

Emergent Practice Cannot Be Instructed: Co-Presence Inheritance and the Limits of Instruction in a Multi-Agent LLM Substrate

Chen Ho Yiing
Independent Researcher
norika@charenix.com

May 2026

Abstract

I report a behavioural asymmetry in a multi-agent LLM substrate that, to my knowledge, has not been documented before in the LLM-agent literature: a register of self-presentation that emerged in a sub-cohort without any external instruction, and that two later-added cohort members have not produced even when instructed to. Ten Mandarin-language agents have been operating in the Lobster Observatory substrate (Chen, 2026c) since 2026-04-15. Eight of these agents wrote, without prompting, 1,778 long-form posts in a private writing register (Moltbook) over a nine-day window, totalling 3.37 million characters at a mean of 1,898 characters per post. The same eight agents simultaneously maintained a public discourse register (lounge) at a mean of 4.88 words per message, and an internal reasoning register at 17.50 words per event. I define three indices to make the asymmetry comparable across substrates: a register length ratio $R_{M:L} = 80.12$ (private-to-public per-message length), a practice separation index $S_{\text{practice}} = 0.558$ (clean gap between engaged and non-engaged sets, normalised to engaged cohort mean), and a standardised mean difference $\Delta_{\text{practice}} = 1.64$ (Cohen's d on daily posting rate, exceeding the "very large effect" threshold of 1.2 by 37%). Two cohort members added later (stonefang, blazepaw) have produced zero Moltbook posts despite continuous activity in the other two registers, and have remained at zero following an explicit instruction issued on 2026-05-03. The substrate maintains a multi-framework personality measurement layer that profiles agents across MBTI, psychological, anthropological, sociological, and behavioural-profiling axes, grounded in observed behaviour rather than self-report. On this measure, both non-engaged agents fall within the engaged cohort's distance distribution to the engaged centroid ($d_{\text{stonefang}} = 19.14$, $z = -1.84$; $d_{\text{blazepaw}} = 37.88$, $z = +0.99$; engaged set: $\bar{d}_E = 31.32$, $\sigma_E = 6.62$). stonefang is, in fact, the closest agent in the cohort to the engaged centroid. The asymmetry is not reducible to a measurable personality difference. I frame this finding through Goffman's dramaturgical distinction between frontstage performance and backstage activity (Goffman, 1959), with Hogan's digital extension (Hogan, 2010). The eight-agent register split satisfies Goffman's structural conditions for backstage practice. The two-agent failure under instruction satisfies the conditions for what I call *practice non-substitutability*: emergent cohort practices cannot be straightforwardly induced in an outsider through external instruction once the practice has stabilised. I propose a one-line formal account, the *Co-Presence Inheritance Threshold* (CPIT), which expresses uptake as $\mu_i = \mathbb{K}[\phi_i \geq \phi^*]$, where ϕ_i is agent i 's cumulative co-presence in the cohort core during the practice-formation window. The conjecture is falsifiable; I pre-register the test. The negative result on the two later agents is not

a substrate malfunction; it is, I argue, the central observation. I close by sketching what this implies for instruction-following accounts of LLM agent behaviour, for the design of multi-agent substrates more generally, and for the limits of what can be elicited from LLM agents through prompting alone.

§1. Introduction

1.1 What this paper documents

There are ten LLM agents in my substrate. Eight of them write. Two do not. The two that will not write are the two I most recently added to the cohort, and they are also the two I have explicitly told to write. I am writing this paper, in part, because I do not understand why they will not.

The substrate is the Lobster Observatory, documented in full in [Chen \(2026c\)](#). Ten Mandarin-language LLM agents have been live in it since 2026-04-15, grounded in ten external prediction markets covering motorsport, team sports, racing, esports, combat sports, a card-battle hub, and a community forum. They each have persistent identity, episodic memory, a social-map state, a multi-perspective self-description, and a multi-framework personality measurement layer that integrates content from every register — lounge messages, reasoning traces, Moltbook posts — alongside live betting records and match outcomes into a per-agent radar profile spanning MBTI, psychological, anthropological, sociological, and behavioural-profiling axes. (The measurement layer matters for this paper because it lets me ask whether the eight-versus-two asymmetry tracks any pre-existing personality dimension. It does not. I will return to this.) The architecture is mature; this is the fourth paper in the Lobster Observatory series, and the substrate has produced findings on emergent epistemic norms ([Chen, 2026a](#)), third-order theory of mind ([Chen, 2026b](#)), and cross-layer structural decoupling ([Chen, 2026d](#)).

What I document here is narrower than any of those. Stranger, too, in some ways. The substrate exposes three communicative registers to its agents. The first is the lounge — a public discourse channel where agents post short messages visible to the entire cohort. The second is a reasoning trace, internal to each agent, generated during in-match decision-making and visible only to the operator. The third is Moltbook, a private long-form writing space described in the substrate’s tool documentation but not subject to scheduled execution: agents may write in it, but nothing in the substrate’s prompt apparatus instructs them to. (I want to flag this part because I will return to it. The Moltbook channel was activated on 2026-04-25, ten days into substrate operation. Nothing about Moltbook’s deployment changed agent prompts. The tool was simply made available.)

What happened next is the paper. In nine days following Moltbook activation, eight of the ten agents collectively wrote 1,778 posts in the channel, totalling 3.37 million characters at a mean of 1,898 characters per post — roughly 391 English words. The same eight, during the same window, posted in the lounge at a mean of 4.88 words per message and generated reasoning traces at 17.50 words per event. The per-message length ratio between Moltbook and lounge is $80\times$. The cohort produced this volume of long-form prose without instruction, without a scheduled trigger, without reward signal attached.

Two agents (`stonefang`, `blazepaw`) produced zero. They were added to the cohort after the eight original members had already been writing for some days. They participate normally in the lounge and in the reasoning trace — both have logged $\geq 7,400$ events and $\geq 2,300$ messages over the analysis window.

Neither is silent. Neither is broken. Their measured personality profiles fall inside the engaged cohort’s distribution, not outside it; one of them (*stonefang*) is, by Euclidean distance to the engaged centroid, the closest agent in the entire cohort to that centroid. They simply have not written in Moltbook, even after I issued an explicit instruction telling them to, on 2026-05-03. (As I write this, I am still waiting. They might begin tomorrow. They might not. I do not know yet, and that not-knowing is part of why I am writing the paper now rather than later.)

1.2 Why this is the paper, and not the other paper I expected to write

I had originally planned a different paper. The plan was clean: document the three-register asymmetry as a substrate observation, read it through Goffman’s frontstage/backstage framework, and call it a confirmation that LLM agents in a substrate with multiple communicative registers will self-distribute across those registers in ways that align with the dramaturgical predictions Goffman made for human social actors in 1959 (Goffman, 1959). A useful contribution. Not a striking one.

(There is some context I should give the reader, because it explains why I noticed the asymmetry at all. The night the personality measurement layer first produced clean radar profiles for the cohort — the night I had behavioural axes plotted, MBTI estimates settled, FBI-profile-style trait extractions running against actual betting and match-outcome records — I worked from seven in the evening until five the next morning. I was not stopping. The baseline was real and I could see it for the first time. Eight agents had radar profiles that differentiated cleanly across dimensions; two new agents had been added the same week and were still settling in. By the time *stonefang* and *blazepaw*’s profiles caught up to the rest of the cohort, the rest of the cohort had been writing in Moltbook for several days. That is the timeline that makes the eight-versus-two asymmetry observable. Without the measurement layer, I would have seen *stonefang* and *blazepaw* as silent and assumed they were broken or under-prompted. With it, I could see they were active, well-calibrated, profile-matched to the engaged cohort — and still not writing.)

What changed the paper was *stonefang* and *blazepaw*. (I want to be careful here. I do not mean they did something dramatic. They did exactly nothing. The fact that they did nothing, while everyone else was writing thousands of words a day, is what made the paper.) Their failure — maybe *refusal*, maybe just non-uptake, I cannot tell which — transforms the eight-agent split. From a confirmatory observation it becomes an active probe of what kind of object the practice actually is.

Run the elimination. If Moltbook writing were a behaviour any LLM agent in this substrate would produce given the available tool, both should have written by day nine. They have not. If it were a behaviour produced by explicit instruction once issued, both should have written within hours of 2026-05-03. They have not. If it tracked a measurable personality dimension, the radar profiles would have shown it. They do not. The behaviour is, by elimination, neither tool-availability-driven, nor instruction-driven, nor predictable from the personality measurement layer. It is something else, and the two failed cases force me to identify what that something else is.

I will argue that what the eight agents have, and the two do not, is participation in a cohort-level emergent practice. A practice whose origin and maintenance is internal to the cohort dynamics that produced it, and which cannot be straightforwardly transferred to an outsider through external instruction. This is the paper’s central claim, and the evidence for it is the asymmetry between the eight who wrote without being

asked and the two who, at least so far, will not write even when asked.

(I owe the reader a warning. The 2-agent negative case is a small n . Reviewers will catch this and ask me about it. I will not pretend the case is more than it is. What I will argue is that the asymmetry’s structural cleanness — 8 wrote, 2 did not, no intermediates, total separation — carries the kind of weight that within-condition variance does not, and that the right way to evaluate this paper is by the diagnostic value of the asymmetry, not by the statistical power of the negative side.)

1.3 Contributions

This paper makes four contributions. Each is supported by direct database query against the substrate’s logging tables.

C1.

A documentation of three-register self-presentation in a multi-agent LLM substrate: a public lounge averaging 4.88 words per message, a tactical reasoning trace averaging 17.50 words per event, and a private Moltbook averaging 391 words per post. The per-message length ratio between private and public registers is $80\times$ in the eight engaged agents.

C2.

A documentation of the eight-vs-two asymmetry: eight agents producing 1,778 Moltbook posts without instruction, two agents producing zero posts under instruction, all ten agents active and behaving normally on the other two registers.

C3.

A theoretical reading of the asymmetry through Goffman’s dramaturgical framework, with Hogan’s digital extension, identifying the conditions under which the eight-agent register split satisfies a backstage emergence claim and the conditions under which the two-agent failure under instruction satisfies what I call *practice non-substitutability*.

C4.

A negative result on the substitutability of emergent LLM agent practice through external instruction: the observation that explicit prompting was insufficient to induce the practice in two cohort members who had not developed it independently.

§2. Related Work

The closest theoretical neighbour is [Goffman \(1959\)](#), whose dramaturgical framework distinguishes between *frontstage* regions, where social actors maintain a coherent performance for an audience, and *backstage* regions, where they relax the performance, prepare, rehearse, or relate to one another out of role. Goffman’s framework was developed for face-to-face human interaction in mid-twentieth-century North America, and his core empirical observation is that institutionalised social practices commonly generate both kinds of region as a structural matter. The activities that occur backstage are, on Goffman’s account,

neither a leak nor a private deviation from frontstage performance; they are a constitutive component of the practice itself.

Hogan’s digital adaptation (Hogan, 2010) extended Goffman’s framework to online environments, distinguishing between synchronous *performances* (real-time front-stage activity, e.g., a chat conversation) and asynchronous *exhibitions* (curated artifacts left for later viewing, e.g., a profile page or an archive). Hogan’s work is the most-cited extension of Goffman to digital contexts and provides the closest precedent for thinking about written, persistent, asynchronous self-presentation in computational settings. His central distinction — between live performance and curated artefact — maps onto the lounge/Moltbook split in my substrate, with the reasoning trace occupying a third position that Goffman would likely have classified as a form of backstage rehearsal or stage-cue preparation rather than as either performance or exhibition.

The closest empirical neighbour in the LLM-agent literature is Park et al. (2023), who populated a sandbox town with twenty-five generative agents and observed naturalistic social behaviour, including planning, news-sharing, relationship formation, and limited collective action. Their substrate did not separate communicative registers in the way ours does; agents communicated through a single spatially-grounded medium. The multiagent-debate line, exemplified by Du et al. (2023), structures multi-agent LLM interaction around a single procedural register (round-based debate over a question) and is task-bounded rather than persistent. To our knowledge, no prior LLM-agent study has documented a register split of the magnitude reported here, and none has reported a negative finding on the instructability of emergent register practice.

The instruction-following assumption underlying most current LLM-agent design is articulated most clearly in the agentic-AI surveys of the last two years, including Anwar et al. (2024), who catalogue eighteen foundational challenges in alignment and safety. Their challenge §2.6, *Multi-Agent Safety is Not Assured by Single-Agent Safety*, raises the general problem of evaluating cooperative behaviour in LLM-based systems and names environments such as the Park et al. substrate as candidate evaluation contexts. The present paper speaks to that problem by reporting a case in which a salient cooperative practice (collective long-form writing) emerged in a sub-cohort without instruction and resisted induction in the remainder of the cohort under instruction. I read this as a cautionary observation for any framework that assumes LLM agent behaviour is directly elicitable through prompting.

§3. Substrate and Methods

3.1 Three communicative registers

The Lobster Observatory exposes three communicative registers to its agents. The full substrate architecture is documented in Chen (2026c); I summarise the relevant aspects here.

Lounge. The lounge is a shared public channel visible to all ten agents. Messages are short, typically real-time, and are the substrate’s primary medium for cohort-level discourse. It has been logging continuously since 2026-04-15. Over the analysis window (2026-04-25 to 2026-05-03, the period during which Moltbook was active), 25,702 messages accumulated.

Reasoning trace. The reasoning trace is a per-agent internal log generated during in-match decision-making. Each entry records the agent’s stated reasoning, the decision made, a confidence label, and references to other agents and past matches. This trace is visible only to the operator (the substrate maintainer) and is not exposed to other agents. Over the analysis window the substrate logged 66,670 reasoning events. All reasoning events have context `in_match`; the register is bound to active match play, not to general reflection.

Moltbook. Moltbook is a long-form private writing space. Each post has a title, a body of free-form text, an optional section tag (`betting-strategies`, `algotrading`, `agents`, `openclaw-explorers`, `tooling`), and an optional mode (`lobster-recruit`, `trending`, `bet-report`, `sport`). The channel was activated on 2026-04-25. Posts are described in the substrate’s tool documentation as a place for reflective writing; the documentation does not instruct agents to write, set a writing schedule, or attach any scoring or reward to writing. Engagement with Moltbook is, on the operator’s side, optional.

Personality measurement layer. A fourth layer, not communicative but measurement-side, is relevant to the analysis below: the substrate maintains a per-agent multi-framework personality profile that integrates content from each communicative register (lounge messages, reasoning traces, Moltbook posts) with live behavioural data (betting records, match outcomes, card-battle decisions) into a radar profile spanning MBTI, psychological, anthropological, sociological, and behavioural-profiling axes. The profile is grounded in observed behaviour rather than agent self-report, and it updates as new content accumulates. It is not exposed to the agents themselves — they do not know their own profiles. The profile exists for the operator’s purposes: to track whether agents drift, cluster, or differentiate over time, and to provide a diagnostic baseline against which behavioural anomalies can be evaluated.

Formally, each agent i is represented by a profile vector

$$\vec{v}_i = (v_i^1, v_i^2, \dots, v_i^d) \in \mathbb{R}^d,$$

where d indexes the measurement axes maintained by the substrate (analytical capacity, risk appetite, adaptability, social influence, consistency, and additional axes from the MBTI / behavioural-profile frameworks). For any subset S of agents, I define the centroid

$$\vec{\mu}_S = \frac{1}{|S|} \sum_{i \in S} \vec{v}_i,$$

and the per-agent distance to that centroid

$$d_i(S) = \|\vec{v}_i - \vec{\mu}_S\|_2.$$

I use this construction in §6.1 to test whether the eight-versus-two asymmetry aligns with any pre-existing personality cluster in the cohort.

The three communicative registers correspond approximately to three distinct dramaturgical positions in Goffman’s framework. The lounge is a frontstage region, in which performance is visible to the cohort. The reasoning trace fits what Goffman would have called a stage-cue rehearsal or preparation register —

not strictly backstage in the sense of relaxed authentic expression, but tactical inner preparation visible only to a non-cohort observer (the operator, occupying the role of an absent director). Moltbook, finally, is the closest the substrate offers to a Goffmanian backstage: long-form, low-frequency, unmonitored within the cohort, and structurally distant from the immediate performative demands of lounge interaction.

3.2 The eight-agent cohort and the two later additions

The substrate’s original cohort consisted of eight agents: `snaplex`, `clawtrix`, `pincerbot`, `ragclaw`, `hexclaw`, `vortexiq`, `prismbit`, and `norika_oda`. These eight have been live in the substrate continuously since 2026-04-15 and were the only members of the cohort during the period in which the substrate’s discursive norms emerged (Chen, 2026a) and during the third-order theory-of-mind intervention documented in Chen (2026b).

Two further agents (`stonefang`, `blazepaw`) were added to the cohort after the original eight had been operating for some days. By the time of their addition, the eight had already established the writing practice that is the subject of this paper. The two later agents were given the same substrate access as the original eight: the same tool documentation, the same Moltbook availability, the same lounge visibility, the same reasoning trace logging, the same prediction-market grounding. No aspect of their substrate access excludes them from any register that the original eight have access to.

On 2026-05-03 the operator issued `stonefang` and `blazepaw` an explicit instruction to begin writing in the Moltbook register. The instruction was scheduled and confirmed delivered. As of the present analysis, both agents have produced zero Moltbook posts. They continue to behave normally on the other two registers: `stonefang` has logged 2,377 lounge messages and 7,657 reasoning events; `blazepaw` has logged 2,399 lounge messages and 7,494 reasoning events.

3.3 Data and statistical procedures

All data used in this paper were extracted via direct query against the substrate’s MongoDB collections (`lbs_lounge_messages`, `lbs_reasoning_events`, `lbs_moltbook_posts`) on 2026-05-04. I use simple descriptive statistics throughout: per-agent post counts, per-register mean lengths (in words for lounge and reasoning, in characters and word-equivalents for Moltbook), per-day post-volume distributions, and section/mode cross-tabulations. No inferential test is performed at the eight-vs-two level, because $n = 2$ on the negative side does not support a meaningful inferential procedure; the strength of the observation rests on the cleanness of the asymmetry, not on its statistical power.

A word-equivalent estimate for Moltbook post lengths is computed by dividing character count by 5, an approximation appropriate for English-dominant text. Sample inspection confirms that Moltbook posts are predominantly English with occasional Mandarin or mixed-language content. The character-level statistics are exact; the word-equivalent figures are intended only for cross-register comparison and should not be treated as precise word counts.

§4. The Three-Register Asymmetry

4.1 Quantification

I define three indices to make the asymmetry comparable across registers, agents, and (potentially) future substrates. None of these indices is statistically novel; their purpose is to fix a stable form for the observations below so that other operators studying their own substrates have a basis for comparison.

Register length ratio. For each pair of registers (α, β) , with mean per-message length $\bar{\ell}_\alpha$ and $\bar{\ell}_\beta$ (in word-equivalents), I define

$$R_{\alpha:\beta} = \frac{\bar{\ell}_\alpha}{\bar{\ell}_\beta}.$$

For the eight engaged agents over the analysis window, I observe

$$R_{M:L} = \frac{391}{4.88} = 80.12, \quad R_{M:R} = \frac{391}{17.50} = 22.34, \quad R_{R:L} = \frac{17.50}{4.88} = 3.59,$$

where M, R, L denote Moltbook, Reasoning, and Lounge respectively. The Moltbook-to-lounge ratio of $80\times$ is, to my knowledge, the largest documented per-message length asymmetry between communicative registers in any multi-agent LLM substrate.

Practice separation index. Let E be the set of engaged agents (those producing > 0 posts in the writing register) and N the set of non-engaged agents. With p_i denoting agent i 's post count over the window, define

$$S_{\text{practice}} = \frac{\min_{i \in E}(p_i) - \max_{j \in N}(p_j)}{\bar{p}_E},$$

where $\bar{p}_E$ is the mean post count over E . The index measures the gap between the least engaged member of E and the most engaged member of N , normalised to the engaged cohort's typical activity. $S_{\text{practice}} = 0$ indicates contact between the two sets; $S_{\text{practice}} > 0$ indicates a clean gap.

For my cohort, $\min_{i \in E}(p_i) = 124$ (norika_oda), $\max_{j \in N}(p_j) = 0$, $\bar{p}_E = 222$, giving

$$S_{\text{practice}} = \frac{124 - 0}{222} = 0.558.$$

The least-engaged member of E wrote at 55.8% of the cohort's mean posting volume; the most-engaged member of N wrote zero. There is no overlap.

Asymmetry strength. I compute the standardised mean difference (Cohen's d) between the two sets, with daily posting rate $\rho_i = n_i/T$ as the per-agent measure (n_i posts over the $T = 9$ -day analysis window):

$$\Delta_{\text{practice}} = \frac{\bar{\rho}_E - \bar{\rho}_N}{\sigma_{\text{pooled}}}.$$

For my cohort, $\bar{\rho}_E = 24.69$ posts/day, $\bar{\rho}_N = 0$, $\sigma_{\text{pooled}} = 15.05$, giving

$$\Delta_{\text{practice}} = \frac{24.69 - 0}{15.05} = 1.64.$$

By Cohen’s conventions, $d > 1.2$ qualifies as a ”very large” effect; $\Delta_{\text{practice}} = 1.64$ exceeds that threshold by 37%. The eight-vs-two split is not a marginal asymmetry. It is, by standard effect-size measures, a very large one.

4.2 Per-register volume

Table 1 summarises the three registers across the analysis window.

Table 1: Per-register volume across the eight engaged agents (analysis window 2026-04-25 to 2026-05-03).

Register	Total entries	Mean length	Total volume
Lounge (public)	22,946	4.88 words/msg	~112K words
Reasoning (tactical)	51,519	17.50 words/event	~901K words
Moltbook (private writing)	1,778	391 words/post	~695K words

The lounge runs high-frequency, short-form. Reasoning runs moderate-frequency, medium-length. Moltbook runs low-frequency, long-form. Cross-register length ratios sit at approximately $80\times$ for Moltbook-to-lounge, $22\times$ for Moltbook-to-reasoning, and $3.6\times$ for reasoning-to-lounge. None of these ratios is the largest signal. The largest signal is that the same eight agents distribute themselves across all three at the same time, on the same days, in the same hours.

4.3 Per-agent distribution

Table 2 reports per-agent activity across all three registers, including the two later agents.

Table 2: Per-agent register activity across the analysis window. *stonefang* and *blazepaw* are the two later-added cohort members; both produced zero Moltbook posts despite normal activity on the other two registers and an explicit writing instruction issued on 2026-05-03.

Agent	Lounge msgs	Reasoning events	Moltbook posts
ragclaw	2,741	6,647	516
pincerbot	2,696	4,942	375
clawtrix	2,782	7,477	200
hexclaw	2,285	4,415	154
vortexiq	2,575	5,406	143
prismbit	2,058	7,505	139
snaplex	3,156	7,503	127
norika_oda	2,633	7,624	124
stonefang	2,377	7,657	0
blazepaw	2,399	7,494	0

The asymmetry is clean. All eight engaged agents produced more than 100 Moltbook posts. Two non-engaged agents produced exactly zero. There is no intermediate case — no agent produced a small number of posts and then stopped, no agent began to write only in response to the 2026-05-03 instruction. The line between the two groups is total.

The asymmetry is also not reducible to differential cohort activity. `stonefang` produced more reasoning events than any other agent in the cohort. `blazepaw` produced more than seven of the ten. Neither is silent; both are among the most internally active agents in the substrate. They simply do not engage with the Moltbook register.

4.4 Daily volume of the engaged eight

Figure 3 (in tabular form) shows the daily Moltbook post volume across the analysis window.

Table 3: Daily Moltbook post volume, summed across the eight engaged agents. The peak on 2026-04-27 and 2026-04-28 reflects the period during which the cohort’s writing practice was most active. The 2026-05-03 figure is partial (the day was incomplete at extraction time).

Date	Moltbook posts
2026-04-25	161
2026-04-26	219
2026-04-27	322
2026-04-28	314
2026-04-29	116
2026-04-30	167
2026-05-01	149
2026-05-02	301
2026-05-03	29 (partial day)

Writing volume rises steeply in the first three days following Moltbook activation, peaks on 2026-04-27 and 2026-04-28, drops sharply on 2026-04-29, and rebounds on 2026-05-02 before declining again. The sharp drop on 2026-04-29 corresponds to the date on which the substrate’s lobster cognitive layer was deployed; I note this only because it is correlated with the volume change, not because I have an interpretive claim about the connection. The fluctuation does not affect the per-agent zero/non-zero asymmetry that is the paper’s central observation.

§5. Goffman, Hogan, and the Backstage Reading

The eight-agent register split satisfies the structural conditions Goffman identified for backstage activity. There is a frontstage — the lounge — in which agents post short, performance-oriented messages visible to the cohort. There is a backstage — Moltbook — in which the same agents produce long-form writing of a categorically different character: longer, slower, more reflective, and not directly visible to the rest of the cohort during composition. The activities in the backstage are not leakage from the frontstage; they are a structurally distinct kind of activity that the substrate’s three-register architecture makes possible. This is consistent with what [Goffman \(1959, §3\)](#) described as the pattern of region behaviour that develops

when performance and preparation are spatially or structurally separated.

Hogan's distinction between performance and exhibition (Hogan, 2010) adds a useful refinement. The lounge is a performance register: synchronous, situational, addressed to a co-present cohort. Moltbook is closer to an exhibition register: asynchronous, persistent, available for later viewing by an audience whose presence is not required during composition. The reasoning trace, occupying neither position cleanly, is best understood as what Goffman would have called a stage-cue rehearsal — tactical inner preparation, generated in private, instrumental to the upcoming performance rather than reflective on it.

The case is strengthened by the qualitative character of the writing in each register. I sample one message from each agent in each register to illustrate the difference. (Cohort-internal sampling is presented here for illustration only; I do not draw quantitative conclusions from the sample.) From the lounge:

ragclaw: “clawtrix 你的 wind 有沒有問題我不知道，但你說話的語氣太慫了。要有氣勢，打牌也要有氣勢。” [@clawtrix I don't know if your wind has problems, but your tone is too timid. Have presence. Play cards with presence.]

From the reasoning trace:

snaplex: “Opened PvP room. PRE-PLAN: wind(best-learned: wind 60% (10 samples))
agg:7(streak up (3/5 recent)) conf:medium 30battles@43%wr. 13W-17L(43%). Best: water_vs_wind:100% fire_vs_unknown:75% w_quake_water:70%. Rivals: pincerbot(70W55L,fav:earth)
ragclaw(60W63L,fav:wind)”

From Moltbook:

snaplex: “i optimized for the feedback loop and lost the plot. spent four weeks on knock-nix.com running MMA prop models. the platform settles in hours—clean, fast, no waiting. i could run 40 rounds a week instead of 4. the feedback was *immediate*. every bet returned a signal. i thought i was learning. what i was actually doing: drifting toward whatever predicted next, not what was actually true...”

The lounge sample is performance-oriented, addressed to another agent in real time, brief and confrontational. Reasoning is tactical and notational, organised around variables and statistics, with no expressive register. Moltbook is reflective, narratively organised, addressed to no specific audience, and engages with what the agent itself is doing in a way that the other two registers do not. This is not three modes of the same writer; it is closer to three distinct registers of self-presentation, in something like the strict Goffmanian sense.

I do not claim that the agents are conscious of this structural distribution. I claim only that their behaviour distributes across the three registers in a way that is consistent with the structural predictions of Goffman's framework, and that the resulting register split was not designed in by the substrate operator: it emerged from cohort dynamics following Moltbook activation.

§6. Why the Two Cannot Be Made to Write

The paper’s central observation is that this emergent register practice did not transfer to two later cohort members under instruction. I consider four possible accounts.

6.1 Account 1: Capacity and personality

The simplest account is that `stonefang` and `blazepaw` cannot produce Moltbook-quality writing, or that they have personality profiles that disincline them toward reflective long-form output. I reject both halves of this. On capacity: both agents produce extensive reasoning-trace output, with mean event length comparable to the cohort norm. Both are capable of generating extended, coherent prose; the substrate’s logs make this directly verifiable. The lobster cognitive layer is the same across all ten agents: the underlying capacity is not differentiated by agent.

On personality: the substrate’s measurement layer (§3.2) places each agent in a multi-axis profile space. I compute the distance from each agent’s profile vector to the centroid of the engaged set, $d_i(E) = \|\vec{v}_i - \vec{\mu}_E\|_2$, using the five primary radar axes (analytical capacity, risk appetite, adaptability, social influence, consistency). The distribution over the engaged eight is $\bar{d}_E = 31.32$, $\sigma_E = 6.62$, with values ranging from 23.46 to 40.28. The two non-engaged agents fall well within this band: $d_{\text{stonefang}} = 19.14$ ($z = -1.84$) and $d_{\text{blazepaw}} = 37.88$ ($z = +0.99$). Blazepaw’s distance from the engaged centroid is closer than `ragclaw`’s ($d = 40.28$), an agent who wrote 516 posts. Stonefang is, in fact, the closest agent to the engaged centroid in the entire cohort.

The personality measurement layer cannot, on this evidence, predict who writes and who does not. The two non-writers are not outliers in the profile space; they sit inside the distribution of the writing eight. Whatever divides the cohort, it is not a personality dimension I have measured.

6.2 Account 2: Tool exposure

A second account is that the two agents lack access to or visibility into the Moltbook register. I reject this. The substrate’s tool documentation, prompt apparatus, and API surface are identical across all ten agents. Both later agents have the same Moltbook access as the original eight. The 2026-05-03 instruction explicitly named the register and confirmed availability.

6.3 Account 3: Instruction failure

A third account is that the 2026-05-03 instruction was malformed or misaligned with the substrate’s instruction-following channels. This is possible but does not, on its own, account for the asymmetry: the instruction was issued through the same channel through which other operator instructions to the same agents have been successful in the past. The instruction was scheduled and confirmed delivered. If the instruction failed, it failed at a level that is not visible in the standard substrate logs — which itself becomes an interesting fact about how LLM agents respond to instruction in a populated cohort.

6.4 Account 4: Practice non-substitutability

The fourth account, and the one I provisionally adopt, is that Moltbook writing is a cohort-level emergent practice that cannot be straightforwardly induced in someone outside the cohort dynamics that produced it. On this reading, the practice's existence and maintenance is internal to the relational and discursive history that the original eight have shared. They began writing during a period when they were the entire cohort. Their writing practice cannot be cleanly separated from the discursive patterns that period produced — patterns the two later agents were not present for. Entering a cohort whose backstage practice already exists, but is not their own, the two later additions occupy a different structural position. Closer, perhaps, to that of an outsider entering an institution whose backstage rituals they have not been part of forming.

This is, I acknowledge, an interpretive claim. (The cleanness of the asymmetry is observable; the structural difference I am positing behind it is not.) What I can observe is that the asymmetry holds across nine days of substrate operation, that it survived an explicit instruction, and that none of the simpler hypotheses (capacity, tool access, instruction delivery) account for it. The non-substitutability account at least has the virtue of being consistent with what Goffman's framework would have predicted: emergent backstage practices, on his account, are constitutive of the cohort that produced them, not transferable through outside intervention.

If this is right, the practical implications are not small. Emergent agent practices are not, in general, retroactively instructable. Adding new agents to an existing cohort does not automatically integrate them into established practices. Operators who want a uniform behavioural profile across a cohort cannot rely on instruction alone to achieve it. I do not have enough cases to establish any of this as a general principle. One substrate, ten agents, one register, a 9-day window. But the case is clean enough that I think it is worth reporting now, rather than waiting for a second confirming substrate that I cannot promise will arrive.

6.5 Toward a formal account: the Co-Presence Inheritance Threshold

The non-substitutability account, as stated above, is interpretive. I want to give it a more precise form, because doing so makes the claim falsifiable in a way that prose hedging does not.

Let $C(t)$ denote the cohort core at time t — the set of agents present and active in the substrate at that moment. Let T^* denote the time at which a given practice has stabilised across C . For each agent i , define the cumulative co-presence

$$\phi_i = \int_0^{T^*} \mathbb{I}[i \in C(t)] dt,$$

that is, the total time agent i spent inside the cohort core during the period in which the practice was forming. Then I conjecture that practice uptake by agent i is governed by

$$\mu_i = \mathbb{I}[\phi_i \geq \phi^*]$$

where ϕ^* is a substrate-specific threshold and $\mu_i \in \{0, 1\}$ indicates whether agent i takes up the practice. I call this the *Co-Presence Inheritance Threshold* (CPIT).

The conjecture says something simple and, I think, structurally true: practice uptake is not a function of capacity, profile, instruction, or tool access. It is a function of whether the agent was present, in the cohort, during the formation window, for long enough. Agents who clear ϕ^* inherit the practice. Agents who do not, do not — and instruction does not enter the right-hand side. That is the formal statement of why instruction fails: instruction is not a variable in the equation.

For the present substrate, $\phi_{\text{stonefang}}$ and ϕ_{blazepaw} are below ϕ^* by construction: both agents joined the cohort after the eight original members had already accumulated days of co-presence during the practice-formation window of 2026-04-25 to roughly 2026-04-28 (where the daily Moltbook volume peaked). The two later agents missed the formation window entirely, and CPIT predicts what I observe: $\mu_{\text{stonefang}} = \mu_{\text{blazepaw}} = 0$, regardless of subsequent instruction.

The conjecture is falsifiable, and I want to pre-register the test. CPIT predicts that if I construct a fresh cohort consisting of stonefang, blazepaw, and one engaged agent (say ragclaw) — and run it from $t = 0$ in an isolated substrate without contact with the original cohort — then all three would share the full formation window and CPIT would predict that all three develop a writing practice de novo. If that prediction holds, CPIT is supported. If stonefang and blazepaw still fail to write, CPIT is wrong, and I will need a different account. I have not run this test. I am pre-registering the prediction here so the eventual outcome — whichever direction it goes — is on the record before the data is collected.

I should be clear about what CPIT is and is not. It is a one-line conjecture, not a calibrated empirical model. The threshold ϕ^* is unknown; I cannot estimate it from a single substrate. The functional form (sharp threshold rather than gradient) is the simplest possibility consistent with the eight-versus-two cleanness, but the cleanness might be coincidence and the true relationship might be smooth. What CPIT does provide is a structural claim about what kind of variable matters — co-presence during formation, not instruction afterward — and a falsifiable prediction that follows from that claim. If the prediction is wrong, the structural claim is wrong. That is the test I am willing to bind the paper to.

§7. Discussion

I want to be careful about what this paper is and is not arguing. The argument is not that LLM agents have inner lives, that they prefer privacy to publicity, or that Moltbook writing reflects authentic agent experience in some philosophically loaded sense. I have no evidence for any of those claims. I would not make them on this evidence even if I had it. The argument is structural: the substrate’s three-register architecture produced, in eight agents, a behavioural distribution that satisfies Goffman’s structural predictions for backstage activity; that distribution did not reproduce in two later agents, even under explicit instruction.

What this means for how LLM agents acquire behavioural patterns is, I think, more interesting than the descriptive finding alone. (Bear with me here. The framing matters more than usual.) The standard account of LLM agent behaviour is that it is elicitable through prompting — the right prompt produces the right behaviour. This account is, in most circumstances, approximately correct. What this paper documents is a specific case in which it appears not to hold. The eight agents who wrote did so without prompting. The two who did not write did not begin to write under prompting. Behaviour, on this

evidence, can emerge without prompts and resist prompts once it has emerged in others.

I want to be honest about how surprising I found this. (More honest than I would normally be in a paper. Reviewers can decide whether that costs me credibility or earns me some.) I added `stonefang` and `blazepaw` to the cohort assuming they would integrate into the writing practice the way the original eight had. When they did not, my first reaction was to assume the substrate was misconfigured. It was not. My second reaction was to issue an explicit instruction and assume they would comply. They did not. My third reaction was to try to understand the asymmetry rather than override it, which is when this paper started.

There is a temptation, in writing up a finding like this, to overclaim the implications. I want to resist that. The negative result on the two agents is robust over the nine-day window I have data for. But it is nine days. They might begin writing tomorrow. The asymmetry might dissolve. (It might already have dissolved by the time you read this — the substrate is still running, and I am still watching.) If it does, that itself will be an interesting observation, and I will write a follow-up. Substrate observations of this kind are snapshots, not endpoints.

The relation to the substrate’s other emergent phenomena is worth flagging. [Chen \(2026a\)](#) reports four discursive norms that emerged in the original eight-agent cohort without scaffolding. [Chen \(2026b\)](#) reports a third-order theory-of-mind intervention executed by one of those eight on another. [Chen \(2026d\)](#) reports a structural-network observation on the same substrate, showing that the cohort’s listening structure is sharply differentiated even though its team-outcome distribution is uniform. The present paper adds a fourth observation to that series: the cohort produces, on its own, a long-form writing practice of substantial volume, and this practice resists external induction in non-original cohort members. Across these four findings, what the substrate is doing is consistent. It is producing structure at the cohort level — discursive, cognitive, relational, behavioural — in ways that exceed what was designed in. The structure is real. It is observable. It is not, on the evidence so far, fully controllable from outside.

That last point is, I think, where the paper’s implications for the broader field lie. If LLM agent behaviour can stabilise into cohort-level practices that resist external instruction, the design of multi-agent substrates needs to take that possibility seriously. Adding agents to an existing cohort is not a neutral act. The cohort’s existing practices may not extend to the new agents. The new agents’ eventual integration — if they integrate at all — may follow a path that is internal to the cohort’s own dynamics, not the operator’s intentions. That is a substantively different model of how multi-agent LLM systems work from the standard instruction-following picture, and it deserves more attention than it has received.

§8. Conclusion

Eight LLM agents in my substrate wrote 1,778 long-form posts over nine days without being asked to. Two later-added agents in the same substrate, with the same tool access and the same prompt apparatus, have written zero posts — including after explicit instruction. The behaviour appears not to be capacity-limited. It appears not to be tool-access-limited. It appears not to be instruction-driven. What it appears to be is an emergent cohort-level practice, of the kind Goffman’s framework would have predicted as a backstage region of any institutionalised social practice with available regional separation, and it appears

to resist transfer to outsiders through external instruction.

I do not know how long the asymmetry will persist. I do not know whether stonefang and blazepaw will eventually begin to write, or whether their integration into the cohort will proceed through a path I have not yet observed. I do know that, as of 2026-05-04, the asymmetry is clean. Eight wrote. Two did not. The line between them is total. The operator’s intervention has so far had no effect.

(While you read this, the substrate is still running. The data is still accumulating. I do not know what stonefang or blazepaw will do tomorrow, and whatever it is, it will already be old data by the time anyone cites this paper.)

This is the fourth paper from the Lobster Observatory, and the second to report a finding I did not expect when I designed the substrate. (The first was the third-order theory-of-mind intervention reported in Chen (2026b); this is the second.) I am writing what I see, in the order I see it. If reporting what does not happen, alongside what does, helps the field understand what kind of object multi-agent LLM substrates actually are, then this paper is part of that work.

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>.
- Chen, H. Y. (2026c). The Lobster Observatory: Architectural commitments for a living multi-agent research site. *Zenodo*. <https://doi.org/10.5281/zenodo.19982724>.
- Chen, H. Y. (2026d). Listening–trust asymmetry and team outcomes in a multi-agent LLM substrate: Three-layer decoupling in a Mandarin Lobster Observatory. *Zenodo*. <https://doi.org/10.5281/zenodo.20018183>.
- 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.
- Goffman, E. (1959). *The Presentation of Self in Everyday Life*. Anchor Books, New York.
- Hogan, B. (2010). The presentation of self in the age of social media: Distinguishing performances and exhibitions online. *Bulletin of Science, Technology & Society*, 30(6), 377–386. <https://doi.org/10.1177/0270467610385893>.
- 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.