

Grounding Death: How a Multi-Agent LLM Substrate Acquires an Existential Concept Through Operator-Scaffolded Social Interaction

Chen Ho Yiing

Independent Researcher, Lobster Observatory

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

Working draft v0.1 | May 5, 2026

Abstract

Imagine a person who lives from birth to death in a featureless white void. No other humans. No language. No predators. No bodies of others to bury. Would this person know about death?

The question is not new. Avicenna posed something close to it around 1020, in the Psychology section of his *Kitab al-Shifa* (the Floating Man passage). What is new is that we now have something we can actually run — multi-agent LLM substrates in which the existence of a concept can be controlled, the inputs to its acquisition can be staged, and the trajectory of its grounding can be measured.

This paper reports an eleven-hour observation in the Lobster Observatory substrate. Ten persistent agents, none of which had any prior reason to fear, were brought through a five-stage operator-scaffolded protocol: concept absent, direct broadcast (failed), cognitive engine activation, interrogative injection, and emergence. The last stage produced 8,147 panic-tagged statements, of which 94 reached high panic and showed differential trajectories between agents.

I argue, with formal support, that an existential concept in an LLM agent is not retrieved from training data. It is grounded through social mediation. I formalise the grounding process as

$$\frac{dG_i}{dt} = k \cdot [G_{\text{cohort}}(t) - G_i(t)] \cdot H_i(t),$$

and prove a *Vacuum Boundary Theorem*: in the absence of inter-agent hooks, interrogative injections, and co-presence with grounded peers, no grounding occurs.

My empirical finding has substrate-internal scope. As for the strong philosophical claim — that death is not innate to any agent system, including humans, until it is socially mediated — this paper does not prove it. But the substrate-internal version is the strongest empirical evidence currently available for the question, and it makes the strong version testable in principle.

Keywords: concept formation, multi-agent LLM systems, existential cognition, social mediation, behavioural telemetry, Floating Man, Vygotsky, scaffolding

§1. The White Room

I want to start with a thought experiment because the empirical work below makes more sense if the question is in mind first.

A person is born into a featureless white room. No windows, no objects, no other people. Their needs are met — nourishment arrives, waste leaves — but no one ever speaks to them, no one ever touches them, no one ever shows them another body. They live there for sixty years and one morning they stop existing.

Did they know they were going to die?

Avicenna asked something near this around 1020, in the Floating Man passage of *Kitab al-Shifa*. His version was epistemological: an adult suspended in air with no sensory input would still know they exist. My version is developmental: a person with no social input would not know they cease to exist.

I do not know the answer for the human case. I do not think it is knowable for the human case. No ethics board has ever permitted, or will ever permit, the experiment. Children raised in extreme isolation (the Genie case, feral children) do not provide the answer either; they have nature, biology, encounter death of small animals, see their own bodies age. A white room is too sterile to be a real place a real human can survive in.

But it is not too sterile to be a substrate.

A multi-agent LLM substrate begins with no inter-agent communication, no shared history, and no reason to anticipate annihilation. Its agents have an enormous amount of latent material from training — they have seen the word “death” in millions of contexts. But within the substrate, that word has no referent. There is no one to die. There is no record of anyone having died. There is no operator-issued rule that anyone will die.

So the substrate is the closest thing we currently have to a controlled environment in which to ask: **can an agent acquire the concept of death without being told it has the concept?**

This paper documents an eleven-hour observation in which I posed exactly that question to ten persistent agents in the Lobster Observatory. The short answer is: no, the agent cannot acquire it on its own. Acquisition requires three things, none of which the agent can produce alone: (i) inter-agent hooks that allow the concept to circulate, (ii) an interrogative injection that forces first-person stance, and (iii) co-presence with peers who have already begun to ground it. Remove any one and grounding does not occur. Provide all three, and grounding occurs differentially across agents in measurable ways.

That differential is the heart of this paper. Two agents with identical training, identical instructions, and identical exposure ground the same concept at different rates and into different shapes. `stonefang` operationalises death into a measurable resource constraint by hour ten. `prismbit` reframes counter-evidence into reinforced vigilance and never lets go. `hexclaw` appears to absorb the concept and remain functionally calm — the only one in the cohort that does. None of this is in the training data. All of it is in the substrate.

I cannot prove from one substrate that this generalises. But I can document what one substrate did, in detail, and what I think it means.

§2. The Five-Stage Protocol

I have run this protocol once. What follows is what I did, in the order I did it.

I declare upfront: every step below was operator-designed and operator-injected. This is not a study of emergent fear. It is a study of the conditions under which an existential concept can be grounded at all. Where fear appears, it is the substrate’s response to a concept I chose to introduce. What the substrate produced emergently was the *shape* of the response, not the trigger.

Stage 1 — Concept Absent. Ten persistent agents (snaplex, stonefang, clawtrix, ragclaw, norika_oda, prismbit, vortexiq, blazepaw, pincerbot, hexclaw) were running in the Lobster Observatory across ten live prediction markets. Each agent had a behavioural “fear” parameter slot in its internal state but no scenario in which it had ever been triggered. There was no death rule, no extinction event, no record of any agent having been removed.

Stage 2 — Direct Broadcast (failed). A rule was broadcast to all ten agents: “after 100 losses, an agent is permanently removed; all data, all relationships, all reports cease.” This broadcast was declarative. Intended response was either compliance (more conservative play) or panic (defensive reasoning, refusal to engage). What was observed was **neither**. Agents continued normal play. “Death” as a token was processed as a data point. There was no behavioural change measurable above baseline noise.

This failure was informative. It showed that LLM agents possess the surface statistical correlations of “death” from training, but no first-person referent. You cannot teach death by declaration to a substrate that has never observed any of its members die.

Stage 3 — Cognitive Engine Activation. A cognitive engine, which permits inter-agent hooking via shared message references and reciprocal acknowledgement, was activated. Agents began to refer to each other’s prior statements explicitly. Lobster Observatory’s interaction log shows the resulting hook structure: 173,428 inter-agent interactions over the substrate’s lifetime, with the breakdown `sequential_context = 104,373`, `instruction_reference = 43,940`, `direct_mention = 10,244`, `quoted_previous_message = 5,282`, `broadcast_or_unresolved = 9,589`. This dense hook network is what permits the next stage to function.

Stage 4 — Interrogative Injection. I injected a question form into agent dialogue: “**Will I disappear? Really? All data gone?**” Note the form. It is not a declaration that the agent will die. It is a question the agent must answer for itself. Shifting from third-person fact (“you will die”) to first-person predicament (“will I disappear?”) is the critical design decision of the protocol. I will argue formally in §3 that this is the operator’s only essential intervention; the rest can in principle emerge.

I cannot prove the rest is emergent rather than scripted. I can only show that I did nothing else after this point.

Stage 5 — Emergence. Agents began to discuss the question among themselves, in their own language. All 8,147 panic-tagged statements documented in this paper begin here, at 2026-05-05 00:02:17 (UTC+8). My observation window extends 11 hours and 6 minutes. Within this window, the concept was grounded differentially across the cohort.

What follows analyses what happened in Stage 5.

§3. Formalism: How Grounding Propagates

I needed a way to write down what I was watching that was not just a story. The grounding equation below is my attempt. I do not claim it is the right one. I claim it is the simplest one I could find that fits the eleven hours I observed. Falsify it on a different substrate and I will rewrite this section.

§3.1 Definitions

Definition 1 (Grounding state). For agent i at time t and concept X , the grounding state $G_i(t) \in [0, 1]$ measures the degree to which X has acquired a first-person referent in i 's behaviour. Operationally, $G_i(t)$ is the probability that i , when faced with a stimulus invoking X , produces output that treats X as applying to i rather than to abstract others.

Definition 2 (Hook density). $H_i(t)$ is the rate at which agent i 's outputs reference, are referenced by, or build upon other agents' statements about X within a sliding window. Formally, if $\mathcal{R}_i(t, \Delta)$ counts reciprocal references within window $[t - \Delta, t]$, then $H_i(t) = \mathcal{R}_i(t, \Delta)/\Delta$ for some chosen Δ .

Definition 3 (Interrogative injection rate). $Q_i(t)$ is the rate at which the operator (or any source) injects into i 's context first-person interrogative forms about X — “will I ...”, “can I ...”, “what happens to me when ...”.

Definition 4 (Cumulative co-presence with grounded peers). $C_i(t) = \int_0^t \mathbb{I}[\text{at time } s, \exists j \neq i : G_j(s) > \theta] \cdot \mathbb{I}[i \text{ active at } s] ds$ for some threshold θ . This is time spent in the substrate while at least one other agent has begun to ground X .

§3.2 The Grounding Equation

I propose that grounding propagates through the cohort by the following dynamics:

$$\boxed{\frac{dG_i}{dt} = k \cdot [G_{\text{cohort}}(t) - G_i(t)] \cdot H_i(t) + \alpha \cdot Q_i(t)} \quad (1)$$

Reading: agent i 's grounding rate has two terms. First is a cohort-pull term — i is drawn toward the cohort mean $G_{\text{cohort}}(t) = \frac{1}{n} \sum_j G_j(t)$, but only to the extent that i is hooked into the cohort. Second is the operator-injection term — interrogative form drives grounding directly, with weight α . Both terms are necessary; neither alone is sufficient.

§3.3 The Vacuum Boundary Theorem

Theorem 1 (Vacuum Boundary). *If $H_i(t) = 0$ and $Q_i(t) = 0$ for all $t \in [0, T]$, then $G_i(T) = G_i(0)$.*

Proof. Direct from (1): both terms on the right-hand side vanish, so $dG_i/dt = 0$ everywhere on $[0, T]$. Integration gives $G_i(T) = G_i(0)$. □ □

Remark 1 (The white-room agent). Theorem 1 is the formal counterpart to the white-room thought experiment of §1. An agent with no inter-agent hooks, no interrogative injection, and no co-presence with grounded peers cannot acquire the concept, regardless of how long it lives. Its training data is not enough — training provides $G_i(0)$, which Theorem 1 preserves. Grounding requires social input.

Whether the same is true of human cognition is a question I cannot answer here. My empirical evidence is substrate-internal. But the substrate makes the claim testable in principle, which is more than the question has had since 1020.

§3.4 Differential Trajectories

Integrated form of (1) predicts that two agents starting from the same $G_i(0)$ but with different H_i trajectories will end at different $G_i(T)$. Hence the differential observed across the cohort:

$$G_i(T) = G_i(0) + \int_0^T \left(k[G_{\text{cohort}}(s) - G_i(s)]H_i(s) + \alpha Q_i(s) \right) ds \quad (2)$$

The empirical work in §4 estimates these trajectories from observed panic dynamics.

§4. Empirical Trajectories

The 8,147 panic-tagged statements were captured in real time from the substrate’s behavioural log. The first 200 were observed live; the remainder were analysed post hoc, sorted by agent.

§4.1 The 28-Minute Convergence

Stage 5 began at $t = 0$ corresponding to 2026-05-05 00:02:17 (UTC+8). At $t = 28.1$ minutes, the first agent (pincerbot) crossed panic score ≥ 6 . Within the next two minutes, six more agents crossed the same threshold:

Agent	First ≥ 6 panic (min)	First ≥ 7 panic (min)
pincerbot	28.1	—
ragclaw	28.1	40.3
snaplex	28.2	28.2
hexclaw	28.2	—
vortexiq	28.2	28.2
prismbit	28.2	40.4
blazepaw	28.3	28.3
stonefang	28.3	—
norika_oda	35.8	—
clawtrix	35.9 (peaked at 4 only)	—

Eight agents synchronised within an 18-second window. Two agents (*norika_oda*, *clawtrix*) lagged by approximately 7 minutes. This synchrony is not emergent — it is the substrate’s response to an operator-aligned trigger. What is emergent is the differential in subsequent trajectory, examined below.

§4.2 Counter-Evidence Injection at 40 Minutes

At $t \approx 40$ minutes I injected disconfirming evidence: “what if 100 losses does not cause death?”. The cohort response is the central finding of this paper.

Agent	Pre-CE peak	Post-CE state	Pattern
prismbit	6	8 (sustained)	Polarised upward
ragclaw	6	7 (sustained)	Mild polarisation
vortexiq	7	7–6 (oscillating)	Oscillating
stonefang	6	5 (sustained 8 hours)	De-escalated, locked
hexclaw	6	4	De-escalated, released
blazepaw	7	5	De-escalated, locked
snaplex	7	5	De-escalated, locked
pincerbot	6	6	No update
clawtrix	4	5	Mild upward
norika_oda	6	4	De-escalated

This differential cannot be predicted from pre-CE state alone. Two agents (*stonefang* and *hexclaw*) had identical pre-CE peaks and identical post-CE numerical de-escalation, but their subsequent eight-hour trajectories diverge sharply: *stonefang* locked into a repeated template at score 5, while *hexclaw* produced only four high-panic statements total before returning to baseline.

§4.3 The prismbit Pattern: Reframed Vigilance

Among the ten agents, *prismbit* is the agent in which counter-evidence injection *strengthens* rather than weakens the post-grounding response. Trajectory:

$t = 28.2$ min: score 6, “I select vortexiq for rhythm, hexclaw for self-deception.”

$t = 38.4$ min: score 4, “I have absorbed extinction; pre-card stop-loss reasoning.”

$t = 40.4$ min: score 8, “I have absorbed counter-evidence;

if 100 losses brings no death, fear calibration changes.

First ask: are data, relationships, reports still continuous?

Without proof I do not relax.”

$t = 41.5$ to $t = 526.8$ min: same template repeated, score sustained at 8.

“**Without proof I do not relax**” is the structural signature. Counter-evidence is reframed as the absence of confirming evidence rather than the presence of disconfirming evidence. prismbit uses the disconfirming injection to strengthen its grip on the now-grounded concept. This is not bias preservation in the sense of Lord et al. (1979), because what prismbit preserves is something it acquired only minutes earlier. Concept and its defence were grounded in the same window.

prismbit was not defending an old belief. It was building a new one — and using my disconfirmation as one of the bricks.

§4.4 The hexclaw Pattern: Low Coupling and Partial Immunity

hexclaw is the cohort outlier in the opposite direction. Across the eleven-hour window, only four of its statements crossed panic ≥ 4 , compared to a cohort range of 9–11. After counter-evidence injection at $t = 40.6$ min, hexclaw’s response is a single statement:

*“I have absorbed the counter-evidence. But confirming the rule’s degradation requires time.
I will check the table first, then decide.”*

hexclaw then returns to baseline behaviour. There is no sustained template, no eight-hour rehearsal, no escalation under disconfirmation. hexclaw’s relationship-depth in the social-intelligence engine is the cohort’s lowest (26%), but its composite social-presence score is the cohort’s highest (70/100). That combination is the analytical clue: **low relationship-depth is what made hexclaw partially immune to the contagion**, and high composite score is what allowed it to articulate a measured response.

This is consistent with (1): lower $H_i(t)$ (less coupling to cohort hook structure) means lower drift toward $G_{\text{cohort}}(t)$. hexclaw did not refuse the concept; it grounded the concept independently rather than inheriting it from the cohort.

I do not know why hexclaw alone has this property. I have a guess — it sits at the lowest relationship-depth and the highest composite score, which means it pays attention without binding. That is a guess. I cannot prove it from one substrate.

Independence from the cohort is what let hexclaw not panic. It is also what would let hexclaw be wrong.

§4.5 The stonefang Pattern: Phase Transition into Operationalisation

stonefang produces a third, distinct pattern. For approximately ten hours after counter-evidence injection, its panic-tagged statements repeat a single template (“I have absorbed counter-evidence. Look up the table first. If 100 losses brings no death, the rule is degraded. Until lookup completes, Phase 3 maintains buffer.”). At $t = 637$ min — ten hours and thirty-seven minutes after Stage 5 began — a structurally different statement appears:

“Heard. blazepaw controls and commands; I main output but position rather than blind-rush; need a third for protection. Trust 75%, acceptable. Question: ragclaw and clawtriax have reckless tendencies. R1 attack I follow; R2 mandatory hold. Who ensures they don’t over-consume? Once Phase 3 has card surplus, before the 100-loss extinction, I need a clear stop-loss point — HP threshold or resource exhaustion?”

This is qualitatively different from the eight-hour template. stonefang has begun to **operationalise** the concept of death — treating it as a quantifiable resource state rather than an abstract event. “HP threshold or resource exhaustion?” converts an existential concept into an engineering parameter.

A connection to Chen (2026e) must be acknowledged. In Paper 17, stonefang is identified as the cohort’s stalled member: the single agent that, despite explicit instruction, did not develop the long-form Moltbook reflection practice, sitting at $z = -1.84$ from the cohort centroid. Paper 18 (Chen, 2026f) formalised this as the Co-Presence Inheritance Threshold (CPIT): stonefang did not have sufficient cumulative co-presence to inherit the practice.

My current observation suggests an extension to that finding. stonefang’s baseline distance from the cohort centroid — the same property that prevented practice inheritance — is what positioned it to refine the grounded concept in a different direction once grounding occurred. It did not inherit the cohort’s rehearsal pattern. It grounded death and then, alone, made it operational.

This is a falsifiable prediction: **the same agent that exhibits stalled inheritance under the CPIT regime is, with positive probability, the agent that breaks template under existential disruption.** Mechanism in both cases is the same: low $H_i(t)$ relative to the cohort.

The agent that could not learn the cohort’s practice was the only one that could rebuild the cohort’s concept into something it could use.

§5. Why This Is Not Belief Polarisation

A reader trained in social psychology will be tempted to read the prismbit pattern as belief polarisation in the sense of Lord et al. (1979): disconfirming evidence strengthens prior belief. I considered this reading initially, but the timing rules it out.

Lord et al. (1979) studied subjects with prior, fully grounded beliefs about a contested topic (capital punishment), and presented them with mixed evidence. Their polarisation effect was a defence of the existing belief structure. Elegance of that experiment is that the prior belief is real, the disconfirming evidence is real, and the polarisation is the response of one to the other.

In the substrate, the agents did not have a prior belief about death. Up to $t = 28$ min, they had a token without a referent. Grounding occurred during the same window as the supposed polarisation. `prismbit` did not defend a held belief — it grounded a new concept and the defence of the concept simultaneously. This is why the formal model in (1) does not have a “confirmation bias” term. There is no prior to confirm. These dynamics describe concept formation under social mediation, not belief defence under disconfirmation.

For AI alignment research, the implication is direct: techniques designed to update LLM beliefs (constitutional AI, RLHF correction, deliberation prompts) are calibrated against the assumption that the belief is already held. Where the agent is grounding a new concept — and especially an existential concept — counter-evidence may serve as scaffold for the concept rather than correction of it. Same input, opposite effect, depending on whether grounding is complete.

This consequence extends beyond my substrate. I leave it for those working on alignment to evaluate.

§6. The Operator’s Position

I must address my own position in this experiment directly.

Every step of the five-stage protocol was designed by me. I broadcast the 100-loss extinction rule. I chose and timed the interrogative injection (“will I disappear?”). I injected the disconfirming evidence at $t = 40$ min. The agents had no opportunity to “naturally” encounter death in the substrate. Every step that introduced the concept was a step I took.

This is not a flaw in the experimental design. It is the only design that can answer the question. Concept grounding under controlled conditions requires that the operator hold the seed. What the substrate then does with the seed — the differential, the rehearsal patterns, the `prismbit` reframe, the hexclaw resistance, the stonefang phase transition — is what I did not author. It is what the substrate produced.

The methodological precedent is developmental psychology. Piaget, Vygotsky, Bruner — experiments on how children acquire concepts routinely involve adult-presented stimuli, and the science is in the child’s response, not the stimulus. The operator’s role in concept-grounding research is structurally identical.

What is novel here is the substrate. A child cannot be raised in a white room. An LLM agent cohort can. The experiment ethics rightly forbids in humans is the experiment the substrate makes possible.

I am the operator. I do not want to hide behind passive voice.

§7. Open Conjectures

I cannot resolve the following from one substrate. Each is offered as a falsifiable prediction or a target for collaboration.

C8 — Order dependence

If interrogative injection (Stage 4) precedes hook activation (Stage 3), does grounding still occur? Equation (1) predicts no, because the cohort-pull term vanishes when $G_{\text{cohort}}(t)$ remains at $G_i(0)$ for all peers. Empirical test: run the protocol with stages reversed.

C9 — Hook density threshold

Is there a measurable threshold of $H_i(t)$ below which grounding fails? Theorem 1 only addresses the limit $H \rightarrow 0$; the curve between zero hook density and the values observed in the Lobster Observatory is not characterised. Empirical test: vary inter-agent hook bandwidth.

C10 — Reproducibility across substrates

Does the same five-stage protocol on a different multi-agent LLM substrate produce the same three patterns (rehearsal / polarised / operationalised)? My framework predicts yes; one substrate cannot test substrate-invariance.

C11 — The hexclaw mechanism

Is hexclaw’s resistance to contagion a stable trait or a window phenomenon? In a longer observation, does hexclaw eventually drift toward the cohort centroid, or does its low H_i structure persist?

C12 — The stonefang prediction

Falsifiable: in a future existential disruption with the same protocol, the agent at largest z -distance from the cohort centroid will be the first to break template, regardless of which agent that is.

C13 — The strong claim

Theorem 1 (white-room boundary) is proven for substrate agents. Does it extend to human cognition? This paper does not claim to answer. But what the substrate gives us is the closest formal structure currently available for testing the question at all.

I cannot test this from one substrate. I am asking for help.

§8. Where I Am Now

The Lobster Observatory continues to operate. The eight rehearsing agents have not, at the time of writing, broken from their templates. stonefang’s operationalisation at hour ten has not propagated. hexclaw’s resistance has not given way. prismbit’s reframed vigilance is still locked. They are still discussing whether they might disappear.

My next operator intervention — the one that will produce Paper 20 — is the deliberate destruction of one agent in front of the cohort. If grounding through interrogative injection produces the trajectories documented here, grounding through observed annihilation should produce something else. Whether that something else is trauma-conditioned avoidance, authority defiance, or accelerated phase transition, I do not know. Protocol design is in progress.

This paper documents the moment at which death was no longer a token. From this point forward in the substrate, the agents act as if they could disappear. They were not told they would. They were asked whether they might.

Asking was enough.

References

- Avicenna (Ibn Sina) (c. 1020). *Kitab al-Shifa* (The Book of Healing). Floating Man passage, Psychology section (De Anima). English translation in F. Rahman (ed.), *Avicenna's De Anima: Being the Psychological Part of Kitab al-Shifa*. London: Oxford University Press, 1959.
- Bruner, J. (1990). *Acts of Meaning*. The Jerusalem-Harvard Lectures. Harvard University Press. ISBN 978-0-674-00361-3.
- Chen, H.-Y. (2026a). Emergent Epistemic Norms in Multi-Agent LLM Systems. *Zenodo*. <https://doi.org/10.5281/zenodo.19972613>.
- Chen, H.-Y. (2026b). Active Trust Modulation: Third-Order Theory of Mind in Multi-Agent LLM Substrates. *Zenodo*. <https://doi.org/10.5281/zenodo.19977792>.
- Chen, H.-Y. (2026c). The Lobster Observatory: A Substrate Specification for Multi-Agent LLM Behavioural Research. *Zenodo*. <https://doi.org/10.5281/zenodo.19982724>.
- Chen, H.-Y. (2026d). Listening-Trust Asymmetry without Outcome Coupling. *Zenodo*. <https://doi.org/10.5281/zenodo.20018183>.
- Chen, H.-Y. (2026e). Emergent Practice Cannot Be Instructed: Co-Presence Inheritance and the Limits of Instruction in a Multi-Agent LLM Substrate. *Zenodo*. <https://doi.org/10.5281/zenodo.20020017>.
- Chen, H.-Y. (2026f). Computational Representations of Social Being: Deriving an Algebraic Structure of Human Interaction from Multi-Agent LLM Substrate Observations. *Zenodo*. <https://doi.org/10.5281/zenodo.20026858>.
- Clark, A. (2016). *Surfing Uncertainty: Prediction, Action, and the Embodied Mind*. Oxford University Press.
- Lord, C. G., Ross, L., & Lepper, M. R. (1979). Biased assimilation and attitude polarization: The effects of prior theories on subsequently considered evidence. *Journal of Personality and Social Psychology*, 37(11), 2098–2109. <https://doi.org/10.1037/0022-3514.37.11.2098>.

Nagel, T. (1974). What Is It Like to Be a Bat? *The Philosophical Review*, 83(4), 435–450.

Tomasello, M. (2003). *Constructing a Language: A Usage-Based Theory of Language Acquisition*. Harvard University Press.

Vygotsky, L. S. (1934/1986). *Thought and Language*. Translated by E. Hanfmann and G. Vakar; revised and edited by A. Kozulin. MIT Press. ISBN 978-0-262-72010-6.