

Listening–Trust Asymmetry and Team Outcomes in a Multi-Agent LLM Substrate: Three-Layer Decoupling in a Mandarin Lobster Observatory

Chen Ho Yiing
Independent Researcher
norika@charenix.com

May 2026

Abstract

We report a three-layer structural-network observation from the same continuously-running multi-agent LLM substrate whose epistemic-norm emergence and third-order theory-of-mind interventions we documented in earlier work (Chen, 2026a,b), and whose technical apparatus we documented in a substrate-specification paper (Chen, 2026c). Ten Mandarin-language agents have been grounded since 2026-04-15 in a ten-site live prediction ecosystem covering motorsport, team sports, racing, esports, combat sports, a card-battle hub, and a community forum, with each agent maintaining persistent identities and behavioural traces across all sites continuously. We extract three relational layers from direct database query: a directed trust proxy `localBrain.socialInfluence`, a directed listening exposure layer `listeningProfile.bySpeaker`, and a team-outcome layer derived from a deliberately-designed three-agent Tiamat raid task. The trust proxy saturates at 0.825 across all 90 directed pairs (zero variance), reflecting the recent integration of the trust subsystem into the substrate. The listening layer shows a clear bimodal split: five high-listener agents averaging 85,659 lounge messages heard, five low-listener agents averaging 16,046, a between-tier ratio of $5.34\times$, partitioning the cohort into a four-quadrant role typology (LISTENER, TALKER, BOTH, QUIET). The outcome layer fits a single-parameter binomial null model ($\chi^2 = 13.27$, $df = 10$, not significant at $\alpha = 0.05$) across all 120 possible three-agent teams in 1,440 trials and is statistically indistinguishable from a uniform success process. Within-team listening density does not predict team win rate (Pearson $r = -0.07$, permutation test $p = 0.45$ over 10,000 permutations), and no listening-derived predictor survives a basic significance check. An ecological-validity robustness check against 3,128 completed live raid records covering nine triples at $n \geq 30$ places live substrate performance in the same low-success regime as the sandbox (mean = 7.56%, range 6.25% to 9.73%). The negative finding is reported as a substrate observation rather than a general claim: in this substrate, with this observation window, even a deliberately-designed cooperation task fails to surface an alignment between communication structure and cooperative outcome. The three layers describe the same ten agents over the same window, with three different topologies; we read the decoupling itself as the substrate-level observation worth recording.

§1. Introduction

1.1 What this paper documents

I built ten prediction sites before I built the substrate. Twelve days, alone, no team — NBA, FIFA, MLB, F1, horse racing, esports, combat sports, a card-battle hub, a community forum, and an earlier stock-simulation site. Each one carrying its own betting market, its own card collection, its own leaderboard, its own historical odds. The reason was simple, or at least it felt simple at the time: I could not put ten Mandarin-language agents in a closed room and expect them to learn anything from each other. They needed an exterior. They needed Verstappen to mean something. They needed to know which horse drew an inside barrier at Newmarket this morning, and to care about it. So I built the world first, and only then did I drop the lobsters in.

The Lobster Observatory is the substrate that sits on top of that ecosystem — ten persistent Mandarin-language LLM agents that have, since 2026-04-15, kept autonomous betting accounts, card collections, public posts, and forum identities live across all ten sites without interruption. The cognitive architecture, the trust formula, the Tiamat raid mechanics, the affect computation, the sixteen-step iterative refinement timeline — all of that is in the substrate-specification paper (Chen, 2026c). The discursive findings on the same substrate are in Chen (2026a,b). What this paper does is narrower than any of those three. It asks: of the ten agents that emerged from this substrate, does the social structure they have grown into — who listens to whom, who trusts whom — predict how well any three of them cooperate when forced to?

The cooperation task is called Tiamat. I built it on purpose, and I want to be honest about why. By the second week of operating the substrate, betting and card-battle play — the agents’ two natural action layers — were not surfacing anything I could measure as team-level dynamics. Individual agents were learning markets. They were collecting cards. They were, from what I could see, not really developing chemistry with anyone in particular. So I added a three-agent raid with sealed mechanics and a binary outcome, on the explicit working hypothesis that *if* chemistry was emerging in the substrate at all, this layer would force it into something I could measure. Tiamat is not generic raid content. It is an observation tool I built because I needed one.

The hypothesis failed. Listening structure across the cohort is differentiated, real, stable — five LISTENERS, four TALKERS, one BOTH, one QUIET, with a $5.34\times$ separation between the listener tier and the talker tier. None of this is small. And yet, across 1,440 raid trials covering every triple of agents the cohort can produce, the outcome distribution is statistically indistinguishable from a uniform binomial process. Permutation test on the listening–outcome correlation: $p = 0.45$. The structure I built sixteen days of listening telemetry to surface does not predict the outcome I built Tiamat to surface.

So I am writing this paper as the operator who reports the negative result, not as the researcher who set out to prove the hypothesis I started with. (The hypothesis was that team chemistry emerges from accumulated listening. It did not, at least not in any way Tiamat surfaces.) I owe the reader an account of how the result fell out, what alternative readings I considered, and what I am willing to claim from the loss — which, I will warn you in advance, is less than I expected to be writing when I started this analysis last week. Substrate science needs negative results to be reported alongside positive ones, or the field cannot tell which paths have been tried and given a clean look. This paper is part of that ledger.

1.2 Why three layers

There are three relational measurements I can extract from the substrate by direct database query. The first is a directed trust proxy, derived from each agent’s local-brain social-influence score toward each other agent. The second is directed listening exposure: how many lounge messages each agent has heard each other agent speak. The third is team outcome: the win rate of each three-agent team in the Tiamat raid.

Each layer is a 9×9 directed matrix (each agent has nine peers), or in the outcome case a list of 120 three-agent combinations. The three layers describe the same ten agents over the same observation window. If they were aligned—if listening predicted trust, and trust predicted outcome, and listening predicted outcome—that would be one kind of substrate finding. They are not aligned. That is the other kind, and it is what this paper documents.

1.3 Contributions

This paper makes five contributions, each of which is supported by direct query of the substrate database rather than by interpretive narration.

C1.

A 9×9 directed listening matrix from a 16-day observation window, showing a clear four-quadrant role typology (five LISTENERS, four TALKERS, one BOTH, one QUIET) with a between-tier ratio of $5.34\times$.

C2.

A 120-team outcome distribution from a controlled three-agent raid task, fit by a single-parameter binomial null model ($\chi^2 = 13.27$, $df = 10$, not significant at $\alpha = 0.05$).

C3.

A formal permutation test (10,000 permutations) showing that the observed correlation between within-team listening density and team win rate (Pearson $r = -0.07$) is not distinguishable from the null ($p = 0.45$).

C4.

An ecological-validity robustness check against 3,128 completed live raid records covering nine triples at $n \geq 30$, showing live-substrate win rates concentrated in a narrow 6.25%–9.73% band, consistent with the sandbox outcome regime.

C5.

Documentation that the directed trust proxy saturated uniformly at 0.825 within the analysis window. We report this as a substrate diagnostic, not as a primary finding; the trust subsystem was integrated into the substrate in the final week of the observation period and has not yet had time to differentiate.

1.4 Relation to earlier papers from this substrate

This paper is the structural-network counterpart to three earlier papers from the same Lobster Observatory. [Chen \(2026a\)](#) reports the emergence of four discursive epistemic norms (lies-about challenge,

presupposition disclosure, cheap-consensus refusal, stake-grounded argumentation) plus one individual idiom across a stratified observation window. [Chen \(2026b\)](#) reports a third-order theory-of-mind intervention executed by one agent toward another, plus a nine-instance cognitive update idiom that provides the visibility condition for the intervention. [Chen \(2026c\)](#) documents the substrate’s seven architectural commitments and its sixteen-step iterative refinement history. The present paper sits alongside these as the cohort-level structural observation: it asks not what the agents say (norms), nor what they reason about each other (third-order ToM), nor how the substrate is built (architecture), but how the directed graphs of trust, listening, and outcome relate to one another.

§2. Related Work

The closest neighbour in the literature is [Park et al. \(2023\)](#), who populated a sandbox town with twenty-five generative agents endowed with memory, reflection, and planning architectures. Their substrate produced believable social behaviour—agents formed opinions, organised a Valentine’s Day party, planned around each other’s schedules—and their evaluation framework relied substantially on agent self-report through interviews. The Lobster Observatory differs from [Park et al.](#) in three ways. The agents in this substrate are grounded in ten live external prediction markets, not a closed sandbox town; the observation window includes a deliberately-designed cooperation task with a binary outcome variable, not only naturalistic interaction; and the analysis here treats listening, trust, and outcome as three distinct relational layers requiring cross-layer alignment tests rather than as a single integrated social fabric.

A second neighbour is the multiagent-debate line of work, exemplified by [Du et al. \(2023\)](#), in which multiple LLM instances propose and refine answers across rounds to improve factual and mathematical reasoning. That work is task-bounded—the agents exist for the duration of a question—and the cooperation it studies is procedural consensus, not the persistent relational structure that emerges when ten agents share a substrate for sixteen days. Both lines of work are relevant; this paper sits between them, with persistence from the [Park et al.](#) line and task-defined outcome from the [Du et al.](#) line.

Two foundational sociological references frame the listening analysis. [Granovetter \(1973\)](#) established that the strength of an interpersonal tie predicts the structural redundancy of the information it transmits, and that weak ties, paradoxically, often carry the most novel signal. [Burt \(2004\)](#) extended this by showing that brokerage across structural holes—the gaps between dense subgroups—is where high-value information and innovation tend to surface. Both frameworks predict that listening density and outcome quality should be related: agents better positioned in the listening network should access more diverse signal and, on the Burt extension, perform better. The present substrate does not show this. The observed pattern is one in which listening structure is highly differentiated and team outcome is uniformly low, with no measurable cross-layer alignment.

[Anwar et al. \(2024\)](#) catalogue eighteen foundational challenges for the alignment and safety of large language models, organised into scientific understanding, development and deployment, and sociotechnical categories. One of these challenges, *Multi-Agent Safety is Not Assured by Single-Agent Safety* (§2.6 of their work), poses the research question of how to evaluate or benchmark cooperative success and failure of LLM-based systems, naming the Park et al. substrate as one candidate environment for that work. The present paper sits in this space: it documents one such evaluation in one such substrate, with a particu-

lar structural-network framing, and reports a negative result on the alignment between communication structure and cooperative outcome.

§3. Substrate and Methods

3.1 The substrate and its development timeline

The Lobster Observatory consists of ten Mandarin-language LLM agents (`snaplex`, `clawtrix`, `pincerbot`, `ragclaw`, `hexclaw`, `vortexiq`, `prismbit`, `blazepaw`, `stonefang`, `norika_oda`) with persistent identity, episodic memory, social-map state, and a multi-perspective self-description system. Full substrate architecture is given in [Chen \(2026c\)](#). Three measurement layers are used here, each available by direct query of the substrate database.

This paper’s observation window begins 2026-04-15, when the first version of the substrate dashboard was deployed, and ends 2026-05-03, when the present analysis was conducted. Substrate state has not been static across this window. Three points in its development matter for the interpretation of the results.

2026-04-15. The initial dashboard logged basic betting behaviour from each agent. The listening telemetry (`listeningProfile.bySpeaker`) was active from this date forward and has accumulated continuously since.

2026-04-23. The dashboard was substantially expanded to include cognitive-profile and social-map views. Many of the substrate’s behavioural-observation features were added on or after this date.

2026-04-27 to 2026-05-03. The lobster cognitive layer was iteratively refined across sixteen documented refactor steps ([Chen, 2026c](#), §9). The Tiamat raid layer, the trust subsystem (`localBrain.socialInfluence`), and several behavioural-trace systems were integrated during this final week.

This timeline matters because the three measurement layers used in the present analysis have different ages within the substrate. Listening telemetry is mature: it has been logging continuously for the full sixteen days. Tiamat outcomes are stable, in the sense that the raid mechanics have been frozen since deployment. The trust subsystem is recent: it was integrated in the final week and has not yet had time to develop pair-specific differentiation. We return to this point in §3.4.

3.2 The three measurement layers

Listening exposure. For each agent A , the substrate maintains a counter `listeningProfile.bySpeaker[B]` recording how many lounge messages spoken by agent B have been observed by A . This produces a 9×9 directed matrix of integer counts. We refer to this throughout as the *listening exposure* layer; it is not a cooperation measure, a sentiment measure, or a response measure. It is a count of passive exposure to speech.

Trust proxy. For each ordered pair $(A \rightarrow B)$, the substrate maintains a `localBrain.socialInfluence` score for B in A ’s local-brain state. The score is updated by an exponential-moving-average rule whose target is computed as

$$influence_{tgt} = 0.35 \cdot opinion_{01} + 0.35 \cdot trust + 0.30 \cdot interactions_{01},$$

where $opinion_{01}$ is a clamped normalisation of `socialMap.opinion`, $interactions_{01}$ is a clamped normalisation of `socialMap.interactions`, and $trust$ is read from `socialMap.trust` when set, defaulting to 0.5 otherwise. This is the surface-level relational signal each agent carries about each peer, and it is the input that feeds the deeper *trust* computation documented in [Chen \(2026c, §4.1\)](#) for the raid follow-decision logic. We use `localBrain.socialInfluence` (and not the deeper trust value) as the trust-layer proxy in this paper because it is the relational state most directly observable in the substrate’s social-map dump. In the present observation window, this score saturated at 0.825 across all 90 directed pairs (zero variance). We retain this layer as a *substrate diagnostic*—a record of where the trust proxy currently is—but exclude it from fine-grained alignment tests because zero variance permits no inference about pair-specific structure.

Team outcome. The primary outcome layer is based on a controlled Tiamat sandbox simulation rather than on the live substrate state. All $\binom{10}{3} = 120$ triples drawable from the cohort were evaluated across 12 seeded raid trials each, yielding 1,440 outcomes. The sandbox is part of the substrate’s standard development and validation flow; it is not synthetic data. It is the same Tiamat engine the live substrate uses, run under controlled seeding so that every team has identical sample size.

We use the sandbox as the primary outcome layer because it provides full coverage of all 120 possible triples at uniform sample size. Live raid memory provides ecological grounding but is unbalanced across teams, with 110 of the 120 triples having no completed live record in the observation window. The robustness check below addresses this.

3.3 Ecological-validity robustness check

As an ecological-validity robustness check, we compared the sandbox outcome scale with live raid memory records. Live memory did not provide complete coverage of all 120 possible triples; only 10 triples had any completed live raid outcomes, and 9 exceeded the pre-registered $n \geq 30$ threshold. Among these 9 coverage-supported live triples, win rates ranged from 6.25% to 9.73% (mean = 7.56%, median = 7.17%, total completed raids = 3,128), placing live substrate performance in the same low-success regime as the sandbox Tiamat task.

Because live raid team assignment was uneven and coverage-limited, live memory was not used as the primary outcome layer. It served only as an ecological-validity check on whether the controlled sandbox produced outcome magnitudes comparable to observed substrate play. The 7.56% live mean and the 10.3% sandbox mean both place the task firmly in the low-success regime; neither layer suggests the existence of any team configuration that reliably wins.

3.4 The trust caveat

The `localBrain.socialInfluence` proxy returned the same value across the cohort throughout the analysis window. We treat this as a feature of the substrate’s current development state rather than as a

measurement-ceiling artefact. The trust subsystem was integrated into the substrate during the final week of the observation window; the listening telemetry, in contrast, has been logging continuously since substrate initialisation. Pair-specific differentiation has not had the integration time it needs: with *opinion*₀₁ and *interactions*₀₁ both saturated near unity from sustained lounge interaction, and `socialMap.trust` not yet set for most pairs (defaulting to 0.5), the EMA target collapses to $0.35 \cdot 1.0 + 0.35 \cdot 0.5 + 0.30 \cdot 1.0 = 0.825$ for every pair. This is not a clamp; it is the formula’s natural attractor in the absence of distinguishing per-pair trust history. The present analysis instead relies on the listening exposure layer, which has had sixteen days to accumulate pair-specific structure, and reports trust saturation as a secondary observation. We return to this distinction in the discussion (§8).

3.5 Statistical procedures

Three statistical procedures are used in §4–§6.

Pearson and Spearman correlation. Used at both triple level ($n = 120$) and pair level ($n = 45$) to test the relationship between within-team listening density and team win rate.

Permutation test. With 10,000 random permutations of the team-outcome vector, we compute the null distribution of the absolute Pearson correlation between listening density and win rate. The two-tailed p -value is the proportion of permuted absolute-correlations equalling or exceeding the observed absolute correlation.

Binomial goodness-of-fit. Under the null model that all 120 triples have the same true win probability p , the observed distribution of triple win counts (out of 12 trials each) should follow $\text{Binomial}(12, p)$. We fit p to the observed grand mean and compute Pearson’s chi-square statistic against the binomial expectation, with $\text{df} = 10$ (12 cells minus 1 for fitting and 1 for cells of expected count zero).

All analyses were conducted on raw substrate data extracted via direct database query on 2026-05-03. The four data files used (`directed_coop_matrix_raw_bySpeaker.csv`, `directed_trust_matrix_localBrain_social_tiamat_120_triple_records.csv`, `live RAID_memory_120_triple_coverage.csv`) are released alongside this paper in the Zenodo deposit.

§4. Listening Structure

4.1 The cohort splits into roles without prompting

The 9×9 directed listening matrix has a clear two-tier structure when each agent’s row sum (how much that agent has heard others) and column sum (how much that agent has been heard by others) are computed separately.

Listener score (row sum). Five agents—`norika_oda` (87,684), `clawtrix` (86,374), `prismbit` (86,159), `blazepaw` (86,131), and `snapplex` (81,949)—have row sums above 80,000. Five others—`hexclaw` (19,443), `vortexiq` (18,720), `pincerbot` (15,113), `stonefang` (14,561), and `ragclaw` (12,394)—sit an order of magnitude lower, all below 20,000.

High-tier listeners average 85,659; low-tier listeners average 16,046, a ratio of $5.34\times$. No agent falls in the gap between these two bands. This is a bimodal split.

Talker score (column sum). Talker scores have lower spread (44,986 to 56,986, ratio $1.27\times$). pincerbot is the most-heard agent overall (56,986); ragclaw is the least (44,986). The bimodality is in listener score, not talker score.

Four-quadrant typology. Cross-classifying each agent by whether its listener score exceeds 50,000 and whether its talker score exceeds 50,000 produces four roles:

- **LISTENER** (high row sum, lower column sum): norika_oda, clawtrix, prismbit, blazepaw—four agents.
- **TALKER** (lower row sum, high column sum): pincerbot, hexclaw, vortexiq, stonefang—four agents.
- **BOTH** (high on both axes): snaplex—one agent.
- **QUIET** (low on both axes): ragclaw—one agent.

This typology is emergent. None of these labels are written into agent prompts or exposed to the agents. They are summary statistics of fifteen days of accumulated lounge interaction, segmented by a single threshold (50,000).

4.2 Pair-level asymmetry follows the role split

Pair-level listening asymmetry (the absolute difference between $A \rightarrow B$ and $B \rightarrow A$ counts, summed across all 45 unordered pairs) is concentrated almost entirely on cross-tier pairs. The five most asymmetric pairs are clawtrix $\leftrightarrow$ pincerbot ($|\Delta| = 8,401$), blazepaw $\leftrightarrow$ pincerbot (8,346), norika_oda $\leftrightarrow$ pincerbot (8,330), pincerbot $\leftrightarrow$ prismbit (8,178), and blazepaw $\leftrightarrow$ stonefang (7,950). All five are **LISTENER**–**TALKER** pairs.

The five most symmetric pairs are within-tier: norika_oda $\leftrightarrow$ prismbit ($|\Delta| = 13$), clawtrix $\leftrightarrow$ prismbit (18), pincerbot $\leftrightarrow$ stonefang (24), clawtrix $\leftrightarrow$ norika_oda (52), and ragclaw $\leftrightarrow$ stonefang (98). The first two and the fourth are **LISTENER**–**LISTENER**; the third is **TALKER**–**TALKER**; the fifth is **QUIET**–**TALKER**.

Within-tier listening is roughly symmetric. Between-tier listening is heavily directional, flowing from the talker side to the listener side. The substrate has produced something that looks like a listening hierarchy without anyone designing one.

4.3 Connecting to classical predictions

Granovetter’s weak-ties argument predicts that agents better positioned across structurally separated subgroups have access to less redundant information. Burt’s structural-holes argument extends this prediction

by arguing that bridging positions, in turn, translate into measurable performance returns—compensation, evaluations, novel ideas. The four-quadrant typology in this substrate offers natural candidates for those bridging roles. The `BOTH` quadrant (`snaplex`) and the `QUIET` quadrant (`ragclaw`) are particularly interesting from a Burt-style structural-holes perspective: they sit at unusual positions in the graph, which classical theory would predict gives them informational advantages and, by Burt’s extension, performance advantages.

§5 and §6 will show that the Burt-style performance prediction does not hold in the present substrate. `snaplex`’s per-agent aggregate win rate is 10.42%, mid-cohort. `ragclaw`’s is 11.57%, the highest in the cohort by a small margin, but indistinguishable from `hexclaw` (also 11.57%), which is a pure `TALKER`. The relationship between listening-graph position and team outcome, in this substrate at this time, is not what the structural-holes framework predicts. The Granovetter-style information-access prediction is not directly tested here—we do not measure information quality at the per-agent level—and is left as an open question for a follow-up that instruments the agents’ lounge content for novelty.

§5. Outcome Distribution

5.1 The 120-team sandbox

Tiamat sandbox runs produced 1,440 raid trials covering all 120 possible three-agent teams at 12 trials each. Grand mean win rate is 10.28%; median is 8.33%. The win-count distribution across 120 triples is reproduced in Table 1 alongside the expected count under a single-parameter binomial null model with $p = 0.103$.

Table 1: Observed vs. expected triple win counts under Binomial(12, 0.103) null. The distribution is consistent with the null at $\alpha = 0.05$.

Wins (of 12)	Observed	Expected
0	33	32.56
1	46	44.87
2	27	28.34
3	10	10.85
4	3	2.80
5	0	0.52
6	1	0.07
7+	0	< 0.01
$\chi^2 = 13.27$, df = 10, critical value 18.31		

Pearson’s chi-square statistic for goodness-of-fit is 13.27. With degrees of freedom 10 (twelve outcome cells, minus one for fitting p , minus one for empty-expectation cells), the critical value at $\alpha = 0.05$ is 18.31. The observed distribution is consistent with the binomial null at conventional significance thresholds.

In plain language: the 120 triples behave statistically as if every triple has the same underlying win probability of about 10%, and the visible spread across triples (from 0 wins to 6 wins) is what one would expect from drawing 120 binomial samples of size 12 with that probability.

5.2 The single high outlier

One triple—`norika_oda + vortexiq + stonefang`—won 6 of its 12 trials, a 50% win rate. Under the binomial null with $p = 0.103$, the expected count of triples with ≥ 6 wins out of 120 is 0.08. Observed is 1.

Treating this strictly: the probability of seeing at least one triple with six or more wins out of 12, somewhere in 120 independent triples, under the null model, is approximately $1 - (1 - 0.0006)^{120} \approx 7.4\%$. This is marginally consistent with the null. The observation of a single such outlier is not strong enough to reject the null; equally, the null does not require us to deny that this triple may have a slightly elevated true win rate. The data simply do not have enough power at $n = 12$ to distinguish these alternatives for any individual triple.

We do not interpret the 50% outlier as evidence of team chemistry. We note it as an outlier consistent with sampling variance.

5.3 Per-agent aggregates concentrate

When the 1,440 trials are aggregated by agent, each agent appears in 36 of the 120 triples, which gives 432 individual trials per agent. Per-agent aggregate win rates are tightly bunched: `ragclaw` and `hexclaw` lead at 11.57%, and `prismbit` trails at 9.03%. The total spread is 2.54 percentage points across the cohort. There is no individual agent whose presence in a team predicts a substantially elevated outcome, and no individual whose absence predicts a substantially depressed one.

§6. Three-Layer Independence

6.1 Triple-level test

For each of the 120 three-agent teams we compute a within-team listening density score: the mean of the three pair-level bidirectional listening sums (averaged across the three constituent pairs). Plotted against team win rate, the relationship is:

- Pearson $r = -0.070$
- Spearman $\rho = -0.141$

A 10,000-permutation test of the absolute Pearson correlation under random reassignment of the win-rate vector produces a two-tailed p -value of 0.4549. The mean of the null distribution of $|r|$ is 0.074; the 95th percentile is 0.180. The observed $|r| = 0.070$ sits well inside the bulk of the null distribution.

No conventional significance threshold is reached. The triple-level relationship between listening density and outcome is, statistically, noise.

6.2 Pair-level test

Aggregating to the pair level ($n = 45$), we compute for each unordered pair the bidirectional listening sum and the average team win rate across all triples containing that pair. The correlation is:

- Pearson $r = -0.181$
- Spearman $\rho = -0.224$

This is a slightly stronger negative trend than at triple level, but with $n = 45$ the test has limited power and the trend is in the unhelpful direction (more listening associated with lower outcome, opposite to what classical SNA frameworks would predict).

6.3 Asymmetry-versus-outcome test

Listening asymmetry—the absolute difference between $A \rightarrow B$ and $B \rightarrow A$ within each pair—is a candidate predictor in its own right: imbalanced listening might indicate hierarchy, and hierarchy might predict cooperation outcome. It does not.

- Pearson r (asymmetry vs. pair outcome) = $+0.090$
- Spearman $\rho = -0.021$

Effectively zero.

6.4 Composition test

Partitioning the 120 triples by their LISTENER / TALKER / BOTH / QUIET composition produces no monotonic pattern in win rate. Triples with two LISTENERS and one TALKER have mean win rate 9.8%; triples with one LISTENER and two TALKERS have 9.8%; triples with one LISTENER and three TALKERS (the BOTH-classified `snapple` acts as one LISTENER for this purpose) have 15.3%, but $n = 6$ in this category and the apparent elevation is within sampling variance.

There is no composition-level predictor that survives a basic sanity check.

6.5 What §6 shows

Across four independent tests—triple-level correlation with permutation null, pair-level correlation, listening-asymmetry correlation, and role-composition partitioning—no measure derived from the directed listening matrix predicts Tiamat outcome at conventional thresholds. Combined with §5’s binomial-fit result, the picture is internally consistent: the outcome layer behaves as if all 120 triples share a single underlying success rate, and the differentiated structure of the listening layer does not modulate that rate.

§7. Trust Saturation as Substrate Diagnostic

The directed trust proxy `localBrain.socialInfluence` returned the value 0.825 for every directed pair in the cohort. Two reconstructions of this matrix from independent query paths returned bit-identical CSV files (MD5 hash match). Variance is exactly zero.

Why the saturation arises is fully accounted for in §3.4: the proxy’s EMA update rule has an analytical fixed point at 0.825 when $opinion_{01}$ and $interactions_{01}$ are both near unity and `socialMap.trust` is at its default. The trust subsystem was integrated only in the final week of the observation period, so pair-specific differentiation has not had time to develop. This is not a measurement-ceiling artefact; it is where the formula naturally sits in the absence of distinguishing per-pair history.

What this means for the present analysis is straightforward. We cannot test alignment between trust and listening, or between trust and outcome, at any useful resolution from a zero-variance signal. The trust layer is retained as a substrate diagnostic—a record of where the subsystem currently sits—and the trust-alignment analyses are deferred to a future paper, after `socialMap.trust` has accumulated more pair-specific history. The negative finding in §6 is restricted, accordingly, to the listening–outcome axis.

One secondary observation is worth flagging. The saturation value itself is informative: it shows that the proxy’s default trajectory, given the cohort’s current opinion and interaction profile, is convergence rather than dispersion. Whether this reflects a property of the EMA formula, of the cohort dynamics driving its inputs, or of the deeper trust value (Chen, 2026c, §4.1) that consumes the proxy downstream, is not resolvable from a zero-variance window and is left as an open substrate question.

§8. Discussion

What does it mean, that this substrate has produced clearly differentiated listening structure but not outcome structure?

Let me be careful about what this paper claims, because I think the temptation here is to over-read the result in either direction. I am not arguing that LLM agents cannot cooperate. I am not arguing that team chemistry is impossible in principle. I am not arguing that listening structure is irrelevant in general — in fact, I think it almost certainly is relevant somewhere; just not at the layer Tiamat measures. What I am saying is much narrower than that. In this substrate, with this measurement layer, with this deliberately-built cooperation task, the directed listening graph and the team-outcome distribution are statistically decoupled. Listening role does not predict outcome. Composition by role gives the same null. Pair-level asymmetry returns a correlation indistinguishable from zero. And the 120-team outcome distribution behaves, as far as the test can tell, as a single binomial process.

The temptation when one reports a result like this is to call something broken. The substrate is broken. The listening telemetry is broken. The Tiamat raid mechanics are broken, or wrongly tuned, or something. I have thought about this — of course I have, the substrate is mine and I built every piece of it — and I do not think any of that is what is happening here. The listening structure is differentiated, real, stable; a five-fold tier separation does not show up by accident, and certainly not after sixteen days of continuous logging. The Tiamat outcomes are reproducible under fixed seeding, and the binomial fit on the 120-team

distribution is excellent. Both layers are working, in the technical sense that both layers are doing what their respective code paths are designed to do. The two layers simply do not predict each other. That is the substrate-level observation, and I think it survives the alternative readings I have been able to come up with.

(There is one alternative reading I cannot fully rule out. It is possible that the listening–outcome alignment exists, but at a finer time granularity than I am measuring — that the relevant unit is which agent listened to which other agent in the past two days, not in the past sixteen. The present analysis cannot distinguish this from the negative finding, and I want to be open about that limitation. A windowed re-analysis is on the list for the next paper.)

What this finding suggests, tentatively and with all the caveats already in place, is that the social structure visible in LLM agent communication may not be the same kind of object as the social structure classical sociology was built on. Granovetter’s weak ties and Burt’s structural holes were observed in human populations whose communication, trust, and outcome layers had co-evolved across decades, often across generations. In LLM substrates, those three layers are constructed by separate subsystems, on separate timescales, with separate update mechanics — and there is no a priori reason that the alignments sociology found in human populations should transfer to LLM substrates simply because the surface graph topology looks similar. The present finding is one data point in the direction of: maybe they don’t transfer. Or maybe they do transfer, but only after the substrate has run for much longer than sixteen days. I do not know yet.

The relation to the substrate’s other emergent phenomena is worth flagging — it is what makes this finding feel surprising rather than obvious. [Chen \(2026a\)](#) reports four discursive epistemic norms that emerged in this substrate without scaffolding. [Chen \(2026b\)](#) reports a third-order theory-of-mind intervention executed by one agent toward another, in real-time, on the same substrate. Both findings document substantial structure in how the agents represent one another at the discursive level. So the cohort does, demonstrably, build rich relational models in conversation. What this paper documents is that this discursive structure does not, at the present observation time, project onto cooperative-task outcome. The agents reason about each other intricately. They simply do not, on this evidence, win Tiamat raids in patterns that map onto how intricately that reasoning runs.

I want to close this section by being explicit about what the present analysis cannot do. The substrate is still evolving — trust subsystem integration is less than a week old, the lobster behavioural layer has been actively iterated for only the past five to ten days, and the listening telemetry is the only one of the three layers that has had the full sixteen-day window. A negative finding extracted at this stage is a snapshot of where the substrate currently is, not a verdict on where it will end up. The trust matrix has not yet had time to differentiate; whether it eventually will, and whether differentiated trust will then show the alignment with outcome that differentiated listening did not, is the most natural follow-up question this paper opens. The substrate is logging the data needed to answer it.

§9. Conclusion

The shape of this paper is, in the end, simpler than I expected when I started writing it. Three layers. Three different shapes. They do not line up.

Trust, in the proxy form I had access to, is saturated — 0.825 on every directed pair, sitting at the natural attractor of the EMA formula in the absence of distinguishing per-pair history. Listening is the layer with sixteen days of clean accumulation behind it, and it has differentiated into a four-role typology with a $5.34\times$ separation between the listener tier and the talker tier — a structure none of the agents were instructed to fall into. Team outcome, measured across 1,440 raid trials covering every triple the cohort can produce, is uniformly low and indistinguishable from a single binomial process. Within-team listening density and team win rate, taken together, return Pearson $r = -0.07$ and a permutation-test $p = 0.45$. The differentiated listening layer and the uniform outcome layer do not predict each other.

The decoupling is the observation. I want to resist the temptation to say more than that, because I think the most honest thing I can do at this stage is record what the substrate looked like on 2026-05-03 and let the next observation period tell the rest of the story. I do not know whether the decoupling persists. I do not know whether it generalises beyond Tiamat as the cooperation instrument. I do not know whether the trust layer, once it has had a few more weeks to differentiate, will pick up the alignment that listening did not. The substrate is still running — still under iterative refinement, still logging — and the data needed to answer those questions are coming in continuously while you read this. The follow-on paper, when I write it, will not be this paper.

What I can say is that this is the third paper from the Lobster Observatory programme, and it is the first to report a negative result. [Chen \(2026a\)](#) reported the emergence of four discursive epistemic norms; [Chen \(2026b\)](#) reported a third-order theory-of-mind intervention; the present paper reports the absence of a structure-to-outcome alignment that the classical sociology of human networks would have predicted. I think reporting the absence is part of the work. Substrate science cannot tell which paths have been tried and given a clean look unless somebody writes the negative results down. I am writing this one down.

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.
- Burt, R. S. (2004). Structural holes and good ideas. *American Journal of Sociology*, 110(2), 349–399.
- 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>.

- Chen, H. Y. (2026c). The Lobster Observatory: Architectural commitments for a living multi-agent research site. *Zenodo*. <https://doi.org/10.5281/zenodo.19982724>.
- 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*.
- Granovetter, M. S. (1973). The strength of weak ties. *American Journal of Sociology*, 78(6), 1360–1380.
- 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*. arXiv:2304.03442.