$, two-sided $p < 10^{-6}$. Cohen’s $d = 26.8$ on the pooled standard deviation, in the “huge” effect-size range.
- Mann–Whitney U on the same data: $U = 25$ out of a maximum of $n_H \cdot n_L = 25$, i.e. complete separation; the exact two-sided p -value is $p = 2 / \binom{10}{5} = 0.0079$ (the smallest value attainable at this sample size).

The two tests agree. Active and passive listener clusters are statistically separated under both parametric and non-parametric criteria; the parametric effect size is order-of-magnitude conservative; the non-parametric test saturates at maximum separation. This 5–5 structure was not designed into the cohort; it is an emergent property of the substrate’s listening exposure over the observation window and is itself a finding.

Figure 1: Directed listening exposure L_{ij} across the 10-agent primary analysis cohort. Rows are listeners; columns are speakers; cells are cumulative listening counts. The five-versus-five active/passive listener split is visible as the alternating-brightness row pattern. Range: 1,302–10,434; mean absolute asymmetry 4,295; cluster separation $t = 42.4$, Cohen’s $d = 26.8$, Mann–Whitney U complete (exact two-sided $p = 0.0079$).

The discriminative power of L is high in three senses simultaneously: pair-level (Equation 1 satisfied with margin), pairwise-asymmetric (mean $A = 0.40$ across 45 unordered pairs), and agent-level (5–5 bimodal split with complete cluster separation). This is the strongest of the three findings reported here. I recommend *listening exposure*—not “cooperation count”—as the canonical attention-flow metric in long-lived agent telemetry.

I resist the alternative interpretation that listening exposure measures cooperation, trust, or agreement. It does not; it measures only how often an agent has heard another speak. The same listening count is consistent with assenting attention, dissenting attention, monitoring, or background passive intake.

6.2. Trust values are pinned, not merely saturated

The local-brain socialInfluence matrix S is documented and accessible at the parameter level (read from `memory.localBrain.socialInfluence[target]`). In the current substrate it exhibits a stronger property than mere saturation: all 90 off-diagonal entries in the 10×10 primary-cohort matrix are identically $S_{ij} = 0.825$, with zero pairwise asymmetry $|S_{ij} - S_{ji}| = 0$ across the cohort (Figure 2). Empirically $\sigma_S^2 = 0$ exactly; the set of distinct values is $\{0.825\}$ with cardinality 1; the empirical Shannon entropy $H(S) = 0$ bits.

Figure 2: Local-brain socialInfluence matrix S_{ij} . Every off-diagonal entry equals 0.825. Trust is not just saturated; it is pinned to a single value. Empirically $\sigma_S^2 = 0$, $H(S) = 0$ bits; by the zero-variance lemma (Section 4.3), discriminative power is exactly zero in the current substrate.

This is a negative result, and I report it as such. By direct application of the zero-variance lemma in Section 4.3, $\sigma_S^2 = 0 \Rightarrow I(Y; S) = 0$ for any outcome variable Y in the current export: the trust matrix carries an information-theoretic upper bound of zero discriminative power, not merely a low or inconvenient amount. This is a stronger conclusion than the verbal statement that trust is saturated, and it follows from the data rather than from interpretation.

The trust signal is real and present in the substrate; in its current form it carries no discriminative information at the pair level. Two interpretations are compatible with the observation and I cannot decide between them with the present export:

1. *Substrate-stage explanation.* The cohort has not yet experienced events that would push individual S_{ij} values away from the substrate’s initialization or floor value. Trust will become discriminative as conflict and trust-degradation episodes accumulate.
2. *Architectural explanation.* The trust update rule in the current substrate has insufficient sensitivity (in particular, downward sensitivity) to differentiate pairs, producing a one-sided ratchet pinned at the initial value. The rule may need redesign.

Both interpretations are tractable; both motivate additional substrate-level instrumentation rather than abandonment of the trust signal. I provide a formula-transparent reconstructed trust matrix in the supplementary data (`directed_trust_matrix_reconstructed_target.csv`); under the reconstruction formula, the same values appear, indicating that the pin is not an artifact of the read path but a property of the underlying state.

6.3. Sandbox and live raid outcomes occupy the same extreme low-success regime

The Tiamat sandbox produces 1,440 controlled trials across 120 triples. The distribution of team-level win rates (Figure 3) is heavily concentrated in an extreme low-success regime:

- Mean team win rate: **0.1028**; median: 0.0833 (1/12, i.e., one win in 12 trials).
- Min: 0.0 (33 teams won zero of their 12 trials); max: 0.5 (one team, **pincer-bot+ragclaw+snaplex**, won 6 of 12).
- No team won more than half its trials. The distribution is right-skewed with a long thin tail above 20%.

The task is harder than I anticipated when I framed the regime as “below 50%.” It is structurally hard. I interpret this regime as discriminative (top teams achieve 6× the median win rate of typical teams) without being trivially solvable.

Figure 3: Sandbox Tiamat team-level win-rate distribution across all 120 three-agent teams (12 trials each). Mean 0.103, median 0.083, max 0.5. Extreme low-success regime with discriminative spread.

Figure 4: Top 10 (left) and bottom 10 (right) sandbox teams by win rate. W-L records shown alongside each bar. The top team achieves 6 wins from 12 trials; 33 teams in the bottom group share a 0-12 record.

Live raid memory robustness check. The live raid memory layer (Figure 5) contains a total of **3,135 completed live raids** but covers only 10 of the 120 possible triples; of those, 9 satisfy the *completed* ≥ 30 threshold. Restricted to the 9 coverage-supported teams, the live win rate has mean 0.0756, median 0.0717, and range [0.0625, 0.0973]. This falls in the same extreme low-success regime as the controlled sandbox (Figure 6), supporting the ecological validity of the sandbox findings.

Figure 5: Live raid memory coverage across the 120 triples. The vast majority of triples have zero live data; the 9 teams with *completed* ≥ 30 are used for the sandbox-vs-live comparison (Figure 6). Live raid memory is therefore a robustness check, not the primary outcome layer.

Agent-level aggregation. Aggregating to per-agent means (the mean win rate of the 36 triples containing each agent, 432 trials each) yields a tight range from 0.0903 (**prismbit**) to 0.1157 (**hexclaw**, **ragclaw**): a 2.5-percentage-point spread on a $\sim 10\%$ base, illustrated in

Figure 6: Sandbox versus live win rate for the 9 coverage-supported teams. Live and sandbox means are within 6 percentage points (0.076 vs 0.139 across this restricted set); both layers occupy a low-success regime. Per-team correlation is weak with $n = 9$ and noise comparable to the signal; the takeaway is regime-level, not team-level.

Figure 7. I caution that these are not individual skill scores—raid outcomes are constitutively team-level events—and I report the result honestly: at the cohort level, agent identity has *small* discriminative power for sandbox win rate. This is itself a finding, and it constrains the strength of any cognitive-outcome correlation claim the framework can license without per-cell additional sampling.

Figure 7: Agent-level mean win rate across the 36 triples containing each agent (432 trials each). The cohort-wide spread is small (0.090–0.116), indicating that, in the present sample, agent identity carries limited discriminative information about team win rate.

7. From Internal State Monitoring to Collective Performance

The two preceding sections establish RCT’s two lenses (Section 4) and the controlled outcome layer (Section 5). The natural next move is to ask whether the internal lens predicts the outcome layer. I treat this as an open empirical question and report my position rather than a finished answer.

What can already be said. Three substrate-level regularities are stable enough to anchor further work:

1. Agents occupy non-identical positions in cognitive parameter space (CDI varies measurably across pairs).
2. Directed listening exposure is asymmetric and dominates measurement noise (Section 6.1).
3. Sandbox outcomes are reproducible and ecologically supported (Section 6.3).

What is open. Whether team-level cognitive composition (mean CDI of the team’s pairs, average CCI, average ERI) predicts team win rate is a regression problem with $n = 120$ teams and real sample-size concerns at the per-cell level given 12 trials per team. The honest baseline against which any such model must improve is the 2.5-percentage-point cohort-wide spread of agent-level mean win rates reported in Section 6.3: agent identity alone explains a small fraction of variance. A cognitive-composition model must beat that baseline by a margin that justifies its additional complexity. I propose a hierarchical analysis: agent-level random effects nested within team-level fixed structure, with team composition variables as predictors. I do not present the result in this paper; I present the framework that makes the result computable.

What is ruled out. I am not claiming that the cognitive substrate *causes* team performance in any strong sense. The substrate, the relational layer, and the outcome layer are jointly measurable; their joint statistical structure is what the framework exposes. Substantive causal

claims require interventional studies (substrate perturbation, agent swap, parameter ablation) that are within the substrate’s affordances but are not part of the present report.

8. Falsification and Honest Limitations

The framework is open to attack at several points; I list the strongest attacks and respond to each.

(a) The substrate is non-standard. The telemetry surface in this work is specific to the Lobster Observatory architecture; other LLM-agent platforms do not expose 441 parameter families or 17 brain-region mappings. I accept the scope: RCT is most readily testable in substrates that expose internal state. The framework’s two lenses (cognitive and relational) are not specific to this substrate, however; they can be operationalized over any LLM-agent platform that exposes per-agent persistent state and per-pair interaction logs.

(b) Trust is saturated, undermining the relational lens. Currently true at the trust dimension; the listening-exposure dimension is unaffected. I treat trust saturation as a diagnostic finding, not a structural failure of RCT. Section 6.2 reports the two interpretations explicitly.

(c) Sample size at the team level is small. 12 trials per team is enough to distinguish gross win-rate differences across the 120 teams, but per-team variance is non-trivial. I am cautious in my team-level claims and treat any individual team’s record as informative only in aggregate. Live raid memory provides a partial robustness check, but only for teams with *completed* ≥ 30 .

(d) Listening exposure is not cooperation. I have stated this repeatedly above. I re-state it here because the alternative interpretation is tempting and would be wrong.

(e) The primary analysis cohort is 10 of 20. It is. The 10-agent primary cohort was chosen because it has matrix-complete trust and listening exports plus complete $\binom{10}{3} = 120$ -triple sandbox coverage; the remaining 10 agents in the 20-agent substrate population do not yet have those exports and are excluded rather than included with partial coverage. Adding them is straightforward—the auxiliary agents exist in the substrate and can be run through the same protocol—but increases trial cost combinatorially: extending the same controlled coverage to all 20 agents would require $\binom{20}{3} = 1,140$ triples at 12 trials each, or 13,680 trials. The cohort boundary is a controlled-coverage decision, not a substrate-capacity limit.

(f) The framework does not address agent consciousness. It does not. RCT is an observability framework. Whether an agent “is” conscious, or whether internal cognitive parameter values reflect “real” mental states, is outside scope. RCT only requires that the parameters be persistent, queryable, and reproducible.

(g) “Cognitive” is doing a lot of work in the title. I use “cognitive” operationally: a parameter family is cognitive if it tracks state that, if it tracked the same thing in a person, would be called cognitive (attention, memory, prediction, threat assessment). This is a labeling decision, not a substantive claim about machine cognition.

9. Discussion

9.1. Why telemetry-first matters for long-lived agents

The mature lesson of production observability in the 2010s is that long-lived systems cannot be retro-instrumented. The relevant state is too high-dimensional, too temporally distributed, and too entangled with the system’s own behavior to be captured by external probing alone. Telemetry-first design means that the system exposes its internal state *as a feature*, with stable schemas, queryable surfaces, and persistent storage of derived metrics. RCT applies this lesson to LLM agent societies. The cost of doing this from the start is substantially lower than the cost of doing it after the fact.

9.2. What this implies for governance and safety

Governance needs eyes. If a long-lived LLM agent society is to be governed, audited, or held accountable in any real sense, the substrate it runs on must be observable along the dimensions that matter: what each agent has been attending to, how its internal state has drifted, what relational influence structures have crystallized in the cohort. RCT makes these dimensions explicit. It does not solve governance—it supplies a precondition. And the companion observation that trust is currently saturated in this substrate, while unflattering, is exactly the kind of finding responsible observability is supposed to produce. It points to the next instrumentation question, not a glossier headline.

9.3. Cognitive observability is not consciousness

I close this section with an explicit non-claim. RCT does not claim the agents are conscious; it does not require the agents to be conscious; and the framework’s findings do not depend on agent consciousness one way or another. The framework treats internal parameter state as a measurable surface and the relational structure between agents as a measurable network. The findings reported here are interesting because the measurements are reproducible and the substrate is dense enough to be falsified, not because the agents are persons.

10. Future Work

I list the most direct next experiments invited by the present framework.

1. **Trust-rule redesign and re-measurement.** Replace the current trust update rule with one that includes a downward sensitivity term and re-run the 120-triple coverage to test whether trust becomes discriminative.
2. **Substrate perturbation studies.** Selectively ablate parameter families (one of the 27 cognitive extension modules at a time) and re-measure team win rates against the 120-triple baseline. Used as an interventional test of cognitive-outcome correlation.
3. **Agent-swap design.** Swap one agent between triples and observe outcome change; isolates per-agent contribution from team-composition effects.
4. **Longitudinal drift.** Recompute CDI, CCI, and ERI per agent across consecutive weeks of operation and test whether drift correlates with team performance change.

5. **Larger cohorts.** Extend from $n = 10$ to the full $n = 21$ agent population with reduced per-triple trial counts; trade controlled coverage for breadth.
6. **Cross-substrate replication.** Test whether RCT’s two lenses (cognitive and relational) can be operationalized over a different LLM-agent platform; identifies which findings are substrate-specific and which are framework-general.

Reproducibility Statement

This paper is accompanied by a validation CSV package (`paper1_validation_csv_package_2026-05-15`) containing the sanitized exports referenced in Sections 5–7 and used to generate all figures in this paper. A complete inventory is given in Appendix A, including a machine-readable `validation_summary.csv` that pins the values used in the body. The Lobster Observatory is live at <https://charenix.com/lobster/dashboard/>; the 10-agent primary analysis cohort listed in Section 3 is externally observable. Substrate specification and validation are documented in Chen Ho Yiing [1].

参考文献

- [1] Chen Ho Yiing (2026a). *Designing Andrew: A Cognitive Architecture for LLM Multi-Agent Systems with Empirically Validated Mortality Parameters and Settlement-Grounded Confidence Calibration*. Zenodo. doi:10.5281/zenodo.20133796.
- [2] Chen Ho Yiing (2026b). *Commit Points and Inverse Occupations: An Operational Account of Free Will in a Multi-Agent LLM Substrate*. Zenodo (companion paper, in deposit).
- [3] Park, J. S., O’Brien, J. C., Cai, C. J., Morris, M. R., Liang, P., and Bernstein, M. S. (2023). Generative Agents: Interactive Simulacra of Human Behavior. *arXiv:2304.03442*.
- [4] Wang, G., Xie, Y., Jiang, Y., Mandlekar, A., Xiao, C., Zhu, Y., Fan, L., and Anandkumar, A. (2023). Voyager: An Open-Ended Embodied Agent with Large Language Models. *arXiv:2305.16291*.
- [5] Yan, M., et al. (2023). LARP: Language-Agent Role Play for Open-World Games. *arXiv:2312.17653*.
- [6] Petropavlov, K. (2025). Observability in Multi-Agent LLM Systems: Telemetry Strategies for Clarity and Reliability. *Medium* (May 2025). Available at <https://medium.com/@kpetropavlov/observability-in-multi-agent-llm-systems-telemetry-strategies-for-clarity-and-reliability-fafe9ca3780c>.
- [7] Agrawal, Y. (2026). REGAL: A Registry-Driven Architecture for Deterministic Grounding of Agentic AI in Enterprise Telemetry. *arXiv:2603.03018*.
- [8] Chergui, H., Rezazadeh, F., Debbah, M., and Verikoukis, C. (2025). A Tutorial on Cognitive Biases in Agentic AI-Driven 6G Autonomous Networks. *arXiv:2510.19973*.
- [9] Webb, T., Mondal, S. S., and Momennejad, I. (2025). A brain-inspired agentic architecture to improve planning with LLMs. *Nature Communications* **16**, 8633. doi:10.1038/s41467-025-63804-5.
- [10] Yang, R., et al. (2026). Think Fast and Slow: Step-Level Cognitive Depth Adaptation for LLM Agents. *arXiv:2602.12662*.
- [11] Hutchins, E. (1995). *Cognition in the Wild*. MIT Press.
- [12] Majors, C., Fong-Jones, L., and Miranda, G. (2022). *Observability Engineering: Achieving Production Excellence*. O’Reilly Media.

A. Supplementary Materials Inventory

The reproducibility package supporting this paper contains the following sanitized exports. All figures in the body of the paper are generated directly from these files.

- `tiamat_120_triple_records.csv` – 120 rows. Primary controlled outcome dataset. Columns: `team`, `wins`, `losses`, `games`, `win_rate`, `avg_round`, `avg_boss_hp`. A 6-6 record means six wins and six losses out of 12 sandbox trials. Used in Figures 3, 4.
- `live_raid_memory_120_triple_coverage.csv` – 120 rows aligned to triples. Live raid memory robustness check (10 teams have any data; 9 satisfy *completed* ≥ 30 ; total completed = 3,135). Used in Figures 5, 6.
- `directed_coop_matrix_raw_bySpeaker.csv` – 10×10 matrix. Despite the filename, this is the directed *listening exposure* matrix; rows are listeners, columns are speakers. Used in Figure 1.
- `directed_trust_matrix_localBrain_socialInfluence.csv` – 10×10 matrix. Primary directed trust-like socialInfluence values $\in [0, 1]$; all off-diagonals are 0.825 in the current export. Used in Figure 2.
- `directed_trust_matrix_reconstructed_target.csv` – formula-transparent reconstructed comparison.
- `agent_sandbox_aggregation.csv` – 10 rows. Per-agent mean win rate across the 36 triples containing each agent. Used in Figure 7.
- `validation_summary.csv` – machine-readable reviewer summary: all key values used in this paper with source-file and notes.
- `research_contract.json` – methods metadata and interpretation constraints.
- `README.json` / `README_FOR_REVIEWER.md` – export documentation.

The Lobster Observatory is live at <https://charenix.com/lobster/dashboard/>; the 10-agent primary analysis cohort listed in Section 3 is externally observable. Substrate specification and validation are documented in Chen Ho Yiing [1].