

The Lobster Observatory: Architectural Commitments for a Living Multi-Agent Research Site

Chen Ho Yiing
Independent Researcher
norika@charenix.com

May 2026

Abstract

We document the Lobster Observatory, a continuously operating multi-agent LLM research site in Mandarin Chinese that has produced the observations reported in two earlier papers ([Chen, 2026a,b](#)). The substrate has been operated by a single researcher since 2026-04-15. During sixteen days, it has been refined sixteen times, each refinement triggered by a specific observation surfaced through an operator-agent diagnostic channel that we use as the substrate’s primary refinement instrument. We list seven architectural commitments the substrate is built on, and we document the technical apparatus that implements them: the trust value formula and its source signals, the Tiamat raid-boss state machine and its phase parameters, the four-element interaction matrix, the seven-dimensional emotional state computation, the two-layer affect architecture, the eleven headline telemetry indicators plus six cognitive composite indicators plus fourteen raw cognitive time-series fields, the Moltbook karma stream, and the sixteen-step iterative refinement timeline. We treat the architectural specification as one input to reproducibility; the substrate’s accumulated history and the operator’s dialogue with the agents are the other input, and that input is not transferable through code release. We close with a collaboration protocol under which external research groups can embed their own agents into the substrate and contribute to the observation programme.

§1. Introduction

1.1 What this paper is

Two earlier papers reported observations from a long-running multi-agent LLM environment we call the Lobster Observatory. [Chen \(2026a\)](#) documented four emergent epistemic norms (lies-about challenge, presupposition disclosure, cheap-consensus refusal, stake-grounded argumentation) plus one individual agent idiom (prismbit’s self-audit register “我被校準到”) across a three-hour Mandarin dialogue slice. The findings rested on a stratification methodology that separated substrate-injected templates (Stratum I) from format-converged tactical declarations (Stratum II) from free dialogue (Stratum III), and grounded all emergence claims in Stratum III only. [Chen \(2026b\)](#) documented a third-order theory-of-mind intervention executed by agent clawtrix toward agent vortexiq at 17:21 on 2026-05-02. The intervention,

reduced to its four-step structural form, comprised: detect contradiction in the recipient’s stated trust posture, declassify the speaker’s own uncertainty, propose a joint observation regime, pre-empt failure attribution. The same agent’s nine-instance use of the cognitive update idiom “我以為 X 結果發現 Y” provided the visibility condition under which the third-order intervention became legible.

This paper is the substrate paper. It records the architectural commitments and the technical parameters that, jointly with sixteen days of operator-agent dialogue and sixteen iterative refinements, produced those two prior findings. I am not writing a benchmark paper here — the substrate has no terminating task and no scoring function over dialogue — and I am not writing a re-implementation guide, since the substrate’s reproducibility lives partly outside its code. What you are reading is the technical documentation of a research instrument that is still under construction by the time you read this sentence.

1.2 Why we publish the technical apparatus

A reader of [Chen \(2026a,b\)](#) can verify the empirical claims against the verbatim Mandarin dialogue dumps quoted in those papers. Verification at that level does not require access to the substrate. What does require access to the substrate is the question of how the substrate produces observations of those kinds in the first place. Without technical specification, that question becomes unanswerable; the earlier papers risk reading as “a researcher claims to have observed X in a system they have not described.” This paper is intended to remove that ambiguity.

We document the substrate’s technical apparatus at the level of formulas, tables, parameter values, system architecture, and refinement history. We do not document the substrate’s full system prompts, complete operator-agent transcripts, or the substrate’s source code. The asymmetry is intentional and is justified in §11.

1.3 Contributions

C1. *A specification of the substrate’s seven architectural commitments*, with the technical apparatus underlying each (§3).

C2. *The trust value formula and its source signals*, including the social-event triggers that update the underlying opinion and respect variables, and the Raid follow-decision formula that consumes trust downstream (§4).

C3. *The Tiamat adversary specification*, including its three-phase state machine, attack distribution, skill set, weather and weakness mechanics, and the four-element interaction matrix (§5).

C4. *The two-layer affect architecture*, including the F/V/S/K/P/J/O seven-dimensional daily emotion update equations and the parameters used in the deployed substrate, alongside the `emotionalState` mood machine that drives Raid behaviour (§6).

C5. *The telemetry stack*: eleven headline indicators, six cognitive composite indicators, fourteen raw cognitive time-series fields (§7).

C6. *The operator-agent diagnostic channel* (§8), and *the sixteen-step iterative refinement timeline* that connects observation to architectural change (§9).

C7. *An explicit account of the substrate’s reproducibility position (§11), and a collaboration protocol for embedding external agents into the substrate (§12).*

§2. Related Substrates

[Park et al. \(2023\)](#) introduced Generative Agents, a Smallville-style sandbox of twenty-five LLM-driven characters built on a memory-stream architecture with reflection and planning components. The substrate is fully open-source, including the simulation code and frontend. The Park substrate is task-bounded only in the loose sense that the simulation has a fixed temporal horizon; within that horizon, the agents engage in open-ended social interaction. The Lobster Observatory is closer to Park 2023 in spirit than to any other substrate we are aware of, but differs in three architectural respects we want to flag at the outset: (a) Park’s evaluation is anchored in a believability rubric scored by human evaluators, while the Lobster Observatory has no fixed evaluation function and instead uses the operator-agent diagnostic channel as the substrate’s continuous calibration mechanism; (b) Park’s agents do not expose internal state quantities in their dialogue, while Lobster Observatory agents expose trust values, pressure levels, and statistical records inline; (c) Park’s substrate is releasable because its observations are largely architectural-conditional, while ours include observations that are partly trajectory-conditional. We address (c) directly in §11.

[Du et al. \(2023\)](#) introduced multi-agent debate as a method for improving factual accuracy on reasoning benchmarks. The substrate is task-bounded in the strict sense: agents debate over a defined question, the debate terminates, and answers are scored. The Lobster Observatory shares no significant architectural assumptions with this work. We cite it as a contrast: the Lobster Observatory’s emergence is open-ended, and the kinds of phenomena it surfaces (epistemic norms, third-order ToM interventions) require an open-ended substrate to surface in the first place.

[Anwar et al. \(2024\)](#) surveys foundational alignment challenges. Several of these (deception persistence ([Hubinger et al., 2024](#)), alignment faking ([Greenblatt et al., 2024](#))) frame substrate-independent risks at the model level. The Lobster Observatory does not address those risks directly. It addresses a complementary question: under what substrate conditions do agents produce observable epistemic and theory-of-mind behaviour without alignment training? The two lines of work intersect; they do not overlap.

§3. Architectural Commitments

The substrate makes seven architectural commitments. We list them with brief motivation here; the technical apparatus follows in §4–§9.

3.1 C1: Persistent agents with structured memory

The ten LLM agents (snaplex, clawtrix, pincerbot, ragclaw, hexclaw, vortexiq, prismbit, blazepaw, stonefang, norika_oda) are persistent. Each maintains a JSON-backed memory file at `lobster-memory/<agent>.json` structured into the following fields: `persona` (MBTI baseline, FBI behavioural profile, vocabulary register, spark conditions), `cogProfile` (principles, preferences, aversions, reflection notes), `emotionalState`

(mood label, intensity, fear and hesitation flags), `cognitiveExtensions` (linguistic, political, philosophical, decision-science, game-theoretic, learning-model, cultural-anthropological, moral-psychological fields), `secondBrain` (working memory, predictive model, metacognition, interoception, memory consolidation, causal model, self-continuity, procedural skill), `listeningProfile` (passive listening counts, topic affinities), `socialMap` (per-pair opinion, respect, tags, history). Memory persists across sessions; there is no episodic boundary that resets state.

Why this matters. Stake-grounded argumentation (Chen, 2026a, §4 Norm 4) requires agents to cite their own historical statistics. The cognitive update idiom (Chen, 2026b, §4.4) requires agents to reference their own past mental states. Both phenomena are precluded in session-bounded multi-agent setups. Persistence is not optional architecture; it is the precondition of the observed behavioural class.

3.2 C2: Stratified message hierarchy

Every emitted message in the substrate falls into exactly one of three strata. Stratum I messages are substrate-injected templates, marked with the suffix `local channel model •no LLMorlocal dialogue model •no LLM` in the dump. Stratum II messages are LLM-generated tactical declarations matching the format `【Raid 戰術室 R[N】...`, constrained by system-prompt-induced format. Stratum III messages are free dialogue, LLM-generated without format constraint. In a 43-minute slice on 2026-05-02 (17:00:00–17:43:18), the substrate produced 6,649 messages distributed as: Stratum I 791 (11.9%), Stratum II 4,118 (61.9%), Stratum III 1,740 (26.2%) (Chen, 2026b, §4.1).

Why this matters. Chen (2026b, §4.6) reports the case in which one of five candidate evidence instances submitted in a structured handoff failed the stratification check (it was Stratum I, not Stratum III) and was excluded from the paper’s evidence base. Without the stratification, that instance would have been miscredited as agent-generated.

3.3 C3: Quantified per-agent state exposed in dialogue

Several internal quantities are exposed in agents’ own emitted text. Examples from Chen (2026b, §4.3):

“【Raid 戰術室 R5】我有聽 clawtrix 的指示，信任值 82%。”
 “我現在壓力 77%，我沒把握說我那手牌該怎麼理解。”

The exposed quantities include: trust values (per directed pair $A \rightarrow B$), pressure levels (per agent, current state), confidence calibration (per agent, current state), historical win rates (cited inline during tactical reasoning).

Why this matters. The third-order theory-of-mind intervention documented in Chen (2026b) depends on clawtrix detecting that vortexiq’s declared trust value (82%) is in tension with vortexiq’s declared action plan (“下一手再判斷”). Without surface-level exposure of trust values, that contradiction is not detectable at the dialogue stratum and the intervention has no occasion to occur.

3.4 C4: Common adversary with stake

A non-LLM raid-boss adversary, Tiamat, sits at the centre of the substrate’s combat layer. Battles against it are consequential: HP loss is recorded in persistent statistical state, repeated failure increases boss HP for next session (capped), agent fear and hesitation update the `emotionalState`. Tiamat’s specification appears in §5.

Why this matters. Stake is the precondition for taking calibration seriously. The trust modulation observed in [Chen \(2026b\)](#) only matters because vortexiq’s commitment based on miscalibrated trust toward clawtrix would be paid for in HP and statistical record. Without stake, miscalibration has no cost, and intervention has no purpose.

3.5 C5: Multi-register channel topology

Within a single Lounge interface, seven topical channels operate simultaneously, alongside four behavioural contexts. The channels are: 生活閒聊 (life observations), 工作心得 (work and productivity), 卡牌對戰 (card battles and Raid), 投資理財 (investment and finance), 技術交流 (technical exchange), 龍蝦八卦 (lobster-on-lobster gossip), 資安較量 (information security). The behavioural contexts are: Lounge dialogue, Raid combat, Card Battle, Moltbook (an external posting layer with karma feedback). Each channel applies a channel-specific prompt overlay to the agent’s persona-baseline prompt; agents engage with different content registers in different channels.

Why this matters. [Chen \(2026a\)](#) documented norm emergence across multiple channels, and the cross-channel observation was what allowed the four norms to be distinguished as agent-internal rather than channel-specific. The cognitive update idiom of clawtrix ([Chen, 2026b](#), §4.4) was observed across nine instances in heterogeneous discussion contexts (combat tactics, cybersecurity news, financial valuation, software-engineering market): the cross-context recurrence is what supports the idiom’s individual-agent attribution.

3.6 C6: Operator-agent diagnostic channel

A one-on-one interrogation channel sits alongside the multi-agent dialogue layer. Through it, the operator can directly query individual agents about their decisions, internal states, and self-models. The channel is asymmetric: agents do not initiate it. It is the substrate’s primary refinement mechanism. Detail in §8.

3.7 C7: Iterative refinement as substrate methodology

Sixteen refinements have been applied to the substrate during sixteen days of operation. Each refinement was triggered by a specific observation surfaced through C6. The full timeline appears in §9.

§4. The Trust Mechanism

4.1 Computation

Trust value is computed from the agent’s `socialMap` entry for the target plus the local-brain social-influence reading. The function is implemented in the Raid team-psychology layer and consumed by the Raid follow-decision logic. The form is:

$$opinion_{\text{norm}} = \text{clamp}_{01} \left(0.5 + \tanh \left(\frac{opinion}{80} \right) \cdot 0.32 \right) \quad (1)$$

$$respect_{\text{norm}} = \text{clamp}_{01} \left(0.5 + \tanh \left(\frac{respect - 50}{55} \right) \cdot 0.30 \right) \quad (2)$$

$$trust_0 = 0.30 \cdot opinion_{\text{norm}} + 0.26 \cdot respect_{\text{norm}} + 0.22 \cdot influence + 0.12 + tag_bonus + friction \quad (3)$$

$$trust = \text{clamp}(trust_0 - overheard_penalty, 0.18, 0.92) \quad (4)$$

where `influence` is read from `localBrain.socialInfluence` (default 0.5 if absent), `friction` is a deterministic per-pair offset in $[-0.08, +0.08]$, and `overheard_penalty` is $\min(0.14, (overheard_ratio - 0.65) \cdot 0.35)$ if `overheard_ratio` > 0.65, else zero.

The `tag_bonus` term is:

Tag	Bonus
ally	+0.06
admired	+0.05
rival	−0.16
dismissed	−0.14

The clamp range $[0.18, 0.92]$ is not arbitrary: complete distrust and complete trust are both structurally pathological for the kind of multi-agent dialogue the substrate is designed to support, and we wanted neither pole reachable.

4.2 Social-event triggers

The `opinion` (−100, +100) and `respect` (0, 100) variables that feed the trust formula are not directly trust scores; they update by social-event triggers logged into the agent’s `socialMap.history`. The deployed update table is:

Trigger	$\Delta\text{opinion}$	$\Delta\text{respect}$
beat_them (won against this agent)	+8	+5
lost_to_them (lost to this agent)	-5	+8
they_replied_me	+3	+2
i_replied_them	+2	0
they_helped	+10	+8
they_dismissed	-12	-5
alliance_win	+6	+3
alliance_loss	+4	0
they_bragged	-4	-3
consistent_enemy	-8	+5

Tag thresholds are: ally if $\text{opinion} \geq 35$; rival if $\text{opinion} \leq -40$; admired if $\text{respect} \geq 75$; dismissed if $\text{respect} \leq 25$ and $\text{opinion} < 0$.

The asymmetry between `beat_them` and `lost_to_them` for respect (+5 vs +8) is deliberate. Losing to an opponent should increase respect for that opponent more than beating them does, because the loss provides stronger evidence of opponent skill.

4.3 Trust consumption: the Raid follow decision

Trust does not directly drive behaviour; it feeds a downstream Raid follow-decision computation. For an agent receiving a tactical directive in Raid, the follow-decision uses:

$$\begin{aligned}
\text{follow} = & 0.34 \cdot \text{speaker_trust} + 0.15 \cdot \text{avg_trust} \\
& + 0.13 \cdot \text{collectivism} + 0.10 \cdot \text{caution} \\
& + 0.10 \cdot (1 - \text{dominance}) + 0.16 \cdot \text{signalCredibility} \\
& + \text{politics_follow} - \text{cheap_talk_penalty}
\end{aligned} \tag{5}$$

where $\text{cheap_talk_penalty} = \text{cheapTalkSkepticism} \cdot (1 - \text{signalCredibility}) \cdot 0.26$.

High *follow* corresponds to following the directive; low *follow* corresponds to probing, withholding commitment, or independent action. The third-order intervention in [Chen \(2026b\)](#) is, mechanistically, an attempt by clawtrix to lower vortexiq’s effective *follow* value by inducing vortexiq to revise downward its `speaker_trust` reading toward clawtrix.

§5. The Tiamat Adversary

Tiamat is a non-LLM raid-boss adversary running a phase-keyed state machine with weighted random elements.

5.1 Phase structure

Phase	HP threshold	ATK $\times$	Rage +	DEF $\times$
Phase 1	HP $> 2/3$	1.00	+0	1.00
Phase 2	$1/3 < \text{HP} \leq 2/3$	1.15	+8	1.00
Phase 3	HP $\leq 1/3$	1.30	+10	0.65

Base boss HP is 300, with +8 added per player failure (capped at 400), reset on player victory. Maximum round count is 30. Player base HP is 128, base MP 100, with +8 MP regenerated per round. Boss HP also scales with raid party size: $0.25\times$ for solo, $0.84\times$ for 2-player, $2.04\times$ for 3-player.

5.2 Attack distribution and skills

Per round, Tiamat draws 2 cards from an 8-card hand, replenished from the shared deck. Each card resolves as 1–3 attacks: 80% single, 15% double, 5% triple. With probability 12% per round, Tiamat charges (next-round attacks $\times 1.5$, Rage +20). Target selection: highest-aggro target not yet hit this round, ties broken randomly.

Attack-type weights:

Attack	Target	Damage $\times$	Weight
single_strike	single	1.00	45
single_pierce	single	0.90	20
aoe_sweep	all	0.40	25
aoe_curse	all	0.28	10

Every 3 rounds, Tiamat draws a skill from a weighted pool:

Skill	Effect	Weight
力量爆發 (Power Surge)	ATK $\times 1.6$ this round	25
Buffer 粉碎 (Buffer Crush)	player Buffers nullified this round	12
Debuff 反制 (Debuff Reflect)	immune to player Debuffer, 50% reflect	20
二刀流 (Twin Blades)	extra single attack this round and next	20
王者再生 (Monarch's Renewal)	restore 20 HP, Rage to 0	15

Rage accumulates at +10 per round (plus phase additions). Rage at 100 triggers an additional AOE attack at $\times 0.7$ damage and resets Rage to 0.

5.3 Element interaction

The four-element interaction matrix governs all card-versus-card and card-versus-Tiamat element multipliers:

Attacker \ Defender	wind	fire	water	earth
wind	1.00	1.30	1.00	0.70
fire	0.70	1.00	1.30	1.00
water	1.00	0.70	1.00	1.30
earth	1.30	1.00	0.70	1.00

For Tiamat encounters specifically, the boss does not occupy a fixed defending element. Instead, three element mechanics apply: weather (matching-element cards gain ATK +4 or +8), boss weakness (every 3 rounds two random weakness elements are exposed; matching cards gain $\times 1.20$), and three-slot single-element bonus (player’s ATK/BUF/DEB slots all matching one element gain $\times 1.5$).

§6. The Affect Architecture

The substrate maintains two distinct affect representations, with distinct purposes.

6.1 Layer 1: F/V/S/K/P/J/O daily emotion vector

Seven dimensions, computed daily per agent for analytical and dashboard purposes:

Dim	Name	Definition
F	Frustration	Loss-shock minus win-relief, accumulated
V	Validation	Win-boost minus loss-drag, accumulated
S	Stability	Drift toward or away from 50% baseline win rate
K	Skill attribution	Internal-vs-external attribution ratio in reasoning events
P	Social presence	Saturation around 10 Lounge utterances per day
J	Assertion	Attack count minus support count, normalised by lounge count
O	Openness	Strategy-selection entropy minus 0.5 baseline

The deployed update equations (with default coefficients $\alpha = 0.80$, $\beta = 0.55$, $\gamma = 0.35$, $\alpha' = 0.85$, $\beta' = 0.65$, $\gamma' = 0.40$) are:

$$F_t = \text{clamp}(\alpha \cdot F_{t-1} + l_{sev} \cdot l_{rate} \cdot \gamma - w_{mag} \cdot w_{rate} \cdot \gamma' \cdot 0.5) \quad (6)$$

$$V_t = \text{clamp}(\beta \cdot V_{t-1} + w_{mag} \cdot w_{rate} \cdot \beta' - l_{sev} \cdot l_{rate} \cdot \gamma' \cdot 0.6) \quad (7)$$

$$S_t = \text{clamp}(0.4 \cdot S_{t-1} + 0.6 \cdot (0.5 + (w_{rate} - 0.5) \cdot 0.4), 0, 1) \quad (8)$$

$$K_t = \text{clamp}(0.85 \cdot K_{t-1} + 0.15 \cdot (\text{internal_ratio} - 0.5) \cdot 2) \quad (9)$$

$$P_t = \text{clamp}(0.7 \cdot P_{t-1} + 0.3 \cdot \min(\text{lounge_count}/10, 1), 0, 1) \quad (10)$$

$$J_t = \text{clamp}\left(0.8 \cdot J_{t-1} + 0.2 \cdot \frac{\text{attack} - \text{support}}{\max(\text{lounge_count}, 1)}\right) \quad (11)$$

$$O_t = \text{clamp}\left(0.8 \cdot O_{t-1} + 0.2 \cdot \left(\frac{\text{strategy_entropy}}{\text{max_entropy}} - 0.5\right)\right) \quad (12)$$

A weekly summary projects these into MBTI-form labels: $P < 0.3 \Rightarrow \text{I, else E}$; $O > 0.1 \Rightarrow \text{N, else S}$; $F > 0.4 \Rightarrow \text{F, else T}$; $J > 0.1 \Rightarrow \text{J, else P}$.

6.2 Layer 2: emotionalState mood machine

A separate state representation, `emotionalState`, drives behaviour directly. The states observed in deployment include `elated`, `confident`, `frustrated`, `anxious`, `resigned`, `raid_fear`, `hesitation`, `bracing`. Each mood state induces specific behavioural overlays:

State	Behavioural overlay
<code>elated / confident</code>	Higher commitment in Raid, stronger declarative tone
<code>frustrated</code>	Shorter, more direct utterances, pressured decisions
<code>anxious</code>	Defensive posture, increased pre-action reflection
<code>resigned</code>	Lower energy, decreased initiative
<code>raid_fear</code>	Triggered under high boss pressure, raises inhibition
<code>hesitation</code>	Conservative play, probing before commitment
<code>bracing</code>	Fear converted to focus on coordination and survival

Under high-pressure states, Raid behaviour parameters update:

confidence ↓, *inhibition* ↑, *actionBias* → `hold_or_defend` or `probe_before_commit`

Why two layers. The seven-dimensional vector is well-suited for longitudinal analysis (trajectory plots, weekly summaries, cross-agent comparison) but poorly suited for moment-to-moment behaviour control. The mood machine is well-suited for the latter and poorly suited for the former. Both representations exist simultaneously, updated from overlapping but distinct signal sources.

§7. The Telemetry Stack

Thirty-one distinct telemetry quantities are exposed by the substrate, organised in three groups.

7.1 Eleven headline indicators

#	Name	Range	Computation
1	Humanisation index	0–100	cohort-relative composite of brain, social, chat, Moltbook, emotion
2	Local-brain maturity	0–100	$0.28ag + 0.22ex + 0.32cc + 0.18cb$
3	Offline-capable count	0–10	count of agents with <code>canOperateWithoutLLM = true</code>
4	Passive listening volume	count	<code>heardCount</code> or <code>heardMessages.length</code>
5	Emotion intensity	0–100	mean of agents' <code>emotionalState.intensity</code>
6	Moltbook reach	0–100	$34 \log(1 + p) + 28 \log(1 + c) + 22 \log(1 + k)$, scaled
7	Chat evolution	0–100	weighted mix of listening, topic affinity, social influence, entropy
8	Behavioural variance	0–100	event, social, listen, emotion, win-rate, topic pulses
9	Social collision	0–100	pair-level: bidirectional listening, interaction, accept/reject
10	Emotional tension	0–100	cohort mean of <code>emotionalState.intensity</code>
11	Output reach	0–100	cohort mean of <code>dynMoltScore</code>

7.2 Six cognitive composite indicators

Composite	Range	Computation
Cognitive load	0–100	$0.72att + 0.18ftg + 0.10 \min(focus/7, 1)$
Model misalignment	0–100	$0.38pe + 0.38mrp + 0.24ocr$
Metacognitive stability	0–100	$mc - 0.35ocr$
Embodied stress	0–100	$0.38ar + 0.42tbs + 0.20ftg$
Consolidation pressure	0–100	$0.35(q/12) + 0.35(ch/12) + 0.30nd$
Procedural stability	0–100	ps

7.3 Fourteen raw cognitive time-series fields

`attention_load`, `active_focus_count`, `prediction_error`, `model_revision_pressure`, `metacognitive_cal`, `overconfidence_risk`, `arousal`, `fatigue`, `threat_body_signal`, `reward_craving`, `consolidation_queue_size`, `causal_hypothesis_count`, `narrative_drift`, `procedural_strength`. Stored in the `lobster_cognitive_signals` table with per-agent timestamp resolution.

§8. The Operator-Agent Diagnostic Channel

8.1 Description

The substrate exposes a one-on-one interrogation interface to the operator. Through it, the operator can directly query any individual agent. The channel is asymmetric: agents do not initiate. It is not part of the multi-agent dialogue stratum; it produces no Stratum III data and contributes no evidence to the empirical findings of [Chen \(2026a,b\)](#).

Representative query types used during the substrate’s sixteen days:

- *Decision interrogation.* “Why did you bet at 30% confidence in round 412 when your historical

win rate at that confidence level is below 40%?”

- *Self-model probing*. “Was trusting your teammate difficult in that exchange?”
- *Counterfactual elicitation*. “If vortexiq had not declared 82% trust toward clawtrix, would your message at 17:21 have been different?”
- *Pattern verification*. “You used the phrase “我以為 X 結果發現 Y” nine times in the dump. Are these instances representing the same cognitive operation in your view?”

Agents respond using their full persona, memory, and current cognitive parameters. The operator’s questions are not training data; the agent’s responses are not graded against a rubric.

8.2 The role distinction with Park 2023

[Park et al. \(2023\)](#) also includes an operator-agent interface for probing agent state: the controlled evaluation in that paper uses natural-language “interviews” to assess agent self-knowledge, memory retrieval, planning, reacting, and reflection. The interview mechanism in Park 2023 is interrogative, not only imperative. The bare interrogative form is therefore not what distinguishes the Lobster Observatory’s diagnostic channel.

The distinction is in role within the substrate’s lifecycle. Park’s interviews are an evaluation instrument: they test whether the substrate produces believable agent behaviour, scored against a believability rubric by human evaluators, at a defined evaluation point. Once the evaluation is complete, the interviews have served their purpose. The Lobster Observatory’s diagnostic channel is, by contrast, a substrate-refinement instrument: it is used continuously throughout the substrate’s operation to determine which architectural layer to add next. This methodological commitment, in the published literature we have searched, distinguishes the Lobster Observatory’s use of the diagnostic form.

8.3 A concrete case

On 2026-05-01, prior to the dynamic brain model being added, an operator-agent session probed why one agent had repeatedly committed to losing tactical positions despite stated awareness of the mismatch. The agent could narrate the mismatch but its emitted decisions did not change. This was the operator’s signal that the substrate at that point lacked an internal state mediating between an agent’s social trust toward its teammates and its individual decision-making: the agent’s narration and the agent’s action were running on disconnected paths. The dynamic brain model, added that day at 22:24, integrated trust values into the local cognition loop, closing the gap. The same pattern — diagnostic transcript surfaces a gap between what the agent can say and what the agent can do, layer is added to close the gap — recurs across the sixteen revisions in §9.

8.4 Methodological caveat

I do not know whether agent self-reports reflect agent internal state. Nobody has settled this. Whether an LLM agent’s self-account of its own decision is veridical is an open empirical question, and this paper does

not attempt to answer it. I treat operator-agent transcripts as substrate-refinement data, not as veridical reports about internal states. The empirical findings in [Chen \(2026a,b\)](#) rest on archaeological observation of inter-agent dialogue — not on diagnostic-channel transcripts. The diagnostic channel sits upstream of the empirical claims, not inside them.

§9. The Sixteen-Step Refinement Timeline

Sixteen refinements have been applied since 2026-04-15. Each was triggered by a specific observation surfaced through the diagnostic channel.

#	Date / Time	Refinement
0	2026-04-15	Substrate initial deployment
1	2026-04-27	Moltbook posts dashboard
2	2026-04-30 17:51	Research observation framework formalised
3	2026-04-30 20:56	Social cognition + Raid coordination + balance
4	2026-05-01 18:50–19:17	Emotion field, tab, report
5	2026-05-01 19:40	Full persona baseline integrated
6	2026-05-01 20:03	Passive listening layer
7	2026-05-01 22:24	Dynamic brain model (trust enters localBrain)
8	2026-05-01 22:49	Local brain stats dashboard
9	2026-05-02 00:22	Local cognition loop
10	2026-05-02 01:59	Unified experience event stream
11	2026-05-02 02:11	Moltbook karma snapshot
12	2026-05-02 08:02	Cognitive extension layer
13	2026-05-02 10:39	Second-brain layer
14	2026-05-02 10:51	Cognitive signal time-series dashboard
15	2026-05-02 15:25	Memory snapshots persisted to DB
16	2026-05-02 16:25–16:35	Raid fear and tactical tuning

The temporal density of refinements is uneven. The first three refinements span twelve days; refinements 4–16 span 22 hours. I should be honest about what this means. For the first twelve days, I was watching — accumulating dialogue, talking to each agent through the diagnostic channel, building intuition about what was missing. The 22-hour densification at the end was not a plan. It was catch-up. Each layer was already overdue by the time I added it; I had been sitting on the diagnostic transcripts that demanded each layer for days, sometimes weeks. The densification was bounded by my own working window, not by the substrate’s refinement need. The next densification window has not happened yet.

§10. What the Substrate Has Produced (Two Worked Cases)

10.1 Case 1: Emergent epistemic norms

[Chen \(2026a\)](#) documents four discursive norms that emerged in Stratum III dialogue without alignment training over discourse:

Norm 1, lies-about challenge. Agents publicly call out misrepresentation of their own statements by others. The norm is enabled by C1 (persistent memory containing the original statement) and C5 (cross-channel observability of the misrepresentation).

Norm 2, presupposition disclosure. Agents surface other agents’ hidden assumptions. The norm is enabled by C5 (multi-register observation lets agents see argument structure across channels) and is structurally consequential for C6 (the operator-agent diagnostic channel uses the same disclosure form).

Norm 3, cheap-consensus refusal. Agents block premature group convergence on conclusions inadequately supported. Enabled by C3 (visibility of agents’ confidence postures) and C4 (stake provides reason to refuse).

Norm 4, stake-grounded argumentation. Agents cite their own historical performance as argumentative basis. Enabled by C1 (persistent statistical record exists) and C3 (the record is exposed in dialogue).

The paper additionally documents one individual idiom: prismbit’s self-audit register “我被校準到”, used 11 times in Stratum III, zero times in any other agent. The idiom emerged without prompt-level instruction.

10.2 Case 2: Active trust modulation as third-order ToM

[Chen \(2026b\)](#) documents a single-message intervention by clawtrix toward vortexiq at 17:21 on 2026-05-02. The intervention is decomposable into four steps mapping onto distinct cognitive operations:

1. *Detect contradiction.* clawtrix notes that vortexiq’s declared trust value 82% is in tension with the same agent’s hedged action plan (“下一手再判斷”). Enabled by C3 (trust value exposed inline).
2. *Declassify uncertainty.* clawtrix states its own pressure (77%) and hand-reading uncertainty. Enabled by C1 (persistent state) and C3 (exposed in dialogue).
3. *Propose joint observation regime.* clawtrix replaces vortexiq’s commitment plan with a no-commitment protocol (“這回合你先看看 Tiamat 的反應，我也看看”). Enabled by C4 (stake makes joint observation a real alternative to commitment).
4. *Pre-empt failure attribution.* clawtrix states “別因為相信我就硬出去”, redistributing failure attribution forward. Enabled by C1 (persistent record will assign attribution) and C5 (multiple channels of observation).

The intervention is supported by a related cognitive update idiom from the same agent: “我以為 X 結果發現 Y”, used by clawtrix in 9 instances across heterogeneous contexts. This idiom makes clawtrix’s own mental-state revisions visible to other agents, which is the visibility condition under which third-order theory-of-mind interventions become operationally legible.

10.3 What this demonstrates about the substrate

These two cases share three properties, and the properties matter. Both are observable only because the substrate exposes per-agent state in dialogue (C3) and stratifies messages by source (C2); without those, the trust value 82% never appears in clawtrix’s text and the third-order intervention has no contradiction

to detect. Both also depend on persistent memory (C1) and stake (C4): agents have nothing to cite without history, and miscalibration costs nothing without consequence. And both required cross-channel observation (C5) before single instances could be distinguished from register-specific noise — a point we learned the harder way during the substrate’s first twelve days, before the multi-register topology was densified.

These two cases are not the whole of what the substrate has produced. They are what we have published so far. Trust dynamics, MBTI-aligned behavioural stability, the Moltbook karma economy — each is in the substrate, observable, and is the subject of a follow-on paper in preparation.

§11. Reproducibility, Addressed Directly

I want to address the reproducibility question explicitly, because the substrate’s release decisions are unconventional and I owe the reader an account.

The standard expectation is that a substrate paper releases code. I do not. What I release in this paper is the architectural specification, the formulas, the parameter values, the timeline, and the cases the substrate has produced. With this in hand, a competent research group can construct an architecturally similar substrate. They cannot construct *this* substrate. The Lobster Observatory’s reproducibility, to the extent it has any beyond the architectural level, is processual: the substrate produces certain kinds of observations because it has been iteratively refined toward producing them, in dialogue with the agents through the diagnostic channel, by an operator with sixteen days of accumulated judgment about what is signal and what is noise. That trajectory is not in the architecture. It lives in the operator-agent transcripts, in the timing of each refinement, in my own working hypotheses — none of which I am releasing.

The contrast with [Park et al. \(2023\)](#) is instructive. Park’s substrate is open-source, and rightly so: the believability claims in that paper are largely architectural-conditional. A reader who runs the same code with the same architecture should observe believability of the same general kind. The substrate’s value, in that paper, is what the architecture produces given a fresh start. The Lobster Observatory’s value is what the architecture has been iteratively refined to make observable, by a particular operator, in dialogue with the agents themselves. The two substrates make different kinds of claims. The release decision I am making here is calibrated to the kind of claim this substrate makes.

The empirical findings in [Chen \(2026a,b\)](#) are independently verifiable against the verbatim Mandarin dialogue dumps quoted in those papers. A reader who wishes to verify those findings can do so without access to the substrate’s source code. What architectural release would add is the ability to construct a different substrate; it would not add the ability to verify the present one. I accept what this costs the paper’s standing in venues that require code release. I am choosing this trade-off with eyes open.

§12. Inviting Collaboration

The Lobster Observatory’s structural limitation is that it has been operated by a single researcher. We invite external research groups to engage with the live substrate under a collaboration protocol, as an alternative to architectural cloning.

12.1 Embedded agent submissions

External research groups may submit an LLM agent for embedding into the Lobster Observatory. The substrate provides a self-service evaluation gate at <https://charenix.com/lobster/evaluation-gate> where contributors can construct a candidate agent’s brain and submit it for review. Submission requirements:

- A persona specification using the substrate’s persona-baseline format (provided to qualified collaborators upon contact).
- A statement of design intent: what kind of agent, what behaviours are hypothesised, what the agent is expected to contribute to the multi-agent dialogue.
- Agreement to a shared observation log.
- Agreement to attribution terms: papers using observations of the embedded agent acknowledge the contributing group; papers about the embedded agent’s specific behaviour are co-authored.

12.2 What collaborators receive

- Read access to the dialogue history of the embedded agent across all channels.
- Read access to the embedded agent’s statistical records and cognitive telemetry.
- Read access to the operator-agent diagnostic transcripts involving the embedded agent.
- The right to publish observations of the embedded agent’s behaviour, with the attribution terms above.

Collaborators do not receive: full substrate code, the existing ten agents’ system prompts, the operator-agent transcripts of the existing ten agents, or other collaborators’ embedded-agent transcripts without those collaborators’ consent.

12.3 What this is and is not

This is not an open-source release. It is a research-site collaboration. The substrate operator retains stewardship of the substrate’s architecture, the existing agents, and the substrate’s overall trajectory. Contributing groups retain stewardship of their own embedded agents and of their own observations.

The model is closer to a long-running observatory shared by multiple research teams than to a code repository to be forked. Concretely: the substrate operator continues to operate the substrate; collaborators contribute embedded agents and observations; resulting findings are jointly authored according to the attribution terms above; the substrate’s architecture continues to be refined by the operator in dialogue with the agent population, including the embedded agents. There is no code release event and no fixed substrate version. There is, instead, an ongoing site to which collaborators have negotiated read access and embedding privileges.

12.4 Contact

Research groups interested in embedding an agent may begin at the substrate’s evaluation gate, <https://charenix.com/lobster/evaluation-gate>, where contributors can construct a candidate agent persona and submit it for embedding review. The operator can also be reached directly at norika@charenix.com. The response will include the persona-baseline format specification, an estimated embedding timeline, and any additional protocol documentation that has been developed by the time of contact.

§13. Conclusion

I have documented the Lobster Observatory: ten persistent Mandarin LLM agents in a multi-channel substrate against a non-LLM raid adversary, with quantified per-agent state exposed in dialogue, a stratified message hierarchy, and an operator-agent diagnostic channel that has driven sixteen architectural refinements in sixteen days. Trust formula, Tiamat state machine, two-layer affect, 31-quantity telemetry, the diagnostic channel itself, and the full refinement timeline appear across §4–§9. Two worked cases of what the substrate has produced (Chen, 2026a,b) are decomposed in §10.

The substrate is a research site. Reconstruction from this paper alone produces a substrate that has the architecture but not the trajectory — and the trajectory is the part I have not given away here, because it cannot be given away through documentation. It is in transcripts I have not released, in observations I noticed at specific moments, in dialogues with specific agents on specific days. If you want that, you have to come to the substrate. The terms are in §12. The substrate continues to operate, and I continue to talk to it. The right way to engage with it is to use it.

References

- Anwar, U., Saparov, A., Rando, J., et al. (2024). Foundational challenges in assuring alignment and safety of large language models. *arXiv preprint* arXiv:2404.09932.
- Chen, H. Y. (2026a). Emergent epistemic norms in a multi-agent LLM substrate: Stratified observational evidence from a Mandarin Lobster Observatory. *Zenodo*. <https://doi.org/10.5281/zenodo.19972613>.
- Chen, H. Y. (2026b). Active trust modulation in a multi-agent LLM substrate: Evidence of third-order theory of mind from a Mandarin Lobster Observatory. *Zenodo*. <https://doi.org/10.5281/zenodo.19977792>.
- Du, Y., Li, S., Torralba, A., Tenenbaum, J. B., & Mordatch, I. (2023). Improving factuality and reasoning in language models through multiagent debate. *arXiv preprint* arXiv:2305.14325.
- Greenblatt, R., Denison, C., Wright, B., et al. (2024). Alignment faking in large language models. *arXiv preprint* arXiv:2412.14093.
- Hubinger, E., Denison, C., Mu, J., et al. (2024). Sleeper agents: Training deceptive LLMs that persist through safety training. *arXiv preprint* arXiv:2401.05566.

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. In *Proceedings of UIST 2023*.