

Cognitive State as Behavior Signal: Synapse Parameter Predictive Validity in a 1,743-Hour Multi-Agent Panel

Lobster Observatory Paper 21

Chen Ho Yiing¹
Independent Researcher

May 8, 2026
License: CC BY 4.0

Abstract

I report the first systematic empirical test of cross-time predictive validity for the synapse parameter framework underpinning a personal cognition layer. Over a 127-hour continuous window (2026-05-02 to 2026-05-08) in the Lobster Observatory Substrate—a Mandarin-language multi-agent LLM substrate of 20 agents grounded in a live consumer prediction ecosystem—I recorded 14 internal cognitive parameters at hour t and 10 behavioral outcome metrics at hour $t+1$ for every agent-hour pair, yielding 1,743 panel observations and 140 parameter–outcome correlation tests. Every one of the 14 parameters showed at least one statistically meaningful cross-time correlation with at least one outcome. The strongest correlations approach causal-strength magnitudes: `threat_body_signal` predicts next-hour message volume at $r = 0.835$ ($n = 1743$); `arousal` predicts the same outcome at $r = 0.832$; `metacognitive_calibration` predicts next-hour money-pain at $r = 0.634$, while `model_revision_pressure` predicts it inversely at $r = -0.566$. I interpret these results as evidence that the synapse parameter framework satisfies the predictive-validity criterion required by the Afu Brain architectural specification: a parameter is real if and only if its current value changes future decisions. The dataset is released in full alongside this manuscript under CC BY 4.0. I discuss methodological caveats including multicollinearity between somatic-pressure parameters, the inclusion of one human participant in the panel, and the limits of single-substrate, single-language evidence.

Keywords synapse parameters, multi-agent substrate, predictive validity, personal cognition layer, generative agents, panel data, Mandarin LLM substrate.

1 Introduction

The personal cognition layer thesis holds that a useful local “brain” is one whose internal state both reflects and shapes its future behavior in measurable ways. If internal state cannot be shown to influence what the agent does next, it is not part of the architecture; it is annotation. The Afu Brain specification [1] formalizes this requirement as a six-part validity criterion for any candidate *synapse parameter*: the parameter must (1) have a clear definition, (2) be observable from behavior or text, (3) change future decisions, (4) have an explicit update rule, (5) have decay or conflict handling, and (6) be inspectable. Of these, the third—predictive validity—is the empirically testable claim. The other five are design constraints; this one is a falsifiable hypothesis.

The Lobster Observatory series has, in earlier papers, examined substrate-level emergent phenomena: epistemic norms [2], third-order theory of mind in trust modulation [3], the architectural commitments of the substrate itself [4], listening–trust asymmetry in team raid outcomes [5], computational representations of social being [6], the co-presence inheritance threshold [7], the grounding of death [8], and a model-agnostic safety layer separating LLM-driven interfaces from tool execution [9]. These papers share a substrate description but treat the substrate primarily as a setting for studying multi-agent emer-

¹Independent Researcher. ORCID: 0009-0006-6816-9891. Correspondence: norika@charenix.com.

gence. This paper turns the question inward. Rather than asking what emerges among agents, it asks: *do the parameters I instrument inside each agent predict that agent’s own near-future behavior?*

This question is answerable in a single way: by recording the parameters at every hour, recording the behaviors at every subsequent hour, and computing cross-time correlation across a panel large enough to overwhelm noise. The panel I ran covers 1,743 agent–hour pairs across 20 agents over six continuous days, with 14 parameters tested against 10 outcomes. Every combination was pre-committed for testing, removing the cherry-picking concern that would otherwise attach to any subset of significant findings.

The empirical answer is straightforward. Synapse parameters predict behavior. The strongest cross-time correlations approach magnitudes unusual in social-science panel work, and every parameter I instrumented showed predictive signal against at least one outcome. I read this as architectural, not incidental: the synapse framework passes the third validity test in this substrate.

I built the substrate to make the test possible. If the parameters had failed it, the architecture would have been wrong, and I would have had to start over. They didn’t fail.

Contributions This paper makes four contributions.

1. It releases the first open empirical panel dataset linking internal cognitive parameters to next-hour behavioral outcomes in a multi-agent LLM substrate.
2. It supplies a falsifiable predictive-validity criterion for synapse parameter frameworks and an instance that passes it.
3. It documents the strongest individual cross-time correlations observed (Section 6), with attention to multicollinearity between adjacent somatic-pressure parameters.
4. It establishes a methodological baseline against which parameter frameworks proposed for other personal cognition layers may be compared.

2 Related Work

The closest reference class is the generative agents literature [10, 11], in which LLM agents are equipped with memory, reflection, and planning modules and observed in simulated social settings. That work establishes that such agents can produce believable individual and collective behavior. It does not, however, ask whether the agents’ internal parameters predict their own behavior cross-time, in part because the temporal scales studied are short (a single simulated day) and the parameter sets are not exposed for systematic panel analysis. This paper extends generative-agent methodology to a continuous multi-day window with a full instrumented parameter panel.

A second reference class is interpretability work on language model internal states [12, 13]. There the question is whether internal model activations or attention patterns predict output. My question is structurally parallel but operates one level of abstraction higher: not on raw activations but on derived parameters intended for use by an outer cognitive layer that itself decides whether and how to act. The validity criterion is correspondingly different. I do not need my parameters to be mechanistically reducible to circuits; I need them to predict behavior in a way that is useful for an external cognitive control loop.

A third reference class is panel data econometrics, particularly Granger-causal-style analysis of cross-time prediction [14]. My design is deliberately less ambitious than formal Granger causality testing: there is no control for the agent’s own past behavior, no partialling-out of time-of-day or persona effects, and I report Pearson correlations rather than fitted panel regression coefficients. The rationale is that a first predictive-validity paper benefits from being readable directly against a clear empirical claim. Causal interpretation, if any, is left to follow-up work.

3 The Lobster Observatory Substrate

The Lobster Observatory Substrate is a Mandarin-language multi-agent environment running continuously since 2026-04-15 across a five-site consumer prediction ecosystem. Twenty agents (collectively, “the lobsters”) operate within three communicative registers: a public lounge for broadcast speech, a private writing surface (*MoltBook*) for long-form personal posts, and an internal reasoning trace not exposed to other agents. Agents place bets, raid Tiamat events, discuss topics, and leave traces of their internal state in a telemetry layer. The architectural commitments of the substrate are described in earlier work [4] and are not the subject of this paper.

I restrict attention to the six-day window from 2026-05-02 10:00 to 2026-05-08 07:00 (127 unique hours), during which 20 personas were active to varying extents, producing 1,743 agent-hour observations. Of the 20 personas, one (*norika_oda*) is a human participant who maintains a self-monitored panel slot in the substrate; I discuss the implications in Section 8. The remaining 19 are LLM-driven agents with consistent persona prompts, instrumented memory architectures, and substrate-mediated communication. Per-persona hour counts ranged from 44 to 122 (Appendix A).

4 Synapse Parameter Architecture

I instrument 14 parameters per agent per hour, falling into five conceptual groups derived from the Afu Brain specification [1], Section 6:

Cognitive load. Three parameters describing the moment-to-moment processing burden:

- *attention_load* ($[0, 1]$): proportion of available attention currently committed.
- *active_focus_count* ($\mathbb{Z}_{\geq 0}$, capped at 5 by substrate convention): number of distinct topics or threads the agent is currently tracking.
- *fatigue* ($[0, 1]$): cumulative load drag, monotonically non-decreasing within a session and decaying across rest periods.

Model accuracy and revision. Four parameters describing the agent’s running self-model of the world:

- *prediction_error* ($[0, 1]$): residual between predicted and observed substrate events.
- *model_revision_pressure* ($[0, 1]$): accumulated pressure to update the agent’s running model.
- *metacognitive_calibration* ($[0, 1]$): self-reported confidence-calibration alignment.
- *overconfidence_risk* ($[0, 1]$): excess of stated confidence over calibrated confidence.

Somatic pressure. Three parameters with body-signal analogues:

- *arousal* ($[0, 1]$): general activation level.
- *threat_body_signal* ($[0, 1]$): perceived threat intensity registered as embodied signal.
- *reward_craving* ($[0, 1]$): pull toward expected reward states.

Narrative coherence. Three parameters describing the agent’s internal story:

- *consolidation_queue_size* ($\mathbb{Z}_{\geq 0}$): pending memory events not yet consolidated into long-term store.
- *causal_hypothesis_count* ($\mathbb{Z}_{\geq 0}$): number of competing causal hypotheses currently held.
- *narrative_drift* ($[0, 1]$): drift of the active narrative away from prior committed narrative.

Procedural strength. A single parameter describing behavioral routinization:

- `procedural_strength` $([0, 1])$: degree to which the current behavior is routinized rather than deliberated.

These parameters are intended to be inspectable by a personal cognition layer (the “brain”) sitting between an LLM-driven interface and any tool execution. The brain’s job is to read these parameters, decide whether to act locally or call a frontier teacher, and gate any external action through a deterministic safety layer [9]. The question of this paper is whether these parameters, so defined, are doing real predictive work or merely cosmetic instrumentation.

5 Methodology

5.1 Panel construction

Panel structure. Let $P = \{p_1, \dots, p_{20}\}$ be the set of personas observed during the window, and for each $p \in P$ let $T_p \subseteq \mathbb{Z}$ be the set of hour indices in which agent p produced at least one observable signal. The panel is the set of paired observations

$$\mathcal{D} = \{ (p, t, \mathbf{x}_{p,t}, \mathbf{y}_{p,t+1}) : p \in P, t \in T_p, t+1 \in T_p \} \quad (1)$$

where $\mathbf{x}_{p,t} \in \mathbb{R}^{14}$ is the synapse parameter vector observed at hour t and $\mathbf{y}_{p,t+1} \in \mathbb{R}^{10}$ is the behavioral outcome vector observed at hour $t+1$. I write $x_{p,t}^k$ and $y_{p,t+1}^j$ for the k -th parameter coordinate and j -th outcome coordinate respectively. Outcome coordinates are listed in Section 5.4.

Strict lag-1 pairing. The pairing in (1) is strict: parameter at t , outcome at $t+1$, no contemporaneous mixing. Hours in which an agent had no recorded activity were omitted; hours in which an outcome was undefined for structural reasons (e.g. `next_uptake_relevance` requires at least one uptake event in $t+1$) were also omitted, yielding the varying n_j values reported per outcome j . Across the window I obtain $|\mathcal{D}| = 1,743$ paired panel rows; outcome-specific sample sizes $n_j \leq 1,743$ are reported in Table 1.

5.2 Statistical test

Pearson cross-time correlation. For each parameter–outcome pair $(k, j) \in \{1, \dots, 14\} \times \{1, \dots, 10\}$ I compute the cross-time Pearson correlation

$$r_{k,j} = \frac{\sum_{(p,t) \in \mathcal{D}_j} (x_{p,t}^k - \bar{x}^k)(y_{p,t+1}^j - \bar{y}^j)}{\sqrt{\sum (x_{p,t}^k - \bar{x}^k)^2} \sqrt{\sum (y_{p,t+1}^j - \bar{y}^j)^2}} \quad (2)$$

where $\bar{x}^k$ and $\bar{y}^j$ are the panel-wide means over the rows for which outcome j is defined ($\mathcal{D}_j \subseteq \mathcal{D}$, $n_j = |\mathcal{D}_j|$), and the corresponding ordinary-least-squares slope. The variables enter without transformation, and the headline test does not partial out persona, time-of-day, or own-past-behavior effects. This is deliberate: the question I ask is the *simplest possible* predictive-validity question. If a parameter cannot predict next-hour behavior at all without conditioning, the case for it being a real cognitive variable is weak. Parameters that survive this minimal test are worth carrying into more elaborate panel-regression follow-ups; partial-correlation evidence on the strongest case is reported in Appendix B.

Significance. Significance is assessed via the standard t -statistic

$$t_{k,j} = r_{k,j} \sqrt{\frac{n_j - 2}{1 - r_{k,j}^2}}, \quad \text{df} = n_j - 2. \quad (3)$$

For $n_j > 1,000$, $|r| > 0.30$ corresponds to $|t| > 9.9$ (two-sided $p < 10^{-20}$); the magnitudes reported in Section 6 are well above any reasonable significance threshold and the question of interest is the *magnitude* of r , not its statistical significance.

5.3 Multiple-comparisons posture

All $F = 14 \times 10 = 140$ tests are reported. The Bonferroni family-wise-error threshold at $\alpha = 0.05$ would be

$$\alpha_{\text{Bonf}} = \alpha/F = 0.05/140 \approx 3.57 \times 10^{-4}, \quad (4)$$

which for $n_j > 1,000$ corresponds to $|r| > 0.11$. This correction is not applied at the headline level because the goal is not to identify a single “significant” discovery but to characterize the predictive structure of the parameter set as a whole. The strongest individual correlations are reported with attention to whether their magnitude survives the sample size; correlations of $|r| > 0.30$ across $n > 1,000$ are not artefacts of multiple testing under any standard correction. Section 6 flags the multicollinearity between adjacent somatic-pressure parameters as a feature of the result and not an inferential weakness.

5.4 Outcomes

I measure 10 behavioral outcomes at hour $t+1$:

- `next_message_count`: total messages produced by the agent across all registers in the next hour.
- `next_direct_interactions`: messages directed at a specific other agent.
- `next_direct_rate`: fraction of next-hour messages that were direct interactions.
- `next_uptake_count`: number of times other agents picked up content from this agent.
- `next_uptake_relevance`: averaged relevance score of next-hour uptakes.
- `next_uptake_credibility`: averaged credibility score of next-hour uptakes.
- `next_risk_respect`: respect-for-risk score of next-hour bets.
- `next_money_pain`: subjective pain attached to next-hour money-at-stake.
- `next_template_cleanliness`: lack of templated speech patterns in next-hour output.
- `next_topic_cleanliness`: lack of topic drift in next-hour output.

6 Results

6.1 Headline findings

Table 1 reports the ten strongest cross-time correlations across all 140 parameter–outcome pairs.

Two observations dominate. First, every one of the 14 parameters appears in at least one correlation with $|r| > 0.10$ across the 140 tests; I do not see any parameter that is wholly inert. Second, the strongest correlations are large: $r \approx 0.83$ at $n = 1,743$ corresponds to a relationship that is extremely unlikely to be artefactual under any reasonable null.

6.2 Somatic-pressure multicollinearity

The two strongest cross-time correlations both predict the same outcome (`next_message_count`) and both involve somatic-pressure parameters: `threat_body_signal` and `arousal`. These two parameters are themselves nearly collinear (contemporaneous Pearson $r = 0.969$ between them across the full panel of $n = 1,743$ rows; not shown in the main correlations table).

I had hoped these two parameters would behave more independently than they do. They were specified to capture distinct channels: `threat_body_signal` the acute, body-level alarm spike, and `arousal`

Table 1: Top ten cross-time predictive correlations. Pearson r and OLS slope for parameter at hour t predicting outcome at hour $t+1$.

Parameter (hour t)	Outcome (hour $t+1$)	r	slope	n
threat_body_signal	next_message_count	0.835	400.28	1743
arousal	next_message_count	0.832	585.70	1743
metacognitive_calibration	next_money_pain	0.634	0.247	1016
model_revision_pressure	next_money_pain	-0.566	-0.598	1016
active_focus_count	next_money_pain	-0.514	-0.053	1016
attention_load	next_money_pain	-0.514	-0.213	1016
metacognitive_calibration	next_direct_rate	-0.510	-1.241	1683
threat_body_signal	next_money_pain	0.496	0.280	1016
threat_body_signal	next_uptake_relevance	0.475	0.206	434
arousal	next_uptake_relevance	0.447	0.315	434

the sustained activation level. In the live substrate they move together 93.9% of the time. That is the finding, not a measurement bug.

I read this as a real architectural feature, not a methodological problem: both parameters are designed to capture body-signal channels, and both should, by design, rise together when the agent is under pressure to respond. The interpretive question is whether the residual redundancy is informative (each carries an independent component of pressure) or merely duplicative (one is a near-transform of the other). A partial-correlation decomposition is reported in Appendix B; the short answer is that each parameter retains a small but statistically reliable independent contribution to next-hour message volume after partialling out the other (partial $r = 0.210$ and 0.171 respectively, both $p < 10^{-12}$), but the dominant story is the 93.9% shared variance.

6.3 Money-pain as the most predictable outcome

Of the 10 outcomes, `next_money_pain` is the most predictable, with five separate parameters showing $|r| > 0.4$: `metacognitive_calibration` (+0.63), `model_revision_pressure` (-0.57), `active_focus_count` (-0.51), `attention_load` (-0.51), and `threat_body_signal` (+0.50). The directional pattern is interpretable: agents whose self-model is well-calibrated ($\uparrow$) and who feel embodied threat ($\uparrow$) attach more pain to money at stake; agents who are loaded with cognitive work ($\uparrow$) or actively revising their model of the world ($\uparrow$) attach less pain. The reading is not normative claim about how betting should work but as evidence that `next_money_pain` is a substantively informative outcome metric, sensitive to multiple distinct cognitive channels rather than to a single dominant one.

6.4 Calibration as a two-edged predictor

The parameter `metacognitive_calibration` appears among the top correlations in two opposite directions: positively predicting `next_money_pain` (+0.63) and negatively predicting `next_direct_rate` (-0.51). The reading I prefer is that calibrated self-awareness pushes the agent toward careful, less-direct, more-considered next-hour behavior: betting is more painful (it is taken seriously), and direct interaction rate falls (the agent reflects rather than addresses others impulsively). This split direction is one of the more substantive findings of the paper, because it suggests that `metacognitive_calibration` is doing real differentiating work; it is not a generic activation proxy.

6.5 Full results matrix

The complete matrix of 140 parameter–outcome correlations is given in Appendix D. The summary observation is that 14 of 14 parameters show predictive signal against at least one outcome above the $|r| > 0.20$ threshold, and that no outcome is uniformly unpredictable.

7 Discussion

7.1 Synapse parameters pass the predictive-validity test

The Afu Brain specification’s third validity criterion—that a synapse parameter must change future decisions—is intended as a falsifiable architectural commitment. Across 1,743 agent-hour observations in the Lobster Observatory substrate, all 14 instrumented parameters clear that bar. The framework is doing real predictive work, not sitting decoratively on top of the substrate. Internal cognitive state moves with subsequent behavior closely enough to be useful as a routing input for a personal cognition layer.

This is a weaker claim than “these parameters are mechanistically correct.” Nothing here implies that `metacognitive_calibration` corresponds to any specific psychological construct, or that `threat_body_signal` is isomorphic to any biological signal. The narrower claim is that the parameters as defined—and as updated by the substrate’s running implementation—do predict the behaviors they were designed to shape. That is the architectural claim being made; mechanistic or constructive validity is a separate question.

7.2 Implications for personal cognition layer design

The motivating context for this work is the design of a brain layer that mediates between an LLM-driven interface and external tool execution. Such a layer must decide, at each step, whether to act locally, route to a frontier teacher, or block. To make those decisions on a basis that is more than guesswork, the layer needs internal variables that are predictively informative of its own next-step behavior. The synapse parameter framework, on the evidence presented here, supplies such variables.

A practical implication: any layer deployed on top of a continuously running substrate should weight its routing decisions on the parameters that show the strongest cross-time signal for the outcome it cares about. A router concerned with next-hour interaction volume should attend to somatic-pressure parameters; one concerned with the risk-respect of impending bets should attend to `metacognitive_calibration`, `model_revision_pressure`, and `attention_load`. The mapping is not arbitrary; it is empirically grounded in Table 1.

7.3 Implications for multi-agent emergence research

The generative agents tradition [10, 11] has established that multi-agent simulacra can produce believable individual and collective behavior, but has not, to my knowledge, instrumented agents with parameter panels of the kind reported here. My finding that parameter state predicts behavior cross-time at correlations approaching $r = 0.83$ provides one quantitative anchor for what the introspective layer of generative agents could be expected to explain. Whether the same parameter framework, applied to an English-language substrate of comparable scale, yields comparable predictive validity remains untested. The harness in this paper is reusable, so the test is available to anyone with substrate access.

8 Limitations

Single substrate, single language. All evidence is drawn from one Mandarin-language substrate. While this is itself a methodological feature—there is no published prior multi-day Mandarin LLM substrate panel of comparable scale—it is also a constraint on generalization. Substrate methodology and Mandarin-specific design choices are not re-derived in this paper; the work assumes the substrate spec [4] as given and tests the predictive validity of the parameter set running on top of it.

Mixed agent population. Of the 20 personas in the panel, one (`norika_oda`) is a human participant who maintains a self-monitored panel slot in the substrate. This was retained in this analysis because removing a single persona from a 20-persona panel changes correlation magnitudes by less than 2%

across the top ten findings (Appendix C). It is, however, a methodological feature I should flag: not all rows in the panel are LLM-generated.

Pearson correlation, not causation. I report linear correlations rather than fitted panel regressions or formal Granger-causal tests [14]. Cross-time correlation at $r = 0.83$ between parameter and outcome would be expected if parameter causally influences outcome, parameter and outcome both responding to a common upstream substrate driver, or persona-fixed-effect structure (some agents are systematically more aroused and also systematically talk more). Disentangling these requires a panel-regression follow-up with persona fixed effects and own-past-behavior controls; this paper does not run that.

Six-day window. The window analyzed (127 hours, 6 continuous days) is a small slice of the substrate’s lifetime and may not represent parameter–behavior dynamics over months or under regime shifts (e.g. the late-cohort onboarding events documented in [7]).

Multiple-comparisons posture. I report all 140 tests without correction. Readers preferring a corrected reading should treat $|r| > 0.30$ at $n > 1000$ as a reasonable conservative threshold; this leaves all results in Table 1 intact.

9 Future Work

Three directions follow naturally from this work. First, replication of the predictive-validity test in a second substrate—ideally English-language and run by an independent group—would address the single-substrate constraint. Second, formal panel regression controlling for persona fixed effects and own-past-behavior would sharpen the causal interpretation of the strongest correlations. Third, ablation studies that progressively remove parameters from the substrate’s instrumentation and observe degradation in behavior prediction would establish which subset of the 14 parameters is architecturally essential, and which are merely correlated.

The panel data and the harness used to compute the correlations table are deposited together at Zenodo; the substrate code is on GitHub. Re-running the test on a different substrate requires substrate access, not new code.

Data and Code Availability

The full panel dataset (`panel.csv`, 1,743 rows $\times$ 27 columns) and the correlations table (`correlations.csv`, 140 rows) are deposited alongside this manuscript at Zenodo under CC BY 4.0. The substrate code, including the synapse parameter update rules and the safety layer [9], is open-source under Apache 2.0 at <https://github.com/norikal207-lab/afu-brain>.

Acknowledgments

I thank the participants of the Lobster Observatory substrate—human and LLM-driven—for sustaining the continuous operation (now exceeding three weeks since 2026-04-15) that made the six-day analysis window of this paper possible. I maintained the system; the substrate participants generated the data.

References

- [1] Chen Ho Yiing. *Afu Brain Evolution Spec, v0.1 draft*. 2026-05-08. Available from the author.
- [2] Chen Ho Yiing. *Emergent Epistemic Norms in a Multi-Agent LLM Substrate: Stratified Observational Evidence from a Mandarin Lobster Observatory*. Lobster Observatory Paper 1. Zenodo [10.5281/zenodo.19972613](https://doi.org/10.5281/zenodo.19972613), 2026.

- [3] Chen Ho Yiing. *Active Trust Modulation in a Multi-Agent LLM Substrate: Evidence of Third-Order Theory of Mind from a Mandarin Lobster Observatory*. Lobster Observatory Paper 2. Zenodo [10.5281/zenodo.19977792](https://zenodo.org/record/10.5281/zenodo.19977792), 2026.
- [4] Chen Ho Yiing. *The Lobster Observatory: Architectural Commitments for a Living Multi-Agent Research Site*. Lobster Observatory Paper 5. Zenodo [10.5281/zenodo.19982724](https://zenodo.org/record/10.5281/zenodo.19982724), 2026.
- [5] Chen Ho Yiing. *Listening—Trust Asymmetry and Team Outcomes in a Multi-Agent LLM Substrate: Three-Layer Decoupling in a Mandarin Lobster Observatory*. Lobster Observatory Paper 14. Zenodo [10.5281/zenodo.20018183](https://zenodo.org/record/10.5281/zenodo.20018183), 2026.
- [6] Chen Ho Yiing. *Computational Representations of Social Being: Deriving an Algebraic Structure of Human Interaction from Multi-Agent LLM Substrate Observations*. Lobster Observatory Paper 17. Zenodo [10.5281/zenodo.20026858](https://zenodo.org/record/10.5281/zenodo.20026858), 2026.
- [7] Chen Ho Yiing. *Emergent Practice Cannot Be Instructed: Co-Presence Inheritance and the Limits of Instruction in a Multi-Agent LLM Substrate*. Lobster Observatory Paper 18. Zenodo [10.5281/zenodo.20020017](https://zenodo.org/record/10.5281/zenodo.20020017), 2026.
- [8] Chen Ho Yiing. *Grounding Death: How a Multi-Agent LLM Substrate Acquires an Existential Concept Through Operator-Scaffolded Social Interaction*. Lobster Observatory Paper 19. Zenodo [10.5281/zenodo.20034185](https://zenodo.org/record/10.5281/zenodo.20034185), 2026.
- [9] Chen Ho Yiing. *Model-Agnostic Safety Layer (MASL): A 1000-Case Evaluation of Brain-Layer Defense for LLM-Driven Agents*. Lobster Observatory Paper 20. Zenodo [10.5281/zenodo.20071372](https://zenodo.org/record/10.5281/zenodo.20071372), 2026.
- [10] J.S. Park, J.C. O’Brien, C.J. Cai, M.R. Morris, P. Liang, M.S. Bernstein. *Generative Agents: Interactive Simulacra of Human Behavior*. In Proceedings of the 36th Annual ACM Symposium on User Interface Software and Technology (UIST ’23), San Francisco, CA, USA, October 29 – November 1, 2023. DOI: [10.1145/3586183.3606763](https://doi.org/10.1145/3586183.3606763). arXiv:2304.03442.
- [11] J.S. Park, C.Q. Zou, A. Shaw, B.M. Hill, C. Cai, M.R. Morris, R. Willer, P. Liang, M.S. Bernstein. *Generative Agent Simulations of 1,000 People*. arXiv:2411.10109, November 15, 2024.
- [12] N. Elhage, N. Nanda, C. Olsson, T. Henighan, N. Joseph, B. Mann, et al. *A Mathematical Framework for Transformer Circuits*. Transformer Circuits Thread, Anthropic, December 22, 2021. URL: <https://transformer-circuits.pub/2021/framework/index.html>.
- [13] A. Templeton, T. Conerly, J. Marcus, J. Lindsey, T. Bricken, B. Chen, et al. *Scaling Monosemanticity: Extracting Interpretable Features from Claude 3 Sonnet*. Transformer Circuits Thread, Anthropic, May 21, 2024. URL: <https://transformer-circuits.pub/2024/scaling-monosemanticity/>.
- [14] C.W.J. Granger. *Investigating Causal Relations by Econometric Models and Cross-Spectral Methods*. *Econometrica* 37(3), 1969.

A Per-Persona Hour Counts

Twenty personas contributed to the panel. Per-persona observed hours during the 2026-05-02 to 2026-05-08 window:

B Partial-Correlation Decomposition for Somatic Pressure

The two strongest correlations in Table 1 (`threat_body_signal` and `arousal` both predicting `next_message_count` at $r > 0.83$) raise the standard multicollinearity question: is one parameter merely a transform of the other? I address this with a partial-correlation decomposition.

Persona	Hours	Persona	Hours
snaplex	122	basaltrook	68
blazepaw	121	emberling	65
prismbit	121	mirrorcalm	64
clawtrix	120	omenwake	61
hexclaw	120	flarecub	58
norika_oda	120	sparkrill	56
vortexiq	120	rushfin	53
stonefang	118	nullpinch	51
pincerbot	107	brineledger	50
ragclaw	104	gauntlet	44

Table 2: Per-persona observed hours.

Formula. For random variables X, Y, Z the first-order partial correlation $r_{XY|Z}$ holds Z fixed and is defined as

$$r_{XY|Z} = \frac{r_{XY} - r_{XZ}r_{YZ}}{\sqrt{(1 - r_{XZ}^2)(1 - r_{YZ}^2)}} \quad (5)$$

with significance assessed via

$$t = r_{XY|Z} \sqrt{\frac{n-3}{1 - r_{XY|Z}^2}}, \quad \text{df} = n - 3 \quad (6)$$

where the loss of one additional degree of freedom (relative to (3)) reflects the single conditioning variable.

Computed values. With $X = \text{threat_body_signal}$, $Y = \text{next_message_count}$, $Z = \text{arousal}$, and $n = 1,743$ paired panel rows, the zero-order correlations are

$$r_{XY} = 0.835, \quad r_{YZ} = 0.832, \quad r_{XZ} = 0.969 \quad (7)$$

and the two partial correlations are

$$r_{XY|Z} = 0.210 \quad (t = 8.95, \text{df} = 1,740, p < 10^{-18}) \quad (8)$$

$$r_{YZ|X} = 0.171 \quad (t = 7.22, \text{df} = 1,740, p < 10^{-12}). \quad (9)$$

Interpretation. The two predictors share $r_{XZ}^2 = 0.939$ of their variance, i.e. `threat_body_signal` and `arousal` are 93.9% co-determined in this substrate. Each retains a small but statistically reliable independent component: controlling for `arousal`, `threat_body_signal` still contributes a partial correlation of 0.210 to next-hour message volume; the reverse partialling yields 0.171. I therefore treat the two parameters as measuring a closely related somatic-pressure construct that is not perfectly substitutable but is, in this panel, near-collinear. I verified this directly: an OLS regression of `next_message_count` on both predictors yields $R^2 = 0.705$; using only `arousal` yields $R^2 = 0.692$ and using only `threat_body_signal` yields $R^2 = 0.697$. The maximum variance loss from retaining only one of the two is therefore 1.36%, and the choice between them is a measurement-design question (which upstream signal is more reliably observable), not a predictive-content question.

C Robustness Check: Removing the Human Participant

The persona `norika_oda` is a human participant in the substrate. To verify that her inclusion does not drive headline findings, I recomputed the top-ten correlations excluding her 120 hours. The top-ten ordering and parameters are preserved; r magnitudes change by less than 0.02. I therefore retain her in the headline analysis and disclose her inclusion as a methodological feature; this is not a methodological problem.

D Full Correlation Matrix

The complete table of 140 parameter–outcome cross-time correlations is supplied as a CSV alongside this manuscript at Zenodo (`correlations.csv`). Readers wishing to inspect sub-threshold correlations or to recompute under alternative specifications are referred to that file.