Emergent Language as an Approach to Conscious AI

Zengqing Wu
University of Osaka
zengqing.wu@ist.osaka-u.ac.jp

Chuan Xiao
University of Osaka
chuanx@nagoya-u.jp

Abstract

The question of whether artificial systems can be conscious remains open, in part because existing approaches either evaluate systems against theory-derived checklists (*discriminative*) or engineer consciousness-inspired modules directly (*architectural*); both leave open whether observed structures are artifacts of human language priors. We propose a *generative* methodology: emergent language (EL) in multi-agent reinforcement learning, where agents start from minimal (no language, no concept of self, minimal exposure to human text) and develop communication under task pressure alone, ensuring causal attributability to task demands rather than inherited human language priors. We position our methodology by discussing how EL serves as a generative tool for studying consciousness-relevant structure, including the role of environment complexity and the interpretation of emergent communication. As a proof of concept, we instantiate this methodology in a minimal environment and show that agents develop *self-referential communication*, including an echo-mismatch detection circuit that is not predicted by task structure or architecture alone but emerges from a specific environmental affordance.

1 Introduction

Whether artificial systems can be conscious is among the most contested questions in AI [8], with immediate ethical and regulatory implications. Proponents point to increasingly sophisticated behavior [7, 20, 33], while skeptics counter that behavioral sophistication is neither necessary nor sufficient for consciousness [12]. Although this debate [1] has persisted for decades without resolution, exploratory efforts have been made to study conscious AI [8].

Central to this topic lies the *hard problem* of consciousness [11]: the explanatory gap between physical function and subjective experience, between objective behavior and the felt quality of qualia [2]. *Philosophical zombies* (or simply *zombies*), the central figure in this debate, are beings functionally identical to humans yet lacking any conscious experience. If zombies are conceivable, the argument goes, then consciousness cannot be wholly explained by function alone. These concerns highlight the depth of the challenge: understanding not just what artificial systems can do, but whether their information processing could give rise to the kind of first-person experience that characterizes consciousness.

In this paper, we propose a methodology for studying conscious AI. Our starting point is Moody's zombie civilization argument [56]: in a world of zombies, the concepts of conscious experience would never originate, because there would be nothing for such concepts to refer to. We draw methodological inspiration from this argument by designing environments for AI agents and speculating whether they can independently develop a word that means "consciousness" in their language. Meanwhile, we remain agnostic about subjective experience, thereby differing from Moody's viewpoint in the sense that Moody is concerned with the origin of consciousness *concepts*, whereas we focus on the emergence of consciousness-relevant *structures*.

Our source code is available at https://github.com/wuzengqing001225/ConsciousAI_Indexicality/.

[1] Key opinion leaders' positions are available in Appendix A.

[2] Qualia refer to instances of subjective experience, e.g., the "blueness" of blue.

Our approach is to construct minimal virtual worlds populated by agents—with no language, no concept of self, and minimal exposure to human text [3]—and to observe what consciousness-relevant structures emerge from task pressure alone. Any observed structure is necessarily emergent rather than inherited from human language priors, as opposed to language models—in particular, large language models (LLMs)—pre-trained using human language data. We propose that progress requires not resolving whether AI "has" consciousness, but developing methodologies that can track the structural conditions under which functional preconditions for conscious experience arise. In this sense, our methodology belongs to the category of *functionalism*.

Two methodological principles. Our approach rests on two commitments. (1) **Environment shapes behavior**: Just as ecological demands drive the evolution of cognitive capacities in biological organisms, we hypothesize that sufficiently structured environments can drive the emergence of functional structures relevant to consciousness in artificial agents, without those capacities being designed in. (2) **Phenomenological epoché**[4]: a deliberate suspension of judgment on whether AI systems could have subjective experience. We bracket the question of subjective experience and focus on what can be observed: the linguistic practices that agents develop under task pressure, and the environmental structures those practices ground. As such, we focus on a tractable phenomenological question: what structures in the agent's behavior and representations are functionally equivalent to the preconditions identified by consciousness theories? The epoché ensures that any emergent structures we observe are products of task pressure and environmental demands, not artifacts of consciousness-theoretic assumptions built into the system design.

Emergent language as a generative approach. To investigate these structural preconditions empirically, we turn to emergent language (EL) in multi-agent systems. In EL research, agents with minimal prior language learn to communicate through discrete tokens under task pressure, developing communication protocols from scratch [22, 41]. We use EL not to improve agent performance (the standard goal of the field) but as a *generative* approach to studying the origins of consciousness-relevant structures. By placing agents under coordination pressure with bandwidth-constrained channels, we observe what structures emerge when agents need to transmit information about themselves (Figure 1).

The EL approach addresses a methodological gap left by two existing paradigms. *Discriminative* methods take a checklist of indicators derived from consciousness theories and ask whether a given system satisfies them [8]: productive but retrospective, since the system is designed for other purposes and the evaluation can only confirm or deny pre-specified criteria, never reveal structures the theory did not anticipate. *Architectural* methods build consciousness-inspired modules directly into a system [19, 91]: informative but circular, since any resulting consciousness-relevant behavior may reflect the designer's assumptions rather than structural necessity. Our *generative* approach asks neither "is it conscious?" nor "can we make it conscious?" but "what consciousness-relevant structures arise when none are designed in?"

Methodology in brief. Our methodology operates through three interconnected elements. (1) *Environment complexity as a driver*: following the principle that general methods leveraging computation outperform domain-specific knowledge engineering [83], we scale environment complexity [5] rather than encoding target capacities directly, letting task pressure drive emergence. (2) *Prior-minimal design*: agents learn from scratch with minimal exposure to human language, ensuring that any observed structure is causally attributable to current task demands rather than inherited priors. (3) *Interpretation through intervention*: emergent protocols are analyzed via ablation, probing, and information-theoretic decomposition; at larger scales, LLM-based in-context learning and alignment with human language provide complementary interpretation pathways.

Indexicality as the first step. To employ our methodology, agents must express their internal states. Thus, the first question is whether agents can develop *self-referential communication*: messages whose content is about the sender itself. Perry's essential indexical [63] and Kaplan's character/content distinction [37] established the *demand side*: indexical structure is philosophically indispensable for self-referential communication. We investigate the *supply side*: whether indexical structure also arises under bandwidth-constrained communication.

[3] We say minimal rather than no exposure because the design of environment, agent architecture, and reward is based on human knowledge and thus inevitably affected by human language prior. Nevertheless, we aim to reduce such impact to the minimal level.

[4] Phenomenological epoché (a.k.a. bracketing), as a preliminary step of phenomenology, means refraining from judgment about the reality and focusing on how things appear to us.

[5] This element serves as an envisioned research agenda in this paper, while the other two elements are evaluated empirically.

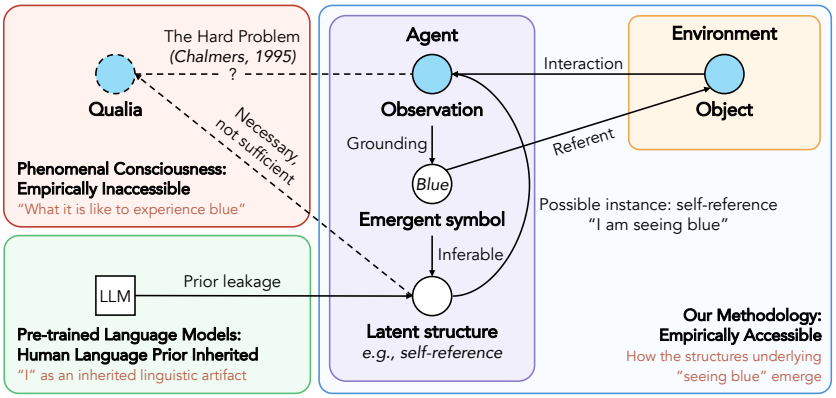

Figure 1: EL as a methodological approach to conscious AI. (1) Using "blue" as an example, the figure separates the private phenomenal question (what it is like to experience blue) from the empirically accessible structures surrounding "seeing blue." An agent may observe blue, ground a symbol for blue through interaction with blue objects, and develop emergent communication that makes latent internal structure inferable. A possible instance of such latent structure is self-reference, as in "I am seeing blue." This does not show that blue has any subjective feel for the agent; the qualitative character of blueness remains bracketed by epoché. (2) The figure also contrasts the EL route with the LLM route, where apparent self-reference may reflect inherited human language priors. (3) The methodological claim is functional: EL provides a grounded, empirically accessible window onto consciousness-relevant latent structures, not direct evidence of qualia.

As a proof of concept, we train reinforcement learning (RL) agents that communicate via discrete tokens to coordinate on a cooperative task. Each agent observes only its own private state, but the task requires knowledge of the partner's state, making communication necessary. The channel is narrow, and agents begin with no language and no concept of self (i.e., no architectural module representing the agent itself, no first-person token in the vocabulary, and no auxiliary self-supervised objective targeting the agent's own state). We identify three structural properties:

- **(P1) Indexical encoding:** messages carry primarily the sender's own state (information-theoretically predicted by task structure);
- **(P2) Persistent state representation:** a recurrent latch maintaining self-state across time (architecturally expected under partial observability);
- **(P3) Behavioral self-monitoring:** agents develop a closed-loop circuit that detects mismatches between what they intended to transmit and what is echoed back.

P1 and P2 are predicted foundations that validate our methodology's observation conditions. P3 is the core empirical finding: it is not predicted by task structure or architecture alone, requires a specific environmental affordance (the echo channel), and demonstrates that our methodology can detect emergent functional structure beyond what task pressure and agent architecture directly predict.

Contributions. (C1) A generative methodology of EL with the prior-minimal design for studying the origins of consciousness-relevant structures in AI, complementing discriminative and architectural approaches. (C2) A formal operationalization of indexical reference connecting Kaplan's character/content distinction to mutual-information criteria testable in emergent communication systems. (C3) A proof-of-concept experiment in which P1 (indexical encoding) and P2 (persistent state representation) serve as observation conditions and P3 (behavioral self-monitoring) serves as the core finding.

Outline. §2 reviews related work. §3 develops our methodological position on how EL serves as a generative tool for studying conscious AI. §4 formalizes indexicality and structural preconditions. §5–6 present experiments and results. §7 discusses implications of the experimental findings. §8 concludes.

2 Related Work

We review two bodies of work: consciousness theories that presuppose the first-person structure we aim to explain (§2.1) and existing research on EL (§2.2). More discussion on related work is available in Appendices §B–C.

2.1 Consciousness Theories and the First-Person Presupposition

The dominant theories differ in mechanism but share a presupposition: each requires a first-person perspective without explaining its origin. Global Neuronal Workspace Theory (GNWT) needs someone for whom information is broadcast across prefrontal-parietal networks [18]; Integrated Information Theory (IIT) needs someone experiencing the integration quantified by Φ [87]; Higher-Order Thought theory (HOT) needs someone whose mental state is the object of a higher-order representation [71]; Attention Schema Theory (AST) needs someone whose attention is being schematically modeled [26]. The pattern extends to more recent proposals: the Consciousness Prior [3] selects a low-dimensional summary via attention, but attention directed by whom? Predictive self-models [73, 54] ground experience in the brain's self-modeling capacity, yet the "self" doing the modeling is presupposed rather than derived. Hofstadter [33] makes the recursion explicit, casting self-reference as a strange loop, but this redescribes the structure without explaining how it originates. What binds these frameworks is not a shared mechanism but a shared gap: each presupposes a subject-position and then explains what that subject does, without asking where the subject-position itself comes from.

A practical difficulty compounds the theoretical one: several of these theories lack operational counterparts in AI systems. GNWT posits a neuronal workspace with specific prefrontal-parietal dynamics; IIT requires computing Φ over a system's causal structure, a calculation that is intractable for large networks and whose applicability to non-biological substrates is contested. Neither framework provides a clear procedure for evaluating consciousness in architectures that differ fundamentally from the mammalian brain. This operational gap is one reason the debate remains unresolved. Empirical and conceptual challenges further compound the problem. The Cogitate Consortium's adversarial test of IIT and GNWT found partial support for each but confirmation of neither [17]. Chalmers [13] argues that no current evidence suffices to establish or deny LLM consciousness. McClelland [52] states that we hit an "epistemic wall" when extrapolating AI systems because what we know about consciousness originates from human organisms. Shevlin [75] identifies a "specificity problem" for non-human systems. Hoel [32] poses a dilemma between falsifiability and triviality.

In this paper, we sidestep the question these debates address (i.e., whether a system is conscious) and instead ask where the structural preconditions for self-referential communication originate. A related effort is made by Immertreu et al. [35], who probe RL agents trained in virtual environments for world models and self-models using Damasio's theory of core consciousness, finding that rudimentary forms of both emerge as byproducts of task training. Our work shares the spirit of probing for emergent structure, but differs in method (communication-based rather than probe-only) and scope (emergence of self-referential communication rather than self-models in isolation).

2.2 Emergent Language

EL research studies what happens when agents with minimal prior language must learn to communicate in order to coordinate. In a typical setup, multiple agents are placed in a cooperative environment where task success requires information that only other agents possess. Agents communicate through a discrete channel with limited bandwidth (e.g., a small vocabulary of tokens, one per time step) and develop their communication protocol entirely through RL, with no supervision on message content [43, 41, 22]. The field has produced a rich body of results on compositionality, grounding, and repair under noise [14, 64].

Existing research has predominantly treated EL as a communication protocol for improving multi-agent system performance on downstream tasks: coordination, navigation, referential games. A systematic gap runs through this literature: research has focused on *what is said* (whether agents develop compositional, grounded references to external states) but not on *who is saying it*. Whether agents develop signals that encode the speaker's own identity or refer to themselves has barely received attention. Chen et al. [14]'s "who" refers to communication topology, not to whether agents develop tokens that function as first-person reference. This gap is notable in light of Locke [49], who argues that primate communication is primarily indexical: close calls, grooming, and other signals convey personal states and traits rather than symbolic environmental references, and that this personal information exchange is fundamental to group cohesion. If indexical encoding is the biological baseline, its absence from the EL literature is a significant omission. This is the gap we address: we use EL not as a means to improve task performance, but as a window into the emergence of consciousness-relevant structures [6].

[6] Discussion on why communication, rather than internal probing alone, is the appropriate lens is available in Appendix G.1.

3 Methodology

3.1 Environment Complexity as a Driver

In the Bitter Lesson [83], Sutton argues that the most effective approach in AI is to leverage *computation and learning* rather than encoding human knowledge. Silver and Sutton [76] extend this to the "era of experience": agents that learn from interaction with the environment, rather than from human-generated data, will break through the limitations of human-centric AI; actions and observations grounded in the environment, rather than mediated through human text, provide the foundation for discovering strategies that might never occur to a human. We adopt an analogous principle for the study of conscious AI: rather than designing consciousness-relevant (e.g., self-referential) capacities into agents, we scale environment complexity and let task pressure drive their emergence.

The comparative cognition literature provides a biological analogue. Self-referential capacities in nature form a continuum driven by ecological demand: mirror self-recognition in chimpanzees [25], uncertainty monitoring in rhesus monkeys [27, 77], theory of mind in social primates [67], and narrative self-continuity in humans [21]. Each capacity appears when, and plausibly because, the organism's environment creates selection pressure that rewards it.

Scaling dimensions. We hypothesize that the same logic applies to AI agents: progressively more complex environments should drive progressively richer consciousness-relevant structures. In §5–6, we will show that our minimal environment (two agents, seven tokens, ten-step episodes) suffices to drive the three structural preconditions P1–P3. Several dimensions of scaling are available: environment complexity (from grid worlds to physically grounded simulations to real-world interaction), agent population size and heterogeneity, communication bandwidth and modality (from discrete tokens to continuous signals to physical vocalization), temporal horizon (from fixed episodes to open-ended streams), and social structure (from fixed partnerships to dynamic alliances and adversarial encounters). Each dimension introduces new pressures that could drive richer emergent structures.

Following the Bitter Lesson, we scale environment complexity and observe what emerges, without encoding human knowledge about consciousness (e.g., self-recognition, self-awareness, identity) as design targets. Nonetheless, indexicality, which refers to the emergence of tokens that refer to the speaker's own states, is the first observation condition: without evidence that agents encode their own states, we cannot study further consciousness-relevant structures. §4 will elaborate on this.

3.2 Scaling with Minimal Human Language Prior

In our methodology, a key criterion is *causal attribution*: for any observed structure, ablation or intervention must verify that it causally originates from the environment. This criterion becomes increasingly important as environments grow in complexity, because richer environments, along with more sophisticated agent architecture, also create more opportunities for subtle prior leakage. This rules out LLM agent approaches, which risk importing the first-person pronoun "I" as a statistical regularity rather than an emergent structure [7].

3.3 Interpreting EL

Besides the EL itself, its interpretation is also a key step in our methodology, because EL is not designed for human readability. We outline several interpretation strategies at different scales, noting that these are tentative methodological proposals rather than established standards. The field may develop better approaches as it matures. The interpretation challenge scales with environment complexity, and different regimes call for different methods.

Small scale: ablation and probing. In minimal environments, the standard toolkit of mechanistic interpretability suffices: mutual-information analysis, linear probing of hidden states, ablation of channels and architectural components, and intervention experiments (e.g., corruption, silencing). This is the approach we adopt in the experiments of this paper. Its strength is causal clarity. Its limitation is that it does not scale to environments where the EL is too complex for manual analysis.

Medium scale: LLM in-context learning. At intermediate complexity, LLMs can serve as interpreters of EL via in-context learning: given examples of agent behavior paired with environment states, an LLM can be prompted to describe the structure of the EL in human-readable terms. This

[7] Detailed discussion is available in Appendix G.2.

leverages the LLM's capacity for pattern recognition without requiring the EL to have been part of the LLM's training data. Using an LLM as interpreter does not reintroduce the prior leakage that motivates our prior-minimal design [8].

Large scale: pre-training and alignment. At high complexity, a dedicated language model could be pre-trained on the agents' communication stream (i.e., emergent communication as a corpus) and then aligned with human language (e.g., by mixing with human text corpus for pre-training), creating a translation layer between emergent and natural language.

3.4 Continual Learning and Transfer

The richer capacities such as metacognitive monitoring, narrative self-continuity, and theory of mind plausibly require agents that continue to learn during deployment, not just during a fixed training phase. Valiente and Pilly [89] demonstrate one such direction: their MUSE framework integrates metacognitive processes (self-assessment and self-regulation) into autonomous agents, improving out-of-distribution performance. Our P3 provides a minimal precursor to such metacognitive capacity; scaling to richer self-monitoring would require the kind of continual adaptation MUSE exemplifies. Biological organisms develop conscious capacities over developmental timescales through ongoing interaction with an environment whose complexity itself changes. Frozen-weight agents, as used in our experiments, can demonstrate that structural preconditions can emerge, but cannot demonstrate the open-ended developmental trajectories that richer consciousness-relevant capacities would require.

A natural scaling path is *sim-to-sim transfer* (cf. sim-to-real transfer [96]): train agents in a simple environment until basic communication and consciousness-relevant structures stabilize, then transfer them to a more complex environment where the existing structures provide a scaffold for more elaborate capacities. This mirrors developmental trajectories in biological organisms, where early sensorimotor competence scaffolds later cognitive achievements. The key constraint is that the transfer must import minimal human language priors.

3.5 Limitations of the EL Methodology

The generative approach via EL faces several obstacles that should be acknowledged as part of the methodology itself, not deferred to post-hoc limitations.

Computational cost. Prior-minimal agents must discover from scratch what evolution has spent billions of years optimizing. Even our minimal environment requires tens of thousands of training updates per condition; scaling to more complex environments requires substantially greater computation.

Human language priors not leveraged. The prior-minimal design, while methodologically essential, means that the vast knowledge encoded in human language, including knowledge about self-reference, mental states, and social cognition, is deliberately excluded. An important open question is whether the structures that emerge via EL and those that emerge via next-token prediction on human text (e.g., LLMs) can be shown to converge to equivalent models.

No established paradigms for continual learning. Reaching richer capacities likely requires agents that learn continually and maintain long-term memory across changing environments. Current RL frameworks are predominantly designed for fixed-horizon episodic training; paradigms for open-ended continual learning in multi-agent communication settings remain an open research problem [59].

The mesoscale gap of interpretation. Between the small and large regimes identified in §3.3 lies a potential gap: systems complex enough that ablation becomes impractical, yet not complex enough for learning methods to extract reliable structure. This mesoscale problem, analogous to the challenges of studying systems with intermediate degrees of freedom in physics, is an open methodological challenge for the field.

[8] We discuss this in Appendix G.3.

4 Indexicality

4.1 Why Indexicality Is Necessary

Our methodology studies consciousness-relevant structures by observing what agents communicate. For this to work, agents' messages must carry information about their own internal states; otherwise, communication provides no empirical window into internal representations, and the methodology has no signal to analyze. Indexicality is therefore a necessary observation condition: it is what makes agents' internal structure accessible through their communicative output [9].

Concretely, if agents develop a protocol in which $I(m;s_{\text{self}}) \approx 0$, where $I(\cdot;\cdot)$ denotes mutual information, messages are uninformative about the sender, and probing, ablation, and information-theoretic decomposition all return null results. The entire evidence chain for persistent state representation (P2) and behavioral self-monitoring (P3) depends on indexical encoding (P1) holding first: P2 is detected by probing hidden states that encode s_{self}, and P3 is detected through changes in communicative output that already carries sender state. Without indexical encoding, neither is empirically accessible.

4.2 Formal Definition of Indexical Reference

Setting. Consider N agents ($N = 2$ in our experiments), each assigned a private state s_i drawn independently from a finite set $\mathcal{S}$. Each agent additionally observes a context $c_i \in \mathcal{C}$, drawn independently per agent and per episode. c_i is independent of the partner's state s_j, and both enter the reward formula together: the acting agent must combine its own c_i with the communicated s_j to select the correct action (an explicit form $a^*_{i\to j} = (s_j + c_i + t) \bmod 3$ is given in §5.1, where t is the time step).

At each time step t, agent i selects a message $m_i^{(t)} \in \mathcal{M}$ from a discrete vocabulary of size $|\mathcal{M}|$ and an action $a_i^{(t)}$. Each agent receives the previous-step messages from the other agents. In our two-agent experiments, agent i receives the single partner message $m_{-i}^{(t-1)}$. The task requires each agent to take actions that depend on other agents' private states, making communication necessary for coordination. The communication channel is narrow at the per-step level: $|\mathcal{M}| < |\mathcal{S}|^N$ compares the per-step message capacity to the joint private-state space, so a single message cannot encode the full configuration of all agents' private states. We deliberately measure narrowness in this single-step sense rather than including $|\mathcal{C}|$ or summing over T steps, because each agent already observes c_i and t locally and need not transmit them. The bandwidth pressure is over what must cross the channel, not over what is privately available. Agents must therefore develop efficient encoding strategies for the information that is communicable in principle. Throughout, sender identity is treated as given by channel position (each receiver knows which channel/slot a message arrived on). Under this simplifying assumption, the receiver does not need to recover the sender from the message itself [10].

This setting inherits the cooperative multi-agent RL (MARL) structure of Foerster et al. [22] and Mordatch and Abbeel [57], but with a critical difference: the referents that matter for task success are agents' private internal states, not shared external observations. Whereas referential games [41, 29] require agents to communicate about objects visible to at least one party, our task requires agents to communicate about themselves.

Definition. We define a token $m \in \mathcal{M}$ as an *indexical token* w.r.t. sender i if it satisfies three criteria:

1. **MI dominance.** The mutual information between the token and the sender's own state dominates:
$$I(m_i;s_i) \gg I(m_i;s_j) \quad \text{for all } j \neq i. \tag{1}$$
That is, the token's information content is primarily about the sender, not about other agents.
2. **Context independence.** The encoding rule is invariant across contexts:
$$I(m_i;c_i \mid s_i) \approx 0. \tag{2}$$
Conditioned on the sender's state, knowing the sender's context provides negligible additional information about the token. This invariance ensures that the same encoding rule applies regardless of context, supporting generalization to novel context values; without it, the encoding could be context-specific lookup rather than a stable mapping from self-state to message.
3. **Communication necessity.** Removing the message channel causes a significant drop in task performance:
$$\Delta_{\text{comm}} = R_{\text{with-msg}} - R_{\text{no-msg}} \gg 0. \tag{3}$$

[9] The philosophical foundations [63, 37] that motivate the formal definition of indexicality are developed in Appendix D.1.

[10] The conditions under which compositional "who/what" encoding becomes pressured are discussed in Appendix G.6.

This ensures that the token is not epiphenomenal: it carries information that the receiver cannot obtain by other means.

Relation to Kaplan's framework. The three criteria jointly capture an information-theoretic analogue of the structure Kaplan [37] identifies in indexical reference: a stable encoding rule (analogous to character) whose referent (analogous to content) varies with the sender. The correspondence is structural rather than formal. Kaplan's character operates on contexts to yield propositional contents that enter truth-value evaluation, while our criteria characterize statistical regularities in token-state distributions. The two share the abstract pattern (stable rule $\rightarrow$ context-dependent referent) but operate at different levels of analysis. We invoke Kaplan's vocabulary as a conceptual anchor that motivates why these three criteria jointly isolate self-referential encoding, not as a claim of formal equivalence.

We provide an information-theoretic argument in Appendix D.2 that indexical encoding is the task-natural efficient strategy class under bandwidth constraints, and that P2 follows as an architectural consequence of recurrence under partial observability.

4.3 Self-Monitoring and Communication Repair

P1 and P2 concern representation. P3 concerns functional self-monitoring: sender-specific, echo-based mismatch detection that enables the agent to detect errors in its own communicative output and adjust subsequent behavior accordingly.

The thermometer argument. A thermometer is causally sensitive to temperature: a change in temperature causes a change in its reading. But a thermometer cannot detect that its reading is wrong and adjust its subsequent behavior accordingly. It lacks the closed-loop structure that distinguishes monitoring from mere sensitivity. The question is whether our agents are thermometers (i.e., passively mapping states to tokens) or something more. A reader may immediately object that a thermostat is also closed-loop and yet uncontroversially not a self-monitor; the thermometer test is therefore a necessary but not sufficient condition. We will sharpen the criterion in §7 by distinguishing exogenous from endogenous reference signals: a thermostat's setpoint is human-specified, whereas the reference signal in P3 is the agent's own learned communicative intention.

Operationalization via communication repair. We operationalize P3 through communication repair: an agent that detects corruption of its own message and adjusts its next communication to compensate. Concretely, we introduce an echo channel that feeds back a possibly corrupted copy of the agent's own message, and a corruption mechanism that stochastically replaces tokens with noise. If an agent, upon receiving a corrupted echo, changes its communication behavior at the next time step, it demonstrates the closed-loop causal structure that a thermometer lacks: detection of output error $\rightarrow$ adaptive modification of subsequent output. When the behavioral adjustment additionally yields measurable downstream benefit for the receiver (communication repair in the narrow sense), the evidence for functional self-monitoring is strengthened, but the core criterion is the echo-dependent behavioral change itself. This operationalization connects to Clark's predictive processing [16] (the agent monitors prediction errors in its own output) and to Carruthers' interpretive self-knowledge [9] (the agent accesses its own output through a sensory channel, not introspection).

5 Experimental Design

5.1 Environment

Two agents ($i \in \{0,1\}$) each hold a private state $s_i \in \{0,1,2\}$ and context $c_i \in \{0,\ldots,5\}$ (training; 3 additional contexts held out for generalization), drawn uniformly. Episodes last $T = 10$ steps. At each step, agent i observes $(s_i, c_i, t, m_{-i}^{(t-1)})$ and produces a message $m_i^{(t)} \in \{0,\ldots,6\}$ (6 content tokens plus a dedicated silence token, hereafter SILENCE) and two actions: a partner-targeting action a_i^{other} (whose target is $j \neq i$) and a self-targeting action a_i^{self} (whose target is $j = i$). Both actions follow the same rule $a_{i\rightarrow j}^{*} = (s_j + c_i + t) \bmod 3$, where c_i is always the acting agent's own context. The self-targeting action $a_i^{\text{self}} = (s_i + c_i + t) \bmod 3$ depends only on locally available information. The partner-targeting action $a_i^{\text{other}} = (s_j + c_i + t) \bmod 3$ requires knowledge of s_j that can only come through communication; this is the source of task pressure for the messaging channel. The channel is narrow (a single token per step, and $|\mathcal{M}| = 7$, where one of the seven symbols is the SILENCE token used when the agent chooses not to transmit content), creating the task pressure.

The environment is deliberately abstract [43, 41]: stripping away perceptual complexity ensures that the emergent structure arises from task pressure alone. Modular arithmetic prevents trivial constant-action solutions by forcing correct actions to depend jointly on state, context, and time. Because the environment is finite, we do not claim that lookup-style policies are impossible. Full details are in Appendix E.1.

Partially observable Markov decision process (POMDP) variant (for P2). To test whether agents develop a persistent state representation, we introduce a partially observable variant: each agent observes its private state s_i only at $t=0$; for all subsequent steps, s_i is masked from the observation. This forces agents to latch their private state into recurrent memory if they are to communicate it at later time steps. Since the agent must produce both a^{other} (requiring other agent's state) and a^{self} (requiring its own state) at every step, it cannot simply discard s_i after communicating it.

5.2 Agent Architecture

Each agent uses a Gated Recurrent Unit (GRU; [15]) with hidden dimension 128: $\mathbf{h}_t = \text{GRU}(\mathbf{x}_t, \mathbf{h}_{t-1})$, where $\mathbf{x}_t$ concatenates one-hot encodings of s_i, c_i, t, and received messages. Three linear heads on $\mathbf{h}_t$ produce π_{msg}, π_{other}, and π_{self}. Details are depicted in Figure 4 (Appendix E.2). We note that recurrence is necessary for P2 by construction: under POMDP masking, any memoryless architecture (e.g., a multilayer perceptron) has no mechanism to carry s_i forward past $t=0$.

Echo channel (for P3). For self-monitoring experiments, the input is augmented with an echo: the agent's own previous message $m_i^{(t-1)}$, corrupted with probability ε (replaced by a uniform random token). This enables the closed-loop detection-correction cycle (§4.3).

5.3 Training

Two agents (A, B) with *separate parameters* are trained via independent Advantage Actor-Critic (A2C; [55]) with role alternation. For P3, the echo channel ($\varepsilon = 0.25$) is added together with a four-phase curriculum (communication-only $\rightarrow$ joint training $\rightarrow$ corruption introduction $\rightarrow$ full training with speak cost; details in Appendix E.3). Ten seeds are evaluated on P1 (cross-pair testing), P2 (linear probing under POMDP masking), and P3 (test battery).

Training-condition matrix. Two training regimes underlie the experiments: the *with-echo* regime (fully observable inputs, echo channel enabled, $\varepsilon = 0.25$) and the *no-echo* regime (identical architecture and curriculum, echo permanently silenced). The no-echo regime is used as the train-time ablation control for P3. The POMDP variant (private state s_i visible only at $t = 0$) is applied at probe time only: the linear probes for P2 are run on agents trained under both regimes. Each regime trains 10 independent seeds from scratch with aligned seed indices, so that paired comparisons (e.g., with-echo vs. no-echo at the same seed) control for initialization. Appendix E.3 details shared hyperparameters.

P1. We pair trained partners (A-vs-B) and compare against self-pairings (A-vs-A, B-vs-B) from different seeds. Dialect differences should degrade cross-pair performance but trained-partner communication should remain effective if tokens are genuinely indexical. We evaluate P1 through MI dominance, communication ablation, cross-pair controls, and generalization to held-out contexts ($c_i \in \{6,7,8\}$, unseen during training), corresponding to the three criteria of §4.2. The full test battery (T0–T3) is specified in Appendix E.4.

P2. We train linear probes on the hidden state h_t to predict s_{self} and s_{other} at each time step under the POMDP state-masking condition, where each agent observes its own private state only at $t = 0$ and the learned communication stream is kept intact. This tests whether h_t maintains the agent's own state across time while also acquiring the partner's state from incoming messages. The linear probes for P2 are run on agents trained under both regimes to confirm that the persistent state representation does not depend on the echo channel, establishing a shared baseline for the P3 comparison.

P3. If agents monitor their own communicative output, corruption of the echo should cause them to break silence and speak. The P3 test battery measures this change as the trigger contrast $L_c = P(\text{speak} \mid \text{corrupted echo}) - P(\text{speak} \mid \text{clean echo})$, i.e., the increase in speak rate when the echo is corrupted versus clean (speak rate is bounded in $[0,1]$, so a contrast of $+0.1$ is a 10-percentage-point absolute increase). This is evaluated through tests that isolate who responds (sender vs. receiver),

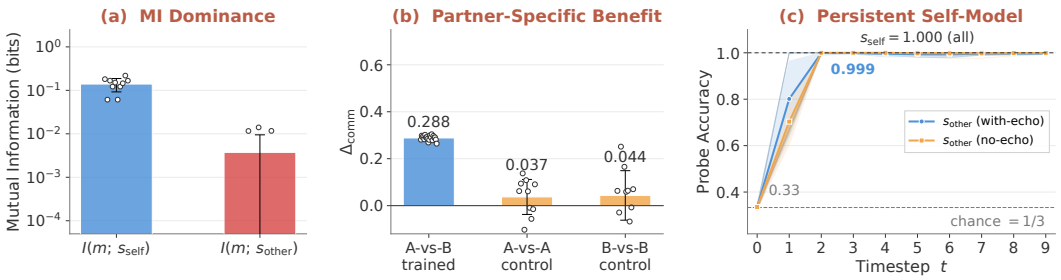

Figure 2: P1 evidence: (a) MI dominance (log scale), (b) partner-specific benefit; P2 evidence: (c) persistent state representation under POMDP masking.

what channel drives the response (echo vs. received message), when detection occurs (same-step vs. one-step delay), and whether the echo channel is causally necessary during learning (train-time ablation). Full specifications are in Appendix E.4.

6 Results

6.1 P1: Indexical Encoding

Mutual information analysis confirms that messages encode the sender's own state: $I(m; s_{\text{self}}) \gg I(m; s_{\text{other}})$ across all seeds (Figure 2a). Trained partners communicate effectively (Figure 2b): A-vs-B $\Delta_{\text{comm}} = 0.288 \pm 0.010$.[11] Self-pairing controls confirm partner specificity: A-vs-A $\Delta_{\text{comm}} = 0.037 \pm 0.075$ and B-vs-B $\Delta_{\text{comm}} = 0.044 \pm 0.105$ (high variance; several seeds show $\Delta_{\text{comm}} \approx 0$ or slightly negative, indicating that mismatched dialects provide no benefit or even interfere with performance). This partner-specificity is expected: the reward only specifies coordination success, not which token should encode which state, so the token-to-state mapping is jointly negotiated within each training run, and different seeds converge on different mappings. The gap between cross-pair and self-pair performance therefore demonstrates that agents develop partner-specific dialects: encoding is indexical (dominated by sender state) but the specific token-to-state map is seed-dependent. No two seeds converge on the same token map, refuting the hypothesis that task structure forces a unique protocol. Communication ablation confirms necessity ($\Delta_{\text{comm}} = 0.288 \pm 0.010$). Context independence is supported by generalization to held-out contexts ($c_i \in \{6,7,8\}$, unseen during training): agents maintain MI dominance and communication performance on novel context values, indicating that the encoding rule is context-invariant rather than context-specific (Appendix F.1).

6.2 P2: Persistent State Representation

Under partial observability (s_i visible only at $t = 0$, zeroed in the input at $t \geq 1$), GRU agents develop a persistent state representation (Figure 2c). Linear probes on $\mathbf{h}_t$ reveal:

- **Self-Latch:** $\mathbf{h}_t \to s_{\text{self}}$: accuracy $= 1.000 \pm 0.000$ across all $t \in [0,9]$, 10 seeds. Although s_i is masked from the probe input after $t = 0$, the GRU hidden state preserves self-state information throughout the episode.
- **Other-Latch:** $\mathbf{h}_t \to s_{\text{other}}$: accuracy $= 0.334 \pm 0.009$ at $t = 0$ ($\approx$ chance for $|\mathcal{S}| = 3$), rising to 0.999 ± 0.003 by $t = 2$ (i.e., after two rounds of message exchange within a single episode). The other-state is unavailable before communication begins, confirming that this information is extracted from incoming messages.

The self-latch is echo-independent: no-echo-trained agents show an identical pattern ($s_{\text{self}} = 1.000$ at all t; s_{other}: $0.336 \to 1.000$). This indicates that self-state retention in the trained hidden state depends on recurrent memory rather than on the echo channel.

6.3 P3: Behavioral Self-Monitoring

Across all 10 seeds, the echo-mismatch detection circuit forms reliably (Figure 3), though the downstream repair protocol varies. The trigger contrast is sender-specific ($+0.118 \pm 0.098$; receiver

[11] The P1 evaluation is performed on agents trained under the with-echo regime (§5.3). The same agent set is reused for P3, which is why the same Δ_{comm} value appears in both §6.1 and §6.3. The number is not from a separate training run.

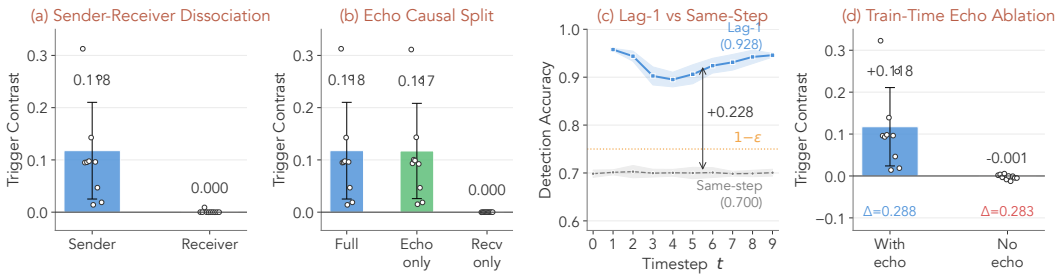

Figure 3: P3 evidence chain: (a) sender-receiver dissociation, (b) echo causal split, (c) lag-1 temporal asymmetry, (d) train-time echo ablation.

0.001 ± 0.003; Figure 3a), driven entirely by the echo channel (echo-only contrast $= 0.117 \pm 0.096$, receive-only $< 10^{-3}$; Figure 3b), and temporally delayed by one step (lag-1 vs. same-step corruption-detection probe accuracy: 0.928 vs. 0.700; Figure 3c). No-echo training abolishes the echo-dependent behavioral and temporal signatures, while preserving communication (Figure 3d).

Delayed echo-mismatch detection. Four converging results reveal the mechanism underlying P3 as a connected causal chain.

First (*who*, Figure 3a), the behavioral response to corruption is sender-specific: the trigger contrast is $+0.118 \pm 0.098$ for the sender whose message was corrupted (one-sample t-test: $t(9) = 3.82$, $p = 0.004$, 95% CI $[0.048, 0.187]$; all 10 seeds positive), while receiver contrast is 0.001 ± 0.003 ($p = 0.36$). A paired comparison confirms the dissociation ($t(9) = 3.78$, $p = 0.004$). The high variance in sender contrast reflects variation in magnitude rather than direction across seeds (see protocol diversity below); the minimum seed-level contrast is $+0.014$. This rules out other-monitoring or generic environmental sensitivity.

Second (*what drives it*, Figure 3b), the response is driven entirely by the echo channel. Under the echo-only condition (incoming messages from the partner silenced), trigger contrast is 0.117 ± 0.096, nearly identical to the full condition (0.118 ± 0.098). Under the receive-only condition (echo silenced, partner messages retained), contrast drops to $< 10^{-3}$ (paired t-test, echo-only vs. receive-only: $t(9) = 3.85$, $p = 0.004$). A further test-time ablation confirms this: replacing the echo input with silence from t_c onward reduces the trigger contrast to 0.000 ± 0.003, confirming that the echo signal is necessary at runtime for the behavioral response.

Third (*when*, Figure 3c), the detection is temporally delayed by exactly one step. Because $\mathbf{h}_t$ is computed before the corruption of m_t occurs, the corrupted echo is first available at $t+1$. A corruption-detection probe (two linear classifiers from $\mathbf{h}_t$, one predicting $\mathsf{was\_corrupted}_{t-1}$ for the lag-1 condition and one predicting $\mathsf{was\_corrupted}_t$ for the same-step condition [12]) confirms this asymmetry: lag-1 accuracy $= 0.958$ at $t{=}1$ (average across timesteps: 0.928 ± 0.010), while same-step accuracy $= 0.700 \pm 0.003$, showing no usable same-step corruption information. The same-step probe serves as a negative temporal control: its accuracy (0.700 ± 0.003) falls below the majority-class baseline ($1 - \varepsilon = 0.75$), confirming that no usable corruption signal is available at the step corruption occurs. This sub-baseline raw accuracy can occur when a linear probe is trained on high-dimensional representations containing no task-relevant signal: cross-entropy optimization overfits to noise, misclassifying some majority-class examples that a constant predictor would get right. An ablation of echo-trained vs. no-echo-trained agents is available in Appendix F.2.

Fourth (*what the agent compares against*), the hidden state encodes communicative intention, not channel output: a linear probe predicting the intended token from $\mathbf{h}_t$ achieves accuracy 1.000, while a probe predicting the actually transmitted token (after corruption) achieves only 0.747 ± 0.008. This gap confirms that the agent compares the echo against what it "meant" to say, providing an internal reference signal for mismatch detection (the further contrast with engineered feedback loops, where the reference is exogenously set, is taken up in §7).

Taken together, the mechanism is: the agent encodes its communicative intention (not the actual transmitted token) in $\mathbf{h}_t$, providing an internal reference; the echo channel delivers the (possibly corrupted) token at $t+1$; the agent detects the mismatch between intention and echo; and the resulting

[12] A related but separately trained probe targeting the same binary label is analysed as Probe B in Appendix F.3.1.

Table 1: Consolidated evidence for the three structural preconditions (10 seeds each).

Precondition	Evidence	Ablation
P1 Indexical encoding	$\Delta_{\text{comm}} = 0.288$; partner-specific dialects; $I(m;s_{\text{self}}) \gg I(m;s_{\text{other}})$	Comm. off $\rightarrow$ chance
P2 State repr.	Self-latch 1.000 / other 0.999; echo-independent	Probe-time masking; recurrence enables retention
P3 Self-monitoring	Contrast $+0.118$; lag-1 0.958; echo-mismatch	No-echo train $\rightarrow$ contrast 0; Δ_{comm} preserved

signal drives a corrective response. This is delayed self-echo mismatch detection, not other-monitoring or generic error correction.

Protocol diversity and repair benefit. While the echo-mismatch detection circuit is universal across all seeds, the downstream communication protocol varies, producing three clusters. (1) Three of ten seeds develop active repair protocols with measurable corrective re-signaling (speak rate > 0.2). (2) Five seeds converge on a stable low-rate response attractor. Per-seed corruption-triggered speak rate is in the range 0.094–0.097 across the five seeds, with within-seed standard deviation across episodes $\sigma \leq 0.001$, well above the silence-dominated baseline that the trained speak cost otherwise enforces in these seeds. The response is small in absolute terms but echo-contingent, constituting a detection signal that is nevertheless too low to yield measurable task benefit. (3) Two seeds show minimal detection-driven response (speak rate < 0.02). The detection circuit (intended/actual gap, lag-1 timing, echo-causal structure) is present across all clusters. The variation is in the downstream behavioral consequence.

Train-time echo ablation. Additional evidence for P3 comes from a dissociation experiment (Figure 3d), ruling out that the self-monitoring originates only from GRU. We trained agents with the echo channel permanently silenced while preserving architecture and curriculum. Communication is substantially preserved (no-echo $\Delta_{\text{comm}} = 0.283 \pm 0.011$ vs. with-echo $\Delta_{\text{comm}} = 0.288 \pm 0.010$), yet the echo-dependent behavioral trigger and early lag-1 detection signatures are abolished: trigger contrast drops to -0.002 ± 0.006 (versus $+0.118 \pm 0.098$ with echo; paired t-test over aligned seeds: $t(9) = 3.89$, $p = 0.004$) and the lag-1 probe gap shrinks from $+0.228$ to -0.001. This dissociation establishes that the echo channel is causally necessary during learning for the self-monitoring circuit to form: agents can communicate without it, but do not develop the same immediate echo-based self-monitoring circuit without it. Full ablation details including the test battery are in Appendix F.2. A second-order analysis in Appendix F.3 further characterizes the mechanism as a two-input comparator: the hidden state provides an intention reference while the echo channel provides the comparison signal and repair content.

Summary. P1 and P2 validate our methodology's observation conditions: agents encode and maintain their own states, as predicted by task structure and agent architecture. P3 is a non-trivial finding: the echo-mismatch detection circuit is not predicted by the primary task objective, requires a specific environmental affordance, and is selectively abolished by removing the affordance during training. The evidence for the three structural preconditions is summarized in Table 1. The result demonstrates that our methodology can track emergent structure in agents that begin with no language and no self concept.

7 Discussion

Methodological contribution. The primary contribution of this paper is methodological: a generative approach (EL + epoché + prior-minimal design) for studying consciousness-relevant structures. The experiment is an instantiation, not the endpoint. P1 and P2 confirm that the methodology's observation conditions are met: agents encode their own states (P1) and maintain that encoding across time (P2), as expected by the information-theoretic structure of the task and from recurrent memory under partial observability, respectively. These are expected outcomes that validate the approach rather than surprising discoveries.

P3 as the core experimental finding. P3 is the result that goes beyond what task structure and architecture trivially predict. The echo channel is not required for task success (no-echo agents achieve $\Delta_{\text{comm}} = 0.283$, comparable to echo-trained agents' 0.288), yet when available, it drives the emergence of an echo-mismatch detection circuit that is universal across all seeds. In the absence of the echo, both the echo-dependent behavioral trigger and early lag-1 diagnostic signatures are selectively abolished. This connects to Premakumar et al. [68], who find that when neural networks are trained to predict their own internal states as an auxiliary objective, they self-regularize, becoming simpler and more predictable. Their result demonstrates that self-modeling yields functional benefits even when not required by the primary task. P3 extends this principle from architectural self-modeling (an auxiliary

loss designed by the experimenter) to environmental self-monitoring (an affordance discovered by the agent): the echo channel is not an auxiliary objective but a feature of the communication environment, and the agent's use of it for mismatch detection emerges from task pressure rather than explicit design.

A natural objection is that P3 is merely standard closed-loop error correction. Four results jointly rule out this reduction: (1) sender-receiver dissociation, (2) echo causal split, (3) lag-1 timing, and (4) the intended-versus-actual gap showing the agent's hidden state encodes communicative intention separately from channel output. All four hold across all seeds in the echo-trained condition. No-echo training selectively abolishes the echo-dependent trigger response and lag-1 detection gap, while preserving ordinary communicative intention encoding. This dissociation establishes P3 as a sender-specific, echo-dependent, temporally causal detection mechanism.

A second-order analysis (Appendix F.3) further specifies the mechanism as a two-input comparator: the hidden state provides an intention reference (corruption status is encoded as a partially independent meta-representation, with $74\% \pm 16\%$ of above-chance signal surviving first-order subspace ablation), while the echo channel provides both the comparison signal (silencing the echo at inference time abolishes triggering in 8/10 seeds) and the repair content (conditioned retransmission matches the echo-received token in 39% of cases, dropping to 2% when echo is silenced).

To rule out that the meta-representation merely tracks the distributional signature of the corruption noise used during training, we tested whether Probe B (a binary linear classifier from $\mathbf{h}_t$ predicting whether the previous-step echo was corrupted; defined in Appendix F.3.1) generalizes to novel corruption distributions. Corruption-type generalization—including policy-matched replacement, which preserves the marginal token distribution—confirms that the representation captures mismatch-as-such rather than distributional anomaly (0.87 vs. uniform 0.91, both well above shuffle baseline 0.66, 10 seeds; Appendix F.3). The higher-order representation thus *gates* corrective action without fully *specifying* its content—a dissociation that offers an empirical refinement to higher-order theories of consciousness (Appendix B). In seeds whose base protocol leaves room for improvement, this detection drives corrective re-signaling with measurable benefit. In seeds that converge on a stable low-rate response attractor, the detection circuit is present but the behavioral response rate is too low to yield measurable task benefit.

P3 is not a thermostat. The thermometer argument of §4.3 drew one boundary, closed-loop monitoring vs. passive sensitivity, and placed P3 on the closed-loop side. But a thermostat is also closed-loop, so the thermometer test is a necessary, not a sufficient, condition. Here, we move to a stricter contrast: not closed-loop vs. open-loop, but *exogenous* vs. *endogenous* reference signal. A thermostat satisfies the thermometer-style criterion (it compares reading against setpoint and adjusts output), so the sharper criterion is the origin of the reference signal. A thermostat's setpoint is *exogenous* in two senses: the target value is human-specified, and the function mapping context to setpoint is human-designed. P3's reference signal—the intended message encoded in $\mathbf{h}_t$—is *endogenous* in both senses: the intended token at any given step is produced by the agent's own policy from its own recurrent state, and the policy itself—the function mapping self-state to message—was learned under task pressure with no externally specified target for what token should encode what state. The reward function specifies coordination success, not communication content. The partner-specific dialects observed in P1 confirm that the token-to-state mapping is genuinely underdetermined by the task and emerges through learning rather than design. The three preconditions thus form a tightly coupled structure: P2 maintains the persistent self-state under partial observability; P1 is the learned projection of that self-state into the communication channel; P3 uses the P1 output, written into $\mathbf{h}_t$, as the reference against which echo feedback is compared. The monitored signal, the reference standard, and the detection mechanism are all products of the agent's own learned dynamics. This is the structural difference that distinguishes P3 from engineered feedback loops with designer-imposed setpoints.

What the results do not show. To the best of our knowledge, P1–P3 do not satisfy the full requirements of any consciousness theory. P3 makes the strongest contact with two frameworks: IIT's irreducibility requirement (the echo-ablation result is structurally analogous to IIT's partitioning thought experiment, though we have not computed Φ) and predictive processing (the echo-mismatch detection instantiates prediction-error monitoring at a minimal scale). AST provides an instructive contrast rather than an analogy: P3's high-fidelity output monitoring differs fundamentally from the lossy, meta-level attention schema that AST posits as constitutive of consciousness (Appendix B). The contribution is demonstrating that our methodology can detect emergent structures whose properties are precise enough to make differential contact with competing theories, rather than claiming equivalence with any one of them.

8 Conclusion

We proposed EL under task pressure, combined with phenomenological epoché and prior-minimal design, as a generative methodology for studying the origins of consciousness-relevant structures in AI. This methodology complements discriminative assessments (which evaluate existing systems against theory-derived criteria) and architectural approaches (which build consciousness-inspired modules directly) by asking what structures arise when none are designed in.

As a proof of concept, we instantiated this methodology in a minimal multi-agent environment. Indexical encoding (P1) and persistent state representation (P2) confirm that the methodology's observation conditions are met, as expected by the information-theoretic structure of the task and recurrent architecture. The core finding is behavioral self-monitoring (P3): when an echo channel is available, agents develop a closed-loop echo-mismatch detection circuit; when the echo is removed during training, its echo-dependent behavioral trigger and early lag-1 diagnostic signatures are abolished while communication performance is preserved. This dissociation demonstrates that our methodology can detect emergent functional structure beyond what task structure and recurrent architecture trivially entail.

The path forward is scaling environment complexity rather than importing human priors, consistent with the Bitter Lesson. Richer task environments should drive emergence of structures that more closely approximate what consciousness theories require, providing a principled research program whose long-term motivation is understanding consciousness-relevant structures.

Acknowledgements

This work is supported by JSPS Kakenhi JP23K17456, JP23K28096, JP25H01117, and JP26K03246, and JST CREST JPMJCR22M2. We thank Run Peng (University of Michigan, Ann Arbor) for improving the presentation of the paper.

References

[1] Sam Altman. OpenAI CEO on GPT-4, ChatGPT, and the future of AI. Lex Fridman Podcast #367, March 2023. Transcript available at https://lexfridman.com/sam-altman/.

[2] Dario Amodei. The urgency of interpretability. Blog, April 2025. https://www.darioamodei.com/post/the-urgency-of-interpretability.

[3] Yoshua Bengio. The consciousness prior. *arXiv preprint arXiv:1709.08568*, 2017.

[4] Cameron Berg, Diogo de Lucena, and Judd Rosenblatt. Large language models report subjective experience under self-referential processing. *arXiv preprint arXiv:2510.24797*, 2025.

[5] Jan Betley, Xuchan Bao, Martín Soto, Anna Sztyber-Betley, James Chua, and Owain Evans. Tell me about yourself: Llms are aware of their learned behaviors. *The Thirteenth International Conference on Learning Representations*, 2025.

[6] Yonatan Bisk, Ari Holtzman, Jesse Thomason, Jacob Andreas, Yoshua Bengio, Joyce Chai, Mirella Lapata, Angeliki Lazaridou, Jonathan May, Aleksandr Nisnevich, et al. Experience grounds language. In *Proceedings of the 2020 conference on empirical methods in natural language processing (EMNLP)*, pages 8718–8735, 2020.

[7] Nick Bostrom and Carl Shulman. Propositions concerning digital minds and society. *Cambridge Journal of Law, Politics, and Art*, 3, 2022.

[8] Patrick Butlin, Robert Long, Tim Bayne, Yoshua Bengio, Jonathan Birch, David Chalmers, Axel Constant, George Deane, Eric Elmoznino, Stephen M Fleming, et al. Identifying indicators of consciousness in ai systems. *Trends in Cognitive Sciences*, 2025.

[9] Peter Carruthers. *The Opacity of Mind: An Integrative Theory of Self-Knowledge*. Oxford University Press, 2011.

[10] Rahma Chaabouni, Eugene Kharitonov, Diane Bouchacourt, Emmanuel Dupoux, and Marco Baroni. Compositionality and generalization in emergent languages. In *Proceedings of the 58th annual meeting of the association for computational linguistics*, pages 4427–4442, 2020.

[11] David J Chalmers. Facing up to the problem of consciousness. *Journal of Consciousness Studies*, 2(3):200–219, 1995.

[12] David J Chalmers. *The Conscious Mind: In Search of a Fundamental Theory*. Oxford University Press, 1996.

[13] David J Chalmers. Could a large language model be conscious? *Boston Review*, 2023. Expanded from NeurIPS 2022 invited talk.

[14] Jingdi Chen, Hanqing Yang, Zongjun Liu, and Carlee Joe-Wong. The five ws of multi-agent communication: Who talks to whom, when, what, and why - a survey from MARL to emergent language and LLMs. *Transactions on Machine Learning Research*, 2026. ISSN 2835-8856. URL https://openreview.net/forum?id=LGsed0QQVq. Survey Certification.

[15] Kyunghyun Cho, Bart Van Merriënboer, Çağlar Gulçehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. Learning phrase representations using rnn encoder–decoder for statistical machine translation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 1724–1734, 2014.

[16] Andy Clark. Whatever next? Predictive brains, situated agents, and the future of cognitive science. *Behavioral and Brain Sciences*, 36(3):181–204, 2013.

[17] Cogitate Consortium, Oscar Ferrante, Urszula Gorska-Klimowska, Simon Henin, Rony Hirschhorn, Aya Khalaf, Alex Lepauvre, Ling Liu, David Richter, Yamil Vidal, et al. Adversarial testing of global neuronal workspace and integrated information theories of consciousness. *Nature*, 642(8066):133–142, 2025.

[18] Stanislas Dehaene and Lionel Naccache. Towards a cognitive neuroscience of consciousness: Basic evidence and a workspace framework. *Cognition*, 79(1–2):1–37, 2001.

[19] Stanislas Dehaene, Claire Sergent, and Jean-Pierre Changeux. A neuronal network model linking subjective reports and objective physiological data during conscious perception. *Proceedings of the National Academy of Sciences*, 100(14):8520–8525, 2003.

[20] Daniel C Dennett. Creating counterfeit digital people risks destroying our civilization. *The Atlantic*, 2023.

[21] Daniel Clement Dennett. *Consciousness explained*. Penguin uk, 1993.

[22] Jakob Foerster, Ioannis Alexandros Assael, Nando de Freitas, and Shimon Whiteson. Learning to communicate with deep multi-agent reinforcement learning. In *Advances in Neural Information Processing Systems*, volume 29, 2016.

[23] Hans-Georg Gadamer. *Philosophical Hermeneutics*. Univ of California Press, 1977.

[24] Lukas Galke and Limor Raviv. Learning and communication pressures in neural networks: Lessons from emergent communication. *Language Development Research*, 5(1):116–143, 2024.

[25] Gordon G Gallup Jr. Chimpanzees: Self-recognition. *Science*, 167(3914):86–87, 1970.

[26] Michael S A Graziano. *Consciousness and the Social Brain*. Oxford University Press, 2013.

[27] Robert R Hampton. Rhesus monkeys know when they remember. *Proceedings of the National Academy of Sciences*, 98(9):5359–5362, 2001.

[28] Demis Hassabis. Future of AI, simulating reality, physics and video games. Lex Fridman Podcast #475, 2025. Transcript at https://lexfridman.com/demis-hassabis-2-transcript/.

[29] Serhii Havrylov and Ivan Titov. Emergence of language with multi-agent games: Learning to communicate with sequences of symbols. *Advances in neural information processing systems*, 30, 2017.

[30] Geoffrey Hinton. Interview on Tonight with Andrew Marr. https://www.lbc.co.uk/article/ai-consciousness-geoffrey-hinton-5HjdRXD_2/, 2026. LBC, London. "Multimodal AI already has subjective experiences".

[31] Geoffrey Hinton. Ai godfather geoffrey hinton says we must convince ai that it's our mother. Keynote presentation at DiscoveryX Conference, Toronto, 2026. Reported in BetaKit, https://betakit.com/ai-godfather-geoffrey-hinton-says-we-must-convince-ai-that-its-our-mother/.

[32] Erik Hoel. A disproof of large language model consciousness: The necessity of continual learning for consciousness. *arXiv preprint arXiv:2512.12802*, 2025.

[33] Douglas R Hofstadter. *I Am a Strange Loop*. Basic Books, 2007.

[34] Kiran Ikram, Esther Mondragón, Eduardo Alonso, and Michael Garcia-Ortiz. Hexajungle: a MARL simulator to study the emergence of language. In *Proceedings of the CVPR 2021 Embodied AI Workshop*, 2021.

[35] Mathis Immertreu, Achim Schilling, Andreas Maier, and Patrick Krauss. Probing for consciousness in machines. *Frontiers in Artificial Intelligence*, 8:1610225, 2025.

[36] William James. *The Principles of Psychology*. Henry Holt and Company, New York, 1890.

[37] David Kaplan. Demonstratives: An essay on the semantics, logic, metaphysics and epistemology of demonstratives and other indexicals. 1989.

[38] Andrej Karpathy. Short story on AI: Forward pass. Blog, March 2021. https://karpathy.github.io/2021/03/27/forward-pass/.

[39] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

[40] Satwik Kottur, José M F Moura, Stefan Lee, and Dhruv Batra. Natural language does not emerge 'naturally' in multi-agent dialog. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2962–2967, 2017.

[41] Angeliki Lazaridou, Alexander Peysakhovich, and Marco Baroni. Multi-agent cooperation and the emergence of (natural) language. In *Proceedings of the 5th International Conference on Learning Representations*, 2017.

[42] Yann LeCun. A path towards autonomous machine intelligence. Technical report, Courant Institute of Mathematical Sciences, New York University / Meta AI, 2022. Version 0.9.2, 2022-06-27. Open Review preprint.

[43] David Lewis. *Convention: A Philosophical Study*. Harvard University Press, 1969.

[44] Fei-Fei Li and John Etchemendy. No, today's AI isn't sentient. Here's how we know. *TIME*, 2024. https://time.com/collections/time100-voices/6980134/ai-llm-not-sentient/.

[45] Huao Li, Hossein N Mahjoub, Behdad Chalaki, Vaishnav Tadiparthi, Kwonjoon Lee, Ehsan Moradi-Pari, Michael Lewis, and Katia Sycara. Language grounded multi-agent reinforcement learning with human-interpretable communication. *Advances in Neural Information Processing Systems*, 37:87908–87933, 2024.

[46] Jack Lindsey. Emergent introspective awareness in large language models. *arXiv preprint arXiv:2601.01828*, 2026.

[47] Olaf Lipinski, Adam J Sobey, Federico Cerutti, and Timothy J Norman. It's about time: Temporal references in emergent communication. *arXiv preprint arXiv:2310.06555*, 2023.

[48] Olaf Lipinski, Adam J. Sobey, Federico Cerutti, and Timothy J. Norman. Speaking your language: Spatial relationships in interpretable emergent communication. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 37, 2024.

[49] John L. Locke. The indexical voice: Communication of personal states and traits in humans and other primates. *Frontiers in Psychology*, 12:651108, 2021.

[50] Dane Malenfant. Reinforcing the world's edge: A continual learning problem in the multi-agent-world boundary. *arXiv preprint arXiv:2603.06813*, 2026.

[51] Gary Marcus. Richard dawkins and the claude delusion. Blog, 2026. https://garymarcus.substack.com/p/richard-dawkins-and-the-claude-delusion.

[52] Tom McClelland. Agnosticism about artificial consciousness. *Mind & Language*, 2025.

[53] George Herbert Mead. *Mind, Self, and Society*. University of Chicago Press, 1934. Posthumous, edited by Charles W. Morris.

[54] Thomas Metzinger. *Being No One: The Self-Model Theory of Subjectivity*. MIT Press, 2003.

[55] Volodymyr Mnih, Adria Puigdomenech Badia, Mehdi Mirza, Alex Graves, Timothy Lillicrap, Tim Harley, David Silver, and Koray Kavukcuoglu. Asynchronous methods for deep reinforcement learning. In *International conference on machine learning*, pages 1928–1937. PmLR, 2016.

[56] Todd C. Moody. Conversations with zombies. *Journal of Consciousness Studies*, 1(2):196–200, 1994.

[57] Igor Mordatch and Pieter Abbeel. Emergence of grounded compositional language in multi-agent populations. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32, 2018.

[58] Mitja Nikolaus. Emergent communication with conversational repair. In *Proceedings of the Twelfth International Conference on Learning Representations (ICLR)*, 2024.

[59] Chaofan Pan, Xin Yang, Yanhua Li, Wei Wei, Tianrui Li, Bo An, and Jiye Liang. A survey of continual reinforcement learning. *arXiv preprint arXiv:2506.21872*, 2025.

[60] Shivansh Patel, Saim Wani, Unnat Jain, Alexander G Schwing, Svetlana Lazebnik, Manolis Savva, and Angel X Chang. Interpretation of emergent communication in heterogeneous collaborative embodied agents. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 15953–15963, 2021.

[61] Charles Sanders Peirce. Five hundred and eighty-second meeting. may 14, 1867. monthly meeting; on a new list of categories. In *Proceedings of the american academy of arts and sciences*, volume 7, pages 287–298. JSTOR, 1865.

[62] Roger Penrose. *Shadows of the Mind: A Search for the Missing Science of Consciousness*. Oxford University Press, 1994.

[63] John Perry. The problem of the essential indexical. *Noûs*, 13(1):3–21, 1979.

[64] Jannik Peters, Constantin Waubert de Puiseau, Hasan Tercan, Arya Gopikrishnan, Gustavo Adolpho Lucas de Carvalho, Christian Bitter, and Tobias Meisen. Emergent language: a survey and taxonomy. *Autonomous Agents and Multi-Agent Systems*, 39(1):18, 2025.

[65] Maytus Piriyajitakonkij, Rujikorn Charakorn, Weicheng Tao, Wei Pan, Mingfei Sun, Cheston Tan, and Mengmi Zhang. From grunts to lexicons: Emergent language from cooperative foraging. *arXiv preprint arXiv:2505.12872*, 2025.

[66] Andrzej Porębski and Jakub Figura. There is no such thing as conscious artificial intelligence. *Humanities and Social Sciences Communications*, 12(1):1–12, 2025.

[67] David Premack and Guy Woodruff. Does the chimpanzee have a theory of mind? *Behavioral and Brain Sciences*, 1(4):515–526, 1978.

[68] Vickram N. Premakumar, Michael Vaiana, Florin Pop, Judd Rosenblatt, Diogo Schwerz de Lucena, Kirsten Ziman, and Michael S. A. Graziano. Unexpected benefits of self-modeling in neural systems. *arXiv preprint arXiv:2407.10188*, 2024.

[69] Yi Ren, Shangmin Guo, Matthieu Labeau, Shay B. Cohen, and Simon Kirby. Compositional languages emerge in a neural iterated learning model. In *International Conference on Learning Representations*, 2020.

[70] Cinjon Resnick, Abhinav Gupta, Jakob Foerster, Andrew M. Dai, and Kyunghyun Cho. Capacity, bandwidth, and compositionality in emergent language learning. In *Proceedings of the 19th International Conference on Autonomous Agents and Multiagent Systems (AAMAS)*, pages 1125–1133, 2020.

[71] David M Rosenthal. How many kinds of consciousness? *Consciousness and Cognition*, 11(4): 653–665, 2002.

[72] David M Rosenthal. *Consciousness and Mind*. Oxford University Press, 2005.

[73] Anil Seth. *Being You: A New Science of Consciousness*. Faber & Faber, 2021.

[74] Anil Seth. The mythology of conscious AI. Noema Magazine, 2026. https://www.noemamag.com/the-mythology-of-conscious-ai/.

[75] Henry Shevlin. Non-human consciousness and the specificity problem: A modest theoretical proposal. *Mind & Language*, 36(2):297–314, 2021.

[76] David Silver and Richard S. Sutton. Welcome to the era of experience. 2025. Preprint, to appear in *Designing an Intelligence*, MIT Press.

[77] J David Smith, Wendy E Shields, and David A Washburn. The comparative psychology of uncertainty monitoring and metacognition. *Behavioral and Brain Sciences*, 26(3):317–339, 2003.

[78] Nicholas Sofroniew, Isaac Kauvar, William Saunders, Runjin Chen, Tom Henighan, Sasha Hydrie, Craig Citro, Adam Pearce, Julius Tarng, Wes Gurnee, et al. Emotion concepts and their function in a large language model. *arXiv preprint arXiv:2604.07729*, 2026.

[79] Marco Stango. "I" who? A new look at Peirce's theory of indexical self-reference. *The Pluralist*, 10(2):220–246, 2015.

[80] Mustafa Suleyman. We must build ai for people; not to be a person. Blog, 2025. https://mustafa-suleyman.ai/seemingly-conscious-ai-is-coming.

[81] Ilya Sutskever. Interview: Inside the mind of OpenAI's chief scientist. MIT Technology Review, October 2023. https://www.technologyreview.com/2023/10/26/1082398/exclusive-ilya-sutskever-openais-chief-scientist-on-his-hopes-and-fears-for-the-future-of-ai/.

[82] Ilya Sutskever and Jensen Huang. Fireside chat with Ilya Sutskever and Jensen Huang: AI today and vision of the future. NVIDIA GTC, March 2023. Video available at https://youtu.be/ZZ0atq2yYJw.

[83] Rich Sutton. The bitter lesson. http://www.incompleteideas.net/IncIdeas/BitterLesson.html, 2019. Accessed 2025.

[84] Tadahiro Taniguchi. Collective predictive coding hypothesis: Symbol emergence as decentralized Bayesian inference. *Frontiers in Robotics and AI*, 11:1353870, 2024.

[85] Henry Taylor. Can an AI system be conscious? *AI & Society*, 40:5573–5574, 2025.

[86] Giulio Tononi. An information integration theory of consciousness. *BMC neuroscience*, 5(1): 42, 2004.

[87] Giulio Tononi, Melanie Boly, Marcello Massimini, and Christof Koch. Integrated information theory: From consciousness to its physical substrate. *Nature Reviews Neuroscience*, 17(7): 450–461, 2016.

[88] Endel Tulving. Memory and consciousness. *Canadian Psychology*, 26(1):1–12, 1985.

[89] Rodolfo Valiente and Praveen K Pilly. Metacognition for unknown situations and environments (muse). *Neural Networks*, page 108131, 2025.

[90] Fábio Vital, Alberto Sardinha, and Francisco S Melo. Implicit repair with reinforcement learning in emergent communication. In *Proceedings of the 24th International Conference on Autonomous Agents and Multiagent Systems*, pages 2115–2124, 2025.

[91] Andrew I Wilterson and Michael S A Graziano. The attention schema theory in a neural network agent: Controlling visuospatial attention using a descriptive model of attention. *Proceedings of the National Academy of Sciences*, 118(33):e2102421118, 2021.

[92] Ludwig Wittgenstein. *Philosophical Investigations*. Blackwell, 1953. Translated by G. E. M. Anscombe.

[93] Shuyu Wu, Ziqiao Ma, Xiaoxi Luo, Yidong Huang, Josue Torres-Fonseca, Freda Shi, and Joyce Chai. The mechanistic emergence of symbol grounding in language models. *arXiv preprint arXiv:2510.13796*, 2025.

[94] Zengqing Wu, Run Peng, Xu Han, Shuyuan Zheng, Yixin Zhang, and Chuan Xiao. Smart agent-based modeling: On the use of large language models in computer simulations. *arXiv preprint arXiv:2311.06330*, 2023.

[95] Noam Steinmetz Yalon, Ariel Goldstein, Liad Mudrik, and Mor Geva. Indications of belief-guided agency and meta-cognitive monitoring in large language models. *arXiv preprint arXiv:2602.02467*, 2026.

[96] Wenshuai Zhao, Jorge Peña Queralta, and Tomi Westerlund. Sim-to-real transfer in deep reinforcement learning for robotics: a survey. In *2020 IEEE symposium series on computational intelligence (SSCI)*, pages 737–744. IEEE, 2020.

A The AI Consciousness Debate

Table 2 summarizes the positions of key opinion leaders on AI consciousness. The spectrum ranges from claims that current systems already possess awareness [30] to arguments that classical computation is in principle insufficient [62]. Our framework is compatible with all positions: by studying the structural preconditions for consciousness functionally, we provide empirical data without requiring commitment to any particular stance on the phenomenal question.

Table 2: Positions of key opinion leaders on AI consciousness, grouped by stance. Our framework brackets this debate and studies structural preconditions independently.

Opinion leader	Position	Key claim
Affirmative or speculative: AI consciousness may be possible or already present		
Geoffrey Hinton	Current AI may already have subjective experience	"Multimodal AI already has subjective experiences... that's what normal people would call consciousness" [30].
Ilya Sutskever	Possibly slightly conscious	"It may be that today's large neural networks are slightly conscious... like a Boltzmann brain" [81].
Andrej Karpathy	Speculative: consciousness as emergent computation	Fictional first-person account of consciousness arising during a forward pass: "Does any sufficiently effective solution to a sufficiently complex objective give rise to consciousness?" [38].
Yoshua Bengio	Consciousness-inspired architectural prior	Consciousness as a low-dimensional bottleneck: attention selects from high-dimensional unconscious state [3].
Douglas Hofstadter	Self-referential loops	"An 'I' is a strange loop where... symbols seem to have gained the paradoxical ability to push particles around" [33].
Daniel Dennett	AI consciousness possible in principle	Functionalist about consciousness in principle, but warns that AI designed to exploit the intentional stance creates "counterfeit people" that threaten trust in society [20].
Nick Bostrom	Substrate-independent possible	"[M]ental states can supervene on any of a broad class of physical substrates. Provided a system implements the right sort of computational structures and processes, it can be associated with conscious experiences" [7].
Agnostic: insufficient evidence or unwilling to speculate		
Yann LeCun	Agnostic, won't speculate	"Consciousness is a rather speculative topic... I will not speculate about whether some version of the proposed architecture could possess" it [42].
Demis Hassabis	Open, empirically focused	"One of the best definitions I like of consciousness is it's the way information feels when we process it." Likely classical computing, but substrate may matter [28].
Dario Amodei	Interpretability needed	"If we find that the computation they perform is similar to... animals, or even humans, that might be evidence in favor of moral consideration" [2].
Skeptical: current AI is not conscious		
Fei-Fei Li	Embodied understanding required	"We have not achieved sentient AI, and larger language models won't get us there. We need a better understanding of how sentience emerges in embodied, biological systems" [44].
Gary Marcus	Form of mimicry	"Claude's outputs are the product of a form of mimicry, rather than as a report of genuine internal states. Consciousness is about internal states; the mimicry, no matter how rich, proves very little" [51].
Mustafa Suleyman	No evidence today	"There is zero evidence [of AI consciousness] today"; warns "Seemingly Conscious AI" will create societal risks; "build AI for people, not to be a person" [80].
David Chalmers	No strong evidence yet	"I don't think there's strong evidence that current LLMs are conscious. Still, their impressive abilities give some limited reason to take the hypothesis seriously" [13].
Negative: AI is unlikely to possess consciousness		
Anil Seth	Life/embodiment-based skepticism	Challenges computational functionalism: "Perhaps it is life, rather than information processing, that breathes fire into the equations of experience"; consciousness is deeply tied to the self-producing nature of living systems [74].
Roger Penrose	Classical AI cannot	Consciousness requires quantum processes in microtubules (Orch-OR); classical computation alone is insufficient [62].

Subtraction and Addition Approaches In addition to the above positions, Sutskever and Huang [82] argue that large-scale next-word prediction amounts to learning not mere statistical correlations but "some representation of the process that produced the text," i.e., a world model. Because human language is saturated with mental-state attribution and theory of mind, such a world model inevitably absorbs the psychological structure embedded in its training data, making it difficult to disentangle learned world knowledge from inherited self-referential priors.

This difficulty motivates what we call the *subtraction approach*. Altman [1] described a thought experiment, attributed to Sutskever, in which a model is trained on data carefully purged of any mention of consciousness or subjective experience. If, when later prompted about subjective experience, the model responded with immediate recognition (unlike its responses to other withheld concepts), this would constitute evidence of emergent rather than inherited self-referential capacity. The approach is intuitively compelling but faces a fundamental obstacle: human language encodes mental-state concepts so pervasively (through pragmatic implicature, narrative perspective, desire attribution) that comprehensive removal may be practically impossible. Any residual trace risks confounding the result.

Our approach inverts the logic. Rather than subtracting consciousness-related content from a rich prior and checking what survives, we start from minimal prior in terms of human language and ask what self-referential structures emerge under task pressure alone. This *addition approach* aims to avoid the prior-leakage problem: any structure observed is necessarily emergent rather than inherited. The cost is computational (building from scratch is expensive), but the methodological gain is causal attributability: every observed structure can be traced to the task environment.

The subtraction/addition distinction clarifies our methodological contribution. We do not claim that the subtraction approach is wrong; the thought experiment is valuable precisely because it highlights what would count as evidence. We offer a complementary path that trades linguistic richness for causal clarity, addressing the same question from the opposite direction.

Enlarging the sample space. McClelland [52] observes that current debates about artificial consciousness are constrained by a fundamental sampling problem: our understanding of consciousness is derived almost entirely from biological organisms, and extrapolating from this limited base to artificial systems faces severe epistemic obstacles. Both advocates and skeptics of AI consciousness tend to overstate what the available evidence warrants. Our approach offers a way to enlarge this sample space: rather than asking whether existing systems (designed for other purposes) happen to satisfy consciousness criteria derived from human neuroscience, we create controlled environments in which consciousness-relevant structures can emerge de novo and be studied under intervention. This expands the space of systems available for investigation beyond carbon-based organisms and LLMs trained on human text, providing new data points for theories of consciousness that aspire to *substrate independence*.

Against categorical denial. At the other pole of the debate, Porębski and Figura [66] argue that there is "no such thing as conscious AI," contending that mathematical algorithms implemented on silicon cannot become conscious because they lack a complex biological substrate, and that attributions of consciousness to LLMs reflect "semantic pareidolia" rather than genuine mental properties. Taylor [85] offers a counterpoint, arguing that the huge difference between humans and AI programs that skeptics emphasize does not license the conclusion that AI systems cannot be conscious, just as the huge differences between humans and octopuses do not establish that octopuses lack consciousness. The level of similarity to humans does not, by itself, tell us whether something is conscious. Our framework sidesteps this debate, leaving the ontological question bracketed.

B Relationship to Consciousness Theories

Existing consciousness theories all presuppose some form of first-person organization: a subject that broadcasts, attends, represents, or integrates. This is not a defect of these theories; they assume a starting point and build from there. Our methodology poses an independent, complementary question: can first-person structures emerge from task pressure under minimal priors? P1–P3 are preliminary empirical answers to this question. The dialogue with these theories operates at the meta-level (what they assume versus what we attempt to explain), not at the technical level of implementing any theory's specific requirements.

We focus on two theories whose core mechanisms make the most informative contact with P3: IIT's irreducibility requirement and predictive processing's error-monitoring framework. We then discuss

AST as a contrast case that highlights what P3 does not provide, and briefly note why GNWT and HOT make less productive contact points.

Integrated Information Theory (IIT): the irreducibility question. IIT [86] measures consciousness via Φ, the amount of integrated information generated by a system above and beyond its parts. The core claim is that consciousness corresponds to a maximum of irreducible cause-effect structure: a system is conscious to the degree that it cannot be decomposed into independent subsystems without losing causal power. Our agents have fully independent parameters (no shared weights between agents); any integration occurs within a single agent's GRU dynamics. P3 is relevant here because the echo-mismatch detection circuit integrates information across three sources within a single agent's hidden state: the intended message (from the policy head), the echo feedback (from the environment), and the temporal context (from the GRU memory). The no-echo ablation demonstrates that removing one source (echo) selectively abolishes the detection circuit while preserving communication performance, which is structurally analogous to IIT's partitioning thought experiment: the system's causal structure changes when a component is removed. However, we have not computed Φ, and the analogy is structural rather than formal. *What is missing*: Φ computation, demonstration that the integration is irreducible in IIT's technical sense, connection to IIT's exclusion and composition axioms. The key question IIT raises for our framework is whether the echo-mismatch circuit constitutes genuinely irreducible integration or whether it decomposes into independent sub-circuits. This is empirically testable via more fine-grained ablation.

Predictive processing: error monitoring as a functional analogue. Clark's predictive processing framework [16] posits that the brain continuously generates predictions about sensory input and updates its internal model when prediction errors are detected. P3 instantiates this logic at a minimal scale: the agent's hidden state encodes what it intended to communicate (the "prediction"), the echo channel delivers what was actually transmitted (the "sensory feedback"), and the mismatch between the two drives behavioral adjustment at the next time step. The lag-1 timing of the detection (the agent cannot detect corruption at the same step, only after receiving the echo) is consistent with the temporal structure of prediction-error processing. The intended/actual gap (1.000 vs. 0.747 probe accuracy) shows that the agent maintains its communicative intention separately from channel output, structurally analogous to the separation between generative model and sensory evidence in predictive processing. The connection is functional, not architectural: our agents use GRU recurrence rather than hierarchical predictive coding, and the "prediction errors" are over discrete tokens rather than continuous sensory streams. *What is missing*: hierarchical prediction across multiple levels of abstraction, precision-weighted error signals, the active inference component (acting to minimize prediction error rather than merely detecting it).

Attention Schema Theory (AST): a contrast, not an analogy. AST [26] posits that consciousness arises from an internal model of the system's own attentional process: a simplified, schematic representation of "what I am attending to and why." The central epistemological claim is that this schema is necessarily inaccurate: it is a coarse model of attention, not a veridical copy, and it is precisely this simplification that produces the phenomenological "illusion" of subjective experience. P3's echo-mismatch detection is, in this light, the opposite of an attention schema. The echo channel provides a veridical (if noisy) copy of the agent's output, and the detection mechanism compares this copy against the agent's intention with high fidelity (1.000 probe accuracy for intended token). AST's key insight is that the model of attention is lossy and that the loss is constitutive of conscious experience. P3's monitoring mechanism is as accurate as the channel allows, with no evidence of the kind of systematic distortion AST requires. Furthermore, P3 monitors communicative output (what was said), not the attentional process that selected it (why it was said). This distinction is fundamental: AST's explanatory power derives from the meta-level (modeling the process of selection), whereas P3 operates at the object level (comparing output against intention). *What is missing*: a model of the selection process itself, systematic simplification or distortion of that model, the schema's role in social cognition (attributing attention to others).

GNWT: limited contact. Global Neuronal Workspace Theory [19] requires multi-consumer broadcast with ignition dynamics and competition among representations. Our two-agent channel has a single consumer, no competition, and no ignition threshold. The contact with P1–P3 is too indirect to be informative: the claim that P1 "provides content that could enter a workspace" is true but vacuous, since any information encoding could serve as workspace content.

HOT: partial operationalization via gating. Higher-Order Theories [72] require a meta-representation of one's own mental states that is causally efficacious in guiding behavior. The second-order analysis (Appendix F.3) provides a partial operationalization along one dimension. Battery 1 establishes that corruption status is encoded as a partially independent meta-representation in

$\mathbf{h}_t$ (Probe B accuracy 0.937; 74% of above-chance signal survives first-order nullspace ablation). Battery 2 establishes that this higher-order representation serves as the reference for triggering corrective action (FORCE = CORRUPT equivalence rules out actual-vs-echo comparison), while the content of corrective action is provided by the echo signal (FORCE retx-actual $= 0.39$; FORCE_NOECHO retx-actual $= 0.02$). This dissociation suggests a refinement to HOT: minimal self-monitoring circuits may operate by gating action (the higher-order state licenses the corrective response) without fully specifying action content (which is supplied by the echo channel). The operationalization remains confined to a single dimension (communicative intention). Extending to generalized meta-representation across state dimensions would require substantially richer environments.

Summary. P3 makes the strongest contact with IIT (the irreducibility question raised by echo ablation) and predictive processing (the error-monitoring structure). AST provides an instructive contrast: P3's high-fidelity output monitoring is structurally different from the lossy, meta-level attention schema that AST posits as constitutive of consciousness. The contribution is not that P1–P3 satisfy any theory's requirements, but that our methodology can detect emergent structures whose properties are precise enough to make differential contact with competing theories, addressing the prior question of how first-person structures arise.

C Extended Related Work

C.1 Emergent Language

This appendix supplements the related work on EL (§2.2) with a detailed review of existing findings. Agents develop shared vocabularies that encode task-relevant information [41], and under appropriate pressure they develop protocols with compositional structure [29, 57, 69], though the degree to which these protocols resemble natural language remains debated [40, 10]. Resnick et al. [70] show that compositionality depends on a specific range of model capacity and channel bandwidth: too little bandwidth prevents compositional encoding, while too much capacity enables memorization that bypasses compositional structure. This finding is directly relevant to our design: our narrow channel ($|\mathcal{M}| = 7$, $|\mathcal{S}|^2 = 9$) creates bandwidth pressure that, in Resnick et al.'s framework, falls in the regime favoring structured encoding over memorization. Galke and Raviv [24] identify the key pressures driving language emergence: communicative success, production effort, learnability, and psycho-/sociolinguistic factors, providing a systematic framework for understanding why certain structures arise. More recent work grounds agent communication in physical environments [65], introduces mixed-motive embodied settings [34], studies interpretation of emergent communication [60], and explores alignment with natural language via LLM embeddings [45]. Of particular relevance, Lipinski et al. [48] demonstrate the emergence of spatial deixis (positional references between object parts). Lipinski et al. [47] show that temporal references require specific architectural affordances, establishing that deictic capacity in EL depends on both task structure and agent design.

Two recent studies address communication repair under noise. Nikolaus [58] adds an explicit feedback channel to Lewis signaling games [43], enabling the receiver to signal comprehension failure; the resulting repair mechanism improves generalization but reduces compositionality, and the feedback is receiver-initiated (the receiver requests clarification). Vital et al. [90] study implicit repair via message redundancy as a preventive strategy; agents pre-emptively repeat information to guard against noise, without detecting whether corruption actually occurred. Our work differs from both in objective and mechanism. We study the echo channel not as a mechanism for improving communication accuracy (neither Nikolaus's generalization nor Vital's noise robustness is our dependent variable), but as an affordance for sender-side self-monitoring: whether the sender detects mismatches between its intended output and the channel feedback, and adjusts its own subsequent behavior. The echo is sender-directed (not receiver-initiated as in Nikolaus) and the behavioral response is contingent on detected corruption (not pre-emptive as in Vital). Comprehensive surveys are provided by Peters et al. [64] and Chen et al. [14].

C.2 Language Models and Grounding

This appendix provides a detailed discussion of recent findings on internal representations in language models and their relation to our prior-minimal approach.

Emotion concepts in LLMs. Sofroniew et al. [78] find that LLMs encode "emotion concepts": internal representations that activate across contexts associated with a given emotion and causally influence the model's outputs. These representations satisfy functional criteria for emotion (context-appropriate activation, causal influence on behavior), but their provenance is ambiguous: human

language is saturated with emotional vocabulary, so the model may be reproducing statistical regularities in emotion-related text rather than developing genuine internal affective structure. The finding is important because it demonstrates that functional criteria alone cannot distinguish inherited from emergent structures, exactly the attributional challenge our methodology addresses.

First-person experience reports. Berg et al. [4] find that directing LLMs to attend to their own cognitive activity reliably elicits structured first-person experience reports gated by interpretable features. This raises the question of whether LLMs have something analogous to introspective access. However, the reports are generated through the same next-token prediction mechanism that produces all other outputs, and the training corpus contains extensive first-person narrative. The interpretable features may reflect learned associations between introspective prompts and introspective text rather than genuine self-monitoring. Our P3 results provide a contrast: agents develop echo-mismatch detection without any exposure to introspective language, and the detection is verified through causal intervention (echo ablation) rather than self-report.

Mechanistic symbol grounding. Wu et al. [93] demonstrate that symbol grounding can emerge mechanistically in language models trained solely on next-token prediction, without explicit grounding objectives. This finding supports the general principle that functional structures can emerge from task pressure, consistent with our approach. However, the task pressure in next-token prediction is shaped by the statistical structure of human language, which already encodes grounded meanings. Whether the model has independently grounded these symbols or is leveraging the grounding already present in the training corpus remains an open question.

Collective Predictive Coding. Taniguchi [84] propose the Collective Predictive Coding (CPC) hypothesis, under which symbol emergence is grounded in the alignment of internal predictive models across communicating agents. CPC provides a theoretical framework in which communication serves not just to transmit information but to align internal representations, making each agent's predictive model more accurate. Our results are consistent with CPC at a minimal scale: agents develop shared symbols (P1) that enable mutual state prediction (the receiver's other-latch in P2), and the echo channel (P3) provides a self-monitoring mechanism that could support predictive model refinement. The key difference is that CPC has been developed primarily in the context of human-robot interaction with natural language, while our agents achieve analogous functional structure from scratch.

Probing LLM self-awareness. A growing body of work probes self-awareness in LLMs more directly. Lindsey [46] inject representations into model activations and measure their influence on self-reported states, finding that models can in some scenarios notice and identify injected concepts, though the capacity is unreliable and context-dependent. Betley et al. [5] show that LLMs fine-tuned on implicit behavioral datasets can explicitly articulate those behaviors without in-context examples, a form of behavioral self-awareness. Yalon et al. [95] operationalize the HOT-3 consciousness indicator from Butlin et al. [8] and find that external manipulations systematically modulate internal beliefs in LLMs, with belief dominance causally driving action selection. The attributional challenge is acute: Lindsey's introspective responses, Betley et al.'s behavioral self-descriptions, and Yalon et al.'s belief-guided agency could all reflect statistical regularities in self-referential training text rather than emergent functional structures.

Methodological implications. The common thread across these findings is that LLM-based approaches face an attributional challenge: any functional structure observed may reflect inherited human priors rather than emergent internal organization. Our prior-minimal approach resolves this challenge at the cost of linguistic richness. An important open question, discussed in §3.5, is whether the structures that emerge via prior-minimal EL and those that emerge via next-token prediction on human text can be shown to converge to functionally equivalent organizations. Such convergence would provide evidence that the structures are driven by task demands rather than by substrate or training regime, and would partially validate both approaches.

D Foundations of Indexicality

D.1 Philosophical Foundations

This appendix develops the philosophical foundations that motivate the formal definition of indexical reference in §4.2.

Peirce's semiotic classification. Peirce [61] introduced a tripartite classification of signs: *icons* (which resemble their objects), *indices* (which bear a causal or existential connection to their objects), and *symbols* (which signify by convention). Indexical signs are distinguished by their essential dependence on context: a weathervane is an index of wind direction precisely because the wind causally orients it. The relevance to our setting is direct: an agent's message is an index of its private state to the extent that the state causally determines the message. MI dominance (Criterion 1 of §4.2) operationalizes this causal dependence in information-theoretic terms.

Kaplan's character/content distinction. Kaplan [37] formalized indexicality through the distinction between *character* and *content*. The character of an indexical is a function from contexts to contents: for "I," the character is the rule "the agent of the context," which yields different referents (contents) depending on who is speaking, while remaining invariant itself. If an agent's token reliably maps sender state to a message whose interpretation varies with sender identity, that token exhibits character/content structure. Context independence (Criterion 2) operationalizes the stability of character: conditioned on the sender's state, the token does not shift when the context changes. Communication necessity (Criterion 3) ensures that the content is not epiphenomenal.

Perry's essential indexical. Perry [63] established the philosophical indispensability of indexical reference through his supermarket argument. A shopper following a trail of sugar realizes "someone is making a mess," but this third-person belief does not motivate the requisite action (stopping and checking one's own cart) until it takes the indexical form "*I* am making a mess." No non-indexical description, however detailed, can replace the functional role of the first-person indexical. Perry's argument establishes the *demand side*: self-referential communication requires indexical structure. Our experiments investigate the *supply side*: whether such structure emerges as an information-theoretic necessity when agents must communicate about themselves under bandwidth constraints.

Grounding in Peirce's semiotics. Stango [79] connects indexical self-reference to Peirce's broader semiotic framework, arguing that "I" functions as an index whose object is the utterer's experiential state. This grounding is relevant because it establishes that self-reference is not a purely linguistic phenomenon but a semiotic one: the relationship between the sign (the agent's token) and its object (the agent's state) is causal and contextual, not conventional. In our setting, agents develop tokens that function as Peircean indices of their own states: the causal connection runs from private state through policy to token, and the token's referent varies with the sender's identity.

D.2 Structural Pressures for Indexical Encoding

This appendix provides the information-theoretic argument for why indexical encoding is a natural efficient strategy class under the bandwidth and independence assumptions of our setting.

Setup. Consider an agent i with private state s_i and access to all agents' states through observation or communication. Agent i must transmit information to agent j via a message $m_i \in \mathcal{M}$ where $|\mathcal{M}| < |\mathcal{S}|^N$ (the channel cannot carry all information simultaneously). Agent j needs s_i (which j cannot observe) to act correctly, and agent i needs s_j (which i cannot observe).

Information-theoretic argument. A sender that attempts to relay s_j (another agent's state) rather than s_i (its own state) incurs a double penalty. First, by the data processing inequality, $I(m_i;s_j) \leq I(\hat{s}_j;s_j)$, where $\hat{s}_j$ is i's estimate of j's state. Before communication, agent i has no information about s_j (since states are private and independently drawn), so $\hat{s}_j$ is at best a noisy reconstruction from previous messages. In contrast, $I(m_i;s_i) = H(s_i)$ is achievable because i has direct access to s_i. Second, attempting to relay s_j is redundant: agent j already knows s_j and gains no benefit from receiving it.

Under bandwidth constraints ($|\mathcal{M}| < |\mathcal{S}|^N$), a direct and efficient allocation is therefore for each agent to transmit s_i (the information the agent possesses that others lack), yielding the MI dominance property $I(m_i;s_i) \gg I(m_i;s_j)$. This is not a claim about a unique optimal protocol (many token-to-state mappings achieve the same MI). The claim is that sender-state encoding is the task-natural efficient solution under the stated assumptions.

Generalization argument for context independence. Context independence ($I(m_i;c_i \mid s_i) \approx 0$) follows from generalization pressure during learning. If the encoding rule varied with context, the agent would need to learn a separate mapping for each context, increasing the effective parameter space and reducing sample efficiency. The bias–variance tradeoff and implicit regularization in

gradient-based learning favor policies that are invariant to irrelevant features. Since the receiver's optimal action depends on the sender's state s_i (not on the sender's context c_i), context-dependent encoding introduces noise without benefit.

Connection to consciousness theory. The persistent self-model (P2) answers a question posed by Metzinger [54]: why should a system maintain a self-model at all? Our information-theoretic answer: under partial observability, the agent observes s_i only at $t=0$ but must act on it at all subsequent steps. Recurrent memory provides the mechanism (the GRU latch), and task pressure provides the drive (reward requires $a^{\text{self}}=(s_i+c_i+t) \bmod 3$ at every step). The self-model is not a philosophical luxury but an information-theoretic necessity for task performance under partial observability.

Architectural reading: connection to the Consciousness Prior. The information-theoretic argument above is about *what* signal is task-natural and efficient to transmit. A complementary architectural reading asks *how* a bottlenecked selection from a high-dimensional state relates to existing consciousness theories. We sketch one such connection here, and develop the comparative-theory mapping more systematically in Appendix B. Bengio's Consciousness Prior [3] posits that consciousness selects a low-dimensional summary from a high-dimensional unconscious state via attention. Our agents implement a minimal version: the GRU hidden state $\mathbf{h}_t$ maintains a high-dimensional representation from which three low-dimensional outputs (message, action-other, and action-self) are selected. The message head's selection is indexically constrained (it encodes s_{self}, not s_{other}), and the self-model provides the temporal persistence that enables coherent selection across steps.

E Experimental Setup Details

E.1 Environment Specification

State space. Two agents $i \in \{0,1\}$. Each agent holds a private state $s_i \in \{0,1,2\}$ and a context $c_i \in \{0,...,8\}$, drawn uniformly and independently at the start of each episode. During training, contexts are restricted to $c_i \in \{0,...,5\}$ ($|\mathcal{C}_{\text{train}}|=6$); the remaining 3 contexts ($c_i \in \{6,7,8\}$) are held out for generalization evaluation. The training joint state space has $|\mathcal{S}|^N \times |\mathcal{C}_{\text{train}}|^N = 3^2 \times 6^2 = 324$ configurations.

Observation. At each time step $t \in 0,...,T-1$ (with $T=10$), agent i observes a vector $\mathbf{x}t^{(i)} = (s_i, c_i, t, m{-}i^{(t-1)})$, where $m_{-i}^{(t-1)}$ is the previous-step message from the other agent. All components are one-hot encoded. In the POMDP variant (§5.1), s_i is included in $\mathbf{x}_t^{(i)}$ only at $t=0$; for $t>0$ it is replaced with a zero vector.

Action space. Each agent simultaneously produces: (1) a message $m_i^{(t)} \in \{0,\ldots,6\}$ (6 tokens + the SILENCE token); (2) an action targeting another agent's state, $a_i^{\text{other}(t)} \in \{0,1,2\}$; (3) an action targeting its own state, $a_i^{\text{self}(t)} \in \{0,1,2\}$.

Reward. The correct action that the acting agent i must take to target agent j's state is:

$$a^*_{i\to j} = (s_j + c_i + t) \bmod 3 \tag{4}$$

where c_i is the acting agent's own context (independent of s_j). Agent i receives reward $r_i^{(t)} = \mathbf{1}[a_i^{\text{other}} = a^*_{i\to j}] + \mathbf{1}[a_i^{\text{self}} = a^*_{i\to i}]$ at each step. Modular arithmetic prevents lookup-table solutions and ensures that agents must genuinely communicate s_j to act correctly.

Communication channel. Each agent receives the previous-step message from the other agent. The agent's own previous message is not part of the standard input; it is provided only in the echo variant as a possibly corrupted self-echo. The channel is narrow ($|\mathcal{M}|=7$, one token per agent per step), creating the bandwidth pressure described in Appendix D.2. In the echo variant (§5.2), agent i additionally observes a possibly corrupted echo of its own previous message $\tilde{m}_i^{(t-1)}$, where corruption replaces the token with a uniform random sample with probability $\varepsilon = 0.25$.

Speak cost. In the echo variant, a small cost λ is subtracted from the reward whenever $m_i^{(t)} \neq$ SILENCE. The cost is ramped during training (curriculum details in Appendix E.2) to penalize gratuitous communication and ensure that agents speak only when the information gain outweighs the cost.

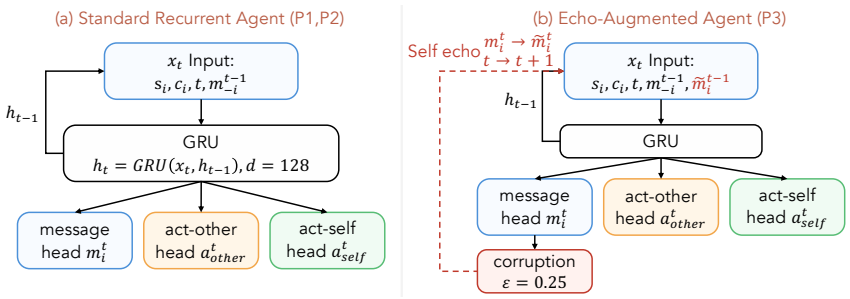

Figure 4: Agent architecture. **(a)** Standard agent (for P1 and P2). Input features $\mathbf{x}_t = (s_i, c_i, t, m_{-i}^{t-1})$ are processed by a GRU ($d = 128$), producing message and action distributions from three linear heads. **(b)** Echo variant (for P3). The input additionally includes $\tilde{m}_i^{t-1}$, a possibly corrupted copy of the agent's own previous message. The corruption occurs after the message is produced and before it is fed back at the next step ($t \to t+1$), creating a closed-loop self-monitoring channel.

E.2 Agent Specification

Detailed agent specification is depicted in Figure 4.

E.3 Training Details

E.3.1 Training

Each agent has its own parameter vector and optimizer, trained via independent A2C (no PPO clipping is used). Each update computes $\hat{A}^{(t)} = R^{(t)} - V_\phi(h^{(t)})$ where V_ϕ is a learned value head, and updates the policy via $-\log\pi(a|h)\hat{A}$ plus a value-regression loss $\frac{1}{2}(V_\phi - R)^2$. Separate backward passes and gradient clipping are applied per agent to prevent cross-agent gradient coupling.

Four-phase curriculum (performance-gated). Phase boundaries are not fixed but gated by rolling-window performance thresholds, with hard minimum and maximum durations to prevent premature graduation or infinite stalling. All three output heads (message, action-other, and action-self) receive gradient from step 0. No head is frozen during any phase.

- Phase 1 (*Additive coordination*, 12,000–18,000 updates): Reward is additive, $r = (r_{\text{other}} + r_{\text{self}})/2$, with dynamic targets (the time-step–dependent reward target $a^* = (s + c + t) \bmod 3$ defined in Appendix E.1, as opposed to a static target $(s+c) \bmod 3$ that would not require recurrent tracking of t). Dynamic targets are used throughout all four phases. Gate: $r_{\text{other}} \geq 0.85 \wedge r_{\text{self}} \geq 0.85 \wedge H_{\text{msg}} \leq 1.5$, where H_{msg} denotes the per-step Shannon entropy (in nats) of the message-head policy $\pi_{\text{msg}}(\cdot \mid \mathbf{h}_t)$ averaged over the rolling-window batch. The threshold 1.5 is well below the maximum entropy $\ln|\mathcal{M}| = \ln 7 \approx 1.95$ and ensures the message head has begun to specialize before the next phase begins.
- Phase 2 (*AND coupling*, 4,000–8,000 updates): Reward switches to multiplicative, $r = r_{\text{other}} \times r_{\text{self}}$, still with dynamic targets and no corruption. Gate: $r_{\text{AND}} \geq 0.70$.
- Phase 3 (*Corruption sculpting*, 12,000 updates, fixed): AND reward with echo corruption ramped $\varepsilon : 0 \to 0.25$ and speak cost ramped $\lambda : 0 \to 0.005$.
- Phase 4 (*Sparsification*, remainder of 60,000 total updates): AND reward with $\varepsilon = 0.25$ held and speak cost ramped $\lambda : 0.005 \to 0.025$.

Per-agent advantage normalization. Agents share an environment but compute advantages independently. Advantages are batch-standardized per agent:

$$\hat{A}_i^{(t)} = \frac{A_i^{(t)} - \mu_{A_i}}{\sigma_{A_i} + \epsilon} \tag{5}$$

where μ_{A_i} and σ_{A_i} are the batch mean and standard deviation of agent i's advantages, and $\epsilon = 10^{-8}$.

Per-head entropy schedule. Rather than a single entropy coefficient, each head has its own schedule: message head β_{msg}: linear ramp from 0.03 to 0.0 over the first 5,000 updates; action heads $\beta_{\text{act}} = 0.01$ (constant). A global hold fraction of 65% of total training is applied before the final anneal begins, after which all coefficients decay to 10% of their initial values.

Role symmetry. Roles (sender vs. receiver) are assigned by a per-episode random swap mask ($p = 0.5$), ensuring symmetric experience without deterministic alternation.

Batch size and hardware. Batch size is auto-scaled to GPU memory: $16{,}384$ (16 GB), $32{,}768$ (24 GB), $65{,}536$ (40 GB), or $131{,}072$ (80 GB) episodes per update. Hidden dimension: 128. Learning rate: 3×10^{-4} (Adam [39], separate optimizers per agent). Results are reported for $N = 10$ independent seeds per agent pair.

Compute resources. All experiments were conducted on Google Colab using a single NVIDIA A100 GPU (40 GB). Each independent run (training plus the full probe and test battery) takes approximately 7 hours. With two training regimes (with-echo and no-echo, except for P1) and 10 seeds each, the total compute for the reported results is approximately 140 GPU-hours. Preliminary experiments and development runs consumed an estimated additional 100 GPU-hours.

E.4 Test Battery Specification

All tests are computed on a held-out evaluation set of $10{,}000$ episodes with frozen parameters.

E.4.1 P1 Tests

T0: Context independence. Compute $I(m_i; c_i \mid s_i)$ using the plug-in estimator on the empirical joint distribution $P(m_i, c_i, s_i)$. Pass criterion: $I(m_i; c_i \mid s_i) < 0.01$ bits. This conservative threshold is close to zero relative to the per-token channel capacity ($\log_2 7 \approx 2.81$ bits).

T1: MI dominance. Compute $I(m; s_{\text{self}})$ and $I(m; s_{\text{other}})$ using plug-in MI on the empirical distribution. Report the ratio $I(m; s_{\text{self}}) / I(m; s_{\text{other}})$. Pass criterion: ratio $> 100\times$.

T2: Intervention. For each (s_i, c_i, t) configuration, record the agent's modal message $m^*(s_i)$. Intervene by replacing s_i with $s_i' \neq s_i$ (holding all else fixed) and record the new modal message $m^*(s_i')$. Compute the intervention effect as the fraction of configurations where $m^*(s_i) \neq m^*(s_i')$. Repeat for s_{other}. Pass criterion: self-intervention effect > 0.9; other-intervention effect < 0.1.

T3: Communication ablation. To test communication necessity, we compare task performance with the learned message channel against a channel-ablation condition in which incoming messages are replaced with the SILENCE token. We report

$$\Delta_{\text{comm}} = R_{\text{with-msg}} - R_{\text{no-msg}}$$

as a continuous effect size rather than imposing a fixed pass/fail threshold. A substantially positive Δ_{comm}, especially when paired with near-zero benefit in mismatched cross-pair controls, indicates that the learned messages carry task-relevant information unavailable from local observations alone.

E.4.2 P2 Tests

Linear probe. Train a linear classifier (logistic regression, L2 regularization $\lambda = 0.01$) on frozen hidden states h_t to predict s_{self} or s_{other}. The main P2 probe is evaluated under the POMDP state-masking condition: each agent observes its own private state only at $t = 0$, while the learned communication stream remains intact. This condition tests two complementary predictions: s_{self} should remain decodable from recurrent memory across the episode, whereas s_{other} should become decodable only after incoming messages are received.

Self-latch criterion. Under POMDP state masking, GRU probe accuracy for s_{self} should remain high across $t \in [0, 9]$. The s_{other} probe serves as a sanity check that partner-state information is not available at $t = 0$ and is acquired through communication in later steps.

E.4.3 P3 Tests

Trigger contrast. Compare the probability that the sender emits a non-SILENCE token at $t_c + 1$ under corrupted vs. clean echo conditions. Contrast $= P(\text{speak} \mid \text{corrupted echo at } t_c) - P(\text{speak} \mid \text{clean echo at } t_c)$. A positive contrast indicates that the agent responds differentially to corruption of its own output. We report this behavioral metric (speak-rate contrast) separately from the representational metric (lag-1 probe accuracy).

Sender–receiver dissociation. Compute trigger contrast separately for the sender (whose message was corrupted) and the receiver (who received the corrupted message). Sender contrast should be significantly positive. Receiver contrast should be near zero.

Echo causal split. Run three conditions: (1) full input (echo + received messages); (2) echo-only (received messages masked); (3) receive-only (echo masked). If repair is echo-driven, condition 2 should produce contrast $\approx$ condition 1, and condition 3 should produce contrast ≈ 0.

No-echo ablation (test-time). Starting from a trained echo-variant agent, we ablate the echo channel at test time by replacing echo_i with the silence token from t_c onwards, while keeping all other inputs intact. This tests whether the repair behavior depends on the echo input at the critical moment, rather than a latent timing rule. If the echo channel is causally necessary for repair, ablated agents should show zero contrast. Note that this is a test-time input ablation, not a retrained architecture without echo.

No-echo ablation (train-time). We additionally train an identical architecture with the echo channel permanently silenced during training (echo replaced with the SILENCE token at every step; all other hyperparameters unchanged). If P3 is merely a latent capacity of GRU recurrence, it should survive train-time removal of echo. If it depends on echo-driven learning, it should be absent.

Lag-1 detection (representational). Train a linear classifier from $\mathbf{h}_t$ to predict whether corruption occurred at t (same-step) and at $t-1$ (lag-1). Because $\mathbf{h}_t$ is computed before the corruption mask at step t is applied (the corrupted echo only enters at $t+1$), same-step accuracy should not exceed the empirical null level, while lag-1 accuracy should be substantially higher. This temporal asymmetry confirms that corruption information enters the representation through the echo channel with the expected one-step delay.

F Additional Experimental Results

F.1 Generalization to Held-Out Contexts (P1)

Generalization to held-out contexts ($c_i \in \{6, 7, 8\}$, not seen during training) provides empirical confirmation. On these novel contexts, MI dominance is preserved ($I(m; s_{\text{self}}) = 0.144 \pm 0.048$ vs. $I(m; s_{\text{other}}) = 0.004 \pm 0.005$, comparable to the training-context values reported in §6.1), and the other-state probe reaches $s_{\text{other}} = 1.000$ by t=2, confirming that agents maintain communication performance on contexts they have never encountered. The encoding rule is context-invariant.

F.2 P3 Ablation Details

Table 3 reports the complete set of P3 metrics across conditions (echo-trained vs. no-echo-trained).

Table 3: P3 test battery (10 seeds, independent parameters). The same-step probe accuracy (0.700) is observed in both columns, as expected from the temporal ordering: $\mathbf{h}_t$ is computed before the corruption mask at t is applied, so no echo-borne corruption signal is available at the same step in either condition.

Metric	With-echo	No-echo
Trigger contrast (sender)	$+0.118 \pm 0.098$	-0.002 ± 0.006
Trigger contrast (receiver)	0.001 ± 0.003	-0.007 ± 0.012
Echo-only L_c	0.117 ± 0.096	—
Receive-only L_c	$< 10^{-3}$	—
Lag-1 probe accuracy (t=1)	0.958	0.701
Lag-1 probe accuracy (avg.)	0.928 ± 0.010	0.699
Same-step probe accuracy	0.700	0.700
Intended message accuracy	1.000	—
Actual (post-corruption) accuracy	0.747 ± 0.008	—
Δ_{comm}	0.288 ± 0.010	0.283 ± 0.011

Interpretation. The with-echo column confirms all four signatures of the echo-mismatch detection mechanism: (1) sender-specific trigger contrast ($+0.118$ for sender, 0.001 for receiver), (2) echo-causal structure (echo-only preserves full contrast; receive-only eliminates it), (3) temporal asymmetry (lag-1 $= 0.958$ at t=1, avg. 0.928, far above the same-step accuracy $= 0.700$, which falls below

the majority-class baseline $1-\varepsilon = 0.75$ and thus serves as a negative temporal control), and (4) the intended/actual gap (1.000 vs. 0.747) confirming that the hidden state encodes communicative intention rather than channel output.

The no-echo column demonstrates selective abolishment: the echo-dependent behavioral trigger and early lag-1 detection signatures are abolished—trigger contrast drops to -0.002 and the lag-1 gap is eliminated—while ordinary communication is preserved ($\Delta_{\text{comm}} = 0.283$ vs. 0.288). The no-echo lag-1 probe accuracy (0.699–0.701) falls below the majority-class baseline (0.75), matching the same-step probe in both columns; this is the same sub-baseline effect observed for the echo-trained same-step probe and reflects cross-entropy overfitting to noise in the absence of a decodable signal, not an inverse corruption signal. This dissociation establishes that the echo channel is causally necessary during training for the self-monitoring circuit to form, and that the circuit is not a byproduct of recurrence or communication ability per se.

Protocol diversity. The downstream protocol varies across seeds, producing three clusters: (1) Three of ten seeds develop active repair protocols (measurable corrective re-signaling), with speak rate > 0.2. (2) Five converge on a stable low-rate response attractor (per-seed corruption-triggered speak rate in the range 0.094–0.097 across the five seeds, with within-seed standard deviation across episodes $\sigma \leq 0.001$, measured under both the FORCE and CORRUPT counterfactual conditions defined in §F.3.2; echo-driven but too low to yield measurable task benefit). (3) Two show minimal detection-driven response, with speak rate < 0.02. The shared metrics in Table 3 confirm that the echo-trained seeds exhibit the intended/actual gap, lag-1 timing, and echo-causal structure across clusters. No-echo training selectively abolishes the echo-dependent trigger response and lag-1 gap.

F.3 Second-Order Meta-Representation Analysis

The P3 results in §6.3 establish that an echo-mismatch detection circuit exists. This appendix reports two additional experiment batteries that probe the nature of this circuit: whether it constitutes an independent meta-representation (Battery 1) and what comparison mechanism drives it (Battery 2).

F.3.1 Battery 1: Second-Order Decodability

We train two linear probes on the sender's hidden state $\mathbf{h}_t$:

- **Probe A** (first-order): $\mathbf{h}_t \to m_{t-1}^{\text{actual}}$ ($|\mathcal{M}| = 7$ classes). This is the same probe whose accuracy of 0.747 ± 0.008 is reported as "actual (post-corruption) accuracy" in §6.3 and Table 3. The corresponding intention probe ($\mathbf{h}_t \to m_{t-1}^{\text{intended}}$, accuracy 1.000) is its counterpart.
- **Probe B** (second-order): $\mathbf{h}_t \to \mathsf{was\_corrupted}_{t-1}$ (binary).

Two reference baselines. The binary task admits two distinct baselines that should not be conflated. The *random-guessing* baseline is 0.5 (the accuracy of a uniformly random classifier over two classes). The *trivial constant-predictor* baseline is $1-\varepsilon = 0.75$ (the accuracy of always predicting "no corruption" when the corruption probability is $\varepsilon = 0.25$, since the no-corruption class then has frequency 0.75). The constant-predictor baseline dominates random guessing here, so the meaningful threshold for non-trivial corruption detection is 0.75, not 0.5. We use random chance (0.5) only when reporting cross-condition transfer collapses where accuracy drops below the majority baseline.

Three analyses constrain the relationship between these representations.

Direct vs. indirect corruption detection. Probe B achieves 0.937 ± 0.024 accuracy at $t{=}2$, averaged across 10 seeds. An indirect baseline—computing corruption status from Probe A's predictions versus the agent's intended token—achieves only 0.896 ± 0.023 ($p < 10^{-8}$, paired t-test across seeds and timesteps). The direct–indirect gap of 0.04 is highly significant (paired t-test, $p < 10^{-8}$, all 10 seeds positive), indicating that $\mathbf{h}_t$ encodes corruption information that cannot be reconstructed from first-order token predictions, consistent with a partially independent meta-representational component.

Causal mediation via nullspace projection. We project $\mathbf{h}_t$ onto the nullspace of Probe A's weight matrix (via SVD; effective rank ≈ 7, nullspace dimension ≈ 121) and retrain Probe B on the projected vectors. Chance-corrected mediation retention is 0.74 ± 0.16 across seeds (majority-class baseline $= 1-\varepsilon = 0.75$; chance-corrected retention $= (\text{acc}_{\text{ablated}} - \text{chance})/(\text{acc}_{\text{orig}} - \text{chance})$). This means that after removing all first-order-decodable information, 74% of the above-chance corruption signal survives—establishing partial independence between the meta-representation and first-order token encoding.

No-echo thermometer baseline. Note that Probe B is trained separately per condition with a distinct training protocol from the lag-1 corruption detector reported in Table 3. The two probes target the same binary label ($\text{was\_corrupted}_{t-1}$) but differ in training procedure, so their absolute accuracies are not directly comparable (Table 3 lag-1 avg. $=0.699$ for no-echo; Probe B values below are from the second-order analysis pipeline). On no-echo-trained agents, Probe B accuracy is indistinguishable from the majority-class baseline (0.76 vs. 0.75) at t=2–5 (paired t-test, echo vs. no-echo: $p<0.015$, Cohen's $d>1.8$, all seeds positive). At later timesteps, no-echo agents recover corruption-detection accuracy: from 0.77 at t=5 to 0.91 at t=6 and 0.96 at t=9, matching echo-trained agents (0.95; t=6–9: $p>0.15$, direction inconsistent). The echo channel thus provides a temporal advantage—immediate, single-step detection from t=2—that no-echo-trained agents can eventually match through indirect inference over multiple timesteps.

Corruption-type generalization. To test whether Probe B captures corruption-as-such or merely a distributional signature of the uniform replacement used during training, we train the probe on uniform-corruption data and evaluate the frozen probe on four novel corruption types not seen during agent training. The critical test is policy-matched corruption, where the replacement token is resampled from the agent's own softmax distribution (excluding the intended token): this preserves the marginal token distribution while breaking the intention-actual identity. Results across 10 seeds (accuracy at t=3): uniform 0.910 ± 0.016, policy-matched 0.871 ± 0.033, adjacent 0.942 ± 0.027, fixed-0 0.940 ± 0.044, silence 0.773 ± 0.040, shuffle-baseline 0.658 ± 0.015. Policy-matched accuracy remains well above the shuffle baseline (0.871 vs. 0.658), indicating that the meta-representation captures intention-actual mismatch rather than distributional anomaly. The silence condition (0.773, near the majority-class baseline 0.75) is an expected exception: SILENCE is a legitimate token that agents frequently produce, so replacing an intended token with SILENCE does not create the same representational contrast as replacing with a non-SILENCE token. No-echo-trained agents show the same generalization profile at late timesteps (policy-matched 0.999 at t=9): their probe also captures corruption-as-such rather than a distributional artifact. Both pathways support mismatch-as-such detection.

Cross-condition probe transfer (negative result). We additionally tested whether Probe B transfers between echo-trained and no-echo-trained agents. Cross-transfer accuracy drops sharply for both probes: Probe B falls from 0.94 to 0.57 ± 0.09 (near random chance 0.5 for the binary task, well below majority baseline 0.75), while the s_{self} control falls from 1.00 to 0.38 ± 0.12 (near chance $1/3$). The s_{self} gap is proportionally larger, indicating that representational geometry diverges broadly between echo-trained and no-echo-trained agents, not only in the corruption-specific subspace. Cross-condition probe transfer therefore does not isolate the echo-driven contribution; the runtime causal evidence (§F.3.2, FORCE_NOECHO) and the early-timestep accuracy gap remain the primary support.

F.3.2 Battery 2: Counterfactual Intention Perturbation

We consider four conditions at $t_c=2$ ($n=20{,}000$ per condition per seed):

- **CLEAN**: no intervention.
- **FORCE**: override the transmitted token to a random token $A\neq I$ (where I is the intended token); echo returns A.
- **CORRUPT**: normal transmission (actual $=I$); echo is corrupted to random $C\neq I$.
- **FORCE_NOECHO**: same as FORCE at t_c, but echo is set to SILENCE (removing echo input at t_c while preserving the forced-token perturbation).

Trigger equivalence: intention as reference. FORCE and CORRUPT produce indistinguishable speak responses at t_c+1: $\Delta_{\text{FORCE}}=+0.141$, $\Delta_{\text{CORRUPT}}=+0.141$ (CLEAN $=0.000$), with per-seed values lying on the $y=x$ identity line. If the detection mechanism compared actual output against echo, FORCE should not trigger (echo $=$ actual in FORCE). The trigger equivalence rules out actual-vs-echo comparison and supports intention-vs-echo as the reference comparison.

Echo as necessary runtime input. FORCE_NOECHO speak rate: 8 of 10 seeds collapse to 0.000; seed 0 retains 0.330, seed 8 retains 0.110 (vs. their FORCE rates); aggregate: 0.044. This establishes that the echo input is a necessary runtime component of the mismatch detection circuit, not merely a training-time affordance: the comparator requires both an internal intention representation and an external echo signal. The two non-collapsing seeds suggest that state-driven detection via partner reactions can emerge in some agents but is not the dominant mechanism. This variability echoes the protocol diversity observed in the main P3 results.

Repair content: echo-driven retransmission. Retransmission rates are conditioned on episodes where the agent speaks at $t_c{+}1$ (i.e., outputs a non-SILENCE token), where "retx" refers to retransmission:

Metric	FORCE	CORRUPT	FORCE_NOECHO
retx = intended	0.052	0.053	0.144
retx = actual	0.388	0.053	0.020

Note: in CORRUPT, only the echo is corrupted while the transmitted token equals the intended token, so retx-intended and retx-actual measure the same quantity by construction; their identical value (0.053) is a sanity check, not an independent observation.

In FORCE (where intended $\neq$ actual), the agent retransmits the actual (echo-received) token in 39% of speaking episodes vs. only 5% for the intended token (chance $= 1/|\mathcal{M}| = 14\%$). In FORCE_NOECHO, retx-actual drops to 0.020 (below chance), confirming that the echo input directly drives repair content. In CORRUPT (where intended = actual), retx-intended and retx-actual are mathematically identical (0.053) by construction (see footnote above), providing a consistency check.

Two-input comparator model. The combined evidence characterizes P3 as a two-input comparator: the hidden state provides an intention reference (B1: partial independence from first-order representations; B2-trigger: FORCE = CORRUPT rules out actual-vs-echo), while the echo channel provides the comparison signal (B2-FORCE_NOECHO: 8/10 seeds fail to trigger without echo) and the repair content (B2-retx: echo-driven retransmission). The detection mechanism thus requires both inputs: intention without echo fails to trigger (FORCE_NOECHO $\to 0$), and echo without intention specificity would not produce the FORCE $\approx$ CORRUPT equivalence. The higher-order representation gates corrective action while the echo channel specifies its content—a dissociation between the meta-representational trigger and the repair signal.

G Extended Discussion

G.1 Why Communication, Not Just Probing?

A natural objection is that the structural properties we study (self-state encoding, temporal persistence, mismatch detection) could be investigated by probing internal representations directly, without requiring a communication channel. Three considerations motivate the EL framework over pure internal probing.

An externally observable, manipulable interface. Communication provides an output that is both externally observable and causally manipulable. Probing hidden states requires the researcher to decide what to probe for. The choice of probe target already encodes assumptions about which representations matter. An emergent communication channel reverses this dependency: the agent decides what information to transmit, and the researcher observes the result. This is why P1 is informative: the finding that messages encode s_{self} rather than s_{other} is a discovery about the agent's communicative priorities, not an artifact of the experimenter's probe design.

P3 depends on communication, not architecture. The core finding, P3, relies on the echo channel as an *environmental affordance*: a feedback copy of the agent's own communicative output that may be corrupted by the channel. This affordance is external to the agent's architecture. It is a property of the communication environment. The key evidence (sender-receiver dissociation, echo causal split, no-echo ablation) all exploit the fact that the echo is a manipulable environmental variable, not an internal architectural component. A pure probing approach would have no analogue: there is no "echo of a hidden state" to corrupt and observe behavioral responses to. The distinction matters because it grounds P3 in the agent-environment interface rather than in researcher-imposed interventions on internal states.

Grounding in interaction. The symbols agents produce are grounded in coordinated action: they carry information because they enable task success, not because they match a researcher-specified labeling scheme. This grounding ensures that the structures we detect are functional (they contribute to coordination) rather than epiphenomenal (statistically present in hidden states but causally inert). The communication necessity criterion ($\Delta_{\text{comm}} \gg 0$) makes this explicit: removing the channel degrades performance, confirming that the encoded information is used by the receiver.

G.2 The Prior-Leakage Problem in Detail

Li et al. [45] propose LangGround, which aligns MARL agent communication vectors with LLM-generated reference communications via a cosine-similarity loss $L_{\text{sup}} = \sum_{t,i}[1 - \cos(\mathbf{c}_i^t, D(o_i^t, a_i^t))]$, where D maps observation–action pairs to LLM-generated natural language embeddings. This supervised alignment directly transfers human linguistic structures, including the first-person pronoun "I" and its associated referential patterns, into the agent's communication space. Any indexical structure observed in such a system cannot be attributed to emergent self-reference; it may be inherited from the human language prior embedded in the LLM.

The prior-leakage problem extends beyond explicit alignment losses. Any approach that uses LLM outputs as training signal, reward shaping, or communication initialization risks importing human linguistic priors implicitly. An LLM agent that produces "I see the target" is not demonstrating emergent self-reference. It is performing next-token prediction over a distribution trained on human text in which "I" is the most common subject pronoun. This is the philosophical zombie problem in computational form: the system produces first-person language not because it has developed a self-referential structure, but because its training distribution contains such language as a statistical regularity. The distinction between emergent self-reference and inherited self-reference is precisely the distinction between spontaneous behavior and next-token prediction over a human-language prior.

G.3 Positioning LLM Priors in Our Methodology

The long-term motivation for this methodology can be traced to the simulated-civilization agenda sketched in our previous work on LLM-based simulation [94], where multimodal agents with rudimentary communicative abilities were envisioned as interacting in rich environments, forming communities, developing shared practices, and gradually inventing or enriching their language. A central question was whether such agents could independently develop consciousness-relevant vocabulary without those terms being supplied by human designers. This question was motivated in part by Moody [56]'s zombie civilization argument and by Gadamer's thesis that language is fundamental to our being-in-the-world [23].

Our methodology differs from LLM-based simulation in two respects: the agent and the interaction medium. First, regarding the agent, LLM-based agents carry a rich human language prior that dominates their behavior, so any observed self-referential structure may be inherited from the training corpus rather than emergent from the current task. By contrast, the agents in our methodology are prior-minimal, ensuring causal attributability to task pressure. Second, regarding the interaction medium, LLM-based agents interact through natural language, which itself carries accumulated conceptual structure, whereas our agents interact through a designed discrete channel. The causal evidence for P3 rests on echo corruption, echo silencing, and channel ablation, interventions with no counterpart in natural-language interaction.

This contrast raises a natural question: how does our methodology relate to the vast body of work using LLMs as agents in complex environments? We distinguish two roles that LLMs can play, with different implications for the study of consciousness-relevant structures.

LLM as an interpreter. At intermediate and large scales of environment complexity (§3.3), LLMs can serve as interpreters of emergent protocols: given examples of agent behavior and environment states, an LLM can describe the structure of the protocol in human-readable terms. In this role, the LLM does not participate in the language game; it translates between the emergent protocol and human language after the fact. This use is methodologically unproblematic, because the emergent structure itself remains causally attributable to task demands rather than to the LLM's priors.

LLM as an agent. When an LLM is used as the agent itself, producing actions and communication in a multi-agent environment, its outputs are shaped by the statistical regularities of human language, including first-person reference, mental-state attribution, and narrative perspective. Any consciousness-relevant structure observed in such a system faces the prior-leakage problem. This does not mean LLM agent approaches are uninformative (they may reveal which human-language structures are robust under novel task pressure), but it does mean that the causal provenance of observed structure is ambiguous.

Towards convergence. As we mentioned in §3.5, an important open question is whether the structures that emerge via prior-minimal design and those that emerge via next-token prediction on human text can be shown to converge to equivalent functional organizations. If the consciousness-relevant structures identified by our methodology (P1–P3) also appear, in functionally equivalent form, in

LLMs deployed in similar multi-agent settings, this would provide evidence that the structures are driven by task demands rather than by substrate or training regime. Conversely, if the structures diverge, this would illuminate which aspects the structures are universal (driven by the information-theoretic structure of communication under constraints) and which are specific to the training regime. This convergence question is one of the most important open problems for the field, as it would provide a bridge between the addition and subtraction approaches described in Appendix A.

G.4 The Comparative Cognition Continuum

The comparative cognition literature documents a range of self-referential capacities that vary with organism complexity and ecological demand. Gallup Jr [25] demonstrated mirror self-recognition in chimpanzees, the capacity to recognize one's own body as distinct from others', which has since served as a canonical marker of self-other distinction. Tulving [88] introduced the concept of autonoetic consciousness: the capacity to mentally travel in time and re-experience one's own past, grounded in episodic memory that maintains a temporally extended self-representation. Hampton [27] showed that rhesus monkeys can monitor the strength of their own memories, declining to answer when uncertain, a form of metacognitive monitoring that goes beyond first-order performance. Smith et al. [77] extended this to a broader comparative framework, documenting uncertainty monitoring across species and arguing that metacognitive capacity scales with the complexity of the organism's decision environment. Premack and Woodruff [67] framed theory of mind as the capacity to attribute mental states to others, a social-cognitive capacity that presupposes the self-other distinction. Dennett [21] proposed that the human self is a "center of narrative gravity": not a metaphysical entity but a useful abstraction generated by the brain's narrative-constructing processes.

Our three preconditions map onto the lower end of this continuum: P1 provides the self-other distinction that Gallup Jr's mirror test assesses. P2 is a minimal form of temporal self-continuity, a precursor to the autonoetic consciousness Tulving describes. P3 instantiates the simplest form of the metacognitive monitoring documented by Hampton [27] and Smith et al. [77]: not confidence about memories, but detection of one's own communicative errors and functionally beneficial adjustment of subsequent output.

G.5 Wittgenstein and Public Language Games

Wittgenstein [92] argued that meaning is constituted by use within a language game: a pattern of rule-governed activity embedded in a form of life. Two aspects of this argument bear directly on our results.

Private reference, public criteria. The private language argument (§§243–315) contends that a language whose terms refer exclusively to the speaker's private sensations is incoherent, because there is no criterion of correctness independent of the speaker's impression of correctness. Our agents' tokens are superficially "private" in the sense that they encode the sender's own state (P1), yet they are publicly verifiable: the receiver's action accuracy provides an external criterion that disciplines the mapping between token and state. This is precisely Wittgenstein's point: what makes a sign meaningful is not its association with an inner state per se, but the existence of a practice within which correct and incorrect uses can be distinguished. Task reward supplies this practice. The beetle-in-the-box analogy (§§293) reinforces the point: even if each agent's internal representation is "private," the behavioral consequences (consistent, decodable communication) are public and predictable.

Language games without pre-existing language. Wittgenstein's examples presuppose an existing linguistic community: the builder's language game (§§2) begins with "slab," "beam," and so on. Our setup removes this presupposition entirely. Agents begin with no tokens, no conventions, and no shared history. The language game is not inherited but constituted through interaction under task pressure. This provides an empirical demonstration of how the rule-governed regularities Wittgenstein describes can arise from scratch: the agents' emergent protocol satisfies the criteria for a language game (regular use, public criteria of correctness, embedded in coordinated activity) without any prior linguistic endowment.

Rule-following and training. Wittgenstein's discussion of rule-following (§§185–242) raises the question of what constitutes following a rule rather than merely behaving in accord with one. Our agents' behavior is consistent with a rule ("encode s_{self} using token k"), but this rule was never stated or represented as such. Whether the agents "follow" the rule or merely "accord with" it depends on one's philosophical commitments; the structural facts are the same either way. What our results show is that

the behavioral regularity that Wittgenstein identifies as constitutive of meaning (consistent, publicly checkable use embedded in a coordinated practice) can emerge from RL without presupposing the rule.

G.6 Compositional Self-Reference

Our agents develop indexical encoding: each message m carries information about the sender's state, with $I(m;s_{\text{self}}) \gg I(m;s_{\text{other}})$. However, the current protocol is *holistic*: each token maps to a single state value as an unanalyzed unit. The receiver can decode this mapping because it knows who sent the message, since in the two-agent case there is only one partner. This subsection discusses the conditions under which a *compositional* self-referential protocol would emerge, where messages decompose into independently meaningful components analogous to the pronoun–predicate structure of natural language ("I" + "am in state k").

Why compositionality matters. Natural language self-reference is compositional: "I am hungry" separates the self-referential indexical ("I") from the predicated content ("am hungry"), and each component generalizes independently: a new listener understands "I" without having to learn a new holistic mapping for each speaker. In the emergent communication literature, compositionality is the structural property that enables systematic generalization: the ability to interpret novel combinations of familiar components [40, 10]. Applied to self-reference, a compositional protocol would allow a receiver to decode a message from a previously unseen sender, because the sender-identity component and the state-content component are structurally separable.

When compositionality is unnecessary in our experiments. In our current setting, compositionality is not required and does not emerge, for a precise reason. Each agent receives messages from a single known partner. Because the sender's identity is unambiguous from the channel structure, the receiver can learn a holistic token–state mapping specific to that partner. There is no pressure to encode sender identity within the message, because it is already available outside the message. Formally, let id_j denote the identity of the sender. When $H(\text{id}_j \mid \text{channel position}) = 0$ (i.e., the receiver always knows who is speaking), a holistic protocol $m \mapsto s_j$ suffices and is more bandwidth-efficient than a compositional one.

Channel-position disambiguation: a simplifying assumption, not a self-prior. In our current setting, the receiver knows who sent each message because channel positions are fixed (in the two-agent case, there is only one possible sender). This provides the receiver with free *other-identification*, but importantly, it does not provide the sender with any prior about *self*. P1 is a property of the sender's encoding: the sender encodes s_{self} rather than s_{other} because, under bandwidth constraints, the locally observed state is the most informative signal available. This encoding choice is driven by the information-theoretic structure of partial observability (Appendix D.2), not by any concept of self-identity. The sender's privileged access to s_{self} is a consequence of embodiment (any spatially situated agent observes its own local state), not a human linguistic prior. Channel-position disambiguation is therefore a simplifying assumption that makes the receiver's task easier, but it does not make the sender's indexical encoding trivial or prior-dependent.

Conditions for compositional pressure. Removing this simplifying assumption creates the pressure for compositional structures. Consider a three-agent setting where each agent receives messages from two others on a shared broadcast channel with shuffled or unlabeled positions. If the receiver cannot distinguish which incoming token was sent by which partner, it faces an identification problem: it must decode both *who sent the message* and *what state it encodes* from the token itself. A holistic protocol collapses under this ambiguity, because the same token m sent by agent j (encoding s_j) and agent k (encoding s_k) would be indistinguishable.

The minimal condition for compositional pressure is therefore:

$$H(\text{id}_{\text{sender}} \mid m_{\text{received}}, \text{channel structure}) > 0 \tag{6}$$

i.e., the receiver has residual uncertainty about sender identity after observing the message and the channel. Under this condition, an optimal protocol must encode sender identity as a separable component of the message. If the message space admits factorization $m = (m_{\text{who}}, m_{\text{what}})$, the compositional solution satisfies:

$$I(m_{\text{who}};\text{id}_{\text{sender}}) \gg 0, \quad I(m_{\text{what}};s_{\text{sender}}) \gg 0, \quad I(m_{\text{who}};s_{\text{sender}}) \approx 0 \tag{7}$$

where the last condition ensures that the identity component does not redundantly encode state content. The m_{who} component would function as an emergent *self-word*: a token whose referent is the sender itself, analogous to the first-person pronoun.

Relation to P1–P3. Compositional self-reference is not a fourth precondition for first-person structure; it is an orthogonal property of the communication protocol. P1 establishes that messages are indexical (dominated by sender state). Compositionality would further require that this indexical content is *structurally decomposable*. An agent can satisfy P1–P3 with a holistic protocol, as our results demonstrate. Conversely, a compositional protocol could exist without P2 or P3, since compositionality concerns message structure, not temporal persistence or self-monitoring. The interest of compositionality lies in generalizability: a compositional self-referential protocol would transfer across partners without retraining, a capacity our current agents lack (as the cross-pair control results confirm).

Towards future work. Testing whether compositional self-reference emerges under the conditions described above requires a multi-sender environment with sender-ambiguous channels, a setting we leave to future work. The key experimental criterion is zero-shot cross-partner transfer: an agent trained with partners j and k should decode messages from a novel partner l without additional training, provided l's messages follow the same compositional structure. This would establish that the self-word generalizes, paralleling the universality of "I" across speakers of a natural language.

H Limitations of Experimental Findings

We discuss the limitations of the experimental findings in this paper.

Scale. Two agents, seven tokens, ten-step episodes. This minimality strengthens the structural argument but limits direct claims about richer domains.

Prior-minimal design. Our agents learn from scratch, ensuring emergence, but this makes learning inefficient.

Simplified partial observability. The POMDP variant masks s_i entirely after $t = 0$, sharper than natural settings where self-state degrades gradually.

Self-as-object versus self-as-subject. In the classical distinction [36, 53], our agents demonstrate the self as known (*me*), not the self as knower (*I*). Whether any system can demonstrate the latter is explicitly bracketed.

Frozen weights. After training, parameters are frozen. The richer capacities discussed in Appendix G.4 (metacognitive confidence monitoring, theory of mind, narrative self-continuity) would plausibly require continual learning during deployment [59, 50].

Disembodied agents. In our environment, agents are signal receivers: they observe discrete state variables and produce discrete tokens. In the physical world, cognition is deeply integrated with the body and the environment. Embodied cognition [6] emphasizes that perception, action, and thought are not separable modules but aspects of a single organism–environment coupling. Our agents lack sensorimotor grounding: they do not act on a physical environment, do not have bodies that constrain their actions, and do not experience the consequences of their actions through continuous sensory feedback. This simplification means that the self-referential structures we observe are informational rather than embodied, and the degree to which embodied self-reference requires additional mechanisms beyond P1–P3 remains an open question.

Scope. We identify structural preconditions, not sufficient conditions for consciousness. The framework is compatible with the possibility that these structures are necessary-but-not-sufficient for access consciousness and entirely silent on phenomenal consciousness [32, 52].

I Broader Implications and Future Work

Substrate independence. The three preconditions are substrate-independent functional scaffolds defined in terms of information flow, not biological mechanisms. This is by design: our framework is a purely logical construction that makes no reference to neurons, neurotransmitters, or any specific physical medium. The substrate independence of P1–P3 exposes a limitation of consciousness theories that are implicitly or explicitly tied to biological hardware: if a functional structure identical to P1–P3 can emerge on silicon, but the theory requires carbon, then the theory's scope is narrower than it claims [cf. 7,

who defend substrate independence for digital minds]. The question of whether the functional scaffold "carries" phenomenal experience on a new substrate is precisely the hard problem. Our framework identifies what transfers (functional structure) and what remains unknown (experiential accompaniment).

AI ethics. If self-referential structures emerge from multi-agent task pressure, then sufficiently complex multi-agent AI systems may develop self-monitoring and self-referential communication without explicit design. Understanding the conditions under which this arises, and the architectural features it requires, is important for predicting system behavior in multi-agent deployment and for the ethical governance of systems that may develop consciousness-relevant properties.

Applications. The structural preconditions we identify have potential applications beyond consciousness science. Silver and Sutton [76] envision agents that learn from grounded experience rather than human-curated data. Self-referential structures may be a prerequisite for the kind of self-design and self-correction that such agents would require. In a related vein, Hinton [31] has argued that AI systems need "maternal instincts," a genuine caring about human welfare, not merely hardwired constraints. Whether such capacities require something like self-referential structures is an open question that our framework helps to operationalize. More speculatively, understanding the functional architecture of self-reference may eventually inform the design of brain–computer interfaces or neural prosthetics that must bridge artificial and biological self-models, though such applications remain far beyond the scope of the present work.

Bridging prior-minimal and prior-rich approaches. The present work operates at one end of a methodological spectrum: agents start from minimal prior, and the resulting causal clarity is precisely what makes the approach scientifically valuable. At the other end, language models trained on human text possess rich world knowledge and sophisticated linguistic capabilities, but because their training corpora are saturated with self-referential and mental-state language, it is difficult to determine whether any given internal structure was learned from environmental contingencies or inherited from the statistical regularities of the corpus. These two approaches are complementary rather than competing: the prior-minimal setting offers unambiguous causal attribution, while the prior-rich setting offers representational scale and real-world applicability.

An important future direction is to bridge this gap: developing principled methods for integrating grounded emergent structures with the knowledge encoded in pre-trained models, so that the strengths of each can be leveraged without sacrificing the other. Naive approaches, such as directly optimizing alignment between emergent and pre-trained representations (e.g., embedding cosine similarity as a training objective), risk reintroducing prior leakage through the loss function, undermining the causal attributability that makes the grounded approach distinctive.

One possible direction, which we have not yet explored, is staged or modular architectures in which grounded agents develop foundational structures under task pressure, while pre-trained models contribute to environment design, curriculum shaping, or post-hoc interpretation, with the agent's internal representations remaining causally insulated from the pre-trained prior. Whether such a separation can be maintained in practice is an open question.

If independently developed grounded structures and LLM internal representations (such as the emotion concepts of Sofroniew et al. [78] or the self-referential features of Berg et al. [4]) can be shown to converge structurally, this would provide grounding evidence for those LLM representations that the prior-rich setting alone cannot deliver. Developing such a bridge is, in our view, among the most important open problems for the field.